数学为机器学习

Marc Peter Deisenroth
A.Aldo Faisal
Cheng Soon Ong

内容

前言

机器学习是将人类知识和推理提炼成适合构建机器和工程自动化系统的形式的最新尝试。随着机器学习变得越来越普遍，它的软件包也越来越容易使用，自然而然地，低层次的技术细节就会被抽象出来，不被从业者发现，这是可取的。然而，这也带来了一个危险，那就是从业者变得不知道设计决策，从而不知道机器学习算法的极限。

有兴趣了解成功的机器学习算法背后的魔力的热心从业者，目前面临着一套令人生畏的前提知识。

- 编程语言和数据分析工具
- 大规模计算和相关框架 数学和统计学以及机器学习如何建立在它
- 的基础上

在大学里，机器学习的入门课程往往会在课程的早期部分涵盖这些先决条件。由于历史原因，机器学习的课程往往是在计算机科学系教授的，那里的学生往往在前两个领域的知识中得到训练，但在数学和统计学方面却没有那么多。

目前的机器学习教科书主要关注机器学习的算法和方法，并假定读者对数学和统计学有一定的了解。因此，这些书只用了一到两章来介绍背景数学，要么在书的开头，要么作为附录。我们发现很多想深入研究基本机器学习方法基础的人，在阅读机器学习教科书时对所需的数学知识感到困惑。在大学教授本科和研究生课程后，我们发现高中数学与阅读标准机器学习教科书所需的数学水平之间的差距对许多人来说太大。

本书将基本机器学习概念的数学基础凸显出来，并将信息收集在一个地方，以便缩小甚至消除这种技能差距。

为什么又是一本关于机器学习的书？

机器学习建立在数学语言的基础上，以表达那些在直觉上似乎很明显，但却出奇地难以形式化的概念。一旦形式化得当，我们就可以深入了解我们想要解决的任务。全球数学专业学生的一个共同抱怨是，所涉及的主题似乎与实际问题关系不大。我们认为，机器学习是人们学习数学的一个明显而直接的动机。

"在大众心目中，数学与恐惧症和焦虑症相联系。你会认为我们在讨论蜘蛛"。(斯特罗格茨，第2014,281页)

本书旨在成为构成现代机器学习基础的大量数学文献的指导手册。我们通过直接指出数学概念在基本机器学习问题中的作用，来激发人们对数学概念的需求。为了保持本书的简短，许多细节和更高级的概念都被遗漏了。掌握了这里介绍的基本概念，以及它们如何融入机器学习的大背景，读者可以找到许多进一步学习的资源，我们在各章的结尾处提供了这些资源。对于有数学基础的读者来说，本书提供了机器学习的简要但精确的一瞥。与其他专注于机器学习方法和模型的书籍相比（MacKay,2003；Bishop,2006；Alpaydin,2010；Bar- ber,2012；Murphy,2012；Shalev-Shwartz and Ben-David,2014；Rogers和Girolami，2016）或机器学习的编程方面（M üller和Guido，2016；Raschka和Mirjalili，2017；Chollet和Allaire，2018），我们只提供机器学习算法的四个代表性例子。相反，我们专注于模型本身背后的数学概念。我们希望读者能够更深入地了解机器学习的基本问题，并将使用机器学习产生的实际问题与数学模型中的基本选择联系起来。

我们的目的不是要写一本经典的机器学习书。相反，我们的目的是提供数学背景，应用于四个典型的机器学习问题，以使其更容易阅读其他机器学习教科书。

谁是目标受众？

随着机器学习在社会上的应用越来越广泛，我们认为每个人都应该对其基本原理有一些了解。本书是以学术性的数学风格来写的，这使我们能够精确地描述机器学习背后的概念。我们鼓励不熟悉这种看似简洁的风格的读者坚持下去，并牢记每个主题的目标。我们把评论和意见洒在整个文本中，希望它能在大局方面提供有用的指导。

本书假设读者具有通常的数学知识

在高中数学和物理学中都有涉及。例如，读者以前应该见过导数和积分，以及二维或三维的几何向量。从这里开始，我们概括了这些概念。因此，本书的目标读者包括大学本科生、夜间学习者和参加在线机器学习课程的学习者。

与音乐类比，人们与机器学习的互动有三种类型。

精明的听众 通过开源软件、在线教程和基于云的工具的支持，机器学习的民主化使用户不必担心管道的具体细节。用户可以专注于使用现成的工具从数据中提取洞察力。这使不懂技术的领域专家也能从机器学习中受益。这类似于听音乐；用户能够选择和辨别不同类型的机器学习，并从中受益。更有经验的用户就像乐评人一样，对机器学习在社会中的应用提出重要的问题，如道德、公平和个人的特权。我们希望这本书为思考机器学习系统的认证和风险管理提供基础，并让他们利用自己的领域专长来建立更好的机器学习系统。

经验丰富的艺术家 熟练的机器学习从业者可以将不同的工具和库插入和发挥到分析管道中。最典型的从业者是数据科学家或工程师，他们了解机器学习界面及其使用案例，并能够从数据中进行精彩的预测。这类似于一个演奏家在演奏音乐，技术高超的从业者可以将现有的乐器带入生活，给观众带来享受。利用这里介绍的数学知识作为入门，从业者将能够了解他们最喜欢的方法的好处和限制，并扩展和概括现有的机器学习算法。我们希望这本书能够为机器学习方法的更严格和更有原则的发展提供推动力。

随着机器学习被应用到新的领域，机器学习的开发者需要开发新的方法并扩展现有的算法。他们通常是研究人员，需要了解机器学习的数学基础，发现不同任务之间的关系。这类似于音乐的作曲家，他们在音乐理论的规则和结构中，创造出新的、令人惊叹的作品。我们希望这本书能为那些想成为机器学习作曲家的人提供其他技术书籍的高水平概述。社会上非常需要新的研究人员，他们能够提出和探索新的方法来攻克从数据中学习的诸多挑战。

鸣谢

我们感谢许多人，他们看了本书的早期草稿，并忍受了痛苦的概念论述。我们试图实现他们的想法，我们并不强烈反对。我们要特别感谢Christfried Webers，感谢他仔细阅读了本书的许多部分，以及他对结构和表述的详细建议。许多朋友和同事也非常友好，为每一章的不同版本提供了他们的时间和精力。我们很幸运地受益于网上社区的慷慨解囊，他们通过https://github.com，提出了改进建议，大大改善了本书的内容。

以下人士通过https://github.com 或个人交流，发现了错误，提出了澄清，并建议了相关文献。他们的名字是按字母顺序排列的。

Abdul-Ganiy Usman
Adam Gaier
阿黛尔-杰克逊
Aditya Menon 阿拉
斯戴尔-特兰
Aleksandar Krnjaic
亚历山大-马克里乔戈斯
阿尔弗雷多-坎齐亚尼
Ali Shafti
Amr Khalifa
Andrew Tanggara
Angus Gruen
Antal A. Buss
Antoine Toisoul Le Cann
Areg Sarvazyan
Artem Artemev
Artyom Stepanov
Bill Kromydas
Bob Williamson
Boon Ping Lim
Chao Qu
李成 克里斯-
夏洛克
Christopher Gray
Daniel McNamara
Daniel Wood
Darren Siegel
David Johnston
Dawei Chen
Ellen Broad
Fengkuangtian Zhu
Fiona Condon
Georgios Theodorou
He Xin
Irene Raissa Kameni
Jakub Nabaglo
James Hensman
Jamie Liu
Jean Kaddour
Jean-Paul Ebejer
Jerry Qiang
Jitesh Sindhare
John Lloyd
Jonas Ngnawe
Jon Martin
Justin Hsi
Kai Arulkumaran
Kamil Dreczkowski
Lily Wang
Lionel Tondji Ngoupeyou
Lydia Knifing
马哈茂德-阿斯兰
Mark Hartenstein
Mark van der Wilk
Markus Hegland
Martin Hewing
Matthew Alger
Matthew Lee

马克西姆斯-麦肯

张梦艳 迈克尔-贝

内特 迈克尔-佩德

森 申敏静

Mohammad Malekzadeh
Naveen Kumar
Nico Montali
Oscar Armas
Patrick Henriksen
Patrick Wieschollek
Pattarawat Chormai
Paul Kelly
Petros Christodoulou
Piotr Januszewski
Pranav Subramani
Quyu Kong

拉吉布-扎曼

张锐

Ryan-Rhys Griffiths
Salomon Kabongo
Samuel Ogunmola
Sandeep Mavadia
Sarvesh Nikumbh
Sebastian Raschka

塞纳纳亚克-塞什-库马尔-卡

尔里-白承宪

Shahbaz Chaudhary
Shakir Mohamed
Shawn Berry

谢赫-阿卜杜勒-拉希姆-阿里-盛

雪

Sridhar Thiagarajan
Syed Nouman Hasany
Szymon Brych
Thomas Kehler
Timur Sharapov
Tom Melamed
Vincent Adam
Vincent Dutordoir
Vu Minh

瓦西姆-阿夫

塔布-温智

Wojciech Stokowiec

庄晓楠 张亚伟 郝义

成 罗亚舟

杨李玉露

程云 黄玉晓

Zac Cranko

曹子健 Zoe

Nolan

通过GitHub的贡献者，他们的真实姓名没有在GitHub个人资料中列出，他们是：。

SamDataMad	insad	胜利者和
bumptiousmonkey	HorizonP	17SKYE
idoamihai	cs-maillist	jessjing1995
deepakiim	kudo23	

我们也非常感谢Parameswaran Raman和许多由剑桥大学出版社组织的匿名审稿人，他们阅读了手稿早期版本中的一个或多个章节，并提出了结构性的批评意见，从而使手稿得到了很大的改进。特别要感谢我们的LATEX支持者Dinesh Singh Negi，他对LATEX相关问题提供了详细而及时的建议。 最后，我们非常感谢我们的编辑Lauren Cowles，她在本书的酝酿过程中一直耐心地指导我们。

符号表

	符号典型含义
$a, b, c, \alpha, \beta, \gamma$	标点符号为小写。
$\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$	矢量为小写粗体
$\boldsymbol{A}, \boldsymbol{B}$, CMatrices为粗体大写字母	
$\boldsymbol{x}^\top$, ⊤	一个向量或矩阵的ATranspose
矩阵的$^{-1}$	倒数
$(\boldsymbol{x}, \boldsymbol{y}$$\boldsymbol{x}$和$\boldsymbol{y}$的内积	
$^\top$xyDot是$\boldsymbol{x}$和$\boldsymbol{y}$的乘积	
$\boldsymbol{\beta} = (\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3)$	(有序的)元组
$\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}]$	$_3$水平叠加的列向量矩阵
B = b, $\boldsymbol{b}_{12}, \boldsymbol{b}_3$	一组向量（无序）。
Z、N $\{\}$ntegers和自然数，分别为	
R，CReal和复数，分别为	
^{n}	$\boldsymbol{R}n$维实数向量空间
$\forall x$	通用量词：对所有x
$\exists x$	存在性量词：存在x
$a :=$	ba被定义为 b
$a =:$	bb被定义为a
$\propto$	aba与b成正比，即a=常数b
gf Function组成。"g在f之后" 当且仅当	
$\Leftrightarrow \Rightarrow$	适用
$\vec{A}$,	套
$C \in$	aa是集合的一个元素
$\emptyset$	空集合
A $A \setminus B$ $A \cap B$ $A \cup B$	不包括：元素 的 集合， 但不包括在
	DN维度数；用$d = 1, \ldots, D$
	NN数据点的数量；用$n = 1, \ldots, N$
m	二、尺寸为 × *毫米*矩阵
$\mathbf{0}_{m,n}$	大小为 × mn的零点矩阵
$\mathbf{1}_{m,n}$	大小为 × mn的1的矩阵
eStandard$_i$	/canonical向量（其中$\boldsymbol{i}$是分量，是 1）。
向量空间的维度dim	
rk($\boldsymbol{A}$)矩阵$\boldsymbol{A}$的等级	
线性映射Φ的Im(Φ)	图像
ker(Φ)	线性映射Φ span[$\boldsymbol{b}_1$] 的内核(空空间)$\boldsymbol{b}$
的跨度(生成集)$_1$	
tr($\boldsymbol{A}$)$\boldsymbol{A}$的轨迹	

$\det(\boldsymbol{A})$	$\boldsymbol{A}$的决定式
$\lvert\text{-}\rvert$	绝对值或决定性的（取决于上下文）
$\lVert\text{-}\rVert$	规范；欧几里得，除非指定
λ	特征值或拉格朗日乘数
E_λ	与特征值λ对应的Eigenspace

	符号典型含义
$\perp$ $\boldsymbol{xy}$	方程$\boldsymbol{x}$和$\boldsymbol{y}$是正交的。
VVector空间	
$V^{\perp}$	Orthogonal complement of vector space V
$\sum_{n=1}^{N} x_n$	x的总和$_n$：x_1 $+ \ldots +$ x_N
$\prod_{n=1}^{N} xn$	x的乘积$_n$ $: x_1 \cdot \ldots \cdot x_N$
$\boldsymbol{\theta}$	参数向量
$\frac{\partial f}{\partial x}$	f对x的部分导数
$\frac{df}{dx}$	f对x的总导数
∇_x	梯度
$f_* = \min_x f(x)$	f的最小函数值
$x_* \in \arg\min_x f(x)$	最小化f的值x_*（注意：arg min返回一组值）。
LLagrangian	
$\binom{n}{k}$	负对数可能性二项式系数，n选择k
$\mathbb{V}_X[\boldsymbol{x}]$	$\boldsymbol{x}$相对于随机变量X的方差
$\mathbb{E}_X[\boldsymbol{x}]$	$\boldsymbol{x}$相对于随机变量X的期望值
$\mathrm{Cov}_{X,Y}[\boldsymbol{x}, \boldsymbol{y}]$	$\boldsymbol{x}$和$\boldsymbol{y}$之间的协方差。
$X \perp Y \mid Z$	X是有条件地独立于Y的，给定Z
$X \sim p$	随机变量X的分布是根据p
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	高斯分布，均值为$\boldsymbol{\mu}$，协方差为$\boldsymbol{\Sigma}$
Ber(μ)参数为μ的伯努利分布	Bin(N, μ) 参数为N, μ的二项分布
Beta(α, β)	参数为α, β的β分布

缩写和首字母缩略词表

	首字母缩写含义
	例如 Exempligratia（拉丁文：例如）。
GMM	高斯混合模型
	i.e.Idest（拉丁文：这意味着）。
	i.i.d.独立、相同的分布
	MAPMaximuma posteriori
	MLEMaximumlikelihood
estimation/estimator	ONBOrthonormalbasis
	PCAP主要成分分析
	PPCAProbabilisticprincipal component
analysis	REFRow-echelonform
	SPDS对称、正定

SVMS支持向量机

第一部分

数学基础

1

简介和动机

机器学习是关于设计算法，从数据中自动提取有价值的信息。这里强调的是 "自动"，也就是说，机器学习关注的是可以应用于许多数据集的通用方法，同时产生的东西是有意义的。有三个概念是机器学习的核心：数据、模型和学习。

由于机器学习在本质上是由数据驱动的，所以*数据*是机器学习的核心 数据。机器学习的目标是设计通用的方法，从数据中提取有价值的模式，最好是不需要很多特定领域的专业知识。例如，给定一个大型语料库的文件（例如许多图书馆中的书籍），机器学习方法可以用来自动寻找在文件中共享的相关主题（Hoffman等人，2010）。为了实现这一目标，我们设计了*模*

*其他*通常与产生数据的过程有关，类似于模型。
我们得到的数据集。例如，在回归环境中，模型将描述一个将输入映射到实值输出的函数。用Mitchell(1997)的话来说。如果一个模型在考虑了数据之后，其在特定任务上的表现有所改善，那么这个模型就可以说是从数据中学习。我们的目标是找到对未见过的数据有良好概括性的好模型。
我们在未来可能会关心的问题。*学习*可以理解为一种 通过优化模型的参数来自动寻找数据的模式和结构的学习方式。

虽然机器学习已经有了很多成功的案例，软件也很容易设计和训练丰富而灵活的机器学习系统，但我们认为机器学习的数学基础是很重要的，这样才能理解更复杂的机器学习系统所依据的基本原理。理解这些原

理可以促进创建新的机器学习解决方案，理解和调试现有的方法，并了解我们正在使用的方法的内在假设和限制。

1.1 为直觉找话说

我们在机器学习中经常面临的一个挑战是，概念和词语是很滑稽的，机器学习系统的一个特定组成部分可以被抽象为不同的数学概念。例如，"算法 "这个词在机器学习的概念中至少有两种不同的含义。在第一种意义上，我们用 "机器学习算法 "指的是一个基于输入数据进行预测的系统。我们把这些算法称为*预测器*。在第二种意义上，我们用完全相同的短语 "机器学习算法 "指的是一个系统，它调整了预测器的一些内部参数，使其在未来未见的输入数据上表现良好。这里我们把这种适应性称为*训练*系统。

预测者

培训

本书不会解决模棱两可的问题，但我们想预先说明，根据上下文，同样的表达方式可以有不同的含义。然而，我们试图使上下文足够清晰，以减少歧义的程度。

本书的第一部分介绍了谈论机器学习系统的三个主要组成部分所需的数学概念和基础：数据、模型和学习。我们将在此简要概述这些组成部分，一旦我们讨论了必要的数学概念，我们将在第八章中再次重温这些组成部分。

虽然不是所有的数据都是数字的，但考虑数字格式的数据往往是有用的。在本书中，我们假设*数据*已经被适当地转换为适合读入计算机程序的数字表示。因此，我们将数据视为向量。为了说明词语的微妙性，有（至少）三种不同的方式来思考向量：向量是一个数字阵列（计算机科学的观点），向量是一个有方向和大小的箭头（物理学的观点），向量是一个服从加法和缩放的物体（数学的观点）。

作为向量的数据

模型

一个*模型*通常用于描述一个生成数据的过程，与手头的数据集相似。因此，好的模型也可以被认为是真实（未知）数据生成过程的简化版本，捕捉到与数据建模相关的方面，并从中提取出隐藏的模式。然后，一个好的模型可以用来预测现实世界中会发生什么，而不需要进行真实世界的体验。

学习

现在我们来看看问题的关键，即机器学习的*学习*部分。假设我们得到了一个数据集和一个合适的模型。*训练*模型意味着使用现有的数据来优化模型的一些参数，这些参数与评估模型对训练数据的预测程度的效用

喻中，山顶对应的是某些参数的最大值。

函数有关。大多数训练方法可以被认为是一种类似于爬山的方法，以达到其峰值。在这个比

所期望的性能指标。然而，在实践中，我们感兴趣的是模型在未见过的数据上表现良好。在我们已经看过的数据（训练数据）上表现良好，可能只意味着我们找到了一个记忆数据的好方法。然而，这可能不能很好地推广到未见过的数据，在实际应用中，我们经常需要将我们的机器学习系统暴露在它以前没有遇到过的情况下。

让我们总结一下我们在本书中涉及的机器学习的主要概念。

- 我们用向量来表示数据。
- 我们选择一个适当的模型，要么使用概率论的观点，要么使用操作化的观点。
- 我们通过使用数值优化方法从现有数据中学习，目的是使模型在未用于训练的数据上表现良好。

1.2 阅读此书的两种方式

我们可以考虑两种策略来理解机器学习的数学。

- **自下而上。**从基础概念到更高级的概念。这通常是技术性较强的领域的首选方法，如数学。这种策略的好处是，读者在任何时候都能依靠他们以前学过的概念。不幸的是，对于从业人员来说，许多基础概念本身并不特别有趣，而且缺乏动力意味着大多数基础定义很快就会被遗忘。
- **自上而下。**从实际需要到更多的基本要求，逐层深入。这种以目标为导向的方法的优点是，读者在任何时候都知道他们为什么需要学习某个特定的概念，而且有一条清晰的所需知识路径。这种策略的缺点是，知识建立在可能不稳定的基础上，读者必须记住一套他们没有办法理解的单词。

我们决定以模块化的方式写这本书，把基础（数学）概念和应用分开，这样这本书就可以用两种方式阅读。本书分为两部分，第一部分奠定了数学基础，第二部分将第一部分的概念应用于一组基本的机器学习问题，如图1.1所示，这些问题构成了机器学习的四大支柱：回归、降维、密度估计和分类。第一部分的各章大多建立在前面的基础上，但如果有必

要，也可以跳过某一章，向后学习。第二部分的各章只是松散地联系在一起，可以按任何顺序阅读。有许多向前和向后的指向性

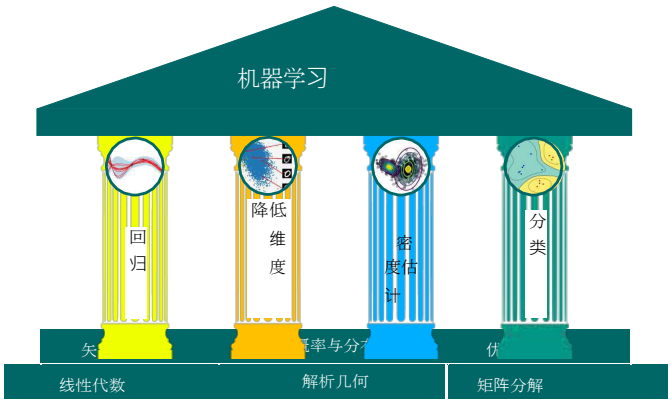


图 机器学习的**1.1** 基础和四大支柱。

这本书的两部分之间，将数学概念与机器学习算法联系起来。

*当然，阅读这本书的方法不止两种。*大多数读者采用自上而下和自下而上相结合的方式学习，有时在尝试更复杂的概念之前，先积累基本的数学技能，但也会根据机器学习的应用来选择课题。

第一部分是关于数学

我们在本书中所涉及的机器学习的四个支柱（见图1.1）需要一个坚实的数学基础，这在第一部分中有所阐述。

线性代数

我们用向量来表示数字数据，用矩阵来表示这些数据的表格。对向量和矩阵的研究被称为*线性代数*，我们在第二章中介绍。那里也描述了作为矩阵的向量的集合。

解析几何学

给定两个代表现实世界中两个物体的向量，我们想对它们的相似性做出说明。我们的想法是，相似的向量应该被我们的机器学习算法（我们的预测器）预测为有相似的输出。为了正式确定向量之间的相似性概念，我们需要引入一些操作，将两个向量作为输入，并返回一个代表其相似性的数值。相似性和距离的构造是*解析几何*的核心，将在第三章讨论。

基体

在第四章中，我们介绍了一些关于矩阵和*矩阵分解的*基本概念。矩阵上的一些操作是非常

分解

在机器学习中很有用，它们允许对数据进行直观的解释和更有效的学习。

我们经常认为数据是对一些真正的底层信号的嘈杂观察。我们希望，通过应用机器学习，我们可以从噪声中识别出信号。这就要求我们有一

*1.2阅读此书的两种方式*种语言来量化 "噪音 "的含义。我们通常也希望能有预测器，以便

允许我们表达某种不确定性，例如，量化我们对某一特定测试数据的预测值的置信度。

点。对不确定性的量化是*概率论*和 概率论的范畴。
在第6章中涉及。

为了训练机器学习模型，我们通常要找到能使某些性能指标最大化的参数。许多优化技术都需要一个梯度的概念，它告诉我们应该朝哪个方向前进。

寻找解决方案。第五章是关于 *向量微积分的*，详细介绍了向量微积分的内容。
梯度的概念，我们随后在第七章中使用了这一概念。

谈论*优化*，寻找函数 的最大值/最小值。
优化

第二部分是关于机器学习

本书的第二部分介绍了*机器学习的四个支柱*，如图所示。 1.1.我们说明了本书第一部分所介绍的数学概念是如何成为每个支柱的基础。大体上说，各章是按难度排序的（按升序排列）。

在第八章中，我们以数学的方式重申了机器学习的三个组成部分（数据、模型和参数估计）。此外，我们还提供了一些建立实验装置的指南，以防止对机器学习系统进行过于乐观的评估。回顾一下，我们的目标是建立一个在未见过的数据上表现良好的预测器。

在第九章中，我们将仔细研究*线性回归*，其中我们的 线性回归
目标是找到将输入$\boldsymbol{x} \in \mathbb{R}^D$映射到相应的函数值$\boldsymbol{y} \in \mathbb{R}$的函数，我们可以将其解释为各自输入的标签。我们将讨论通过最大似然和最大后验估计的经典模型拟合（参数确定），以及贝叶斯线性回归，其中我们把参数整合出来，而不是优化它们。

第10章重点讨论*降维*问题，这是图中的第二个支柱。
减少
1.1,使用主成分分析法对数据进行降维。降维的主要目的是为高维数据$\boldsymbol{x} \in \mathbb{R}^D$找到一个紧凑的、低维的表示$\boldsymbol{x} \in$，这通常比原始数据更容易分析。与回归不同，降维只关注数据的建模--没有与数据点$\boldsymbol{x}$相关的标签。

在第11章，我们将进入第三个支柱：*密度估计*。密度估计
密度估计的目的是找到一个描述给定数据集的概率分布。为此，我们将关注高斯混合模型，并将讨论一个迭代方案来寻找这个模型的参数。与

降维 *1.2 阅读指南* 没有为数据点 $\boldsymbol{x}$ 提供有用的相关的标签 D。然而，我们并不寻求数据的低维表示。相反，我们感兴趣的是一个描述数据的密度模型。

第12章是本书的结尾，深入讨论了第四章的内容。

分类

支柱：*分类*。我们将在支持向量机的背景下讨论分类。与回归（第九章）类似，我们有输入$\boldsymbol{x}$和相应的标签y。然而，与回归不同的是，标签是实值的，分类中的标签是整数，这需要特别注意。

1.3 练习和反馈

我们在第一部分提供了一些练习，这些练习大部分可以通过纸笔完成。对于第二部分，我们提供了编程教程（jupyter笔记本）来探索我们在本书中讨论的机器学习算法的一些特性。

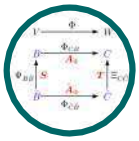
我们感谢剑桥大学出版社大力支持我们实现教育和学习民主化的目标，将此书免费提供给大家下载。

https://mml-book.com

在这里可以找到教程、勘误表和其他材料。可以使用前面的URL报告错误和提供反馈。

2

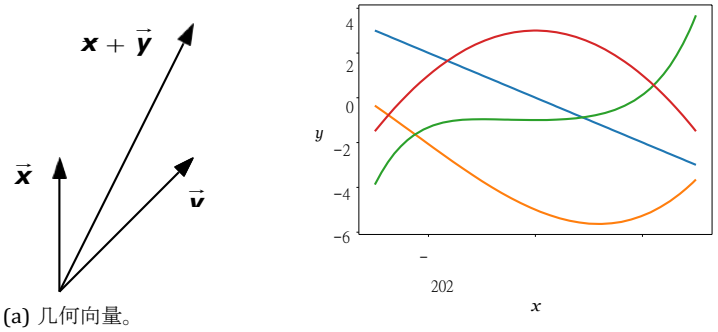
线性代数



在对直观概念进行形式化时，一个常见的方法是构建一套对象（符号）和一套操作这些对象的规则。这被称为*代数*。线性代数是对向量和某些操作向量的规则。我们许多人在学校认识的向量被称为 "几何向量"，通常用一个小箭头表示字母上方，例如，$\vec{x}$和$\vec{y}$。在本书中，我们讨论的是更一般的一般来说，向量是一些特殊的对象，它们可以相加并与标量相乘以产生另一个同类的对象。从抽象的数学观点来看，任何满足这两个属性的对象都可以被认为是一个矢量。下面是一些这样的例子向量对象。

代数的研究。

1. 几何向量。这个矢量的例子可能在高中数学和物理学中很熟悉。几何向量--见图2.1(a)
   - 是有方向的线段，可以被画出来（至少在二维上）。两个几何向量$\vec{x}$，$\vec{y}$可以相加，这样$\vec{x}+\vec{y}=\vec{z}$是另一个几何向量。此外，乘以一个标量$\lambda\vec{x}, \lambda \in \mathbb{R}$，也是一个几何向量。事实上，它是原始矢量
   因此，几何向量是前面介绍的向量概念的实例。将向量解释为几何向量使我们能够利用我们对方向和大小的直觉来推理数学运算。
2. 多项式也是向量，见图2.1(b)。两个多项式可以



(a) 几何向量。

(b) 多项式。

图 不同**2.1**类型的向量。向量可以是令人惊讶的对象，包括(a)几何向量和(b)多项式。

它们可以相加，结果是另一个多项式；它们可以乘以一个标量$\lambda \in \mathbb{R}$，结果也是一个多项式。因此，多项式是矢量的（相当不寻常的）实例。请注意，多项式与几何向量有很大不同。几何向量是具体的 "图纸"，而多项式是抽象的概念。然而，它们都是前面描述的意义上的向量。

3. 音频信号是矢量。音频信号被表示为一系列的数字。我们可以把音频信号加在一起，它们的总和就是一个新的音频信号。如果我们对一个音频信号进行缩放，我们也会得到一个音频信号。因此，音频信号也是矢量的一种。

4. $\mathbb{R}$的元素^n（n个实数的图元）是向量。$\mathbb{R}^n$比多项式更抽象，它是我们在本书中关注的概念。比如说。

$$\boldsymbol{a} = \begin{bmatrix} 1 \\ 0 \\ 3 \end{bmatrix} \in \mathbb{R}^3 \tag{2.1}$$

是一个三联数的例子。将两个向量$\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$分量相加会产生另一个向量：$\boldsymbol{a} + \boldsymbol{b} = \boldsymbol{c} \in \mathbb{R}^n$。此外，将$\boldsymbol{a} \in \mathbb{R}^n$乘以$\lambda \in \mathbb{R}$会产生一个比例向量$\lambda\boldsymbol{a} \in \mathbb{R}^n$。许多编程语言都支持数组操作，这使得涉及向量操作的算法可以方便地实现。

在计算机上实现时要注意检查数组操作是否真正执行了矢量操作。

线性代数关注的是这些矢量概念之间的相似性。我们可以将它们相加，并将它们与标量相乘。我们将主要关注$\mathbb{R}^n$中的向量，因为线性代数中的大多数算法都是在$\mathbb{R}^n$中模拟的。我们将在第8章中看到，我们经常认为数据在$\mathbb{R}^n$中被表示为向量。

Pavel Grinfeld的线性代数系列：http://tinyurl.com/nahclwm

吉尔伯特-斯特朗的线性代数课程：http://tinyurl.com/29p5q8j

3Blue1Brown线性代数系列：https://tinyurl.com/h5g4kps

数学的一个主要思想是 "封闭 "的思想。这是个问题。从我提出的操作中可以产生的所有事物的集合是什么？以向量为例。从一个小的向量集开始，把它们互相加起来并按比例排列，能产生的向量集是什么？这就产生了一个向量空间（第2.4).矢量空间的概念及其属性是许多机器学习的基础。本章所介绍的概念在图中作了总结2.2.

本章主要基于Drumm和Weil(2001)、Strang(2003)、Hogben(2013)、Liesen和Mehrmann(2015)的讲义和书籍，以及Pavel Grinfeld的线性代数系列。其他优秀的

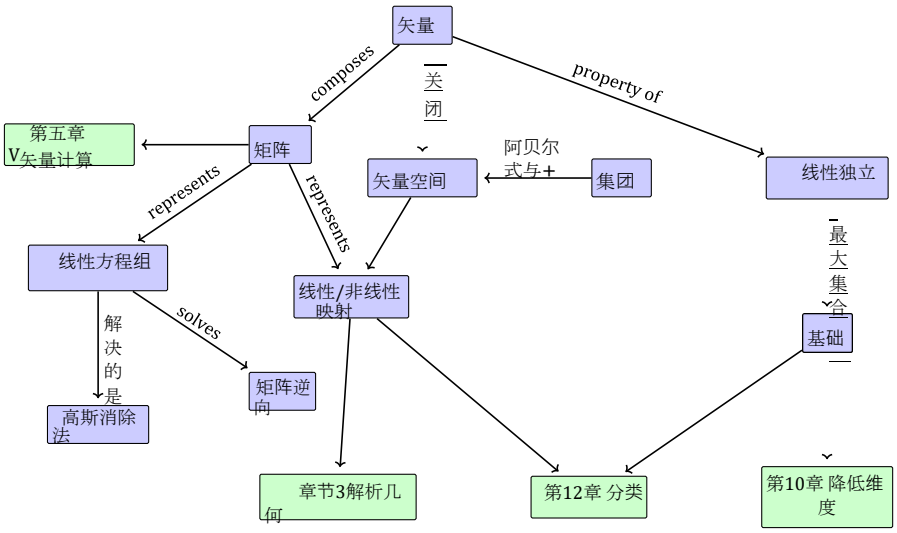


图 本章介绍的概念的**2.2**思维导图，以及这些概念在本书其他部分的使用情况。

resources are Gilbert Strang's Linear Algebra course at MIT and the Linear Algebra Series by 3Blue1Brown.

线性代数在机器学习和普通数学中发挥着重要作用。本章介绍的概念将在第三章中进一步扩展，包括几何学的概念。在第五章中，我们将讨论矢量微积分，其中矩阵运算的原则性知识是必不可少的。在第十章中，我们将使用投影（将在第3.8节中介绍）来进行主成分分析（PCA）的降维。在第九章中，我们将讨论线性回归，其中线性代数在解决最小二乘法问题中起着核心作用。

2.1 线性方程组

线性方程组是线性代数的核心部分。许多问题可以被表述为线性方程组，而线性代数为我们提供了解决这些问题的工具。

例子 2.1

一家公司生产产品$N_1, \ldots, N_n$个，这些产品需要资源$R_1, \ldots, R_m$是需要的。生产一个单位的产品N_j，需要一个$_{ij}$单位的资源R_i，其中$i = 1, \ldots, m$和$j = 1, \ldots, n$.

目标是找到一个最佳的生产计划，即如果总共有b_i个单位的资源R_i，并且（理想情况下）没有剩余的资源，应该生产多少单位的产品N_j的$_j$计划。

如果我们生产$x_1, \ldots, x_n$个单位的相应产品，我们需要

共有

$$i_1 轴_1 + \cdots + i_n 轴_n \tag{2.2}$$

许多单位的资源R_i。一个最佳的生产计划（$x_1, \ldots, x_n$）$\in \mathbb{R}^n$。因此，必须满足以下方程组。

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ &\vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m \end{aligned} \tag{2.3}$$

其中$a_{ij} \in \mathbb{R}$和$b_i \in \mathbb{R}$。

系统的线性方程

解决办法

方程(2.3)是一个*线性方程组的*一般形式，而$x_1, \ldots, x_n$是这个系统的*未知数*。每一个n元组（$x_1, \ldots, x_n$）都$\in$了$\mathbb{R}^n$，满足(2.3)是线性方程组的*解*。

例子 2.2

线性方程组

$$\begin{aligned} x_1 + x_2 + x_3 &= 3 \quad (1) \\ x_1 - x_2 + 2x_3 &= 2 \quad (2) \\ 2x_1 \quad\quad + 3x_3 &= 1 \quad (3) \end{aligned} \tag{2.4}$$

*没有解。将*前两个方程相加得到$2x_1 + 3x_3 = 5$，这与第三个方程（3）相矛盾。

让我们看看线性方程组的情况

$$\begin{aligned} x_1 + x_2 + x_3 &= 3 \quad (1) \\ x_1 - x_2 + 2x_3 &= 2 \quad (2) \\ x_2 + x_3 &= 2 \quad (3) \end{aligned} . \tag{2.5}$$

从第一个和第三个方程可以看出，$x_1=1$。从（1）+（2），我们得到$2x_1+3x_3=5$，即$x_3=1$。从（3）中，我们又得到$x_2=1$。因此，（1，1，1）是唯一可能和*唯一的解*（验证一下

(1,1 , 1)是一个通过插入的解)。

作为第三个例子，我们考虑

$$\begin{aligned} x_1 + x_2 + x_3 &= 3 \quad (1) \\ x_1 - x_2 + 2x_3 &= 2 \quad (2) \\ 2x_1 \quad\quad + 3x_3 &= 5 \quad (3) \end{aligned} . \tag{2.6}$$

由于(1)+(2)=(3)，我们可以省略第三个方程（多余的）。从(1)和(2)，我们得到$2x_1=5-3x_3$和$2x_2=1+x_3$。我们定义$x_3=a \in \mathbb{R}$作为一个自由变量，这样，任何三联体

$$\left(\frac{5}{2} - \frac{3}{2}a, \frac{1}{2} + \frac{1}{2}a, a \right), \quad a \in \mathbb{R} \tag{2.7}$$

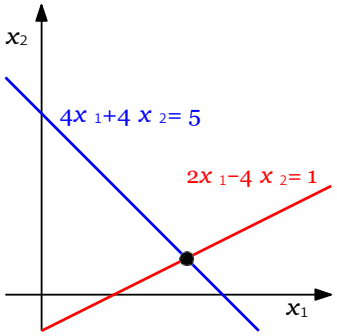


图 由两个变量组成的线性方程组的**2.3** 解空间在几何学上可以解释为两条线的交点。每个线性方程都代表一条线。

是线性方程组的解，也就是说，我们得到一个包含*无限*多解的解集。

一般来说，对于一个实值的线性方程组，我们要么没有解，要么正好有一个解，要么有无限多个解。线性回归（第9章）解决了例题的一个版本2.1当我们不能解决线性方程组的时候。

备注（线性方程组的几何解释）。在一个有两个变量x_1，x_2的线性方程组中，每个线性方程在xx_{12}平面上定义了一条线。由于线性方程组的解必须同时满足所有方程，所以解集是这些线的交点。这个交集可以是一条线（如果线性方程描述的是同一条线），一个点，或空（当线是平行的）。图中给出了一个说明2.3所示，该系统

$$\begin{aligned} 4x_1+4x_2&=5\\ 2x_1-4x_2&=1 \end{aligned} \tag{2.8}$$

其中解空间是点（x_1，x_2）=（1，）[1]。同样地，对于三个变量，每个线性方程都决定了三维空间中的一个平面。当我们与这些平面相交时，即同时满足所有的线性方程，我们可以得到一个解集，它是一个平面、一条线、一个点或空（当平面没有共同的交点时）。 ◆

为了系统地解决线性方程组的问题，我们将引入一个有用的紧凑符号。我们将系数a收集$_{ij}$成向量，将向量收集成矩阵。换句话说，我们把系统从(2.3)的形式。

$$\begin{matrix} \text{是1} \\ \vdots \\ \text{是2} \end{matrix} x_1 + \begin{matrix} \\ \vdots \\ \end{matrix} x_2 + \cdots + \begin{matrix} \text{1一个} \\ \vdots \\ amn \end{matrix} x_n = \begin{matrix} b_1 \\ \vdots \\ b_m \end{matrix} \tag{2.9}$$

$$\Longleftrightarrow \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix} \tag{2.10}$$

在下文中，我们将仔细研究这些*矩阵*和组化的计算规则。我们将在第二节回到解决线性方程的问题。2.3.

2.2 矩阵

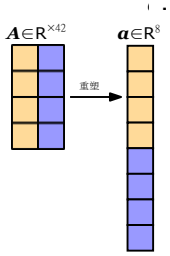
矩阵在线性代数中起着核心作用。它们可以用来表示线性方程组，但它们也可以表示线性函数（线性映射），我们将在第2.7节中看到。在讨论这些有趣的话题之前，让我们先定义一下什么是矩阵，以及我们可以用矩阵做什么样的运算。我们将在第四章中看到矩阵的更多属性。

基体

定义2.1（矩阵）。以m，n N为例,∈一个实值（m，n）*矩阵*$\boldsymbol{A}$是一个由元素a_{ij}，$i = 1, \ldots, m, j = 1, \ldots, n$，根据由$m$行和$n$列组成的矩形方案排序。

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & 1\text{一个} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \in \mathrm{R}. \tag{2.11}$$

行
列行向量
列向量 图 通过2.4
堆叠其列，一个矩阵可以表示为一个$\boldsymbol{A}$长向量



注意矩阵的大小。

C =
np.einsum('il, lj', A, B)

按照惯例，$(1, n)$-矩阵被称为*行*，$(m, 1)$-矩阵被称为*列*。这些特殊的矩阵也被称为*行/列向量*。

$\mathrm{R}^{m\times n}$是所有实值（m，n）矩阵的集合。一个$\mathrm{R}^{m\times n}$可以^{mn}通过将矩阵的所有n列堆叠成一个长矢量来等效地表示为一个R；见图2.4.

2.2.1 矩阵加法和乘法

两个矩阵$\boldsymbol{A}$ $\mathrm{R}^{m\times n}$、$\boldsymbol{B}$ R之^{m}∈^{n}和被定义为元素明智之和，即：。

$$\boldsymbol{A} + \boldsymbol{B} \in \begin{bmatrix} a_{11} + b_{11} & \cdots & a_{1n} + b_{1n} \\ \vdots & & \vdots \\ a_{m1} + b_{m1} & \cdots & a_{mn} + b_{mn} \end{bmatrix} \in \mathrm{R}^{m\times n}. \tag{2.12}$$

对于矩阵$\boldsymbol{A}\in\mathrm{R}^{m\times n}$，$\boldsymbol{B}\in\mathrm{R}^{n\times k}$，乘积的元素$c_{ij}$
$\boldsymbol{C}=\boldsymbol{A}\boldsymbol{B}\in\mathrm{R}^{m\times k}$的计算方法为

$$c_{ij} = \sum_{l=1}^{n} a_{il} b_{lj}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, k. \tag{2.13}$$

这意味着，为了计算元素c_{ij}，我们将第i个元素相乘。

解释$\boldsymbol{A}$的第i行和$\boldsymbol{B}$的第j列，并将它们相加。在后面的章节中3.2，我们将其称为相应行和列的*点积*。在一些情况下，我们需要明确指出我们正在进行乘法运算。我们用符号$\boldsymbol{A}\cdot\boldsymbol{B}$来表示乘法（明确显示"-"）。

中和n行中我们可以在$\boldsymbol{B}$ Aso计算一个ij为$l=1,\ldots,n$。通常情况下，两个向量之间的点积，是$\boldsymbol{aTb}$ $\boldsymbol{a}\,\boldsymbol{b}$用表示，或$(\boldsymbol{a},\boldsymbol{b})$.

备注。矩阵只有在其 "相邻 "维度上才能相乘匹配。例如，一个$n\times k$矩阵$\boldsymbol{A}$可以与一个$k\times m$-矩阵$\boldsymbol{B}$相乘，但只能从左边开始。

$$\underset{n\times k}{\boldsymbol{A}}\times\underset{k\times m}{\boldsymbol{B}}=\underset{n\times m}{\boldsymbol{C}} \tag{2.14}$$

如果$m\neq n$，则乘积$\boldsymbol{BA}$没有定义，因为相邻的尺寸不匹配。 ◆

备注。矩阵乘法*并不*定义为对矩阵元素的明智操作，即$c_{ij}=a_{ij}b_{ij}$（即使$\boldsymbol{A}$、$\boldsymbol{B}$的大小被适当选择）。在编程语言中，当我们对（多维）数组进行乘法运算时，这种从元素角度出发的乘法运算经常出现。

彼此之间的关系，被称为*哈达玛德积*。 ◆ 哈达玛德积

例子 2.3

对于$\boldsymbol{A}=\begin{bmatrix}1&2&3\\3&2&1\end{bmatrix}\in\mathbb{R}^{2\times 3}$, $\boldsymbol{B}=\begin{bmatrix}0&2\\1&-1\\0&1\end{bmatrix}\in\mathbb{R}^{3\times 2}$，我们得到

$$\boldsymbol{AB}=\begin{bmatrix}1&2&3\\3&2&1\end{bmatrix}\begin{bmatrix}0&2\\1&-1\\0&1\end{bmatrix}=\begin{bmatrix}2&3\\2&5\end{bmatrix}\in\mathbb{R}^{2\times 2}. \tag{2.15}$$

$$\boldsymbol{BA}=\begin{bmatrix}0&2\\1&-1\\0&1\end{bmatrix}\begin{bmatrix}1&2&3\\3&2&1\end{bmatrix}=\begin{bmatrix}6&4&2\\-2&0&1\\3&2&1\end{bmatrix}\in\mathbb{R}^{3\times 3}. \tag{2.16}$$

图 2.5 即使$\boldsymbol{A}$和$\boldsymbol{B}$可能是这样的，矩阵乘法被定义，结果的尺寸不同。

从这个例子中，我们已经可以看出，矩阵乘法不是交换性的，也就是说，$\boldsymbol{AB}\neq\boldsymbol{BA}$；另见图2.5来说明。

定义（2.2*身份矩阵*）。在$\mathbb{R}^{n\times n}$中，我们定义*身份矩阵*

$$\boldsymbol{I}_n:=\begin{bmatrix}1&0&\cdots&0&0\\0&1&\cdots&0&0\\\vdots&\vdots&\ddots&\vdots&\vdots\\0&0&\cdots&1&0\\0&0&\cdots&0&1\end{bmatrix}\in\mathbb{R}^{n\times n} \tag{2.17}$$

作为包含在对角线上1和其他地方0的$n \times n$矩阵。

现在我们已经定义了矩阵乘法、矩阵加法和识别矩阵，让我们来看看矩阵的一些特性。

关联性

- *关联性。*

$$\forall \boldsymbol{a} \in \mathsf{r}^{m \times n},\ \boldsymbol{b} \in \mathsf{r}^{n \times p},\ \boldsymbol{c} \in \mathsf{r}^{p \times q}:\ (\boldsymbol{ab})\boldsymbol{c} = \boldsymbol{a}(\boldsymbol{bc}) \tag{2.18}$$

分布性

- *分配性。*

$$\forall \boldsymbol{A}, \boldsymbol{B} \in \mathsf{R}^{m \times n},\ \boldsymbol{C}, \boldsymbol{D} \in \mathsf{R}^{n \times p}:\ (\boldsymbol{A} + \boldsymbol{B})\boldsymbol{C} = \boldsymbol{AC} + \boldsymbol{BC} \tag{2.19a}$$

$$\boldsymbol{A}\ (\boldsymbol{C} + \boldsymbol{D}) = \boldsymbol{AC} + \boldsymbol{AD} \quad (2.19b)\ 。$$

- *与身份矩阵相乘。*

$$\forall \boldsymbol{a} \in \mathsf{r}^{m \times n}:\ {}_{m}\boldsymbol{ia} = \boldsymbol{ai}_{n} = \boldsymbol{a} \tag{2.20}$$

请注意，${}_{n}$对于$m \neq n$，$\boldsymbol{I}_{m} \neq \boldsymbol{I}$。

一个正方形矩阵拥有相同数量的列和行。

2.2.2 逆向和转置

定义2.3（逆向）。考虑一个正方形矩阵$\boldsymbol{A} \in \mathsf{R}^{n \times n}$，让矩阵$\boldsymbol{B} \in \mathsf{R}^{n \times n}$具有$\boldsymbol{AB} = \boldsymbol{I}_{n} = \boldsymbol{BA}$的性质。$\boldsymbol{B}$被称为$\boldsymbol{A}$的*逆*，用$\boldsymbol{A}^{-1}$表示。

不幸的是，并不是每个矩阵$\boldsymbol{A}$都拥有一个逆$\boldsymbol{A}^{-1}$。

正规的可倒置的非星形单一的不可逆的

矩阵的逆值确实存在，则$\boldsymbol{A}$被称为*正则*/*可逆*/非*正则*，否则称为*奇异*/不*可逆*。当矩阵的逆存在时，它是唯一的。在本节中，2.3,我们将讨论一种通过解决线性方程组来计算矩阵逆的一般方法。

备注（a×22-矩阵的逆的存在）。考虑一个矩阵

$$\boldsymbol{A} := \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \in \mathsf{R}^{2 \times 2}。 \tag{2.21}$$

如果我们将$\boldsymbol{A}$乘以

$$\boldsymbol{A}' := \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} \tag{2.22}$$

we obtain

$$\boldsymbol{AA}' = \begin{bmatrix} a_{11}a_{22} - a_{12}a_{21} & 0 \\ 0 & a_{11}a_{22} - a_{12}a_{21} \end{bmatrix} = (a_{11}a_{22} - a_{12}a_{21})\boldsymbol{I}_{2}. \tag{2.23}$$

因此。

$$\boldsymbol{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} \tag{2.24}$$

26 当且仅当 $a_{11}a_{22} - a_{12}a_{21} \neq 0$。在第4.1节，我们将看到行列式可以用来计算

aa_{1221}是a×22-矩阵的行列式。此外，我们一般可以用行列式来检查一个矩阵是否可逆。 ◆

例子（2.4逆矩阵）

矩阵

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 1 \\ 4 & 45 & 7 \\ 6 & 7 & 7 \end{bmatrix} \quad \boldsymbol{B} = \begin{bmatrix} -7 & -7 & 6 \\ 2 & 1 & -1 \\ 4 & 5 & -4 \end{bmatrix} \tag{2.25}$$

因为$\boldsymbol{AB}=\boldsymbol{I}=\boldsymbol{BA}$，*所以*彼此互为反比。

定义（2.4转置）。对于AR来说$^{m\times n}$, ∈矩阵BR$^{n\times m}$有$b_{ij}=a_{ji}$被称为$\boldsymbol{A}$的转置，我们写$\boldsymbol{B}=\boldsymbol{A}^{\mathsf{T}}$。

一般来说，$\boldsymbol{A}^{\mathsf{T}}$可以通过将$\boldsymbol{A}$的列写成$\boldsymbol{A}$的T行来获得。以下是反转和转置的重要属性。

$$\boldsymbol{aa}^{-1}=\boldsymbol{i}=\boldsymbol{aa} \tag{2.26}$$

$$(\boldsymbol{ab})^{-1}=\boldsymbol{ba}(^{-1} \tag{2.27}$$

$$(\boldsymbol{a}+\boldsymbol{b})^{-1} \quad \boldsymbol{a}^{-1}+\boldsymbol{b}^{-1} \tag{2.28}$$

$$(\boldsymbol{A}^{\mathsf{T}})^{\mathsf{T}}=\boldsymbol{A} \tag{2.29}$$

$$(\boldsymbol{A}+\boldsymbol{B})=\boldsymbol{A}^{\mathsf{T}}+\boldsymbol{B}^{\mathsf{T}} \tag{2.30}$$

$$(\boldsymbol{ab})^{\mathsf{T}}=\boldsymbol{ba}(^{\mathsf{T}} \tag{2.31}$$

定义（2.5对称矩阵）。一个矩阵AR$^{n\times n}$是*对称的*, ∈如果$\boldsymbol{A}=\boldsymbol{A}^{\mathsf{T}}$。

请注意，只有(n, n)-矩阵可以是对称的。一般来说，我们把(n, n)-矩阵也是*方形矩阵*，因为它们拥有相同的num行和列的数量。此外，如果$\boldsymbol{A}$是可逆的，那么$\boldsymbol{A}$也T是可逆的，并且$(\boldsymbol{A}^{-1})^{\mathsf{T}}=(\boldsymbol{A})^{\mathsf{T}-1}=。\boldsymbol{A}^{-\mathsf{T}}$.

备注（对称矩阵的和与积）。对称矩阵$\boldsymbol{A}$，$\boldsymbol{B}$ R的$^{n\times n}$和总是对称的。然而，尽管它们的乘积总是被定义的，但它通常不是对称的。

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}=\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}。 \tag{2.32}$$

◆

2.2.3 乘以一个标量的乘法

主对角线（有时称为"主对角线"、"主要对角线"、"主导对角线"或"主要对角线"）。

一个 矩阵的 A 词条集 A_{ij}，其中$i=j$。的标量情况是(2.28)的标量情况_是 $\frac{1}{2+4}=\frac{1}{6}\quad\frac{1}{2}+\frac{1}{4}$

对称矩阵

-square矩阵

让我们来看看当矩阵乘以标量$\lambda \in \mathbb{R}$时会发生什么$_{ij}$。实际上，λ对$\boldsymbol{A}$的 *线性代数* 每个元素都有标度，对于$\lambda, \psi \in \mathbb{R}$，以下情况成立。

关联性

- *Associativity*。
  $(\lambda\psi)\boldsymbol{C} = \lambda(\psi\boldsymbol{C}), \quad \boldsymbol{C} \in \mathrm{R}^{m\times n}$
- $\lambda(\boldsymbol{B}\boldsymbol{C}) = (\lambda\boldsymbol{B})\boldsymbol{C} = \boldsymbol{B}(\lambda\boldsymbol{C}) = (\boldsymbol{B}\boldsymbol{C})\lambda$, $\boldsymbol{B} \in \mathrm{R}^{m\times n}$, $\boldsymbol{C} \in \mathrm{R}^{n\times k}$。
  请注意，这允许我们移动标量值。

分布性

- $(\lambda\boldsymbol{C})^\top = \boldsymbol{C}^\top\lambda^\top = \boldsymbol{C}^\top\lambda = \lambda\boldsymbol{C}$，因为$\top$对于所有$\lambda \in \mathrm{R}$，$\lambda^\top = \lambda$。
- *分配性*。
  $(\lambda + \psi)\boldsymbol{C} = \lambda\boldsymbol{C} + \psi\boldsymbol{C}, \quad \boldsymbol{C} \in \mathrm{R}^{m\times n}$
  $\lambda(\boldsymbol{B} + \boldsymbol{C}) = \lambda\boldsymbol{B} + \lambda\boldsymbol{C}, \quad \boldsymbol{B}, \boldsymbol{C} \in \mathrm{R}^{m\times n}$

例子（2.5分配性）

如果我们定义

$$\boldsymbol{C} := \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \tag{2.33}$$

那么对于任何$\lambda, \psi \in \mathrm{R}$，我们可以得到

$$(\lambda + \psi)\boldsymbol{C} = \begin{bmatrix} (\lambda+\psi)1 & (\lambda+\psi)2 \\ (\lambda+\psi)3 & (\lambda+\psi)4 \end{bmatrix} = \begin{bmatrix} \lambda+\psi & 2\lambda+2\psi \\ 3\lambda+3\psi & 4\lambda+4\psi \end{bmatrix} \tag{2.34a}$$

$$= \begin{bmatrix} \lambda & 2\lambda \\ 3\lambda & \lambda \end{bmatrix} + \begin{bmatrix} \psi & 2\psi \\ \psi & 4\psi \end{bmatrix} = \lambda\boldsymbol{C} + \psi\boldsymbol{C} \tag{2.34b}$$

2.2.4 线性方程系统的紧凑表示

如果我们考虑线性方程组

$$\begin{aligned} 2x_1 + 3x_2 + 5x_3 &= 1 \\ 4x_1 - 2x_2 - 7x_3 &= 8 \\ 9x_1 + 5x_2 - 3x_3 &= 2 \end{aligned} \tag{2.35}$$

并使用矩阵乘法的规则，我们可以把这个方程组写成一个更紧凑的形式，即

$$\begin{bmatrix} 2 & 3 & 5 \\ 4 & -2 & -7 \\ 9 & 5 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 8 \\ 2 \end{bmatrix} \tag{2.36}$$

请注意，x_1代表第一列，x_2代表第二列，而x_3代表第三列。

一般来说，一个线性方程组可以用矩阵形式紧凑地表示为$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$；见(2.3)，乘积$\boldsymbol{A}\boldsymbol{x}$是$\boldsymbol{A}$列的（线性）组合。我们将在第二节详细讨论线性组合。2.5.

2.3 解决线性方程组的问题

在(2.3)中，我们介绍了方程组的一般形式，即：。

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ &\vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m \end{aligned} \tag{2.37}$$

。

其中$a_{ij} \in \mathbb{R}$和$b_i \in \mathbb{R}$是已知常数，x_j是未知数，$i = 1, \ldots, m$，$j = 1, \ldots$至此，我们看到矩阵可以作为一种紧凑的方式来表述线性方程组，这样我们就可以写出$\boldsymbol{Ax} = \boldsymbol{b}$，见(2.10).此外，我们定义了基本的矩阵运算，如矩阵的加法和乘法。在下文中，我们将重点讨论解决线性方程组的问题，并提供一种寻找矩阵逆的算法。

2.3.1 特定的和一般的解决方案

在讨论如何普遍解决线性方程组之前，让我们先看看一个例子。考虑方程组

$$\begin{bmatrix} 1 & 0 & 8 & -4 \\ 0 & 1 & 2 & 12 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 42 \\ 8 \end{bmatrix} . \tag{2.38}$$

The system has two equations and four unknowns. Therefore, in general we would expect infinitely many solutions. This system of equations is in a particularly easy form, where the first two columns consist of a 1 and a 0. Remember that we want to find scalars $x_1, \ldots, x_4$, such that $\sum_{i=1}^{4} x_i \boldsymbol{c}_i = \boldsymbol{b}$，其中我们定义$\boldsymbol{c}_i$ 是矩阵的第i列，而$\boldsymbol{b}$的右手边(2.38).在(2.38)中的问题，可以通过取第一列的42倍和第二列的8倍立即找到，因此

$$\boldsymbol{b} = \begin{bmatrix} 42 \\ 8 \end{bmatrix} = 42 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 8 \begin{bmatrix} 0 \\ 1 \end{bmatrix} 。 \tag{2.39}$$

因此，一个解是$[42, 8, 0, 0]^\top$。这个解被称为*特定解*（particular solution）或*特殊解*。然而，这并不是这个特殊解的线性方程组的唯一解。为了捕捉所有其他的解，我们需要要创造性地使用矩阵的列来**0**生成非琐碎的方式。在我们的特解中加入**0**并不改变特解。为此，我们用前两列来表示第三列(这些都是非常简单的形式)

$$\begin{vmatrix} 8 & 1 \\ 2 & 0 \end{vmatrix} = 8 + \begin{vmatrix} 0 & 2 \\ 0 & 1 \end{vmatrix} \tag{2.40}$$

以便=**0**8$\boldsymbol{c}_1$+2$\boldsymbol{c}_2$−1$\boldsymbol{c}_3$+0$\boldsymbol{c}_4$和（x_1，x_2，x_3，x_4）=（8，2，−1，0）。事实上，这个解决方案的任何缩放比例由$\lambda_1 \in \mathbb{R}$产生的**0**矢量，即：。

$$\begin{bmatrix}0&1&2\\1&0&8\\-4&&\end{bmatrix}\left(\lambda_1\begin{bmatrix}8\\2\\-1\\0\end{bmatrix}+\lambda_2\begin{bmatrix}8\\0\\-1\\0\end{bmatrix}\right)=\boldsymbol{0}. \quad (2.41)$$

按照同样的推理，我们用 (2.38)使用前两列，并生成另一组的非琐碎版本，因为**0**

$$\begin{bmatrix}1&0&8&-4\\0&1&2&\end{bmatrix}\lambda_{12}\begin{bmatrix}4\\-4\\1\\-1\end{bmatrix}=\lambda_{12}(-4\boldsymbol{c}_1+\boldsymbol{c}_2)=\boldsymbol{0} \quad (2.42)$$

对于任何$\lambda_{12} \in \mathbb{R}$。把所有的东西放在一起，我们可以得到( )中方程组的所有解，这被称为一般解。2.38)中的所有解，也就是所谓的*一般解*，因为这组 ∈

一般解决方案

$$\boldsymbol{x}\in\mathbb{R}^4:\boldsymbol{x}=\begin{bmatrix}8\\0\\0\\0\end{bmatrix}+\lambda_1\begin{bmatrix}8\\0\\-1\\0\end{bmatrix}+\lambda_2\begin{bmatrix}4\\-4\\1\\-1\end{bmatrix},\lambda_1,\lambda_2\in\mathbb{R}. \quad (2.43)$$

*备注。*我们遵循的一般方法包括以下三个步骤。

1.找出$\boldsymbol{Ax}=\boldsymbol{b}$的一个特定解。 2.找出
$\boldsymbol{Ax}=\boldsymbol{0}$.的所有解。
3.将步骤和1.的解决方案结合起来，2.得出一般解决方案。

一般解和特殊解都不是唯一的。

前面的例子中的线性方程组很容易解决，因为 (2.38)中的矩阵具有这种特别方便的形式，这使得我们可以通过反省找到特殊和一般的解。然而，一般的方程组并不具有这种简单的形式。幸运的是，存在一种建设性的算法，可以将任何线性方程组转换为这种特别简单的形式。高斯消除法。高斯消除法的关键是线性方程组的基本转换，它将方程组转换为简单的形式。然后，我们可以将这三个步骤应用于简单形式，即我们刚才在（）的例子中讨论过。2.38).

2.3.2 初级变换

初级的转换

解决线性方程组的关键是*基本变换*，保持解集不变，但将方程组转换成更简单的形式。

- 两个方程的交换（代表方程组的矩阵中的行）。
- 方程（行）与常数$\lambda \in \mathbb{R}\setminus\{0\}$的乘法关系。
- 两个方程的加法（行）

例子 2.6

对于$a \in \mathbb{R}$，我们寻求以下方程组的所有解。

$$\begin{array}{rrrrrrr} -2x_1 & +4x_2 & -x_3 & -x_4 & +4x_5 & = & -3 \\ 4x_1 & -8x_2 & +3x_3 & -3x_4 & +x_5 & = & 2 \\ x_1 & -2x_2 & +x_3 & -x_4 & +x_5 & = & 0 \\ x_1 & -2x_2 & & -3x_4 & +4x_5 & = & a \end{array} . \tag{2.44}$$

我们首先将这个方程组转换为紧凑的矩阵符号$\boldsymbol{Ax} = \boldsymbol{b}$。我们不再明确提及变量$\boldsymbol{x}$，而是建立一个*增强的矩阵*（形式为$\boldsymbol{A} \mid \boldsymbol{b}$）

增强的矩阵

$$\left[\begin{array}{cccc|c} 1 & -2 & -1 & 4 & -3 \\ 4 & -8 & -3 & 1 & 2 \\ 1 & -2 & -1 & 0 & 0 \\ 1 & -2 & 0 & -3 & a \end{array}\right] \begin{array}{l} \\ -4R_1 \\ -R_1 \\ -R_1 \end{array}$$

与R_3互换 与R_2互换

其中我们用垂直线将左手边和右手边分开，在(2.44).我们用$\rightsquigarrow$来表示使用基本变换对增量矩阵进行变换。

扩增后的矩阵紧凑地$[\boldsymbol{A}|\boldsymbol{b}]$表示线性方程组$\boldsymbol{Ax} = \boldsymbol{b}$.

交换行和1导致3

$$\left[\begin{array}{cccc|c} 1 & -2 & -1 & 4 & -3 \\ 4 & -8 & -3 & 1 & 2 \\ 1 & -2 & -1 & 0 & 0 \\ 1 & -2 & 0 & -3 & a \end{array}\right] \begin{array}{l} \\ -4R_1 \\ +2R_1 \\ -R_1 \end{array}$$

当我们现在应用指定的转换（例如，减去行 1 从行2）的四次，我们得到

$$\left[\begin{array}{cccc|c} 1 & -2 & -1 & 4 & -3 \\ 0 & 0 & -1 & -3 & 6 \\ 0 & 0 & -1 & -2 & 3 \\ 1 & -2 & 1 & -1 & 1 \\ 0 & 0 & -1 & -3 & 6 \\ 0 & 0 & -1 & -3 & 0 \end{array}\right] \begin{array}{l} \\ \\ -R_2 - R_3 \\ \\ \\ a+1-(-1) \end{array}$$

$$\left[\begin{array}{cccc|c} 1 & -2 & -1 & 4 & -3 \\ 0 & 0 & -1 & -3 & 6 \\ 0 & 0 & 0 & 0 & 3 \\ 0 & 0 & -1 & -3 & 6 \\ 0 & 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 0 & a+1 \end{array}\right]$$

行-歇尔形式

这个（增强的）矩阵是一个方便的形式，即*行-歇尔形式*（REF）。将这一紧凑的符号还原为我们所寻求的变量的显式符号，我们可以得到

$$\begin{array}{rrrrrrrrr} x_1 & -2x_2 & & +x_3 & -x_4 & + & x_5 & = & 0 \\ & & & x_3 & - & x_4 & +3x_5 & = & -2 \\ & & & & & x_4 & -2x_5 & = & 1 \\ & & & & & & 0 & = & a+1 \end{array} . \tag{2.45}$$

特定解决方案

只有对于一个$a=-1$这个系统可以被解决。一个*特殊的解决方案*是

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ -1 \\ 1 \\ 0 \end{bmatrix} . \tag{2.46}$$

一般解决方案

捕捉到所有可能解决方案的集合的*一般解决方案*是

$$\boldsymbol{x} \in \mathbb{R}^5 : \boldsymbol{x} = \begin{bmatrix} 2 \\ 0 \\ -1 \\ -1 \\ 0 \end{bmatrix} + \lambda_1 \begin{bmatrix} 2 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 2 \\ 0 \\ 1 \\ 2 \\ 1 \end{bmatrix} \quad \lambda_1, \lambda_2 \in \mathbb{R} \tag{2.47}$$

。

在下文中，我们将详细介绍获得线性方程组的特定和一般解决方案的建设性方法。

枢轴

备注（支点和阶梯结构）。一行的前导系数（从左边开始的第一个非零数）被称为*支点*，并且总是严格地在它上面一行的支点的右边。因此，任何等式

行-歇尔形式

行-歇尔形式的分配系统总是有一个 "阶梯 "结构。 ◆

定义（2.6行-梯队形式）。在下列情况下，一个矩阵是*行-梯形形式*的

- 所有只包含零的行都在矩阵的底部；相应地，所有包含至少一个非零元素的行都在只包含零的行的上面。

枢轴

- 只看非零行，从左边开始的第一个非零数（也称为*枢轴*或*前导系数*）总是严格地指向

在它上面一行的枢轴的前导系数。

在其他文本中，它是

有时要求支点是1

备注（基本变量和自由变量）。对应于行--歇尔形式中的枢轴的变量被称

。

为*基本变量*，而其他variables是*自由变量*。例如，在(2.45)中，x_1、x_3、x_4是基本，而x_2，x_5是自由变量。 ◆

基本变量
自由变量

备注（获得特定解）。行-echelon形式使

当我们需要确定一个特定的解决方案时，我们的生活会更容易。为了做到这一点，我们用支点来表达方程组的右侧列，使得$\boldsymbol{b} = \sum_{i=1}^{P} \lambda_i \boldsymbol{p}_i$，其中$\boldsymbol{p}_i$，$i=1, \ldots, P$，是支点列。如果我们从最右边的支点列开始，然后向左走，λ_i的确定是最容易的。

在前面的例子中，我们将试图找到λ_1、λ_2、λ_3以便

$$\lambda_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \lambda_3 \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}. \tag{2.48}$$

从这里，我们相对直接地发现$\lambda_3=1$，$\lambda_2=1$，$\lambda_1=2$。当我们把所有的东西放在一起时，我们不能忘记非支点列，我们把它们的系数隐含地设置为0。因此，我们得到特定的解决方案$\boldsymbol{x} = [2, 0, -1, 1, 0]^\top$。 ◆

备注（减行梯形形式）。一个方程组是在*减缩的*减缩的*行-梯形*（又称：*行-还原梯形或行-典型形式*），如果 行-歇尔形式

- 它是以行-歇尔形式存在的。
- 每个支点都是1.
- 枢轴是其列中唯一的非零条目。

◆

减少的行-歇尔形式将在后面的章节中发挥重要作用。2.3.3因为它允许我们确定一个系统的一般解。

以一种直接的方式 对线性方程进行分析。

备注（高斯消除法）。*高斯消除法*是一种执行基本转换的算法，将线性方程组转化为简化的行-歇尔形式。 ◆ 高斯消除

例子（2.7减少行的梯队形式）

验证以下矩阵是否为减行-减列形式（枢轴为**黑体**）。

$$\boldsymbol{A} = \begin{bmatrix} \mathbf{1} & 3 & 0 & 3 \\ 0 & 0 & \mathbf{1} & 9 \\ 0 & 0 & 0 & -4 \end{bmatrix}. \tag{2.49}$$

寻找$\boldsymbol{Ax}=\mathbf{0}$的解决方案的关键想法是看一下*非枢轴列*，我们需要将其表达为一个（线性）的组合枢轴列。缩减的行梯形使之变得相对简单，我们用其左边的枢轴列的和和倍数来表示非枢轴列。第二列是第一列的倍数3（我们可以忽略左边的枢轴列）。

第二列的右边）。因此，为了得到$\mathbf{0}$，我们需要减去

第二列是第一列的三倍。现在，我们看一下第五列，这是我们的第二个非枢纽列。第五列可以表示为第一支点列的倍数3，第二支点的倍数9列，和−4倍的第三支点列。我们需要跟踪的是枢轴列的指数，并将其转化为第一列的倍数3，第二列（非枢轴列）的倍数0，第三列的倍数9。第三列（也就是我们的第二支点列），以及−4倍的第四列（也就是第三支点列）。然后我们需要减去第五列，得到 $\mathbf{0}$。最后，我们仍然在解决一个同质方程组。

综上所述，$\boldsymbol{Ax}=\mathbf{0}$，$\boldsymbol{x}\in\mathbb{R}^5$的所有解由以下公式给出

$$\boldsymbol{x}\in\mathbb{R}^5 : \boldsymbol{x}=\lambda_1\begin{bmatrix}3\\0\\0\\0\\0\end{bmatrix}+\lambda_2\begin{bmatrix}9\\0\\-4\\-1\\0\end{bmatrix}\qquad \lambda_1,\lambda_2\in\mathbb{R} \tag{2.50}$$

2.3.3 减去1的伎俩

在下文中，我们将介绍一个实用的技巧，以读出同质线性方程组$\boldsymbol{Ax}=\mathbf{0}$，其中$\boldsymbol{A}\in\mathbb{R}^{k\times n}$，$\boldsymbol{x}\in\mathbb{R}^n$的解$\boldsymbol{x}$。

首先，我们假设$\boldsymbol{A}$是简化的行-歇尔形式，没有任何只包含零的行，也就是说。

$$\boldsymbol{A}=\begin{bmatrix}0\cdots0\,1 & * & \cdots & * & 0 & * & \cdots & * & 0 & * & \cdots & * \\ \vdots & & & & 0\cdots0\,1 & * & \cdots & * & 0 & * & \cdots & * \\ & & & & & & & & 0\cdots0\,1 & * & \cdots & * \\ & & & & & & & & & & & \end{bmatrix}, \tag{2.51}$$

其中可以是一个任意的实数，约束条件是每行的第一个非零条目必须是1，而相应列中的所有其他条目必须是0。列$j_1,\ldots,j_k$与枢轴（标记为黑体）是标准单位向量$\boldsymbol{e}_1,\ldots$。我们通过增加$n-k$行的形式将这个矩阵扩展为$n\times n$矩阵$\tilde{\boldsymbol{A}}$。

$$\begin{bmatrix}0 & \cdots & 0 & -1 & 0 & \cdots & 0\\ & & & & & & \\ & & & & & & \end{bmatrix} \tag{2.52}$$

因此，增强后的矩阵$\tilde{\boldsymbol{A}}$的对角线包含或1-1。

那么，$\tilde{\boldsymbol{A}}$中包含1-的列作为支点，是解的

更准确地说，这些柱子构成了$\boldsymbol{Ax}=\mathbf{0}$的解空间的基础（第2.6.1节）$\mathbf{0}$。
我们稍后将称其为*内核或空空间*（见第2.7.3）。 内核
空旷的空间

例子（2.8减去1的窍门）

让我们重新审视( )中的矩阵。2.49)中的矩阵，它已经在减少的REF中了。

$$\boldsymbol{A}=\begin{bmatrix}1&3&0&0&3\\0&0&1&0&9\\0&0&0&1&-4\end{bmatrix}. \tag{2.53}$$

现在，我们5通过增加以下的行，将这个矩阵增强为×5矩阵形式(2.52)，在对角线上的支点缺失的地方，得到

$$\tilde{\boldsymbol{A}}=\begin{bmatrix}1&3&0&&03\\0&-1&000&&\\0&0&1&0&9\\0&0&0&1&-4\\0000&&&&-1\end{bmatrix}. \tag{2.54}$$

从这个表格中，我们可以立即读出$\boldsymbol{Ax}$=by$\mathbf{0}$的解决方案取$\tilde{\boldsymbol{A}}$的列， 其对角线上包含-1。

$$\boldsymbol{x}\in\mathbb{R}^5:\boldsymbol{x}=\lambda_1\begin{bmatrix}1\\0\\0\\0\end{bmatrix}+\lambda_2\begin{bmatrix}\upsilon\\9\\-4\\-1\end{bmatrix},\quad \lambda_1,\lambda_2\in\mathbb{R} \tag{2.55}$$

。

这与我们通过 "洞察力 "得到的（） 中的解决方案相同。2.50)中，我们通过 "洞察力 "得到的解决方案。

计算倒数

为了计算$\boldsymbol{A}$ R^{-1}的逆$\boldsymbol{A}^{n\times n}$，我们需要找到一个满足$\boldsymbol{AX}=\boldsymbol{I}_n$的矩阵$\boldsymbol{X}$，然后，$\boldsymbol{X}=\boldsymbol{A}^{-1}$。我们可以把这写成一组同步线性方程$\boldsymbol{AX}=\boldsymbol{I}_n$，其中我们求解$\boldsymbol{X}=[\boldsymbol{x}_1\ \boldsymbol{x}]_n$。我们使用增强的矩阵符号来表示一个紧凑的
表示这组线性方程组，并得到

$$\boldsymbol{A}|\boldsymbol{I}]_n\ \text{--}\!\!-\!\!-\ \boldsymbol{I}|_n\boldsymbol{A}^{-1}]. \tag{2.56}$$

这意味着，如果我们将增强的方程组转化为简化的行-歇尔形式，我们可以在方程组的右侧读出逆。因此，确定矩阵的逆就等同于解决线性方程

组。

例子（2.9通过高斯消除法计算反矩阵）。

为了确定

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 2 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \tag{2.57}$$

我们写下增强的矩阵

$$\left[\begin{array}{cccc|cccc} 1 & 0 & 2 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 2 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \end{array}\right]$$

并使用高斯消除法将其转化为简化的行-歇尔形式

$$\left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & -1 & 2 & -\frac{1}{2} & 0 \\ 0 & 1 & 0 & 0 & 0 & -\frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 & 0 & -\frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & -\frac{1}{2} & 1 \end{array}\right],$$

这样，所需的逆是作为其右手边给出的。

$$\boldsymbol{A}^{-1} = \begin{bmatrix} -1 & 2 & -\frac{1}{2} & 0 \\ 0 & -\frac{1}{2} & \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} & \frac{1}{2} & 0 \\ -1 & 0 & -\frac{1}{2} & 1 \end{bmatrix}. \tag{2.58}$$

我们可以验证(2.58)确实是逆的，通过执行多乘法$\boldsymbol{A}\boldsymbol{A}^{-1}$并观察到我们恢复了$\boldsymbol{I}_4$。

2.3.4 解决线性方程组的算法

在下文中，我们将简要地讨论解决$\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$形式的线段方程组的方法。如果没有解，我们需要求助于近似解，这一点我们在本章中没有涉及。解决近似问题的一种方法是使用线性回归的方法，我们在第九章中详细讨论。

在特殊情况下，我们可能能够确定逆$\boldsymbol{A}^{-1}$，从而使$\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$的解被赋予$\boldsymbol{x}=^{-1}\boldsymbol{A}\boldsymbol{b}$。然而，这只有在$\boldsymbol{A}$是正方形矩阵且可反转的情况下才有可能，但这往往不是

的情况。否则，在温和的假设下（即，$\boldsymbol{A}$需要有线性独立的列），我们可以使用转换

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \iff {}^{\top}\boldsymbol{A}\boldsymbol{A}\boldsymbol{x} = {}^{\top}\boldsymbol{A}\boldsymbol{b} \iff \boldsymbol{x} = ({}^{\top}\boldsymbol{A}\boldsymbol{A})^{-1}{}^{\top}\boldsymbol{A}\boldsymbol{b} \tag{2.59}$$

并使用*摩尔-彭罗斯伪逆*（$\boldsymbol{A}^\top\boldsymbol{A}$）$^{-1}\boldsymbol{A}^\top$来确定摩尔-彭罗斯的解 (2.59)，解决了$\boldsymbol{Ax}=\boldsymbol{b}$，这也对应于最小规范的最小二乘解。这种方法的缺点是，它需要对矩阵-矩阵乘积进行多次计算，并且需要计算摄取$\boldsymbol{A}^\top\boldsymbol{A}$的逆值。此外，由于数字精度的原因，一般不建议计算逆或伪逆。

伪逆

因此，在下文中，我们将简要地讨论解决线性方程组的其他方法。

高斯消去法在计算解数（第4.1节）、检查一组向量是否是线性下垂的（第2.5)，计算矩阵的逆（第2.2.2节），计算矩阵的秩（第2.6.2)，以及确定一个向量空间的基（第2.6.1).高斯消除法是解决有数千个变量的线性方程组的一种直观和建设性的方法。然而，对于有数百万个变量的系统来说，它是不切实际的，因为所需的算术运算数量与同时发生的方程数量成立方体。

在实践中，许多线性方程组被间接地解决，如Richardson方法、Ja-cobi方法、Gauß-Seidel方法和逐次过度放松方法，或Krylov子空间方法，如共轭梯度、生成最小残差或双共轭梯度。我们参考Stoer和Burlirsch(2002)、Strang(2003)以及Liesen和Mehrmann(2015)的书中的进一步细节。

让$\boldsymbol{x}_*$是$\boldsymbol{Ax}=\boldsymbol{b}$的一个解。这些迭代方法的关键思想是建立一个迭代的形式

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{C}\boldsymbol{x}^{(k)} + \boldsymbol{d} \tag{2.60}$$

对于合适的$\boldsymbol{C}$和$\boldsymbol{d}$，在每次迭代中减少残余误差$\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}_*\|$，并收敛到$\boldsymbol{x}_*$。我们将在第二节介绍规范，它允许我们计算向量间的相似性。3.1.

2.4 矢量空间

到目前为止，我们已经研究了线性方程组和如何解决它们（第2.3).我们看到，线性方程组可以用矩阵-向量符号(2.10).在下文中，我们将仔细研究矢量空间，即矢量所在的结构化空间。

在本章开始时，我们非正式地将向量描述为可以加在一起并与标量相乘的对象，而且它们仍然是同一类型的对象。现在，我们准备将其正式化，我们将首先介绍群的概念，它是一个元素的集合，以及在这些元素上定义的、保持集合的某些结构不变的操作。

2.4.1 群体

群在计算机科学中发挥着重要作用。除了为集合的操作提供基本框架外，它们还被大量用于密码学、编码理论和图形学。

定义（2.7群）。考虑一个集合G和一个操作$\otimes: G \times G \to G$

组
封闭
关联性
中性元素 反向元素

那么$G := (G, \otimes)$被称为*群*，如果以下情况成立。

1. G在$\otimes$下的*封闭*：$\forall x, y \in G : x \otimes y \in G$
2. *关联性*：$\forall x, y, z \in G : (x \otimes y) \otimes z = x \otimes (y \otimes z)$
3. *中性元素*：$\exists e \in G \forall x \in G : x \otimes e = x$ and $e \otimes x = x$
4. *反元素* $\forall x \in G \exists y \in G : x \otimes y = e$ 和 $y \otimes x = e$，其中e是中性元素。
   我们经常写x^{-1}来表示x的逆元素。

*备注。*反元素是相对于操作$\otimes$而定义的。
并不一定意味着. [1] ◆

阿贝尔群

如果另外有$\forall x, y \in G : x \otimes y = y \otimes x$，那么$G = (G, \otimes)$是一个*阿贝尔群*（换元）。

例子（2.10群体）

让我们看看一些带有相关操作的集合的例子，看看它们是否是组。

- $(\mathbb{Z}, +)$是一个阿贝尔群。
- $(\mathbb{N}_0, +)$不是一个群。虽然$(\mathbb{N}_0, +)$拥有一个中性元素(0)，但缺少逆向元素。

$\mathbb{N}_0 := \mathbb{N} \cup \{0\}$

- $(\mathbb{Z}, -)$不是一个群。虽然（$\mathbb{Z}$，-）包含一个中性元素（1），但对于任何$z \in \mathbb{Z}$，$z \neq \pm 1$，都缺少逆向元素。
- $(\mathbb{R}, -)$不是一个群，因为它0不具备一个逆元素。
- $(\mathbb{R} \setminus \{0\}, -)$是阿贝尔的
- $(\mathbb{R}^n, +), (\mathbb{Z}^n, +), n \in \mathbb{N}$是阿贝尔的，如果+是分量定义，即。

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n). \tag{2.61}$$

  那么，$(x_1, \ldots, x_n)^{-1} := (-x_1, \ldots, -x_n)$是反元素，并且$e = (0, \ldots, 0)$是中性元素。
- $(\mathbb{R}^{m \times n}, +)$，$m \times n$矩阵的集合是阿贝尔的(与分量上的中定义的加法。2.61)).
- 让我们仔细看看$(\mathbb{R}^{n \times n}, -)$，即.,$n \times n$矩阵的集合，其中有矩阵乘法的定义为(2.13).
  - 闭合性和关联性直接来自于矩阵乘法的定义。
  - 中性元素。对于$(\mathbb{R}^{n \times n}, -)$中的矩阵乘法"-"，身份矩阵$\boldsymbol{I}_n$是中性元素。

- 逆元素。如果反向存在（$\boldsymbol{A}$是有规律的），那么$\boldsymbol{A}^{-1}$就是$\boldsymbol{A}\in\mathbb{R}$的$n\times n$逆元素，而在这种情况下，$(\mathbb{R}^{n\times n},\cdot)$是一个群，称为*一般线性群*。

定义（2.8一般线性群）。正则（可逆）矩阵的集合$A\mathbb{R}^{n\times n}$是一个关于矩阵乘法的群，$\in$即
定义在(2.13)，被称为*一般线性群*$GL(n,\mathbb{R})$。然而，一般线性群
因为矩阵乘法不是换元的，所以这个群不是阿贝尔群。

2.4.2 矢量空间

当我们讨论组的时候，我们看了集合和内部的操作。
在下文中，我们将考虑除了内运算+之外还包含外运算的集合，即向量$\boldsymbol{x}$与标量$\lambda\mathbb{R}$的乘法。请注意，内/外运算与内/外积没有关系。$\in$

定义（2.9矢量空间）。一个实值*向量空间*$V=(V,+,\cdot)$是向量空间
一个有两个操作的集合V

$$+:V\times V\to V \tag{2.62}$$
$$\cdot:\mathbb{R}\times V\to V \tag{2.63}$$

其中

1. $(V,+)$是一个阿贝尔群 2.分
布性。
   1.$\forall\lambda\in\mathbb{R},\boldsymbol{x},\boldsymbol{y}\in V:\lambda\cdot(\boldsymbol{x}+\boldsymbol{y})=\lambda\cdot\boldsymbol{x}+\lambda\cdot\boldsymbol{y}$
   2.$\forall\lambda,\psi\in\mathbb{R},\boldsymbol{x}\in V:(\lambda+\psi)\cdot\boldsymbol{x}=\lambda\cdot\boldsymbol{x}+\psi\cdot\boldsymbol{x}$
3. 关联性（外部操作）：$\forall\lambda,\psi\in\mathbb{R},\boldsymbol{x}\in V:\lambda\cdot(\psi\cdot\boldsymbol{x})=(\lambda\psi)\cdot\boldsymbol{x}$
4. 关于外部操作的中性元素：$\forall\boldsymbol{x}\in V:1\cdot\boldsymbol{x}=\boldsymbol{x}$

元素$\boldsymbol{x}\in V$被称为*向量*。$(\cdot,+)$的中性元素是向量
零向量$=[0,\ldots,0]^T$，内部操作+称为向量加法
加法。$\lambda\in\mathbb{R}$的元素被称为*标量*，外在操作scalar
$\cdot$是*标量的乘法*。请注意，标量乘法是指以标量为单位的乘法。
不同的，我们将在第二节中讨论这个问题。3.2.来处理

备注。一个 "向量乘法 "$\boldsymbol{ab}$，$\boldsymbol{a}$，$\boldsymbol{b}\in\mathbb{R}^n$，并没有定义。从理论上讲，我们可以定义一个从元素上看的乘法，如$\boldsymbol{c}=\boldsymbol{ab}$，$c_j=_j\ ab_j$。这种 "数组乘法 "在许多程序中是常见的。

在数学上，使用矩阵乘法的标准规则，意义有限。通过把向量当作$n\times$矩阵1

(我们通常这样做)，我们可以使用矩阵乘法，正如在(2.13).然而，这样一来，向量的尺寸就不匹配了。只有
外部产品 定义了以下向量的乘法：$\boldsymbol{a}\boldsymbol{b}^\top \in \mathbb{R}^{n\times n}$（*外积*），$^\top\boldsymbol{a}\boldsymbol{b}\in\mathbb{R}$（内积/标量/点积）。 ◆

例子（2.11向量空间）

让我们看一下一些重要的例子。

- $V=\mathbb{R}^n$，$n\in\mathbb{N}$是一个矢量空间，其操作定义如下。
  - 加法：$\boldsymbol{x}+\boldsymbol{y}=(x_1,\ldots,x_n)+(y_1,\ldots,y_n)=(x_1+y_1,\ldots,x_n+y_n)$ 对于所有的$\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$
  - 标量的乘法：$\lambda\boldsymbol{x}=\lambda(x_1,\ldots,x_n)=(\lambda x_1,\ldots,\lambda x_n)$ 对于所有$\lambda\in\mathbb{R}$，$\boldsymbol{x}\in\mathbb{R}^n$
- $V=\mathbb{R}^{m\times n}$，m，$n\in\mathbb{N}$是一个矢量空间，有
  - 加法。$\boldsymbol{A}+\boldsymbol{B}=\begin{bmatrix} a_{11}+b_{11} & \cdots & a_{1n}+b_{1n} \\ \vdots & & \vdots \\ a_{m1}+b_{m1} & \cdots & a_{mn}+b_{mn} \end{bmatrix}$ 被定义为对所有的$\boldsymbol{A}$，$\boldsymbol{B}\in V$
  - 标量的乘法：$\lambda\boldsymbol{A}=\begin{bmatrix} \lambda a_{11} & \cdots & \lambda a_{1n} \\ \vdots & & \vdots \\ \lambda a_{m1} & \cdots & \lambda a_{mn} \end{bmatrix}$ 中所定义的节 2.2.记住，$\mathbb{R}^{m\times n}$等同于$\mathbb{R}^{mn}$。
- $V=\mathbb{C}$，有复数加法的标准定义。

备注。 在下文中，我们将用V表示一个矢量空间（V，$+$，$\cdot$），当$+$和$\cdot$是标准的矢量加法和标量乘法时。此外，我们将使用符号$\boldsymbol{x}\in V$来表示V中的向量，以简化符号。 ◆

*备注。*矢量空间$\mathbb{R}^n$、$\mathbb{R}^{n\times 1}$、$\mathbb{R}^{1\times n}$只是在我们写矢量的方式上有所不同。在下文中，我们将不对$\mathbb{R}^n$和$\mathbb{R}$进行区分$^{n\times 1}$，这使得我们可以将$\boldsymbol{n}$*个*元组写成*列向量*
列向量

$$\boldsymbol{x}=\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \tag{2.64}$$

这简化了有关向量空间操作的符号。然而，我们确实区分了$\mathbb{R}^{n\times 1}$和$\mathbb{R}^{1\times n}$（*行向量*），以避免与矩阵乘法相混淆。默认情况下，我们用$\boldsymbol{x}$*表示*一个列向量，行向量用$\boldsymbol{x}$表示，$^\top$即$\boldsymbol{x}$的*转置*。◆
行向量转置

2.4.3 矢量子空间

在下文中，我们将介绍矢量子空间。直观地说，它们是包含在原始矢量空间中的集合，其特性是当我们对这个子空间中的元素进行矢量空间操作时，我们将永远不会离开它。在这个意义上，它们是"封闭的"。矢量子空间是机器学习的一个关键思想。例如，第十章展示了如何使用矢量子空间进行降维。

定义（2.10矢量子空间）。设$V = (\mathcal{V}, +, \cdot)$是一个矢量空间和$\mathcal{U} \subseteq \mathcal{V}, \mathcal{U} \neq \emptyset$。那么，如果$U$是一个矢量空间，其矢量空间操作$+$线性子空间和$\cdot$限于$\mathcal{U} \times \mathcal{U}$和$\mathbb{R} \times \mathcal{U}$，则$U = (\mathcal{U}, +, \cdot)$被称为$V$的矢量子*空间*（或矢量子空间*线性子空间*）。我们写$U \subseteq V$来表示一个子空间

如果$U \subseteq V$是一个矢量空间，那么U自然直接从V继承了许多特性，因为它们对所有的$\boldsymbol{x} \in \mathcal{V}$都是成立的，特别是对所有的$\boldsymbol{x} \in \mathcal{U} \subseteq \mathcal{V}$。这包括阿贝尔群的特性、分布性、关联性和中性元素。为了确定是否$(\mathcal{U}, +, \cdot)$是V的一个子空间，我们仍然需要证明

1. $\mathcal{U} \neq \emptyset$特别是。$\mathbf{0} \in \mathcal{U}$
2. U的关闭。
   a. 关于外部操作：$\forall \lambda \in \mathbb{R} \forall \boldsymbol{x} \in \mathcal{U} : \lambda \boldsymbol{x} \in \mathcal{U}$。
   b. 关于内部操作：$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{U} : \boldsymbol{x} + \boldsymbol{y} \in \mathcal{U}$。

例子（2.12矢量子空间）

让我们看一下一些例子。

- 对于每一个矢量空间V，琐碎的子空间是V本身和$\{\mathbf{0}\}$。
- 只有图中的例子D2.6是$\mathbb{R}$的一个子空间[2]（有通常的内/外操作）。在A和C中，封闭属性被违反；B不包含$\mathbf{0}$.
- 同质线性方程组的解集$\boldsymbol{A}\boldsymbol{x} = \mathbf{0}$有$n$个未知数的$\boldsymbol{x} = [x_1, \ldots, x_n]^\top$是$\mathbb{R}$的$n$一个子空间。
- 非均质线性方程组$\boldsymbol{A}\boldsymbol{x} =$的解$\boldsymbol{b}$，$\boldsymbol{b} \neq \mathbf{0}$不是$\mathbb{R}$的$n$一个子空间
- 任意多个子空间的交集本身就是一个子空间。

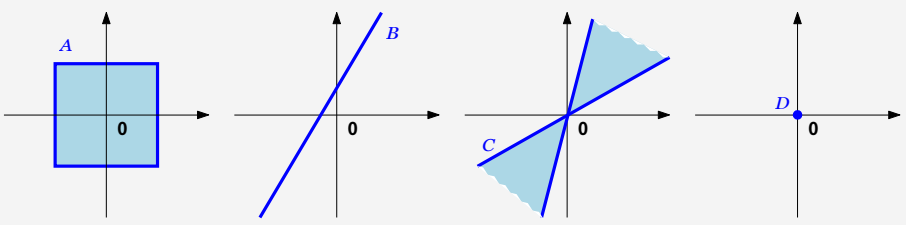


图2.6 不是所有$\mathbb{R}$的子集[2]都是子空间。在A和C中，封闭属性被违反；B不包含$\mathbf{0}$.只有D是一个

子空间。

备注。 每个子空间$U \subseteq (\mathbb{R}^n, +, \cdot)$都是同基因线性方程组$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$的解空间，对于$\boldsymbol{0} \in \mathbb{R}^n$。

2.5 线性独立

在下文中，我们将仔细研究我们能用向量（向量空间的元素）做什么。特别是，我们可以将向量相加，并与标量相乘。闭合属性保证我们最终在同一向量空间中得到另一个向量。有可能找到一组向量，我们可以用这组向量来代表向量空间中的每一个向量，把它们加在一起并按比例计算。这组向量是一个*基数*，我们将在第5节讨论它们。2.6.1.在这之前，我们需要介绍一下线性组合和线性独立的概念。

定义（2.11线性组合）。 考虑一个向量空间V和有限数量的向量$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in V$。那么，每个$\boldsymbol{v} \in V$的形式为

$$\boldsymbol{V} = \lambda_1 \boldsymbol{x}_1 \quad \text{--- ++ } \lambda_k \boldsymbol{x} = \sum_{i=1}^{k} {}_k \lambda_{ii} \quad xV \qquad (2.65)$$

线性组合

与$\lambda_1, \ldots, \lambda_k \in \mathbb{R}$是矢量$\boldsymbol{x}_1, \ldots \ldots \in$的*线性组合。*$, \boldsymbol{x}_k$.

$\boldsymbol{0}$向量总是可以被写成$\boldsymbol{k}$个向量$\boldsymbol{x}$, . . .因为$= \boldsymbol{0}\boldsymbol{k}\boldsymbol{x}0$总是真的。在下文中。

我们对^{1}代表$\boldsymbol{0}$一^{k}组向量的非线性组$_{i=}$合感$_i$兴趣，即向量$\boldsymbol{x}_1, \ldots \ldots$的线性组合。$, \boldsymbol{x}_k$的线性组合，其中并非所有系数$\boldsymbol{\lambda}_i$在(2.65)中的所有系数都是0.

定义（2.12线性（内）依赖性）。 让我们考虑一个矢量空间V的$k \in \mathbb{N}$和$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in V$。如果存在一个非显著的线性组合，那么就有一个非显著的线性组合。搭配，这样，$= \boldsymbol{0}$ $\sum_{i=1}^{k} \lambda \boldsymbol{x}_i$ 至少有一个$\lambda_i \neq 0$，向量

线性依赖

$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$是*线性依赖的。*如果只有琐碎的解存在，即。

线性独立

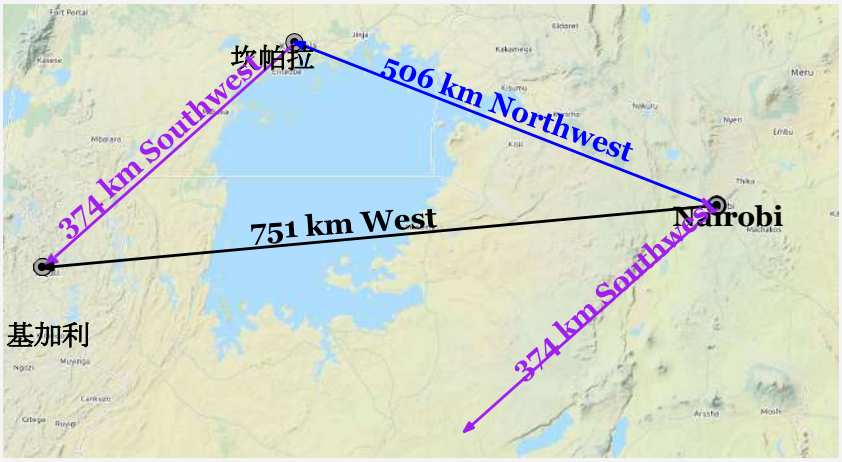
$\lambda_1 = \ldots = \lambda_k = 0$矢量$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$是*线性独立的。*

线性独立是线性代数中最重要的概念之一。直观地说，一组线性独立的向量由没有冗余的向量组成，也就是说，如果我们从这组向量中删除任何一个，我们就会失去一些东西。在接下来的章节中，我们将更多地把这个直觉正式化。

例子 (2.13线性依赖的矢量)

一个地理上的例子可能有助于澄清线性独立的概念。一个在内罗毕（肯尼亚）的人在描述基加利（卢旺达）的位置时可能会说，"你可以通过先向西北方向走506公里到Kam- pala（乌干达），然后374向西南方向走几公里来到达基加利"。这就是足够的信息

来描述基加利的位置，因为地理坐标系统可以被认为是一个二维的矢量空间（忽略了高度和地球的弯曲表面）。这个人可能会补充说："它在这里以西大约一公里751"。虽然最后这句话是真的，但鉴于前面的信息，没有必要找到基加利（见图2.7图中有一个说明）。在这个例子中，"西北506公里 "向量（蓝色）和 "西南374公里 "向量（紫色）是线性独立的。这意味着西南方向的矢量不能用西北方向的矢量来描述，反之亦然。然而，第三个 "751西部公里 "矢量（黑色）是其他两个矢量的线性组合，它使这组矢量具有线性依赖性。等价地，给定的 "751西部公里 "和 "374西南公里 "可以线性组合，得到 "506西北公里"。



图：地理学**2.7**上的例子（有粗略的心轴方向的近似值），在一个线性依赖的向量中的二维空间（平面）。

*备注。*以下属性对找出向量是否线性独立很有用。

- k个向量要么是线性依赖，要么是线性独立。没有第三种选择。
- 如果至少有一个矢量$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$是，那么$\mathbf{0}$它们就是线性去垂线的。如果两个向量是相同的，也是如此。
- 矢量$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$: $\boldsymbol{x}_i = \mathbf{0}, i = 1, \ldots, k, k \geqslant 2$，是线性依赖的，当且仅当（至少）其中一个是线性组合的时候的倍数。特别是，如果一个向量是另一个向量的倍数，即$\boldsymbol{x}_i = \lambda \boldsymbol{x}_j, \lambda \in \mathbb{R}$，那么集合$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k : \boldsymbol{x}_i = \mathbf{0}, i = 1, \ldots, k\}$是线性依赖的。
- 检查向量$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$是线性独立的是使用高斯消除法将所有的向量写成矩阵$\boldsymbol{A}$的列，然后进行高斯消除，直到矩阵为行梯形形式（这里不需要缩小的行梯形形式）。

- 支点列表示向量，它们与左边的向量是线性独立的。注意，在建立矩阵时，有一个矢量的排序。
- 非支点列可以表示为其左侧支点列的线性组合。例如，行-歇尔形式

$$\begin{matrix} 1 & 3 & (20 \\ 0 & 0 & 2 \end{matrix} \qquad .66)$$

告诉我们，第一列和第三列是枢纽列。第二列是一个非枢纽列，因为它是第一列的三倍。

所有的列向量都是线性独立的，当且仅当所有的列都是支点列。如果至少有一个非支点列，那么这些列（以及相应的向量）是线性独立的。

◆

例子 2.14
考虑$\mathbb{R}^4$与

$$\boldsymbol{x}_1 = \begin{matrix} 2 \\ -3 \\ 4 \end{matrix}, \quad \boldsymbol{x}_2 = \begin{matrix} 1 \\ 0 \\ 2 \end{matrix}, \quad \boldsymbol{x}_3 = \begin{matrix} -2 \\ 1 \\ 1 \end{matrix}. \qquad (2.67)$$

为了检查它们是否是线性相关的，我们遵循一般的方法，求解

$$\lambda_1\boldsymbol{x} + \lambda\boldsymbol{x}_2 + \lambda\boldsymbol{x}_{23} = \lambda \begin{matrix} 2 \\ -3 \\ 4 \end{matrix} + \lambda_2 \begin{matrix} 1 \\ 0 \\ 2 \end{matrix} + \lambda_3 \begin{matrix} -2 \\ 1 \\ 1 \end{matrix} = (\mathbf{0} \quad 2.68)$$

对于$\lambda_1, \ldots$我们把向量$\boldsymbol{x}_i, i = 1, 2, 3,$写成矩阵的列，并应用基本的行操作，直到我们确定支点列$_3$。

$$\begin{matrix} 2 & 11-1 \\ -3 & 1-2 \\ & 0 \\ & 1 \\ & 421 \end{matrix} \quad \text{-- —} \quad \begin{matrix} 1 & 1 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{matrix} \quad . \qquad (2.69)$$

这里，矩阵的每一列都是支点列。因此，不存在非琐碎的解决方案，我们要求$\lambda_1=0$，$\lambda_2=0$，$\lambda_3=0$来解决这个方程组。因此，向量$\boldsymbol{x}_1$、$\boldsymbol{x}_2$、$\boldsymbol{x}_3$是线性独立的。

*备注。*考虑一个具有k个线性独立向量的向量空间V
$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k$和$m$个线性组合

$$\boldsymbol{x} = \sum_{i=1}^{k} \lambda_{1i}\boldsymbol{b}_i$$
$$\vdots \tag{2.70}$$
$$\boldsymbol{x} = \sum_{i=1}^{k} \lambda_{mi}\boldsymbol{b}_i \,。$$

定义$\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k]$为矩阵，其列是线性独立的向量$\boldsymbol{b}_1, \ldots, b_。, \boldsymbol{b}_k$，我们可以写成

$$\boldsymbol{x}_j = \boldsymbol{B}\boldsymbol{\lambda}_j, \quad \boldsymbol{\lambda}_j = \begin{bmatrix} \lambda_{1j} \\ \vdots \\ \lambda_{kj} \end{bmatrix}, \quad j = 1, \ldots, m, \tag{2.71}$$

以一种更紧凑的形式。

我们想测试$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$是线性独立的。对于这个目的，我们遵循一般的方法，当$\sum_{j=1}^{m} \psi_j \boldsymbol{x}_j = \boldsymbol{0}$.

通过(2.71)，我们得到

$$\sum_{j=1}^{m} \psi_j \boldsymbol{x}_j = \sum_{j=1}^{m} \psi_j \boldsymbol{B}\boldsymbol{\lambda}_j = \boldsymbol{B}\sum_{j=1}^{m} \psi_j \boldsymbol{\lambda}_j 。 \tag{2.72}$$

这意味着，$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$是线性独立的，当且仅当列向量$\{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m\}$是线性独立的。

◆

*备注。*在一个矢量空间V中，k个矢量$\boldsymbol{x}_1, \ldots$的m个线性组合。$, \boldsymbol{x}_k$是线性相关的，如果$m>k$。 ◆

例子 2.15

考虑一组线性独立向量$\boldsymbol{b}_1$，$\boldsymbol{b}_2$，$\boldsymbol{b}_3$，$\boldsymbol{b}_4 \in \mathbb{R}^n$和

$$\begin{aligned} \boldsymbol{x}_1 &= \boldsymbol{b}_1 - 2\boldsymbol{b}_2 + \boldsymbol{b}_3 - \boldsymbol{b}_4 \\ \boldsymbol{x}_2 &= -4\boldsymbol{b}_1 - 2\boldsymbol{b}_2 + 4\boldsymbol{b}_4 \\ \boldsymbol{x}_3 &= 2\boldsymbol{b}_1 + 3\boldsymbol{b}_2 - \boldsymbol{b}_3 - 3\boldsymbol{b}_4 \\ \boldsymbol{x}_4 &= 17\boldsymbol{b}_1 - 10\boldsymbol{b}_2 + 11\boldsymbol{b}_3 + \boldsymbol{b}_4 \end{aligned} \,. \tag{2.73}$$

矢量$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_4 \in \mathbb{R}^n$是线性独立的吗？为了回答这个问题问题，我们调查了列向量是否

$$\begin{bmatrix} 1 \\ -2 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} -2 \\ 3 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 2 \\ -1 \end{bmatrix}, \begin{bmatrix} 17 \\ -10 \\ 11 \\ 1 \end{bmatrix} \tag{2.74}$$

是线性独立的。相应的线性方程组的减行-减值形式，其系数矩阵为

$$\boldsymbol{A} = \begin{bmatrix} 1 & -4 & 7 & 21 \\ -2 & -2 & 3 & -10 \\ 1 & 0 & -1 & 11 \\ -1 & 4 & -3 & 1 \end{bmatrix} \tag{2.75}$$

给出的结果是

$$\begin{bmatrix} 1 & 0 & 0 & -7 \\ 0 & 1 & 0 & -15 \\ 0 & 0 & 1 & -18 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{2.76}$$

我们看到，相应的线性方程组是非暴力解决的
能。最后一列不是支点列，且$\boldsymbol{x}_4 = -7\boldsymbol{x}_1 - 15\boldsymbol{x}_2 - 18\boldsymbol{x}_3$。
因此，$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_4$是线性依赖的，因为$\boldsymbol{x}_4$可以表示为$\boldsymbol{x}_1, \ldots \ldots$的线性组合。$, \boldsymbol{x}_3$.

2.6 基准和等级

在一个矢量空间V中，我们特别感兴趣的是那些拥有以下特性的矢量集：任何矢量$\boldsymbol{v} \in V$都可以由其中的矢量线性组合得到。这些向量是特殊的向量，在下文中，我们将描述它们的特征。

础

生成集跨度

最低限

度的基

可以表示为$\boldsymbol{x}_1, \ldots\ldots$的线性组合。，$\boldsymbol{x}_k$，称为$V$的*生成集*和*基数*

2.6.1 生成集和基数

考虑一个矢量空间$V = (\mathcal{F}, +, \cdot)$和矢量集$\mathcal{A} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\} \subseteq V$。如果$V$中的所有线性组合的集合是$V$，则称$\mathcal{A}$为$V$的*生成集*。$V$的生成集的集合称为其*跨度*。如果$\mathcal{A}$跨越矢量空间$V$，我们写$V = \text{span}[\mathcal{A}]$或$V = \text{span}[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k]$。

生成集是跨越向量（子）空间的向量集，也就是说，每个向量都可以表示为生成集中的向量的线性组合。现在，我们将更具体地描述跨越一个向量（子）空间的最小生成集的特征。

定义（2.14基数）。考虑一个矢量空间$V = (\mathcal{F}, +, \cdot)$和$\mathcal{A} \subseteq V$。如果不存在更小的集合，那么$\mathcal{A}$的生成集被称为*最小*。

定义（2.13生成集和跨度）。考虑一个矢量空间$V = (\mathcal{F}, +, \cdot)$和矢量集$\mathcal{A} = \{\boldsymbol{x}_1, \ldots$如果每个向量$\boldsymbol{v}$跨越了$V$。$V$的每个线性独立生成集是最小的，被称为$V$的*基*。

设$V=(V,+,\cdot)$是一个向量空间，$B\subseteq V$，$B=\emptyset$。 那么，在以下说法是等价的：

一个基础是一个最小生成集，一个最大的线性独立向量集。

- B是V的一个基。
- B是一个最小生成集。
- B是V中最大的线性独立的向量集，也就是说，向这个集子添加任何其他的向量都会使其成为线性依赖。
- 每个向量$x\in V$都是来自，的向量的线性组合，每个线性组合都是唯一的，即，有 $\in B$

$$x=\sum_{i=1}^{k}\lambda_i b_i=\sum_{i=1}^{k}\psi_i b_i \tag{2.77}$$

而$\lambda_i,\ \psi_i\in R,\ b_i\in B$，由此可知，$\lambda_i=\psi_i,\ i=1,\ \ldots,\ k$。$,k$.

例子 2.16

典型的基础

- 在R^3中，*典范/标准基础*是

$$B=\left\{\begin{bmatrix}1\\0\\0\end{bmatrix},\begin{bmatrix}0\\1\\0\end{bmatrix},\begin{bmatrix}0\\0\\1\end{bmatrix}\right\} \tag{2.78}$$

- R^3中的不同基数是

$$b_1=\left\{\begin{bmatrix}1\\0\\0\end{bmatrix},\begin{bmatrix}1\\1\\0\end{bmatrix},\begin{bmatrix}1\\1\\1\end{bmatrix}\right\},\quad b_2=\left\{\begin{bmatrix}0.8\\0.3\\-1.3\end{bmatrix},\begin{bmatrix}0.5\\0.8\\0.4\end{bmatrix},\begin{bmatrix}1.8\\-2.2\\3.5\end{bmatrix}\right\} \tag{2.79}$$

- 这一套

$$A=\left\{\begin{bmatrix}\frac{1}{2}\\2\\0\end{bmatrix},\begin{bmatrix}1\\-\frac{3}{2}\\4\end{bmatrix},\begin{bmatrix}1\\0\\-4\end{bmatrix}\right\} \tag{2.80}$$

是线性独立的，但不是R^4生成集（也不是基）。
例如，向量$[1,0,0,0]^T$不能由A中元素的线性组合得到。

备注。每个向量空间V都有一个基。前面的例子B表明，一个向量空间V可以有许多基，也就是说，没有唯一的基。然而，所有基都拥有相同数量的元素。

◆ 基数向量

的*基向量*。

我们只考虑有限维向量空间V。在这种情况下，
V的*维度*是V的基向量的数量，我们写成$\dim(V)$。
如果$U \subseteq V$是V的一个子空间，那么$\dim(U) \leqslant \dim(V)$和$\dim(U) =$

dim(*V*)，当且仅当*U*=*V*。直观地说，一个矢量空间的维度可以被认为是这个矢量空间中独立方向的数量。

矢量空间的维度对应于其基向量的数量。

*备注。*矢量空间的维数不一定是指的元素的向量。例如，矢量空间$V = \text{span}[\begin{bmatrix}0\\1\end{bmatrix}]$是一维的，尽管基向量拥有两个元素。◆

*备注。*子空间$U = \text{span}[\boldsymbol{x}_1, \ldots, \boldsymbol{x}] \subseteq {}_m\mathrm{R}$的^{n}基可以通过执行以下步骤找到。

1. 把跨度向量写成矩阵$\boldsymbol{A}$的列
2. 确定$\boldsymbol{A}$的行-歇尔形式。
3. 与支点列相关的跨度向量是一个基础，即 U.

◆

例子（2.17确定一个基础）
对于一个矢量子空间$U \subseteq \mathrm{R}^5$，由矢量跨越的是

$$\boldsymbol{x}_1 = \begin{bmatrix}1\\2\\-1\\-1\\-1\end{bmatrix}, \quad \boldsymbol{x}_2 = \begin{bmatrix}2\\-1\\1\\2\\-2\end{bmatrix}, \quad \boldsymbol{x}_3 = \begin{bmatrix}3\\-4\\3\\5\\-3\end{bmatrix}, \quad \boldsymbol{x}_4 = \begin{bmatrix}-1\\8\\-5\\-6\\1\end{bmatrix} \in \mathrm{R}^5, \tag{2.81}$$

我们感兴趣的是找出哪些向量$\boldsymbol{x}_1, \ldots$.为此，我们需要检查$\boldsymbol{x}_{41}, \ldots$.*为此，我们需要检查*$x, \ldots, \boldsymbol{x}_4$是否是线性独立的。因此，我们需要解决

$$\sum_{i=1}^{4} \lambda_i \boldsymbol{x} = \boldsymbol{0}, \tag{2.82}$$

这导致了一个同质方程组，矩阵为

$$[\boldsymbol{X}_2, \boldsymbol{X}_1, \boldsymbol{X}_4, \boldsymbol{X}] = \begin{bmatrix}1 & 2 & 3 & -1\\2 & -1 & - & \\-1 & 1 & 3 & 54\\-1 & 2 & 85 & -6\\-1 & -2 & - & \end{bmatrix}. \tag{2.83}$$

利用线性方程组的基本转换规则，我们可以得到行-歇尔的形式

$$\begin{bmatrix}1 & & 23-1\\2 & -1 & -\\-1 & 1 & 3-5\ 4\\-1 & 2 & 8\ 5-6\\1- & -2 & -\end{bmatrix} \text{--} \begin{bmatrix}1 & 2 & 3-1\\0 & 1 & 2-2\\0 & 0 & 01\\0 & 0 & 0\\0 & 0 & 0\end{bmatrix}.$$

由于枢轴列表明哪一组向量是线性不分离的，我们从行-歇尔形式中看到，$\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_4$是线性不分离的。

垂线（因为线性方程组$\lambda_1\boldsymbol{x}_1 + \lambda_2\boldsymbol{x}_2 + \lambda_4\boldsymbol{x}_4 = \boldsymbol{0}$只能用$\lambda_1 = \lambda_2 = \lambda_4 = 0$来求解。因此，$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_4\}$是一个$U$的基础。

2.6.2 级别

矩阵$\boldsymbol{A} \in \mathbb{R}^{m\times n}$的线性独立列的数量等于线性独立行的数量，称为$\boldsymbol{A}$的秩，用$\operatorname{rk}(\boldsymbol{A})$表示。

*备注。*矩阵的等级有一些重要的属性。

- $\operatorname{rk}(\boldsymbol{A}) = \operatorname{rk}(\boldsymbol{A}^\top)$，即列级等于行级。
- $\boldsymbol{A} \in \mathbb{R}^{m\times n}$横跨一个子空间$U \subseteq \mathbb{R}^m$，$\dim(U) = \operatorname{rk}(\boldsymbol{A})$。以后我们将称这个子空间为*图像*或*范围*。一个基的U可以通过对$\boldsymbol{A}$应用高斯消除法来确定支点列来找到。
- $\boldsymbol{A} \in \mathbb{R}^{m\times n}$行横跨一个子空间$W \subseteq \mathbb{R}^n$，$\dim(W) = \operatorname{rk}(\boldsymbol{A})$。通过对$\boldsymbol{A}^\top$进行高斯消除，可以找到$W$的一个基。
- 对于所有的$\boldsymbol{A} \in \mathbb{R}^{n\times n}$来说，当且仅当$\operatorname{rk}(\boldsymbol{A}) = n$时，$\boldsymbol{A}$是有规律的（可倒置的）。
- 对于所有$\boldsymbol{A} \in \mathbb{R}^{m\times n}$和所有$\boldsymbol{b} \in \mathbb{R}^m$，当且仅当$\operatorname{rk}(\boldsymbol{A}) = \operatorname{rk}(\boldsymbol{A}|\boldsymbol{b})$时，线性方程组$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$可以被解决，其中$\boldsymbol{A}|\boldsymbol{b}$表示增强的系统。
- 对于$\boldsymbol{A} \in \mathbb{R}^{m\times n}$，$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$的解的子空间拥有$n - \operatorname{rk}(\boldsymbol{A})$。以后，我们将称这个子空间为*内核*或*空核空间*。
- 如果一个矩阵$\boldsymbol{A} \in \mathbb{R}^{m\times n}$的秩等于最大可能的相同维度的矩阵的等级。这意味着一个全等级矩阵的行数和列数的较小者，即。$\operatorname{rk}(\boldsymbol{A}) = \min(m, n)$。如果一个矩阵不存在有完整的等级。

◆

例子(2.18排名)

- $\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$.

  $\boldsymbol{A}$有两个线性独立的行/列，所以$\operatorname{rk}(\boldsymbol{A}) = 2$.

- $\boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ -2 & -3 \\ 3 & 0 \end{bmatrix}$.

我们使用高斯消除法来确定等级。

$$\begin{bmatrix} 1 & 1 \\ -2 & -3 \\ 3 & 0 \end{bmatrix} \rightsquigarrow \cdots \rightsquigarrow \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \tag{2.84}$$

在这里，我们看到，线性独立的行和列的数量是这样2,的：rk($\boldsymbol{A}$)= 2.

2.7 线性映射

在下文中，我们将研究矢量空间上保留其结构的映射，这将使我们能够定义坐标的概念。在本章开始时，我们说过，向量是可以加在一起并与标量相乘的对象，所得对象仍然是一个向量。我们希望在应用映射时保留这一属性。考虑两个实向量空间V，W。一个映射$\Phi : V \to W$保留了向量空间的结构，如果

$$\Phi(\boldsymbol{x}+\boldsymbol{y}) = \Phi(\boldsymbol{x}) + \Phi(\boldsymbol{y}) \tag{2.85}$$
$$\Phi(\lambda\boldsymbol{x}) = \lambda\Phi(\boldsymbol{x}) \tag{2.86}$$

对于所有的$\boldsymbol{x}, \boldsymbol{y} \in V$和$\lambda \in \mathbb{R}$，我们可以用以下定义来概括。

线性映射 矢量空间同态性 线性转换

定义（2.15线性映射）。对于矢量空间V，W，一个映射$\Phi : V \to W$被称为*线性映射*（或*矢量空间同构*/*线性变换*），如果

$$\forall \boldsymbol{x}, \boldsymbol{y} \in V, \lambda, \psi \in \mathbb{R} : \Phi(\lambda\boldsymbol{x} + \psi\boldsymbol{y}) = \lambda\Phi(\boldsymbol{x}) + \psi\Phi(\boldsymbol{y}) \text{。} \tag{2.87}$$

事实证明，我们可以用矩阵来表示线性映射（第二节2.7.1).回顾一下，我们也可以将一组向量收集为矩阵的列。在处理矩阵时，我们必须记住矩阵代表什么：是线性映射还是向量集合。我们将在第四章看到更多关于线性映射的内容。在我们继续之前，我们将简单介绍一下特殊映射。

定义（2.16射入、射出、双射）。考虑一个映射$\Phi : V \to W$，其中V，W可以是任意的集合。那么Φ就被称为

注入式、射出式、双射式

- 如果$\forall \boldsymbol{x}, \boldsymbol{y} \in V$则为$\Phi(\boldsymbol{x}) = \Phi(\boldsymbol{y}) \Rightarrow \boldsymbol{x} = \boldsymbol{y}$*注入式*。
- 如果$\Phi(\ ) =$，则是*射影的*。

如果它是注入性和抛射性的，则是*双射性的*。

如果Φ是射影的，那么每个元素 W inc都可以用Φ来 "到达"。一个双射的Φ可以被 "撤消"，也就是说，存在一个映射Ψ：$W \to V$ 这个映射Ψ被称为Φ的逆映射，通常用Φ^{-1}表示。

有了这些定义，我们介绍以下线性的特殊情况 V 向量空间V和W之间的映射。

同构性
内形态
同构

- *同构性*。$\Phi : V \to W$的线性和双射性
- *内形态*。$\Phi : V \to V$线性
- *自动变形*。$\Phi : V \to V$的线性和双射性
- 我们定义$\mathrm{id}_V : V \to V, \boldsymbol{x} \mapsto \boldsymbol{x}$为*身份映射*或*身份*

同一性映射
身份自定型

V中的*自动变形*。

例子（2.19同态性）

映射$\Phi : \mathbb{R}^2 \to \mathbb{C}, \Phi(\boldsymbol{x}) = x_1 + ix_2$，是一个同态的。

$$\begin{aligned}\Phi\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}+\begin{bmatrix}y_1\\y_2\end{bmatrix}\right) &= (x_1+y_1)+i(x_2+y_2) = x_1+ix_2+y_1+iy_2\\ &= \Phi\begin{bmatrix}x_1\\x_2\end{bmatrix}+\Phi\begin{bmatrix}y_1\\y_2\end{bmatrix}\\ \Phi\left(\lambda\begin{bmatrix}x_1\\x_2\end{bmatrix}\right) &= \lambda x_1+\lambda i x_2 = \lambda(x_1+ix_2) = \lambda\Phi\begin{bmatrix}x_1\\x_2\end{bmatrix}.\end{aligned} \tag{2.88}$$

这也证明了为什么复数可以在$\mathbb{R}$中被表示为图元^2。有一个双射的线性映射，可以将$\mathbb{R}$中图元的加法转换^{2}为具有对应加法的复数集。请注意，我们只展示了线性，但没有展示双射。

定理（2.17 Axler（2015）中3.59的定理）。*当且仅当*$\dim(V) = \dim(W)$时，*有限维向量空间V和W才是同构的。*

该定理2.17指出，在两个相同维度的向量空间之间存在一个线性的、双射的映射。直观地说，这意味着相同维度的向量空间是同一种东西，因为它们可以相互转换而不产生任何损失。

该定理2.17也使我们有理由将 $\mathbb{R}^{m\times n}$（ mn的向量空间）和$\mathbb{R}^{mn}$（长度为1.5米的向量的向量空间）处理为一个整体 。

mn）相同，因为它们的维度是mn，而且存在一个线性的、双向的映射，将一个转化为另一个。

*备注。考虑*向量空间V，W，X，那么。

- 对于线性映射$\Phi :$ $\to VW$和$\Psi :$ WX，映射

- 如果$\Phi: V \to W$是一个同构，那么$\Phi^{-1}: W \to V$也是一个同构。

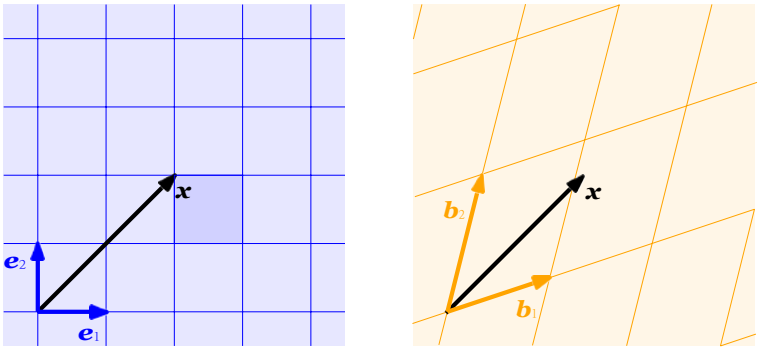


图 由两组基向量定义的两个2.8不同的坐标系。一个矢量$\boldsymbol{x}$有不同的坐标表示取决于选择哪个坐标系。

- 如果$\Phi: V \to W$, $\Psi: V \to W$是线性的，那么$\Phi + \Psi$和$\lambda\Phi$, $\lambda \in \mathbb{R}$, 也是线性的。

◆

2.7.1 线性映射的矩阵表示

任何n维向量空间都与$\mathbb{R}^n$同构（定理2.17).我们考虑一个基础$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$的$n$维向量空间$V$。在下文中，基向量的顺序将是重要的。因此，我们写成

$$B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n) \tag{2.89}$$

有序的基础

并称这个n元组为V的*有序基础*。

备注（记号）。我们正处于符号化变得有点棘手的阶段。因此，我们在此总结了一些部分。$B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)$ 是一个有序的基，$\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n\}$是一个（无序的）基础，而$\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n]$ 是一个矩阵，其列是向量$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$. ◆

定义2.18（坐标）。考虑一个矢量空间V和V的有序基B=（$\boldsymbol{b}_1$，...，$\boldsymbol{b}_n$）。对于任何$\boldsymbol{x} \in V$，我们得到一个唯一的代表（线性组合）

$$\boldsymbol{x} = \alpha_1\boldsymbol{b}_1 + \ldots + \alpha_n\boldsymbol{b}_n \tag{2.90}$$

协调

那么$\alpha_1, \ldots, \alpha_n$是$\boldsymbol{x}$相对于$B$的*坐标*，而矢量

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \in \mathbb{R}^n \tag{2.91}$$

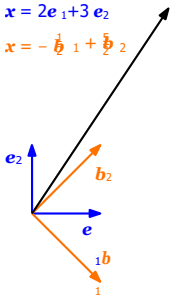
坐标向量坐标表示

是$\boldsymbol{x}$的*坐标矢量/坐标表示*，相对于有序基B。

一个基点有效地定义了一个坐标系。我们熟悉二维的直角坐标系，它是由典型的基向量$\boldsymbol{e}_1$，$\boldsymbol{e}_2$跨越的。在这个坐标系中，一个向量$\boldsymbol{x} \in \mathbb{R}^2$有一个表示，告诉我们如何将$\boldsymbol{e}_1$和$\boldsymbol{e}_2$线性地结合起来得到$\boldsymbol{x}$。然而，$\mathbb{R}^2$的任何基都定义了一个有效的坐标系，以前的同一个向量$\boldsymbol{x}$可能有一个不同的坐标代表
在$(\boldsymbol{b}_1, \boldsymbol{b}_2)$的基础上进行表达。在图2.8中，$\boldsymbol{x}$的坐标相对于标准基（$\boldsymbol{e}_1$，$\boldsymbol{e}_2$）是$[2, 2]^\top$。然而，相对于$(\boldsymbol{b}_1, \boldsymbol{b}_2)$基，同一个向量$\boldsymbol{x}$被表示为$[1.09, 0.72]^\top$，即$\boldsymbol{x} = 1.09\boldsymbol{b}_1 + 0.72\boldsymbol{b}_2$。在下面的章节中，我们将发现如何获得这种表示。

例子 2.20
让我们来看看一个坐标为$[2, 3]$的几何向量$\boldsymbol{x} \in \mathbb{R}^2$。
关于$\mathbb{R}^2$的标准基础（$\boldsymbol{e}_1$，$\boldsymbol{e}_2$），这意味着，我们可以写出
$\boldsymbol{x} = 2\boldsymbol{e}_1 + 3\boldsymbol{e}_2$。然而，我们不必选择标准基础来
代表这个向量。如果我们使用基础向量$\boldsymbol{b}_1 = [1, -1]^\top$，$\boldsymbol{b}_2 = [1, 1]^\top$。
我们将得到坐标$[-\frac{1}{2}, \frac{5}{2}]^\top$来表示同一个矢量的
关于（$\boldsymbol{b}_1$，$\boldsymbol{b}_2$）（见图2.9）。

图 2.9
不同的坐标表示的一个向量，取决于关于基础的选择。



*备注。*对于一个n维向量空间V和V的有序基B，映射$\Phi : \mathbb{R}^n \to V$，$\Phi(\boldsymbol{e}_i) = \boldsymbol{b}_i$，$i = 1, \ldots, n$，是线性的（并且由于定理2.17同构），其中（$\boldsymbol{e}_1$，…，$\boldsymbol{e}_n$）是$\mathbb{R}^n$的标准基。

◆

现在我们准备在矩阵和有限维向量空间之间的线性映射之间建立明确的联系。

定义（2.19变换矩阵）。考虑矢量空间V，W与相应的（有序）基$B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)$和$C = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m)$。此外，我们考虑一个线性映射$\Phi : V \to W$。对于$j \in \{1, \ldots, n\}$，

$$\Phi(\boldsymbol{b}_j) = \alpha_{1j}\boldsymbol{c}_1 + \cdots + \alpha_{mj}\boldsymbol{c}_m = \sum_{i=1}^{m} \alpha_{ij}\boldsymbol{c}_i \tag{2.92}$$

是$\Phi(\boldsymbol{b}_j)$相对于C的唯一表示，那么，我们称$m \times n$矩阵$\boldsymbol{A}_\Phi$，其元素由以下公式给出

$$A_\Phi(i, j) = \alpha_{ij}, \tag{2.93}$$

Φ的*变换矩阵*（相对于Vtransformation的有序基B*而言*和W的C）。

基体

$\Phi(\boldsymbol{b}_j)$相对于W的有序基C的坐标是$\boldsymbol{A}$的Φ第j列。考虑（有限维）矢量空间V，W的有序基B，C和线性映射$\Phi : V \to W$，其中有

如果$\hat{\boldsymbol{x}}$是$\boldsymbol{x}$ V相对于B的坐标向量，$\hat{\boldsymbol{y}}$是$\boldsymbol{y} = \Phi(\boldsymbol{x})$ W相对于C的坐标向量，那么，

$$\hat{\boldsymbol{y}} = \boldsymbol{A}\hat{\boldsymbol{x}} \text{ 。} \tag{2.94}$$

这意味着变换矩阵可以用来将相对于V中的有序基础的坐标映射到相对于W中的有序基础的坐标。

例子(2.21变换矩阵)
考虑一个同态性$\Phi : V \to W$和有序基$B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_3)$的$V$和$C = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_4)$的$W$。随着

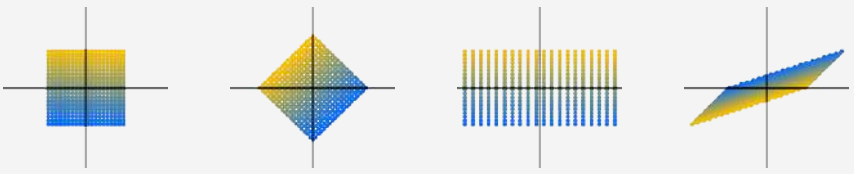
$$\begin{aligned}\Phi(\boldsymbol{b}_1) &= \boldsymbol{c}_1 - \boldsymbol{c}_2 + 3\boldsymbol{c}_3 - \boldsymbol{c}_4\\ \Phi(\boldsymbol{b}_2) &= 2\boldsymbol{c}_1 + \boldsymbol{c}_2 + 7\boldsymbol{c}_3 + 2\boldsymbol{c}_4\\ \Phi(\boldsymbol{b}_3) &= 3\boldsymbol{c}_2 + \boldsymbol{c}_3 + 4\boldsymbol{c}_4\end{aligned} \tag{2.95}$$

矩阵$\boldsymbol{A}_\Phi$相对于B和C的变换满足$\Phi(\boldsymbol{b}_k) = \sum_{i=1}^{4} \alpha_{ik}\boldsymbol{c}_i$为$k=1, \ldots$并3给定为

$$\boldsymbol{A}_\Phi = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3] = \begin{bmatrix} 1 & 2 & 0 \\ -1 & 1 & 3 \\ 3 & 7 & 1 \\ -1 & 2 & 4 \end{bmatrix}, \tag{2.96}$$

其中$\boldsymbol{\alpha}_j$，$j=1$，2，3，是$\Phi(\boldsymbol{b}_j)$相对于C的坐标向量。

例2.22（矢量的线性变换）



(a) 原始数据。 (b)通过45°旋转。 (c) 沿着水平轴。 (d) 一般线性拉伸 绘图。

我们考虑$\mathbb{R}^2$中一组向量的三种线性变换，变换矩阵为

$$\boldsymbol{A}_1 = \begin{bmatrix} \cos(\frac{\pi}{4}) & -\sin(\frac{\pi}{4}) \\ \sin(\frac{\pi}{4}) & \cos(\frac{\pi}{4}) \end{bmatrix}, \boldsymbol{A}_2 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{A}_3 = \frac{1}{2}\begin{bmatrix} 3 & -1 \\ 1 & -1 \end{bmatrix}. \tag{2.97}$$

Figure 2.10 Three examples of linear transformations of the vectors shown as dots in(a); (b)Rotation by 45°; (c)Stretching of the horizontal coordinates by 2; (d)Combination of reflection, rotation and stretching.

宇宙空间(

图2.10给出了一组矢量的线性变换的三个例子。图2.10(a)显示了$\mathbb{R}^2$中的400向量，每个向量在相应的(x_1, x_2)坐标上用一个点表示。这些向量被排列在一个正方形中。当我们使用矩阵$\boldsymbol{A}_1$在(2.97)对这些向量进行线性变换，我们得到图2.10(b)中的旋转的正方形。如果我们应用$\boldsymbol{A}$所代表的线性映射，我们得到图2.10(c)中的矩形，其中每个x_1坐标被拉伸了2。图2.10(d)显示了图2.10(a)中的原始正方形在使用$\boldsymbol{A}_3$进行线性变换时的情况，它是反射、旋转和拉伸的组合。

2.7.2 基准变化

在下文中，我们将仔细研究如果我们改变V和W中的基数，线性映射$\Phi : V \to W$的转换矩阵如何变化。考虑到两个有序的基数

$$B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n), \quad \tilde{B} = (\tilde{\boldsymbol{b}}_1, \ldots, \tilde{\boldsymbol{b}}_n) \tag{2.98}$$

V和两个有序的基数

$$C = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m), \quad \tilde{C} = (\tilde{\boldsymbol{c}}_1, \ldots, \tilde{\boldsymbol{c}}_m) \tag{2.99}$$

的。此外，$\boldsymbol{A}_\Phi \in \mathbb{R}^{m \times n}$是线性映射$\Phi : V \to W$相对于基数$B$和$C$的变换矩阵，$\tilde{\boldsymbol{A}}_\Phi \in \mathbb{R}^{m \times n}$是相对于$\tilde{B}$和$\tilde{C}$的相应变换映射。
在下文中，我们将研究$\boldsymbol{A}$和$\tilde{\boldsymbol{A}}$的关系，也就是说，如何/
如果我们选择从B，C到$\tilde{B}$，$\tilde{C}$进行基础改变，我们是否可以将$\boldsymbol{A}$转化$_\Phi$为$\tilde{\boldsymbol{A}}_\Phi$。

*备注。*我们实际上得到了身份映射id的$_V$不同坐标表示。在图的背景下2.9,这意味着在不改变向量$\boldsymbol{x}$的情况下将关于$(\boldsymbol{e}_1, \boldsymbol{e}_2)$的坐标映射到关于$(\boldsymbol{b}_1, \boldsymbol{b}_2)$的坐标上。通过改变基和相应的向量表示，关于这个新基的变换矩阵可以有一个特别简单的形式，允许
用于直接计算。 ◆

例子（2.23基数变化）

考虑一个变换矩阵

$$\boldsymbol{A} = \begin{vmatrix} 2 & 1 \\ 1 & 2 \end{vmatrix} \tag{2.100}$$

如果我们定义一个新的基础

$$B = (\begin{vmatrix} 1 \\ 1 \end{vmatrix}, \begin{vmatrix} 1 \\ -1 \end{vmatrix}) \tag{2.101}$$

我们得到一个对角线转换矩阵

$$\tilde{\boldsymbol{A}} = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \tag{2.102}$$

相对于$\boldsymbol{B}$，它比$\boldsymbol{A}$更容易操作。

在下文中，我们将研究将关于一个基的坐标向量转化为关于另一个基的坐标向量的映射。我们将首先说明我们的主要结果，然后提供一个解释。

定理（2.20基数变化）。 *对于线性映射*$\Phi : V \to W$*，有序基数*

$$B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n), \quad \tilde{B} = (\tilde{\boldsymbol{b}}_1, \ldots, \tilde{\boldsymbol{b}}_n) \tag{2.103}$$

的V*和*

$$C = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m), \quad \tilde{C} = (\tilde{\boldsymbol{c}}_1, \ldots, \tilde{\boldsymbol{c}}_m) \tag{2.104}$$

和Φ*相对于*B*和*C*的变换矩阵*$\boldsymbol{A}_\Phi$*，相应的变换矩阵*$\tilde{\boldsymbol{A}}_\Phi$*相对于基数*$\tilde{B}$*和*$\tilde{C}$*给定为*

$$\tilde{\boldsymbol{A}}_\Phi = \boldsymbol{T}^{-1} \boldsymbol{A}_\Phi \boldsymbol{S}。 \tag{2.105}$$

这里，$\boldsymbol{S} \in \mathbb{R}^{n \times n}$*是*id*的变换矩阵*$_V$*，将*$\tilde{B}$*的坐标映射到*$B$*的坐标上，*$\boldsymbol{T} \in \mathbb{R}^{m \times m}$*是*id*的*$_W$*变换矩阵，将*$\tilde{C}$*的坐标映射到*$C$*的坐标。*$\boldsymbol{A}_\Phi$*相对于*$C$*的坐标。*

证明 根据Drumm和Weil(2001)，我们可以把V的新基$\tilde{B}$的向量写成B的基向量的线性组合，这样

$$\tilde{\boldsymbol{b}}_j = s_{1j}\boldsymbol{b}_1 + \cdots + s_{nj}\boldsymbol{b}_n = \sum_{i=1}^{n} s_{ij}\boldsymbol{b}_i, \quad j = 1, \ldots, n. \tag{2.106}$$

同样地，我们把W的新基向量$\tilde{C}$写成C的基向量的线性组合，这就得到了

$$\tilde{\boldsymbol{c}}_k = t_{1k}\boldsymbol{c}_1 + \cdots + t_{mk}\boldsymbol{c}_m = \sum_{l=1}^{m} t_{lk}\boldsymbol{c}_l, \quad k = 1, \ldots, m. \tag{2.107}$$

我们定义$\boldsymbol{S} = ((s_{ij})) \in \mathbb{R}^{n \times n}$为转换矩阵，它映射了围绕着$\tilde{B}$的坐标和围绕着$B$的坐标，以及$\boldsymbol{T} = ((t_{lk}))$是$\tilde{C}$的第$j$列是$\tilde{\boldsymbol{b}}_j$相对于$B$的坐标表示，而$\mathbb{R}^{m \times m}$则是将相对于$\tilde{C}$的坐标映射到相对于$C$的坐标。

$\boldsymbol{T}$的第k列是$\tilde{\boldsymbol{c}}_k$的坐标表示，相对于C.请注意，$\boldsymbol{S}$和$\boldsymbol{T}$都是有规律的。

我们将从两个角度来看$\Phi(\tilde{\boldsymbol{b}}_j)$。首先，应用映射$\Phi$，我们可以得到，对于所有$j = 1, \ldots, n$

$$\Phi(\tilde{\boldsymbol{b}}_j) = \sum_{k=1}^{m} \tilde{a}_{kj} \tilde{\boldsymbol{c}}_k \in W \overset{(2.107)}{=} \sum_{k=1}^{m} \tilde{a}_{kj} \sum_{l=1}^{m} t_{lk} \boldsymbol{c}_l = \sum_{l=1}^{m} \sum_{k=1}^{m} t_{lk} \tilde{a}_{kj} \boldsymbol{c}_l, \quad (2.108)$$

其中，我们首先将新的基向量$\tilde{\boldsymbol{c}}_k \in W$表示为基向量$\boldsymbol{c}_l \in W$的线性组合，然后交换了求和的顺序。

另外，当我们把$\tilde{\boldsymbol{b}}_j \in V$表示为线性组合时，就会出现以下情况$\boldsymbol{b}_j \in V$, we arrive at

$$\Phi(\tilde{\boldsymbol{b}}_j) \overset{(2.106)}{=} \Phi\left(\sum_{i=1}^{n} s_{ij} \boldsymbol{b}_i\right) = \sum_{i=1}^{n} s_{ij} \Phi(\boldsymbol{b}_i) = \sum_{i=1}^{n} s_{ij} \sum_{l=1}^{m} a_{li} \boldsymbol{c}_l \quad (2.109a)$$

$$= \sum_{l=1}^{m} \sum_{i=1}^{n} a_{li} s_{ij} \boldsymbol{c}_l, \quad j = 1, \ldots, n, \quad (2.109b)$$

其中我们利用了Φ的线性。对比(2.108)和(2.109b)，可以看出对于所有$j = 1, \ldots, n$和$l = 1, \ldots, m$，则

$$\sum_{k=1}^{mn} t_{lk} \tilde{a}_{kj} = \sum_{i=1}^{mn} a_{li} s_{ij} \quad (2.110)$$

因此，。

$$\boldsymbol{t} \tilde{\boldsymbol{a}}_{\varphi} = {}_{\Phi}\boldsymbol{a}\boldsymbol{s} \in \mathsf{r}^{m \times n}。 \quad (2.111)$$

以致于

$$\tilde{\boldsymbol{A}}_{\Phi} = \boldsymbol{T}^{-1}{}_{\Phi} \boldsymbol{A}\boldsymbol{S}。 \quad (2.112)$$

这证明了定理2.20. □

该定理2.20告诉我们，随着V（B被替换为$\tilde{B}$）和W（C被替换为$\tilde{C}$）的基础改变，线性映射$\Phi : V \to W$的变换矩阵$\boldsymbol{A}_{\Phi}$被一个等价的矩阵$\tilde{\boldsymbol{A}}_{\Phi}$所替换，该矩阵为

$$\tilde{\boldsymbol{A}}_{\Phi} = \boldsymbol{T}^{-1}{}_{\Phi} \boldsymbol{A}\boldsymbol{S}。 \quad (2.113)$$

图2.11说明了这种关系。考虑一个同态性$\Phi : V \to W$矩阵，以及V的有序基B，$\tilde{B}$和W的有序基C，$\tilde{C}$。映射Φ_{CB}是Φ的实例化，将B的基向量映射到C的基向量的线性组合上。假设我们知道Φ_{CB}相对于有序基B、C的变换矩阵$\boldsymbol{A}_{\Phi}$。

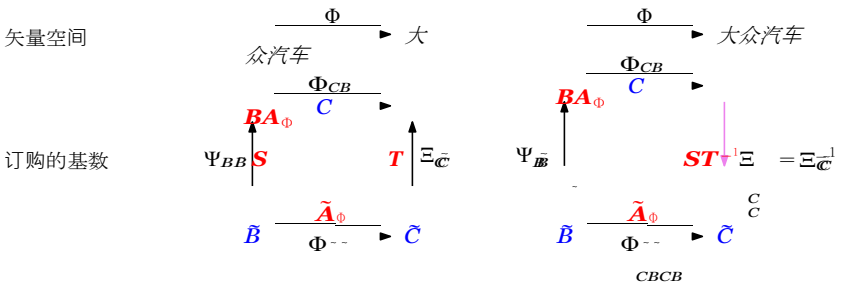


图2.11 对于一个同态的 $\Phi: V \to W$ 和有序基 $B, \tilde{B}$ V 的 C，W 的 $\tilde{C}$ (蓝色标记)，我们可以表示出映射 $\Phi_{\tilde{C}\tilde{B}}$ 与关于基地 $\tilde{B}, \tilde{C}$ 等效为同态的组成，$\Phi_{\tilde{C}\tilde{B}} = \Xi_{\tilde{C}C} \Phi_{CB} \Psi_{B\tilde{B}}$ 关于的基础上，在子标。的对应的变换矩阵为红色。

相应的变换矩阵 $\tilde{A}_\Phi$ 如下。首先，我们找到线性映射 Ψ_{BB} 的ma-trix表示：$V \to V$，它将相对于新基 $\tilde{B}$ 的坐标映射到（唯一的）坐标上，并具有相对于 "旧 "基础 B（在 V）。然后，我们使用变换ma-Φ_{CB} 的三角形 A_Φ：$V \to W$，将这些坐标映射到坐标上。

就 W 中的 C 而言。最后，我们使用一个线性映射 $\Xi_{\tilde{C}C}: W \to W$ 来把关于 C 的坐标映射到关于 $\tilde{C}$ 的坐标上。因此，我们可以把线性映射 $\Phi_{\tilde{C}\tilde{B}}$ 表达为涉及 "旧 "基础的线性映射的组合。

$$\Phi_{\tilde{C}\tilde{B}} = \Xi_{\tilde{C}C} \circ \Phi_{CB} \circ \Psi_{B\tilde{B}} = \Xi^{-1}_{C\tilde{C}} \circ \Phi_{CB} \circ \Psi_{B\tilde{B}}. \tag{2.114}$$

具体来说，我们使用 $\Psi_{B\tilde{B}} = \mathrm{id}_V$ 和 $\Xi_{\tilde{C}C} = \mathrm{id}_W$，即把向量映射到自己身上的同一映射，但关于不同的基。

相当于

定义（2.21等价性）。 两个矩阵 A、$\tilde{A} \in \mathbb{R}^{m \times n}$ 是*等价的*，如果存在规则矩阵 $S \in \mathbb{R}^{n \times n}$ 和 $T \in \mathbb{R}^{m \times m}$，那么 $\tilde{A} = T^{-1} A S$。

类似的

定义（相似性2.22）。 两个矩阵 A、$\tilde{A} \in \mathbb{R}^{n \times n}$ 是*相似的*，如果存在一个有规律的矩阵 $S \in \mathbb{R}^{n \times n}$，$\tilde{A} = S^{-1} A S$

备注。 相似矩阵总是等价的。然而，相等的矩阵不一定是相似的。 ◆

备注。 考虑向量空间 V，W，X。2.17,我们已经知道，对于线性映射 $\Phi: V \to W$ 和 $\Psi: W \to X$，映射 $\Psi \circ \Phi: V \to X$ 也是线性的。随着相应映射的变换矩阵 A_Φ 和 A_Ψ，整体变换矩阵为 $A_{\Psi \circ \Phi} = A_\Psi A_\Phi$。 ◆

考虑到这一点，我们可以从构成线性映射的角度来看待基数变化。

- A_Φ 是线性映射 Φ 的变换矩阵 $_{CB}$：$\to VW$ 相对于基数 B、C。
- $\tilde{A}_\Phi$ 是线性映射 $\Phi_{\tilde{C}\tilde{B}}$ 的变换矩阵：$V \to W$ 相对于基数 $\tilde{B}$，$\tilde{C}$。
- S 是一个线性映射 $\Psi_{BB}: V \to V$ 的变换矩阵。(通常，$\Psi = \mathrm{id}_V$ 是 V 中的身份映射。

- $\boldsymbol{T}$是线性映射$\Xi_{CC}: W \to W$的变换矩阵。
  (通常情况下，$\Xi = \mathrm{id}_W$是W中的身份映射，它代表$\tilde{C}$。

如果我们（非正式地）只用基数来写出变换。
那么$\boldsymbol{A}_\Phi: B \to C, \tilde{\boldsymbol{A}}_\Phi: \tilde{B} \to \tilde{C}, \boldsymbol{S}: \tilde{B} \to B, \boldsymbol{T}: \tilde{C} \to C$和
$\boldsymbol{T}: {}^{-1}C \to \tilde{C}$, 和

$$\tilde{b} \to \tilde{c} = \tilde{b} \to b \to c \to \tilde{c} \tag{2.115}$$

$$\tilde{\boldsymbol{A}}_\Phi = \boldsymbol{T}^{-1}{}_\Phi \boldsymbol{A}\boldsymbol{S} \text{。} \tag{2.116}$$

在右手边相乘，所以$\boldsymbol{x}$ 1 $\to \boldsymbol{S}\boldsymbol{x}$ 1 $\to \boldsymbol{A}({}_\Phi \boldsymbol{S}\boldsymbol{x})$ 1 $\to$
$\boldsymbol{T}\boldsymbol{A}({}^{-1}{}_\Phi \boldsymbol{S}\boldsymbol{x}) = \tilde{\boldsymbol{A}}_\Phi \boldsymbol{x}$

。

例子（2.24基数变化）
考虑一个线性映射$\Phi: \mathrm{R}^3 \to \mathrm{R}^4$其变换矩阵为

$$\boldsymbol{A}_\Phi = \begin{matrix} 1 & 2 & 0 \\ -1 & 1 & 3 \\ 3 & 7 & 1 \\ -1 & 2 & 4 \end{matrix} \tag{2.117}$$

相对于标准基数

$$B = (0, 1, 0), \quad 001 \qquad C = (,,,) \begin{matrix} 1000 \\ 0100 \\ 0010 \\ 0001 \end{matrix} \tag{2.118}$$

我们寻求Φ相对于新基数的变换矩阵$\tilde{\boldsymbol{A}}_\Phi$

$$\boldsymbol{B} = (1, 1, 0) \in \mathrm{R}^3, \quad C = (,,,) \begin{matrix} 1101 \\ 1010 \\ 0110 \\ 0001 \end{matrix} \tag{2.119}$$

011

然后

。

$$\boldsymbol{S} = \begin{matrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{matrix}, \qquad \boldsymbol{T} = \begin{matrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{matrix}, \tag{2.120}$$

其中，$\boldsymbol{S}$的第i列是$\tilde{\boldsymbol{b}}_i$的坐标表示，在
因为$\boldsymbol{B}$是标准基，所以共轨表示法很容易找到。对于一般的基$\boldsymbol{B}$，我们需要解决一个线性方程组来找到λ_i，以便

请注意，在(2.116)中的执行顺序是从右到左，因为vec-

$\sum_{i=1}^{3} \lambda_i \boldsymbol{b}_i = \tilde{\boldsymbol{b}}_j, j = 1, \ldots, 3$.类似地， $\boldsymbol{T}$的第j列是coordi-的。在$\boldsymbol{C}$的基向量方面对$\tilde{\boldsymbol{c}}_j$的表示， 因此，我们得到

$$\tilde{\boldsymbol{A}}_\Phi = \boldsymbol{S}^{-1} \boldsymbol{A}_\Phi \boldsymbol{T} = \begin{bmatrix} \frac{1}{2} & 1 & \frac{3}{2} \\ 1 & 1 & \frac{1}{2} \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{bmatrix} \tag{2.121a}$$

$$= \begin{bmatrix} -4 & -4 & -2 \\ 6 & 0 & 0 \\ 4 & 8 & 4 \\ 1 & 6 & 3 \end{bmatrix} . \tag{2.121b}$$

在第四章中，我们将能够利用基点变化的概念来找到一个基点，在这个基点上，一个en-domorphism的变换矩阵有一个特别简单的（对角线）形式。在第十章中，我们将研究一个数据压缩问题，并找到一个方便的基，我们可以把数据投射到上面，同时使压缩损失最小。

2.7.3 图像和内核

线性映射的图像和内核是具有某些重要特性的矢量子空间。在下文中，我们将更仔细地描述它们的特性。

定义（2.23图像和内核）。

内核无效空间

对于$\Phi : V \to W$，我们定义*内核/空洞空间*

$$\ker(\Phi) := \Phi^{-1}(\boldsymbol{0}_W) = \{\boldsymbol{v} \in V : \Phi(\boldsymbol{v}) = \boldsymbol{0}_W\} \tag{2.122}$$

图像 范围

和*图像/范围*

$$\mathrm{Im}(\Phi) := \Phi(V) = \{\boldsymbol{w} \in W \mid \exists \boldsymbol{v} \in V : \Phi(\boldsymbol{v}) = \boldsymbol{w}\} 。 \tag{2.123}$$

域名代码

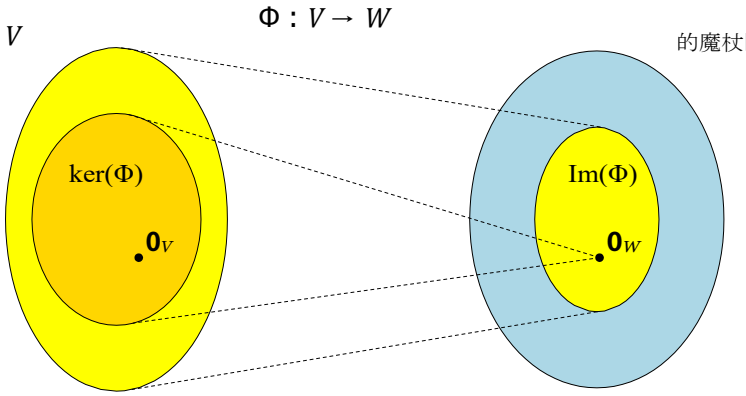
我们也把V和W分别称为Φ的*域*和密码*域*。

直观地说，内核是Φ映射到中性元素W $\boldsymbol{0}_W$*的向量*$\boldsymbol{v}$ V*的*集合。图像是Φ从V中的任何向量可以 "到达 "的向量$\boldsymbol{w}$ W*的*集合。图中给出了一个说明图2.12.

*备注。考虑*一个线性映射$\Phi : V \to W$，其中V, W是向量空间。

- 始终认为，$\Phi(\boldsymbol{0}_V) = \boldsymbol{0}_W$因此，$\boldsymbol{0}_V \in \ker(\Phi)$。特别是，空空间从来不是空的。

$\operatorname{Im}(\Phi) \subseteq W$ 是 W 的一个子空间，$\ker(\Phi) \subseteq V$ 是 V 的一个子空间。



图中的2.12内核线性映射 $\Phi : V \to W$ 的魔杖图像。

- 当且仅当ker(Φ)={**0**}时，Φ是注入性的（一对一）。

◆

备注（空洞空间和列空间）。让我们考虑$\boldsymbol{A} \in \mathbb{R}^{m \times n}$和一个线性映射$\Phi : \mathbb{R}^n \to \mathbb{R}^m$，$\boldsymbol{x} \mapsto \boldsymbol{A}\boldsymbol{x}$。

- For $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]$, where $\boldsymbol{a}_i$ are the columns of $\boldsymbol{A}$, we obtain

$$\mathrm{Im}(\Phi) = \{\boldsymbol{A}\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^n\} = \sum_{i=1}^{n} x_i \boldsymbol{a}_i : x_1, \ldots, x_n \in \mathbb{R} \tag{2.124a}$$

$$= \mathrm{span}[\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n] \subseteq \mathbb{R}^m, \tag{2.124b}$$

即，图像是$\boldsymbol{A}$的列的跨度，也称为*列空间*。因此，列空间（图像）是$\mathbb{R}^m$的一个子空间，其中m是矩阵的 "高度"。

与列的空间

- $\mathrm{rk}(\boldsymbol{A}) = \dim(\mathrm{Im}(\Phi))$。
- 核/空空间ker(Φ)是同构线性方程组$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$的一般解，它捕捉了产生$\mathbb{R}^n$的元素的所有可能线性组合$\boldsymbol{0}^m$。
- 核心是$\mathbb{R}^n$的一个子空间，其中n是矩阵的 "宽度"。 ∈
- 内核侧重于列之间的关系，我们可以用它来确定是否/如何将一个列表达为其他列的线性组合。

◆

例子（线性映射的2.25图像和内核）。

绘图

$$\Phi : \mathbb{R}^4 \to \mathbb{R}^2, \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \mapsto \begin{bmatrix} 1 & 2 & -1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} x_1 + 2x_2 - x_3 \\ x_1 + x_4 \end{bmatrix} \tag{2.125a}$$

。

$$= x_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + x_2 \begin{bmatrix} 2 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} -1 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad (2.125b)$$

是线性的。为了确定Im(Φ)，我们可以取变换矩阵的列的跨度，得到

$$\mathrm{Im}(\Phi) = \mathrm{span}[\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}]. \quad (2.126)$$

为了计算Φ的内核（无效空间），我们需要解决$\boldsymbol{Ax} = \mathbf{0}$，即。
我们需要解决一个同质方程组。要做到这一点，我们使用
高斯消除法将$\boldsymbol{A}$转化为简化的行-歇尔形式。

$$\begin{bmatrix} 1 & 2 & -1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \rightsquigarrow \cdots \rightsquigarrow \begin{bmatrix} 1 & 0 & -\frac{1}{2} & 0 \\ 0 & 1 & -\frac{1}{2} & 1 \end{bmatrix}. \quad (2.127)$$

这个矩阵是简化的行-歇尔形式，我们可以使用减1技巧来计算内核的一个基（见第2.3.3).另外，我们可以将非枢纽列（第3列和4）表示为枢纽列（第和12）的线性组合。第三列$\boldsymbol{a}_3$是
相当于第二列$\boldsymbol{a}_2$的$-\frac{1}{2}$倍。 因此，$\mathbf{0} = \boldsymbol{a}_3 + \frac{1}{2}\boldsymbol{a}_2$.在
同样的方法，我们看到$\boldsymbol{a}_4 = \boldsymbol{a}_1 + \frac{1}{2}\boldsymbol{a}_2$，因此，$\mathbf{0} = \boldsymbol{a}_1 + \frac{1}{2}\boldsymbol{a}_2 - \boldsymbol{a}_4$。
总的来说，这给我们的内核（空空间）为

$$\ker(\Phi) = \mathrm{span}[\begin{bmatrix} 0 \\ \frac{1}{2} \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ \frac{1}{2} \\ 0 \\ 1 \end{bmatrix}]. \quad (2.128)$$

rank-nullity

定理（Rank-**2.24**Nullity定理）。*对于矢量空间V，W和一个线耳映射$\Phi : V \to W$，它认为*

$$\dim(\ker(\Phi)) + \dim(\mathrm{Im}(\Phi)) = \dim(V)。 \quad (2.129)$$

线性映射的基本定理

等级空性定理也被称为*线性映射的基本定理*（Axler,2015，定理3.22）。以下是定理2.24的直接后果。

- 如果$\dim(\mathrm{Im}(\Phi)) < \dim(V)$，那么$\ker(\Phi)$就是非琐碎的，也就是说，内核包含超过和$\mathbf{0}_V \dim(\ker(\Phi)) \geq 1$。
- 如果$\boldsymbol{A}_\Phi$是Φ相对于一个有序基础的变换矩阵且$\dim(\mathrm{Im}(\Phi)) < \dim(V)$，则线性方程组$\boldsymbol{A}_\Phi \boldsymbol{x} = \mathbf{0}$有无限多的解。
- 如果$\dim(V) = \dim(W)$，那么下面的三方等价关系成立。
  - Φ是注入式的
  - Φ是射出的
  - 由于$\mathrm{Im}(\Phi) \subseteq W$，*所以*Φ是双射的。

2.8 仿生空间

在下文中，我们将仔细研究那些从原点偏移的空间，即不再是矢量子空间的空间。此外，我们将简要讨论这些仿生空间之间的映射的性质，这些映射类似于线性映射。

*备注。*在机器学习文献中，线性和仿生之间的区别有时并不明确，因此我们可以找到仿生空间/映射作为线性空间/映射的参考。 ◆

2.8.1 仿生子空间

定义（2.25仿生子空间）。设V是一个矢量空间，$\boldsymbol{x}_0 \in V$且$U \subseteq V$一个子空间。那么这个子集

$$L = \boldsymbol{x}_0 + U := \{\boldsymbol{x}_0 + \boldsymbol{u} : \boldsymbol{u} \in U\} \quad (2.130a)$$

$$= \{\boldsymbol{v} \in V \mid \exists \boldsymbol{u} \in U : \boldsymbol{v} = \boldsymbol{x}_0 + \boldsymbol{u}\} \subseteq V \quad (2.130b)$$

被称为V的*仿射子空间*或*线性流形*。U被称为*方向*或*方向的空间*，而$\boldsymbol{x}_0$被称为*支持点*。在第十二章中，我们把这样的子空间称为*超平面*。

affine子空间
线性流形的方向性空间
支持点
超平面

请注意，仿生子空间的定义排除了如果$\mathbf{0}\boldsymbol{x}_0 \notin U$。因此，一个仿生子空间不是$V$的一个（线性）子空间（矢量子空间），因为$\boldsymbol{x}_0 \notin U$。

仿射子空间的例子是$\mathbb{R}$中的^{3}点、线和平面，它们是不（一定）要经过原点。

*备注。*考虑一个向量空间V的两个仿生子空间$L = \boldsymbol{x}_0 + U$和$\tilde{L} = \tilde{\boldsymbol{x}}_0 + \tilde{U}$。那么，$L \subseteq \tilde{L}$当且仅当$U \subseteq \tilde{U}$和$\tilde{\boldsymbol{x}}_0 - \boldsymbol{x}_0 \in \tilde{U}$。

仿生子空间通常由*参数*来描述。考虑一个k-dimen-的*名义*仿射空间$L = \boldsymbol{x}_0 + U$。如果$(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k)$是V的一个有序基础$U$，那么每个元素$\boldsymbol{x} \in L$可以被唯一地描述为

$$\boldsymbol{x} = \boldsymbol{x}_0 + \lambda_1 \boldsymbol{b}_1 + \ldots + \lambda_k \boldsymbol{b}_k, \quad (2.131)$$

其中$\lambda_1, \ldots, \lambda_k \in \mathbb{R}$.这种表示法被称为*参数方程*$L$的方向向量$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k$和*参数*$\lambda_1, \ldots, \lambda_k$. ◆

参数方程
参数

例子（2.26仿生子空间）

- 一维仿射子空间被称为*线*，可以写成$\boldsymbol{y} = \boldsymbol{x}_0 + \lambda \boldsymbol{b}_1$，其中$\lambda \in \mathbb{R}$，$U = \text{span}[\boldsymbol{b}_1] \subseteq \mathbb{R}^n$是一个一维子空间。这意味着一条线是由一个上端口点$\boldsymbol{x}_0$和一个定义方向的矢量$\boldsymbol{b}_1$定义的^n。参见图2.13来说明。

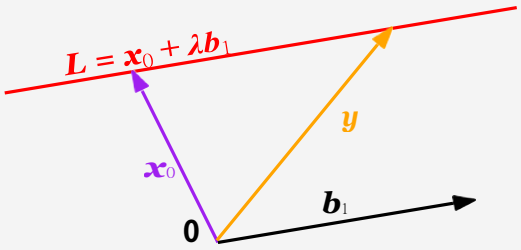
线

plane

- R的n二维仿射子空间被称为*平面*。平面的准度量方程是$\boldsymbol{y} = \boldsymbol{x}_0 + \lambda_1\boldsymbol{b}_1 + \lambda_2\boldsymbol{b}_2$，其中$\lambda_1, \lambda_2 \in \mathbb{R}$和$U = \text{span}[\boldsymbol{b}_1, \boldsymbol{b}_2] \subseteq \mathbb{R}^n$.这意味着一个平面的定义是由a支持点$\boldsymbol{x}_0$和两个线性独立的矢量$\boldsymbol{b}_1$，$\boldsymbol{b}_2$横跨方向空间。

超平面

- 在R中n，$(n-1)$-维仿生子空间被称为*超平面*。而相应的参数方程为$\boldsymbol{y} = \boldsymbol{x}_0 + \sum_{i=1}^{n-1} \lambda_i\boldsymbol{b}_i$，其中$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{n-1}$构成一个$(n-1)$维子空间的基础这意味着一个超平面是由一个支持点定义的$\boldsymbol{x}_0$和$(n-1)$个线性独立向量$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{n-1}$，它们横跨了方向空间。在R中^2，一条线也是一个超平面。在R中^3，一个平面也是一个超平面。



图线**2.13**是仿生子空间。线上的矢量$\boldsymbol{y} = \boldsymbol{x}_0 + \lambda\boldsymbol{b}_1$位于一个支持点为$\boldsymbol{x}_0$和方向为$\boldsymbol{b}_1$的仿生子空间中。

备注（非均质线性方程组和仿生子空间）。对于$\boldsymbol{A} \in \mathbb{R}^{m\times n}$和$\boldsymbol{x} \in \mathbb{R}^m$，线性方程组的解$\boldsymbol{A}\boldsymbol{\lambda} = \boldsymbol{x}$要么是空集，要么是维数为$\text{nrk}(\boldsymbol{A})$的R的^{n}仿射子空间。特别是，线性方程$\lambda_1\boldsymbol{b}_1 + \ldots + \lambda_n\boldsymbol{b}_n = \boldsymbol{x}$，其中$(\lambda_1, \ldots, \lambda_n) = (0, \ldots, 0)$，是R的^{n}一个超平面。

在R中^n，每个k维仿生子空间都是一个不均匀的线性方程组$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$的解，其中$\boldsymbol{A} \in \mathbb{R}^{m\times n}$，$\boldsymbol{b}$ R^{m}和$\text{rk}(\boldsymbol{A}) = nk$。回顾一下，对于同质方程组来说$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$解决方案是一个矢量子空间，我们也可以把它看作是一个特殊的仿生空间，支持点$\boldsymbol{x}_0 = \boldsymbol{0}$。

2.8.2 仿生映射

类似于我们在第二节2.7,中讨论的矢量空间之间的线性映射，我们可以在两个仿生空间之间定义仿生映射。线性映射和仿生映射是密切相关的。因此，我们从线性映射中已经知道的许多属性，例如，线性映射的组合是线性映射，也适用于仿生映射。

定义 (2.26 Affine Mapping)。对于两个向量空间V，W，一个线性的

映射$\Phi: V \to W$，和$\boldsymbol{a} \in W$，映射

$$\varphi: V \to W \tag{2.132}$$

$$\boldsymbol{x} \mapsto \boldsymbol{a} + \Phi(\boldsymbol{x}) \tag{2.133}$$

是一个从V到W的*仿生映射*。向量$\boldsymbol{a}$被称为*平移*φ的*向量*。

affine映射

翻译向量

- 每一个仿生映射$\varphi: V \to W$也是一个线性映射$\Phi: V \to W$和一个平移$\tau: W \to W$的组成，这样，$\varphi = \tau \circ \Phi$. 这些映射$\Phi$和$\tau$是唯一确定的。
- 仿生映射$\varphi: V \to W$，$\varphi(\vec{x}) = \Phi(\vec{x}) + \vec{a}$的组合$\varphi'$是仿生的。
- 仿生映射保持了几何结构的不变性。它们也预先为维度和平行度服务。

2.9进一步阅读

有许多学习线性代数的资源，包括Strang(2003)、Golan(2007)、Axler(2015)以及Liesen和Mehrmann(2015)的文本书籍。还有一些在线资源，我们在本章的介绍中提到过。我们在这里只介绍了高斯消除法，但还有很多其他的方法来解决线性方程组，我们可以参考Stoer和Burlirsch(2002)、Golub和Van Loan(2012)以及Horn和Johnson(2013)的数值线性代数教科书来深入讨论。

在本书中，我们区分了线性代数的主题（例如，向量、矩阵、线性独立、基）和与向量空间的几何有关的主题。在第三章中，我们将介绍内积，它诱导了一个规范。这些概念使我们能够定义角度、长度和距离，我们将使用这些概念进行正交投影。投影是许多机器学习算法的关键，如线性回归和主成分分析，我们将在第9章和第10章分别介绍这两种算法。

练习

2.1 我们考虑 （R\{−1}, *），其中

$$a * b := ab + a + b, \quad a, b \in \mathbb{R}\setminus\{-1\} \tag{2.134}$$

a.证明（R/{−1}, *）是一个阿贝尔群。 b.求解

$$3 * x * x = 15$$

在阿贝尔群（R{−1}, *）中，其中*的定义见(2.134).

2.2 让n是在N中。 0我们将整数k的同构类$\bar{k}$定义为以下集合
}

$$\begin{aligned}\bar{k} &= \{x \in \mathbb{Z} \mid x - k = 0 \ (mod n)\} \\ &= \{x \in \mathbb{Z} \mid \exists a \in \mathbb{Z}: \ (x - k = n - a)\}。\end{aligned}$$

我们现在将Z/nZ（有时写成Z_n）*定义*为所有同调类的集合，欧几里得除法意味着这个集合是一个包含n个元素的有限集合。

$$Z_n = \{0, \bar{1}, \ldots, \overline{n-1}\}$$

对于所有$\bar{a}$, $b \in Z_n$，我们定义

$$\bar{a} \oplus \bar{b} := \overline{a + b}$$

a.证明（Z_n, ⊕）是一个群。它是阿贝尔群吗？
b.现在我们对Z_n中的所有a和b定义另一个操作⊗为

$$\bar{a} \otimes \bar{b} = \overline{a \times b}, \tag{2.135}$$

其中$a \times b$代表Z中通常的乘法。
让$n = 5$.....画出$\mathbb{Z}_5\setminus\{0\}$中元素在⊗下的次数表，即计算$\mathbb{Z}_5\setminus\{0\}$中所有$a$和$b$的乘积$a \otimes b$。
因此，请证明$\mathbb{Z}_5\setminus\{0\}$在⊗下是封闭的，并且对⊗拥有一个中性元素。显示$\mathbb{Z}_5\{0\}$中所有元素在⊗下的倒数。结论是（$\mathbb{Z}_5\setminus 0\{\}$, ⊗）是一个阿贝尔群。
c.证明（$\mathbb{Z}_8\setminus\{0\}$, ⊗）不是一个群。
d.我们记得，B ézout定理指出，两个整数a和b是相对素数（即$gcd(a, b) = 1$），当且仅当存在两个整数u和v，使得$au + bv = 1$。 证明（$Z_n\setminus\{0\}$, ⊗）是一个群，如果并且
只有当$n \in \mathbb{N}\setminus\{0\}$是素数时，才有可能。

2.3 考虑到定义如下的×3矩阵3的集合 G。

$$G = \left\{ \begin{bmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{bmatrix} \in \mathbb{R}^{3\times 3} \right\}, \quad x, y, z \in \mathbb{R}$$

我们定义为标准的矩阵乘法。

(,)G是一个群吗？如果是的话，它是阿贝尔群吗？请说明你的答案。

2.4 如果可能的话，请计算以下矩阵积。

a.

211 1 0

7 81 0 1

b.

1 2 3 1 1 0

7 8 9 1 0 1

c.

1 1 01 2 3

1 0 17 8 9

d.

1 0 3
1 1 −1
212 21
4 1 -1 −4 52

e.

11- 2 1 212
1 4 1 -1 −4
52

2.5 找出以下非均质线性系统$\boldsymbol{Ax} = \boldsymbol{b}$的$\boldsymbol{x}$中所有解的集合，其中$\boldsymbol{A}$和$\boldsymbol{b}$的定义如下。

a.

$$\boldsymbol{A} = \begin{bmatrix} 11 & 11 & - \\ 25 & -75 & - \\ 2- & 113 & \\ 52 & - & 42 \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} 2- \\ \\ \\ 6 \end{bmatrix}$$

b.

$$\boldsymbol{A} = \begin{bmatrix} -1 & 1 & 00 & 1 \\ & 110- & 30 & \\ 2 & -1 & 01 & - \\ - & 12021 & & \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} 6 \\ \\ 15 \\ --1 \end{bmatrix}$$

2.6 使用高斯消除法，找出非均质方程组$\boldsymbol{Ax}=\boldsymbol{b}$的所有解，其中包括

$$\boldsymbol{A} = \begin{bmatrix} 0 1 0 0 102 \\ 0 0 0 1 1 \\ 0 1 0 0 0 1 \end{bmatrix}, \quad 0\boldsymbol{b} = \begin{bmatrix} \\ 1 - \\ 1 \end{bmatrix}。$$

2.7 找到$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \in \mathbb{R}^3$的方程组$\boldsymbol{A}\boldsymbol{x}=12\boldsymbol{x}$的所有解

其中

$$\boldsymbol{A} = \begin{bmatrix} 6 & 4 & 3 \\ 6 & 0 & 9 \\ 0 & 8 & 0 \end{bmatrix}$$

和 $\sum_{i=1}^{3} x_i = 1$.

2.8 如果可能的话，请确定以下矩阵的反数。

a.

$$\boldsymbol{A} = \begin{bmatrix} 2 & 3 & 4 \\ 3 & 4 & 5 \\ 4 & 5 & 6 \end{bmatrix}$$

b.

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

2.9 以下哪些集合是$\mathbb{R}^3$的子空间？

a. $A = \{(\lambda, \lambda + \mu^3, \lambda - \mu^3) \mid \lambda, \mu \in \mathbb{R}\}$

b. $B = \{(\lambda^2, \lambda^2, 0) \mid \lambda \in \mathbb{R}\}$

c.让γ在$\mathbb{R}$中。$\}$

$C = \{(\xi_1, \xi_2, \xi_3) \in \mathbb{R}^3 \mid \xi_1 - 2\xi_2 + 3\xi_3 = \gamma\}$。

d.$D = \{(\xi_1, \xi_2, \xi_3) \in \mathbb{R}^3 \mid \xi_2 \in \mathbb{Z}\}$。

2.10 下列各组向量是否线性独立？

a.

$\boldsymbol{x}_1 = [1, -3, 0]^\top$ $\boldsymbol{x}_2 = [-2, 1, 3]^\top$ $\boldsymbol{x}_3 = [8, -3, 2]^\top$

。

b.

$$\boldsymbol{x}_1 = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{x}_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{x}_3 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

2.11 撰写

$$\boldsymbol{y} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$$

的线性组合

$$\boldsymbol{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{x}_2 = \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \quad \boldsymbol{x}_3 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

2.12 考虑$\mathbb{R}^4$的两个子空间。

$$U_1 = \operatorname{span}\left[\begin{pmatrix}1\\1\\1\\-1\end{pmatrix}, \begin{pmatrix}-1\\1\\-1\\0\end{pmatrix}\right], \quad U_2 = \operatorname{span}\left[\begin{pmatrix}-1\\2\\0\\1\end{pmatrix}, \begin{pmatrix}-2\\3\\-2\\0\end{pmatrix}, \begin{pmatrix}-1\\6\\-1\\0\end{pmatrix}\right].$$

确定$U_1 \cap U$的一个基础$_2$。

2.13 考虑两个子空间U_1和U_2， 其中U_1是同质方程组$_1\boldsymbol{Ax}=$的解空间， $\mathbf{0}U_2$是同质方程组$_2\boldsymbol{Ax}=$的解空间，其中有$\mathbf{0}$

$$\boldsymbol{A}_1 = \begin{pmatrix}1&-2&1&-1\\1&0&1&-3\\1&3&1&3\end{pmatrix}, \quad \boldsymbol{A}_2 = \begin{pmatrix}7&-5&12&3\\3&-1&2&1\end{pmatrix}.$$

a.确定U_1、U的$_2$维度。 b.确定U_1和U的$_2$基。 c.确定$U_1 \cap U$的$_2$基。

2.14 考虑两个子空间U_1和U_2， 其中U_1是由以下列所跨越的$\boldsymbol{A}_1$和U_2是由$\boldsymbol{A}$的列跨越的， $_2$有

$$\boldsymbol{A}_1 = \begin{pmatrix}1&-2&1&-1\\1&0&1&-3\\1&3&1&3\end{pmatrix}, \quad \boldsymbol{A}_2 = \begin{pmatrix}7&-5&12&3\\3&-1&2&1\end{pmatrix}.$$

a.确定U_1、U的维数 $_2$b.确定U_1和U的基数 $_2$c.确定$U_1 \cap U$的基数$_2$

2.15设$F=\{(x, y, z)\in\mathbb{R}^3 | x+y-z=0\}$， $G=\{(a-b, a+b, a3-b) | a, b\in\mathbb{R}\}$。 a.证明$F$和$G$是$\mathbb{R}$的^{3}子空间。

b.计算$F\cap G$而不借助于任何基向量。

c.找到一个F和一个G的基，用之前找到的基向量计算$F\,G$，并将你的结果与之前的问题核对。

2.16 下列映射是线性的吗？ a.设a,

$b\in\mathbb{R}$。

$$\Phi : L^1([a, b]) \to \mathbb{R}$$
$$f \to \Phi(f) = \int_a^b f(x)dx \text{。}$$

其中$L(^1[a, b])$表示$[a, b]$上可整定函数的集合。

b.

$$\Phi : C^1 \to C^0$$
$$f \to \Phi(f) = f' \text{。}$$

其中对于$k\geq 1$， C^k表示k次连续可差函数的集合， 而C^0表示连续函数的集合。

c.

$$\Phi : \mathbb{R} \to \mathbb{R}$$
$$x \to \Phi(x) = \cos(x)$$

d.

$$\Phi : \mathbb{R}^3 \to \mathbb{R}^2$$
$$\boldsymbol{x} \to \begin{bmatrix} 1 & 2 & 3 \\ 1 & 4 & 3 \end{bmatrix} \boldsymbol{x}$$

e.假设θ在$[0, 2\pi]$中，并且

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^2$$
$$\begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \boldsymbol{x}$$

2.17 考虑线性映射

$$\Phi : \mathbb{R}^3 \to \mathbb{R}^4, \quad \Phi \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{bmatrix} x_1 + x_2 + x_3 \\ x_1 - 3x_2 \\ 2x_1 + 3x_2 + x_3 \\ x_3 \end{bmatrix}$$

- 求变换矩阵$\boldsymbol{A}_\Phi$。确定$\operatorname{rk}(\boldsymbol{A}_\Phi)$。
- 计算Φ的内核和图像。什么是$\dim(\ker(\Phi))$和$\dim(\operatorname{Im}(\Phi))$?

2.18 设E是一个矢量空间。设f和g是E上的两个自变形，使得$f \circ g = id_E$（即$f \circ g$是身份映射id_E）。证明$\ker(f) = \ker(g \circ f)$，$\operatorname{Im}(g) = \operatorname{Im}(g \circ f)$，并且$\ker(f) \cap \operatorname{Im}(g) = \{0_E\}$。

2.19 考虑一个内形态$\Phi : \mathbb{R}^3 \to \mathbb{R}^3$，其变换矩阵（相对于$\mathbb{R}^3$的标准基）为

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

a.确定$\ker(\Phi)$和$\operatorname{Im}(\Phi)$。 b.确定变换矩阵$\tilde{\boldsymbol{A}}_\Phi$就基础而言，

$$B = \left(\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \right),$$

即对新的基础B进行基础变更。

2.20 让我们考虑$\boldsymbol{b}_1$，$\boldsymbol{b}_2$，$\tilde{\boldsymbol{b}}_1$，$\tilde{\boldsymbol{b}}_2$，$\mathbb{R}$的4向量在$\mathbb{R}^2$标准基础上表示为

$$\boldsymbol{b}_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad \boldsymbol{b}_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \quad \tilde{\boldsymbol{b}}_1 = \begin{bmatrix} 2 \\ -2 \end{bmatrix}, \quad \tilde{\boldsymbol{b}}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

并让我们定义$\mathbb{R}^2$两个有序基$B = (\boldsymbol{b}_1, \boldsymbol{b}_2)$和$\tilde{B} = (\tilde{\boldsymbol{b}}_1, \tilde{\boldsymbol{b}}_2)$。

a.证明B和$\tilde{B}$是$\mathbb{R}^2$的两个基，并画出这些基向量。 b.计算从$\tilde{B}$到B进行基改变的矩阵$\boldsymbol{P}$。 c.我们考虑$\boldsymbol{c}_1$，$\boldsymbol{c}_2$，$\boldsymbol{c}_3$，$\mathbb{R}^3$的三个向量定义在标准基上。的$\mathbb{R}^3$为

$$\boldsymbol{c}_1 = \begin{bmatrix} 1 \\ -2 \\ 0 \end{bmatrix}, \quad \boldsymbol{c}_2 = \begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix}, \quad \boldsymbol{c}_3 = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

我们定义$C = (\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{c}_3)$。

(i) 证明C是$\mathbb{R}^3$的一个基础，例如，通过使用行列式（见第4.1节）。

(ii) 让我们称$\tilde{C} = (\tilde{\boldsymbol{c}}_1, \tilde{\boldsymbol{c}}_2, \tilde{\boldsymbol{c}}_3)$为$\mathbb{R}^3$的标准基。确定$\Phi$执行从$C$到$\tilde{C}$的基改变的矩阵$\boldsymbol{P}$。

d.我们考虑一个同态性$\Phi : \mathbb{R}^2 \to \mathbb{R}^3$，以便

$$\Phi(\boldsymbol{b}_1 + \boldsymbol{b}_2) = \boldsymbol{c}_2 + \boldsymbol{c}_3$$
$$\Phi(\boldsymbol{b}_1 - \boldsymbol{b}_2) = 2\boldsymbol{c}_1 - \boldsymbol{c}_2 + 3\boldsymbol{c}_3$$

其中$B = (\boldsymbol{b}_1, \boldsymbol{b}_2)$和$C = (\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{c}_3)$分别是$\mathbb{R}^2$和$\mathbb{R}^3$有序基。

确定Φ相对于指定基数B和C的变换矩阵$\boldsymbol{A}_\Phi$。

e. 确定$\tilde{B}$和$\tilde{C}$，即Φ相对于基数的变换矩阵。

f.让我们考虑矢量$\boldsymbol{x} \in \mathbb{R}^2$其在$\tilde{B}$中的坐标为$[2, 3]^T$。换句话说，$\boldsymbol{x} = 2\tilde{\boldsymbol{b}}_1 + 3\tilde{\boldsymbol{b}}_2$。

(i) 计算$\boldsymbol{x}$在B中的坐标。

(ii) 在此基础上，计算用C表示的$\Phi(\boldsymbol{x})$的坐标。 (iii)然后，用$\tilde{\boldsymbol{c}}_1, \tilde{\boldsymbol{c}}_2, \tilde{\boldsymbol{c}}_3$写出$\Phi(\boldsymbol{x})$。

(iv)用$\tilde{B}$中$\boldsymbol{x}$的表示法和矩阵$\tilde{A}$来求出这一点直接的结果。

3

解析几何

在第二章中，我们在一般但抽象的水平上研究了向量、向量空间和线性映射。在本章中，我们将为所有这些概念添加一些几何解释和直觉。特别是，我们将研究几何向量，计算它们的长度和两个向量之间的距离或角度。为了能够做到这一点，我们为向量空间配备一个内积，以诱导向量空间的几何。内积及其相应的规范和度量捕捉了相似性和距离的直观概念，我们在第12章中用它来开发支持向量机。然后，我们将使用向量之间的长度和角度的概念来讨论正交投影，这将在我们讨论第十章的主成分分析和第九章的最大似然估计的回归时发挥核心作用。图3.1概述了本章的概念之间的关系以及它们与本书其他章节的联系。

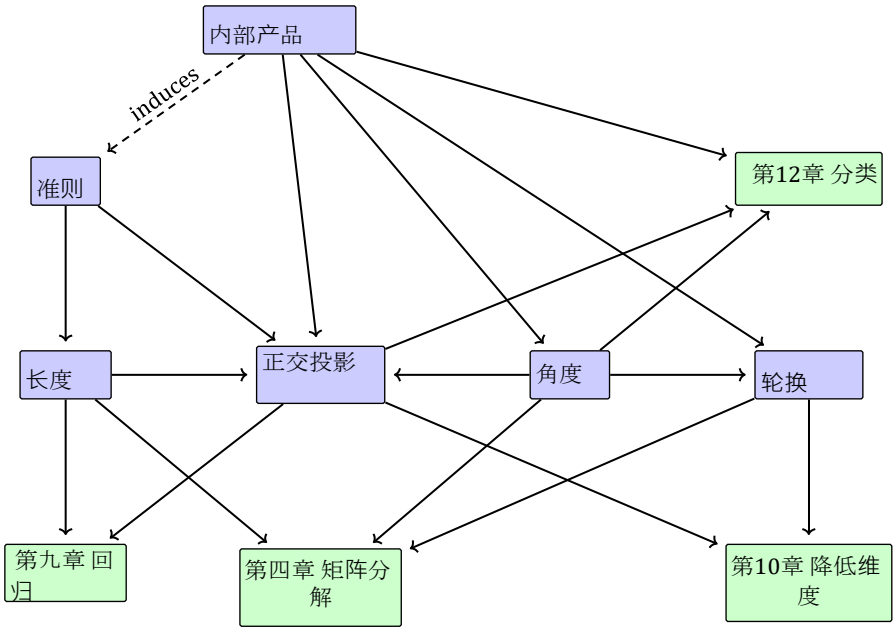


图 本章介绍的概念的**3.1**思维导图，以及它们在本书其他部分的使用时间。

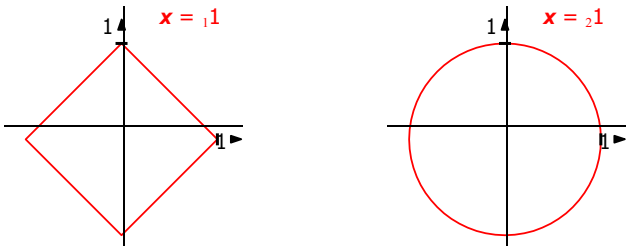


图 对于3.3不同的规范，红线表示具有规范1的向量集。左边：曼哈顿准则；右边。欧几里德距离。

3.1 规范

当我们想到几何向量时，即以原点为起点的有向线段，那么直观地说，一个向量的长度就是这个有向线段的 "终点 "与原点的距离。在下文中，我们将用规范的概念来讨论向量的长度概念。

定义（3.1规范）。向量空间V上的*规范*是一个函数

norm

$$\|\cdot\| : V \to \mathbb{R}, \tag{3.1}$$

$$\boldsymbol{x} \mapsto \|\boldsymbol{x}\|, \tag{3.2}$$

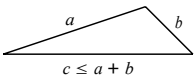
它为每个向量$\boldsymbol{x}$分配其*长度*$\|\boldsymbol{x}\| \in \mathbb{R}$，这样，对于所有$\lambda \in \mathbb{R}$length且$\boldsymbol{x}, \boldsymbol{y} \in V$，以下情况成立。

- *绝对同质：* $\|\lambda\boldsymbol{x}\| = |\lambda|\|\boldsymbol{x}\|$
- *三角形不等式：* $\|\boldsymbol{x}+\boldsymbol{y}\| \leq \|\boldsymbol{x}\|+\|\boldsymbol{y}\|$
- *正定：* $\|\boldsymbol{x}\| \geq 0$和$\|\boldsymbol{x}\| = 0 \iff \boldsymbol{x} = \boldsymbol{0}$

绝对同质化

三角形不等式正定

图3.2 三角形不平等。



在几何学方面，三角形不等式指出，对于任何三角形，任何两边的长度之和必须大于或等于剩余边的长度；见图3.2的说明。定义3.1是以一般向量空间V为单位的（第2.4回顾一下，对于一个向量$\boldsymbol{x} \in \mathbb{R}^n$我们用一个下标来表示该向量的元素，即$x_i$是向量$\boldsymbol{x}$的$i^{\text{th}}$个元素。

例子（3.1曼哈顿规范）

对于$\boldsymbol{x} \in \mathbb{R}^n$的$n$*曼哈顿准则*定义为

Manhattan norm

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|, \tag{3.3}$$

其中$|\cdot|$是绝对值。图中的左侧面板3.3显示了所有向量$\boldsymbol{x} \in \mathbb{R}^2$，$\|\boldsymbol{x}\|_1 = 1$。曼哈顿规范也被称为$\ell_1$*规范*。

例子（3.2欧几里德准则）

Euclidean norm

$\boldsymbol{x} \in \mathbb{R}^n$的*欧几里得规范*被定义为

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}} \tag{3.4}$$

欧几里得距离

并计算出$\boldsymbol{x}$与原点的*欧几里得距离*。图中的右面板3.3显示了所有的向量$\boldsymbol{x} \in \mathbb{R}^2$，$\|\boldsymbol{x}\|_2 = 1$。欧几里得的

$f!_2$规范

准则也被称为$/!_2$*准则*。

备注。在本书中，如果没有特别说明，我们将默认使用欧几里得准则(3.4)，如果没有特别说明的话，我们将默认使用。 ◆

3.2 内部产品

内积允许引入直观的几何概念，如一个向量的长度和两个向量之间的角度或距离。内积的一个主要目的是确定矢量是否相互正交。

3.2.1 点状产品

标量积 点积

我们可能已经熟悉了一种特殊类型的内积，即

在$\mathbb{R}$中的^n*标量积/点积*，其公式为

$$\boldsymbol{x}^\top \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i \tag{3.5}$$

在本书中，我们将把这种特殊的内积称为点积。然而，内积是更普遍的概念，具有特定的属性，我们现在将介绍这些属性。

3.2.2 一般内部产品

双线性映射

回顾一下本节2.7,中的线性映射，我们可以在加法和与标量相乘的问题上重新安排该映射。*双线性映射*Ω是一个有两个参数的映射，它在每个参数中都是线性的，也就是说，当我们看一个向量空间V时，那么它就会认为

对于所有$\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in V$，$\lambda, \psi \in \mathbb{R}$，即

$$\Omega(\lambda x + \psi y, z) = \lambda\Omega(x, z) + \psi\Omega(y, z) \tag{3.6}$$

$$\Omega(x, \lambda y + \psi z) = \lambda\Omega(x, y) + \psi\Omega(x, z) 。 \tag{3.7}$$

这里，(3.6)断言Ω在第一个参数中是线性的，而(3.7)断言Ω在第二个参数中是线性的（也见(2.87)）.

定义 让**3.2.***V*是一个向量空间，Ω：*VVR*是一个双线性映射，它取两个向量并将它们映射到一个实数上。那么

- 如果Ω（x，*y*）=Ω（y，***x***），对所有的x，*yV*来说，∈Ω被称为*对称的*，也就是说，对称的论点的顺序并不重要。
- 如果是正定的，则称为*正定的*。

$$\forall \boldsymbol{x} \in V \setminus \{\mathbf{0}\} : \Omega(\mathrm{x}, \boldsymbol{x}) > 0 \text{ , } \Omega(\mathbf{0}, \mathbf{0}) = 0 \tag{3.8}$$

定义 让**3.3.***V*是一个向量空间，Ω：*VVR*是一个双线性映射，它取两个向量并将它们映射到一个实数上。那么

- 一个正定的、对称的双线性映射Ω：V×V→R被称为*V*上的一个*内积*。我们通常写（***x***，***y***），而不是Ω（x，*y*）。内积
- 这对（*V*，（–，–））被称为*内积空间*或（实）*向量空间与内积*。如果我们使用定义在(3.5)，我们称(*V*,(–,–))一个*欧几里得向量空间*。

  在本书中，我们将把这些空间称为内积空间。

内积空间

有内积的向量空间

欧几里得向量空间

例子（不是点积的3.3内积）。

考虑*V*=R^2，如果我们定义

$$(\boldsymbol{x}, \boldsymbol{y}) := {}_1xy_1 - (xy_{12} + xy_{21}) + 2xy_{22} \tag{3.9}$$

则（–，–）是一个内积，但与点积不同。该证明将是一个练习。

3.2.3 对称、正定矩阵

对称的正定矩阵在机器学习中发挥着重要作用，它们是通过内积来定义的。在第4.3节中，我们将在矩阵分解的背景下重新讨论对称正定矩阵。对称正半定矩阵的概念是内核定义的关键（第12.4节）。

考虑一个***n***维矢量空间*V*，其内积（–，–）：*V*×*V*→R（见定义3.3）和*V*的有序基*B*=（$\boldsymbol{b}_1$，…，$\boldsymbol{b}_n$）。回顾一下2.6.1任何向量***x***，***y***∈V可以写成基础向量的线性组合，因此，$\boldsymbol{x} = \sum_{i=}^{n} \psi_i \boldsymbol{b}$ $i \in V$和$\boldsymbol{y} = \sum_{j=}^{n} \lambda_j$ $j \in V$为合适的ψ_i，λ $j \in$R。由于双线性的内积，对所有的***x***，***y***∈*V*来说，都认为

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \Big\langle \sum_{i=1}^{n} \psi_i \boldsymbol{b}_i , \sum_{j=1}^{n} \lambda_j \boldsymbol{b}_j \Big\rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_i \langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle \lambda_j = \hat{\boldsymbol{x}}^\top \boldsymbol{A} \hat{\boldsymbol{y}}, \quad (3.10)$$

。

其中$A_{ij} := \langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle$，$\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}$是$\boldsymbol{x}$和$\boldsymbol{y}$相对于基$B$的坐标。这意味着内积 $\langle \cdot, \cdot \rangle$ 是通过$\boldsymbol{A}$唯一确定的。

是对称的。此外，内积的正定性意味着

$$\forall \boldsymbol{x} \in V \setminus \{\mathbf{0}\} : \boldsymbol{x}^{\top} \boldsymbol{A} \boldsymbol{x} > 0 \text{。} \tag{3.11}$$

定义（3.4对称、正定矩阵）。一个对称的矩阵满足（）的$\boldsymbol{A} \in \mathbb{R}$被称为对称的、正定的或$n \times n$3.11)被称为*对称*、*正定*、或只是*正确定的*。如果只有$\geq$在(3.11)，那么$\boldsymbol{A}$就被称为*对称的*。*正半定理*。

对称的、正的确定的

正定

symmetric, positive semidefinite

例子（3.4对称的、正定的矩阵）

考虑到矩阵

$$\boldsymbol{A}_1 = \begin{bmatrix} 9 & 6 \\ 6 & 5 \end{bmatrix}, \quad \boldsymbol{A}_2 = \begin{bmatrix} 9 & 6 \\ 6 & 3 \end{bmatrix}. \tag{3.12}$$

$\boldsymbol{A}_1$是正定的，因为它是对称的和

$$\boldsymbol{x}^{\top} \boldsymbol{A}_1 \boldsymbol{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 9 & 6 \\ 6 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{3.13a}$$

$$= 9x_1^2 + 12 x_1 x_2 + 5 x_2^2 = (3x_1 + 2x_2)^2 + x_2^2 > 0 \tag{3.13b}$$

对于所有的$\boldsymbol{x} \in V \setminus \{\mathbf{0}\}$。相比之下，$\boldsymbol{A}_2$是对称的，但不是正定的因为$\boldsymbol{x}^{\top} \boldsymbol{A} \boldsymbol{x} = 9x_1^2 + 12 x_1 x_2 + 3 x_2^2 = (3x_1 + 2x_2)^2 - x_2^2$可以小于比0，例如，对于$\boldsymbol{x} = [2, -3]^{\top}$。

如果$\boldsymbol{A} \in \mathbb{R}^{n \times n}$是对称的、正定的，那么

$$(\boldsymbol{x}, \boldsymbol{y}) = \hat{\boldsymbol{x}}^{\top} \boldsymbol{A} \hat{\boldsymbol{y}} \tag{3.14}$$

定义了一个相对于有序基础$\boldsymbol{B}$的内积，其中$\hat{\boldsymbol{x}}$和$\hat{\boldsymbol{y}}$是$\boldsymbol{x}, \boldsymbol{y} \in V$相对于$\boldsymbol{B}$的坐标表示。

定理3.5。*对于一个实值的有限维向量空间V和V的有序基B，当且仅当存在一个对称的正定矩阵$\boldsymbol{A} \in \mathbb{R}^{n \times n}$时，可以认为$(-, -) : V \times V \to \mathbb{R}$是一个内积。*

$$(\boldsymbol{x}, \boldsymbol{y}) = \hat{\boldsymbol{x}}^{\top} \boldsymbol{A} \hat{\boldsymbol{y}} \text{。} \tag{3.15}$$

如果$\boldsymbol{A} \in \mathbb{R}^{n \times n}$是对称的和正定的，以下属性成立。

- $\boldsymbol{A}$的无效空间（内核）只包括因为$\mathbf{0}^{\top} \boldsymbol{A} \boldsymbol{x} > 0$对于所有$\boldsymbol{x} \neq \mathbf{0}$。这意味着如果$\boldsymbol{0x} \neq \mathbf{0}$，$\boldsymbol{Ax} \neq$。
- $\boldsymbol{A}$的对角线元素a_{ii}是正的，因为$a_{ii} = \boldsymbol{e}_i^{\top} \boldsymbol{A} \boldsymbol{e}_i > 0$，其中$\boldsymbol{e}_i$是$\mathbb{R}$中$n$标准基的第i个向量。

3.3 长度和距离

在第3.1,我们已经讨论了我们可以用来计算向量长度的规范。内积和规范在以下方面密切相关
sense that any inner product induces a norm

Inner products induce norms.

$$\|\boldsymbol{x}\| := \sqrt{(\boldsymbol{x}, \boldsymbol{x})} \tag{3.16}$$

以一种自然的方式，这样我们就可以用内积来计算向量的长度了。然而，并不是每个规范都是由内积引起的。曼哈顿规范 (3.3)就是一个没有相应内积的规范的例子。在下文中，我们将重点讨论由内积引起的规范，并介绍几何概念，如长度、距离和角度。

备注（考奇-施瓦茨不等式）。对于一个内积向量空间
$(V, (\cdot))$的诱导 $\|\cdot\|$ 规范满足*考奇-施瓦茨不等式*，

考奇-施瓦茨不平等

$$|(\boldsymbol{x}, \boldsymbol{y})| \leqslant \|\boldsymbol{x}\| \|\boldsymbol{y}\|。 \tag{3.17}$$

◆

例子（3.5使用内积的矢量长度）

在几何学中，我们经常对向量的长度感兴趣。现在我们可以用内积来计算它们，使用(3.16).让我们取$\boldsymbol{x} = [1, 1]^\top \in$ $\mathbb{R}^2$.如果我们用点积作为内积，用(3.16)我们得到

$$\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}} = \sqrt{1^2 + 1^2} = \sqrt{2} \tag{3.18}$$

。

现在让我们选择一个不同的内积，作为$\boldsymbol{x}$的长度。

$$(\boldsymbol{x}, \boldsymbol{y}) := \boldsymbol{x}^\top \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \boldsymbol{y} = x_1 y_1 - \frac{1}{2}(x_1 y_2 + x_2 y_1) + x_2 y_2 \tag{3.19}$$

。

如果我们计算一个向量的规范，那么如果x_1和x_2有相同的符号（并且$x_1 x_2 > 0$），那么这个内积返回的值比点积小；否则，它返回的值比点积大。通过这个内积，我们可以得到

$$(\boldsymbol{x}, \boldsymbol{x}) = x_1^2 - x_1 x_2 + x_2^2 = 1 - 1 + 1 = 1 \quad \Rightarrow \quad \|\boldsymbol{x}\| = \sqrt{1} = 1 \tag{3.20}$$

这样，$\boldsymbol{x}$在这个内积下比在点积下 "更短"。

定义（3.6距离和公制）。考虑一个内积空间
$(V, (\cdot, \cdot))$. Then

$$d(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{x} - \boldsymbol{y}\| = \sqrt{(\boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y})} \tag{3.21}$$

被称为$\boldsymbol{x}$和$\boldsymbol{y}$之间的*距离*，对于$\boldsymbol{x}, \boldsymbol{y} \in V$。如果我们使用dotdistance的内积，那么这个距离就叫做*欧几里得距离*。

解析欧几得距离

绘图

$$d : V \times V \to \mathbb{R} \tag{3.22}$$
$$(\boldsymbol{x}, \boldsymbol{y}) \mapsto d(\boldsymbol{x}, \boldsymbol{y}) \tag{3.23}$$

公制

被称为*公制*。

*备注。与向量*的长度类似，向量之间的距离不需要内积：有一个规范就足够了。如果我们有一个由内积引起的规范，那么距离可能会因以下因素而变化
内积 的选择。 ◆

正定对称

一个公制d满足以下条件。

1. d是*正定的*，即$d(\boldsymbol{x}, \boldsymbol{y}) \geq 0$对所有$\boldsymbol{x}, \boldsymbol{y} \in V$和$d(\boldsymbol{x}, \boldsymbol{y}) = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$。
2. d是*对称的*，即对于所有的$\boldsymbol{x}, \boldsymbol{y} \in V$，$d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$。

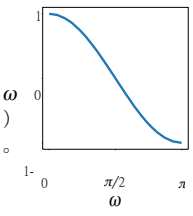
三角不等式

3. *三角不等式*：$d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$，对所有$\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in V$。

*备注。*乍一看，内积和元的属性列表看起来非常相似。然而，通过比较定义3.3和定义3.6我们注意到，$\boldsymbol{x}$、$\boldsymbol{y}$和$d(\boldsymbol{x}, \boldsymbol{y})$的行为方向是相反的。
非常相似的$\boldsymbol{x}$和$\boldsymbol{y}$将导致内积的大值和公积的小值。 ◆

3.4 角度和正交性

图3.4 当限制在$[0, \pi]$时，那么$f(\omega) = \cos(\omega)$在区间[-1, 1]中返回一个唯一的数字。



除了能够定义向量的长度以及两个向量之间的距离之外，内积还通过定义两个向量之间的角度ω来捕捉向量空间的几何。我们使用考奇-施瓦茨不等式(3.17)来定义内积空间中两个向量$\boldsymbol{x}$、$\boldsymbol{y}$之间的角度ω，这个概念与我们的
在$\mathbb{R}^2$和$\mathbb{R}^3$中的直觉，假设$\boldsymbol{x} \neq \boldsymbol{0}, \boldsymbol{y} \neq \boldsymbol{0}$.然后

$$-1 \leq \frac{(\boldsymbol{x}, \boldsymbol{y})}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|} \leq 1 \tag{3.24}$$

因此，存在一个唯一的$\omega \in [0, \pi]$，如图所示3.4,。

$$\cos \omega = \frac{(\boldsymbol{x}, \boldsymbol{y})}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|} . \tag{3.25}$$

角度

数字ω是向量$\boldsymbol{x}$和$\boldsymbol{y}$之间的*角度*。直观地说，两个向量之间的角度告诉

它们的方向是相同的。

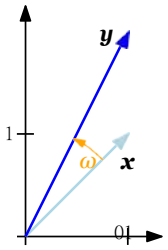
我们它们的方向有多相似。例如，使用点积，$\boldsymbol{x}$和$\boldsymbol{y}=\boldsymbol{4x}$之间的角度，即$\boldsymbol{y}$是$\boldsymbol{x}$的缩放版本，是0。

例子（矢量之间的3.6角度）
让我们计算$\boldsymbol{x} = [1, 1]^\top \in \mathbb{R}^2$和$\boldsymbol{y} = [1, 2]^\top \in \mathbb{R}^2$之间的角度。
见图3.5,，我们用点积作为内积。然后我们得到

$$\cos\omega = \frac{(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{(\boldsymbol{x}, \boldsymbol{x})\,(\boldsymbol{y}, \boldsymbol{y})}} = \frac{\boldsymbol{x}\boldsymbol{y}^\top}{\sqrt{}} = \frac{3}{\sqrt{10}}, \tag{3.26}$$

而两个向量之间的角度为$\frac{3}{\sqrt{10}}\arccos() \approx$ ，其中
相当于约0.32...rad 18°.

图表
两个向量$\boldsymbol{x}, \boldsymbol{y}$是使用内积计算的。



orthogonal
正交的

内积的一个关键特征是，它还允许我们对正交的向量进行定性。

定义（3.7正交性）。两个向量$\boldsymbol{x}$和$\boldsymbol{y}$是*正交的*，如果和
只有当x，$\boldsymbol{y}$=0，我们才写$\boldsymbol{x}\boldsymbol{y}$。如果再加上$\boldsymbol{x} = = 1\boldsymbol{y}$。()⊥11 1
即，矢量是单位矢量，那么$\boldsymbol{x}$和$\boldsymbol{y}$是*正交的*。

这个定义的一个含义是，**0**-向量与向量空间中的每个向量都是正交的
。

*备注。*正交性是对双线性形式的过穹性概念的概括，不一定是点积。在我们的背景下，从几何学的角度来看，我们可以认为正交的矢量有一个
◆ 相对于一个特定的内积来说，是一个直角。 ◆

例子（3.7正交向量）

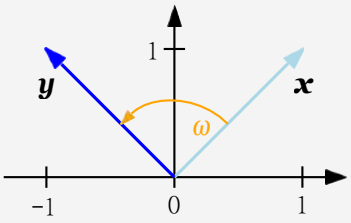


Figure 3.6 The angle ω between
两个向量$\boldsymbol{x}, \boldsymbol{y}$可以变化取决于内积。

考虑两个向量$\boldsymbol{x} = [1, 1]^\top, \boldsymbol{y} = [-1, 1]^\top \in \mathbb{R}^2$；见图3.6.
我们感兴趣的是用两种不同的内积来确定它们之间的角度ω。使用点积作为内积，可以得到
$\boldsymbol{x}$和$\boldsymbol{y}$之间有90°一个角度ω，使得$\boldsymbol{x} \perp \boldsymbol{y}$。
选择内积

$$(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \boldsymbol{y}, \tag{3.27}$$

我们得到，$\boldsymbol{x}$和$\boldsymbol{y}$之间的角度ω由以下公式给出

$$\cos\omega = \frac{(\boldsymbol{x},\boldsymbol{y})}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} = -\frac{1}{3} \Rightarrow \omega \approx 1.91\text{rad} \approx 109.5^\circ, \tag{3.28}$$

而$\boldsymbol{x}$和$\boldsymbol{y}$并不是正交的。因此，就一个内积而言是正交的向量不一定就会与另一个内积正交。

正交矩阵

定义（3.8正交矩阵）。一个方形矩阵$A \in \mathbb{R}^{n\times n}$是一个*正交矩阵，*当且仅当其列是正交的，以便

$$\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{I} = \boldsymbol{A}^\top\boldsymbol{A}, \tag{3.29}$$

这意味着

$$\boldsymbol{A}^{-1} = \boldsymbol{A}^\top, \tag{3.30}$$

按照惯例，这些矩阵被称为 "正交"，但更准确的描述是 "正交"。用正交矩阵进行的变换保留了距离和角度。

即通过简单的转置矩阵来获得逆值。

用正交矩阵进行变换是很特别的，因为用正交矩阵$\boldsymbol{A}$进行变换时，向量$\boldsymbol{x}$的长度不会改变。对于点积，我们得到

$$\|\boldsymbol{A}\boldsymbol{x}\|^2 = (\boldsymbol{A}\boldsymbol{x})^\top(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^\top\boldsymbol{I}\boldsymbol{x} = \boldsymbol{x}^\top\boldsymbol{x} = \|\boldsymbol{x}\|^2. \tag{3.31}$$

此外，任何两个向量$\boldsymbol{x}$、$\boldsymbol{y}$之间的角度，由它们的内积衡量，在使用正交矩阵$\boldsymbol{A}$对它们进行变换时也是不变的。假设点积为内积，图像$\boldsymbol{A}\boldsymbol{x}$和$\boldsymbol{A}\boldsymbol{y}$的角度为

$$\cos\omega = \frac{(\boldsymbol{A}\boldsymbol{x})^\top(\boldsymbol{A}\boldsymbol{y})}{\|\boldsymbol{A}\boldsymbol{x}\|\|\boldsymbol{A}\boldsymbol{y}\|} = \frac{\boldsymbol{x}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{y}}{\sqrt{\boldsymbol{x}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x}\boldsymbol{y}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{y}}} = \frac{\boldsymbol{x}\boldsymbol{y}^\top}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}, \tag{3.32}$$

这意味着$\boldsymbol{A}^\top\boldsymbol{A}=\boldsymbol{I}$的正交矩阵同时保留了角度和距离。事实证明，正交矩阵定义的变换是旋转的。

在这一节中，我们可以看到，有很多人都在使用 "翻转 "这个词，这也是一种 "翻转 "的可能性。在第3.9,我们将讨论关于旋转的更多细节。

3.5 正态基础

在这一节中2.6.1,我们描述了基向量的特性，并发现在一个n维的向量空间中，我们需要n个基向量，也就是n个线性独立的向量。在第3.3和3.4节中，我们用内积来计算向量的长度和向量之间的角度。在下文中，我们将讨论基向量相互正交的特殊情况，每个基向量的长度为 1。

我们将把这种基称为正交基。

让我们更正式地介绍这一点。

定义3.9（正态基）。考虑一个n维向量空间V和V的基$\{\boldsymbol{b}_1$。 $, \boldsymbol{b}_n\}$。如果

$$(\boldsymbol{b}_i, \boldsymbol{b}_j) = 0 \quad \text{对于} \boldsymbol{i} \neq \boldsymbol{j} \tag{3.33}$$

$$(\boldsymbol{b}_i, \boldsymbol{b}_i) = 1 \tag{3.34}$$

对于所有$i, j = 1, \ldots n$，那么这个基就被称为*正态基*（*ONB*）。如果只有(3.33)得到满足，那么这个基就被称为*正交基*。请注意，(3.34)意味着每个基向量都有长度/1规范.

正态基
ONB
正交基础

回顾一下第2.6.1我们可以用高斯消除法为一组向量所跨越的向量空间找到一个基。假设我们得到了一组$\{\tilde{\boldsymbol{b}}_1, \ldots, \tilde{\boldsymbol{b}}_n\}$的非正交和非正常化的基向量。我们将它们串联成一个矩阵$\tilde{\boldsymbol{B}} = [\tilde{\boldsymbol{b}}_1, \ldots, \tilde{\boldsymbol{b}}_n]$，并应用高斯消除法。对增强的矩阵(节)2.3.2$[\tilde{\boldsymbol{B}}\, \tilde{\boldsymbol{B}}^\top]$进行计算，得到一个正态基。这种迭代建立正交基$\{\boldsymbol{b}_1, \ldots \boldsymbol{b}_n\}$被称为*Gram-Schmidt过程*（Strang，2003）。

例子（3.8正交基数）

欧几里得向量空间R的^{n}典型/标准基础是一个正态基础，其中内积是向量的点积。

在R中^2，向量

$$\boldsymbol{b}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{b}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{3.35}$$

构成一个正态基础，因为$\boldsymbol{b}_1^\top \boldsymbol{b}_2 = 0$和$\|\boldsymbol{b}_1\| = 1 = \|\boldsymbol{b}_2\|$。

我们将在第12章和第10章讨论支持向量机和主成分分析时利用正态基的概念。

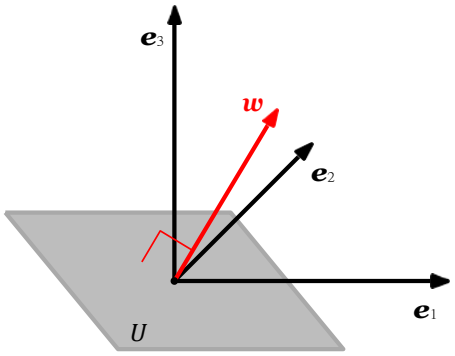
3.6 正交互补

在定义了正交性之后，我们现在将研究相互正交的向量空间。这将在第十章中发挥重要作用，届时我们将从几何学的角度讨论线性降维问题。

考虑一个D维的矢量空间V和一个M维的次空间$U \subseteq V$。那么它的*正交补集*$U^\perp$是一个$(D-M)$维的正交空间。

V的子空间，包含了V中所有与U中每个矢量正交的矢量。 此外，
$U \cap U = \{\mathbf{0}\}$，所以任何矢量$\boldsymbol{x} \in V$都可以是

补充



图A 3.7平面U在一个三维矢量空间可以用其法向量来描述，法向量跨越其正交补数$U^{\perp}$。

唯一地分解为

$$\boldsymbol{x} = \sum_{m=1}^{M} \lambda_m \boldsymbol{b}_m + \sum_{j=1}^{D-M} \psi_j \boldsymbol{b}_j^{\perp}, \quad \lambda_m, \psi_j \in \mathbb{R} 。 \tag{3.36}$$

其中$(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M)$是$U$的一个基，$(\boldsymbol{b}_1^{\perp}, \ldots, \boldsymbol{b}_{D-M}^{\perp})$是$U^{\perp}$的一个基。

因此，正交补数也可以用来描述一个平面U（二维子空间）的三维矢量空间。更具体地说，与平面U正交的矢量$\boldsymbol{w}$ $\|\boldsymbol{w}\| = 1$，是U的基矢量。3.7说明了这一设置。所有与$\boldsymbol{w}$正交的向量都必须（根据结构）位于平面U中。 向量$\boldsymbol{w}$被称为U的*法向量*。

法向量

一般来说，正交补数可以用来描述n维矢量空间和仿生空间中的超平面。

3.7 函数的内积

到目前为止，我们研究了内积计算长度、角度和距离的特性。我们关注的是有限维向量的内积。下面，我们将看一个不同类型向量的内积的例子：函数的内积。

到目前为止，我们所讨论的内积是针对具有有限个条目的向量而定义的。我们可以把一个向量$\boldsymbol{x} \in \mathbb{R}^n$看作是一个有$n$*个函数值*的函数。内积的概念可以推广到有无限个条目的向量（可数无限）和连续值函数（不可数的无限）。然后，向量的各个组成部分的总和（见公式(3.5)变成了一个积分。

两个函数$u : \mathbb{R} \to \mathbb{R}$和$v : \mathbb{R} \to \mathbb{R}$的内积可以定义为定积分

$$(u, v) := \int_a^b u(x)v(x)dx \tag{3.37}$$

分别为下限和上限$a, b < \infty$。就像我们通常的内积一样，我们可以通过观察内积来定义规范和正交性。如果(3.37)的值为 0，则函数u和v是正交的。为了使前面的内积在数学上精确，我们需要照顾到度量和积分的定义，从而导致希尔伯特空间的定义。此外，与有限维向量的内积不同，函数的内积可能会发散（有无限的值）。所有这些都需要深入研究实数和函数分析的一些更复杂的细节，我们在本书中没有涉及。

例子 (函数的3.9内积)

如果我们选择$u = \sin(x)$和$v = \cos(x)$，积分$f(x) = u(x)v(x)$的(3.37)，f如图所示3.8.我们看到过程中，函数是奇数，即$f(-x) = -f(x)$。因此，$\int_{-\pi}^{0} f(x)\,dx = -\int_{0}^{\pi} f(x)\,dx$，即$\pi$的积分因此，sin和cos是正交函数0。

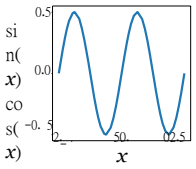


图3.8 $f(x) = \sin(x)\cos(x)$。

备注。 还可以认为，函数的集合

$$\{1, \cos(x), \cos(2x), \cos(3x), \ldots\} \tag{3.38}$$

如果我们从π到π进行积分，则是正交的，即任何一对函数都是相互正交的。在(3.38)中的函数集合跨越了一个很大的函数子空间，这些函数在$[\pi, \pi)$上是偶数和周期性的，将函数投射到这个子空间上是( ) 的基本思想。

傅里叶数列。 ◆

在第6.4.6节中，我们将看看第二类非常规的内积：随机变量的内积。

3.8 正交投射

投影是一类重要的线性变换（除了旋转和反射），在图形学、编码学、统计学和机器学习中发挥着重要作用。在机器学习中，我们经常处理高维的数据。高维数据往往难以分析或可视化。然而，高维数据往往具有这样的特性：只有少数维度包含大多数信息，而大多数其他维度对于描述数据的关键属性并不重要。当我们对高维数据进行压缩或可视化时，我们将失去信息。为了尽量减少这种压缩损失，我们最好能找到数据中最有信息量的维度。正如第一章中所讨论的，"特征 "是一个关模式。 例如，机器数据可以被表示为向量，在本章中，我们将讨论一些数据压缩的基本工具。更具体地说，我们可以将原始的高维数据投射到一个低维的特征空间，并在这个低维空间中工作，以了解更多关于数据集的信息并提取相

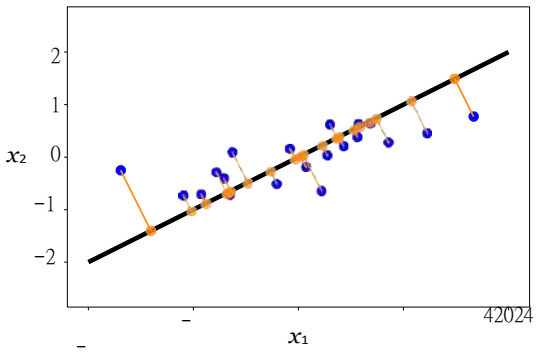


图 正交投影**3.9**（橙色点）的二维数据集(蓝点)上的数据。一维子空间（直线）。

学习算法，如Pear-son(1901)和Hotelling(1933)的主成分分析(PCA)和深度神经网络(如深度自动编码器(Deng等人,2010))，大量利用了降维的思想。在下文中，我们将专注于正交投影，我们将在第十章中用于线性降维，在第十二章中用于分类。即使是我们在第九章讨论的线性回归，也可以用正交投影来解释。对于一个给定的低维子空间，高维数据的正交投影保留了尽可能多的信息，并使原始数据和相应投影之间的差异/误差最小。图中给出了这样一个正交投影的说明 在我们详细说明如何获得这些投影3.9.之前，让我们先定义一下投影到底是什么。

定义3.10（投影）。设V是一个矢量空间，$U \subseteq V$是V的一个子空间。如果$\pi^2 = \pi \circ \pi = \pi$，一个线性映射$\pi : V \to U$被称为*投影*。

投影

由于线性映射可以用变换矩阵来表示（见第2.7节），前面的定义同样适用于一种特殊的变换矩阵，即*投影矩阵*$\boldsymbol{P}_\pi$，它表现出$\boldsymbol{P} = \boldsymbol{P}^2$的特性$\pi$。

在下文中，我们将推导出向量的正交投影。

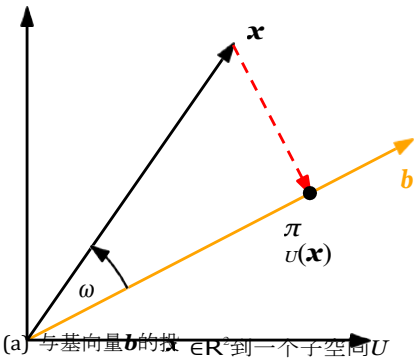
投影矩阵

内积空间$(\mathbb{R}^n, \cdot, \cdot)$到子空间$U \subseteq \mathbb{R}^n$。我们将从一维的子空间开始，它们也被称为*线*。如果没有另外提到，我们假定点积$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}\boldsymbol{y}^\top$为内积。

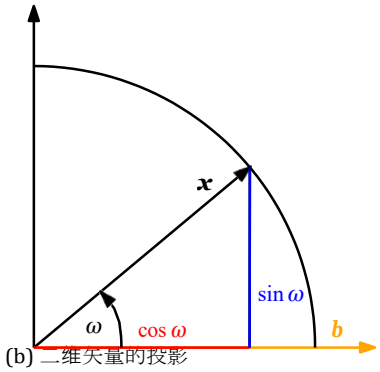
线

3.8.1 投射到一维子空间（线）上

假设我们得到一条通过口的线（一维子空间），其基础向量$\boldsymbol{b} \in \mathbb{R}^n$。该线是$\boldsymbol{b}$所跨越的一维子空间$U \subseteq \mathbb{R}^n$。当我们把$\boldsymbol{x} \in \mathbb{R}$投射^{n}到$U$上时，我们寻求最接近$\boldsymbol{x}$的向量$\pi_U(\boldsymbol{x}) \in U$。



(a) 与基向量$\boldsymbol{b}$的投影$\boldsymbol{x} \in \mathbb{R}^2$到一个子空间$U$影。



(b) 二维矢量的投影$\boldsymbol{x}$与$\|\boldsymbol{x}\| = 1$到一个一维的所跨越的子空间$\boldsymbol{b}$.

图为投射到一维子空间的例子**3.10**。

我们描述一下投影$\pi_U(\boldsymbol{x})$的一些特性（图3.10(a)是一个说明）。

- 投影$\pi_U(\boldsymbol{x})$最接近$\boldsymbol{x}$，其中"最接近"意味着距离$\boldsymbol{x}\pi_U(\boldsymbol{x})$最小。因此，从$\pi(\boldsymbol{x})$到$\boldsymbol{x}$的一段$\pi_U(\boldsymbol{x})\boldsymbol{x}$与$U$正交，因此是$U$的基向量$\boldsymbol{b}$。 正交条件得到$\pi_U(\boldsymbol{x})\boldsymbol{x}$，$\boldsymbol{b}$=因为0角度
- $\boldsymbol{x}$对U的投影$\pi(\boldsymbol{x})$必须是U的一个元素，并且，在这里
- 因此，$\pi_U(\boldsymbol{x}) = \lambda\boldsymbol{b}$，对于某个$\lambda \in \mathbb{R}$。

λ是$\pi(\boldsymbol{x})$的坐标

关于.....的问题。U $\boldsymbol{b}$ $\boldsymbol{x}$

在以下三个步骤中，我们确定坐标λ、投影$\pi_U(\boldsymbol{x}) \in U$，以及将任何$\boldsymbol{x} \in \mathbb{R}^n$映射到$U$的投影矩阵$\boldsymbol{P}$。

1.找出坐标λ，正交条件可得

$$\langle \boldsymbol{x} - \pi_U(\boldsymbol{x}), \boldsymbol{b} \rangle = 0 \iff \langle \boldsymbol{x} - \lambda\boldsymbol{b}, \boldsymbol{b} \rangle = 0 \tag{3.39}$$

现在我们可以利用内积的双线性，得出

$$\langle \boldsymbol{x}, \boldsymbol{b} \rangle - \lambda \langle \boldsymbol{b}, \boldsymbol{b} \rangle = 0 \iff \lambda = \frac{\langle \boldsymbol{x}, \boldsymbol{b} \rangle}{\langle \boldsymbol{b}, \boldsymbol{b} \rangle} = \frac{\langle \boldsymbol{b}, \boldsymbol{x} \rangle}{\|\boldsymbol{b}\|^2} . \tag{3.40}$$

有一般内积的W乘积，我们得到$\lambda = \langle \boldsymbol{x}, \boldsymbol{b} \rangle$女$\|\boldsymbol{b}\| = 1$.

在最后一步，我们利用了内积是对称的这一事实。如果我们选择$\langle -, - \rangle$为点积，我们得到

$$\lambda = \frac{\boldsymbol{b}\boldsymbol{x}\boldsymbol{b}\boldsymbol{x}^{\top}}{\boldsymbol{b}\boldsymbol{T}\boldsymbol{b}} = \frac{\boldsymbol{b}\boldsymbol{x}}{\|\boldsymbol{b}\|^2} \tag{3.41}$$

如果$\|\boldsymbol{b}\| = 1$，那么投影的坐标λ由$\boldsymbol{b}\boldsymbol{x}$给出$^{\top}$。

2. 找到投影点$\pi_U(\boldsymbol{x})U$。 由于$\pi_U(\boldsymbol{x}) = \lambda\boldsymbol{b}$，我们马上就可以得到(3.40)，即

$$\pi_U(\boldsymbol{x}) = \lambda\boldsymbol{b} = \frac{(\boldsymbol{x}, \boldsymbol{b})}{\|\boldsymbol{b}\|^2}\boldsymbol{b} = \frac{\boldsymbol{b}^T\boldsymbol{x}}{\|\boldsymbol{b}\|^2}\boldsymbol{b}, \tag{3.42}$$

其中最后一个等式只对点积成立。我们也可以通过定义来计算π_U（$\boldsymbol{x}$）的长度3.1为

$$\|\pi_U(\boldsymbol{x})\| = \|\lambda\boldsymbol{b}\| = |\lambda|\,\|\boldsymbol{b}\|。 \tag{3.43}$$

因此，我们的投影长度是$\boldsymbol{b}$长度的λ倍。这也增加了一个直觉，即λ是$\pi_U(\boldsymbol{x})$相对于跨越我们一维子空间U的基向量$\boldsymbol{b}$的坐标。

如果我们用点积作为内积，我们可以得到

$$\|\pi_U(\boldsymbol{x})\| \overset{(3.42)}{=} \frac{|\boldsymbol{b}^T\boldsymbol{x}|}{\|\boldsymbol{b}\|^2}\|\boldsymbol{b}\| \overset{(3.25)}{=} |\cos\omega|\,\|\boldsymbol{x}\|\,\|\boldsymbol{b}\|\frac{\|\boldsymbol{b}\|}{\|\boldsymbol{b}\|^2} = |\cos\omega|\,\|\boldsymbol{x}\|。 \tag{3.44}$$

这里，ω是$\boldsymbol{x}$和$\boldsymbol{b}$之间的角度。这个方程在三角学中应该很熟悉。如果$\boldsymbol{x}$=1，那么$\boldsymbol{x}$就位于单位圆上。因此，在$\boldsymbol{b}$所跨越的水平轴上的投影正好是cos ω，相应的矢量π_U（$\boldsymbol{x}$）的长度=cos ω。图3.10(b)中给出了一个说明。

横轴是一个一维的子空间。

3. 找到投影矩阵$\boldsymbol{P}_\pi$。 我们知道，投影是一种线耳映射（见定义3.10).因此，存在一个投影矩阵$\boldsymbol{P}_\pi$，使得$\pi(_U\boldsymbol{x})=\boldsymbol{P}_\pi\boldsymbol{x}$.以点积作为内积和

$$\pi_U(\boldsymbol{x}) = \lambda\boldsymbol{b} = \boldsymbol{b}\lambda = \boldsymbol{b}\frac{\boldsymbol{b}\boldsymbol{x}^T}{\|\boldsymbol{b}\|^2} = \frac{\boldsymbol{b}\boldsymbol{b}^T}{\|\boldsymbol{b}\|^2}\boldsymbol{x}, \tag{3.45}$$

我们立即看到，

$$\boldsymbol{P}_\pi = \frac{\boldsymbol{b}\boldsymbol{b}^T}{\|\boldsymbol{b}\|^2}. \tag{3.46}$$

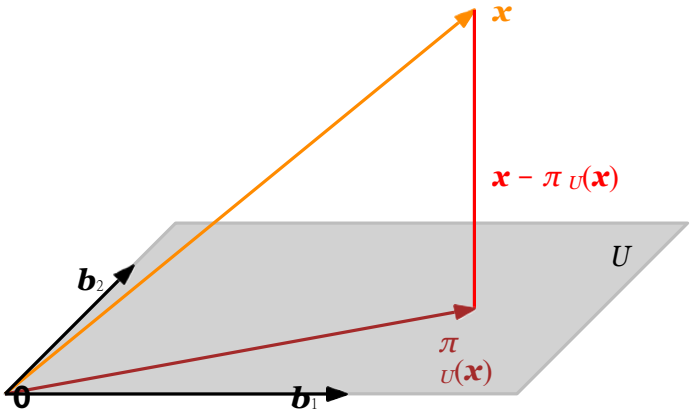
投影矩阵总是对称的。

请注意，$\boldsymbol{b}\boldsymbol{b}^T$（以及，因此，$\boldsymbol{P}_\pi$）是一个对称矩阵（等级为1)，而$\|\boldsymbol{b}\|^2=(\boldsymbol{b}, \boldsymbol{b})$是一个标量。

投影矩阵$\boldsymbol{P}_\pi$将任何矢量$\boldsymbol{x}\in\mathbb{R}^n$投射到通过原点的方向为$\boldsymbol{b}$的直线上（等同于$\boldsymbol{b}$所跨越的子空间$U$）。

备注。 投影$\pi_U(\boldsymbol{x})\in\mathbb{R}^n$仍然是一个$n$维向量，而不是一个标量。然而，我们

不再需要n个坐标来表示投影，而只需要一个坐标，如果我们想相对于跨越子空间U：λ的基向量$\boldsymbol{b}$来表达它



图：投影3.11到一个二维子空间U上，有基础$\boldsymbol{b}_1, \boldsymbol{b}_2$。的。投影$\pi$ (x)的$\boldsymbol{x} \in \mathbb{R}^3$对$U$可表示为线性组合和的$\boldsymbol{b}_1, \boldsymbol{b}_2$位移向量$\boldsymbol{x} - \pi_U(\boldsymbol{x})$是正交于两者$\boldsymbol{b}_1$和$\boldsymbol{b}_2$。

例子（3.10投射到直线上）。

求投影矩阵$\boldsymbol{P}$到$\pi$$\boldsymbol{b}$]=1 2 2所跨越的通过原点的线上。$\boldsymbol{b}$是一个方向和一维子空间（通过原点的线）的基础。

通过(3.46)，我们得到

$$\boldsymbol{P} = \frac{\boldsymbol{b}\boldsymbol{b}^\top}{\boldsymbol{b}^\top\boldsymbol{b}} = \frac{1}{9}\begin{bmatrix}1\\2\\2\end{bmatrix}\begin{bmatrix}1 & 2 & 2\end{bmatrix} = \frac{1}{9}\begin{bmatrix}1 & 2 & 2\\2 & 4 & 4\\2 & 4 & 4\end{bmatrix}. \tag{3.47}$$

现在让我们选择一个特定的$\boldsymbol{x}$，看看它是否位于子空间中对于$\boldsymbol{x}$ = 1 1]T1，投影为

$$\pi_U(\boldsymbol{x}) = \boldsymbol{P}\boldsymbol{x} = \frac{1}{9}\begin{bmatrix}1 & 2 & 2\\2 & 4 & 4\\2 & 4 & 4\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix} = \frac{1}{9}\begin{bmatrix}5\\10\\10\end{bmatrix} \in \operatorname{span}\left[\begin{bmatrix}1\\2\\2\end{bmatrix}\right]. \tag{3.48}$$

请注意，$\boldsymbol{P}$对$\pi$$\pi_U$（$\boldsymbol{x}$）的应用没有任何改变，即：。$\boldsymbol{P}_\pi \pi_U(\boldsymbol{x}) = \pi_U(\boldsymbol{x})$。这是预期的，因为根据定义3.10，我们知道，一个投影矩阵$\boldsymbol{P}$满足$\pi\boldsymbol{P}\boldsymbol{x}^2 = \boldsymbol{P}_\pi\boldsymbol{x}$，对$\pi$于所有的$\boldsymbol{x}$。

*备注。*利用第四章的结果，我们可以证明π_U（$\boldsymbol{x}$）是$\boldsymbol{P}$的π一个特征向量，相应的特征值是1. ◆

3.8.2 投射到一般子空间

在下文中，我们将研究向量$\boldsymbol{x} \in \mathbb{R}^n$在低维子空间$U \subseteq \mathbb{R}^n$上的正交投影，$\dim(U) = m \geqslant 1$。图中给出了一个说明3.11.

假设$(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m)$是$U$的一个有序基，任何投影$\pi_U(\boldsymbol{x})$因此，它们可以被表示为

如果U是由一组跨度向量给出的，这不是一个基，请确保你确定一个基$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$在继续进行之前。

as linear combinations of the basis vectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$ of U, such that $\pi_U(\boldsymbol{x}) = \sum_{i=1}^{m} \lambda_i \boldsymbol{b}_i$。

这些基向量构成了 $\boldsymbol{B} \in \mathbb{R}^{n\times m}$，其中 $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m]$.

与一维情况一样，我们遵循一个三步程序来找到projec- tion $\pi_U(\boldsymbol{x})$ 和投影矩阵$\boldsymbol{P}_\pi$。

1.找出投影的坐标$\lambda_1, \ldots, \lambda_m$的投影（关于$U$的基础），使得线性组合

$$\pi_U(\boldsymbol{x}) = \sum_{i=1}^{m} \lambda_i \boldsymbol{b}_i = \boldsymbol{B}\boldsymbol{\lambda}, \tag{3.49}$$

$$\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m] \in \mathbb{R}^{n\times m}, \quad \boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_m]^\top \in \mathbb{R}^m, \tag{3.50}$$

与一维情况一样，"最接近 "意味着 "最小距离"，这意味着连接$\pi_U(\boldsymbol{x}) \in U$和$\boldsymbol{x} \in \mathbb{R}^n$的矢量必须与$U$的所有基矢量正交。 因此，我们得到$m$个同时存在的条件（假设点积为内积）。

$$\langle \boldsymbol{b}_1, \boldsymbol{x} - \pi_U(\boldsymbol{x}) \rangle = \boldsymbol{b}_1^\top (\boldsymbol{x} - \pi_U(\boldsymbol{x})) = 0 \tag{3.51}$$

$$\vdots$$

$$\langle \boldsymbol{b}_m, \boldsymbol{x} - \pi_U(\boldsymbol{x}) \rangle = \boldsymbol{b}_m^\top (\boldsymbol{x} - \pi_U(\boldsymbol{x})) = 0 \tag{3.52}$$

，在$\pi_U(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\lambda}$的情况下，可写为

$$\boldsymbol{b}_1^\top (\boldsymbol{x} - \boldsymbol{B}\boldsymbol{\lambda}) = 0 \tag{3.53}$$

$$\vdots$$

$$\boldsymbol{b}_m^\top (\boldsymbol{x} - \boldsymbol{B}\boldsymbol{\lambda}) = 0 \tag{3.54}$$

这样，我们得到一个同质线性方程组

$$\begin{bmatrix} \boldsymbol{b}_1^\top \\ \vdots \\ \boldsymbol{b}_m^\top \end{bmatrix} \boldsymbol{x} - \boldsymbol{B}\boldsymbol{\lambda} = \boldsymbol{0} \Leftrightarrow \boldsymbol{B}^\top (\boldsymbol{x} - \boldsymbol{B}\boldsymbol{\lambda}) = \boldsymbol{0} \tag{3.55}$$

$$\Leftrightarrow \boldsymbol{B}^\top \boldsymbol{B}\boldsymbol{\lambda} = \boldsymbol{B}^\top \boldsymbol{x} 。 \tag{3.56}$$

正常方程

最后一个表达式被称为*正常方程*。由于$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$是$U$的基础，因此是线性独立的，所以$\boldsymbol{B}^\top \boldsymbol{B} \in \mathbb{R}^{m\times m}$是规整的，可以被倒置。这使我们能够解决系数/的问题。

坐标

$$\boldsymbol{\lambda} = (\boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{B}^\top \boldsymbol{x} 。 \tag{3.57}$$

伪逆

矩阵$(\boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{B}^\top$也被称为$\boldsymbol{B}$的*伪逆*，对于非平方矩阵$\boldsymbol{B}$可以计算。它只要求$\boldsymbol{B}^\top \boldsymbol{B}$是正定的，如果$\boldsymbol{B}$是全秩的，就会出现这种情况。在实际应用中

在一些应用中（如线性回归），我们通常会添加一个 "抖动项 "$\boldsymbol{I}$到

$\top BB$来保证增加数值稳定性和正确定性。这个 "山脊 "可以用贝叶斯推断法严格地推导出来。详见第九章。

2. 找到投影$\pi_U(\boldsymbol{x})$ U。我们已经确定$\pi_U(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\lambda}$。因此，随着(3.57)

$$\pi_U(\boldsymbol{x}) = \boldsymbol{B}(\top BB)^{-1}\top \boldsymbol{B}\boldsymbol{x}。 \tag{3.58}$$

3. 找到投影矩阵$\boldsymbol{P}_\pi$。从(3.58)，我们可以立即看到，解决$\boldsymbol{P}_\pi \boldsymbol{x} = \pi(U\boldsymbol{x})$的投影矩阵必须是

$$\boldsymbol{P} = \pi \boldsymbol{B}(\top BB)^{-1}\boldsymbol{B}\top。 \tag{3.59}$$

*备注。*投射到一般子空间的解决方案包括作为特例的一维情况。如果$\dim(U)=1$，那么$\top BB$ R是一个标量，我们可以将投影矩阵改写为(3.59)$\boldsymbol{P} = \pi \boldsymbol{B}(\top BB)^{-1}\boldsymbol{B}\top$为
$\boldsymbol{P} = \pi \frac{\boldsymbol{B}\boldsymbol{B}^\top}{\top BB}$ 这正是( )中的投影矩阵。3.46). ◆

例子（3.11投射到二维子空间上）。

对于一个子空间$U = \text{span}[\begin{matrix}1\\1\\0\end{matrix}, \begin{matrix}1\\0\\1\end{matrix}] \subseteq \mathrm{R}$和$\boldsymbol{x} = \begin{matrix}6\\0\\0\end{matrix} \in \mathrm{R}^3$ 请找出就子空间U而言，$\boldsymbol{x}$的坐标$\boldsymbol{\lambda}$,2 投射点$\pi_U(\boldsymbol{x})$和投射矩阵$\boldsymbol{P}_\pi$。

首先，我们看到U的生成集是一个基础（线性不连续）。dence），并将U的基向量写成一个矩阵$\boldsymbol{B} = \begin{matrix}1&1\\1&0\\0&1\end{matrix}$。

其次，我们计算矩阵$\top BB$和矢量$\top Bx$为

$$\boldsymbol{B}\top\boldsymbol{B} = \begin{matrix}1&1&1\\0&1&2\end{matrix}\begin{matrix}1&0\\1&1\\0&2\end{matrix} = \begin{matrix}3&3\\3&5\end{matrix}, \quad \boldsymbol{B}\top\boldsymbol{x} = \begin{matrix}1&1&1\\0&1&2\end{matrix}\begin{matrix}6\\0\\0\end{matrix} = \begin{matrix}6\\0\end{matrix}. \tag{3.60}$$

第三，我们求解法线方程$\boldsymbol{B}\boldsymbol{B}\top\boldsymbol{\lambda} = \boldsymbol{B}\boldsymbol{x}\top$，以找到$\boldsymbol{\lambda}$。

$$\begin{matrix}3&3\\3&5\end{matrix}\begin{matrix}\lambda_1\\\lambda_2\end{matrix} = \begin{matrix}6\\0\end{matrix} \iff \boldsymbol{\lambda} = \begin{matrix}5\\-3\end{matrix}. \tag{3.61}$$

第四，$\boldsymbol{x}$在U上的投影$\pi_U(\boldsymbol{x})$，即进入U的列空间$\boldsymbol{B}$，可以通过以下方式直接计算出

$$\pi_U(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\lambda} = \begin{matrix}2\\-1\end{matrix} \tag{3.62}$$

projection error

投影误差也被称为*重建误差*。

相应的*投影误差*是差异向量的常数

原有矢量与它在U上的投影之间的关系，即：

$$\|\boldsymbol{x} - \pi_U(\boldsymbol{x})\| = \begin{bmatrix} 1 & 1 & -2 & 1 \end{bmatrix}^\top \begin{bmatrix} 1 \\ 1 \\ -2 \\ 1 \end{bmatrix} = \sqrt{7}. \tag{3.63}$$

第五，投影矩阵（对于任何$\boldsymbol{x} \in \mathbb{R}^3$）由以下公式给出

$$\boldsymbol{P} = \boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B})^{-1}\boldsymbol{B}^\top = \frac{1}{6}\begin{bmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{bmatrix}. \tag{3.64}$$

为了验证结果，我们可以（a）检查位移矢量是否$\pi_U(\boldsymbol{x}) - \boldsymbol{x}$与$U$的所有基向量正交，并且 (b) 验证$\boldsymbol{P} = \boldsymbol{P}\pi$（见定义3.10）.

备注。投影$\pi_U(\boldsymbol{x})$仍然是$\mathbb{R}^n$中的向量，尽管它们位于一个m维的子空间$U$$\mathbb{R}^n$中。然而，为了表示一个投影向量，我们只需要$m$个坐标$\lambda_1, \ldots, \lambda_m$相对于基向量$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$的$\boldsymbol{b}$。 ◆

备注。在具有一般内积的向量空间中，我们在计算角度和距离时必须注意，这些角度和距离是通过内积来定义的。 ◆

我们可以用投影法找到无法解决的线性方程组的近似解。

投影使我们能够研究有一个没有解的线性系统$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$的情况。回顾一下，这意味着$\boldsymbol{b}$不在$\boldsymbol{A}$的跨度内，也就是说，矢量$\boldsymbol{b}$不在$\boldsymbol{A}$的列所跨越的子空间内。考虑到线性方程不能被精确解决，我们可以找到一个*近似的解决方案*。其思路是在$\boldsymbol{A}$列所跨越的子空间中找到与$\boldsymbol{b}$最接近的向量，即我们计算$\boldsymbol{b}$在$\boldsymbol{A}$列所跨越的子空间上的*正交投影*。

最小二乘法解是

一个超定常的系统。这将在第9.4节进一步讨论。使用重建误差(3.63)是导出主成分分析的一种可能的方法（第10.3节）。

备注。我们刚刚研究了向量$\boldsymbol{x}$在子空间U上的投影，基向量为$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k$.如果这个基是$\{\}$个ONB，即(3.33)和(3.34)得到满足，投影方程(3.58)大大简化为

$$\pi_U(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{x} \tag{3.65}$$

因为$\boldsymbol{B}^\top\boldsymbol{B} = \boldsymbol{I}$，坐标为

$$\boldsymbol{\lambda} = \boldsymbol{B}\boldsymbol{x}^\top 。 \tag{3.66}$$

这意味着我们不再需要从(3.58)，这就节省了计算时间。 ◆

3.8.3 格拉姆-施密特正交化

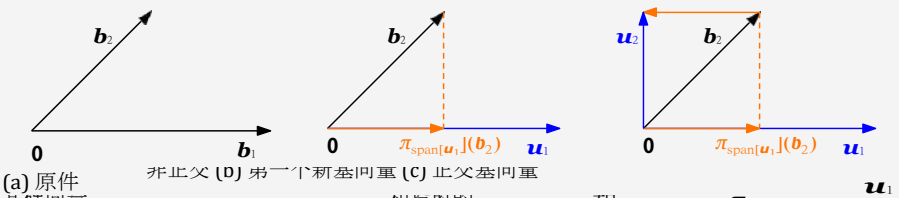
投影是Gram-Schmidt方法的核心，它允许我们将n维向量空间V的任何基（$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$）建设性地转换为$V$的正交/广义基（$\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$）。这个基总是存在的（Liesen和Mehrmann,2015），并且$\operatorname{span}[\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n] = \operatorname{span}[\boldsymbol{u}_1, \ldots, \boldsymbol{u}]_n$。*Gram-Schmidt正交方法*迭代了Gram-Schmidt构建一个正交基$(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$，从任何一个基$(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)$的 正统化 V如下：

$$\boldsymbol{u}_1 := \boldsymbol{b}_1 \tag{3.67}$$

$$\boldsymbol{u}_k := \boldsymbol{b}_k - \pi_{\operatorname{span}[\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{k-1}]}(\boldsymbol{b}_k)。 \quad k = 2, \ldots, n. \tag{3.68}$$

在(3.68)中，第k个基向量$\boldsymbol{b}_k$被投射到由前k 1个构建的正交向量$\boldsymbol{u}_1, \ldots, \boldsymbol{u}$ *所跨越的*子空间上。，$\boldsymbol{u}_{k-1}$；见第3.8.2节。然后从$\boldsymbol{b}$中减去k这个投影，得到一个与$\boldsymbol{u}_1$，.....所跨越的（$k-1$）维子空间正交的向量$\boldsymbol{u}_k$。对所有n个基向量$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$*得到*$V$的正交基（$\boldsymbol{u}_1$，…，$\boldsymbol{u}_n$）。如果我们对$\boldsymbol{u}$进行归一化处理$k$，就会得到一个ONB，其中$\|\boldsymbol{u}_k\| = 1$对于$k = 1, \ldots, n$.

例子（Gram3.12-Schmidt正交化）。



(a) 原件 非正交 (b) 第一个新基向量 (c) 正交基向量
基础向量 $\boldsymbol{b}_1, \boldsymbol{b}_2$. $\boldsymbol{u}_1 = \boldsymbol{b}_1$ 和投射的 $\boldsymbol{u}_2 = \boldsymbol{b}_2 - \pi_{\operatorname{span}[\boldsymbol{u}_1]}(\boldsymbol{b}_2)$.
覆盖的子空间上的 $\boldsymbol{u}_1$.

Figure 3.12 Gram-Schmidt正交化。(a)非正交的$\mathbb{R}^2$基础（，）。(b)首次建造基础向量 $\boldsymbol{u}_1$ 和正交的预测 $\boldsymbol{b}_2$ 到跨度[$\boldsymbol{u}_1$];(c)正交基础 $(\boldsymbol{u}_1, \boldsymbol{u}_2)$ 的$\mathbb{R}^2$。

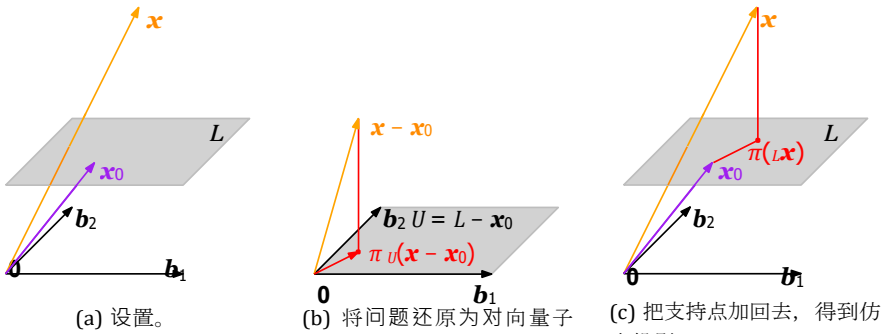
考虑$\mathbb{R}^2$一个基（$\boldsymbol{b}_1$，$\boldsymbol{b}_2$），其中

$$\boldsymbol{b}_1 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \quad \boldsymbol{b}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{3.69}$$

参见图3.12(a)。使用Gram-Schmidt方法，我们构建$\mathbb{R}^2$正交基（$\boldsymbol{u}_1$，$\boldsymbol{u}_2$）如下（假设点积为内积）。

$$\boldsymbol{u}_1 := \boldsymbol{b}_1 = \begin{bmatrix} 2 \\ 0 \end{bmatrix} \tag{3.70}$$

$$\boldsymbol{u}_2 := \boldsymbol{b}_2 - \pi_{\operatorname{span}[\boldsymbol{u}_1]}(\boldsymbol{b}_2) \overset{(3.45)}{=} \boldsymbol{b}_2 - \frac{\boldsymbol{u}_1\boldsymbol{u}_1^\top}{\|\boldsymbol{u}_1\|^2}\boldsymbol{b}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{3.71}$$



(a) 设置。 (b) 将问题还原为对向量子空间的剖析π_U。 (c) 把支持点加回去，得到仿生投影π_L。

图3.13 投影到一个仿生空间。(a) 原始设置；(b)设置移位，$-\boldsymbol{x}_0$可以是$\boldsymbol{x}-\boldsymbol{x}_0$投射到方向空间$U$上；(c) 投射是指译回为$\boldsymbol{x}_0+\pi_U(\boldsymbol{x}-\boldsymbol{x}_0)$，从而得到最终的正交投影$\pi_L(\boldsymbol{x})$。

图3.12(b)和(c)说明了这些步骤。我们立即看到，$\boldsymbol{u}_1$和$\boldsymbol{u}_2$是正交的，即$\boldsymbol{u}_1^\top \boldsymbol{u}_2=0$。

3.8.4 投射到仿生子空间

到目前为止，我们讨论了如何将一个向量投射到一个低维子空间U上。下面，我们提供一个将一个向量投射到一个仿生子空间的解决方案。

考虑一下图3.13(a)中的设置。为了确定$\boldsymbol{x}_0$在L上的正交投影$\pi_L(\boldsymbol{x})$，我们将问题转化为我们知道如何解决的问题：投影到一个矢量子空间。为了达到这个目的，我们从$\boldsymbol{x}$和L中减去支持点$\boldsymbol{x}_0$。因此，$L-\boldsymbol{x}_0=U$正是矢量子空间U。我们现在可以使用正交投射到我们在第3.13节讨论过的子空间上，并得到投影π（x x）。3.8.2并得到投影$\pi_U(\boldsymbol{x}-\boldsymbol{x}_0)$，如图3.13（b）所示。这个投影现在可以通过添加$\boldsymbol{x}_0$而被翻译回L，这样我们就得到了对仿生空间L的正交投影为

$$\pi_L(\boldsymbol{x})=\boldsymbol{x}_0+\pi_U(\boldsymbol{x}-\boldsymbol{x}_0)\,, \tag{3.72}$$

其中π_U（）是对子空间U的正交投影，即L的方向空间；见图3.13（c）。

从图3.13中还可以看出，$\boldsymbol{x}$与仿生空间L的距离与$\boldsymbol{x}-\boldsymbol{x}_0$与$U$的距离相同，即。

$$d(\boldsymbol{x},L)=\|\boldsymbol{x}-\pi_L(\boldsymbol{x})\|=\|\boldsymbol{x}-(\boldsymbol{x}_0+\pi_U(\boldsymbol{x}-\boldsymbol{x}_0))\| \tag{3.73a}$$

$$=d(\boldsymbol{x}-\boldsymbol{x}_0,\pi_U(\boldsymbol{x}-\boldsymbol{x}_0))=d(\boldsymbol{x}-\boldsymbol{x}_0,U)\,。 \tag{3.73b}$$

我们将在第12.1节中使用对仿射子空间的投影来推导分离超平面的概念。

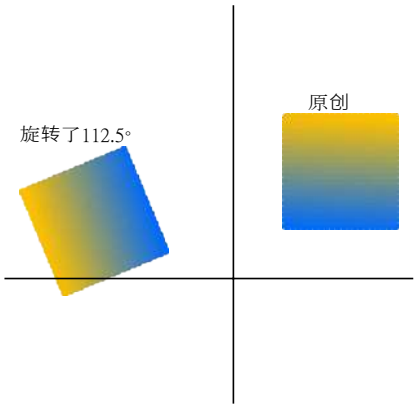


图 3.14旋转使物体在一个平面内围绕原点旋转。如果旋转的角度是正的，我们就逆时针旋转。

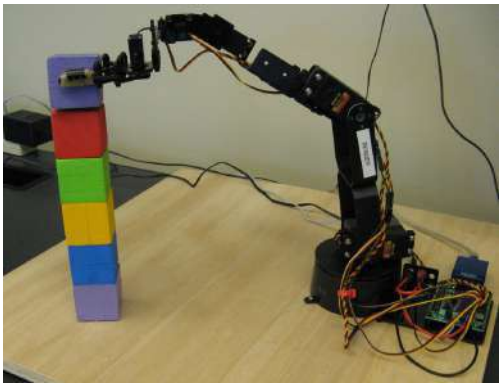


图 3.15机器人手臂需要旋转其关节，以便拿起物体或正确放置物体。图取自（Deisenroth等人，2015）。

3.9 轮换

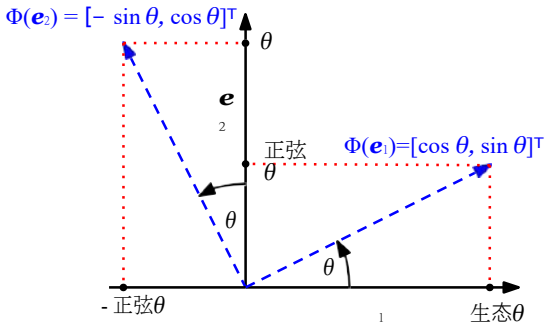
如同在第二节中所讨论的那样，保留长度和角度是正交变换矩阵的两个特征。3.4,是具有正交变换矩阵的线性映射的两个特征。在下文中，我们将仔细研究描述旋转的特定正交变换矩阵。

*旋转*是一种线性映射（更确切地说，是旋转的自动变形一个欧几里得矢量空间），将一个平面围绕原点旋转一个角度θ，也就是说，原点是一个固定点。对于一个正的角度$\theta>0$，按照惯例，我们以逆时针方向进行旋转。图中显示了一个例子，3.14,其中的变换矩阵是

$$\boldsymbol{R} = \begin{bmatrix} 0.38 & -0.92 \\ 0.92 & -0.38 \end{bmatrix}. \tag{3.74}$$

旋转的重要应用领域包括计算机图形学和机器人学。例如，在机器人学

中，知道如何旋转机械臂的关节以拿起或放置一个物体往往很重要，见图3.15.



图：R²中的标准基础旋转3.16一个角度θ。

3.9.1 R²中的旋转

考虑R²的标准基础$\boldsymbol{e}_1 = \begin{bmatrix}1\\0\end{bmatrix}$，$\boldsymbol{e}_2 = \begin{bmatrix}0\\1\end{bmatrix}$，它定义了我们的目标是将这个坐标系旋转一个角度$\theta$，如图所示。3.16.请注意，旋转后的向量仍然是线性独立的，因此是R²的一个基。这意味着旋转进行了一个基的改变。

旋转矩阵

旋转Φ是线性映射，因此我们可以用*旋转矩阵*$\boldsymbol{R}(\theta)$来表示。三角学(见图3.16)使我们能够确定旋转轴的坐标(Φ的图像)相对于R²标准基。

$$\Phi(\boldsymbol{e}_1) = \begin{bmatrix}\cos\theta\\ \text{正弦}\theta\end{bmatrix} \quad \Phi(\boldsymbol{e}_2) = \begin{bmatrix}-\sin\theta\\ \theta\end{bmatrix} \tag{3.75}$$

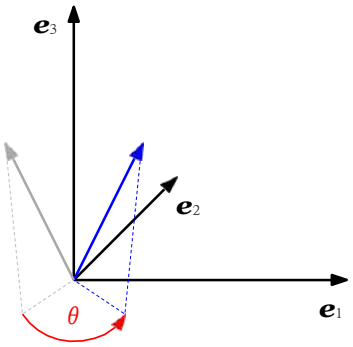
因此，执行基数变化到旋转坐标$\boldsymbol{R}(\theta)$的旋转矩阵给定为

$$\boldsymbol{R}(\theta) = \Phi(\boldsymbol{e}_1)\ \Phi(\boldsymbol{e}_2)\ ] = \begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix} \tag{3.76}$$

3.9.2 R³中的旋转

与R²的情况不同，在R³中我们可以围绕一个一维轴旋转任何二维平面。指定一般旋转矩阵的最简单方法是指定标准基$\boldsymbol{e}_1$、$\boldsymbol{e}_2$、$\boldsymbol{e}_3$的图像应该如何旋转，并确保这些图像$\boldsymbol{Re}_1$、$\boldsymbol{Re}_2$、$\boldsymbol{Re}_3$相互之间是正交的。然后，我们可以通过组合标准基的图像得到一个一般的旋转矩阵$\boldsymbol{R}$。

为了得到一个有意义的旋转角度，我们必须定义 "逆时针 "的含义，当我们在两个维度以上操作时。我们使用的惯例是，围绕一个轴的 "逆时针"（平面）旋转是指当我们 "正面看，从末端看向原点 "时围绕一个轴的旋转。因此，在R³中，有三种围绕三个标准基矢量的（平面）旋转（见图3.17):



图：R中^{3}的一个矢量（灰色）围绕一个角度θ旋转**3.17**。e_3-轴。旋转后的矢量显示为蓝色。

- 关于$_1$e轴的旋转

$$\boldsymbol{R}(_1\theta) = \begin{bmatrix} \Phi(\boldsymbol{e}_1) & \Phi(\boldsymbol{e}_2) & \Phi(\boldsymbol{e}_3) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}. \tag{3.77}$$

这里，$\boldsymbol{e}_1$坐标是固定的，逆时针旋转是在$\boldsymbol{e}\boldsymbol{e}_{23}$平面内进行的。

- 关于$_2$e轴的旋转

$$\boldsymbol{R}(_2\theta) = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix}. \tag{3.78}$$

如果我们将$\boldsymbol{e}\boldsymbol{e}_{13}$平面围绕$\boldsymbol{e}_2$轴旋转，我们需要看一下$\boldsymbol{e}_2$从其 "尖端 "向原点的轴。

- 关于$_3$e轴的旋转

$$\boldsymbol{R}(_3\theta) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3.79}$$

图3.17说明了这一点。

3.9.3 在n个尺寸中的旋转

从二维和三维旋转到n维欧氏向量空间的一般化可以直观地描述为固定n个二维空间，并将旋转限制在n个二维空间的一个二维平面。如同在三维的情况下，我们可以旋转任何平面（R的^{n}二维子空间）。

定义3.11（吉文斯旋转）。设V是一个n维的欧几里得向量空间，$\Phi: V \to V$是一个具有变换ma-的自动形态。

三角形

$$\boldsymbol{R}(ij\theta) := \begin{bmatrix} \boldsymbol{I}_{i-1} & 0 & \cdots & \cdots & 0 \\ 0 & \cos\theta & 0 & -\sin\theta & 0 \\ \vdots & 0 & \boldsymbol{I}_{j-i-1} & 0 & \vdots \\ 0 & \sin\theta & 0 & \cos\theta & 0 \\ 0 & \cdots & \cdots & 0 & \boldsymbol{I}_{n-j} \end{bmatrix} \in \mathrm{R}^{n\times n}, \quad (3.80)$$

吉文斯轮换

对于 $i < j$ 1$\leqslant$ n 和 θ $\in$ R。那么$\boldsymbol{R}(ij\theta)$就被称为*吉文斯旋转*。本质上，$\boldsymbol{R}(ij\theta)$是身份矩阵$\boldsymbol{I}_n$，具有

$$r_{ii} = \cos\theta, \quad r_{ij} = -\sin\theta, \quad r_{ji} = \sin\theta, \quad r_{jj} = \cos\theta。 \quad (3.81)$$

在二维空间（即n=2），我们得到(3.76)作为一个特例。

3.9.4 旋转的属性

旋转表现出一些有用的特性，可以通过将它们视为正交矩阵来推导出这些特性（定义为3.8):

- 旋转保留距离，即$\boldsymbol{x}\ \boldsymbol{y} = \boldsymbol{R}(\theta\boldsymbol{x})$ ℝ$_\theta(\boldsymbol{y})$ 。换句话说]1旋转使任何两点之间的距离在变换后保持不变。
- 旋转保留角度，即$_\theta\boldsymbol{Rx}$和$_\theta\boldsymbol{Ry}$之间的角度等于$\boldsymbol{x}$和$\boldsymbol{y}$之间的角度。
- 三个（或更多）维度的旋转通常是不相通的。因此，旋转的顺序很重要，即使它们围绕同一点旋转。只有在二维的矢量旋转是互换的，如$\boldsymbol{R}(\varphi)\boldsymbol{R}(\theta)=\boldsymbol{R}(\theta)\boldsymbol{R}(\varphi)$，对所有$\varphi, \theta$ [0,2 π)。只有当它们围绕同一点（如原点）旋转时，它们才构成一个阿贝尔群（有乘法）。

3.10 进一步阅读

在这一章中，我们简要介绍了解析几何的一些重要概念，我们将在本书后面的章节中使用这些概念。对于我们提出的一些概念的更广泛和更深入的概述，我们可以参考以下优秀书籍：Axler(2015)和Boyd和Vandenberghe(2018)。

内积使我们能够确定矢量（子）空间的特定基数，其中每个矢量与所有其他矢量正交（正交基数），使用Gram-Schmidt方法。这些基数在解决线性方程组的优化和数字算法中非常重要。例如，Krylov子空间方法，如共轭梯度或广义最小残差法（GMRES），最小化相互正交的残差误差（Stoer and Burlirsch,2002）。

在机器学习中，内部产品在以下方面非常重要

核方法（Schölkopf and Smola， 2002）。核方法利用了这样一个事实：许多线性算法可以纯粹通过内积计算来表达。然后，"内核技巧 "允许我们在一个（可能是无穷大的）特征空间中隐式地计算这些内积，甚至不需要明确知道这个特征空间。这使得许多用于机器学习的算法得以 "非线性化"，如用于降维的核-PCA（Schölkopf等人， 1997）。 高斯过程（Rasmussen和Williams,2006）也属于核方法的范畴，是目前概率再回归（将曲线拟合到数据点）的最先进技术。第12章将进一步探讨核的概念。

投影经常被用于计算机图形学，例如，生成阴影。在优化中，正交投影经常被用来（迭代地）最小化残余误差。这在机器学习中也有应用，例如，在线性回归中，我们想找到一个（线性）函数，使残余误差最小，即数据在线性函数上的正交投影的长度（Bishop,2006）。我们将在第九章中进一步研究这个问题。PCA（Pearson,1901;Hotelling,1933）也使用投影来降低高维数据的维度。我们将在第十章中详细讨论这个问题。

练习

3.1 证明 $(-,-)$对所有$\boldsymbol{x} = [x_1, x]_2^\top \in \mathbb{R}^2$和$\boldsymbol{y} = [y_1, y]_2^\top \in \mathbb{R}^{的2}$定义为

$$(\boldsymbol{x}, \boldsymbol{y}) := xy_{11} - (xy_{12} + xy_{21}) + 2(xy_{22})$$

是一个内积。

3.2 考虑 ^{2} R，对于R中的^{2}所有$\boldsymbol{x}$和$\boldsymbol{y}$，$(-,-)$定义为

$$(\boldsymbol{x}, \boldsymbol{y}) := \text{xT} \begin{bmatrix} 2 & 0 \\ 1 & 2 \end{bmatrix} \boldsymbol{y} \text{。}$$

是内积吗？

3.3 算出

$$\boldsymbol{x} = , \quad 2\boldsymbol{y} = 1 \; 30$$

使用

a. $(\boldsymbol{x}, \boldsymbol{y}) :=$ $\boldsymbol{xTy}$

b. $(\boldsymbol{x}, \boldsymbol{y}) := \boldsymbol{xTAy}$ ， $\boldsymbol{A} := \begin{bmatrix} 21 & 0 \\ 13 & -1 \\ 0- & 12 \end{bmatrix}$

3.4 请计算出

$$\boldsymbol{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} , \quad \boldsymbol{y} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$

使用

a. $(\boldsymbol{x}, \boldsymbol{y}) :=$ $\boldsymbol{xTy}$

b. $(\boldsymbol{x}, \boldsymbol{y}) := \boldsymbol{xTBy}$ ， $\boldsymbol{B} := \begin{bmatrix} 21 \\ 1 \; 3 \end{bmatrix}$

3.5 考虑欧几里得向量空间 $\mathbb{R}^5$与点积。一个子空间 $U \subseteq \mathbb{R}^5$和$\boldsymbol{x} \in \mathbb{R}^5$由以下公式给出

$$U = \text{span}[\,, 2\,, 1, 1]5\,, \qquad \boldsymbol{x} = 1\_\,.$$

a.确定$\boldsymbol{x}$在U上的正交投影$\pi_U(\boldsymbol{x})$

b.确定距离$\text{d}(\boldsymbol{x}, U)$

3.6 考虑 ^{3} R的内积

$$(\boldsymbol{x}, \boldsymbol{y}) := \boldsymbol{x}^\top \begin{bmatrix} 21 & 0 \\ 12 & 1 \\ 0- & 12 \end{bmatrix} \boldsymbol{y}\,.$$

此外，我们定义$\boldsymbol{e}_1$，$\boldsymbol{e}_2$，$\boldsymbol{e}_3$为R中的^{3}标准/经典基础。

a.确定$\boldsymbol{e}$的正交投影π_U（$\boldsymbol{e}_2$）$_2$到

$$U = \text{span}[\boldsymbol{e}_1, \boldsymbol{e}_3].$$

提示：正交性是通过内积定义的。 b. 计算距离d(e$_2$, U)。

c.画出情景：标准基向量和π_U（$\boldsymbol{e}_2$）。

3.7 设 V是一个矢量空间，π是V的内形态。

a.证明π是一个投影，当且仅当$id_V - \pi$是一个投影，其中id_V是V上的同一内形态。

b.现在假设π是一个投影。计算Im($id_V \pi$)和ker($id_V \pi$)。 作为Im(π)和ker(π)的函数。

3.8 用Gram-Schmidt方法，将二维子空间U⊆R的基础 $B = (\boldsymbol{b}_1, \boldsymbol{b}_2)$ 的二维子空间$U \subseteq \mathbb{R}^3$变成U的ONB $C = (\boldsymbol{c}_1, \boldsymbol{c}_2)$，其中

$$\boldsymbol{b}_1 := \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{b}_2 := \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$$

3.9 让 $n \in \mathbb{N}$，让$x_1, \ldots, x_n > 0$为n个正实数，以便$x_1 + \ldots + x_n = 1$. 使用考奇-施瓦茨不等式并表明

a. $n x^2 . \quad \frac{1}{i}$

b. $\sum_{i=1}^{n} \frac{1}{x_i} \geq n^2$

提示：想一想R上的n点积，然后选择特定的向量。

$\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$并应用考奇-施瓦茨不等式。

3.10 旋转向量

$$\boldsymbol{x}_1 := \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \quad \boldsymbol{x}_2 := \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

由 30°.

4

矩阵分解

在第二章和第三章中，我们研究了操作和测量向量、向量的投影和线性映 射 的 方法。向量的映射和转换可以方便地描述为由矩阵进行的操作。此外，数据也经常以矩阵形式表示，例如，矩阵的行代表不同的人，列描述人的不同特征，如体重、身高和社会经济地位。在这一章中，我们介绍了矩阵的三个方面：如何对矩阵进行总结，如何对矩阵进行分解，以及如何将这些分解用于矩阵的逼近。

我们首先考虑那些允许我们只用几个数字来描述矩阵的方法，这些数字描述了矩阵的整体属性。我们将在行列式（第4.1节）和特征值（第4.2节）这两节中，针对方形矩阵这一重要的特殊情况进行研究。这些特征数具有重要的数学意义，使我们能够迅速掌握一个矩阵的有用特性。从这里开始，我们将进入矩阵分解方法。矩阵分解的一个类比是数字的因式分解，如21素数的因式分解3。7由于这个原因，矩阵分解也经常被称为*矩阵因式分解*。矩阵分解是通过使用可解释矩阵的因子的不同表示方法来描述一个矩阵。

矩阵因式分解

我们将首先介绍一个类似于对称正定矩阵的平方根运算，即Cholesky分解（Section4.3).从这里开始，我们将研究两种相关的方法，将矩阵分解为规范的形式。第一种方法被称为矩阵对角化（Section4.4)，如果我们选择一个合适的基础，它允许我们用一个对角线转换矩阵来表示线性映射。第二种方法，奇异值分解（Section4.5)，将这种分解法扩展到非正方形矩阵，它被认为是线性代数的基本概念之一。这些分解很有帮助，因为代表数字数据的矩阵往往非常大，难以分析。在本章的最后，我们以矩阵秩的形式，系统地介绍了矩阵的类型和区别它们的特征属性（第4.7).

我们在本章中所涉及的方法将在以下方面变得非常重要

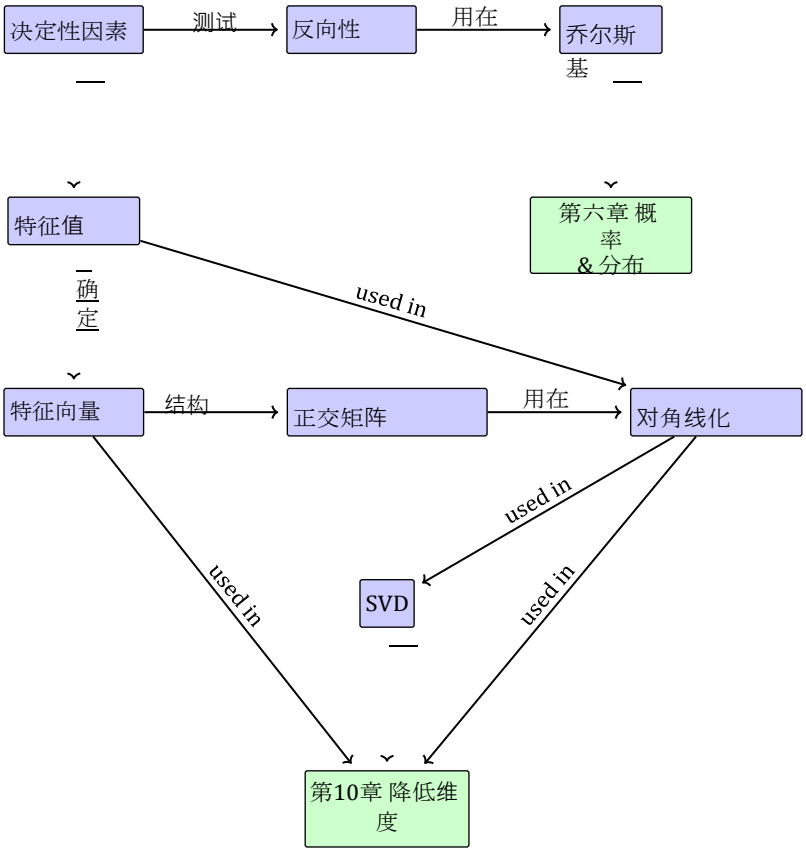


图 本章介绍的概念的**4.1**思维导图，以及这些概念在本书其他部分的使用情况。

在随后的数学章节中，如第6章，也在应用章节中，如第10章的降维或第11章的密度估计。本章的整体结构如图所示。4.1.

4.1 行列式和追踪

行列式符号 $|\boldsymbol{A}$ 不能与绝对值混淆。

决定数是线性代数中的重要概念。决定数是分析和解决线性方程组的一个数学对象。决定数只对方形矩阵$\boldsymbol{A}$ R定义$^{n\times n}$，即行数和列数相同的矩阵。在这本书中。

我们把行列式写成det($\boldsymbol{A}$)，有时也写成$|\boldsymbol{A}|$，以便

$$\det(\boldsymbol{A}) = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} \tag{4.1}$$

方形矩阵$\boldsymbol{A}\in\mathbb{R}^{n\times n}$的*行列式*是一个函数，它映射了Adeterminant

到一个实数上。在提供一般$n \times n$矩阵的行列式定义之前，让我们看看一些激励性的例子，并为一些特殊的矩阵定义行列式。

例子 (4.1测试矩阵的可逆性)

让我们从探索一个正方形矩阵$\boldsymbol{A}$是否可逆开始（见第2.2.2节）。对于最小的情况，即$\boldsymbol{A}$是1×1矩阵，即它是一个标度数，那么$\boldsymbol{A} = a = \Rightarrow \boldsymbol{A}^{-1} = \frac{1}{a}$。因此$a$成立，当且仅当$a \neq 0$。

对于2×2矩阵，根据逆的定义（定义2.3），我们那么，根据（2.24），$\boldsymbol{A}^{-1}$的倒数是

$$\boldsymbol{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}. \tag{4.2}$$

因此，当且仅当$\boldsymbol{A}$是可倒置的。

$$a_{11}a_{22} - a_{12}a_{21} \neq 0 \tag{4.3}$$

这个量是$\boldsymbol{A} \in \mathbb{R}^{2\times 2}$的行列式，即：。

$$\det(\boldsymbol{A}) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}. \tag{4.4}$$

例子4.1已经指出了行列式和逆矩阵的存在之间的关系。下一个定理说明了$n \times n$矩阵的相同结果。

定理 *对于***4.1.***任何方形矩阵*$\boldsymbol{A} \in \mathbb{R}^{n\times n}$*，当且仅当*$\det(\boldsymbol{A}) \neq 0$*时，*$\boldsymbol{A}$*是可逆的*。

我们对小矩阵的行列式有明确的（闭合形式）表达，以矩阵的元素为单位。对于$n = 1$，

$$\det(\boldsymbol{A}) = \det(a_{11}) = a_{11}. \tag{4.5}$$

对于$n = 2$，

$$\det(\boldsymbol{A}) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}, \tag{4.6}$$

这一点我们在前面的例子中已经观察到了。

对于$n = 3$（被称为Sarrus规则）。

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{31}a_{22}a_{13} - a_{32}a_{23}a_{11} - a_{33}a_{21}a_{12}. \tag{4.7}$$

为了帮助记忆Sarrus规则中的乘积项，可以尝试追踪矩阵中的三乘积元素
。

如果T_{ij}=forupper()-triangular，我们称正方形矩阵T为*上三角矩阵*。
$i>j$，也就是说，矩阵在其对角线以下为零。类似地，我们定义一个
*下三角矩阵*是一个对角线上方有零的矩阵。对于一个三下三角的
角度矩阵TR$^{n\times n}$，行列式是对角线元素的乘积，即：。

基体

基体

$$\det(\boldsymbol{T})=\prod_{i=1}^{n}T_{ii}。 \tag{4.8}$$

例子（4.2决定性因素作为数量的衡量标准）

行列式的概念是很自然的，当我们把它视为从一组跨越R中的对象^{n}的n*个向量的*映射时，事实证明，de-terminant det($\boldsymbol{A}$)是由矩阵$\boldsymbol{A}$的列组成的n维平行四边形的有符号体积。

对于n =2，矩阵的列形成一个平行四边形；见图 4.2.随着向量之间的角度变小，平行四边形的面积也会缩小。考虑到两个向量$\boldsymbol{b}, \boldsymbol{g}$构成矩阵$\boldsymbol{A}$ = [$\boldsymbol{b}, \boldsymbol{g}$]的列。那么，$\boldsymbol{A}$的行列式的绝对值就是平行四边形的面积**0**，其顶点为$\boldsymbol{b}, \boldsymbol{g}, \boldsymbol{b}+\boldsymbol{g}$。特别是，如果$\boldsymbol{b}, \boldsymbol{g}$是线性相关的，因此，对于某些λ∈R，$\boldsymbol{b}$=λ$\boldsymbol{g}$，它们不再是形成一个二维的平行四边形。因此，相应的面积是0。相反，如果$\boldsymbol{b}$，$\boldsymbol{g}$是线性独立的，并且是...的倍数。的经典基础向量$\boldsymbol{e}$，那么它们可以写成$\boldsymbol{b}$ = $\begin{bmatrix}b\\0\end{bmatrix}$ 和 $\boldsymbol{g}$ = $\begin{bmatrix}0\\g\end{bmatrix}$，而决定数是 $\begin{vmatrix}b & 0\\0 & g\end{vmatrix}$ = $bg-0=bg$。

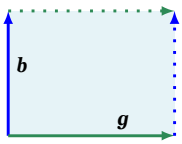
决定数的符号表示跨度向量$\boldsymbol{b}$，$\boldsymbol{g}$相对于标准基（$\boldsymbol{e}_1$，$\boldsymbol{e}_2$）的方向。在我们的图中，将顺序翻转到$\boldsymbol{g}$，$\boldsymbol{b}$交换$\boldsymbol{A}$的列并颠倒方向的阴暗区域。这就成为我们熟悉的公式：面积=高度×长度。这种直觉可以延伸到更高的维度。在R中^3，我们考虑三个向量$\boldsymbol{r}$、$\boldsymbol{b}$、$\boldsymbol{g}$∈R^{3}横跨一个平行四边形的边缘，即a实体，其面是平行的平行四边形（见图4.3).腹面的×3矩阵[3$\boldsymbol{r}, \boldsymbol{b}, \boldsymbol{g}$]的行列式的溶质值是指体积的固体。因此，行列式作为一个函数，衡量由矩阵中的列向量组成的签名体积。

考虑三个线性独立矢量$\boldsymbol{r}$、$\boldsymbol{g}$、$\boldsymbol{b}$∈R，给^{3}定为

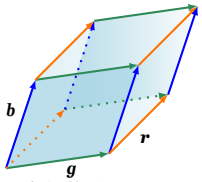
$$\boldsymbol{r}=\begin{bmatrix}0\\-8\end{bmatrix},\quad \boldsymbol{g}=\begin{bmatrix}1\\0\end{bmatrix},\quad \boldsymbol{b}=\begin{bmatrix}4\\-1\end{bmatrix}. \tag{4.9}$$

The determinant is the signed volume of the parallelepiped formed by the columns of the matrix.

图中的平行四边形（阴影区域）所跨越的**4.2**面积是多少？向量，并且是$\boldsymbol{g}$ |det([$\boldsymbol{b}$,$\boldsymbol{bg}$])|.



图：向量所跨越的平行四边形的**4.3**体积（阴影部分）。$\boldsymbol{r}, \boldsymbol{b},$ 是$\boldsymbol{g}$ det([$\boldsymbol{r}$, $\boldsymbol{b}$, $\boldsymbol{g}$])|.



决定数的符号表示
跨度向量的方向。

将这些向量写成矩阵的列

$$\boldsymbol{A} = [\boldsymbol{r}, \boldsymbol{g}, \boldsymbol{b}] = \begin{matrix} 2 & 6 & 1 \\ 0 & 1 & 4 \\ -8 & 0 & -1 \end{matrix} \tag{4.10}$$

让我们可以计算出所需的体积为

$$V = |\det(\boldsymbol{A})| = 186\,. \tag{4.11}$$

计算 $\times n$ 个矩阵的行列式需要一个通用的算法来解决 $n>$ 的3情况，我们将在下文中进行探讨。定理4.2下面的定理将计算 n 个矩阵的行列式的问题简化为计算 $(n\ 1)(n\ 1)$ 矩阵的行列式。通过递归应用拉普拉斯展开（定理4.2), $\times-\times$ 因此，我们可以通过最终计算×2矩阵2的行列式来计算 $n\times n$ 矩阵的行列式。

拉普拉斯扩展

定理（4.2拉普拉斯扩展）。*考虑一个矩阵* $\mathsf{AR}^{n\times n}$。*那么* $\in$ *对于所有* j $=1$ ， *... ..., n:*

被称为 $\det(\boldsymbol{A}_{k,j})$ *未成年人*和 $(-1)^{k+j}\det(\ )\boldsymbol{A}_{,j}$ 一个*辅助因子*。

1. 沿着柱子 j *扩张*

$$\det(\boldsymbol{A}) = \sum_{k=1}^{n} (-1)^{k+j} a_{kj}\ \text{扣除}\ _{k,j})\,. \tag{4.12}$$

2. 沿着行 j *扩张*

$$\det(\boldsymbol{A}) = \sum_{k=1}^{n} (\text{-}1)^{k+j} a_{jk}\ \det(\boldsymbol{A}_{j,k})。 \tag{4.13}$$

这里 $_{k,j}\ \in \mathsf{AR}^{(n-1)\times(n-1)}$ *是我们删除* k *行和* j *列后得到的* $\boldsymbol{A}$ *的子矩阵。*

例子（4.3拉普拉斯扩展）

让我们来计算一下以下的行列式 $1\quad\ \ 3$

$$\boldsymbol{A} = \begin{matrix} & 2 & \\ 3 & 1 & 2 \\ 0 & 0 & 1 \end{matrix} \tag{4.14}$$

使用沿第一行的拉普拉斯展开。应用(4.13)得出的结果是

$$\begin{matrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 0 & 0 & 1 \end{matrix} = (\text{-}1\quad^{1+1})\text{-}\ \begin{matrix} 1 & 2 \\ 0 & 1 \end{matrix} + (\text{-}1\quad^{1+2})\text{-}\ \begin{matrix} 3 & 2 \\ 0 & 1 \end{matrix} + (\text{-}1\quad^{1+3})\text{-}\ \begin{matrix} 3 & 1 \\ 0 & 0 \end{matrix}\ . \tag{4.15}$$

我们用(4.6)来计算所有×2号矩阵2的行列式，得到

$$\det(\boldsymbol{A}) = 1(1-0) - 2(3-0) + 3(0-0) = -5。 \tag{4.16}$$

为了完整起见，我们可以将这一结果与使用Sarrus规则计算决定性因素进行比较(4.7):

$$\det(\boldsymbol{A}) = -111+30-3022013102321+----------=1-6=-5。 \tag{4.17}$$

对于$\boldsymbol{A}\in\mathbb{R}^{n\times n}$来说，行列式表现出以下属性。

- 矩阵乘积的行列式是相应行列式的乘积，$\det(\boldsymbol{AB})=\det(\boldsymbol{A})\det(\boldsymbol{B})$。
- 决定数对转置是不变的，即$\det(\boldsymbol{A})=\det(\boldsymbol{A}^{\mathrm{T}})$。如果$\boldsymbol{A}$是正规的（
- 可倒置的），那么$\det(\boldsymbol{A}^{-1})=\frac{1}{\det(\boldsymbol{A})}$.
- 类似的矩阵（定义2.22）拥有相同的行列式。因此，对于一个线性映射$\Phi : V \to V$，Φ_Φ的所有变换矩阵$\boldsymbol{A}$都具有相同的行列式。因此，行列式对线性映射的基的选择是不变的。
- 将一列/行的倍数添加到另一列/行不会改变$\det(\boldsymbol{A})$.
- 列/行与$\lambda\in\mathbb{R}$的乘法将$\det(\boldsymbol{A})$缩放为λ。特别是，$\det(\lambda\boldsymbol{A})=\lambda^n\det(\boldsymbol{A})$。
- 交换两行/列会改变$\det(\boldsymbol{A})$的符号。

由于后三个属性，我们可以使用高斯消除法（见第2.1节）来计算$\det(\boldsymbol{A})$，将$\boldsymbol{A}$变成行-歇尔形式。当我们有一个三角形的$\boldsymbol{A}$，其中对角线以下的元素都是0.....时，我们可以停止高斯消除法。从(4.8)，三角矩阵的行列式是对角线元素的乘积。

定理 *一个*$\boldsymbol{A}\in\mathbb{R}^{n\times n}$*的*$\det(\boldsymbol{A})=0$*当且仅当*$\mathrm{rk}(\boldsymbol{A})=n$*。换句话说，当且仅当*$\boldsymbol{A}$*是全等级时，它是可倒置的。*

当数学主要由手工完成时，行列式计算被认为是分析矩阵反演的一个重要方法。然而，当代机器学习的方法直接使用数值方法，取代了行列式的明确计算。例如，在第二章中，我们了解到逆矩阵可以通过高斯消除法来计算。因此，高斯消除法可以用来计算一个矩阵的行列式。

决定数将在以下章节中发挥重要的理论作用，特别是当我们通过特征多项式学习特征值和特征向量时（第4.2节）。

定义 方形矩阵$\boldsymbol{A}\in\mathbb{R}^{n\times n}$的**4.4.***迹线*定义为：Astrace

$$\operatorname{tr}(\boldsymbol{A}) := \sum_{i=1}^{n} a_{ii}\,, \tag{4.18}$$

即，轨迹是$\boldsymbol{A}$的对角线元素之和。 轨迹满足以下属性

。

- $\operatorname{tr}(\boldsymbol{A}+\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{B})$ 为$\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$
- $\operatorname{tr}(\alpha\boldsymbol{A}) = \alpha\operatorname{tr}(\boldsymbol{A})$, $\alpha \in \mathbb{R}$ 为 $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ $\operatorname{tr}(\boldsymbol{I}_n) = n$
- 对于$\boldsymbol{A} \in \mathbb{R}^{n\times k}$， $\boldsymbol{B} \in \mathbb{R}^{k\times n}$， $\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA})$

可以证明，只有一个函数能同时满足这四个属性--跟踪（Gohberg等人，2012）。

轨迹在以下条件下是不变的

矩阵乘积的迹线的特性更为普遍。具体来说，轨迹在循环排列下是不变的，也就是说。

$$\operatorname{tr}(\boldsymbol{AKL}) = \operatorname{tr}(\boldsymbol{KLA}) \tag{4.19}$$

对于矩阵$\boldsymbol{A} \in \mathbb{R}^{a\times k}$， $\boldsymbol{K} \in \mathbb{R}^{k\times l}$，$\boldsymbol{L} \in \mathbb{R}^{l\times a}$，这个属性可以推广到任意数量的矩阵的乘积。作为(4.19)的一个特例，可以看出，对于两个向量$\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$\operatorname{tr}(\boldsymbol{xy}^{\mathrm{T}}) = \operatorname{tr}(\boldsymbol{yx}^{\mathrm{T}}) = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{x} \in \mathbb{R}\,。 \tag{4.20}$$

给定一个线性映射$\Phi : V \to V$，其中V是一个向量空间，我们通过使用Φ的矩阵表示的迹来定义这个映射的迹。对于V的一个给定基础，我们可以通过变换矩阵$\boldsymbol{A}$来描述Φ。对于V的不同基来说，Φ的相应变换矩阵$\boldsymbol{B}$可以通过对合适的$\boldsymbol{S} \in \mathrm{GL}(n)$进行基的改变而得到（见第2.7.2节）。对于$\Phi$的相应轨迹，这意味着

$$\operatorname{tr}(\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{S}^{-1}\boldsymbol{AS}) \overset{(4.19)}{=} \operatorname{tr}(\boldsymbol{ASS}^{-1}) = \operatorname{tr}(\boldsymbol{A})\,。 \tag{4.21}$$

特征多项式

为一个方形矩阵。结合我们对行列式和迹线的理解，我们现在可以定义一个重要的方程，用多项式来描述矩阵$\boldsymbol{A}$，我们将在下面的章节中广泛使用。

定义4.5（特征多项式）。对于$\lambda \in \mathbb{R}$和一个方形ma- trix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$

$$p_{\boldsymbol{A}}(\lambda) := \det(\boldsymbol{A} - \lambda \boldsymbol{I}) \tag{4.22a}$$

$$= c_0 + c_1\lambda + c_2\lambda^2 + - - + c_{n-1}\lambda^{n-1} + (-1)^n\lambda^n, \tag{4.22b}$$

$c_0, \ldots, c_{n-1} \in \mathbb{R}$，是$\boldsymbol{A}$的*特征多项式*，特别是。

因此，虽然线性映射的矩阵表示取决于基数，但线性映射Φ的轨迹与基数无关。

这中们了式线函特一，介行和作数性在节我绍列迹为的。表现

$$c_0 = \det(\boldsymbol{A})\,, \tag{4.23}$$

$$c_{n-1} = (-1)^{n-1}\operatorname{tr}(\boldsymbol{A})\,。 \tag{4.24}$$

特征多项式（4.22a）将使我们能够计算特征值和特征向量，在下一节中涉及。

4.2 特征值和特征向量

现在我们将了解一种新的方法来描述一个矩阵及其相关的线性映射的特征。回顾第2.7.1节，每个线性映射都有一个唯一的变换矩阵，给定一个有序的基础。我们可以通过以下方式来解释线性映射及其相关的变换矩阵

进行 "特征 "分析。正如我们将看到的，林

耳朵映射将告诉我们一组特殊的向量，即特征向量，是如何被线性映射所转换的。

*-埃兹*的特征值是一个德意思是 "特点"、"自己 "或 "自己的"。

定义4.6。 设$\boldsymbol{A}\in\mathbb{R}^{n\times n}$是一个方形矩阵。那么$\lambda\in\mathbb{R}$是一个$\boldsymbol{A}$的*特征值*，$\boldsymbol{x}\in\mathbb{R}^n\setminus\{\boldsymbol{0}\}$是$\boldsymbol{A}$的相应*特征向量*，如果

特征值是
特征向量

$$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}\,。 \tag{4.25}$$

我们把(4.25)为*特征值方程*。

特征值方程

*备注。*在线性代数文献和软件中，通常有这样一种习惯：特征值按降序排序，因此最大的特征值和相关的特征向量被称为第一特征值及其相关的特征向量，第二大特征值被称为第二特征值及其相关的特征向量，以此类推。然而，教科书和出版物可能有不同的或没有排序的概念。我们不希望

在本书中，如果没有明确

说明，则推定为一种排序。◆

以下声明是等价的。

- λ是$\boldsymbol{A}\in\mathbb{R}^{n\times n}$一个特征值。
- 存在一个$\boldsymbol{x}\in\mathbb{R}^n\setminus\{\boldsymbol{0}\}$，$\boldsymbol{A}\boldsymbol{x}=\lambda\boldsymbol{x}$，或者等价地，$(\boldsymbol{A}-\lambda\boldsymbol{I})_n\boldsymbol{x}=$可以$\boldsymbol{0}$非三段式求解，即$\boldsymbol{x}\neq\boldsymbol{0}$。
- $\operatorname{rk}(\boldsymbol{A}-\lambda\boldsymbol{I}_n) < n$.
- $\det(\boldsymbol{A}-\lambda\boldsymbol{I}_n)=0$.

定义（4.7共轭和共向）。两个指向的向量在

同方向的两个向量被称为*共向*。如果两个向量是同向的，那么它们指向相同或相反的方向。

就是*相邻的*。

备注（特征向量的非唯一性）。如果$\boldsymbol{x}$是一个与特征值λ相关的$\boldsymbol{A}$的特征向量，那么对于任何$c\in\mathbb{R}\setminus\{0\}$，都可以认为$c\boldsymbol{x}$是一个具有

相同特征值的$\boldsymbol{A}$的特征向量，因为 *矩阵分解*相邻的

$$\boldsymbol{A}(c\boldsymbol{x}) = c\boldsymbol{A}\boldsymbol{x} = c\lambda\boldsymbol{x} = \lambda(c\boldsymbol{x}) \text{。} \tag{4.26}$$

因此，所有与$\boldsymbol{x}$相邻的向量也是$\boldsymbol{A}$的特征向量。

◆

定理 4.8. R是AR的$n\times n$一个特征值，当且仅当λ是$\boldsymbol{A}$的特征多项式$p_{\boldsymbol{A}}(\lambda)$的一个根。

代数

定义 4.9. 让一个方形矩阵$\boldsymbol{A}$有一个特征值λ_i。

λ的*多重性（multiplicitymultiplicity）*$_i$是根在特征多项式中出现的次数。

定义（特征**4.10**空间和特征谱）。对于AR$^{n\times n}$，与特征值λ相关的$\boldsymbol{A}$的所有特征向量的集合跨越了一个子空间

特征空间 特征谱 谱系

$\boldsymbol{A}$的所有特征值的集合被称为$\boldsymbol{A}$的*特征谱*^{n}，或简称为*谱*$_\lambda$。

如果λ是$\boldsymbol{A}\ \boldsymbol{R}$的一个特征值$^{n\times n}$，那么相应的特征空间$\boldsymbol{E}_\lambda$就是同质线性方程组$(\boldsymbol{A}\ \lambda\boldsymbol{I})\boldsymbol{x}=$ 的解空间$\boldsymbol{0}$。 从几何学上讲，非零特征值对应的特征向量指向一个被线性映射拉伸的方向。特征值是它被拉伸的因子。如果特征值是负的，则拉伸的方向是翻转的。

例子(4.4身份矩阵的案例)
身份矩阵$\boldsymbol{I}\in$R$^{n\times n}$有特征多项式$p_{\boldsymbol{I}}(\lambda)=$
$\det(\boldsymbol{I}-\lambda\boldsymbol{I})=(1-\lambda)=$，它只有一个特征值$\lambda=$，即1oc-
诅咒n次。此外，$\boldsymbol{I}\boldsymbol{x}=\boldsymbol{\lambda}\boldsymbol{x}=1\ \boldsymbol{x}$对所有向量$\boldsymbol{x}\in$R$\setminus^n\{\boldsymbol{0}\}$都成立。
正因为如此，身份矩阵的唯一特征空间$\boldsymbol{E}_1$跨越了n个二元空间，而R的^{n}所有n个标准基向量都是$\boldsymbol{I}$的特征向量。

关于特征值和特征向量的有用属性包括以下内容。

- 一个矩阵$\boldsymbol{A}$和它的转置$\boldsymbol{A}^\top$拥有相同的特征值，但不一定是相同的特征向量。
- 特征空间$\boldsymbol{E}_\lambda$是$\boldsymbol{A}-\lambda\boldsymbol{I}$的空空间，因为

$$\boldsymbol{A}\boldsymbol{x}=\lambda\boldsymbol{x}\iff\boldsymbol{A}\boldsymbol{x}-\lambda\boldsymbol{x}=(\boldsymbol{0} \tag{4.27a}$$
$$\iff(\boldsymbol{A}-\lambda\boldsymbol{I})\boldsymbol{x}=\boldsymbol{0}\iff\boldsymbol{x}\in\ker(\boldsymbol{A}-\lambda\boldsymbol{I})\ 。 \tag{4.27b}$$

- 类似的矩阵（见定义2.22）拥有相同的特征值。因此，线性映射Φ的特征值与它的变换矩阵的基的选择无关。这使得特征值，连同行列式和迹线，成为线性映射的关键特征参数，因为它们在基的变化下都是不变的。
- 对称的正定矩阵总是有正的、真实的特征值。

例子4.5(计算特征值、特征向量和特征空间)

Let us find the eigenvalues and eigenvectors of the 2×2 matrix

$$\boldsymbol{A} = \begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix}. \tag{4.28}$$

步骤1：特征多项式。根据我们对$\boldsymbol{A}$的特征向量$\boldsymbol{x} \neq \boldsymbol{0}$和特征值$\lambda$的定义，将有一个向量使$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$，即$(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{x} = \boldsymbol{0}$。由于$\boldsymbol{x} \neq \boldsymbol{0}$，这要求$\boldsymbol{A} - \lambda\boldsymbol{I}$的核（空空间）包含更多的元素，而不仅仅是$\boldsymbol{0}$。这意味着$\boldsymbol{A} - \lambda\boldsymbol{I}$不是可逆的，因此$\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$。因此，我们需要计算特征多项式(4.22a)的根来找到特征值。

第2步：特征值。特征多项式为

$$p_{\boldsymbol{A}}(\lambda) = \det(\boldsymbol{A} - \lambda\boldsymbol{I}) \tag{4.29a}$$

$$= \det\left(\begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right) = \begin{vmatrix} 4-\lambda & 2 \\ 1 & 3-\lambda \end{vmatrix} \tag{4.29b}$$

$$= (4-\lambda)(3-\lambda) - 2 \cdot 1。 \tag{4.29c}$$

我们对特征多项式进行因式分解，得到

$$p(\lambda) = (4-\lambda)(3-\lambda) - 2 \cdot 1 = 10 - 7\lambda + \lambda^2 = (2-\lambda)(5-\lambda) \quad (4.30)。$$

给出根$\lambda_1 = 5$和$\lambda_2 = 2$。

第三步：特征向量和特征空间。我们通过寻找与这些特征值相对应的特征向量，即向量$\boldsymbol{x}$，使得

$$\begin{bmatrix} 4-\lambda & 2 \\ 1 & 3-\lambda \end{bmatrix} \boldsymbol{x} = \boldsymbol{0}. \tag{4.31}$$

对于$\lambda = 5$，我们得到

$$\begin{bmatrix} 4-5 & 2 \\ 1 & 3-5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 & 2 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \boldsymbol{0}。 \tag{4.32}$$

我们解决这个同质系统并得到一个解空间

$$E_5 = \operatorname{span}\left[\begin{bmatrix} 2 \\ 1 \end{bmatrix}\right]。 \tag{4.33}$$

这个特征空间是一维的，因为它拥有一个单一的基向量。

类似地，我们通过2求解同构方程组找到$\lambda = 2$的特征向量

$$\begin{bmatrix} 4-2 & 2 \\ 1 & 3-2 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix} \boldsymbol{x} = \boldsymbol{0}. \tag{4.34}$$

这意味着任何矢量$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$，其中$x_2 = -x_1$，如$\begin{bmatrix} 1 \\ -1 \end{bmatrix}$，是特征向量，其特征值2为。相应的特征空间是这样给出的一个

$$E_2 = \text{span}\left[\begin{bmatrix} 1 \\ -1 \end{bmatrix}\right]. \tag{4.35}$$

例子2中的两个特征空间E_5和E*是一维的，因为它们都被一个矢量所覆盖。*4.5中的两个特征空间E和E是一维的，因为它们都被一个矢量所跨越。然而，在其他情况下，我们可能有多个相同的特征值（见定义4.9），特征空间可能有一个以上的维度。

定义 让**4.11.**λ_i是一个正方形矩阵$\boldsymbol{A}$的特征值。那么

几何

λ的*几何倍数*$_i$是指线性独立特征的数量。

换句话说，它是与λ_i相关的特征向量所跨越的特征空间的维度$_i$。

*备注。*一个具体的特征值的几何倍数必须至少是1，因为每个特征值至少有一个相关的特征向量。一个特征值的几何倍数不能超过其代数倍数。它，但它可能更低。 ◆

例子 4.6

矩阵$\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$有两个重复的特征值$\lambda_1 = \lambda_2 =$和2一个的代数倍数2。然而，该特征值只有一个不同的单位特征向量$\boldsymbol{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$，因此，几何倍数1。

二维的图形直觉

Let us gain some intuition for determinants, eigenvectors, and eigenvalues using different linear mappings. Figure4.4depicts five transformation matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_5$ and their impact on a square grid of points, centered at the origin:

在几何学中，这种平行于轴线的剪切的保面积特性也被称为卡瓦列里的平行四边形的等面积原理

(Katz,2004)

。

- $\boldsymbol{A}_1 = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 2 \end{bmatrix}$.两个特征向量的方向对应于R中的^{2}典范基向量，即对两个心轴。纵轴被扩展了一个系数（2特征值λ_1=2），横轴被压缩了一个系数（^{1}特征值λ_2=）^{1}。该映射的面积为preserving ($\det(\boldsymbol{A}_1) = 1 = 2 \cdot \frac{1}{2}$).
- $\boldsymbol{A}_2 = \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & 1 \end{bmatrix}$对应于一个剪切映射，也就是说，它剪切了

沿着横轴的点，如果它们在正数上，则向右移动。

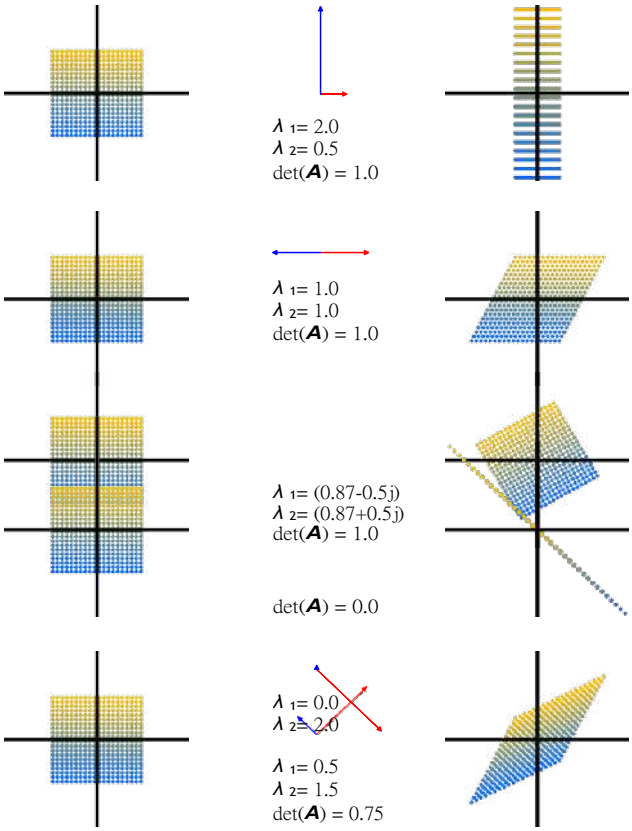


图 决定性**4.4**因素和特征空间。五种线性映射及其相关转换矩阵的概述 $\mathbb{R}^{2\times 2}$ 探测 400 彩色编码的点 $\boldsymbol{x} \in \mathbb{R}^2$（左列）到目标点（右列）。的。每一列描述了第一个特征向量，由其相关的特征值λ_1拉伸，第二个特征向量由其特征值λ_2拉伸。每一行描述了五个变换中的一个的效果 矩阵 与 $\boldsymbol{A}$ 关于标准基础 。

纵轴的一半，而向左则相反。这种映射是保存面积的（$\det(\boldsymbol{A}_2) = 1$）。特征值$\lambda_1 = 1 = \lambda_2$是重复的，特征向量是相通的（这里为了强调，画在两个相反的方向）。这表明，该映射只沿一个方向（地平线轴

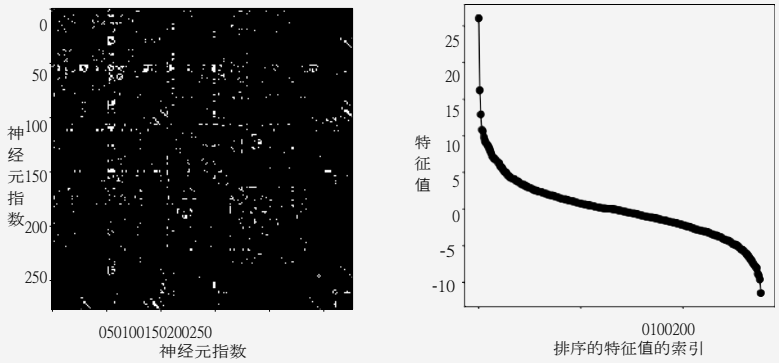
- $\boldsymbol{A}_3 = \begin{bmatrix} \cos(\frac{\pi}{6}) & -\sin(\frac{\pi}{6}) \\ \sin(\frac{\pi}{6}) & \cos(\frac{\pi}{6}) \end{bmatrix} = \frac{1}{2}\begin{bmatrix} \sqrt{3} & -1 \\ 1 & \sqrt{3} \end{bmatrix}$ 矩阵$\boldsymbol{A}_3$旋转了 points by $\frac{\pi}{6}\,\text{rad} = 30°$ counter-clockwise and has only complex eigenvalues, reflecting that the mapping is a rotation (hence, no eigenvectors are drawn). A rotation has to be volume preserving, and so the determinant is 1. For more details on rotations, we refer to Section3.9.
- $\boldsymbol{A}_4 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ 表示标准基础上的映射，该映射的颜色为 将一个二维领域转为一维。由于一个特征-

value is 0, the space in direction of the (blue) eigenvector corresponding to $\lambda_1 = 0$ collapses, while the orthogonal (red) eigenvector stretches space by a factor $\lambda_2 = 2$. Therefore, the area of the image is 0.

- $\boldsymbol{A} = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$ 是一个剪切和拉伸的映射，其空间比例为75%。

因为$\det(\boldsymbol{A}) = \frac{3}{4}$，它沿着$\lambda_2$的（红色）特征向量将空间拉长了一个系数1.5，并沿正交的（蓝色）特征向量压缩了一个系数0.5

示例（4.7生物神经网络的特征谱）。



(a) 连接矩阵。 (b) Eigenspectrum.

图 4.5 Caenorhabditis elegans的神经网络（Kaiser和Hilgetag，2006）。(a)Symmetrized connectivity matrix; (b)Eigenspectrum.

分析和学习网络数据的方法是机器学习方法的一个重要组成部分。理解网络的关键是网络节点之间的连接性，特别是两个节点之间是否有连接。在数据科学应用中，研究捕捉这种连接性数据的矩阵往往是有用的。

我们建立了*C.Elegans*蠕虫的完整神经网络的连接/毗邻矩阵$\boldsymbol{A} \in \mathbb{R}^{277\times277}$。每一行/每一列代表该虫大脑的277个神经元中的一个。如果神经元i通过突触与神经元j对话，连接矩阵$\boldsymbol{A}$的值为$a_{ij}=1$，否则$a_{ij}=0$。连接矩阵不是对称的，这意味着特征值可能不是实值的。因此，我们计算连通性矩阵的对称版本为$\boldsymbol{A}_{sym} := \boldsymbol{A} + \boldsymbol{A}^\top$。这个新的矩阵$\boldsymbol{A}_{sym}$如图4.5(a)所示，如果有一个非零的值$a_{ij}$只有当两个神经元相连时（白色像素），无论连接的方向如何。在图4.5(b)中，我们显示了$\boldsymbol{A}_{sym}$的相应特征谱。横轴显示了特征值的索引，按降序排序。纵轴显示相关的特征值。这个特征谱的S形是许多生物神经网络的典型特征。造成这种情况的基本机制是神经科学研究的一个活跃领域。

定理 *矩阵$A \in \mathbb{R}^{n\times n}$的*$n$*个不同特征值对应的特征向量$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$的矩阵$A \in \mathbb{R}^{n\times n}$的特征向量是线性独立的。*

该定理指出，一个具有n个不同特征值的矩阵的特征向量构成了$\mathbb{R}^n$的一个基础。

定义 一个方形矩阵$A \in \mathbb{R}^{n\times n}$如果拥有少于$n$个线性独立的特征向量，那么它就是*有缺陷的*。

一个无缺陷的矩阵$\boldsymbol{A} \in \mathbb{R}^{n\times n}$不一定需要$n$个不相干的特征值，但它确实需要特征向量构成$\mathbb{R}^n$的一个基。

*备注。*一个有缺陷的矩阵不能有n个不同的特征值，因为不同的特征值有线性独立的特征向量（定理4.12). ◆

定理 *给定一个矩阵$\boldsymbol{A} \in \mathbb{R}^{m\times n}$，我们总是可以通过定义得到一个对称的、正半自由的矩阵$\boldsymbol{S} \in \mathbb{R}^{n\times n}$。*

$$\boldsymbol{S} := \boldsymbol{A}^\top \boldsymbol{A}。 \tag{4.36}$$

*备注。*如果$\mathrm{rk}(\boldsymbol{A}) = n$，那么$\boldsymbol{S} := \boldsymbol{A}^\top \boldsymbol{A}$是对称的、正定的。

◆

了解为什么定理4.14成立的原因，对我们如何使用对称矩阵很有启发。对称性要求$\boldsymbol{S} = \boldsymbol{S}^\top$，通过插入 (4.36)，我们得到$\boldsymbol{S} = \boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{A}(\boldsymbol{A}^\top)^\top = (\boldsymbol{A}\boldsymbol{A})^\top = \boldsymbol{S}^\top$。4.36)，我们得到$\boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}\boldsymbol{A}^\top)(\boldsymbol{A}\boldsymbol{x}) = (\boldsymbol{A}\boldsymbol{x})^\top(\boldsymbol{A}\boldsymbol{x}) \geqslant 0$，因为点积计算的是一个平方之和（其本身就是非负的）。

谱系定理

该定理（光谱定理）。*如果$\boldsymbol{A} \in \mathbb{R}^{n\times n}$是对称的，那么相应的矢量空间$V$就有一个由$\boldsymbol{A}$的特征向量组成的正态基，而且每个特征值都是实数。*

谱系定理的一个直接含义是，对称矩阵$\boldsymbol{A}$的特征分解是存在的（有真实的特征值），而且我们可以找到一个特征向量的ONB，这样$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^\top$，其中$\boldsymbol{D}$是对角线，$\boldsymbol{P}$的列包含特征向量。

例子 4.8

考虑到矩阵

$$\boldsymbol{A} = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix} \tag{4.37}$$

$\boldsymbol{A}$的特征多项式是

$$p_{\boldsymbol{A}}(\lambda)=-(\lambda-1)^2(\lambda-7)\,, \tag{4.38}$$

这样我们就得到了特征值$\lambda_1=1$和$\lambda_2=7$，其中λ_1是一个重复的特征值。按照我们计算特征向量的标准程序，我们得到特征空间

$$E_1=\operatorname{span}\left[\begin{bmatrix}1\\1\\0\end{bmatrix}\right], \quad E_7=\operatorname{span}\left[\begin{bmatrix}1\\-1\\1\end{bmatrix}\right]。 \tag{4.39}$$

我们看到，$\boldsymbol{x}_3$对$\boldsymbol{x}_1$和$\boldsymbol{x}_2$都是正交的。然而，由于$\boldsymbol{x}_1^{\top}\boldsymbol{x}_2=1\neq 0$，它们并不是正交的。谱系定理（定理4.15）指出
指出，存在一个正交的基础，但我们拥有的这个基础不是正交的。然而，我们可以构建一个。

为了构建这样一个基础，我们利用这样一个事实，即$\boldsymbol{x}_1$，$\boldsymbol{x}_2$是与同一个特征值λ相关的特征因子。因此，对于任何α，$\beta\in\mathbb{R}$来说，可以认为

$$\boldsymbol{A}(\alpha_1\boldsymbol{x}+\beta\boldsymbol{x}_2)=\boldsymbol{A}\boldsymbol{x}_1\alpha+\boldsymbol{A}\boldsymbol{x}_2\beta=\lambda(\alpha\boldsymbol{x}_1+\beta_2\boldsymbol{x})\,, \tag{4.40}$$

Gram-Schmidt算法（第3.8.3节）是一种利用这种线性组合从一组基向量中反复构建正交/公法基的方法。因此，即使$\boldsymbol{x}_1$和$\boldsymbol{x}_2$不是正交的，我们也可以应用Gram-Schmidt算法，找到与$\lambda_1=1$相关的彼此正交（以及与$\boldsymbol{x}_3$正交）的特征向量。在我们的例子中，我们将得到

$$\boldsymbol{x}_1=\begin{bmatrix}1\\1\\0\end{bmatrix}, \quad \boldsymbol{x}_2=\begin{bmatrix}\frac{1}{2}\\-\frac{1}{2}\\2\end{bmatrix} \tag{4.41}$$

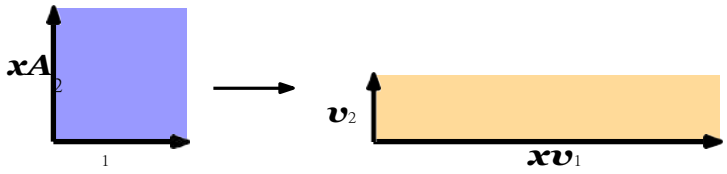
彼此正交，与$\boldsymbol{x}$正交$_3$，并且是x的特征向量。
$\boldsymbol{A}$与$\lambda_1=$相关联1。

在我们结束对特征值和特征向量的考虑之前，将这些矩阵特征与行列式和迹线的概念联系在一起是很有用的。

定理4.16。 *矩阵$\boldsymbol{A}\in\mathbb{R}^{n\times n}$的行列式是其特征值的乘积，即：。*

$$\det(\boldsymbol{A})=\prod_{i=1}^{n}\lambda_i, \tag{4.42}$$

其中$\lambda_i\in\mathbb{C}$是$\boldsymbol{A}$的（可能重复）特征值。



图：特征值的几何解释。的特征向量被相应的特征值拉长。单位方块的面积变化为λ_{12}，即周边变化系数为$\frac{1}{2}(|\lambda_1|+|\lambda_2|)$。

定理4.17。 *矩阵$\boldsymbol{A} \in \mathbb{R}^{n\times n}$的迹线是其特征值之和，即：*

$$tr(\boldsymbol{A}) = \sum_{i=1}^{n} \lambda_i \,, \tag{4.43}$$

其中$\lambda_i \in \mathbb{C}$是$\boldsymbol{A}$的（可能重复）特征值。

让我们为这两个定理提供一个几何学上的直觉。考虑一个拥有两个线性独立特征向量$\boldsymbol{x}_1$，$\boldsymbol{x}_2$的矩阵$\boldsymbol{A} \in \mathbb{R}^{2\times 2}$。在这个例子中，我们假设（$\boldsymbol{x}_1$，$\boldsymbol{x}_2$）是$\mathbb{R}^2$的一个ONB，所以它们是正交的，它们所跨越的正方形的面积是1；见图4.6.从第4.1节中，我们知道行列式可以计算单位面积在变换$\boldsymbol{A}$下的变化。使用$\boldsymbol{A}$对特征向量进行映射，可以得到向量$\boldsymbol{v}_1 = \boldsymbol{A}\boldsymbol{x}_1 = \lambda_1\boldsymbol{x}_1$和$\boldsymbol{v}_2 = \boldsymbol{A}\boldsymbol{x}_2 = \lambda_2\boldsymbol{x}_2$，也就是说，新的向量$\boldsymbol{v}_i$是特征向量$\boldsymbol{x}_i$的缩放版本，缩放因子是相应的特征值$\lambda_i$，$\boldsymbol{v}_1$、$\boldsymbol{v}_2$仍然是正交的，它们所跨越的矩形面积为$\lambda_{12}$。

鉴于$\boldsymbol{x}_1$、$\boldsymbol{x}_2$（在我们的例子中）是正交的，我们可以直接计算出单位正方形的周长为2（1+1）。用$\boldsymbol{A}$来映射特征向量会产生一个矩形，其周长为$2(\lambda_1 + \lambda_2)$。因此，特征值的绝对值之和告诉我们如何在变换矩阵的作用下，单位平方的周长发生变化$\boldsymbol{A}$.

例子（4.9谷歌的PageRank--作为特征向量的网页） 谷歌使用与矩阵$\boldsymbol{A}$的最大特征值相对应的特征向量来确定一个网页的搜索排名。拉里-佩奇和谢尔盖-布林在斯坦福大学开发的PageRank算法的想法是1996,，任何网页的重要性都可以通过链接到它的网页的重要性来接近。为此，他们把所有的网站写成一个巨大的有向图，显示哪个网页链接到哪个网页。PageRank$_i$通过计算指向*某$_i$网站*的网页数量来计算该网站*的*()权重（重要性）x_i。此外，PageRank还考虑了链接到*某$_i$*网站的重要性。

在不同的网站上。矩阵$\boldsymbol{A}$具有这样的特性：对于一个网站的任何初始等级/重要性向量$\boldsymbol{x}$，序列$\boldsymbol{x}$，$\boldsymbol{Ax}$，$^2\boldsymbol{Ax}$，...收敛到一个向量$\boldsymbol{x}^*$。这个向量被称为*PageRank*，并满足以下条件

网页排名

$\boldsymbol{Ax}^* = \boldsymbol{x}^*$，即是说，它是一个特征向量（有相应的特征值1）的$\boldsymbol{A}$。我们可以对$\boldsymbol{x}^*$进行归一化处理，使$\|\boldsymbol{x}^*\| = 1$后，我们可以解释$\boldsymbol{x}^*$中作为概率。关于PageRank的更多细节和不同观点可以在原始技术报告中找到（Page et al.,1999）。

4.3 Cholesky分解法

我们在机器学习中经常遇到的特殊类型的矩阵的因式分解有很多方法。在正实数中，我们有平方根运算，可以将数字分解为相同的部分，例如，$= 93$。3对于矩阵，我们需要注意的是，我们要在正数上计算类似于平方根的运算。对于对称的正定矩阵（见第3.2.3节），我们可以从一些平方根等价运算中选择。*Cholesky*

乔尔斯基分解

*/Cholesky因式分解*提供了一个平方根等效的运算。

Cholesky因式分解

在实践中，对对称、正定矩阵的计算是很有用的。

定理4.18（Cholesky分解）。*一个对称的正定矩阵$\boldsymbol{A}$可以被分解成一个乘积$\boldsymbol{A}=\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}$，其中$\boldsymbol{L}$是一个具有正对角线元素的下三角矩阵。*

$$\begin{matrix} a_{11} & \cdots & a_{1n} \\ a_{n1} & \cdots & a_{nn} \end{matrix} \quad \begin{matrix} l_{11} & \cdots & 0 \\ l_{n1} & \cdots & l_{nn} \end{matrix} \quad \begin{matrix} l_{11} & \cdots & l_{n1} \\ 0 & \cdots & l_{nn} \end{matrix}$$

Cholesky系数

$\boldsymbol{L}$被称为$\boldsymbol{A}$的Cholesky因子，$\boldsymbol{L}$是唯一的。

$$\begin{matrix} \ddots & \vdots \\ \cdots & \cdots \end{matrix} \quad (4.44)$$

例子（4.10Cholesky因式分解）

考虑到一个对称的正定矩阵$\boldsymbol{A}\in\mathbb{R}^{3\times3}$，我们是在在寻找其Cholesky因式分解$\boldsymbol{A}=\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}$，即。

$$\boldsymbol{A} = \begin{matrix} a_{11} & a_{21} & a_{31} \\ a_{21} & a_{22} & a_{32} \\ a_{31} & a_{32} & a_{33} \end{matrix} = \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}} \begin{matrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{matrix} \begin{matrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{matrix} . \quad (4.45)$$

乘以右手边，可以得到

$$\boldsymbol{A} = \begin{matrix} l_{11}^2 & l_{21}l_{11} & l_{31}l_{11} \\ l_{21}l_{11} & l_{21}^2 + l_{22}^2 & l_{31}l_{21} + l_{32}l_{22} \\ l_{31}l_{11} & l_{31}l_{21} + l_{32}l_{22} & l_{31}^2 + l_{32}^2 + l_{33}^2 \end{matrix} \quad (4.46)$$

比较( ) 的左手边和(4.45)的左边和(4.46)显示，在对角线元素l中$_{ii}$存在一个简单的模式。

$$l_{11} = \sqrt{a_{11}} \qquad l_{22} = \sqrt{a_{22} - l_{21}^2} \qquad l_{33} = \sqrt{a_{33} - (l_{32}^2 + l_{31}^2)} \tag{4.47}$$

同样，对于对角线以下的元素（l_{ij}，其中$i>j$），也有一个重复的模式。

$$l_{21} = \frac{1}{l_{11}} a_{21}, \qquad l_{31} = \frac{1}{l_{11}} a_{31}, \qquad l_{32} = \frac{1}{l_{22}} (a_{32} - l_{31} l_{21}). \tag{4.48}$$

因此，我们构建了任何对称的、正--的Cholesky分解。
它的定数×3矩阵3。关键的认识是，我们可以向后
在给定值的情况下，计算出$\boldsymbol{L}$的组件l_{ij}应该是什么？
a_{ij}为$\boldsymbol{A}$和以前计算的l的$_{ij}$值。

Cholesky分解是机器学习基础数值计算的一个重要工具。在这里，对称的正定矩阵需要经常操作，例如，多变量高斯变量的协方差矩阵（见第6.5节）是对称的，正定的。这个协方差矩阵的Cholesky因子化使我们能够从高斯分布中生成样本。它还允许我们对随机变量进行线性转换，这在计算深度随机模型的梯度时被大量利用，比如变异自动编码器（Jimenez Rezende等人，2014；Kingma和Welling，2014）。Cholesky分解也使我们能够非常有效地计算决定因素。鉴于Cholesky分解$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top$，我们知道$\det(\boldsymbol{A}) = \det(\boldsymbol{L})\det(\boldsymbol{L}^\top) = \det(\boldsymbol{L})^2$。由于$\boldsymbol{L}$是一个三角矩阵，行列式只是其对角线项的乘积，所以即$\det(\boldsymbol{A}) = \prod_i l_{ii}^2$。因此，许多数值软件包使用Cholesky分解，使计算更有效率。

4.4 重分解和对角线化

对角线*矩阵*是一个在所有非对角线上的数值为零的矩阵。的形式，即，它们是

$$\boldsymbol{D} = \begin{bmatrix} c_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & c_n \end{bmatrix} \tag{4.49}$$

它们允许快速计算行列式、幂和倒数。行列式是其对角线项的乘积，矩阵幂$\boldsymbol{D}^k$是由每个对角线元素提高到幂k给出的，而逆$\boldsymbol{D}^{-1}$是其对角线元素的倒数，如果它们都是非零的话。

在本节中，我们将讨论如何将矩阵转化为对角线的

形式。这是我们在第2.7.2节讨论的基数变化和第4.2节的特征值的一个重要应用。

回顾一下，如果存在一个可逆矩阵$\boldsymbol{P}$，那么两个矩阵$\boldsymbol{A}$、$\boldsymbol{D}$是相似的（定义2.22），这样$\boldsymbol{D}=\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}$。更具体地说，我们将研究与对角线矩阵$\boldsymbol{D}$相似的矩阵$\boldsymbol{A}$，这些矩阵在对角线上包含了$\boldsymbol{A}$的特征值。

可对角线的

定义4.19（可对角线化）。 如果一个矩阵$\boldsymbol{A}\in\mathbb{R}^{n\times n}$类似于一个对角线矩阵，即如果存在一个可倒置的矩阵$\boldsymbol{P}\in\mathbb{R}^{n\times n}$，使得$\boldsymbol{D}=\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}$，则该矩阵是*可对角线化的*。

在下文中，我们将看到，将矩阵$\boldsymbol{A}\in\mathbb{R}^{n\times n}$对角化是表达相同的线性映射的一种方式，但在另一个基础上（见第2.6.1节），这将是一个由$\boldsymbol{A}$的特征向量组成的基础。

设$\boldsymbol{A}$为$\boldsymbol{R}^{n\times n}$，设$\lambda_1,\ldots,\lambda_n$是一组标量，并让$\boldsymbol{p}_1,\ldots,\boldsymbol{p}_n$我们定义$\boldsymbol{P}:=[\boldsymbol{p}_1,\ldots,\boldsymbol{p}_n]$，让$\boldsymbol{D}\in\mathbb{R}^{n\times n}$是一个对角线矩阵，其对角线条目为$\lambda_1,\ldots,\lambda_n$。那么我们可以证明

$$\boldsymbol{A}\boldsymbol{P}=\boldsymbol{P}\boldsymbol{D} \tag{4.50}$$

当且仅当$\lambda_1,\ldots,\lambda_n$是$\boldsymbol{A}$的特征值，且$\boldsymbol{p}_1,\ldots,\boldsymbol{p}_n$是$\boldsymbol{A}$的特征向量。

我们可以看到，这句话之所以成立，是因为

$$\boldsymbol{A}\boldsymbol{P}=\boldsymbol{A}[\boldsymbol{p}_1,\ldots,\boldsymbol{p}_n]=[\boldsymbol{A}\boldsymbol{p}_1,\ldots,\boldsymbol{A}\boldsymbol{p}_n], \tag{4.51}$$

$$\boldsymbol{P}\boldsymbol{D}=[\boldsymbol{p}_1,\ldots,\boldsymbol{p}_n]\begin{bmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_n\end{bmatrix}=[\lambda_1\boldsymbol{p}_1,\ldots,\lambda_n\boldsymbol{p}_n]. \tag{4.52}$$

。

因此，(4.50) 意味着

$$\boldsymbol{A}\boldsymbol{p}_1=\lambda_1\boldsymbol{p}_1 \tag{4.53}$$

.

$$\boldsymbol{A}\boldsymbol{p}_n=\lambda_n\boldsymbol{p}_n。 \tag{4.54}$$

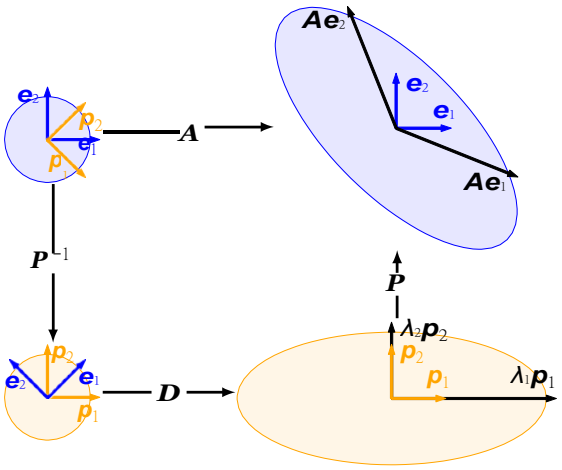
因此，$\boldsymbol{P}$的列必须是$\boldsymbol{A}$的特征向量。

我们对角化的定义要求$\boldsymbol{P}\in\mathbb{R}^{n\times n}$是可倒置的，即$\boldsymbol{P}$具有全等级（定理4.3).这就要求我们有$n$*个*线性独立的特征向量$\boldsymbol{p}_1,\ldots,\boldsymbol{p}_n$，即$\boldsymbol{p}_i$构成$\mathbb{R}^n$的一个基。

定理4.20（Eigendecomposition）。*一个方形矩阵*$\boldsymbol{A}\in\mathbb{R}^{n\times n}$*可以被分解为*

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1} \tag{4.55}$$

其中$\boldsymbol{P} \in \mathbb{R}^{n\times n}$*和*$\boldsymbol{D}$*是一个对角线矩阵，其对角线项是*$\boldsymbol{A}$*的特征值，当且仅当*$\boldsymbol{A}$*的特征向量构成*$\mathbb{R}^n$*的一个基。*



图：作为顺序转换的eigendecomposition背后的直觉4.7。左上角至左下角。P^{-1}执行一个基础变化（这里用$\mathcal{X}$表示[2]，被描述为一个类似旋转的操作），从标准的基数进入特征基数。从左下到右下：沿着重新映射的正交特征向量进行缩放，这里描述的是一个圆被拉伸成一个椭圆。右下角到右上角。撤销基础变化（被描述为反向旋转）并恢复原始坐标框架。

该定理4.20这意味着只有非缺陷矩阵才能被对角化，而且$\boldsymbol{P}$的列是$\boldsymbol{A}$的n个特征向量。对于对称矩阵，我们可以得到更强的特征值分解结果。

定理 *一个***4.21.***对称矩阵*$\boldsymbol{S}\in\mathbb{R}^{n\times n}$*总是可以被对角化。*

该定理4.21直接从光谱定理4.15得出。此外，光谱定理指出，我们可以找到$\mathbb{R}^n$的特征向量的ONB，这使得$\boldsymbol{P}$成为正交矩阵，因此$\boldsymbol{D}=\boldsymbol{P}^\top\boldsymbol{A}\boldsymbol{P}$。

*备注。*矩阵的Jordan法线形式提供了一种适用于缺陷矩阵的分解方法(Lang,1987)，但超出了本书的范围。 ◆

Eigendecomposition的几何直觉

我们可以将一个矩阵的重构解释如下（也见图4.7):假设$\boldsymbol{A}$是相对于标准基$\boldsymbol{e}$的$_i$线性映射的变换矩阵（蓝色箭头）。$\boldsymbol{P}^{-1}$执行一个从标准基到特征基的基数变化。然后，对角线$\boldsymbol{D}$通过特征值λ_i对沿这些轴的向量进行缩放。最后，$\boldsymbol{P}$将这些缩放后的向量转换回标准/经典协同学，得到λ_{ipi}。

示例（4.11Eigendecomposition）。

让我们计算一下$\boldsymbol{A}$=的重构。 $\frac{1}{2}\begin{bmatrix}5 & -2\\ -2 & 5\end{bmatrix}$.

步骤1：计算特征值和特征向量。特征

$\boldsymbol{A}$的多项式是

$$\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = \det \begin{bmatrix} \frac{5}{2} - \lambda & -1 \\ -1 & \frac{5}{2} - \lambda \end{bmatrix} \tag{4.56a}$$

$$= (\tfrac{5}{2} - \lambda)^2 - 1 = \lambda^2 - 5\lambda + \tfrac{21}{4} = (\lambda - \tfrac{7}{2})(\lambda - \tfrac{3}{2}). \tag{4.56b}$$

因此，$\boldsymbol{A}$的特征值是$\lambda_1 = \frac{7}{2}$和$\lambda_2 = \frac{3}{2}$（在A的根上）。特征多项式），相关的（归一化）特征向量通过以下方式获得

$$\boldsymbol{A}\boldsymbol{p}_1 = \frac{7}{2}\boldsymbol{p}_1, \qquad \boldsymbol{A}\boldsymbol{p}_2 = \frac{3}{2}\boldsymbol{p}_2. \tag{4.57}$$

这就产生了

$$\boldsymbol{p}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \boldsymbol{p}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{4.58}$$

第二步：检查是否存在。特征向量$\boldsymbol{p}_1$，$\boldsymbol{p}_2$构成$\mathbb{R}^2$的基础，因此，$\boldsymbol{A}$可以被对角化。

第3步：构建矩阵$\boldsymbol{P}$，对$\boldsymbol{A}$进行对角化。我们在$\boldsymbol{P}$中收集$\boldsymbol{A}$的特征向量，以便

$$\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2] = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}. \tag{4.59}$$

然后我们得到

$$\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P} = \begin{bmatrix} \frac{7}{2} & 0 \\ 0 & \frac{3}{2} \end{bmatrix} = \boldsymbol{D}. \tag{4.60}$$

Figure4.7visualizes the严格意义上的分解$\boldsymbol{A} = \begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix}$作为一连串的线性变换。

等价地，我们得到（利用$\boldsymbol{P}^{-1} = \boldsymbol{P}^{\top}$，因为特征向量本例中的$\boldsymbol{p}_1$和$\boldsymbol{p}_2$构成了一个ONB）

$$\underbrace{\begin{bmatrix} \frac{1}{2} & -\frac{5}{2} \\ -\frac{5}{2} & \frac{1}{2} \end{bmatrix}}_{\boldsymbol{A}} = \underbrace{\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}}_{\boldsymbol{P}} \underbrace{\begin{bmatrix} \frac{7}{2} & 0 \\ 0 & \frac{3}{2} \end{bmatrix}}_{\boldsymbol{D}} \underbrace{\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}}_{\boldsymbol{P}^{\top}}. \tag{4.61}$$

- 对角线矩阵$\boldsymbol{D}$可以有效地被提升到一个幂。因此，我们可以通过特征值分解（如果存在的话）为矩阵$\boldsymbol{A} \in \mathbb{R}^{n\times n}$找到一个矩阵幂，这样就可以

$$\boldsymbol{A}^k = (\boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1})^k = \boldsymbol{P}\boldsymbol{D}^k\boldsymbol{P}^{-1}。 \tag{4.62}$$

  计算$\boldsymbol{D}^k$是有效的，因为我们对任何对角线元素单独应用这个操作。

- 假设存在eigendecomposition $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$。那么。

$$\det(\boldsymbol{A}) = \det(\boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}) = \det(\boldsymbol{P})\det(\boldsymbol{D})\det(\boldsymbol{P}^{-1}) \tag{4.63a}$$

$$= \det(\boldsymbol{D}) = \prod_{i}^{r} d_{ii} \tag{4.63b}$$

可以有效地计算出$\boldsymbol{A}$的行列式。

特征值分解需要方形矩阵。对一般的矩阵进行分解会很有用。在下一节中，我们将介绍一种更通用的矩阵分解技术，即奇异值分解。

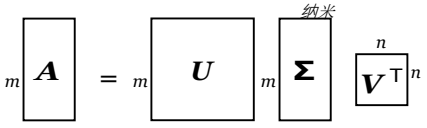
4.5 奇异值分解

矩阵的奇异值分解（SVD）是线性代数中的一种核心矩阵分解方法。它被称为 "线性代数的基本定理"(Strang,1993)，因为它可以应用于所有的矩阵，而不仅仅是方形矩阵，而且它始终存在。此外，正如我们将在下文中探讨的那样，矩阵$\boldsymbol{A}$的SVD代表线性映射$\Phi : V W$，量化了这两个向量空间的基础几何的变化。我们推荐Kalman(1996)以及Roy和Banerjee(2014)的工作，以深入了解SVD的数学原理。

该定理（4.22SVD定理）。*设*$\boldsymbol{A} \in \mathbb{R}^{m \times n}$*是一个矩形矩阵，其特征是*

SVD定理
SVD
奇异值分解

等级$r \in [0, \min(m, n)]$。$\boldsymbol{A}$*的SVD是一种分解形式*



(4.64)

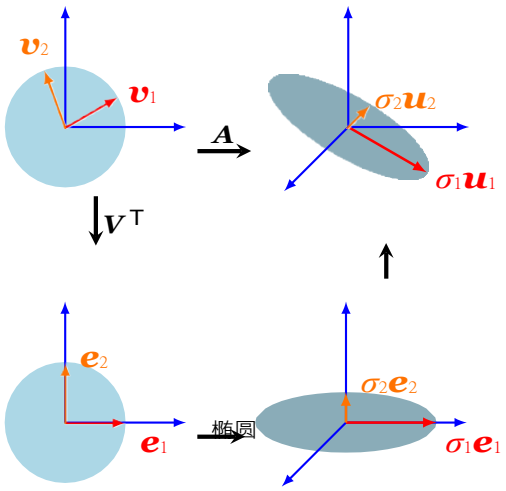
有一个正交矩阵$\boldsymbol{U} \in \mathbb{R}^{m \times m}$，*列向量为*$\boldsymbol{u}_i$，$i = 1, \ldots, m$，*和一个正交矩阵*$\boldsymbol{V} \in \mathbb{R}^{n \times n}$，*列向量为*$\boldsymbol{v}_j$，$j=1, \ldots$*此外，*$\boldsymbol{\Sigma}$*是一个*$m \times n$*矩阵，*$\Sigma_{ii} = \sigma_i \geqslant 0$，$\Sigma_{ij} = 0$，$i \neq j$。

$\boldsymbol{\Sigma}$的对角线项σ_i，$i=1, \ldots \boldsymbol{\Sigma}$的对角线项称为*奇异值*，$\boldsymbol{u}_i$被称为*左弦向量*，而$\boldsymbol{v}_j$被称为*右弦向量*。按照惯例，奇异值是有顺序的，即$\sigma_1 \geqslant \sigma_2 \geqslant \sigma_r \geqslant 0$.

奇异值
左弦向量。
右单数向量

奇异值矩阵$\boldsymbol{\Sigma}$是唯一的，但需要注意一下。观察一下，$\boldsymbol{\Sigma} \in \mathbb{R}^{m \times n}$是矩形的。特别是，$\boldsymbol{\Sigma}$的大小与$\boldsymbol{A}$相同，这意味着$\boldsymbol{\Sigma}$有一个对角线子矩阵，包含奇异值，需要额外的零填充。具体来说，如果$m>n$，那么矩阵$\boldsymbol{\Sigma}$的对角线结构一直到*第*n行，然后由以下部分组成

奇异值
基体



图：A的SVD背后的直觉**4.8** 基体 $\in \mathcal{R}^{3\times 2}$ 作为连续的转变。左上角至左下角。 T 执行一个基础 $\mathcal{R}$的^{2}变化。从左下角到右下角： $\mathbf{\Sigma}$ 比例尺和地图从$\mathcal{R}^2$到$\mathcal{R}^3$。在$\mathbf{\Sigma}$的右下角住着$\mathcal{R}^3$.第三维是正交于椭圆盘的表面。右下角至右上角。执行一个基本 $\mathcal{R}$内^{3}的变化。

$\mathbf{0}^\top$从n+到1m的行向量，以便

$$\mathbf{\Sigma} = \begin{matrix} \sigma_1 & & 00 \\ 的 & \ddots & \\ 0 & & 0 \\ & 00 & \sigma_n \\ 0 & \ldots & \\ & 0 & \\ 0 & \ldots & 0 \end{matrix}. \tag{4.65}$$

如果$m<n$，矩阵$\mathbf{\Sigma}$在m列之前有一个对角线结构，列由$\mathbf{0}m$+到1n组成。

$$\mathbf{\Sigma} = \begin{matrix} \sigma_1 & 0 & 0 & 0 & \ldots & 0 \\ 0 & \ddots & 0 & & & \vdots \\ 00 & \sigma_m & 0 & \ldots & 0 & \end{matrix} \tag{4.66}$$

备注。 SVD对任何矩阵$\boldsymbol{A}\in\mathbb{R}^{m\times n}$都存在。

4.5.1 SVD的几何直觉

SVD提供了描述变换矩阵的几何直觉 $\boldsymbol{A}$.在下文中，我们将把SVD讨论为对基数进行的连续线性变换。在例4.12中，我们将把SVD的变换矩阵应用于R中的^{2}一组向量，这使我们能够更清楚地看到每个变换的效果。

矩阵的SVD可以解释为将相关的线性映射（回顾第2.7.1节）Φ ：^{n} → RR分解^{m}为三个操作；见图4.8. SVD直觉表面上与我们的eigendecomposition直觉有相似的结构，见图4.7：广义上讲，SVD通过$\boldsymbol{V}$进行基础改变，然后$^\top$通过奇异值进行扩展和增加（或降低）维度。

值矩阵$\mathbf{\Sigma}$。最后，它通过$\boldsymbol{U}$进行第二次基础改变。 SVD需要一些重要的细节和注意事项，这就是为什么我们将

更详细地回顾一下我们的直觉。

回顾一下是很有用的基的变化（第2.7.2节），正交矩阵（定义3.8）和正交基（第3.5节）。

假设我们得到一个线性映射$\Phi : \mathbb{R}^n \to \mathbb{R}^m$的变换矩阵，分别与$\mathbb{R}^n$和$\mathbb{R}^m$标准基$\boldsymbol{B}$和$\boldsymbol{C}$有关。 此外， 假设有$\mathbb{R}^n$的第二个基$\tilde{\boldsymbol{B}}$和$\mathbb{R}^m$的$\tilde{\boldsymbol{C}}$。

1. 矩阵$\boldsymbol{V}$在$\mathbb{R}^n$域中从$\tilde{\boldsymbol{B}}$（由图中左上角的红色和橙色矢量$\boldsymbol{v}_1$和$\boldsymbol{v}_2$表示）进行了基数变化。
   $\boldsymbol{V} = \boldsymbol{V}^{-1}$执行从$\boldsymbol{B}$到$\tilde{\boldsymbol{B}}$的基础变化。 红色和橙色的矢量现在与图中左下角的典型基础4.8.
2. 在将坐标系改为$\tilde{\boldsymbol{B}}$后， $\mathbf{\Sigma}$将新的坐标系扩展到了$\tilde{\mathbb{R}}^m$。
   用奇异值σ_i来命名（并增加或减少维度），即。
   $\mathbf{\Sigma}$是Φ相对于$\tilde{\boldsymbol{B}}$和$\tilde{\boldsymbol{C}}$的变换矩阵， 代表
   通过红色和橙色的矢量被拉伸并位于$\boldsymbol{e}_1$-$\boldsymbol{e}_2$平面上可以看出，现在在图的右下角嵌入了一个三维空间。4.8.
3. $\boldsymbol{U}$在编码域$\mathbb{R}^m$中进行基数变化，从$\tilde{\boldsymbol{C}}$到卡诺尼--的变化。
   在$\mathbb{R}^m$的基础上，将红色和橙色矢量旋转出$\boldsymbol{e}_1$-$\boldsymbol{e}_2$平面。这在图的右上角显示。4.8.

SVD表达了域和码域中的基数变化。这与在同一向量空间内操作的eigendecomposition相反，在eigendecomposition中，同样的基数变化被应用，然后又被取消。SVD的特别之处在于，这两个不同的基数同时被奇异值矩阵$\mathbf{\Sigma}$所连接。

例子（4.12向量和SVD）

考虑到一个正方形网格的向量$\boldsymbol{X} \in \mathbb{R}^2$的映射，适合在一个盒子里的大小×2以原点为中心2。使用标准基础，我们将这些使用的向量

$$\boldsymbol{A} = \begin{bmatrix} 1 & -0.8 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} = \boldsymbol{U}\mathbf{\Sigma}\boldsymbol{V}^\top \tag{4.67a}$$

$$= \begin{bmatrix} -0.79 & -0.62 & 0 \\ 0.38 & -0.78 & -0.49 \\ -0.48 & -0.62 & 0.62 \end{bmatrix} \begin{bmatrix} 1.0 & 0 \\ 0 & 1.0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -0.78 & 0.62 \\ -0.62 & -0.78 \end{bmatrix} . \tag{4.67b}$$

我们从一组向量$\boldsymbol{X}$（彩色的点；见图左上角的面板）开始。
ure4.9）排列成一个网格。然后我们应用$\boldsymbol{V}^\top \in \mathbb{R}^{2\times 2}$，使$\boldsymbol{X}$旋转。
旋转后的向量显示在图的左下角，4.9.我们现在用奇异值矩阵$\mathbf{\Sigma}$将这些向量映射到代码域中
$\mathbb{R}^3$（见图中右下角面板4.9).请注意，所有矢量都位于

x_1-x_2平面。第三个坐标总是0.在x_1-x_2中的向量平面已经被奇异值拉长了。

$\boldsymbol{A}$对向量$\mathbf{X}$的直接映射到码域$\mathbf{R}^3$等同于用$\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$对$\mathbf{X}$进行变换，其中$\boldsymbol{U}$在$\mathbf{X}$的内部进行旋转。编码域$\mathbf{R}^3$，使映射的向量不再局限于$x_1$-$x_2$平面；它们仍然在一个平面上，如图右上角的面板所示4.9.

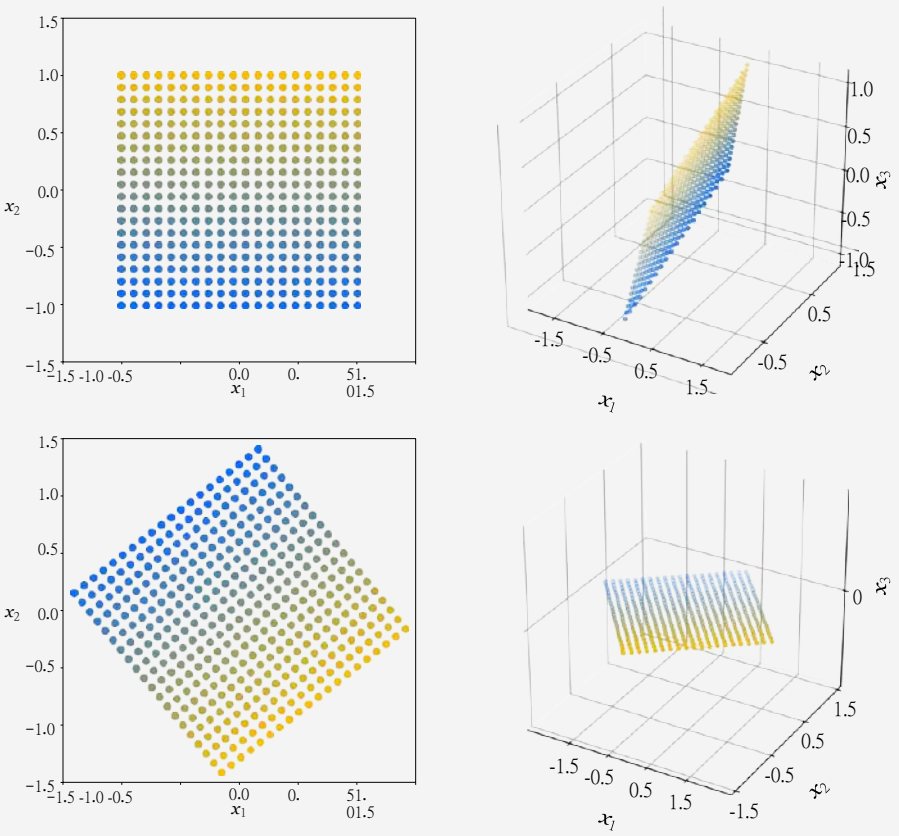


图4.9SVD和向量的映射（用圆盘表示）。各板块遵循相同的图中的逆时针结构4.8.

4.5.2 SVD的构建

接下来我们将讨论SVD存在的原因，并详细说明如何计算它。一般矩阵的SVD与方形矩阵的eigendecomposition有一些相似之处。

*备注。*比较一下SPD矩阵的特征分解

$$\boldsymbol{s} = \boldsymbol{s}^{\mathrm{T}} = \boldsymbol{p}\,\boldsymbol{d}\boldsymbol{p}^{\mathrm{T}} \tag{4.68}$$

与相应的SVD

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}。 \tag{4.69}$$

如果我们设定

$$\boldsymbol{u} = \boldsymbol{p} = \boldsymbol{v}, \quad \boldsymbol{d} = \boldsymbol{\sigma}, \tag{4.70}$$

我们看到，SPD矩阵的SVD是它们的特征分解。 ◆

在下文中，我们将探讨为什么该定理4.22成立的原因以及SVD是如何构建的。计算$\mathbb{R}^{m\times n}$的SVD相当于找到两组正态基$U = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m)$和$V = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$，分别是码域$\mathbb{R}^m$和域$\mathbb{R}^n$。从这些有序基中，我们将构建矩阵$\boldsymbol{U}$和$\boldsymbol{V}$。

我们的计划是，首先构建右奇异向量$\boldsymbol{v}_1, \ldots$.然后我们构建左奇异向量$\boldsymbol{u}_1, \ldots$.此后，我们将把两者联系起来，并要求$\boldsymbol{v}_i$的正交性在$\boldsymbol{A}$的反演下得到保留。这一点很重要，因为我们知道图像$\boldsymbol{A}\boldsymbol{v}_i$

形成一组正交的向量。然后我们将通过标量因子对这些图像进行归一化处理，这将变成奇异值。

让我们从构建右弦向量开始。谱系定理（定理4.15）告诉我们，对称矩阵的特征向量形成一个ONB，这也意味着它可以被对角化。此外，根据定理4.14我们总能从任何矩形矩阵$\boldsymbol{A} \in \mathbb{R}^{m\times n}$构造出一个对称的、正半无限的矩阵$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} \in \mathbb{R}^{n\times n}$，因此，我们总能将$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$对角化，得到

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{\mathrm{T}} = \boldsymbol{P}\begin{bmatrix}\lambda_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \lambda_n\end{bmatrix}\boldsymbol{P}^{\mathrm{T}}, \tag{4.71}$$

。

其中$\boldsymbol{P}$是一个正交矩阵，由正交特征基数组成。$\lambda_i \geqslant 0$是$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$的特征值。让我们假设$\boldsymbol{A}$的SVD存在，并将(4.64)到(4.71).这就得到了

$$\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}})^{\mathrm{T}}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}) = \boldsymbol{V}\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}, \tag{4.72}$$

其中$\boldsymbol{U}, \boldsymbol{V}$是正交矩阵。因此，在$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I}$*的情况*下，我们可以得到

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{V}\begin{bmatrix}\sigma_1^2 & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \sigma_n^2\end{bmatrix}\boldsymbol{V}^{\mathrm{T}}. \tag{4.73}$$

现在比较(4.71)和(4.73)，我们可以确定

$$\boldsymbol{V} = \boldsymbol{P}^{\mathrm{T}}, \tag{4.74}$$

$$\sigma_i^2 = \lambda_i。 \tag{4.75}$$

因此，组成$\boldsymbol{P}$的$\boldsymbol{A}\boldsymbol{A}^\top$的特征向量是$\boldsymbol{A}$的右弦向量$\boldsymbol{V}$（见(4.74)）.$\boldsymbol{A}\boldsymbol{A}^\top$的特征值是$\boldsymbol{\Sigma}$的平方奇异值（见(4.75)）.

为了获得左弦向量$\boldsymbol{U}$，我们遵循类似的程序。我们首先计算对称矩阵$\boldsymbol{A}\boldsymbol{A}^\top \in \mathbb{R}^{m\times m}$的SVD（而不是之前的$\boldsymbol{A}^\top\boldsymbol{A} \in \mathbb{R}^{n\times n}$）。$\boldsymbol{A}$的SVD得到的是

$$\boldsymbol{A}\boldsymbol{A}^\top = (\boldsymbol{U}\boldsymbol{\Sigma}^\top\boldsymbol{V})(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)^\top = \boldsymbol{U}\boldsymbol{\Sigma}^\top\boldsymbol{V}\boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top \tag{4.76a}$$

$$= \boldsymbol{U}\begin{bmatrix}\sigma_1^2 & & 0\\ & \ddots & \\ 0 & & \sigma_m^2\end{bmatrix}\boldsymbol{U}^\top. \tag{4.76b}$$

谱系定理告诉我们，$\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{S}\boldsymbol{D}\boldsymbol{S}^\top$可以被对角化，我们可以找到$\boldsymbol{A}\boldsymbol{A}^\top$的特征向量的ONB，它们被收集在

$\boldsymbol{S}$.$\boldsymbol{A}\boldsymbol{A}^\top$的正态特征向量是左心形向量$\boldsymbol{U}$

并在SVD的编码域中形成一个正态基。

这就留下了矩阵$\boldsymbol{\Sigma}$的结构问题。由于$\boldsymbol{A}\boldsymbol{A}^\top$和$\boldsymbol{A}^\top\boldsymbol{A}$有相同的非零特征值（见第106），两种情况下SVD中$\boldsymbol{\Sigma}$矩阵的非零项必须相同。

最后一步是把我们到目前为止所触及的所有部分联系起来。我们有一个$\boldsymbol{V}$中右弦向量的正态集。为了完成SVD的构造，我们将它们与正交向量$\boldsymbol{U}$连接起来。为了达到这个目的，我们利用$\boldsymbol{A}_i$下的$\boldsymbol{v}$的图像也必须是正交的。我们可以用第3.4节的结果来证明这一点。我们要求$\boldsymbol{A}\boldsymbol{v}_i$和$\boldsymbol{A}\boldsymbol{v}$之间的$_j$内积必须是为0

$i \neq j$. 对于任何两个正交的特征向量$\boldsymbol{v}_i$, $\boldsymbol{v}_j$, $i \neq j$，可以认为

$$(\boldsymbol{A}\boldsymbol{v}_i)^\top(\boldsymbol{A}\boldsymbol{v}_j) = \boldsymbol{v}_i^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{v}_j = \boldsymbol{v}_i^\top(\lambda_j\boldsymbol{v}_j) = \lambda_j\boldsymbol{v}_i^\top\boldsymbol{v}_j = 0. \tag{4.77}$$

对于$m \geqslant r$的情况，可以认为$\boldsymbol{A}\boldsymbol{v}_1, \ldots, \boldsymbol{A}\boldsymbol{v}_r$是$\mathbb{R}^m$一个$r$维子空间的基础。

为了完成SVD结构，我们需要左弦向量是*正交的*。我们将右弦向量$\boldsymbol{A}\boldsymbol{v}$的图像归一化$_i$，得到

$$\boldsymbol{u}_i := \frac{\boldsymbol{A}\boldsymbol{v}_i}{\|\boldsymbol{A}\boldsymbol{v}_i\|} = \frac{1}{\sqrt{\lambda_i}}\boldsymbol{A}\boldsymbol{v}_i = \frac{1}{\sigma_i}\boldsymbol{A}\boldsymbol{v}_i, \tag{4.78}$$

其中最后一个等式是由(4.75)和(4.76b)，表明$\boldsymbol{A}\boldsymbol{A}^\top$的特征值是$\sigma_i^2 = \lambda_i$的。

因此，$\boldsymbol{A}\boldsymbol{A}^\top$的特征向量，我们知道是右边的奇异向量$\boldsymbol{v}_i$，以及它们在$\boldsymbol{A}$下的归一化图像，即左奇异向量$\boldsymbol{u}_i$，形成两个自洽的ONB，通过奇异值矩阵$\boldsymbol{\Sigma}$连接。

奇异值方程

让我们重新排列 (4.78)，得到*奇异值方程*

$$\boldsymbol{A}\boldsymbol{v}_i = \sigma_i\boldsymbol{u}_i, \quad i = 1, \ldots, r. \tag{4.79}$$

这个方程非常类似于特征值方程(4.25)，但左手边和右手边的向量是不一样的。

对于$n<m$，(4.79)只对i:(n成立，但(4.79)对$i>n$的$\boldsymbol{u}$没有说明$_i$。然而，根据结构我们知道它们是正交的。反过来说，对于$m < n$，(4.79对于$i>m$，我们有$\boldsymbol{A}\boldsymbol{v}_i=$，并且$\mathbf{0}$我们仍然知道$\boldsymbol{v}_i$形成了一个正态集。这意味着SVD也为$\boldsymbol{A}$的核（空空间）提供了一个正态基础，即$\boldsymbol{A}\boldsymbol{x}=$的向量集合（$\mathbf{0}$见第2.7.3节）。

将$\boldsymbol{v}_i$作为$\boldsymbol{V}$的列，将$\boldsymbol{u}_i$作为V的列串联起来。

$\boldsymbol{U}$产量

$$\boldsymbol{a}\boldsymbol{v} = \boldsymbol{u}\boldsymbol{\sigma}\,, \qquad (4.80)\text{。}$$

其中$\boldsymbol{\Sigma}$的维度与$\boldsymbol{A}$相同，并且对行1、.....、r有对角线结构。因此，与$\boldsymbol{V}$相乘后$^{\mathrm{T}}$得到$\boldsymbol{A} = \boldsymbol{U}\,\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{V}$，这就是$\boldsymbol{A}$的SVD。

示例（4.13计算SVD）。

让我们来找一找以下的奇异值分解

$$\boldsymbol{A} = \begin{matrix} 1 & 0 & 1 \\ -2 & 1 & 0 \end{matrix}\,. \qquad (4.81)$$

SVD要求我们计算右弦向量$\boldsymbol{v}_j$，奇异值σ_k，以及左弦向量$\boldsymbol{u}_i$。

第1步：将右弦向量作为$^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}$的特征基点。

我们首先计算

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \begin{matrix} 1 & & \\ & -2 & \\ 0 & & 1 \\ & 1 & 0 \end{matrix} \begin{matrix} & & \\ 1 & 1 & \\ -2 & 1 & 0 \end{matrix} = \begin{matrix} 5 & -2 & 1 \\ - & & 210 \\ .1 & & 0 \\ & 1 & \end{matrix} \qquad (4.82)$$

我们$_j$通过$^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}$的特征值分解来计算奇异值和右弦向量$\boldsymbol{v}$，其公式为

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \begin{matrix} \frac{\sqrt{5}}{30} & 0 & \frac{-1}{6} \\ \frac{-2}{30} & \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{6}} \\ \frac{1}{30} & \frac{2}{\sqrt{5}} & \frac{1}{6} \end{matrix} \begin{matrix} 6 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{matrix} \begin{matrix} \frac{\sqrt{5}}{30} & \frac{-2}{30} & \frac{1}{30} \\ 0 & \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{-1}{6} & \frac{-2}{6} & \frac{1}{6} \end{matrix} = \boldsymbol{P}\,\boldsymbol{D}\boldsymbol{P}^{\mathrm{T}} \qquad (4.83)$$

。

我们得到右弦向量作为$\boldsymbol{P}$的列，所以

$$\boldsymbol{V} = \boldsymbol{P} = \begin{matrix} \frac{\sqrt{5}}{30} & 0 & \frac{-1}{6} \\ \frac{-2}{30} & \frac{1}{\sqrt{5}} & \frac{-2}{6} \\ \frac{1}{30} & \frac{2}{\sqrt{5}} & \frac{1}{6} \end{matrix}\,. \qquad (4.84)$$

第2步：单值矩阵。

由于奇异值 σ_i是

$\boldsymbol{A}^\top\boldsymbol{A}$我们直接从$\boldsymbol{D}$中得到它们，因为$\mathrm{rk}(\boldsymbol{A}) = 2$，只有两个非零奇异值：$\sigma_1 = \sqrt{6}$和$\sigma_2 = 1$。奇异值矩阵的大小必须与$\boldsymbol{A}$相同，我们可以得到

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sqrt{6} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{4.85}$$

第三步：左奇异向量作为右奇异向量的归一化图像。

我们通过计算$\boldsymbol{A}$下的右奇异向量的图像来找到左奇异向量，并通过将它们除以相应的奇异值进行归一化。我们得到

$$\boldsymbol{u}_1 = \frac{1}{\sigma_1}\boldsymbol{A}\boldsymbol{v}_1 = \frac{1}{\sqrt{6}}\begin{bmatrix} 1 & 1 & 0 \\ -2 & 1 & 0 \end{bmatrix}\begin{bmatrix} \frac{\sqrt{5}}{30} \\ \frac{-2}{\sqrt{30}} \\ \frac{1}{\sqrt{30}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{bmatrix}, \tag{4.86}$$

$$\boldsymbol{u}_2 = \frac{1}{\sigma_2}\boldsymbol{A}\boldsymbol{v}_2 = \frac{1}{1}\begin{bmatrix} 1 & 1 & 0 \\ -2 & 1 & 0 \end{bmatrix}\begin{bmatrix} 0 \\ \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{bmatrix} = \begin{bmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{bmatrix}, \tag{4.87}$$

$$\boldsymbol{U} = [\boldsymbol{u}_2, \boldsymbol{u}_1] = \frac{1}{\sqrt{5}}\begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix}. \tag{4.88}$$

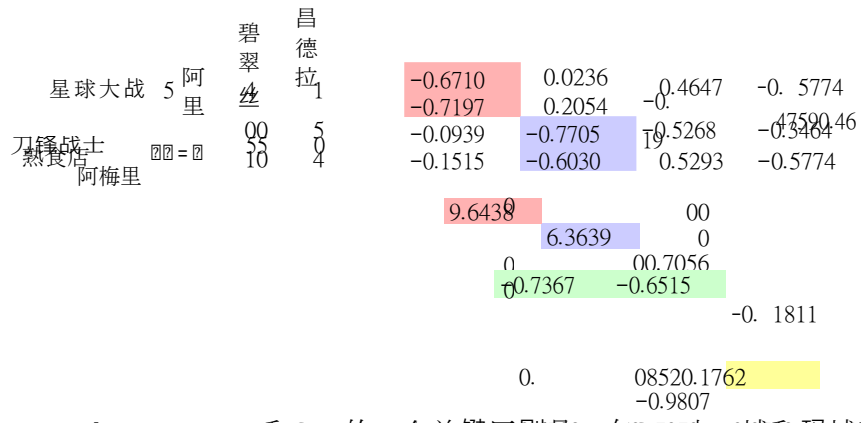
请注意，在计算机上，这里说明的方法有很差的数值行为，通常计算$\boldsymbol{A}$的SVD时，不需要求助于$\boldsymbol{A}^\top\boldsymbol{A}$的特征值分解。

4.5.3 特征值分解与奇异值分解

让我们考虑特征分解$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$和SVD $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$并回顾过去各节的核心内容。

- SVD对于任何矩阵$\mathbb{R}^{m\times n}$都是存在的，而eigendecomposition只对方形矩阵$\mathbb{R}^{n\times n}$定义，并且只有在我们能找到$\mathbb{R}^n$的特征向量的基础时才存在。
- 异质分解矩阵$\boldsymbol{P}$中的向量不一定是正交的，也就是说，基础的改变不是简单的旋转和缩放。另一方面，SVD中的矩阵$\boldsymbol{U}$和$\boldsymbol{V}$中的向量是正交的，所以它们确实代表旋转。
- eigendecomposition和SVD都是三个线性映射的组合。

  1. 领域内基础的改变
  2. 每个新的基向量的独立缩放，以及从做主域到代码域的映射
  3. 改变典范领域的基础



图：三个人对四部电影的**4.10**评分及其SVD分解。

eigendecomposition和SVD的一个关键区别是，在SVD中，域和码域可以是不同维度的向量空间。

- 在SVD中，左弦和右弦矢量矩阵$\boldsymbol{U}$和$\boldsymbol{V}$通常不是相互逆的（它们在不同的矢量空间中进行基改）。在eigendecomposition中，基数变化矩阵$\boldsymbol{P}$和$\boldsymbol{P}$是$^{-1}$相互反的。
- 在SVD中，对角线矩阵$\boldsymbol{\Sigma}$中的条目都是实数，且非负的，这对于eigendecomposition中的对角线矩阵来说一般不是真的。
- SVD和eigendecomposition通过它们的投影密切相关
  - $\boldsymbol{A}$的左弦向量是$\boldsymbol{AA}$的特征向量$^{\mathrm{T}}$
  - $\boldsymbol{A}$的右弦向量是$^{\mathrm{T}}\boldsymbol{AA}$的特征向量。
  - $\boldsymbol{A}$的非零奇异值是$\boldsymbol{AA}^{\mathrm{T}}$和$^{\mathrm{T}}\boldsymbol{AA}$的非零特征值的平方根。
- 对于对称矩阵$\boldsymbol{A}\ \mathrm{R}$来说$^{n\times n}\in$ 特征值分解和SVD是一样的，这从光谱theo-rem中可以看出。4.15.

例子（4.14在电影评分和消费者中寻找结构）让我们通过分析人们和他们喜欢的电影的数据来增加SVD的实际解释。考虑三个观众（Ali、Beatrix、Chandra）对四部不同的电影（《*星球大战*》、《*银翼杀手*》、《*阿梅里*》、《*熱食店*》）。他们的评级是0（最差）和5（最好）之间的数值，并且编码的数据矩阵$\boldsymbol{A}\in\mathbb{R}^{4\times3}$如图所示4.10每一行代表一部电影，每一列代表一个用户。因此，电影评分的列向量，每个观众一个，是$\boldsymbol{x}_{\mathrm{Ali}}$，$x_{\mathrm{Beatrix}}$，$x_{\mathrm{Chandra}}$。

使用SVD对$\boldsymbol{A}$进行因式分解为我们提供了一种方法来捕捉人们如何评价电影的关系，特别是如果有一个结构将哪些人喜欢哪些电影联系起来。在我们的数据矩阵$\boldsymbol{A}$上应用SVD，有一些假设。

1.所有观众使用相同的线性映射对电影进行一致的评分。2.评分中没有错误或噪音。

3.我们把左弦向量$\boldsymbol{u}$解释$_i$为定型电影，右弦向量$\boldsymbol{v}$解释$_j$为定型观众。

因此，SVD域中的一个向量可以被解释为定型观众 "空间 "中的一个观众，而SVD码域中的一个向量则相应地被解释为定型电影 "空间 "中的一个电影。让我们检查一下我们的电影-用户矩阵的SVD。第一个左弦向量$\boldsymbol{u}_1$这两部科幻电影的绝对值很大，而在《大话西游》中则有一个很大的第一奇异值（图中的红色底纹4.10).因此，这将一类用户与一组特定的电影（科幻主题）进行分组。同样地，第一个右单数$\boldsymbol{v}_1$显示了Ali和Beatrix的大绝对值，他们对科幻电影的评价很高（图中的绿色阴影4.10).这表明，$\boldsymbol{v}_1$反映了科幻小说爱好者的概念。

只有当数据本身涵盖了足够多的观众和电影，这两个 "空间 "才会被各自的观众和电影数据有意义地横跨。

同样，$\boldsymbol{u}_2$，似乎抓住了法国艺术电影的主题，而$\boldsymbol{v}_2$in-这表明钱德拉接近于这种电影的理想化爱好者。一个想法是化的科幻爱好者是一个纯粹主义者，只爱科幻电影，所以科幻爱好者$\boldsymbol{v}_1$对除科幻主题外的一切都给予零分--这个逻辑是由奇异值矩阵$\boldsymbol{\Sigma}$的对角线子结构暗示的。因此，一部具体的电影是由它如何（线性地）分解成其定型电影来表示的。同样，一个人也会通过他们如何分解（通过线性组合）为电影主题来表示。

值得简单讨论一下SVD的术语和惯例，因为文献中使用了不同的版本。虽然这些差异可能令人困惑，但数学仍然不受它们影响。

- 为方便记述和抽象，我们使用SVD记述，其中SVD被描述为有两个方形的左和右奇异向量矩阵，但有一个非方形奇异值矩阵。我们对SVD的定义(4.64)的SVD有时被称为*全SVD*。

完整的SVD

- 一些作者对SVD的定义有些不同，他们关注的是方形正弦矩阵。那么，对于$\boldsymbol{A}\in\mathbb{R}^{m\times n}$和$m\geqslant n$。

$$\underset{m\times}{\boldsymbol{A}} = \underset{nm\times n}{\boldsymbol{U}}\ \underset{n\times n}{\boldsymbol{\Sigma}}\ \underset{n\times n}{\boldsymbol{V}}^{\mathrm{T}}。 \tag{4.89}$$

有时，这种表述被称为*还原SVD*（例如，Datta（2010））或SVD（例如，Press等人（2007））。这种替代格式仅仅改变了矩阵的构造方式，但没有改变SVD的数学结构。这种替代形式的便利之处在于，$\boldsymbol{\Sigma}$是对角线，就像在特征值分解中一样。 还原SVD

- 在第4.6,我们将学习矩阵近似技术
  使用SVD，这也被称为*截断SVD*。 截断 的SVD
- 可以定义一个秩-r矩阵$\boldsymbol{A}$的SVD，使$\boldsymbol{U}$是一个$m \; r$矩阵，$\boldsymbol{\Sigma}$是一个尺寸为$r \; r$*的*对角线矩阵，而$\boldsymbol{V}$是一个$n \; r$*矩阵*。这种结构与我们的定义非常相似，并确保对角线矩阵$\boldsymbol{\Sigma}$沿对角线只有非零项。这种替代符号的主要方便之处在于，$\boldsymbol{\Sigma}$是对角线，就像在特征值分解中一样。
- 限制$\boldsymbol{A}$的SVD只适用于m>n的矩阵，实际上是没有必要的。当$m < n$*时*，SVD分解将产生$\boldsymbol{\Sigma}$，其零列多于行，因此，奇异值$\sigma_{m+1}, \ldots, \sigma_n$是0。

SVD被用于机器学习的各种应用中，从曲线拟合中的最小二乘问题到解决线性方程组。这些应用利用了SVD的各种重要特性，它与矩阵等级的关系，以及它用低等级矩阵近似某个等级的矩阵的能力。用SVD替代矩阵的优点往往是使计算对细微的四舍五入错误更加稳定。正如我们将在下一节探讨的那样，SVD能够以一种原则性的方式用 "更简单 "的矩阵来近似矩阵，这为机器学习的应用开辟了道路，从降维和主题建模到数据压缩和聚类。

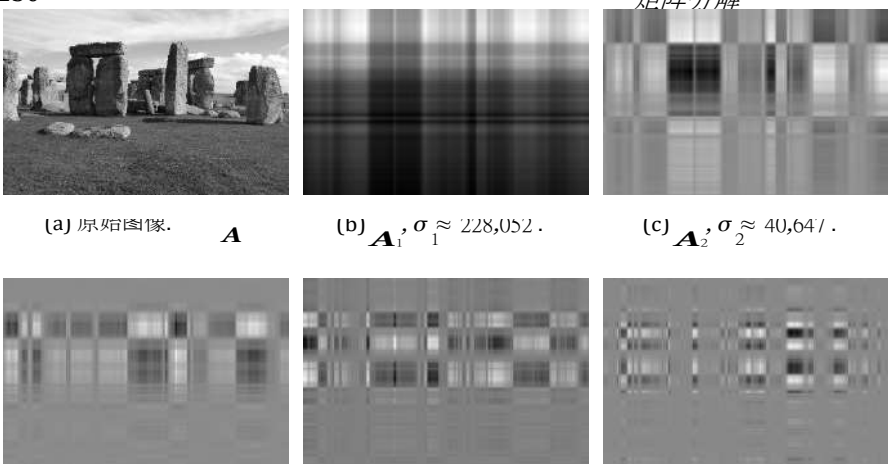
4.6 矩阵逼近

我们认为SVD是将$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \in \mathbb{R}^{m\times n}$分解为三个矩阵的乘积的一种方法，其中$\boldsymbol{U} \in \mathbb{R}^{m\times m}$和$\boldsymbol{V} \in \mathbb{R}^{n\times n}$是正交的，$\boldsymbol{\Sigma}$包含其主对角线上的奇异值。我们现在不做完整的SVD分解，而是研究SVD如何让我们把矩阵$\boldsymbol{A}$表示为更简单的（低秩）矩阵$\boldsymbol{A}_i$的总和，这有利于我们采用比完整SVD更便宜的矩阵逼近方案。

我们构建一个等级1矩阵$\boldsymbol{A}_i \in \mathbb{R}^{m\times n}$为

$$\boldsymbol{A}_i := \boldsymbol{u}_i \boldsymbol{v}_i^\top \tag{4.90}$$

这是由$\boldsymbol{U}$和$\boldsymbol{V}$的第1个正交列向量的外积形成的。 图4.11显示了巨石阵的图像，它可以用矩阵$\boldsymbol{A} \in \mathbb{R}^{1432\times 1910}$和一些外积$\boldsymbol{A}_i$来表示，定义见(4.90).



(a) 原始图像。 $\boldsymbol{A}$ (b) $\boldsymbol{A}_1, \sigma_1 \approx 228{,}052$. (c) $\boldsymbol{A}_2, \sigma_2 \approx 40{,}647$.

(d) $\boldsymbol{A}_3, \sigma_3 \approx 26{,}125$. (e) $\boldsymbol{A}_4, \sigma_4 \approx 20{,}232$. (f) $\boldsymbol{A}_5, \sigma_5 \approx 15{,}436$.

图： 用SVD进行**4.11**图像处理。(a) 原始灰度图像是一个 $1{,}\quad \times 4321{,}910$ 之间的数值矩阵（0黑色）。和（1白色）。(b)-(f)等级-1矩阵 $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_5$ 和其相应的奇异值 $\sigma_1, \ldots, \sigma_5$.该每个秩-1矩阵的网格状结构是由左和右的外积强加的。右弦向量。

一个等级为r的矩阵$\boldsymbol{A} \in \mathbb{R}^{m\times n}$可以写成等级为1的矩阵之和 $\boldsymbol{A}_i$以便于

$$\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}} = \sum_{i=1}^{r} \sigma_i \boldsymbol{A}_i. \tag{4.91}$$

其中外积矩阵$\boldsymbol{A}_i$是由第1个奇异值σ_i加权的。我们可以看到为什么(4.91)成立。奇异值矩阵$\boldsymbol{\Sigma}$的对角线结构只与匹配的左、右奇异值向量 $\boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$相乘，并以相应的奇异值$\sigma_i$为尺度。所有项$\Sigma_{ij}\boldsymbol{u}_i\boldsymbol{v}_j^{\mathrm{T}}$在$i=j$时都消失了，因为$\boldsymbol{\Sigma}$是一个对角线矩阵。任何$i>r$的项都消失了，因为相应的奇异值是0。

在(4.90)中，我们引入了秩-1矩阵$\boldsymbol{A}_i$。我们将r个单独的秩-1矩阵相加，得到一个秩-r矩阵$\boldsymbol{A}$；见(4.91).如果总和不超过所有的矩阵$\boldsymbol{A}_i$，$i = 1, \ldots, r$，但只到一个中间值$k < r$，我们得到一个*秩-k近似值*

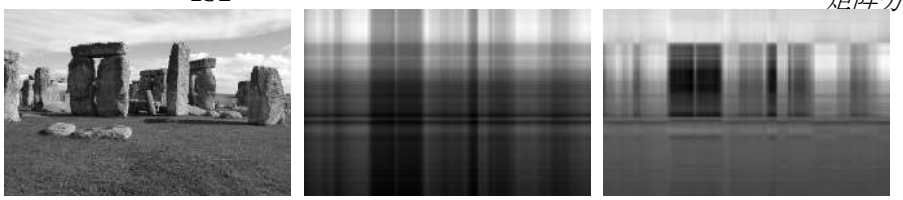
等级-k近似值

$$\hat{\boldsymbol{A}}(k) := \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}} = \sum_{i=1}^{k} \sigma_i \boldsymbol{A}_i \tag{4.92}$$

的$\mathrm{rk}(\hat{\boldsymbol{A}}(k))=k$。图4.12显示了低级别的近似图石阵的原始图像$\boldsymbol{A}$的$\hat{\boldsymbol{A}}(k)$。在Rank-5近似中，岩石的形状越来越明显，可以清楚地识别。While the original image requires 1，1，432= 9102，735，numbers120，the rank-5 approximation requires us-only to store the five sin- gular values and the five left- and right-singular vectors (1，432 and 1，910- dimensional each) for a total of (51，+ 4321，+ 9101) = 16，numbers715 -
- 仅高于0原来的.6%。

为了衡量$\boldsymbol{A}$和它的等级-k近似值$\hat{\boldsymbol{A}}(k)$之间的差异（误差），我们需要

 一个规范的概念。在本节3.1中， 我们已经使用了



(a) 原始图像。 $\boldsymbol{A}$ (b) 等级1的近似值 $\boldsymbol{A}_{(1)}$。 (c) 等级2的近似值 $\boldsymbol{A}_{(2)}$。

(d) 等级3的近似值 $\boldsymbol{A}_{(3)}$。 (e) 等级4的近似值 $\boldsymbol{A}_{(4)}$。 (f) 等级5的近似值 $\boldsymbol{A}_{(5)}$。

图： 用SVD **4.12**重建图像。(a)原始图像。(b)-(f)使用SVD的低秩近似值进行图像重建，其中秩-k近似值由 $\boldsymbol{A}_{(k)} = \sum_{i=1}^{k} \sigma_i \boldsymbol{A}_i$ 。

衡量向量长度的向量上的规范。以此类推，我们也可以定义矩阵的规范。

定义（4.23矩阵的谱系规范）。对于 $\boldsymbol{x} \in \mathbb{R}^n / \{\boldsymbol{0}\}$，*光谱*矩阵 $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ 的*准则*定义为

规范是指

$$\|\boldsymbol{A}\|_2 := \max_{\boldsymbol{x}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}. \tag{4.93}$$

我们在矩阵规范（左侧）中引入了下标的符号，类似于向量的欧几里得规范（右侧），其下标为 2。谱系规范 (4.93)决定了任何向量 $\boldsymbol{x}$ 在与 $\boldsymbol{A}$ 相乘时最多可以变成多长。

定理 $\boldsymbol{A}$ *的* **4.24.** *谱准则是其最大的奇异值* σ_1。

我们把这个定理的证明留作练习。

埃卡特-杨定理

定理（Eckart **4.25**-Young Theorem (Eckart and Young,1936)）。*锥体考虑一个等级为* r *的矩阵* $\boldsymbol{A} \in \mathbb{R}^{m \times n}$，*让* $\boldsymbol{B} \in \mathbb{R}^{m \times n}$ *是一个等级为* r *的矩阵。* k. *对于任何* k:(r *与* $\hat{\boldsymbol{A}}(k) = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$，*则认为*

$$\hat{\boldsymbol{A}}(k) = \operatorname{argmin}_{\operatorname{rk}(\boldsymbol{B})=k} \|\boldsymbol{A} - \boldsymbol{B}\|_2 \text{。} \tag{4.94}$$

$$\|\boldsymbol{A} - \hat{\boldsymbol{A}}(k)\|_2 = \sigma_{k+1} \text{。} \tag{4.95}$$

Eckart-Young定理明确指出了我们使用秩-k*近似*法对 $\boldsymbol{A}$ *进行近似*会带来多大的误差。我们可以把用SVD得到的秩-k近似当作全秩矩阵 $\boldsymbol{A}$ 在最多为 k 个秩矩阵的低维空间上的投影。在所有可能的投影中，SVD使 $\boldsymbol{A}$ 和任何秩-k*近似值*之间的误差最小（相对于谱准则）。

我们可以回溯一些步骤来理解为什么(4.95)应该成立。

我们观察到，$\hat{\boldsymbol{A}}(k)$之间的差异是一个包含其余等级1矩阵之和的矩阵

$$\boldsymbol{A} - \hat{\boldsymbol{A}}(k) = \sum_{i=k+1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}. \tag{4.96}$$

根据该定理，4.24,我们立即得到σ_{k+1}是差分矩阵的谱准则。让我们仔细看一下(4.94).如果我们假设有另一个矩阵$\boldsymbol{B}$，rk($\boldsymbol{B}$):(k，这样

$$\|\boldsymbol{A} - \boldsymbol{B}\|_2 < \|\boldsymbol{A} - \hat{\boldsymbol{A}}(k)\|_2, \tag{4.97}$$

那么存在一个至少($n-k$)-维的空空间$Z \subseteq \mathbb{R}^n$，这样$\boldsymbol{x} \in Z$意味着$\boldsymbol{B}\boldsymbol{x} = \boldsymbol{0}$。那么就可以得出

$$\|\boldsymbol{A}\boldsymbol{x}\|_2 = \|(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{x}\|_2, \tag{4.98}$$

通过使用考奇-施瓦茨不等式的一个版本(3.17)，它包含了矩阵的规范，我们可以得到

$$\|\boldsymbol{A}\boldsymbol{x}\|_2 \leq \|\boldsymbol{A} - \boldsymbol{B}\|_2 \|\boldsymbol{x}\|_2 < \sigma_{k+1}\|\boldsymbol{x}\|_2. \tag{4.99}$$

然而，存在着一个(k+1)维的子空间，其中$\|\boldsymbol{A}\boldsymbol{x}\|_2$。二 $\sigma_{k+1}\boldsymbol{x}$，它被右弦向量$\boldsymbol{v}_j, j \leq k+1$的横跨。$\boldsymbol{A}$.将这两个空间的尺寸相加得出的数字大于$n$，因为在这两个空间中都必须有一个非零矢量。这与第2.7.3节中的等级空性定理（Theorem2.24）相矛盾。

Eckart-Young定理意味着我们可以用SVD来减少一个

在一个原则性的、最佳的（在 "我的 "中）等级-r矩阵$\boldsymbol{A}$到等级-k矩阵$\boldsymbol{A}$谱规范意义上的）方式。我们可以把用k级矩阵对$\boldsymbol{A}$的逼近解释为一种有损压缩的形式。因此，矩阵的低秩近似出现在许多机器学习的应用中，例如，图像处理、噪声过滤和不理想问题的正则化。此外，它在降维和主成分分析中起着关键作用，我们将在第十章中看到这一点。

例子（4.15在电影评分和消费者中寻找结构（续））。

回到我们的电影评级例子，我们现在可以应用低等级近似的概念来近似原始数据矩阵。回顾一下，我们的第一个奇异值抓住了电影和科幻小说爱好者的科幻主题的概念。因此，通过在电影评分矩阵的秩-1分解中只使用第一个奇异值项，我们得到预测的评分

$$\boldsymbol{A}_1 = \boldsymbol{u}_1 \boldsymbol{v}_1^{\mathrm{T}} = \begin{bmatrix} -0.6710 \\ -0.7197 \\ -0.0939 \\ -0.1515 \end{bmatrix} \begin{bmatrix} -0.7367 & -0.6515 & -0.1811 \end{bmatrix} \tag{4.100a}$$

$$= \begin{matrix} 0.49430 & .43720 & .1215 \\ 0.53020 & .46890 & .1303 \\ 0.06920 & .0612 & .00170 \\ 0.11160 & .09870 & .0274 \end{matrix} . \tag{4.100b}$$

这个第一等级-1的近似值$\boldsymbol{A}_1$很有见地：它告诉我们，阿里和贝特里克斯喜欢科幻电影，比如*《星球大战》*和*《黑客帝国》*（条目值>0 .4 ），但未能捕捉到钱德拉对其他电影的评价。这并不奇怪，因为Chandra的电影类型并没有被第一个奇异值所捕获。第二个奇异值为我们提供了一个更好的等级-1-近似于那些电影主题爱好者。

$$\boldsymbol{A} = \boldsymbol{u}_2 \boldsymbol{v}_2^{\mathrm{T}} = \begin{bmatrix} 0.0236 \\ 0.2054 \\ 0.0852 \\ 0.1762 \\ -0.9807 \end{bmatrix} \begin{bmatrix} -0.7705 & -0.6030 \end{bmatrix} \tag{4.101a}$$

$$= \begin{matrix} 0.00200.0042 & -0.0231 \\ 0.01750.0362 & -0.2014 \\ -0.0656 & -0.13580.7556 \\ -0.0514 & -0.10630.5914 \end{matrix} \tag{4.101b}$$

在这个第二等级-1近似值$\boldsymbol{A}$中$_2$，我们很好地捕捉到了Chandra的评分和电影类型，但没有捕捉到科幻电影。这导致我们考虑等级2的近似值$\boldsymbol{A}(2)$，我们将前两个等级1近似值结合起来

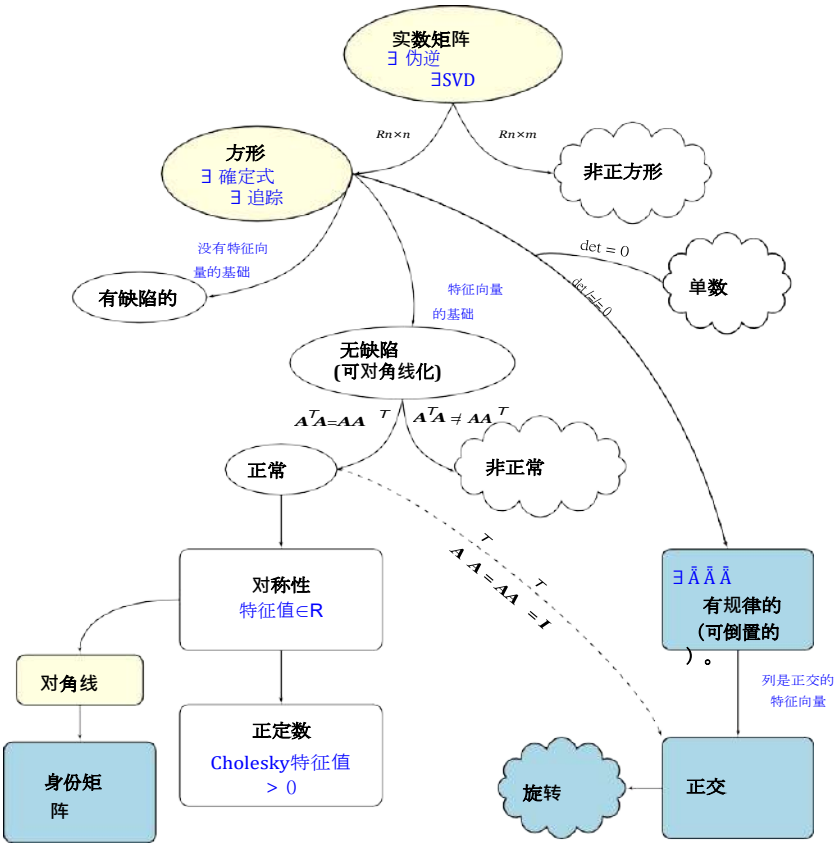
$$\boldsymbol{A}(2) = \sigma_1 \boldsymbol{A}_1 + \sigma_2 \boldsymbol{A}_2 = \begin{matrix} 4.78014. & 24191.0244 \\ 5.22524. & -75220.0250 \\ 0.-24930. & 27434.9724 \\ 0.74950. & 27564.0278 \end{matrix} \tag{4.102}$$

$\boldsymbol{A}(2)$类似于原来的电影评级表

$$\boldsymbol{A} = \begin{matrix} 5 & 5 & .(0 \\ 0 & 0 & 5 \\ 1 & 0 & 4 \end{matrix} \tag{4.103}$$

这表明我们可以忽略$\boldsymbol{A}$的贡献$_3$。我们可以这样解释，在数据表中没有证据表明存在第三个电影主题/电影爱好者类别。这也意味着，在我们的例子中，整个电影主题/电影爱好者的空间是一个由科幻小说和法国艺术电影及爱好者跨越的二维空间。

图为机器学习中遇到的矩阵的**4.13**功能系统发育。



4.7 矩阵系统学

系统发育"这个词描述了我们如何捕捉个体或群体之间的关系，它来自希腊语中的"部落"和"来源"。

在第二章和第三章中，我们介绍了线性代数和解析几何的基础知识。在本章中，我们研究了矩阵和线性映射的基本特征。图4.13描述了不同类型的矩阵（黑色箭头表示"是其子集"）和我们可以对其进行的覆盖操作（蓝色）之间的系统发育树。我们考虑所有的*实数矩阵*$A \in \mathbb{R}^{n \times m}$。对于非正方形矩阵（$n=m$），SVD总是存在的，正如我们在本章看到的那样。关注*正方形矩阵*$A \in \mathbb{R}^{n \times n}$，*行列式*告诉我们一个正方形矩阵是否拥有一个*逆矩阵*，也就是说，它是否属于规则的、可逆的矩阵类别。如果正方形$n \times n$矩阵拥有n个线性独立的特征向量，那么该矩阵是*无缺陷的*，存在一个*特征分解*（定理4.12).我们知道，重复的特征值可能会导致有缺陷的矩阵，这些矩阵不能被对角化。

非星形矩阵和非缺陷矩阵是不一样的。例如，一个旋转矩阵将是可逆的（行列式为非零），但在实数中不可对角化（特征值不能保证是实数）。

我们进一步深入到非缺陷方阵$n \times n$矩阵的分支。如果$\boldsymbol{A}\boldsymbol{A}^{\top} = \boldsymbol{A}^{\top}\boldsymbol{A}$的条件成立，$\boldsymbol{A}$就是*正常的*。此外，如果更严格的条件$\boldsymbol{A}\boldsymbol{A}^{\top} = \boldsymbol{A}^{\top}\boldsymbol{A} = \boldsymbol{I}$成立，那么$\boldsymbol{A}$就被称为*正交*（见定义3.8）。正交矩阵的集合是正交（可逆）矩阵的一个子集，并且满足$\boldsymbol{A}^{\top} = \boldsymbol{A}^{-1}$。

正态矩阵有一个经常遇到的子集，即对称矩阵。对称矩阵只具有实数特征值。对称矩阵的一个子集由满足$\boldsymbol{x}^{\top}\boldsymbol{P}\boldsymbol{x} > 0$条件的正定矩阵$\boldsymbol{P}$组成，对于所有$\boldsymbol{x} \in \mathbb{R}^n \setminus \boldsymbol{0}$。在这种情况下，存在一个唯一的*Cholesky分解*（Theo- rem4.18).正定矩阵只有正的特征值，并且总是可逆的（即有一个非零行列式）。

对称矩阵的另一个子集包括*对角线矩阵*$\boldsymbol{D}$.对角线矩阵在乘法和加法下是封闭的，但不一定形成一个群（只有在所有对角线条目都是非零的情况下才是这样，这样的矩阵是可倒置的）。一个特殊的对角线矩阵是身份矩阵$\boldsymbol{I}$。

4.8进一步阅读

本章的大部分内容建立了基础数学，并将它们与研究映射的方法联系起来，其中许多方法是机器学习的核心，是支撑软件的脚本和几乎所有机器学习理论的构件。使用行列式、特征谱和特征空间的矩阵表征为分类和分析矩阵提供了基本特征和条件。这延伸到了所有形式的数据表示和涉及数据的映射，以及判断这些矩阵上压缩操作的数值稳定性（Press等人，2007）。

决定数是反转矩阵和 "手工 "计算特征值的基本工具。然而，除了最小的实例，几乎所有的高斯消除法的数值计算都优于行列式（Press等人，2007）。然而，行列式仍然是一个强大的理论概念，例如，根据行列式的符号来获得关于基础方向的直观信息。特征向量可以用来进行基数变化，将数据转化为平均正交的特征向量的坐标。同样，矩阵分解方法，如Cholesky分解，在我们计算或模拟随机事件时经常出现（Rubinstein和Kroese，2016）。因此，Cholesky分解使我们能够在我们想要对随机变量进行连续微分的地方计算出*重构技巧，*例如在变异自动编码器中（Jimenez Rezende等人，2014；Kingma和Welling，2014）。

Eigendecomposition是使我们能够提取表征线性映射的平均和可解释信息的根本。

因此，eigendecomposition是一类通用的机器学习算法的基础，这些算法被称为*频谱方法*，对一个正不定核进行eigendecomposi- tion。这些频谱分解方法包含了经典的统计数据分析方法，如：。

主成分分析

- *主成分分析*（PCA（Pearson，1901），另见第10章），其中一个低维的子空间，可以解释大部分的变量。

费希尔判别法分析

- *Fisher判别分析*，其目的是确定一个分离的hy---。

多维度缩放（MDS）

每平面的数据分类（Mika等人，1999）。

- *缩放多维缩放*（MDS）（Carroll和Chang，1970）。

这些方法的计算效率通常来自于找到对称、正半无限矩阵的最佳等级-k*近似值*。更多当代谱系方法的例子有不同的起源，但它们都需要计算特征向量

卫星定位系统(Isomap)

诸如*Isomap*（Tenenbaum等人，2000）、

拉普拉斯的特征图

Laplacian eigenmaps（Belkin和Niyogi，2003）、

Hessian eigenmaps

Hessian eigenmaps（Donoho和Grimes，2003）以及*光谱聚类*（Shi和Malik，

谱系聚类

2000）。这些的核心计算通常由低等级矩阵近似技术支撑（Belabbas和Wolfe,2009），我们在这里通过SVD遇到了。

SVD允许我们发现一些与eigendecomposition相同的信息。然而，SVD更普遍地适用于非方形矩阵和数据表。当我们想通过近似的方式进行数据压缩时，这些矩阵分解方法就变得很重要了，例如，不存储n m值，只存储$(n+m)k$值，或者当我们想要
来进行数据预处理，例如，对 "我的 "和 "我的 "的预测变量进行关联。设计矩阵（Ormoneit等人，2001）。SVD对矩阵进行操作，我们可以将其解释为具有两个索引（行和列）的矩形数组。矩阵类结构对高维数组的扩展被称为张量。事实证明，SVD是对这种张量进行操作的更普遍的分解系列的特例（Kolda和Bader，2009）。类似于SVD的操作和对张量的低秩近似，例如，*Tucker分解*（Tucker,1966）。

塔克分解

或*CP分解*（Carroll and Chang，1970）。

CP分解

SVD低秩近似在机器学习中经常被用于计算效率的原因。这是因为它减少了我们需要对可能非常大的数据矩阵进行的内存和非零乘法的操作（Trefethen and Bau III, 1997）。此外，低秩近似被用来对可能包含缺失值

的矩阵进行操作，以及用于有损压缩和降维的目的（Moonen和De Moor，1995；Markovsky，2011）。

练习

4.1 使用拉普拉斯展开法（使用第一行）和Sarrus法则计算行列式，对于

$$\boldsymbol{A} = \begin{bmatrix} 3 & 1 & 5 \\ 0 & 2 & 4 \\ 2 & 4 & 6 \end{bmatrix}.$$

4.2 有效计算以下行列式。

$$\begin{vmatrix} 0 & 1 & 2 & 2 \\ -1 & 0 & 1 & 2 \\ 1 & 2 & 1 & 0 \\ 0 & 2 & -1 & 2 \\ 0 & 0 & 1 & 2 \end{vmatrix}.$$

4.3 计算a的特征空间。

$$\boldsymbol{A} := \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$$

b.

$$\boldsymbol{B} := \begin{bmatrix} -2 & 2 \\ 2 & 1 \end{bmatrix}$$

4.4 计算所有的特征空间

$$\boldsymbol{A} = \begin{bmatrix} 0 & -1 & 1 & 1 \\ -1 & 1 & -2 & 3 \\ 2 & -1 & 0 & 0 \\ 1 & -1 & 1 & 0 \end{bmatrix}.$$

4.5 矩阵的可对角性与它的可逆性没有关系。确定以下四个矩阵是否可对角化和/或可反转

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}.$$

4.6 计算下列变换矩阵的特征空间。它们是可对角的吗？

a.对于

$$\boldsymbol{A} = \begin{bmatrix} 2 & 3 & 0 \\ 1 & 4 & 3 \\ 0 & 0 & 1 \end{bmatrix}$$

b.对于

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

4.7 以下矩阵是否可对角线化？如果是，请确定它们的对角线形式，以及变换矩阵是对角线的一个基。如果不是，请说明它们不能对角的原因。

a.

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -8 & 4 \end{bmatrix}$$

b.

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

c.

$$\boldsymbol{A} = \begin{bmatrix} 5 & 4 & 2 & 1 \\ 0 & 1 & -1 & -1 \\ -1 & -1 & 3 & 0 \\ 1 & 1 & -1 & 2 \end{bmatrix}$$

d.

$$\boldsymbol{A} = \begin{bmatrix} 5 & -6 & 6 \\ 3 & -6 & -4 \\ 1 & 4 & 2 \end{bmatrix}$$

4.8 寻找矩阵的SVD $\boldsymbol{A} = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix}$.

4.9 Find the singular value decomposition of

$$\boldsymbol{A} = \begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix}.$$

4.10 Find the rank-1 approximation of

$$\boldsymbol{A} = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix}$$

4.11 证明对于任何$\boldsymbol{A} \in \mathbb{R}^{m \times n}$，矩阵$\boldsymbol{A}^\top \boldsymbol{A}$和$\boldsymbol{A}\boldsymbol{A}^\top$拥有相同的非零特征值。

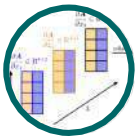
4.12 证明对于$\boldsymbol{x} \neq \boldsymbol{0}$定理4.24成立，也就是说，证明

$$\max_{\boldsymbol{x}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} = \sigma_1$$

其中σ_1是$\boldsymbol{A} \in \mathbb{R}^{m \times n}$的最大奇异值。

5

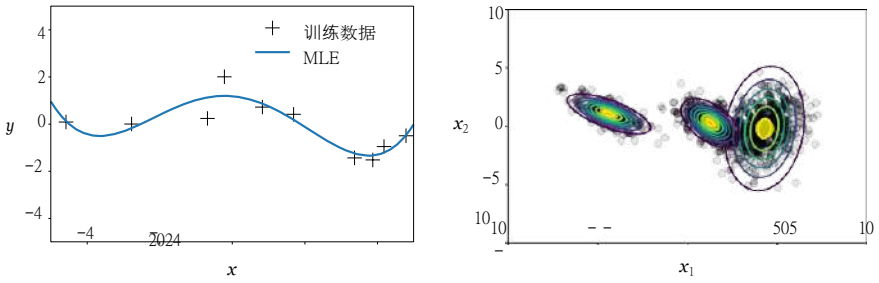
矢量微积分



机器学习中的许多算法都是针对一组理想的模型参数来优化目标函数的，这些参数控制着模型解释数据的程度。寻找好的参数可以被表述为一个优化问题（见第8.2和8.3节）。例子包括。(i) 线性回归（见第9章），我们研究曲线拟合问题并优化线性权重参数以最大化似然；(ii) 用于降维和数据压缩的神经网络自动编码器，参数是每层的权重和偏置，我们通过重复应用链式规则使重建误差最小。(iii) 高斯混合模型（见第11章）用于数据分布建模，我们优化每个混合成分的位置和形状参数，使模型的可能性最大化。图5.1说明了其中一些问题，我们通常使用利用梯度信息的优化算法来解决这些问题（第7.1节）。图5.2概述了本章中的概念是如何关联的，以及它们与本书其他章节的联系。

本章的核心是函数的概念。 一个函数f是一个将两个量相互联系起来的数量。在本书中，这些量通常是输入$\boldsymbol{x} \in \mathbb{R}^D$和目标（函数值）$f(\boldsymbol{x})$，如果没有特别说明，我们假定它们是实值的。这里的$\mathbb{R}^D$是指$f$的*域*，而函数值$f(\boldsymbol{x})$是$f$的*图像/代码域*。 域

图像/域名



(a) 回归问题：寻找参数，使曲线能很好地解释观察结果（交叉点）。 (b) 用高斯混合模型进行密度估计。找到均值和协方差，使数据（点）可以得到很好的解释。

图 矢量**5.1**微积分在(a)回归(曲线拟合)和(b)密度估计，即数据分布建模中起着核心作用。

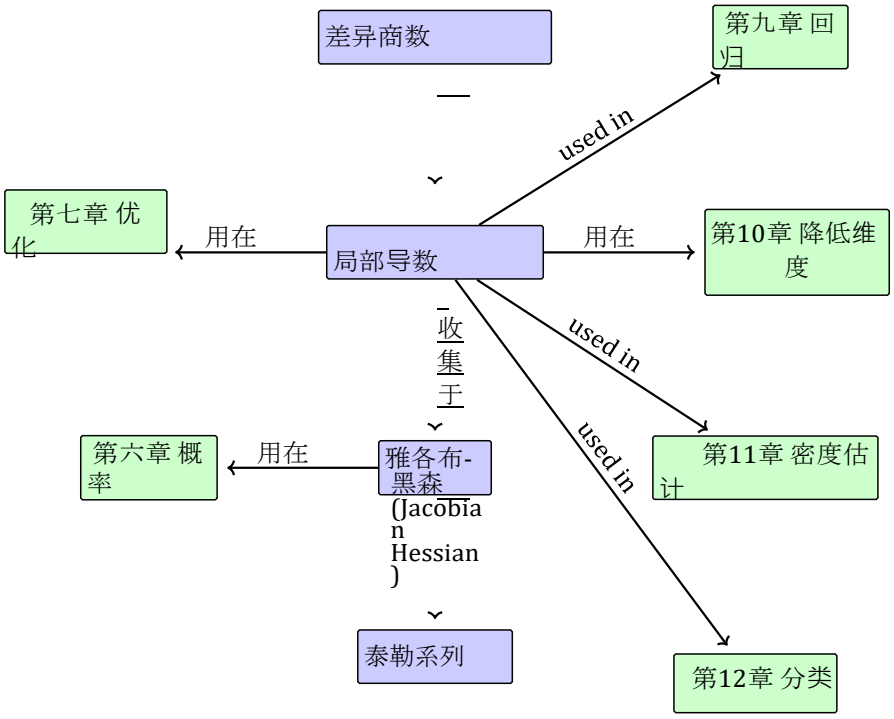


图 本章介绍的概念的**5.2**思维导图，以及它们在本书其他部分的使用时间。

第2.7.3节在线性函数的背景下提供了更详细的讨论。我们经常写

$$f:\mathbb{R}^D \to \mathbb{R} \tag{5.1a}$$

$$\boldsymbol{x} \to f(\boldsymbol{x}) \tag{5.1b}$$

来指定一个函数，其中（5.1a）指定f是一个从$\mathbb{R}^D$到$\mathbb{R}$的映射，（5.1b）指定一个输入$\boldsymbol{x}$到一个函数值$f(\boldsymbol{x})$的明确分配。一个函数f为每个输入$\boldsymbol{x}$精确地分配一个函数值$f(\boldsymbol{x})$。

例子 5.1

记得点积是内积的一个特例（第3.2节）。在前面的符号中，函数$f(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{x}^\top$，$\boldsymbol{x} \in \mathbb{R}^2$，将是指定为

$$f:\mathbb{R}^2 \to \mathbb{R} \tag{5.2a}$$

$$\boldsymbol{x} \to x_1^2 + x_2^2 \tag{5.2b}$$

在本章中，我们将讨论如何计算函数的梯度，这对于促进机器学习模型的学习往往是至关重要的，因为梯度指向最陡峭的上升方向。因此。

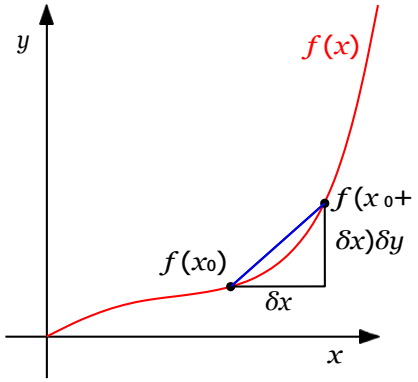


图5.3 函数f在x_0和$x_0+\delta x$之间的平均倾角是通过$f(x_0)$和$f(x_0+\delta x)$的正切（蓝色）的倾角，由$\delta y/\delta x$给出。

向量微积分是我们在机器学习中需要的基本数学工具之一。在本书中，我们假设函数是可微的。通过一些额外的技术定义（我们在此不做介绍），所介绍的许多方法可以扩展到次微分（连续但在某些点不可微分的函数）。我们将在第七章中探讨对有约束条件的函数情况的扩展。

5.1 单变量函数的微分

下面，我们简要地重温一下单变量函数的微分，这在高中数学中可能比较熟悉。我们从单变量函数$y = f(x), x, y \in R$的差商开始，随后我们将用它来定义导数。

定义（5.1差额商）。*差异商数(Difference* Quotientdifference quotient)

$$\frac{\delta y}{} := \frac{f(x+\delta x) - f(x)}{\delta x \delta x} \tag{5.3}$$

计算出通过图上两点的正切线的斜率。
f.在图5.3,中，这些是x坐标为x_0和$x_0+\delta x$的点。

如果我们假设f是一个线性函数，差商也可以被认为是f在x和$x+\delta x$之间的平均斜率。在$\delta 0 x$的极限中，我们得到f在x处的切线，如果f是可微的。那么切线就是f在x处的导数。

定义（5.2导数）。更正式地说，对于$h > 0$ f的*导数*的导数在x处被定义为极限

$$\frac{df}{dx} := \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}, \tag{5.4}$$

而图中的正切5.3变成了正切。

f的导数指向f的最陡峭的上升方向。

例子 (5.2多项式的导数)
我们想计算$f(x) = x^n$的导数，$n \in \mathbb{N}$。
我们知道答案是nx^{n-1}，但我们想用导数的定义来推导这个结果，即差额商的极限。

利用( )中导数的定义，我们可以得到5.4)，我们得到

$$\frac{df}{dx} = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} \tag{5.5a}$$

$$= \lim_{h\to 0} \frac{(x+h)^n - x^n}{h} \tag{5.5b}$$

$$= \lim_{h\to 0} \frac{\sum_{i=0}^{n} \binom{n}{i} x^{n-i} h^i - x^n}{h} \tag{5.5c}$$

我们看到，$x^n = \binom{n}{0} x^{n-0} h^0$通过从总和开始1，$x$-项被抵消。
而我们得到

$$\frac{df}{dx} = \lim_{h\to 0} \frac{\sum_{i=1}^{n} \binom{n}{i} x^{n-i} h^i}{h} \tag{5.6a}$$

$$= \lim_{h\to 0} \sum_{i=1}^{n} \binom{n}{i} x^{n-i} h^{i-1} \tag{5.6b}$$

$$= \lim_{h\to 0} \binom{n}{1} x^{n-1} + \sum_{i=2}^{n} \binom{n}{i} x^{n-i} h^{i-1} \tag{5.6c}$$

$$= \frac{n!}{1!(n-1)!} x^{n-1} = nx^{n-1} \tag{5.6d}$$

5.1.1 泰勒系列

泰勒级数是将一个函数f表示为一个无限大的项之和。这些项是用f在x_0处的导数决定的。

泰勒多项式 我们定义$t^0 := 1$对于所有$t \in \mathcal{R}$。

定义（5.3泰勒多项式）。 *泰勒多项式*的n度为f:在x处的x_0R→R被定义为

$$T_n(x) := \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k, \tag{5.7}$$

其中$f^{(k)}(x_0)$是f在x处的x_0第k次导数（我们假设存在），$\frac{f^{(k)}(x_0)}{k!}$是多项式的系数。

泰勒系列

定义（5.4泰勒系列）。 对于一个平滑函数$f \in C^\infty$，f: RR，f在x处x_0

的*泰勒级数*被定义为

$$T_\infty(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \tag{5.8}$$

对于$x_0 = 0$，我们得到*Maclaurin数列*作为$f \in C^\infty$的一个特殊实例，意味着泰勒数列。如果$f(x) = T_\infty(x)$，那么f就被称为*分析型*。

f是连续可微的无限次。Maclaurin系列分析法

Remark. In general, a Taylor polynomial of degree n is an approximation of a function, which does not need to be a polynomial. The Taylor polynomial is similar to f in a neighborhood around x_0. However, a Taylor polynomial of degree n is an exact representation of a polynomial f of degree $k \leqslant n$ since all derivatives $f^{(i)}$, $i > k$ vanish. ◆

例子（5.3泰勒多项式）

我们考虑多项式

$$f(x) = x^4 \tag{5.9}$$

并寻求泰勒多项式T_6，在$x_0 = 1$处求值。我们首先将$k = 0, \ldots, 6$的系数$f^{(k)}(1)$进行组合：

$$f(1) = 1 \tag{5.10}$$
$$f'(1) = 4 \tag{5.11}$$
$$f''(1) = 12 \tag{5.12}$$
$$f^{(3)}(1) = 24 \tag{5.13}$$
$$f^{(4)}(1) = 24 \tag{5.14}$$
$$f^{(5)}(1) = 0 \tag{5.15}$$
$$f^{(6)}(1) = 0 \tag{5.16}$$

因此，所需的泰勒多项式是

$$T_6(x) = \sum_{k=0}^{6} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \tag{5.17a}$$
$$= 1 + 4(x - 1) + 6(x - 1)^2 + 4(x - 1)^3 + (x - 1)^4 \tag{5.17b}$$

乘出和重新排列的产量

$$\begin{aligned} T_6(x) &= (1 - 4 + 6 - 4 + 1) + x(4 - 12 + 12 - 4) \\ &\quad + x^2(6 - 12 + 6) + x^3(4 - 4) + x^4 \end{aligned} \tag{5.18a}$$
$$= x^4 = f(x), \tag{5.18b}$$

即，我们得到了原函数的精确表示。

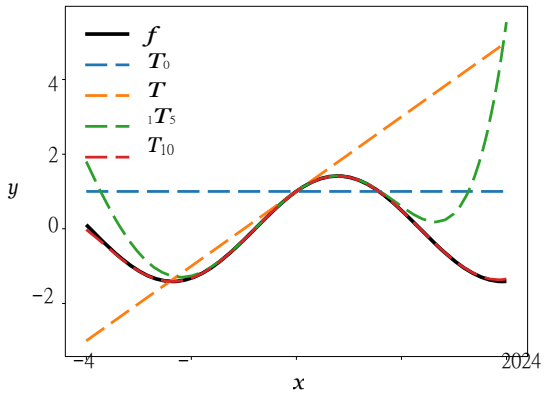


图5.4 泰勒多项式。原始函数$f(x) = \sin(x) + \cos(x)$(黑色，实线)由泰勒多项式(虚线)在$x_0=0$.周围近似。高阶泰勒polynomials approximate the function f better and more globally. T_{10} is already similar to f in $[-4, 4]$.

例子（5.4泰勒系列）

考虑图中的函数5.4中的函数，该函数由

$$f(x) = \sin(x) + \cos(x) \in C^\infty \tag{5.19}$$

我们寻求f在$x_0=$时0的泰勒级数展开，这就是f的Maclaurin级数展开。我们得到以下导数。

$$f(0) = \sin(0) + \cos(0) = 1 \tag{5.20}$$
$$f^{(1)}(0) = \cos(0) - \sin(0) = 1 \tag{5.21}$$
$$f^{(2)}(0) = -\sin(0) - \cos(0) = -1 \tag{5.22}$$
$$f^{(3)}(0) = -\cos(0) + \sin(0) = -(1 \tag{5.23}$$
$$f^{(4)}(0) = \sin(0) + \cos(0) = f(0) = 1 \tag{5.24}$$

.

我们在这里可以看到一个模式。我们的泰勒序列中的系数只有± 1（因为$\sin(0)=0$），每一个都在切换到之前出现了两次另一个。此外，$f^{(k+4)}(0) = f^{(k)}(0)$。

因此，f在$x_0=$处的全泰勒级数展开是由以下公式0给出的

$$T_\infty(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \tag{5.25a}$$
$$= 1 + x - \frac{1}{2!}x^2 - \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \frac{1}{5!}x^5 - \cdots \tag{5.25b}$$
$$= 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 \mp \cdots + x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 \mp \cdots \tag{5.25c}$$
$$= \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k)!}x^{2k} + \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k+1)!}x^{2k+1} \tag{5.25d}$$
$$= \cos(x) + \sin(x) \tag{5.25e}$$

。

其中我们使用了*幂级数表示法* power series representation

$$\cos(x) = \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k)!} x^{2k} \quad (5.26)$$

$$\sin(x) = \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k+1)!} x^{2k+1} \quad (5.27)$$

图中5.4显示了相应的第一泰勒多项式 T_n, $n=0,1,5,10$.

备注。 泰勒级数是幂级数的一个特例

$$f(x) = \sum_{k=0}^{\infty} a_k (x-c)^k \quad (5.28)。$$

其中a_k为系数，c为常数，具有定义中的特殊形式5.4. ◆

5.1.2 差异化规则

在下文中，我们简要说明基本的微分规则，其中我们用f'表示f的导数。

乘积规则： $(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$ (5.29)

商数规则： $\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}$ (5.30)

Sum rule: $(f(x) + g(x))' = f'(x) + g'(x)$ (5.31)

链式规则: $(g(f(x)))' = (g \circ f)'(x) = g'(f(x))f'(x)$ (5.32) 这里，$g \circ f$ 表示函数组成 $x \to f(x) \to g(f(x))$。

例子(5.5连锁规则)
让我们[4]用连锁规则来计算函数$h(x)=(2x+1)^4$的导数。随着

$$h(x) = (2x+1)^4 = g(f(x))。 \quad (5.33)$$

$$f(x) = 2x + 1, \quad (5.34)$$

$$g(f) = f^4。 \quad (5.35)$$

我们得到f'和g'的导数为

$$f'(x) = 2, \quad (5.36)$$

$$g'(f) = 4f^3 \quad (5.37)$$

这样，h的导数被赋予为

$$h'(x) = g'(f)f'(x) = (4f^3) - 2 \overset{(5.34)}{=} 4(2x+1)^3 - 2 = 8(2x+1)^3, \quad (5.38)$$

其中我们使用了连锁规则(5.32)，并将f的定义替换为在(5.34)中的$g'(f)$。

5.2 局部微分和梯度

第5.1节中讨论的微分适用于标量变量$x \in \mathbb{R}$的函数f。在下文中，我们考虑函数f取决于一个或多个变量$\boldsymbol{x} \in \mathbb{R}^n$的一般情况，例如，$f(\boldsymbol{x}) = f(x_1, x_2)$。*梯度*是导数对七个变量函数的概括。

我们通过*每次改变一个变量*并保持其他变量不变来找到函数f相对于$\boldsymbol{x}$的梯度。梯度就是这些*偏导数*的集合。

定义（5.5部分导数）。对于一个函数$f : \mathbb{R}^n \to \mathbb{R}$，$\boldsymbol{x} \to f(\boldsymbol{x})$，$\boldsymbol{x} \in \mathbb{R}^n$的$n$个变量$x_1, \ldots, x_n$我们定义*部分导数*为

偏导数

$$\frac{\partial f}{\partial x_1} = \lim_{h \to 0} \frac{f(x_1 + h, x_2, \ldots, x_n) - f(\boldsymbol{x})}{h}$$

$$\vdots \qquad (5.39)$$

$$\frac{\partial f}{\partial x_n} = \lim_{h \to 0} \frac{f(x_1, \ldots, x_{n-1}, x_n + h) - f(\boldsymbol{x})}{h}$$

and collect them in the row vector

$$\nabla_{\boldsymbol{x}} f = \mathrm{grad} f = \frac{df}{d\boldsymbol{x}} = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} & \frac{\partial f}{(\boldsymbol{x})} & \cdots & \frac{\partial f(\boldsymbol{x})}{\partial x_n} \\ & \partial x_2 & & \end{bmatrix} \in \mathbb{R}^{1 \times n}, \quad (5.40)。$$

其中n是变量的数量，1是f的图像/范围/域的维度。在这里，我们定义了列向量$\boldsymbol{x} = [x_1, \ldots, x_n]^T$。

梯度

雅各布

$\in \mathbb{R}^n$.中的行向量(5.40)中的行向量被称为f的*梯度*或*雅各布系数*和是第1节中导数的一般化。5.1.

*备注。*这个雅各布的定义是矢量值函数的雅各布的一般定义的一个特例，它是矢量值函数的集合偏导数。我们将在第二节中再讨论这个问题。5.3. ◆

我们可以使用标量微分法的结果。每个偏导都是关于一个标量的导数。

例子（5.6使用链式规则的部分派生）。
对于$f(x, y) = (x+2y^3)^2$，我们得到部分导数

$$\frac{\partial f(x, y)}{\partial x} = 2(x+2y^3) \frac{\partial}{\partial x}(x+2y^3) = 2(x+2y^3)。 \quad (5.41)$$

$$\frac{\partial f(x,y)}{\partial y} = 2(x+2y^3)\ \frac{\partial}{\partial y}(x+2y^3) = 12(x+2y^3)y^2。 \tag{5.42}$$

其中，我们使用链式规则(5.32)来计算偏导数。

备注（梯度为行向量）。在文献中，将梯度向量定义为列向量的做法并不罕见，因为按照惯例，向量一般都是列向量。我们将梯度向量定义为行向量的原因有两点。首先，我们可以一致地将梯度概括为矢量值函数$f: \mathbb{R}^n \to \mathbb{R}^m$（那么梯度就变成了一个矩阵）。第二，我们可以立即应用多变量链规则，而不需要注意梯度的维度。我们将在第二节讨论这两点5.3. ◆

例子(梯度)5.7

对于$f(x_1, x_2) = x_1^2 x_2 + x_1 x_2^3 \in \mathbb{R}$导致部分导数（即，导数.....）的变化$f$对$x_1$和$x_2$的主语是

$$\frac{\partial f(x_1, x_2)}{\partial x_1} = 2x_1x_2 + x_2^3 \tag{5.43}$$

$$\frac{\partial f(x_1, x_2)}{\partial x_2} = x_1^2 + 3x_1x_2^2 \tag{5.44}$$

而梯度则是

$$\frac{df}{dx} = \begin{bmatrix}\frac{\partial f(x_1, x_2)}{\partial x_1} & \frac{\partial f(x_1, x_2)}{\partial x_2}\end{bmatrix} = \begin{bmatrix}2x_1x_2 + x_2^3 & x_1^2 + 3x_1x_2^2\end{bmatrix} \in \mathbb{R}^{1\times 2}. \tag{5.45}$$

5.2.1 局部微分的基本规则

在多变量的情况下，在$x \in \mathbb{R}^n$的情况下，我们从学校知道的基本微分规则（例如，求和规则、乘积规则、连锁规则；也见第5.1.2节）仍然适用。然而，当我们计算导数的时候，有了重新对向量$x \in \mathbb{R}^n$的光谱，我们需要注意。我们的梯度现在涉及到向量和矩阵，而矩阵乘法不是换位的（第2.2.1节），也就是说，顺序很重要。

产品规则。$(fg)' = fg' + fg'$，总和规则。$(f+g)' = f' + g'$，连锁规则。$(g(f))' = g'(f)f'$

这里有一般的乘积法则、和法则和链法则。

Product rule: $\frac{\partial}{\partial x} f(x)g(x) = \frac{\partial f}{\partial x} g(x) + f(x)\frac{\partial g}{\partial x}$ (5.46)

总和规则：$f(x) + g(x) = \frac{\partial f}{} + \frac{\partial g}{}$

$$\frac{\partial \boldsymbol{x}}{\boldsymbol{x}} \quad \partial \tag{5.47}$$

链式规则：
$$\frac{\partial}{\partial \boldsymbol{x}}(g \circ f)(\boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{x}} g(f(\boldsymbol{x})) = \frac{\partial g}{\partial f}\frac{\partial f}{\partial \boldsymbol{x}} \tag{5.48}$$

这只是一种直觉，但在数学上并不正确，因为部分导数不是分数。

让我们仔细看一下连锁规则。链式规则(5.48)在某种程度上类似于矩阵乘法的规则，我们说过相邻的维度必须匹配，才能实现矩阵乘法；见第2.2.1节。如果我们从左到右，连锁规则也表现出类似的特性。∂f显示在第一个因子的 "分母 "和第二个因子的 "分子 "中。如果我们把这些因子相乘，乘法就被定义了，也就是说，∂f的维度是匹配的，而且∂f "取消 "了，这样，$\partial g/\partial \boldsymbol{x}$就保留了。

5.2.2 链条规则

考虑一个函数$f : \mathbb{R}^2 \to \mathbb{R}$的两个变量$x_1$，$x_2$。此外，$x_1(t)$和$x_2(t)$本身就是$t$的函数。为了计算$f$相对于$t$*的梯度*，我们需要应用链式规则(5.48)来处理多变量函数，即

$$\frac{df}{dt} = \frac{\partial f}{\partial x_1}\frac{\partial f}{\partial x_2}\begin{bmatrix}\frac{\partial x_1(t)}{\partial t}\\ \frac{\partial x_2(t)}{\partial t}\end{bmatrix} = \frac{\partial f}{\partial x}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x}\frac{\partial x_2}{\partial t}, \tag{5.49}$$

其中d表示梯度，∂为偏导数。

例子 5.8

考虑$f(x_1, x_2) = x_1^2 + 2x_2$，其中$x_1 = \sin t$和$x_2 = \cos t$，那么

$$\frac{df}{dt} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t} \tag{5.50a}$$

$$= 2\sin t \frac{\partial \sin t}{\partial t} + 2\frac{\partial \cos t}{\partial t} \tag{5.50b}$$

$$= 2\sin t \cos t - 2\sin t = 2\sin t(\cos t - 1) \tag{5.50c}$$

是f相对于t的相应导数。

如果$f(x_1, x_2)$是x_1和x_2的函数，其中$x_1(s, t)$和$x_2(s, t)$本身是两个变量s和t的函数，连锁法则得到偏导数

$$\frac{\partial f}{\partial s} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial s} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial s}, \tag{5.51}$$

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t}, \tag{5.52}$$

并通过矩阵乘法得到梯度

$$\frac{\mathrm{d}f}{\mathrm{d}(s,t)} = \frac{\partial f}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{x}}{\partial (s,t)} = \underbrace{\frac{\partial f}{\partial x_1}}_{=\frac{\partial f}{\partial \boldsymbol{x}}} \underbrace{\begin{bmatrix} \frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial s} & \frac{\partial x_2}{\partial t} \end{bmatrix}}_{=\frac{\partial \boldsymbol{x}}{\partial (s,t)}} . \tag{5.53}$$

这种将链式规则写成矩阵乘法的紧凑方式，只需将

链式规则写成矩阵乘法即可。

如果梯度被定义为行向量，那么这个方法就有意义。否则，我们将需要开始转置梯度，使之与矩阵维度相匹配。只要梯度是一个向量或一个矩阵，这可能仍然是直接的；然而，当梯度变成一个张量时（我们将在下面讨论这个问题），转置就不再是一个小事。

可以写成一个矩阵乘法。

备注（验证梯度实现的正确性）。将偏导数定义为相应的差值商的极限（见(5.39)），可以在数值检查时加以利用

计算机程序中梯度的正确性。当我们计算梯度检查时

gradients and implement them, we can use finite differences to numerically test our computation and implementation: We choose the value h to be small (e.g., $h = 10^{-4}$) and compare the finite-difference approximation from (5.39) with our (analytic) implementation of the gradient. If the error is small, our gradient implementation is probably correct. "Small" 可能意味着 $\frac{\sum_i (dh_i - df_i)^2}{\sum_i (dh_i + df_i)^2} < 10^{-6}$，其中 dh_i 是有限差分的近似值，df_i 是 f *相对*于第1个变量 x 的分析梯度 $_i$。 ◆

5.3 矢量值函数的梯度

至此，我们讨论了函数 f 的偏导和梯度。$\mathrm{R}^n \to \mathrm{R}$ 对实数的映射。在下文中，我们将把梯度的概念推广到矢量值函数（矢量场）$\boldsymbol{f} : \mathrm{R}^n \to \mathrm{R}^m$，其中 n 和 $m > 1$。

对于一个函数 $\boldsymbol{f} : \mathrm{R}^n \to \mathrm{R}^m$ 和一个向量 $\boldsymbol{x} = [x_1, \ldots, x_n]^\top \in \mathrm{R}^n$，相应的函数值的向量给定为

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_m(\boldsymbol{x}) \end{bmatrix} \in \mathrm{R}^m . \tag{5.54}$$

以这种方式书写矢量值函数，我们就可以把矢量值函数 $\boldsymbol{f} : \mathrm{R}^n \to \mathrm{R}^m$ 看作是函数 $[f_1, \ldots, f_m]^\top$ 的一个矢量，f_i 的每一个 f_i 微分规则正是我们在第1节

中讨论的那些规则。5.2.

因此，一个矢量值函数$\boldsymbol{f}$的偏导。$\mathrm{R}^n \to$ 关于$x_i \in \mathrm{R}^m$，$i=1$，.....n，给定为矢量

$$\frac{\partial \boldsymbol{f}}{\partial x_i} = \begin{bmatrix} \frac{\partial f_1}{\partial x_i} \\ \vdots \\ \frac{\partial f_m}{\partial x_i} \end{bmatrix} \text{ 肢体 } \lim_{h\to 0} \begin{bmatrix} \frac{f_1(x_1,\ldots,x_{i-1},x_i+h,x_{i+1},\ldots x_n)-f_1(\boldsymbol{x})}{h} \\ \vdots \\ \frac{f_m(x_1,\ldots,x_{i-1},x_i+h,x_{i+1},\ldots x_n)-f_m(\boldsymbol{x})}{h} \end{bmatrix} \in \mathrm{R}^m. \tag{5.55}$$

从(5.40)，我们知道$\boldsymbol{f}$相对于一个向量的梯度是偏导数的行向量。在(5.55)中，每个偏导数$\partial \boldsymbol{f}/\partial_i x$本身是一个列向量。因此，我们得到$\boldsymbol{f}$的梯度。

$\mathrm{R}^n \to \mathrm{R}^m$ with respect to $\boldsymbol{x} \in \mathrm{R}^n$ by collecting these partial derivatives:

$$\frac{d\boldsymbol{f}(\boldsymbol{x})}{d\boldsymbol{x}} = \begin{bmatrix} \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \tag{5.56a}$$

$$= \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \in \mathrm{R}^{m\times n}. \tag{5.56b}$$

定义5.6（Jacobian）。一个矢量值函数$\boldsymbol{f}$的所有一阶偏导的集合。$\mathrm{R}^n \to \mathrm{R}^m$的所有一阶偏导的集合被称为*雅各布式*。雅各邦$\boldsymbol{J}$是一个$m\times n$的矩阵，我们定义和排列如下。

雅各布式

一个函数的梯度是一个矩阵的大小$m\times n$。

$$\boldsymbol{J} = \nabla_{\boldsymbol{x}} \boldsymbol{f} = \frac{d\boldsymbol{f}}{d\boldsymbol{x}} = \begin{bmatrix} \frac{\partial \boldsymbol{f}}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}}{\partial x_n} \end{bmatrix} \tag{5.57}$$

$$= \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}, \tag{5.58}$$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad J(i,j) = \frac{\partial f_i}{\partial x_j}. \tag{5.59}$$

作为()的一个特例。5.58)的特例，一个函数$\boldsymbol{f}:\mathrm{R}^n \to \mathrm{R}^1$，它将a

x_i），拥有矢量激增各布
向量$\boldsymbol{x}\in\mathbb{R}^n$到一个标量（例如，$f(\boldsymbol{x})=\sum_{i=1}^{n}$ 系数
是一个行向量（维数×1n的矩阵）；见(5.40).[1]

分子布局

备注。在本书中，我们使用导数的*分子布局*，即f$\mathbb{R}^m$相对于x$\mathbb{R}^n$导数
$\frac{df}{dx}$是一个m ∈∈
n矩阵，其中f的元素定义为行，f的元素定义为 ×
x定义了相应的雅各布式的列；见(5.58).有

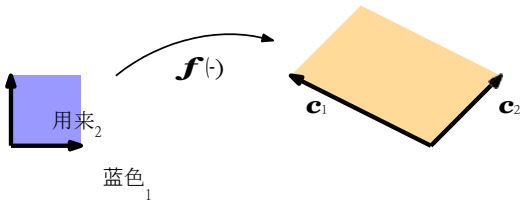


图 坎的雅各布系数的5.5行列式 f 计算放大镜之间和橙色区域。

也存在*分母布局*，它是数字布局。在本书中，我们将使用分母布局。 ◆

与分母布局的转置。

我们将在第6.7节中看到雅各布系数是如何用于概率分布的变量变化方法的。变量变换引起的缩放量是由行列式提供的。

在第4.1,我们看到行列式可以用来计算平行四边形的面积。如果我们给出两个向量$\boldsymbol{b}_1=[1,0]^\top, \boldsymbol{b}_2=[0,1]^\top$作为单位正方形的边（蓝色；见图5.5），这个正方形的面积是

$$\det\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} = 1. \tag{5.60}$$

如果我们取一个边长为$\boldsymbol{c}_1=[2,1]^\top$，$\boldsymbol{c}_2=[1,1]^\top$的平行四边形（图中的橙色5.5），它的面积就会被认为是减数的绝对值(见第4.1节)

$$\left|\det\begin{bmatrix}2 & 1\\ 1 & 1\end{bmatrix}\right| = 3, \tag{5.61}$$

即，其面积正好是单位正方形面积的三倍。我们可以通过找到一个将单位正方形转化为另一个正方形的映射来找到这个比例因子。在线性代数术语中，我们有效地进行了从（$\boldsymbol{b}_1$，$\boldsymbol{b}_2$）到（$\boldsymbol{c}_1$，$\boldsymbol{c}_2$）的变量转换。在我们的例子中，这个映射是线性的，这个映射的行列式的绝对值正好给了我们要找的比例系数。

我们将描述两种方法来识别这种映射。首先，我们假设这个映射是线性的，这样我们就可以使用第二章的工具来识别这个映射。其次，我们将利用本章讨论的工具，用偏导数来找到这个映射关系。

方法 1 为了开始使用线性代数方法，我们把$\boldsymbol{b}_1$，$\boldsymbol{b}_2$和$\boldsymbol{c}_1$，$\boldsymbol{c}$都确定为$\mathbb{R}^2$的基（见第2.6.1节的回顾）。我们所做的实际上是将基数从（$\boldsymbol{b}_1$，$\boldsymbol{b}_2$）改为（$\boldsymbol{c}_1$，$\boldsymbol{c}_2$），我们正在寻找实现基数变化的变换矩阵。利用第2.7.2节的结果，我们确定所需的基础变化矩阵为

$$\boldsymbol{J} = \begin{bmatrix}-2 & 1\\ 5 & -1\end{bmatrix}. \tag{5.62}$$

这样，$\boldsymbol{J}\boldsymbol{b}_1=\boldsymbol{c}_1$和$\boldsymbol{J}\boldsymbol{b}_2=\boldsymbol{c}_2$。

$\boldsymbol{J}$的主量，也就是我们要找的缩放系数，给定为$\det(\boldsymbol{J})=3$，即$(\boldsymbol{c}_1, \boldsymbol{c}_2)$所跨越的正方形面积是$(\boldsymbol{b}_1, \boldsymbol{b}_2)$所跨越面积的三倍。

方法2 线性代数的方法适用于线性转换。对于非线性变换（在第6.7节中成为相关内容），我们采用更一般的方法，使用偏导数。

对于这种方法，我们考虑一个函数$\boldsymbol{f}: \mathbb{R}^2 \to \mathbb{R}^2$执行变量转换。在我们的例子中，$\boldsymbol{f}$将任何矢量$\boldsymbol{x} \in \mathbb{R}^2$相对于$(\boldsymbol{b}_1, \boldsymbol{b}_2)$的坐标表示映射到相对于$(\boldsymbol{c}_1, \boldsymbol{c}_2)$的坐标表示$\boldsymbol{y} \in \mathbb{R}^2$。我们想确定这种映射，这样我们就能计算出当面积（或体积）被$\boldsymbol{f}$转换时，它是如何变化的。为此，我们需要找出如果我们稍微修改一下$\boldsymbol{x}$，$\boldsymbol{f}(\boldsymbol{x})$是如何变化的。这个问题的答案正是由雅各布矩阵$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{x}} \in \mathbb{R}^{2\times 2}$.因为我们可以写出

$$y_1 = -2x_1 + x_2 \tag{5.63}$$
$$y_2 = x_1 + x_2 \tag{5.64}$$

我们得到$\boldsymbol{x}$和$\boldsymbol{y}$之间的函数关系，这使得我们可以得到部分导数

$$\frac{\partial y_1}{\partial x_1} = -2, \quad \frac{\partial y_1}{\partial x_2} = 1, \quad \frac{\partial y_2}{\partial x_1} = 1, \quad \frac{\partial y_2}{\partial x_2} = 1 \tag{5.65}$$

并将雅各布系数组成为

$$\boldsymbol{J} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ 1 & 1 \end{bmatrix}. \tag{5.66}$$

雅各布表示我们正在寻找的坐标变换。如果坐标变换是线性的（如我们的例子），它是精确的，并且(5.66)精确地恢复了()中的基础变化矩阵。5.62).如果坐标变换是非线性的，那么雅各布系数就用线性变换来逼近这个非线性变换。*雅各布行列式*的绝对值$\det(\boldsymbol{J})$是坐标转换时面积或体积被缩放的因素。我们的情况是，$\det(\boldsymbol{J}) = 3$.

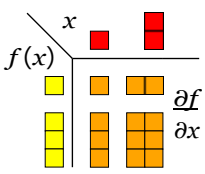
在几何学上，当我们转换一个区域或体积时，雅各布行列式给出了放大/缩放系数。

雅各布决定数

在第6.7节中，当我们转换随机变量和概率能力分布时，雅各布行列式和变量转换将变得相关。在使用*重构技巧*（也称为*无限扰动分析*）训练深度神经网络的背景下，这些变换与机器学习极为相关。

在本章中，我们遇到了函数的导数。图5.6总结了这些导数的尺寸。如果$f: \mathbb{R} \to \mathbb{R}$，梯度就是一个简单的标量（左上角的条目）。对于$f: \mathbb{R}^D \to \mathbb{R}$，梯度是一个$1 \times D$行向量（右上角条目）。对于$\boldsymbol{f}: \mathbb{R} \to \mathbb{R}^E$，梯度是一个$E \times 1$列向量，对于$\boldsymbol{f}: \mathbb{R}^D \to \mathbb{R}^E$的梯度是一个$E \times D$矩阵。

图：（部分）导数的维度**5.6**。



例子（5.9矢量值函数的梯度）

我们被赋予

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}, \quad \boldsymbol{f}(\boldsymbol{x}) \in \mathrm{R}^M, \quad \boldsymbol{A} \in \mathrm{R}^{M\times N}, \quad \boldsymbol{x} \in \mathrm{R}^N.$$

为了计算梯度$d\boldsymbol{f}/d\boldsymbol{x}$，我们首先要确定的是$d\boldsymbol{f}/d\boldsymbol{x}$。因为$\boldsymbol{f}: \mathrm{R}^N \to \mathrm{R}^M$，所获得$d\boldsymbol{f}/d\boldsymbol{x} \in \mathrm{R}^{M\times N}$。第二。为了计算梯度，我们确定$\boldsymbol{f}$对每个$\boldsymbol{x}$的$j$偏导。

$$f_i(\boldsymbol{x}) = \sum_{j=1}^{N} A_{ij}x_j \implies \frac{\partial f_i}{\partial x_j} = A_{ij} \tag{5.67}$$

我们收集Jacobian中的偏导数，得到梯度

$$\frac{d\boldsymbol{f}}{d\boldsymbol{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial f_M}{\partial x_1} & \cdots & \frac{\partial f_M}{\partial x_N} \end{bmatrix} = \begin{bmatrix} A_{11} & \cdots & A_{1N} \\ \vdots & & \vdots \\ A_{M1} & \cdots & A_{MN} \end{bmatrix} = \boldsymbol{A} \in \mathrm{R}^{M\times N}. \tag{5.68}$$

例子(5.10连锁规则)

考虑函数$h: \mathrm{R} \to \mathrm{R}, h(t) = (f \circ g)(t)$，其中

$$f: \mathrm{R}^2 \to \mathrm{R} \tag{5.69}$$

$$g: \mathrm{R} \to \mathrm{R}^2 \tag{5.70}$$

$$f(\boldsymbol{x}) = \exp(x_1 x_2^2) \tag{5.71}$$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = g(t) = \begin{bmatrix} t\cos t \\ t\sin t \end{bmatrix} \tag{5.72}$$

并计算h相对于t的梯度。因为$f: \mathrm{R}^2 \to \mathrm{R}$和$g: \mathrm{R} \to \mathrm{R}^2$，我们注意到

$$\frac{\partial f}{\partial \boldsymbol{x}} \in \mathrm{R}^{1\times 2}, \quad \frac{\partial g}{\partial t} \in \mathrm{R}^{2\times 1}. \tag{5.73}$$

所需的梯度是通过应用连锁规则计算出来的。

$$\frac{dh}{dt} = \frac{\partial f}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial t} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial t} \end{bmatrix} \tag{5.74a}$$

$$= \begin{bmatrix} \exp(x_1 x_2^2)x_2 & 2\exp(x_1 x_2^2)x_1 x_2 \end{bmatrix} \begin{bmatrix} \cos t - t\sin t \\ \sin t + t\cos t \end{bmatrix} \tag{5.74b}$$

$$= \exp(x_1 x_2^2)x_2(\cos t - t\sin t) + 2x_1 x_2(\sin t + t\cos t)。 \tag{5.74c}$$

其中$x_1 = t\cos t$和$x_2 = t\sin t$；见(5.72).

示例（5.11线性模型中最小二乘损失的梯度）。

让我们考虑线性模型

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\theta}, \tag{5.75}$$

其中，$\boldsymbol{\theta} \in \mathbb{R}^D$是参数向量，$\boldsymbol{\Phi} \in \mathbb{R}^{N \times D}$是输入特征和$\boldsymbol{y} \in \mathbb{R}^N$是相应的观测值。我们定义函数

$$L(\boldsymbol{e}) := \|\boldsymbol{e}\|^2, \tag{5.76}$$

$$\boldsymbol{e}(\boldsymbol{\theta}) := \boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}。 \tag{5.77}$$

我们寻求$\frac{\partial L}{\partial \boldsymbol{\theta}}$，为此我们将使用连锁规则。$L$被称为一个*最小二乘法损失*函数。

我们将在第九章线性回归的背景下更详细地讨论这个模型，在这里我们需要最小二乘法损失L关于参数的导数。

$\boldsymbol{\theta}$.

最小二乘法损失

在我们开始计算之前，我们确定梯度的维度为

$$\frac{\partial L}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{1 \times D}. \tag{5.78}$$

链式规则使我们可以计算梯度为

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = \frac{\partial L}{\partial \boldsymbol{e}} \frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}}, \tag{5.79}$$

dLdtheta = np.einsum('n,nd', dLde, dedtheta)

其中第d个元素由以下公式给出

$$\frac{\partial L}{\partial \boldsymbol{\theta}}[1, d] = \sum_{n=1}^{N} \frac{\partial L}{\partial \boldsymbol{e}}[n] \frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}}[n, d]。 \tag{5.80}$$

我们知道，$\|\boldsymbol{e}\|^2 = \boldsymbol{e}\boldsymbol{e}^\top$（见第3.2节），并确定

$$\frac{\partial L}{\partial \boldsymbol{e}} = 2\boldsymbol{e}^\top \in \mathbb{R}^{1 \times N}。 \tag{5.81}$$

此外，我们得到

$$\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}} = -\boldsymbol{\Phi} \in \mathbb{R}^{N \times D}, \tag{5.82}$$

这样，我们期望的导数是

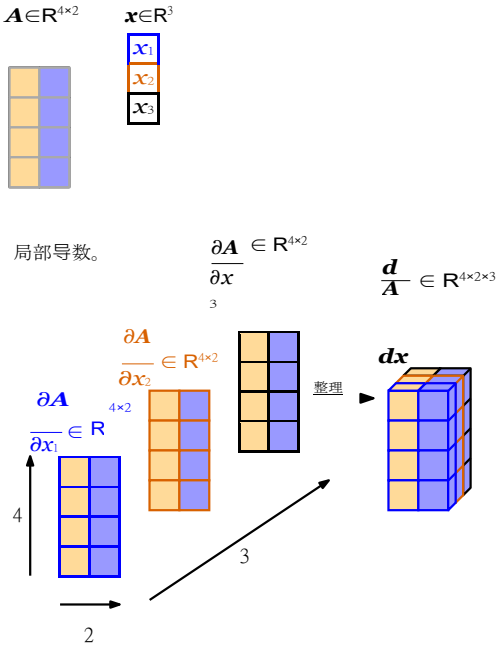
$$\frac{\partial L}{\partial \boldsymbol{\theta}} = 2\boldsymbol{e}^\top(-\boldsymbol{\Phi}) \overset{(5.77)}{=} -2(\boldsymbol{y}^\top - \boldsymbol{\theta}^\top \boldsymbol{\Phi}^\top)\boldsymbol{\Phi} \in \mathbb{R}^{1 \times D}. \tag{5.83}$$

*备注。*如果不使用链式规则，我们也会得到同样的结果，那就是立即看一下函数

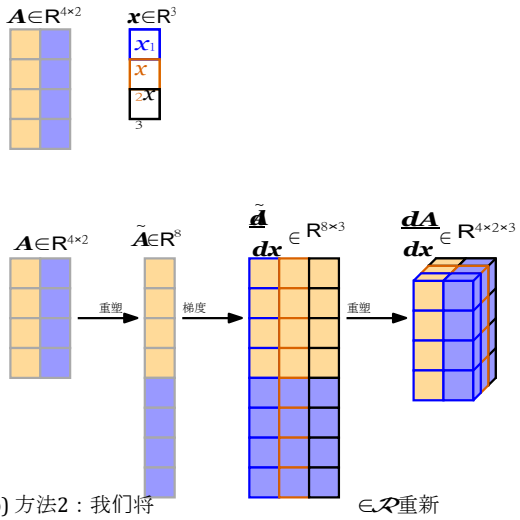
$$L(\boldsymbol{\theta}) := \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\|^2 = (\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta})。 \tag{5.84}$$

这种方法对于像L这样的简单函数仍然实用，但对于深层次的函数组合

就变得不实用了。◆



(a) 方法1：我们计算部分导数 $\frac{\partial A}{\partial x_1}, \frac{\partial A}{\partial x_2}, \frac{\partial A}{\partial x_3}$ each of which is a $\times$ 4matrix2, and col-在一个×4×的23张量中晚期它们。



(b) 方法2：我们将 $A \in \mathbb{R}^{4\times2}$ 重新塑造成（扁平化）的向量。然后，我们计算gAradient $\frac{d\tilde{A}}{dx} \in \mathbb{R}^{8\times3}$. 我们通过重新塑造这个梯度来获得梯度张量，如上图所示。

图：矩阵相对于矢量的梯度计算的可视化**5.7**。我们感兴趣的是计算以下矩阵的梯度 $A \in \mathbb{R}^{4\times2}$ 与相对于一个矢量 $x \in \mathbb{R}^3$。我们知道梯度 $\frac{dA}{dx} \in \mathbb{R}^{4\times2\times3}$ 跟二相当于到达的方法有：（a）整理偏导数变成一个雅各布式张量。(b) 将矩阵扁平化为一个矢量。计算雅各布矩阵，重新塑造为雅各布张量。

5.4 矩阵的梯度

我们可以认为，一个

我们会遇到这样的情况：我们需要获取矩阵相对于向量（或其他矩阵）的梯度，这就会产生一个多维张量。我们可以把这个张量看作是一个多维数组，其中

张量(*tensor*)*微积分* 多维数组。

收集偏导数。例如，如果我们计算一个$m \times n$矩阵$\boldsymbol{A}$相对于$p \times q$矩阵$\boldsymbol{B}$的梯度，得到的Jacobian将是$(m \times n) \times (p \times q)$，即一个四维张量$\boldsymbol{J}$，其条目为$J_{ijkl} = \partial A_{ij} / \partial B_{kl}$。

由于矩阵代表线性映射，我们可以利用这样一个事实：在$\mathbb{R}^{m \times n}$由mn个矩阵组成的空间和由mn个矢量组成$\mathbb{R}^{mn}$的空间之间存在着矢量空间的同构性（线性、可逆映射）。因此，我们可以将我们的矩阵分别重新塑造成长度为mn和pq的向量。使用这些mn向量的梯度会产生一个大小为$mnpq$的雅各布系数。图5.7可视化了这两种方法。在实际应用中，通常希望将矩阵重新塑造成一个向量，然后继续使用这个雅各布矩阵。链式规则(5.48)可以归结为简单的矩阵乘法，而在雅各布张量的情况下，我们需要更加注意我们需要求出哪些维度。

矩阵可以通过叠加矩阵的列（"扁平化"）转化为向量。

例子（5.12矢量相对于矩阵的梯度）

让我们考虑以下例子，其中

$$\boldsymbol{f} = \boldsymbol{A}\boldsymbol{x}, \quad \boldsymbol{f} \in \mathbb{R}^M, \quad \boldsymbol{A} \in \mathbb{R}^{M \times N}, \quad \boldsymbol{x} \in \mathbb{R}^N \tag{5.85}$$

和我们寻求梯度$d\boldsymbol{f}/d\boldsymbol{A}$的地方。让我们再次开始确定梯度的维度为

$$\frac{d\boldsymbol{f}}{d\boldsymbol{A}} \in \mathbb{R}^{M \times (M \times N)} \tag{5.86}$$

根据定义，梯度是偏导数的集合。

$$\frac{d\boldsymbol{f}}{d\boldsymbol{A}} = \begin{bmatrix} \frac{\partial f_1}{\partial \boldsymbol{A}} \\ \vdots \\ \frac{\partial f_M}{\partial \boldsymbol{A}} \end{bmatrix}, \quad \frac{\partial f_i}{\partial \boldsymbol{A}} \in \mathbb{R}^{1 \times (M \times N)} \tag{5.87}$$

为了计算偏导数，明确写出矩阵向量乘法将很有帮助。

$$f_i = \sum_{j=1}^{N} A_{ij} x_j, \quad i = 1, \ldots, M, \tag{5.88}$$

和部分导数，然后给定为

$$\frac{\partial f_i}{\partial A_{iq}} = x_q. \tag{5.89}$$

这使我们能够计算f_i相对于$\boldsymbol{A}$的某一行的偏导数，其结果为

$$\frac{\partial f_i}{\partial A_{i,}} = \boldsymbol{x}^T \in \mathbb{R}^{1 \times N}. \tag{5.90}$$

$$\frac{\partial f_i}{\partial A_{k=i}} = \mathbf{0}^{\mathrm{T}} \in \mathbb{R}^{\times\times 11N} \tag{5.91}$$

其中，我们必须注意正确的维度。由于f_i对应于$\mathbb{R}$，并且$\boldsymbol{A}$的每一行都是大小为$\times 1N$的，我们得到一个$\times 1\times 1N$-----规模的张量作为f_i相对于$\boldsymbol{A}$的某一行的偏导。

我们把偏导数(5.91)，并得到所需的梯度(5.87)中，通过

$$\frac{\partial f_i}{\partial \boldsymbol{A}} = \begin{bmatrix} \mathbf{0}^{\mathrm{T}} \\ \vdots \\ \mathbf{0}^{\mathrm{T}} \\ \mathrm{xT} \\ \mathbf{0}^{\mathrm{T}} \\ \vdots \\ \mathbf{0}^{\mathrm{T}} \end{bmatrix} \in \mathbb{R}^{1\times(M\times N)} \tag{5.92}$$

例子（5.13矩阵相对于矩阵的梯度）

考虑一个矩阵$\boldsymbol{R}\in\mathbb{R}^{M\times N}$和$\boldsymbol{f}:\mathbb{R}^{M\times N}\to\mathbb{R}^{N\times N}$其中

$$\boldsymbol{f}(\boldsymbol{R}) = \boldsymbol{R}\boldsymbol{R}^{\mathrm{T}} = \boldsymbol{K}\in\mathbb{R}^{N\times N}, \tag{5.93}$$

其中我们寻求梯度$\boldsymbol{dK}/\boldsymbol{dR}$。

为了解决这个难题，让我们首先写下我们已经有的东西知道。梯度的尺寸为

$$\frac{\boldsymbol{d}\boldsymbol{K}}{\boldsymbol{d}\boldsymbol{R}} \in \mathbb{R}^{(n\times n)\times(m\times n)} \tag{5.94}$$

这是个张量。此外。

$$\frac{dK_{pq}}{dR} \in \mathbb{R}^{1\times M\times N} \tag{5.95}$$

对于$p, q = 1, \ldots, N$，其中K_{pq}是$\boldsymbol{K} = \boldsymbol{f}(\boldsymbol{R})$的第$(p, q)$个条目。用$\boldsymbol{r}$表示$\boldsymbol{R}$的第i列$_i$，$\boldsymbol{K}$的每个条目都由$\boldsymbol{R}$的两列点乘得到，即。

$$K_{pq} = \boldsymbol{r}_p\boldsymbol{r}_q = \sum_{m=1}^{M} R_{mp}R_{mq} . \tag{5.96}$$

当我们现在计算部分导数$\frac{\partial K_{pq}}{\partial R_{ij}}$时，我们得到

$$\frac{\partial K_{pq}}{\partial R_{ij}} = \sum_{m=1}^{M} \partial_{Rij} R_{mp} R_{pqij} = \partial_{mq} , \tag{5.97}$$

$$\partial_{pqij} = \begin{cases} \text{Ri}\,\text{if}\, j = p, p \neq q \\ \text{裂缝} \quad \text{如果} j = q, p \\ 2\text{Ri}\,\text{if}\, j = p, p = q \\ 0 \quad \text{否则} \end{cases} \neq q \quad . \tag{5.98}$$

从(5.94)，我们知道所需梯度的维度为$(N \times)$。
N）×（$M \times N$），这个张量的每一个条目都是由∂给出的$_{pqij}$
在(5.98)，其中p、q、j =1 , ..., N和$i = 1, \ldots, M$.

5.5 计算梯度的有用特征

在下文中，我们将列出一些在机器学习中经常需要的有用梯度（Petersen and Pedersen,2012）。在这里，我们用tr( )作为迹线（见定义4.4），det( )为行列式(见第4.1节)，$\boldsymbol{f}(\boldsymbol{X})^{-1}$为$\boldsymbol{f}(\boldsymbol{X})$的逆，假设它存在的话。

$$\frac{\partial}{\partial \boldsymbol{X}} \boldsymbol{f}(\boldsymbol{X})^{\mathrm{T}} = \frac{\partial \boldsymbol{f}(\boldsymbol{X})}{\partial \boldsymbol{X}}^{\mathrm{T}} \tag{5.99}$$

$$\frac{\partial}{\partial \boldsymbol{X}} \operatorname{tr}(\boldsymbol{f}(\boldsymbol{X})) = \operatorname{tr} \frac{\partial \boldsymbol{f}(\boldsymbol{X})}{\partial \boldsymbol{X}} \tag{5.100}$$

$$\frac{\partial}{\partial \boldsymbol{X}} \det(\boldsymbol{f}(\boldsymbol{X})) = \det(\boldsymbol{f}(\boldsymbol{X})) \operatorname{tr} \boldsymbol{f}(\boldsymbol{X})^{-1} \frac{\partial \boldsymbol{f}(\boldsymbol{X})}{\partial \boldsymbol{X}} \tag{5.101}$$

$$\frac{\partial}{\partial \boldsymbol{X}} \boldsymbol{f}(\boldsymbol{X})^{-1} = -\boldsymbol{f}(\boldsymbol{X})^{-1} \frac{\partial \boldsymbol{f}(\boldsymbol{X})}{\partial \boldsymbol{X}} \boldsymbol{f}(\boldsymbol{X})^{-1} \tag{5.102}$$

$$\frac{\partial \boldsymbol{a}\boldsymbol{X}\boldsymbol{b}^{\mathrm{T}-1}}{\partial \boldsymbol{X}} = -(\boldsymbol{X}^{-1})^{\mathrm{T}} \boldsymbol{a}\boldsymbol{b}^{\mathrm{T}} (\boldsymbol{X}^{-1})^{\mathrm{T}} \tag{5.103}$$

$$\frac{\partial \boldsymbol{x}\boldsymbol{T}\boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}^{\mathrm{T}} \tag{5.104}$$

$$\frac{\partial \boldsymbol{a}\boldsymbol{T}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}^{\mathrm{T}} \tag{5.105}$$

$$\frac{\partial \boldsymbol{a}\boldsymbol{T}\boldsymbol{X}\boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{a}\boldsymbol{b}^{\mathrm{T}} \tag{5.106}$$

$$\frac{\partial \boldsymbol{x}\boldsymbol{T}\boldsymbol{B}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{B} + \boldsymbol{B})^{\mathrm{T}} \tag{5.107}$$

$$\frac{\partial}{\partial \boldsymbol{s}} (\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^{\mathrm{T}} \boldsymbol{W} (\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s}) = -2 (\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^{\mathrm{T}} \boldsymbol{W}\boldsymbol{A} \quad \text{为对称的} \boldsymbol{W} \tag{5.108}$$

备注。在本书中，我们只涉及矩阵的迹线和转置。然而，我们已经看到，导数可以是更高维的张量，在这种情况下，通常的跟踪和转置都没有定义。在这种情况下，一个$D\ D\ E\ F$张量的迹线将是一个$E\ F$维的矩阵。这是张量收缩的一个特殊情况。$\times\times\times$ 同样地，当我们

"转置 "一个张量，我们指的是交换前两个维度。具体来说，在(5.99)到(5.102)中，当我们处理多变量函数$\boldsymbol{f}(\,)$并计算相对于矩阵的导数时，我们需要与张量相关的计算（并且选择不对它们进行矢量化，如第5.4). ◆

5.6 反向传播和自动差异化

关于反向传播和链式规则的良好讨论，可以在Tim Vieira的博客中找到，网址是https://tinyurl.com/ycfm2yrw。

在许多机器学习应用中，我们通过执行梯度下降来找到好的模型参数（第7.1节），这依赖于我们可以计算出学习目标相对于模型参数的梯度这一事实。对于一个给定的目标函数，我们可以通过微积分和应用链式规则来获得相对于模型参数的梯度；见第5.2.2.7节。5.3我们已经尝到了甜头，当时我们研究了一个平方损失相对于线性回归模型参数的梯度。

Consider the function

$$f(x) = \sqrt{x^2 + \exp(x^2)} + \cos\sqrt{x^2 + \exp(x^2)}\,。 \tag{5.109}$$

通过应用连锁规则，并注意到微分是线性的，我们可以计算出梯度

$$\begin{aligned}\frac{df}{dx} &= \frac{2x + 2x\exp(x^2)}{2\sqrt{x^2 + \exp(x^2)}} - \sin\sqrt{x^2 + \exp(x^2)}\,\frac{2x + 2x\exp(x^2)}{2\sqrt{x^2 + \exp(x^2)}} \\ &= 2x\,\frac{1}{2\sqrt{x^2 + \exp(x^2)}} - \sin\sqrt{x^2 + \exp(x^2)}\,\frac{1 + \exp(x^2)}{\sqrt{x^2 + \exp(x^2)}}\,。\end{aligned} \tag{5.110}$$

以这种明确的方式写出梯度往往是不切实际的，因为它往往导致导数的表达式非常冗长。在实践中，这意味着，如果我们不小心，梯度的实现可能比计算函数要昂贵得多，这就造成了不必要的开销。对于训练深度神经网络模

其他，*反向传播*算法（Kelley，1960；Bryson，1961；Dreyfus，1962;Rumelhart等人,1986)是一种有效的方法来计算关于模型参数的误差函数的梯度。

反向传播）。

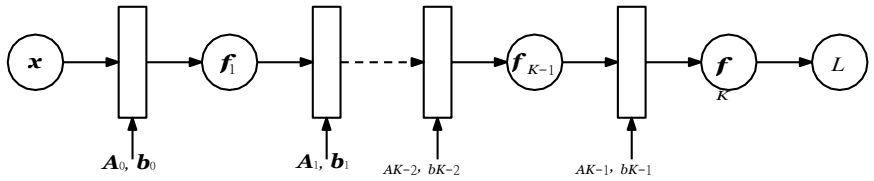
5.6.1 深度网络中的梯度

深度学习是一个将链式规则运用到极致的领域，其中函数值$\boldsymbol{y}$被计算为一个多级函数组成

$$\boldsymbol{y} = (\boldsymbol{f}_K \circ \boldsymbol{f}_{K-1} \circ \cdots \circ \boldsymbol{f}_1)(\boldsymbol{x}) = \boldsymbol{f}_K(\boldsymbol{f}_{K-1}(\cdots(\boldsymbol{f}_1(\boldsymbol{x}))\cdots)), \tag{5.111}$$

其中，$\boldsymbol{x}$是输入（如图像），$\boldsymbol{y}$是观测值（如类标签），每个函数$\boldsymbol{f}_i$，i

$=1, \ldots, K$，拥有它自己的参数。



图为多层神经网络中的前向**5.8**传递，计算损失L的函数。输入和旁门左道A_i，B_i。

我们讨论的情况是，每一层的激活函数都是相同的，以消除混乱的符号。

在具有多层的神经网络中，我们在第i层有函数$f(_i x_{i-1}) = \sigma(_{i-1}Ax$ $_{i-1}+ b_{i-1})$。这里x_{i-1}是第i 1层的输出，σ是一个激活函数，如Logistic sigmoid，$\frac{1}{1+e}$tanh或整流线性单元（ReLU）。为了训练这些模型，我们需要一个损失函数L的梯度，相对于所有的模型参数A_j，b_j来说，j = 1, . .这也要求我们计算L相对于每层输入的梯度。例如，如果我们有输入x和观测值y，网络结构定义为

$$f_0 = x \tag{5.112}$$

$$f_i = \sigma_i(_{i-1}Af_{i1-} + b_{i-1}) 。 \qquad i = 1, \ldots, K, \tag{5.113}$$

也见图5.8的可视化，我们可能有兴趣找到

A_j, b_j为j =0, . . . ， K -1 ， 这样的平方损失

$$L(\theta) = y - f_K(\theta, x)|^2 \tag{5.114}$$

最小化，其中$\theta = A_0, b_0, \ldots, A_{K-1}, b_{K-1}$。

为了获得相对于参数集θ的梯度，我们需要L相对于每个层j=0, ...的参数θ_j=A_j, b_j的偏导数。链式规则允许我们确定部分导数为 {}

更深入的讨论是关于梯度的神经网络可以是在Justin Domke的讲义中可以找到https://tinyurl。com/yalcxgtv 。

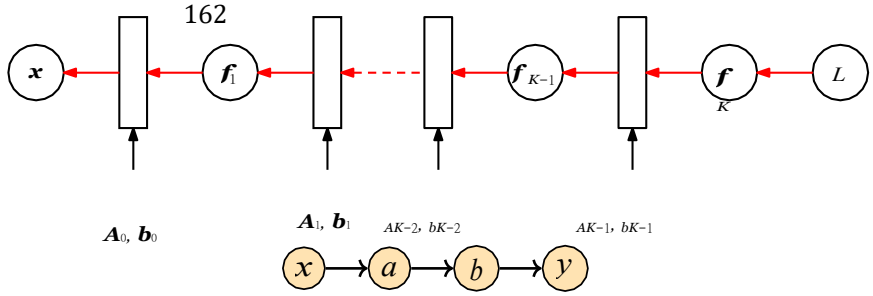
$$\frac{\partial L}{\partial \theta_{K-1}} = \frac{\partial L}{\partial f_K}\frac{\partial f_K}{\partial \theta_{K-1}} \tag{5.115}$$

$$\frac{\partial L}{\partial \theta_{K-2}} = \frac{\partial L}{\partial f_K}\boxed{\frac{\partial f_K}{\partial f_{K-1}}\frac{\partial f_{K-1}}{\partial \theta_{K-2}}} \tag{5.116}$$

$$\frac{\partial L}{\partial \theta_{K-3}} = \frac{\partial L}{\partial f_K}\frac{\partial f_K}{\partial f_{K-1}}\boxed{\frac{\partial f_{K-1}}{\partial f_{K-2}}\frac{\partial f_{K-2}}{\partial \theta_{K-3}}} \tag{5.117}$$

$$\frac{\partial L}{\partial \theta_i} = \frac{\partial L}{\partial f_K}\frac{\partial f_K}{\partial f_{K-1}}\cdots\boxed{\frac{\partial f_{i+2}}{\partial f_{i+1}}\frac{\partial f_{i+1}}{\partial \theta_i}} \tag{5.118}$$

橙色项是一个层的输出相对于其输入的偏导，而蓝色项是一个层的输出相对于其参数的偏导。假设我们已经计算了部分导数$\partial L/\partial \theta_{i+1}$， 那么大部分的计算就可以重新用于计算$\partial L/_i\partial \theta$。额外的条款，我们



图为多层神经网络中计算损失函数梯度的后5.9向传递。

图 表明数据从x到y的流动经过的简单5.10图表一些中间变量a，b。

框中表示的是需要计算的内容。图5.9可视化地表明，梯度通过网络向后传递。

5.6.2 自动差异化

事实证明，反向传播是一般技术的一个特例。

在数值分析中称为*自动微分*。我们可以认为aut-

automatic 差异化

梯度差分法是一套技术，通过处理中间变量和应用链式方法，以数值方式（与符号方式相反）评估一个函数的精确梯度（达到机器精度）。

。

规则。自动微分法应用了一系列的基本算术

微分与符号微分和梯度的数值近似不同，例如使用有限差分。

诸如加法和乘法等运算，以及诸如sin、cos、exp、log等基本函数。通过对这些运算应用连锁法则，可以自动计算出相当复杂的函数的梯度。自动微分适用于一般的计算机程序，有正向和反向模式。Baydin等人（2018）对机器学习中的自动微分做了一个很好的概述。

图中显示了一个简单的图形，表示从输入x到输出y的数据流通过一些中间变量a，b。5.10图中显示了一个简单的图形，表示从输入x到输出y的数据流通过一些中间变量a，b。如果我们要计算导数dy/dx，我们将应用连锁规则，得到

$$\frac{dy}{dx} = \frac{dy}{db}\frac{db}{da}\frac{da}{dx}. \tag{5.119}$$

直观地说，正向模式和反向模式的顺序不同，在一般情况下，正向模式和反向模式的乘数　　不同。

矩阵乘法。由于矩阵乘法的关联性，我们可以在以下方面进行选择

我们使用雅各布，它可以是向量、矩阵或张量。

$$\frac{dy}{dx} = \frac{dy}{db}\frac{db}{da}\frac{da}{dx}, \tag{5.120}$$

$$\frac{dy}{dx} = \frac{dy}{db}\left(\frac{db}{da}\frac{da}{dx}\right)。 \tag{5.121}$$

方程(5.120)将是*逆向模式*，因为梯度是道具

逆向模式

通过图形存储自动差异化

通过图形存储自动差异化与数据流相反。等式梯度 (5.121)将是*正向模式*，其中梯度以通过图表从左到右显示数据。

正向模式流动。

在下文中，我们将重点讨论反向模式的自动区分，也就是反向传播。在神经网络的背景下，输入维度往往比标签的维度高得多，反向模式在计算上要比正向模式便宜得多。让我们从一个有启发性的例子开始。

例子 5.14

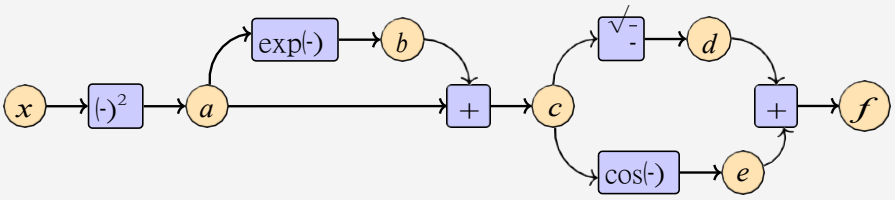
Consider the function

$$f(x) = \sqrt{x^2 + \exp(x^2)} + \cos\sqrt{x^2 + \exp(x^2)} \quad (5.122)$$

从(5.109).如果我们要在计算机上实现一个函数f，我们将能够通过使用*中间变量*来节省一些计算。

中间变量

$$a = x^2, \quad (5.123)$$
$$b = \exp(a), \quad (5.124)$$
$$c = a + b, \quad (5.125)$$
$$d = \sqrt{c}。 \quad (5.126)$$
$$e = \cos(c), \quad (5.127)$$
$$f = d + e。 \quad (5.128)$$



图计算**5.11**图，输入x，函数值f，中间变量a、b、c、d、e。

这与应用连锁规则时的思维过程是一样的。请注意，前面的方程组所需的操作比直接实现函数$f(x)$所定义的(5.109).图中相应的计算图5.11显示了获得函数值f所需的数据流和计算。

包括中间变量的方程组可以被认为是一个计算图，这种表示方法在神经网络软件库的实施中被广泛使用。通过回顾基本函数导数的定义，我们可以直接计算中间变量相对于其相关输入的导数。我们得到以下结果。

$$\frac{\partial a}{\partial x} = 2x \quad (5.129)$$
$$\frac{\partial b}{\partial a} = \exp(a) \quad (5.130)$$

$$\frac{\partial c}{\partial a} = =1 \quad \frac{\partial c}{\partial b} \tag{5.131}$$

$$\frac{\partial d}{\partial c} = \frac{1}{\sqrt{2\,c}} = \tag{5.132}$$

$$\frac{\partial e}{\partial c} = -\sin(c) \tag{5.133}$$

$$\frac{\partial f}{\partial d} = =1 . \quad \frac{\partial f}{\partial e} \tag{5.134}$$

通过查看图中的计算图5.11,我们可以计算出

$\partial f/\partial x$，从输出端向后推导，得到

$$\frac{\partial f}{\partial c} = \frac{\partial f}{\partial d}\frac{\partial}{\partial e} + \frac{d}{\partial e}\frac{}{\partial c} \tag{5.135}$$

$$\frac{\partial f}{\partial b} = \frac{\partial f}{\partial c}\frac{\partial c}{\partial b} \tag{5.136}$$

$$\frac{\partial f}{} = \frac{\partial f}{} + \frac{b}{\partial c}\frac{}{\partial a b} \tag{5.137}$$

$$\frac{\partial a c}{} = \frac{\partial \partial f}{\partial x}\frac{\partial a}{} . \tag{5.138}$$

$$\partial a \quad \partial x$$

请注意，我们隐含地应用了链式规则来获得$\partial f/\partial x$。通过将基本函数的导数结果进行替换，我们可以得到

$$\frac{\partial f}{\partial c} = 1 - \frac{1}{\sqrt{2\,c}} + =1 - (-\sin(c)) \tag{5.139}$$

$$\frac{\partial f}{\partial b} = - \frac{\partial f}{\partial c} \tag{5.140}$$

$$\frac{\partial f}{\partial a} \quad \frac{\partial f}{\partial b} = \exp(a) + \frac{\partial f}{\partial c} - 1 \tag{5.141}$$

$$\frac{\partial f}{\partial a} = \frac{\partial}{\partial x} - 2x . \tag{5.142}$$

通过将上述每个导数视为一个变量，我们观察到计算导数所需的计算与函数本身的计算具有相似的复杂性。这是很反直觉的，因为导数的数学表达式(5.110)
的数学表达要复杂得多。
函数$f(x)$在(5.109).

自动微分是例5.14的形式化。让$x_1, \ldots, x_d$是函数的输入变量，$x_{d+1}, \ldots, x_{D-1}$是中间变量，x_D是输出变量。那么计算图可以表示为：。

$$\text{对于} i = d+1, \ldots, D: \; x_i = g_i(x_{\text{Pa}(x_i)}) \text{。} \tag{5.143}$$

其中$g_i(\ )$是基本函数，$_{x_i \in \mathrm{Pa}(x)}$是图中变量$x$的$i$父节点。给定一个以这种方式定义的函数，我们可以使用链式规则来一步一步地计算该函数的导数。回顾一下，根据定义$f=x_D$，因此

$$\frac{\partial f}{\partial x_D} = 1 \tag{5.144}$$

对于其他变量x_i，我们应用连锁规则

$$\frac{\partial f}{\partial x_i} = \sum_{x_j : x_i \in \mathrm{Pa}(x_j)} \frac{\partial f}{\partial x_j} \frac{\partial x_j}{\partial x_i} = \sum_{x_j : x_i \in \mathrm{Pa}(x_j)} \frac{\partial f}{\partial x_j} \frac{\partial g_j}{\partial x_i} \tag{5.145}$$

其中$\mathrm{Pa}(x_j)$是计算j图中x的父节点集合。方程(5.143)是一个函数的正向传播，而(5.145)是梯度在计算图中的反向传播。对于神经网络的训练，我们反向传播预测与标签有关的误差。

反向模式下的自动分化需要一个解析树。

只要我们有一个可以表达为计算图的函数，其中的元素函数是可微的，上述的自动微分方法就能发挥作用。事实上，该函数甚至可能不是一个数学函数，而是一个计算机程序。然而，并不是所有的计算机程序都可以自动进行微分，例如，如果我们不能找到微分的基本函数。编程结构，如for
循环和if语句，也需要更多的关注。

5.7 高阶衍生品

到目前为止，我们已经讨论了梯度，也就是一阶导数。有时，我们对高阶导数感兴趣，例如，当我们想使用牛顿方法进行优化时，需要二阶导数（Nocedal and Wright,2006）。在第5.1.1节中，我们讨论了用多项式来近似函数的泰勒级数。在多变量的情况下，我们也可以做同样的事情。在下文中，我们将完全这样做。但让我们从一些符号开始。

考虑一个函数f:我们对高阶偏导（和*梯度*）使用以下符号。

- $\frac{\partial^2 f}{\partial x^2}$是$f$对$x$的第二次偏导。
- $\frac{\partial^n f}{\partial x^n}$是$f$对$x$的第$n$次偏导。
- $\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)$是由第一部分差分得到的部分导数。先对$x$进行计算，然后再对$y$进行计算。
- $\frac{\partial^2 f}{\partial x \partial y}$ 是通过第一次偏微分得到的偏导，即y，然后是x。

海西

*Hessian*是所有二阶偏导数的集合。

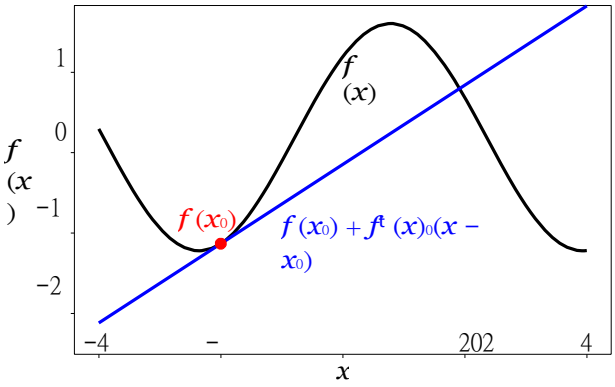


图5.12 一个函数的线性近似。原始函数f在以下位置被线性化$x_0=$使用2一阶泰勒系列扩展。

如果$f(x, y)$是一个两次（连续）可微的函数，那么

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}, \tag{5.146}$$

即，微分的顺序并不重要，而相应的*Hessian矩阵*

Hessian矩阵

$$\boldsymbol{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} \tag{5.147}$$

是对称的。Hessian被表示为$\nabla^2_{x,y} f(x, y)$。一般来说，对于$\boldsymbol{x} \in \mathbb{R}^n$和$f : \mathbb{R}^n \to \mathbb{R}$，Hessian是一个$n \times n$矩阵。Hessian衡量的是函数在$(x, y)$周围的局部曲率。

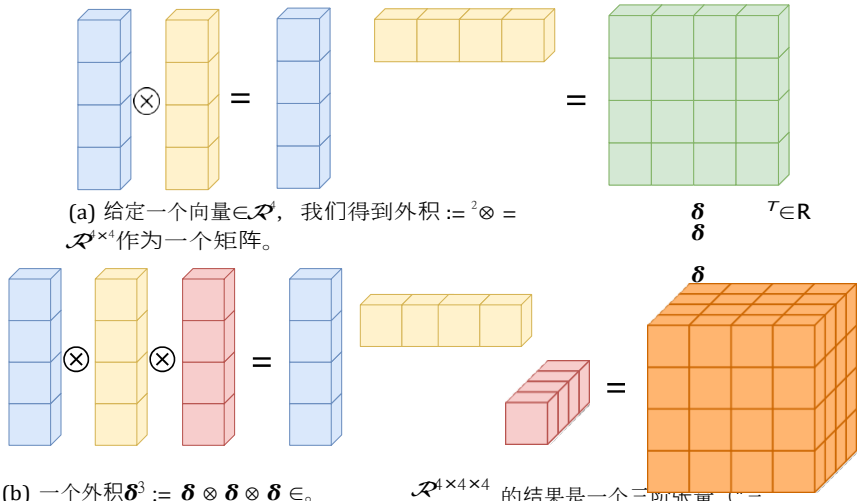
备注（矢量场的Hessian）。如果$f : \mathbb{R}^n \to \mathbb{R}^m$是一个矢量场，Hessian是一个$(m \times n \times n)$张量。 ◆

5.8 线性化和多变量泰勒系列

一个函数f的梯度∇f通常用于f在$\boldsymbol{x}_0$周围的局部线性近似值。

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_0) + (\nabla_{\boldsymbol{x}} f)(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) \,。 \tag{5.148}$$

这里$(\nabla_{\boldsymbol{x}} f)(\boldsymbol{x}_0)$是$f$相对于$\boldsymbol{x}$的梯度，在$\boldsymbol{x}_0$处评估。5.12说明了一个函数$f$在输入$\boldsymbol{x}_0$处的线性近似。这个近似是局部准确的，但是我们离$\boldsymbol{x}_0$越远，近似就越差。方程(5.148)是$f$在$\boldsymbol{x}_0$处的多变量泰勒级数展开的一个特例，在这里我们只考虑前两个项。我们在下文中讨论更一般的情况，这将允许更好的近似。



图：外积的可视化**5.13**。向量的外积会使数组的维数增加一个1项。(a)两个向量的外积产生了一个矩阵；(b)三个向量的外积产生了
一个三阶张量。

(a) 给定一个向量$\in\mathcal{X}^{4}$，我们得到外积$:=\ ^{2}\otimes=$ $\mathcal{X}^{4\times4}$作为一个矩阵。

$\boldsymbol{\delta}$ $^{T}\in\mathrm{R}$

$\boldsymbol{\delta}$

$\boldsymbol{\delta}$

(b) 一个外积$\boldsymbol{\delta}^{3}:=\boldsymbol{\delta}\otimes\boldsymbol{\delta}\otimes\boldsymbol{\delta}\in$。
维矩阵"），即一个有三个索引的数组。

$\mathcal{X}^{4\times4\times4}$ 的结果是一个三阶张量（"三

定义（5.7多变量泰勒系列）。我们考虑一个函数

$$f:\mathrm{R}^{D}\rightarrow\mathrm{R} \tag{5.149}$$

$$\boldsymbol{x}\rightarrow f(\boldsymbol{x}),\quad \boldsymbol{x}\in\mathrm{R}^{D}, \tag{5.150}$$

multivariate Taylor series

当我们定义差异向量$\boldsymbol{\delta}:=\boldsymbol{x}\boldsymbol{x}$时$_{00}$，$f$在$(\boldsymbol{x}_{0})$处的*多变量泰勒级数*被定义为

$$f(\boldsymbol{x})=\sum_{k=0}^{\infty}\frac{D_{\boldsymbol{x}}^{k}f(\boldsymbol{x}_{0})}{k!}\boldsymbol{\delta}^{k}, \tag{5.151}$$

其中Df^{k} $(\boldsymbol{x}_{0})$是f*相对于*$\boldsymbol{x}$*的*第k次（总）导数，在$\boldsymbol{x}$处$_{0}$评估。

泰勒多项式

定义（5.8泰勒多项式）。f在$\boldsymbol{x}$处$_{0}$的n度*泰勒多项式*包含( )中数列的前n+1个分量。5.151)，定义为

$$T_{n}(\boldsymbol{x})=\sum_{k=0}^{n}\frac{D_{\boldsymbol{x}}^{k}f(\boldsymbol{x}_{0})}{k!}\boldsymbol{\delta}^{k}. \tag{5.152}$$

在(5.151)和(5.152)中，我们使用了略显草率的$\boldsymbol{\delta}^{k}$的符号，它对向量$\boldsymbol{x}\in\mathrm{R}^{D}$、$D>1$和$k>$没有定义1。请注意，$^{k}Df$和$\boldsymbol{\delta}^{k}$都是$k$阶张量，也就是$k$维数组。该

矢量可以实现为一个一维数组，矩阵可以实现为一个二维数组。

$\underbrace{\boldsymbol{x},\ldots,\boldsymbol{x}}_{k\text{次}}\in\mathrm{R}^{D}$的$k$倍外积得到$D\times D\times\ldots\times D$的，用⊗表示，例如：

第k阶张量$\boldsymbol{\delta}^{k}\in\mathrm{R}$是作为向量$\boldsymbol{\delta}$。

$$\boldsymbol{\delta}^{2}:=\boldsymbol{\delta}\otimes\boldsymbol{\delta}=\boldsymbol{\delta}\boldsymbol{\delta}^{\mathrm{T}},\qquad \boldsymbol{\delta}^{2}[i,j]=\delta[i]\delta[j] \tag{5.153}$$

。

$$\boldsymbol{\delta}^3 := \boldsymbol{\delta} \otimes \boldsymbol{\delta} \otimes \boldsymbol{\delta}, \quad \boldsymbol{\delta}^3[i,j,k] = \delta[i]\delta[j]\delta[k] 。 \tag{5.154}$$

图5.13显示了两个这样的外积。一般来说，我们得到的条件是

$$Df^k(\boldsymbol{x}_0)\boldsymbol{\delta}^k = \sum_{i_1=1}^{\vec{A}\vec{A}\vec{A}} \cdots \sum_{i_k=1}^{\boldsymbol{x}} Df^k(\boldsymbol{x}_0)[i_1, \ldots, i_k]\delta[i_1] \cdots \delta[i_k] \tag{5.155}$$ 。

在泰勒级数中，$^kDf(\boldsymbol{x}_0)\boldsymbol{\delta}^k$包含$k$-th阶多项式。

现在我们定义了矢量场的泰勒级数，让我们明确地写下泰勒级数展开的第一项$^kDf(\boldsymbol{x}_0)\boldsymbol{\delta}^k$，用于 $k = 0, \ldots$而$3\boldsymbol{\delta} := \boldsymbol{x} - \boldsymbol{x}$:

$$k =: 0 D^0 f(\boldsymbol{x}_0)\boldsymbol{\delta}^0 = f(\boldsymbol{x}_0) \in \mathbb{R} \tag{5.156}$$

$$k =: 1 D^1 f(\boldsymbol{x}_0)\boldsymbol{\delta}^1 = \nabla f(\boldsymbol{x}_0)\boldsymbol{\delta} = \sum_{i=1}^{D} \nabla_i f(\boldsymbol{x}_0)[i]\delta[i] \in \mathbb{R} \tag{5.157}$$

$$k =: 2 D^2 f(\boldsymbol{x}_0)\boldsymbol{\delta}^2 = \operatorname{tr}\left(\boldsymbol{H}(\boldsymbol{x}_0)\boldsymbol{\delta}\boldsymbol{\delta}^T\right) = \boldsymbol{\delta}^T \boldsymbol{H}(\boldsymbol{x}_0)\boldsymbol{\delta} \tag{5.158}$$

$$= \sum_{i=1}^{D}\sum_{j=1}^{D} H[i,j]\delta[i]\delta[j] \quad \mathbb{R} \tag{5.159}$$

$$k =: 3 D^3 f(\boldsymbol{x}_0)\boldsymbol{\delta}^3 = \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D} Df(\boldsymbol{x}_0)[i,j,k]\delta[i]\delta[j]\delta[k] \in \mathbb{R} \tag{5.160}$$

这里，$\boldsymbol{H}(\boldsymbol{x}_0)$是$f$在$\boldsymbol{x}_0$处评估的Hessian。

np.einsum('i,i',Df1,d)
np.einsum('ij,i,j', Df2,d,d)
np.einsum('ijk,i,j,k', Df3,d,d,d)

例5.15（一个有两个变量的函数的泰勒级数展开）。

考虑一下函数

$$f(x,y) = x^2 + 2xy + y^3 。 \tag{5.161}$$

我们想计算f在（x_0，y_0）=（1，2）处的泰勒级数展开。在我们开始之前，让我们讨论一下要期待什么。中的函数(5.161)中的函数是一个度数3为 . 的多项式。我们正在寻找一个泰勒级数展开，它本身就是一个多项式的线性组合。因此，我们不希望泰勒级数展开包含四阶或更高阶的项来表达一个三阶多项式。这就意味着，只要确定(5.151)的前四项就足够了，可以精确地改变(5.161).

为了确定泰勒级数扩展，我们从常数项和一阶导数开始，它们由以下公式给出

$$f(1,2) = 13 \tag{5.162}$$

$$\frac{\partial f}{\partial x} = 2x + 2y \Rightarrow \frac{\partial f}{\partial x}(1,2) = \tag{5.163}$$

$$\frac{\partial f}{\partial y} = 2x + 3y^2 \Rightarrow \frac{\partial f}{\partial y}(1,2) = 14 \tag{5.164}$$

因此，我们得到

$$D^1_{x,y} f(1,2) = \nabla_{x,y} f(1,2) = \begin{bmatrix} \frac{\partial f}{\partial x}(1,2) & \frac{\partial f}{\partial y}(1,2) \end{bmatrix} = \begin{bmatrix} 6 & 14 \end{bmatrix} \in \mathbb{R}^{1\times 2} \tag{5.165}$$

以致于

$$D^1_{x,y} f(1,2)\boldsymbol{\delta} = \begin{bmatrix} 6 & 14 \end{bmatrix} \begin{bmatrix} x-1 \\ y-2 \end{bmatrix} = 6(x-1) + 14(y-2)。 \tag{5.166}$$

请注意，$D^1_{x,y} f(1,2)\boldsymbol{\delta}$只包含线性项，即一阶多义性。

纪念品。

二阶偏导数由以下公式给出

$$\frac{\partial^2 f}{\partial x^2} = 2 \Rightarrow \frac{\partial^2 f}{\partial x^2}(1,2) = 2 \tag{5.167}$$

$$\frac{\partial^2 f}{\partial y^2} = 6y \Rightarrow \frac{\partial^2 f}{\partial y^2}(1,2) = 12 \tag{5.168}$$

$$\frac{\partial^2 f}{\partial y \partial x} = 2 \Rightarrow \frac{\partial^2 f}{\partial y \partial x}(1,2) = 2 \tag{5.169}$$

$$\frac{\partial^2 f}{\partial x \partial y} = 2 \Rightarrow \frac{\partial^2 f}{\partial x \partial y}(1,2) = 2 \tag{5.170}$$

当我们收集二阶偏导数时，我们得到Hes- sian

$$\boldsymbol{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 6y \end{bmatrix}, \tag{5.171}$$

以致于

$$\boldsymbol{H}(1,2) = \begin{bmatrix} 2 & 2 \\ 2 & 12 \end{bmatrix} \in \mathbb{R}^{2\times 2}. \tag{5.172}$$

因此，泰勒数列扩展的下一个项由以下公式给出

$$\frac{D^2_{x,y} f(1,2)}{2!}\boldsymbol{\delta}^2 = \frac{1}{2}\boldsymbol{\delta}^\top \boldsymbol{H}(1,2)\boldsymbol{\delta} \tag{5.173a}$$

$$= \frac{1}{2}\begin{bmatrix} x-1 & y-2 \end{bmatrix} \begin{bmatrix} 2 & 2 \\ 2 & 12 \end{bmatrix} \begin{bmatrix} x-1 \\ y-2 \end{bmatrix} \tag{5.173b}$$

$$= (x-1)^2 + 2(x-1)(y-2) + 6(y-2)^2。 \tag{5.173c}$$

这里，$D^2_{x,y} f(1,2)\boldsymbol{\delta}^2$只包含二次项，即二阶聚

命名。

三阶导数得到的结果是

$$D^3_{x,y}f = \frac{\partial \boldsymbol{H}}{\partial x} \quad \frac{\partial \boldsymbol{H}}{\partial y} \in \mathbb{R}^{2\times 2\times 2}, \tag{5.174}$$

$$D^3_{x,y}f[:,:,1] = \frac{\partial \boldsymbol{H}}{\partial x} = \begin{matrix} \frac{\partial^3 f}{\partial x^3} & \frac{\partial^3 f}{\partial x^2 \partial y} \\ \frac{\partial^3 f}{\partial x \partial y \partial x} & \frac{\partial^3 f}{\partial x \partial y^2} \end{matrix}, \tag{5.175}$$

$$D^3_{x,y}f[:,:,2] = \frac{\partial \boldsymbol{H}}{\partial y} = \begin{matrix} \frac{\partial^3 f}{\partial y \partial x^2} & \frac{\partial^3 f}{\partial y \partial x \partial y} \\ \frac{\partial^3 f}{\partial y \partial^2 x} & \frac{\partial^3 f}{\partial y^3} \end{matrix}. \tag{5.176}$$

由于Hessian中的大多数二阶偏导在(5.171)中的大部分二阶偏导都是常数，唯一非零的三阶偏导是

$$\frac{\partial^3 f}{\partial y^3} = 6 \Rightarrow \frac{\partial^3 f}{\partial y^3}(1,2) = 6. \tag{5.177}$$

高阶导数和度的混合导数（3如：。$\frac{\partial f^3}{\partial x^2 \partial y}$）消失，这样一来

$$D^3_{x,y}f[:,:,1] = \begin{matrix} 0 & \\ 0 & \end{matrix}, \qquad D^{0_3}_{x,y}f[:,:,2] = \begin{matrix} 0 & 0 \\ 0 & 0\ 6 \end{matrix} \tag{5.178}$$

和

$$\frac{D^3_{x,y}f(1,2)}{3!}\boldsymbol{\delta}^3 = (\underline{y}2)^3, \tag{5.179}$$

它收集了泰勒级数的所有立方项。总的来说，f在（x_0，y_0）=（1，2）的（精确）泰勒级数展开是

$$f(x) = f(1,2) + \underline{D}_{x,y}f(1,2)\boldsymbol{\delta} + \frac{D^2_{x,y}f(1,2)}{2!}\boldsymbol{\delta}^2 + \frac{D^3_{x,y}f(1,2)}{3!}\boldsymbol{\delta}^3 \tag{5.180a}$$

$$= f(1,2) + \frac{\partial f(1,2)}{\partial x}(x-1) + \frac{\partial f(1,2)}{\partial y}(y-2) + \frac{1}{2!}\frac{\partial^2 f(1,2)}{\partial x^2}(x-1)^2 + \frac{\partial^2 f(1,2)}{\partial u^2}(y-2)^2 + 2\frac{\partial^2 f(1,2)}{\partial x \partial u}(x-1)(y-2) + \frac{1}{6}\frac{\partial^3 f(1,2)}{\partial u^3}(y-2)^3 \tag{5.180b}$$

$$= 13 + 6(x-1) + 14(y-2) + (x-1)^2 + 6(y-2)^2 + 2(x-1)(y-2) + (y-2)^3。 \tag{5.180c}$$

在这种情况下，我们得到了( )中多项式的精确泰勒级数展开。5.161)，即(5.180c)中的多项式与()中的原始多项式完全相同。5.161).在这个特殊的

例子中，这个结果并不令人惊讶，因为原函数是一个三阶多项式，我们 *矢量微积分*
通过常数项、一阶、二阶和三阶多项式在（5.180c）中的线性组合来表示
。

5.9 进一步阅读

关于矩阵微分的进一步细节，以及对所需线性代数的简短回顾，可以在Magnus和Neudecker（2007）中找到。自动微分有很长的历史，我们可以参考Griewank和Walther（2003），Griewank和Walther（2008），以及Elliott（2009）。
以及其中的参考文献。

在机器学习（和其他学科）中，我们经常需要计算期望值，也就是说，我们需要解决以下形式的积分

$$\mathbb{E}_{\boldsymbol{x}}[f(\boldsymbol{x})] = \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \text{ 。} \tag{5.181}$$

即使$p(\boldsymbol{x})$是一个方便的形式（例如，高斯），这个积分一般也不能用分析法解决。f的泰勒级数展开是找到近似解的一种方法。假设$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

扩展的卡尔曼

是高斯的，那么围绕$\boldsymbol{\mu}$的一阶泰勒级数展开可以使非线性函数f局部线性化。对于线性函数，如果$p(\boldsymbol{x})$是高斯分布，我们可以准确地计算出均值（和协方差）（见第6.5节）。这一特性被*扩展的Kalman*算法所大量利用。

滤波器（Maybeck,1979）用于非线性动态系统（也称为 "状态空间模型"）的在线状态估计。其他决定性的方法

无痕变换拉普拉斯

中的积分进行近似。5.181)中的积分是*不*需要任何梯度的*无痕变换*(Julier and Uhlmann,1997)，或*拉普拉斯变换(Laplace)*。

逼近法

（MacKay，2003；Bishop，2006；Murphy，2012），该法使用了围绕$p(\boldsymbol{x})$模式的局部高斯近似的二阶泰勒级数展开（需要Hessian）。

练习

5.1 计算以下情况的导数 $f^{\mathrm{I}}(x)$

$$f(x) = \log(x^4)\sin(x^3) \text{ 。}$$

5.2 计算logistic sigmoid的导数$f^{\mathrm{I}}(x)$

$$f(x) = \frac{1}{1+\exp(-x)} .$$

5.3 计算函数的导数 $f^{\mathrm{I}}(x)$

$$f(x) = \exp(-\frac{1}{2\sigma^2}(x-\mu)^2) \text{ 。}$$

其中μ、$\sigma \in \mathcal{R}$是常数。

5.4 计算泰勒多项式 $T_n, n = 0, \ldots, 5 f(x) = \sin(x) + \cos(x)$ 在$x_0 = 0$.

考虑到以下函数。

$$f_1(\boldsymbol{x}) = \sin(x_1)\cos(x_2), \quad \boldsymbol{x} \in \mathcal{R}^2$$
$$f_2(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T\boldsymbol{y}, \quad \boldsymbol{x}, \boldsymbol{y} \in \mathcal{R}^n f_3(\boldsymbol{x})$$
$$= \boldsymbol{x}\boldsymbol{x}^T, \quad \boldsymbol{x} \in \mathcal{R}^n$$

a. $\frac{\partial f}{\partial \boldsymbol{x}}$的尺寸是多少？

b.计算雅各布系数。

5.6 对$\boldsymbol{t}$进行微分 f，对$\boldsymbol{X}$进行微分g，其中

$$f(\boldsymbol{t}) = \sin(\log(\boldsymbol{t}\boldsymbol{T}\boldsymbol{t})), \quad t \in \mathcal{R}^D$$
$$g(\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}), \quad \boldsymbol{A} \in \mathcal{R}^{D\times E}, \boldsymbol{X} \in \mathcal{R}^{E\times F}, \boldsymbol{B} \in \mathcal{R}^{F\times D},$$

其中tr( )表示追踪。

5.7 通过使用链式规则计算下列函数的导数$\mathrm{d}f/\mathrm{d}\boldsymbol{x}$。提供每一个偏导的尺寸。详细描述你的步骤。

a.

$$f(z) = \log(1+z), \quad z = \boldsymbol{x}\boldsymbol{T}\boldsymbol{x}, \quad \boldsymbol{x} \in \mathcal{R}^D$$

b.

$$f(\boldsymbol{z}) = \sin(\boldsymbol{z}), \quad \boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \quad \boldsymbol{A} \in \mathcal{R}^{E\times D}, \boldsymbol{x} \in \mathcal{R}^D, \boldsymbol{b} \in \mathcal{R}^E$$

其中sin(·)被应用于$\boldsymbol{z}$的每个元素。

5.8 计算下列函数的导数 $df/d\boldsymbol{x}$。请详细描述你的步骤。

a.使用连锁规则。提供每一个局部派生的尺寸。

$$f(z) = \exp(-\frac{1}{2}z)$$
$$z = g(\boldsymbol{y}) = \boldsymbol{y}\boldsymbol{T}\boldsymbol{S}^{-1}$$
$$\boldsymbol{y} = h(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{\mu}$$

其中$\boldsymbol{x}, \boldsymbol{\mu} \in \mathcal{R}^D$，$\boldsymbol{S} \in \mathcal{R}^{D\times D}$。

b.

$$f(\boldsymbol{x}) = \mathrm{tr}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} + \sigma^2\boldsymbol{I}), \quad \boldsymbol{x} \in \mathcal{R}^D$$

这里tr($\boldsymbol{A}$)是$\boldsymbol{A}$的迹线，即对角线元素A_{ii}的总和。

提示：明确写出外积。

c.使用连锁规则。提供每一个偏导数的尺寸。你不需要明确地计算偏导数的乘积。

$$\boldsymbol{f} = \tanh(\boldsymbol{z}) \in \mathcal{R}^M$$
$$\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \quad \boldsymbol{x} \in \mathcal{R}^N, \boldsymbol{A} \in \mathcal{R}^{M\times N}, \boldsymbol{b} \in \mathcal{R}^M.$$

这里，tanh被应用于$\boldsymbol{z}$的每个分量。

5.9 我们定义

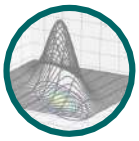
$$g(\boldsymbol{z}, \boldsymbol{v}) := \log p(\boldsymbol{x}, \boldsymbol{z}) - \log q(\boldsymbol{z}, \boldsymbol{v})$$
$$\boldsymbol{z} := t(\boldsymbol{\epsilon}, \boldsymbol{v})$$

对于可微调函数p, q, t和$\boldsymbol{x} \in \mathcal{R}^D, \boldsymbol{z} \in \mathcal{R}^E, \boldsymbol{v} \in \mathcal{R}^F, \boldsymbol{\epsilon} \in \mathcal{R}^G$。通过使用连锁规则，计算梯度

$$\frac{\mathrm{d}g}{\mathrm{d}\boldsymbol{v}}(\boldsymbol{z}, \boldsymbol{v})。$$

6

概率和分布



概率，宽泛地说，是关于不确定性的研究。概率可以被认为是一个事件发生的几率，或者说是对一个事件的相信程度。然后，我们想用这个概率来衡量实验中某事发生的机会。正如第一章所提到的，我们经常对数据的不确定性、机器学习模型的不确定性以及模型所产生的预测的不确定性进行量化。量化不确定性需要一个*随机变量的*概念，这是一个将随机实验的结果映射到我们感兴趣的一系列属性的函数。与随机变量相关联的是一个衡量特定结果（或一组结果）发生概率的函数；这被称为*概率分布*。

随机变量

概率

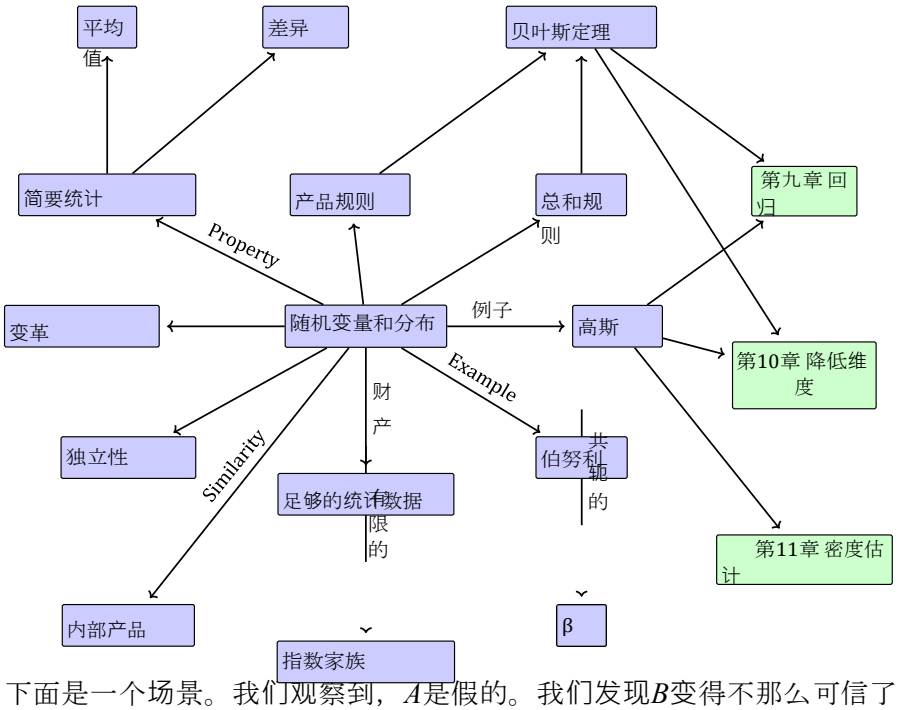
概率分布被用作其他概念的基石，如概率建模（第8.4节）、图形模型（第8.5节）和模型选择（第8.6节）。在下一节中，我们将介绍定义概率空间的三个概念（样本空间、事件和事件的概率），以及它们与第四个概念即随机变量的关系。由于严格的介绍可能会掩盖这些概念背后的直觉，所以介绍时特意略微用手摇晃。本章所介绍的概念的概要见图6.1.

6.1 概率空间的构建

概率论的目的是定义一个数学结构来描述实验的随机结果。例如，在抛掷一枚硬币时，我们无法确定结果，但通过大量抛掷硬币，我们可以观察到平均结果的规律性。使用这种概率的数学结构，目标是进行自动推理，在这个意义上，概率概括了逻辑推理（Jaynes，2003）。

6.1.1 哲学问题

在构建自动推理系统时，经典的布尔逻辑并不允许我们表达某些形式的合理推理。请考虑



图为本章所述的与随机变量和概率分布有关的概念的**6.1**思维导图。

下面是一个场景。我们观察到，A是假的。我们发现B变得不那么可信了，尽管从经典逻辑中无法得出结论。我们观察到B是真的。似乎A变得更可信了。我们每天都在使用这种形式的推理。我们在等一个朋友，并考虑三种可能性。H1，她很准时；H2，她被交通延误了；H3，她被外星人绑架了。当我们观察到我们的朋友迟到了，我们必须从逻辑上排除H1。我们也倾向于认为H2更有可能，尽管我们在逻辑上不需要这样做。最后，我们可能认为H3是可能的，但我们仍然认为它是相当不可能的。

我们如何得出H2是最合理的答案的结论？从这个角度看，"对于合理的推理有必要将离散的真假值扩展到连续的合理性"（Jaynes,2003）。概率论可以被认为是布尔逻辑的一个概括。在机器学习的背景下，它经常以这种方式被应用于自动化推理系统的设计中。关于概率论如何成为推理系统的基础的进一步论证，可以在Pearl（1988）中找到。

考克斯（Jaynes,2003）研究了概率的哲学基础以及它应该如何与我们认为应该是真的（在逻辑意义上）有某种联系。另一种思考方式是，如果我们对我们的常识有精确的认识，我们最终会构建出概率。

E.T. 杰恩斯（1922-1998）确定了三个数学标准，它们必须适用于所有的合理性。

1. 可信度由实数表示。2.这些数字必须是基于常识的规则。

3.由此产生的推理必须是一致的，符合 "一致 "一词的以下三种含义。

(a) 一致性或不矛盾性。当通过不同的手段可以达到相同的结果时，必须在所有情况下找到相同的可信度值。

(b) 诚实。必须考虑到所有可用的数据。(c)可重复性。如果我们对两个问题的知识状况是意思相同，那么我们就必须对 "我 "和 "我 "之间的关系赋予同样的可信度。

他们两个人。

考克斯-杰恩斯定理证明这些合理性足以定义适用于合理性p的普遍数学规则，直至被一个任意的单调函数转化。重要的是，这些规则是概率的规则。

*备注：在机器学习和统计学中，有两种主要的概率解释：贝叶斯解释和频繁解释。*在机器学习和统计学中，有两种主要的概率解释：贝叶斯和频繁主义解释（Bishop，2006；Efron和Hastie，2016）。贝叶斯解释使用概率来说明用户对一个事件的不确定性程度。它有时被称为 "主观概率 "或 "信仰程度"。频繁解释法考虑的是感兴趣的事件与发生的事件总数的相对频率。一个事件的概率被定义为该事件在极限情况下的相对频率，当一个人有无限的数据。 ◆

一些关于概率模型的机器学习文本使用了懒惰的符号和行话，这让人很困惑。这篇文章也不例外。多个不同的概念都被称为 "概率分布"，而读者必须经常从上下文中解读其含义。帮助理解概率分布的一个窍门是检查我们是在试图为某种分类的东西（离散随机变量）还是为某种连续的东西（连续随机变量）建模。我们在机器学习中处理的问题种类与我们是在建立分类模型还是连续模型密切相关。

6.1.2 概率和随机变量

在讨论概率时，有三个不同的概念经常被混淆。首先是概率空间的概念，它允许我们对概率的概念进行量化。然而，我们大多不直接使用这个基本的概率空间。相反，我们使用随机变量（第二个想法），将概率转移到一个更方便的（通常是数字的）空间。第三个想法是与随机变量相

关的分布或规律的想法。我们将在本节中介绍前两个想法，并在第3节中对第三个想法进行扩展。6.2.

现代概率是建立在科尔莫戈罗夫提出的一套公理之上的

(Grinstead and Snell,1997;Jaynes,2003)，介绍了样本空间、事件空间和概率测量这三个概念。概率空间模拟了一个具有随机结果的真实世界过程（称为实验）。

样本空间Ω

*样本空间*是实验的所有可能结果的集合，样本空间

通常用Ω来表示。例如，两次连续掷硬币的样本空间为hh, tt, ht, th，其中 "h "表示 "正面"，"t "表示 "反面"。

活动空间 A

*事件空间*是实验的潜在结果的空间。一个 事件空间

如果在实验结束时我们可以观察到某个特定的结果ω Ω是否在A中，那么样本空间 ω的子集$A \subseteq$ 在事件 空间。事件空间是 通过考虑ω的子集的集合得到的，对于离散概率分布（第6.2.1节），A通常是ω的幂集。

概率*P*

对于每个事件$A \in$ A，我们都有一个数字$P(A)$，用来衡量该事件发生的概率或相信程度。$P(A)$被称为 *A*的*概率*。 概率

单一事件的概率必须位于区间[0，1]内，而样本空间Ω中所有结果的总概率必须是 ，1即P（Ω）=1。给定一个概率空间（Ω， ，*P*），我们想用它来模拟一些现实世界的现象。在机器学习中，我们经常避免明确提及概率空间，而是提及感兴趣的数量的概率，我们用 A 。在本书中，我们指的是

作为*目标空间*，并将 T 其中的元素 称为 状态。我们引入目标空间

函数$X: \Omega \to$ T，它接收Ω的一个元素（一个结果），并返回一个特定的感兴趣的数量x，一个值。 T 这种关联/映射

从Ω T 到 被 称为一个*随机变量*。例如，在抛掷 随机变量的情况下

两枚硬币并计算头数，一个随机变量X映射到三种可能的结果。X(hh)=2，X(ht)=1，X(th)=1，X(tt)=0。 在这种特殊情况下，T={0,1，2}，它的概率为

在我们感兴趣的T的元素上。对于一个有限的样本空间Ω和有限的T，对应于随机变量的函数本质上是一个查找表。对于任何子集$S \subseteq T$，我们将$P(X \in S)$（概率）与随机变量X对应的特定事件发生联系起来。6.1提供了该术语的一个具体说明。

"随机 "这个名字变量 "是造成误解的一个重要原因，因为它既不是随机的，也不是一个变量。它是一个函数。

备注。 不幸的是，上述的样本空间Ω在不同的书中被称为不同的名字。Ω的另一个常用名称是 "状态空间"(Jacod and Protter,2004)，但状态空间有时被保留用于指代动态系统中的状态(Hasselblatt and

Katok,2003）。有时用于描述Ω的其他名称是。"样本描述空间"、"可能性空间 "和 "事件空间"。 ◆

这个玩具的例子本质上是一个有偏见的抛硬币的例子。

例子 6.1

我们假设读者已经熟悉计算事件集的交集和合集的概率。在2Walpole等人(2011)的章节中，可以找到对概率的较 温和的介绍，其中有许多例子。

考虑一个统计学实验，我们建立一个游乐场游戏的模型，包括从一个袋子里抽出两个硬币（有替换）。袋子里有来自美国（表示为$）和英国（表示为£）的硬币，由于我们从袋子里抽出两枚硬币，总共有四个结果。那么这个实验的状态空间或样本空间Ω是（$，$），（$。£), (£, $), (£, £).让我们假设硬币袋的组成是这样的：抽签时随机返回一个$，概率为0.3.....。

我们感兴趣的事件是重复抽签返回$的总次数。让我们定义一个随机变量$X$，将样本空间Ω映射到 ，表示我们从袋子中抽出$的次数。从前面的样本空间我们可以看到，我们可以得到零个$，一个$，或者两个$，因此=0，1 ， 。2随机变量$X$（一个函数或查找表）可以用下面这样的表格表示。

$$X((\$, \$)) = 2 \tag{6.1}$$
$$X((\$, £)) = 1 \tag{6.2}$$
$$X((£, \$)) = 1 \tag{6.3}$$
$$X((£, £)) = .0 \tag{6.4}$$

由于我们在抽第二枚硬币之前先把第一枚硬币还回去，这意味着两次抽奖是相互独立的，我们将在第6节中讨论。 6.4.5.请注意，有两种实验结果，它们映射到同一事件，其中只有一次抽奖能得到$。 因此，$X$的概率质量函数（第6.2.1节）由以下公式给出

$$\begin{aligned} p(x = 2) &= p((\$, \$)) \\ &= p(\$) - p(\$) \\ &= 0.3 - 0.3 = 0.09 \end{aligned} \tag{6.5}$$

$$\begin{aligned} p(x = 1) &= p((\$, £) \cup (£, \$)) \\ &= P((\$, £)) + P((£, \$)) \\ &= 0.3 - (1 - 0.3) + (1 - 0.3) - 0.3 = 0.42 \end{aligned} \tag{6.6}$$

p（x=0）=p（（£，£））。

$$= \qquad 3) = 0.49. \qquad (6.7)$$

$$p(\pounds) - p(\pounds) = (1 - 0.3) - (1 - 0.$$

在计算中，我们将两个不同的概念等同起来，即X的输出概率和Ω中样本的概率。例如，在(6.7)我们说$P(X=0)=P((\text{£}, \text{£}))$。考虑随机变量$X$ ：Ω → 和一个子集S （T 例如， ，的单个元素，如抛郑两个硬币时得到一个头的结果）。S让$X(^{-1}S)$是X对S的前像，即在X下映射到S的Ω元素的集合 $; \omega \{ \Omega : X(\omega)$ S 。理解从Ω中的事件通过随机变量进行概率转换的一种方法是 }
X是与S的前像的概率相关联（Jacod and Protter,2004）。对于$S \subseteq$T，我们有这样的符号

$$P(_X S) = P\,(X \in S) = P\,(X(^{-1}S)) = P\,(\{\omega \in \Omega : X(\omega) \in S\}) \text{。} \quad (6.8)$$

的左手边(6.8)的左边是我们感兴趣的可能结果集（例如，$数=1）的概率。通过随机变量$X$，它将状态映射到结果，我们在(6.8)，这是具有该属性的状态集（在Ω中）的概率（例如$\text{£}, \text{£}$）。我们说，一个随机变量X是按照一个特定的概率分布P_X分布的，它定义了事件和随机变量结果的概率之间的概率映射。换句话说，函数P_X或等同于$P\,X^{-1}$是随机变量X的*规律*或*分布*。 。

备注：。目标空间，也就是随机变量X的范围，用来表示概率空间的种类，即随机变量。当是有限的或可数的无限时，这被称为离散随机变量（第6.2.1).对于连续随机变量（第6.2.2节），我们只考虑T=R或T=R^D。

趣，这些实例与我们已经观察到的实例不完全相同。

◆

6.1.3 统计数据

概率论和统计学经常被放在一起介绍，但它们涉及到不确定性的不同方面。对比它们的一种方法是考虑问题的种类。使用概率论，我们可以考虑一些过程的模型，其中潜在的不确定性是由随机变量捕获的，我们使用概率的规则来推导出发生的事情。在统计学中，我们观察到一些事情已经发生，并试图找出解释观察结果的基本过程。在这个意义上，机器学习与统计学很接近，它的目标是构建一个能充分代表产生数据的过程的模型。我们可以使用概率规则来获得一些数据的 "最佳拟合 "模型。

机器学习系统的另一个方面是，我们对泛化误差感兴趣（见第八章）。这意味着我们实际上对我们的系统在未来观察到的实例上的表现感兴

到目前为止所看到的。这种对未来表现的分析依赖于概率和统计，其中大部分内容超出了本章所要介绍的范围。我们鼓励感兴趣的读者看看Boucheron等人（2013）和Shalev-Shwartz和Ben-David（2014）的书。我们将在第八章看到更多关于统计的内容。

6.2 离散和连续概率

让我们把注意力集中在第6.1节中介绍的描述事件概率的方法上。根据目标空间是离散的还是连续的，指代分布的自然方式是不同的。当目标空间是离散的，我们可以指定一个随机变量X取一个特定值x的概率，表示为$P(X=x)$。离散随机变量X的表达式$P(X=x)$被称为

概率质量

的*概率质量函数*。当目标空间是连续的，例如。

在实线的函数中

，更自然的是指定随机变量X在一个区间内的概率，用$P(a \leqslant X \leqslant b)$表示，因为$a < b$。表达式$P(X \leqslant x)$为

积累性

一个连续的随机变量X被称为*累积分布*

分布函数

函数。我们将在第一节讨论连续随机变量。6.2.2.我们将在第一节中重新审视术语并对比离散和连续随机变量。6.2.3.

单变量

备注。我们将使用*单变量*分布这一短语来指单个随机变量的分布（其状态用非粗体的x表示）。我们将把一个以上的随机变量的分布称为*多变量*分布，并且通常会考虑一个随机变量的向量。

多变量

变量（其状态用黑体$\boldsymbol{x}$表示）。 ◆

6.2.1 离散概率

当目标空间是离散的，我们可以把多个随机变量的概率分布想象成填写一个（多维）数字阵列。图6.2显示了一个例子。联合概率的目标空间是每个随机变量的目标空间的笛卡尔乘积。我们把*联合概率*定义为两个值

联合概率

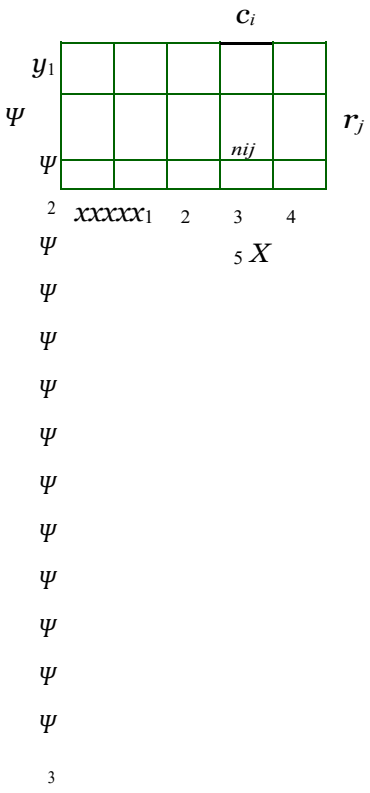
的共同条目

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N} \tag{6.9}$$

其中n_{ij}是具有x_i和y_j状态的事件的数量，N是事件的总数量。联合概率是两个事件之间的概率，即$P(X=x_i, Y=y_j) = P(X=x_i \cap Y=y_j)$。图6.2说明了一个离散概率的*概率质量函数*（pmf）。

概率质量

功能性分布。对于两个随机变量X和Y，其概率为



图为离散双变量概率质量函数的可视化**6.2**，随机变量为*X*和*Y*。此图改编自Bishop（2006）。

X=*x*和*Y*=*y*的*概率*被（懒惰地）写成*p*(*x, y*)，称为联合概率。我们可以把概率看作一个函数，它接收状态*x*和*y*并返回一个实数，这就是我们写*p*(*x, y*)的原因。

*X*取值*x*的*边际概率*，与价值边际概率无关。

的随机变量*Y*被（懒惰地）写成*p*（*x*）。我们写*X* *p*(*x*)表示随机变量*X*是按照*p*(*x*)分布的。如果我们只考虑*X*=*x*的实例，那么，实例的百分比

(*条件概率*)，对于*Y*=*y*来说，可以写成(懒散地)*p*(*y x*)。条件 |

概率

例子 6.2

考虑两个随机变量X和Y，其中X有五个可能的状态，Y有三个可能的状态，如图6.2所示。我们用n_{ij}表示状态为$X=x_i$和$Y=y_j$的事件的数量，用N表示事件的总数。数值c_i是第i列的各个频率之和，即$c_i = \sum_{j=1}^{3} n_{ij}$同样地，值$r_j$是行和，也就是说，$r_j = \sum_{i=1}^{5} n_{ij}$.利用这些定义，我们可以紧凑地表达$X$和$Y$的分布。

每个随机变量的概率分布，即边际概率，可以看做是对一行或一列的总和

$$P(X = x_i) = \frac{c_i}{N} = \frac{\sum_{j=1}^{3} n_{ij}}{N} \tag{6.10}$$

和

$$P(Y = y_j) = \frac{r_j}{N} = \frac{\sum_{i=1}^{5} n_{ij}}{N} \tag{6.11}$$

其中c_i和r_j分别是概率表的第i列和第j行。按照惯例，对于事件数量有限的离散随机变量，我们假设概率之和为1，即：。

$$\sum_{i=1}^{5} P(X = x_i) = 1 \qquad \sum_{j=1}^{3} P(Y = y_j) = 1 \tag{6.12}$$

和

条件概率是指在某一行或某一列中，在某一列中所占的比例。

某个单元。例如，鉴于X，Y的条件概率是

$$P(Y = y_j \mid X = x_i) = \frac{n_{ij}}{c_i} \tag{6.13}$$

而在Y的情况下，X的条件概率为

$$P(X = x_i \mid Y = y_j) = \frac{n_{ij}}{r_j} \tag{6.14}$$

分类变量

在机器学习中，我们使用离散概率分布来模拟*分类变量*，即采取有限的无序值集的变量。它们可以是分类特征，比如用于预测一个人的工资时，在大学里取得的学位，或者是分类的拉贝尔，比如在进行笔迹识别时，字母的数量。离散分布也经常被用来构建概率模型，这些模型结合了有限数量的连续分布（第11章）。

6.2.2 连续概率

在这一节中，我们考虑实值随机变量，也就是说，我们考虑的目标空间是实线R的区间。在这本书中，我们假装可以对实值随机变量进行操作，就像我们有具有有限状态的dis- crete概率空间一样。然而，这种简化在两种情况下是不精确的：当我们无限频繁地重复某件事时，以及当我们想从一个区间中抽出一个点时。第一种情况是在我们讨论机器学习中的泛化错误时出现的（第8章）。第二种情况是在我们想讨论连续分布时出现的，比如高斯（Section6.5).就我们的目的而言，由于缺乏精确性，我们可以对概率进行更简单的介绍。

*备注。*在连续空间中，有两个额外的技术性问题，它们是反直觉的。 首先，所有子集的集合（用于定义第6.1节中 的事件 AA空间）不够乖巧， 需要限制其在集合补足、集合相交和集合联合的情况下表现良好。第二，一个集合的大小（在离散空间中可以通过计算元素得到）被证明是棘手的。一个集合的大小被称为它的*度量*。例如，离散集合的cardinality，R中的一个区间的长度，以及R中^{d}一个区域的体积都是度量。在集合运算下表现良好并具有拓扑结构的集合被称为*Borel σ-algebra*。Betancourt详细介绍了从集合论出发对概率空间的仔细构造，而没有陷入技术性的困境；见https://tinyurl.com/yb3t6mfd。对于更明确的构造，我们参考Billingsley(1995)和Jacod and Protter(2004)。

衡量

Borel σ-代数

在本书中，我们考虑的是实值随机变量与它们的关系。

响应的Borel σ-代数。我们考虑的随机变量的值在 $\mathbb{R}^D$是一个实值随机变量 的向量。 ◆

定义6.1（概率密度函数）。一个函数$f:\mathbb{R}^D \to \mathbb{R}$ 是称为*概率密度函数*（*pdf*），如果

概率密度函数为pdf

1. $\forall \boldsymbol{x} \in \mathbb{R}^D: f(\boldsymbol{x}) \geqslant 0$
2. 它的整体存在和

$$\int_{\mathbb{R}^D} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1. \tag{6.15}$$

编码

对于离散随机变量的概率质量函数（pmf），(6.15)中的积分被替换为一个总和(6.12).

请注意，概率密度函数是任何非负的且积分为1的函数f。我们将随机变量X与这个函数f联系起来，即

$$P(a \leqslant X \leqslant b) = \int_a^b f(x)\mathrm{d}x, \tag{6.16}$$

其中，a、$b\in\mathbb{R}$和$x\in\mathbb{R}$是连续随机变量X的结果，状态$\boldsymbol{x}\in\mathbb{R}^D$的定义类似于考虑一个矢量
的$x\in\mathbb{R}$。这个关联(6.16)被称为*法则*或*分布*的法则
随机变量X。

$P(X=x)$是一组措施为零。

*备注。*与离散随机变量相反，连续随机变量X取一个特定值$P(X=x)$的概率为零。这就像试图指定一个在(6.16)，其中$a=b$。

定义（6.2累积分布函数）。*累积分布函数。*

分布函数

状态为$\boldsymbol{x}\in\mathbb{R}^D$的多变量实值随机变量$X$的cdf*函数*由以下公式给出

$$F_X(\boldsymbol{x}) = P(X_1 \leqslant x_1, \ldots, X_D \leqslant x_D), \tag{6.17}$$

其中$X=[X_1, \ldots, X]_D^\top$，$\boldsymbol{x}=[x_1, \ldots, x_D]^\top$，右手边代表随机变量$X_i$取值小于或等于$x$的$_i$概率。

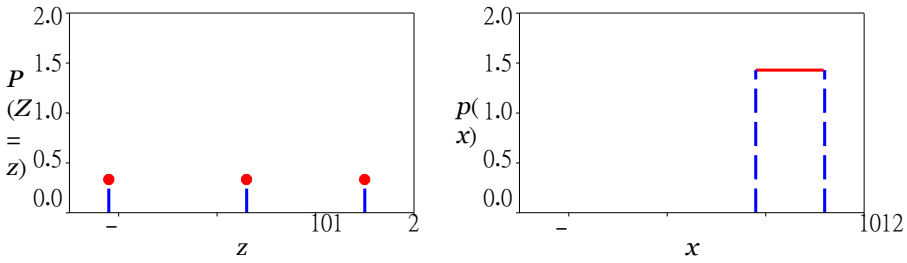
cdf也可以表示为概率密度函数$f(\boldsymbol{x})$的积分，因此

有的cdfs，并没有相应的pdfs。

$$F_X(\boldsymbol{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_D} f(z_1, \ldots, z_D)\mathrm{d}z_1 \cdots \mathrm{d}z_D. \tag{6.18}$$

*备注。*我们重申，在谈论分布的时候，实际上有两个不同的概念。首先是pdf的概念（用$f(x)$表示），它是一个和为1的非负函数。其次是一

随机变量X，也就是随机变量X与pdf$f(x)$的关联。 ◆



(a) 离散分布

(b) 连续分布

图为(a)离散分布和(b)连续均匀分布的例子**6.3**。见例6.3的细节。

分发。

在本书的大部分内容中，我们将不使用$f(x)$和$F(xx)$的符号，因为我们大多不需要区分pdf和cdf。然而，我们需要在第1节中对pdf和cdfs加以注意。6.7.

6.2.3 离散和连续分布的对比

回顾一下6.1.2概率是正数，总概率加起来是1。对于离散随机变量（见(6.12)，这意味着每个状态的概率必须位于区间[0, 1]内。］然而，对于连续随机变量，归一化（见(6.15)并不意味着密度的值对于1所有的值都小于或等于。我们在图中说明了这一点6.3使用离散和连续随机变量的*均匀分布*来说明。

均匀分布

例子 6.3

我们考虑两个均匀分布的例子，其中每个状态发生的可能性相同。这个例子说明了离散和连续概率分布之间的一些区别。

设Z是一个离散的均匀随机变量，有三个状态$\{z = -1.1, z = 0.3, z = 1.5\}$。概率质量函数可以表示为作为一个概率值的表格。

z	-1.1	0.3	1.5
$P(Z = z)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

这些国家并不是的实际值。在这里有意义，我们特意选择数字来促使我们明白，我们不想使用（而且应该忽略）国家的排序。

另外，我们也可以把它看成一个图形，在这里我们使用这样一个事实，即状态可以位于x轴上，而y轴代表一个特定状态的概率。图6.3(a)中的Y轴被故意延长，因此它与图6.3(b)中的相同。

设X是一个连续随机变量，取值范围为$0.9 \leqslant X \leqslant 1.6$，如图6.3(b)所示。观察一下，图中的高度是

类型"点	概率""区间概率"
DiscreteP $(X = x)$ 不适用 概率质量函数	
连续	*sp*(*x*)*P* $(X$ **x** $x)$ 概率密度函数累积分布函数

表 概率分布的命名**6.1** 法。

密度可以大于.1然而，需要坚持的是

$$\int_{0.9}^{1.6} p(x)dx = 1 \tag{6.19}$$

*备注。关于*离散概率分布，还有一个微妙的问题。状态$z_1, \ldots, z_d$原则上没有任何结构，也就是说，通常没有办法对它们进行比较，例如，z_1=红色，z_2=绿色，z_3=蓝色。然而，在许多机器学习的应用中，离散状态采取数值，例如，z_1=1 .1在这里我们可以说$z_{231} < z_2 < z_{153}$。离散状态的数值特别有用，因为我们经常考虑随机变量的期望值（6.4.1节）。

◆

不幸的是，机器学习文献使用的符号和术语隐藏了样本空间Ω、目标空间T和随机变量X之间的区别。对于一个可能的值*x*
我们认为，随机变量*X*的结果，即*x*∈T，*p*(*x*)表示概率
对于离散随机变量，这被写成*P*（*X*=*x*），这被称为概率质量函数。pmf通常被称为 "分布"。对于连续变量，*p*(*x*)被称为概率密度函数（通常被称为密度）。为了进一步混淆视听，累积分布函数*P*(*X*x)也经常被称为 "分布"。
tion"。在本章中，我们将使用*X*这个符号来指代单变量的
和多变量随机变量，并分别用*x*和***x表示***这些状态。我们在表中总结了这些命名方法6.1.

。

我们认为概率论是对逻辑推理的一种延伸。正如我们在第1节中6.1.1,所讨论的，这里提出的概率规则如下

*备注。*我们不仅对离散的概率质量函数使用 "概率分布 "这一表述，而且对连续的概率密度函数也使用这一表述，尽管这在技术上并不正确。根据
在大多数机器学习文献中，我们也依靠上下文来区分概率分布这一短语的不同用途。 ◆

6.3 总和规则、乘积规则和贝叶斯定理

结果x是导致概率$p(x)$的参数。

自然而然地满足了这些要求（Jaynes,2003，第2章）。概率建模（第8.4节）为设计机器学习方法提供了一个原则性的基础。一旦我们定义了与数据和问题的不确定性相对应的概率分布（第6.2节），就会发现只有两条基本规则，即和规则和乘规则。

回顾一下(6.9分布$p(\boldsymbol{x})$和$p(\boldsymbol{y})$是对应的边际分布，而$p(\boldsymbol{y} \mid \boldsymbol{x})$是给定$\boldsymbol{x}$的$\boldsymbol{y}$的条件分布。

介绍概率论中的两条基本规则。

第一条规则，即*总和规则*，指出

这两条规则的产生自然（杰恩斯。2003年），从我们在章节中讨论的要求来看6.1.1.

总和规则

$$p(\boldsymbol{x}) = \begin{cases} \sum_{\boldsymbol{y} \in \mathcal{Y}} p(\boldsymbol{x}, \boldsymbol{y}) & \text{，如果} \boldsymbol{y} \text{是离散的} \\ \int_{\mathcal{Y}} p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} & \text{如果} \boldsymbol{y} \text{是连续的} \end{cases}, \tag{6.20}$$

其中$\mathcal{Y}$是随机变量Y的目标空间的状态。这意味着我们对随机变量Y的状态集进行求和（或积分）。求和规则也被称为*边际化属性*。

边缘化属性

总和规则将联合分布与边际分布联系起来。一般来说，当联合分布包含两个以上的随机变量时，总和规则可以应用于随机变量的任何子集，从而产生一个可能超过一个的随机变量的边际分布。更具体地说，如果$\boldsymbol{x}=[x_1, \ldots, x]_D^{\mathrm{T}}$，我们可以得到边际的

$$p(x_i) = \int p(x_1, \ldots, x)_D d\boldsymbol{x}_{\setminus i} \tag{6.21}$$

通过反复应用求和法则，我们把所有的随机变量整合/求和出来，除了x_i，用i表示，读作"除i之外的所有变量"。

*备注。*概率建模的许多计算挑战都是由于应用了求和规则。当有许多变量或有许多状态的离散变量时，求和规则归结为形成一个高维的和或积分。执行高维求和或积分通常在计算上是困难的，在这个意义上，有没有已知的多项式时间算法来精确计算它们。◆

产品规则

第二条规则，被称为*乘积规则*，通过以下方式将联合分布与条件分布联系起来

$(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x})$ 。 (6.22)

乘积规则可以解释为：两个随机变量的每个联合分布都可以被分解（写成乘积）。

的两个其他分布。这两个因素是第一个随机变量$p(\boldsymbol{x})$的边际分布，以及第二个随机变量在第一个$p(\boldsymbol{y} \mid \boldsymbol{x})$下的条件分布。由于随机变量的排序在$p(\boldsymbol{x}, \boldsymbol{y})$中是任意的，乘积规则也意味着$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x} \mid \boldsymbol{y})p(\boldsymbol{y})$。准确地说，(6.22)是用离散随机变量的概率质量函数来表示的。对于连续随机变量，乘积规则是用概率密度函数表示的（第6.2.3).

在机器学习和贝叶斯统计学中，我们经常对在观察到其他随机变量的情况下对未观察到的（潜在的）随机变量进行推断感兴趣。让我们假设我们有一些关于未观察到的随机变量$\boldsymbol{x}$的先验知识$p(\boldsymbol{x})$，以及$\boldsymbol{x}$和第二个随机变量$\boldsymbol{y}$之间的一些关系$p(\boldsymbol{y} \mid \boldsymbol{x})$，我们可以观察到这些关系。如果我们观察到$\boldsymbol{y}$，我们可以使用贝叶斯定理来得出一些
鉴于$\boldsymbol{y}$的观察值，关于$\boldsymbol{x}$的结论。*贝叶斯定理*（也是Bayes's theorem
*贝叶斯规则*或*贝叶斯法则*）

贝叶斯规则

贝叶斯法则

$$\underbrace{p(\boldsymbol{x} \mid \boldsymbol{y})}_{\text{姿势}} = \frac{\overbrace{p(\boldsymbol{y} \mid \boldsymbol{x})}^{\text{可能性}} \overbrace{p(\boldsymbol{x})}^{\text{先验}}}{\underbrace{p(\boldsymbol{y})}_{\text{识别能力}}} \tag{6.23}$$

是( )中乘积规则的直接结果。6.22)中的乘积规则的直接结果，因为

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x} \mid \boldsymbol{y})p(\boldsymbol{y}) \tag{6.24}$$

和

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x}) \tag{6.25}$$

以致于

$$p(\boldsymbol{x} \mid \boldsymbol{y})p(\boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x}) \iff p(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})}。 \tag{6.26}$$

在(6.23)中，$p(\boldsymbol{x})$是*先验*，它概括了我们的主观先验
在观察任何数据之前，我们需要对未观察到的（潜在的）变量$\boldsymbol{x}$有所了解。我们可以选择任何对我们有意义的先验，但关键是要确保先验在所有可信的$\boldsymbol{x}$上有一个非零的pdf（或pmf），即使它们非常罕见。

可能性$p(\boldsymbol{y} \mid \boldsymbol{x})$描述了$\boldsymbol{x}$和$\boldsymbol{y}$的关系，在可能性中
在离散概率分布的情况下，它是数据$\boldsymbol{y}$的概率

这种可能性有时也被称为 "测量模型"。

注意，可能性不是在$\boldsymbol{x}$中的分布，而只是在$\boldsymbol{y}$中的分布。我们称$p(\boldsymbol{y} \mid \boldsymbol{x})$为 "$\boldsymbol{x}$的可能性（给定$\boldsymbol{y}$）"或 "给定$\boldsymbol{x}$的$\boldsymbol{y}$的概率"，但绝不是$\boldsymbol{y}$的可能性（MacKay,2003）。

后验$p(\boldsymbol{x} \mid \boldsymbol{y})$是贝叶斯统计学

中感兴趣的数量。

因为它准确地表达了我们感兴趣的东西，即我们在观察了$\boldsymbol{y}$之后对$\boldsymbol{x}$的了

解。 188 概率和分布

数量

$$p(\boldsymbol{y}) := \int p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} = \mathbb{E}_{\boldsymbol{x}}[p(\boldsymbol{y} \mid \boldsymbol{x})] \tag{6.27}$$

边际可能性

是*边际可能性/证据*。(的右手边是(6.27)使用的是

6.4.1.根据定义，边际似然将( )的分子与潜在变量x进行整合。6.23因此，边际似然是独立于$\boldsymbol{x}$的，它确保了后验$p(\boldsymbol{x}\ \boldsymbol{y})$是正常化的。边际似然也可以被解释为期望似然，在这里我们取相对于先验$p(\boldsymbol{x})$的期望。除了后验的正常化，边际似然在贝叶斯模型选择中也起着重要作用，我们将在第8.6节中讨论。由于(8.44)中的整合，证据往往很难计算。

贝叶斯定理也被称为 "概率逆向"。 概率逆向

贝叶斯定理(6.23)使我们能够反转由可能性给出的$\boldsymbol{x}$和$\boldsymbol{y}$之间的关系。因此，贝叶斯定理有时被称为*概率逆向*。我们将在第8.4节进一步讨论贝叶斯定理。

*备注。*在贝叶斯统计学中，后验分布是感兴趣的数量，因为它囊括了来自先验和数据的所有可用信息。与其到处携带后验，不如关注后验的某些统计量，例如后验的最大值，我们将在第8.3节中讨论。然而，关注后验的某些统计数字会导致信息的损失。如果我们从更大的角度考虑，那么后验可以在决策系统中使用，拥有完整的后验可能是非常有用的，并导致决策对干扰的稳健。例如，在基于模型的再执行学习的背景下，Deisenroth等人（2015）表明，使用可信过渡函数的完整后验分布会导致非常快速（数据/样本效率）的学习，而关注后验的最大值会导致一致的失败。因此，拥有完整的后验可以对下游的任务非常有用。在第九章中，我们将
继续在线性回归 的背景
下讨论这个问题。 ◆

6.4 简要统计和独立性

我们经常对总结随机变量的集合和随机变量对的组合感兴趣。一个随机变量的统计量是该随机变量的一个终结性函数。分布的汇总统计提供了随机变量行为方式的一个有用的观点，正如其名称所示，提供了概括和描述分布的数字。我们描述了平均数和方差这两个众所周知的汇总统计量。然后我们讨论比较一对随机变量的两种方法：第一，如何说两个随机变量是独立的；第二，如何计算它们之间的内积。

6.4.1 平均数和协方差

平均数和（共）方差通常用于描述概率分布的属性（期望值和扩散）。我们将在本节中看到6.6有一个有用的分布族（称为指数族），其中随机变量的统计量涵盖了所有可能的信息。

期望值的概念是机器学习的核心，而概率的基础概念本身也可以从期望值中推导出来（Whittle,2000）。

定义（6.3期望值）。一个函数g的*期望值*。R→ 预期值
R of a univariate continuous random variable $X \sim p(x)$ is given by

$$\mathsf{E}_X[g(x)] = \int_{\mathsf{X}} g(x)p(x)\,dx。 \tag{6.28}$$

Correspondingly, the expected value of a function g of a discrete random variable $X \sim p(x)$ is given by

$$\mathsf{E}_X[g(x)] = \sum_{x\in\mathsf{X}} g(x)p(x)\,, \tag{6.29}$$

其中 X 是随机变量X的可能结果集（目标空间）。

在这一节中，我们认为离散的随机变量具有数字结果。这可以通过观察函数g取实数的情况看出
数字作为输入。

预期值的函数，有时被称为无意识统计学家的法则（Casella and Berger,2002, Section 2.2）。

*备注。*我们将多变量随机变量X视为单变量随机变量的有限向量$[X_1, \ldots, X]_D^{\mathrm{T}}$。对于多变量随机变量，我们定义期望值元素wise

$$\mathsf{E}_X[g(\boldsymbol{x})] = \begin{bmatrix} \mathsf{E}_{X_1}[g(x_1)] \\ \vdots \\ \mathsf{E}_{X_D}[g(x_D)] \end{bmatrix} \in \mathsf{R}^D, \tag{6.30}$$

其中，下标E_X表示我们要取的是预期值
相对于矢量$\boldsymbol{x}$的第d个元素。 ◆

定义6.3定义了符号E的$_X$含义，即表示我们应该对概率密度（对于连续分布）或所有状态的总和（对于离散分布）进行积分。平均值的定义（定义6.4），是期望值的一个特例，通过选择g为特性函数而得到。

定义（6.4平均值）。一个随机变量X的*平均值*，其状态为 mean

$\boldsymbol{x} \in \mathsf{R}^D$是一个平均值，定义为

$$\mathsf{E}_X[\boldsymbol{x}] = \begin{bmatrix} \mathsf{E}_{X_1}[x_1] \\ \vdots \\ \mathsf{E}_{X_D}[x_D] \end{bmatrix} \in \mathsf{R}^D, \tag{6.31}$$

其中

$$\mathsf{E}_{X_d}[x_d] := \begin{cases} \int_{\mathsf{X}} x_d p(x_d) dx_d & \text{，如果} X \text{是一个连续随机变量} \\ \sum_{x_i \in \mathsf{X}} x_i p(x_d = x_i) & \text{，如果} X \text{是一个离散的随机变量} \end{cases} \tag{6.32}$$

对于$d = 1, \ldots, D$，其中下标d表示$\boldsymbol{x}$的相应二分位数。

在一个维度上，还有两个直观的 "平均 "概念，即*中位数*和*模式*。如果我们对数值进行排序，即50%的数值大于中位数，50%的数值小于中位数，那么*中位数*就是 "中间 "值。对于不对称或有长尾的分布，中位数提供了一个典型值的估计，比均值更接近人类的直觉。此外，中位数比平均值对异常值更加稳健。将中位数推广到更高的维度是不容易的，因为没有明显的方法在一个以上的维度中 "排序"（Hallin等人，2010；Kong和Mizera，2012）。*模式*是最经常出现的值。对于离散随机变量，模式被定义为具有最高出现频率的x值。对于连续随机变量，模式被定义为密度$p(\boldsymbol{x})$中的一个峰值。一个特定的密度$p(\boldsymbol{x})$可能有一个以上的模式，此外，在高维分布中可能有非常多的模式。因此，要找到一个分布的所有模式，在计算上是一个挑战。

中位数

模式

例子 6.4

考虑到图6.4中所示的二维分布。

$$p(\boldsymbol{x}) = 0.4\, \mathcal{N}_{\boldsymbol{x}}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}\right) + 0.6\, \mathcal{N}_{\boldsymbol{x}}\left(\begin{bmatrix} 4 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.8 & 2.0 \\ 0.2 & 1.7 \end{bmatrix}\right). \tag{6.33}$$

我们将在第一节中定义高斯分布$\mathcal{N}(\mu, \sigma^2)$。图6.5.同时显示的是它在每个维度上的相应边际分布。发现该分布是二元的（有两个模式），但其中一个是

边际分布是单模的（有一个模式）。水平双模单变量分布说明了平均数和中位数可以彼此不同。虽然很想把二维中位数定义为每个二元分布中的中位数之和，但我们无法定义二维点的排序，这就很难了。当我们说 "不能定义一个排序 "时，我们
意味着有不止一种方式来定义关系<，以便

$$\begin{matrix}3\\0\end{matrix} \;<.\; \begin{matrix}2\\3\end{matrix}$$

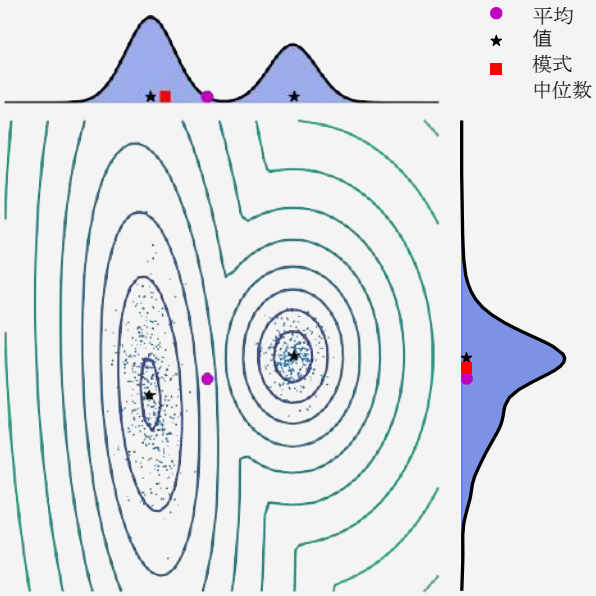


Figure 6.4 Illustration of the mean, mode, and median for a 二维数据集，以及其边际密度。

备注。 预期值（定义6.3）是一个线性算子。例如，给定一个实值函数f（$\boldsymbol{x}$）$=ag$（$\boldsymbol{x}$）$+bh$（$\boldsymbol{x}$），其中a，$b\in\mathbb{R}$，$\boldsymbol{x}\in\mathbb{R}^D$，我们可以得到

$$\mathbb{E}_X[f(\boldsymbol{x})] = f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \tag{6.34a}$$

$$= [ag(\boldsymbol{x}) + bh(\boldsymbol{x})]p(\boldsymbol{x})d\boldsymbol{x} \tag{6.34b}$$

$$= a\int g(\boldsymbol{x})p(\boldsymbol{x})dx + b\int h(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \tag{6.34c}$$

$$= a\mathbb{E}_X[g(\boldsymbol{x})] + b\mathbb{E}_X[h(\boldsymbol{x})]。 \tag{6.34d}$$

◆

对于两个随机变量，我们可能希望描述它们的对应关系。

彼此之间的关联性。协方差直观地代表了随机变量彼此之间的依赖程度的概念。

协方差

定义（6.5协方差（单变量））。两个单变量X、Y $\in \mathbb{R}$之间的*协方差*是由它们偏离各自平均值的预期乘积给出的，即。

$$\mathrm{Cov}_{X,Y}[x, y] := \mathbb{E}_{X,Y}[(x - \mathbb{E}_X[x])(y - \mathbb{E}_Y[y])] 。 \tag{6.35}$$

术语。多变量随机变量的协方差$\mathrm{Cov}[x, y]$有时被称为交叉协方差，协方差指的是$\mathrm{Cov}[x, x]$。

*备注。*当与期望或协方差相关的随机变量由其参数明确时，下标往往被压制（例如，$\mathbb{E}_X[x]$往往被写成$\mathbb{E}[x]$）。 ◆

通过使用期望的线性，定义中的表达式6.5可以改写为产品的期望值减去期望值的乘积，即。

$$\mathrm{Cov}[x, y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] 。 \tag{6.36}$$

差异性

标准差

变量与自身$\mathrm{Cov}[x, x]$的协方差称为*方差*，用$\mathbb{V}_X[x]$表示。方差的平方根称为*标准差*，通常用$\sigma(x)$表示。协方差的概念可以被推广到多变量随机变量。

协方差

定义6.6（协方差（多变量））。如果我们考虑两个状态分别为$\boldsymbol{x} \in \mathbb{R}^D$和$\boldsymbol{y} \in \mathbb{R}^E$的多变量$X$和$Y$，$X$和$Y$之间的*协方差*被定义为

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] = \mathbb{E}[\boldsymbol{x}\boldsymbol{y}^\top] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{y}]^\top = \mathrm{Cov}[\boldsymbol{y}, \boldsymbol{x}]^\top \in \mathbb{R}^{D\times E}。 \tag{6.37}$$

定义6.6可以在两个参数中应用相同的多变量随机变量，这就产生了一个有用的概念，可以直观地捕捉到随机变量的"传播"。对于一个多变量随机变量，方差描述了随机变量各个部分之间的关系。

差异性

定义6.7（方差）。具有状态$\boldsymbol{x} \in \mathbb{R}^D$和均值向量$\boldsymbol{\mu} \in \mathbb{R}^D$的随机变量$X$的*方差*被定义为

$$\mathbb{V}_X[\boldsymbol{x}] = \mathrm{Cov}_X[\boldsymbol{x}, \boldsymbol{x}] \tag{6.38a}$$

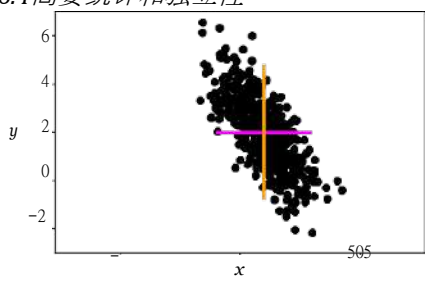
$$= \mathbb{E}_X[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top] = \mathbb{E}_X[\boldsymbol{x}\boldsymbol{x}^\top] - \mathbb{E}_X[\boldsymbol{x}]\mathbb{E}_X[\boldsymbol{x}]^\top \tag{6.38b}$$

$$= \begin{bmatrix} \mathrm{Cov}[x_1, x_1] & \mathrm{Cov}[x_1, x_2] & \dots & \mathrm{Cov}[x_1, x_D] \\ \mathrm{Cov}[x_2, x_1] & \mathrm{Cov}[x_2, x_2] & \dots & \mathrm{Cov}[x_2, x_D] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[x_D, x_1] & \dots & & \mathrm{Cov}[x_D, x_D] \end{bmatrix}。 \tag{6.38c}$$

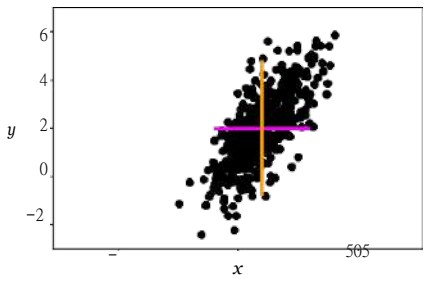
协方差矩阵

(6.38c)中的D D矩阵被称为多变量随机变量X的*协方差矩阵*。协方差矩阵是对称的和正半无限的，它告诉我们一些关于数据传播的信息。在其对角线上，协方差矩阵包含了*边际*变量的方差

边际



(a) x和y是负相关的



。(b)x和y是正相关的。

图 6.5 二维数据集，每个轴上的平均值和方差相同（彩色线条），但协方差不同。

$$p(x_i) = \int p(x_1, \ldots, x_D)\, dx_{\setminus i}\,, \tag{6.39}$$

其中"$\setminus i$"表示"除i外的所有变量"。对角线外的条目是
*交叉方差*项$\mathrm{Cov}[x_i, x_j]$对于$i, j = 1, \ldots, D$，$i \neq j$.

交叉协方差

*备注。*在本书中，我们一般假设协方差矩阵是正定的，以便能够更好地理解。因此，我们不讨论导致正半定（低秩）协方差矩阵的角落情况。

三角形。 ◆

当我们想比较不同对随机变量之间的协方差时，事实证明，每个随机变量的方差会影响协方差的值。协方差的归一化版本被称为相关度。

定义（6.8相关性）。两个随机变量
ables X, Y is given by

之间的*相关性*。

$$\text{珊瑚}[x, y] = \frac{\mathrm{Cov}[x, y]}{\sqrt{\mathrm{V}[x]\mathrm{V}[y]}} \in [-1, 1]\,。 \tag{6.40}$$

相关矩阵是标准化随机变量的协方差矩阵，x/σ（x）。换句话说，每个随机变量在相关矩阵中都被其标准差（方差的平方根）所除。

协方差（和相关）表示两个随机变量的关系；见图6.5。正相关corr $[x, y]$意味着当x增长时，预计y也会增长。负相关意味着当x*增长时*，那么y*就会减少*。

6.4.2 经验平均数和协方差

节中的定义6.4.1中的定义通常也被称为*种群平均数*
和协方差，因为它指的是人口的真实统计数据。在机器学习中，我们需要从数据的经验观察中学习。考虑到一个随机变量X，有两个概念步骤可以从

population mean
和协方差

从人口统计学到经验统计学的实现。首先，我们利用我们有一个有限的数据集（大小为N）这一事实来构造一个经验统计量，它是有限数量的相同随机变量$X_1, ., X_N$.第二，我们观察数据，也就是说，我们观察每个随机变量的实现情况x_1， ... ...X，每个随机变量$_N$的实现，并应用经验统计。

经验平均数 样本平均数

具体来说，对于平均数（定义6.4），给定一个特定的数据集，我们可以得到一个平均值的估计值，这被称为*经验平均值*或*样本平均值*。经验协方差也是如此。

经验平均数

定义6.9（经验平均值和协方差）。*经验*平均数是每个变量的观察值的算术平均数，它被定义为

$$\bar{\boldsymbol{x}} := \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n , \quad (6.41)$$

其中$\boldsymbol{x}_n \in \mathrm{R}^D$。

经验协方差

与经验平均值相似，*经验协方差*矩阵是一个D D $\times$基体

$$\boldsymbol{\Sigma} := \frac{1}{N}\sum_{n=1}^{N} (\boldsymbol{x}_n \bar{\boldsymbol{x}}) \quad (\quad \bar{\boldsymbol{x}}^{\mathrm{T}}). \quad (6.42)$$

在整个

在书中，我们使用经验协方差，这是一个有偏见的估计。无偏的（有时称为修正的）协方差在分母中含有因子N 1，而不是N。

为了计算一个特定数据集的统计数据，我们将使用实现（观测值）$\boldsymbol{x}_1, . ., \boldsymbol{x}, _N$并使用(6.41)和(6.42).经验协方差矩阵是对称的，正半无限的（见3.2.3节）。

本章末尾有推导的练习。

6.4.3 差异的三种表达方式

我们现在关注一个单一的随机变量X，并使用前面的仿真公式推导出方差的三种可能表达方式。下面的推导对人口方差来说是一样的，只是我们需要照顾到积分的问题。方差的标准定义，与协方差的定义相呼应（定义6.5），是随机变量X与它的期望值μ的平方偏差的表达，即。

$$\mathrm{V}[xx] := \mathrm{E}_X[(x - \mu)^2] 。 \quad (6.43)$$

中的期望值(6.43)和平均数μ=E(xx)*的*计算方法是用(6.32)，取决于X是离散的还是连续的随机变量。用( )表示的方差是新的随机变量Z :=(x)的平均数。6.43)表示的方差是一个新的随机变量$Z := (X \mu)$的平均值^2。

当估计(6.43)中的方差时，我们需要采用一种两遍的算法：一遍通过数据，用(6.41)，然后用这个估计值μ*进行*第二遍， 计算出

差异。事实证明，我们可以通过重新排列的方式避免两遍

条款。中的公式(6.43)中的公式可以转换为所谓的*原始分数(raw-*
差异的公式。

scoraw-score)公式
差异性

$$\mathsf{V}[x] = \mathsf{E}[x^2] - (\mathsf{E}[x])^2 。 \tag{6.44}$$

中的表达式(6.44)中的表达式可以被记为 "均值的平方减去均值的平方"。
它可以根据经验一次性计算出数据，因为我们可以把x_i（用来计算平均值
）和x^2

同时，其中x_i是第i个观察值。不幸的是，如果实施- 如果这两个条款
以这种方式管理，它在数字上可能是不稳定的。方差的原始分数版本在
机器学习中很有用，例如，在推导偏置方差分解时（Bishop，2006）。

在(6.44)巨大且近似
相等，我们可能会
在浮点运算中遭受
不必要的数字精度
损失。

第三种理解方差的方法是，它是所有成对观察值之间的成对差异的总和。考虑随机变量X的样本$x_1, \ldots$.我们计算x_{Ni}和x的$_j$成对差异的平方。即N个配对^{2}差异的总和是观测值的经验方差。

$$\frac{1}{N^2}\sum_{i,j=1}^{N}(x_i - x_j)^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2 . \tag{6.45}$$

我们看到，(6.45)是原始分数表达式(6.44).这意味着我们可以将成对距离的总和（其中有N个^2）*表示*为偏离平均值的总和（其中有N个）。从几何学的角度看，这意味着成对距离和从点集合的中心出发的距离之间存在着等价关系。从计算的角度看，这意味着通过计算平均数（求和中的N项），然后计算方差（同样是求和中的N项），我们可以得到一个有N项的表达式（( )的左手边。6.45)），它有N个项^2。

6.4.4 随机变量的总和和变换

我们可能想建立一个无法用教科书上的分布来很好解释的现象的模型（我们在第6.6节中介绍了一些6.5和6.6节），因此可以对随机变量进行简单的操作（如增加两个随机变量）。

考虑两个随机变量X、Y的状态$\boldsymbol{x}$、$\boldsymbol{y}\in\mathsf{R}^D$，那么。

$$\mathsf{E}[\boldsymbol{x}+\boldsymbol{y}] = \mathsf{E}[\boldsymbol{x}] + \mathsf{E}[\boldsymbol{y}] \tag{6.46}$$
$$\mathsf{E}[\boldsymbol{x}-\boldsymbol{y}] = \mathsf{E}[\boldsymbol{x}] - \mathsf{E}[\boldsymbol{y}] \tag{6.47}$$
$$\mathsf{V}[\boldsymbol{x}+\boldsymbol{y}] = \mathsf{V}[\boldsymbol{x}] + \mathsf{V}[\boldsymbol{y}] + \mathrm{Cov}[\boldsymbol{x},\boldsymbol{y}] + \mathrm{Cov}[\boldsymbol{y},\boldsymbol{x}] \tag{6.48}$$
$$\mathsf{V}[\boldsymbol{x}-\boldsymbol{y}] = \mathsf{V}[\boldsymbol{x}] + \mathsf{V}[\boldsymbol{y}] - \mathrm{Cov}[\boldsymbol{x},\boldsymbol{y}] - \mathrm{Cov}[\boldsymbol{y},\boldsymbol{x}] 。 \tag{6.49}$$

当涉及到随机变量的仿生变换时，平均数和（共）方差表现出一些有用的特性。考虑一个随机变量
X的均值为$\boldsymbol{\mu}$，协方差矩阵为$\boldsymbol{\Sigma}$，$\boldsymbol{X}$的（确定性）仿生变换$\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}$，那么$\boldsymbol{y}$本身就是一个随机变量
其均值向量和协方差矩阵由以下公式给出

$$\mathsf{E}_Y[\boldsymbol{y}] = \mathsf{E}_X[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}] = \mathsf{A}\mathsf{E}_X[\boldsymbol{x}] + \boldsymbol{b} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}\,, \tag{6.50}$$

$$\mathsf{V}_Y[\boldsymbol{y}] = \mathsf{V}_X[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}] = \mathsf{V}_X[\boldsymbol{A}\boldsymbol{x}] = \mathsf{A}\mathsf{V}_X[\boldsymbol{x}]\boldsymbol{A}^{\mathsf{T}} = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathsf{T}}, \tag{6.51}$$

这可以通过使用平均值和协方差的定义直接显示出来。

分别。此外。

$$\begin{aligned}\mathsf{Cov}[\boldsymbol{x}, \boldsymbol{y}] &= \mathsf{E}[\boldsymbol{x}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})^{\mathsf{T}}] - \mathsf{E}[\boldsymbol{x}]\mathsf{E}[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}]^{\mathsf{T}} && (6.52a)\\ &= \mathsf{E}[\boldsymbol{x}]\boldsymbol{b}^{\mathsf{T}} + \mathsf{E}[\boldsymbol{x}\boldsymbol{x}]^{\mathsf{T}}\boldsymbol{A}^{\mathsf{T}} - \boldsymbol{\mu}\boldsymbol{b}^{\mathsf{T}} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{A}^{\mathsf{T}} && (6.52b)\text{。}\\ &= \boldsymbol{\mu}\boldsymbol{b}^{\mathsf{T}} - \boldsymbol{\mu}\boldsymbol{b}^{\mathsf{T}} + \left(\mathsf{E}[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}] - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}\right)\boldsymbol{A}^{\mathsf{T}} && (6.52c)\\ &\overset{(6.38b)}{=} \boldsymbol{\Sigma}\boldsymbol{A}^{\mathsf{T}}, && (6.52d)\end{aligned}$$

其中$\boldsymbol{\Sigma}=\mathsf{E}[\boldsymbol{x}\boldsymbol{x}]^{\mathsf{T}}-\boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}$是$X$的协方差。

6.4.5 统计学的独立性

统计

定义（6.10独立）。两个随机变量X、Y是*统计学上的*

当且仅当

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x})\,p(\boldsymbol{y})\text{。} \tag{6.53}$$

直观地说，如果$\boldsymbol{y}$的值（一旦知道）不会增加关于$\boldsymbol{x}$的任何额外信息（反之亦然），两个随机变量X和Y是独立的。如果X、Y是（统计学上）独立的，那么

- $p(\boldsymbol{y}|\boldsymbol{x}) = p(\boldsymbol{y})$
- $p(\boldsymbol{x}|\boldsymbol{y}) = p(\boldsymbol{x})$
- $\mathsf{V}_{X,Y}[\boldsymbol{x} + \boldsymbol{y}] = \mathsf{V}_X[\boldsymbol{x}] + \mathsf{V}_Y[\boldsymbol{y}]$
- $\mathsf{Cov}[_{X,Y}\boldsymbol{x}, \boldsymbol{y}] = \boldsymbol{0}$

最后一点可能反过来不成立，也就是说，两个随机变量的协方差可以为零，但在统计上却不独立。为了理解这个原因，回顾一下协方差只衡量线性依赖性。因此，非线性依赖的随机变量可以有协方差为零。

例子 6.5
考虑一个随机变量X，其均值为零（$\mathsf{E}_X[x]=0$），同时$\mathsf{E}_X[x^3]=0$。让$y=x^2$（因此，Y依赖于X）并考虑X和Y之间的协方差(6.36)，考虑X和Y之间的协方差。但这给出了

$$\mathsf{Cov}[x, y] = \mathsf{E}[xy] - \mathsf{E}[x]\mathsf{E}[y] = \mathsf{E}[x^3] = \text{。} \quad 0 \tag{6.54}$$

在机器学习中，我们经常考虑的问题是可以修改的。

称为*独立同分布*(*i.i.d.*)的随机变量，

独立和同分布i.i.d.

$X_1, \ldots, X_N$.对于两个以上的随机变量，"独立 "一词（定义6.10）通常是指相互独立的随机变量，其中所有子集都是独立的（见Pollard（2002，第4章）和Jacod和Protter（2004，第3章））。"相同分布 "这一短语意味着所有的随机变量都来自相同的分布。

机器学习中另一个重要的概念是条件独立。

定义（6.11条件独立）。两个随机变量X

在给定Z*的情况下，*Y是*有条件独立的，*当且仅当有

条件独立时

$$p(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{z}) p(\boldsymbol{y} \mid \boldsymbol{z}) \text{对于所有} \quad \boldsymbol{z} \in \mathcal{Z}, \tag{6.55}$$

其中，$\mathcal{Z}$ 是随机变量Z的状态集。我们用$X \perp Y \mid Z$来表示，在给定Z的情况下，X与Y是有条件独立的。

定义6.11要求(6.55)中的关系对$\boldsymbol{z}$的每个值都必须成立。6.55)的解释可以理解为 "给定关于$\boldsymbol{z}$的知识，$\boldsymbol{x}$和$\boldsymbol{y}$的分布是因子化的"。独立性可以作为条件独立性的一个特例，如果我们把$X \perp Y \mid \emptyset$.通过使用概率的乘积规则(6.22)，我们可以展开(6.55)的左手边展开，得到

$$p(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z}) p(\boldsymbol{y} \mid \boldsymbol{z}) 。 \tag{6.56}$$

通过比较(6.55)与(6.56)，我们看到$p(\boldsymbol{y} \mid \boldsymbol{z})$出现在两者中，所以

$$p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{z}) . \tag{6.57}$$

方程(6.57这个替代性的表述提供了 "鉴于我们知道$\boldsymbol{z}$，关于$\boldsymbol{y}$的知识不会改变我们对$\boldsymbol{x}$的认识 "的相互假装。

6.4.6 随机变量的内积

回顾一下第3.2节中关于内积的定义。我们可以定义一个随机变量之间的内积，我们在本节中简要介绍。如果我们有两个不相关的随机变量X、Y，那么

内积

在下文中，我们看看是否可以找到不相关的随机变量的方差关系的几何解释，在(6.58).

$$\mathbb{V}[x + y] = \mathbb{V}[x] + \mathbb{V}[y] 。 \tag{6.58}$$

由于方差是以平方单位衡量的，这看起来非常像直角三角形的毕达哥拉斯定理$c^2 = a^2 + b^2$。

可以用类似的方式来处理多变量随机变量之间的关系

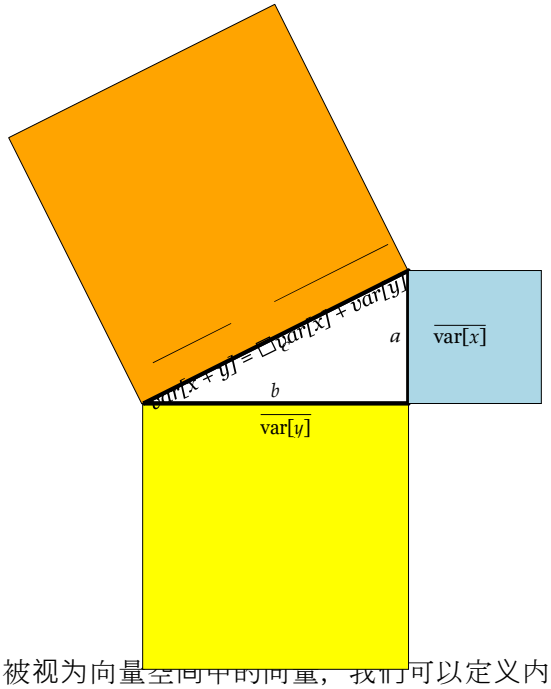


图 随机变量的**6.6**几何学。如果随机变量X和Y是不相关的，它们就是相应矢量空间中的正交矢量，而且勾股定理适用。

随机变量可以被视为向量空间中的向量，我们可以定义内积来获得随机变量的几何属性（Eaton,2007）。如果我们定义

$$(X, Y) := \mathrm{Cov}[x, y] \tag{6.59}$$

对于零平均的随机变量X和Y，我们得到一个内积。我们看到，协方差是对称的、正定的，并且是线性的。

$\mathrm{Cov}[x, x] = 0 \Longleftrightarrow x = 0$

$\mathrm{Cov}[\alpha x + z, y] = \alpha\,\mathrm{Cov}[x, y] + \mathrm{Cov}[z, y]$为$\alpha \in \mathbb{R}$。

参数。一个随机变量的长度是

$$|X| = \sqrt{\mathrm{Cov}[x, x]} = \sqrt{\mathbb{V}[x]} = \sigma[x]\,。 \tag{6.60}$$

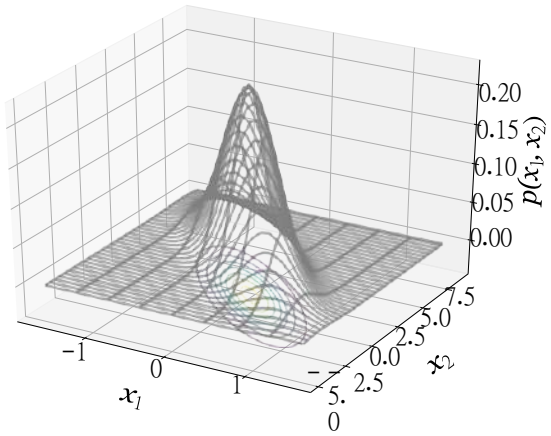
即它的标准偏差。随机变量越"长"，它的不确定性就越大；而一个有长度的随机变量是0确定的。

如果我们看一下两个随机变量X，Y之间的角度θ，我们可以得到

$$\cos\theta = \frac{(X, Y)}{|X|\,|Y|} = \frac{\mathrm{Cov}[x, y]}{\sqrt{\mathbb{V}[x]\mathbb{V}[y]}}, \tag{6.61}$$

这就是两个随机变量之间的相关性（定义6.8）。这意味着，当我们从几何角度考虑两个随机变量时，我们可以把相关性看作是两个随机变量之间的角度的余弦。我们从定义3.7知道，$X \perp Y \Longleftrightarrow (X, Y) = 0$。在我们的例子中，这意味着当且仅当$\mathrm{Cov}[x, y]=0$，即它们不相关时，$X$和$Y$是正交的。图6.6说明了这种关系。

备注。虽然使用欧几里得距离（构建的）是很诱人的。



图为高斯分布**6.7**的两个随机变量x_1和x_2。

从前面的内积定义中可以看出，要比较概率分布，不幸的是，这并不是获得分布之间距离的最佳方法。回顾一下，概率质量（或密度）是正数，需要加起来为1。对这个概率分布空间的研究被称为信息几何学。计算分布之间的距离通常使用Kullback-Leibler分歧，它是距离的概括，考虑了统计流形的属性。就像欧氏距离是公制的一个特例（第3.3节），Kullback-Leibler发散是两类更普遍的发散的特例，称为Bregman发散和f发散。对发散的研究超出了本书的范围，更多细节请参考Amari(2016)最近出版的书，其中一本就是《发散》。

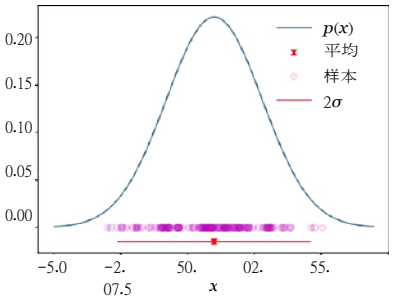
信息几何学 领域的奠基人。 ◆

6.5 高斯分布

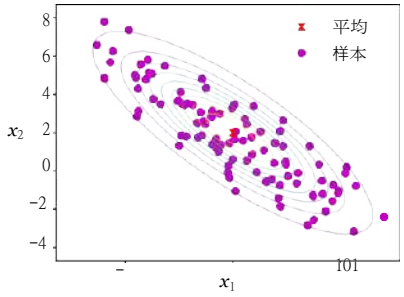
高斯分布是研究得最透彻的概率分布对于连续值的随机变量。它也被称为*正态分布*。它的重要性源于这样一个事实，即它有许多高斯分布。我们将在下文中讨论它在理论上的便利特性。特别是，我们将用它来定义线性回归的似然和先验（第九章），并考虑密度估计的高斯混合物（第十一章）。

当我们考虑独立和相同分布的随机变量的总和时，自然会产生分布。这被称为中心极限定理（Grinstead and Snell,1997）。

机器学习的许多其他领域也得益于高斯分布，例如高斯过程、变异推理和强化学习。它也广泛用于其他应用领域，如信号处理（如卡尔曼滤波）、控制（如线性二次调节器）和统计（如假设检验）。



(a) 单变量（一维）高斯；红叉表示平均值，红线表示方差的范围。



(b) 多变量（二维）Gaus- sian，从顶部看。红色的十字表示平均数，彩色的线表示密度的连线。

图中高斯分布**6.8**与样本100重叠。(a)一维的情况。(b) 二维的情况。

对于一个单变量的随机变量来说，高斯分布的密度是由以下因素决定的

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{6.62}$$

多变量高斯分布平均向量
协方差矩阵

*多变量高斯分布*由*均值向量*$\boldsymbol{\mu}$和*协方差矩阵*$\boldsymbol{\Sigma}$完全表征，定义为

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)。 \tag{6.63}$$

也被称为多变量正态分布。

其中$\boldsymbol{x} \in \mathbb{R}^D$。我们写$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$或$X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$。图6.7显示了一个双变量高斯(网状)，以及相应的锥度

游览图。图中6.8显示了一个单变量的高斯和一个双变量的高斯以及相应的样本。高斯的特殊情况是：零的高斯

标准正常

平均值和同一协方差，即$\boldsymbol{\mu}=\boldsymbol{0}$和$\boldsymbol{\Sigma}=\boldsymbol{I}$，被称为*标准的正态分布*。

高斯分布在统计估计和机器学习中被广泛使用，因为它们有边际和条件分布的闭式表达。在第九章中，我们在线性回归中广泛使用这些闭合式表达式。用高斯分布变量建模的一个主要优点是，通常不需要进行变量转换（第6.7节）。由于高斯分布完全由其均值和协方差指定，我们通常可以通过对随机变量的均值和协方差进行变换来获得变换后的分布。

6.5.1 高斯的边际和条件是高斯的。

在下文中，我们将在多变量随机变量的一般情况下介绍边际化和条件。如果初读时感到困惑，建议读者先考虑两个单变量的随机变量。让X和Y是两个多变量的随机变量，它们可能有

不同的维度。为了考虑应用概率总和法则的效果和条件的影响，我们明确地用连接的状态$[\boldsymbol{x}^\top, \boldsymbol{y}^\top]$来写高斯分布。

$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy}\end{bmatrix}\right). \tag{6.64}$$

其中$\boldsymbol{\Sigma}_{xx} = \mathrm{Cov}[\boldsymbol{x}, \boldsymbol{x}]$和$\boldsymbol{\Sigma}_{yy} = \mathrm{Cov}[\boldsymbol{y}, \boldsymbol{y}]$分别是$\boldsymbol{x}$和$\boldsymbol{y}$的边际协方差矩阵，$\boldsymbol{\Sigma}_{xy} = \mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}]$是$\boldsymbol{x}$和$\boldsymbol{y}$的交叉协方差矩阵。

条件分布$p(\boldsymbol{x} \mid \boldsymbol{y})$也是高斯的（如图6.9(c)所示），并由（2.3Bishop,2006的章节中得出）给出

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}\right) \tag{6.65}$$

$$\boldsymbol{\mu}_{x|y} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y) \tag{6.66}$$

$$\boldsymbol{\Sigma}_{x|y} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yx}. \tag{6.67}$$

请注意，在计算(6.66)中，$\boldsymbol{y}$值是一个观察值，不再是随机的。

*备注。*条件高斯分布出现在许多地方，在那里我们对后验分布感兴趣。

- 卡尔曼滤波器（Kalman,1960）是信号处理中状态估计的最核心算法之一，它只是计算联合分布的高斯条件（Deisenroth and Ohlsson, 2011; S ̈arkk ̈a,2013）。
- 高斯过程（Rasmussen和Williams，2006）是函数分布的一种实际实现方式。在高斯过程中，我们对随机变量的联合高斯性进行了假设。通过对观察到的数据进行（高斯）调节，我们可以确定一个函数的后向分布。
- 潜伏线性高斯模型（Roweis和Ghahramani，1999；Mur-phy, 2012），其中包括概率主成分分析（PPCA）（Tipping和Bishop, 1999）。我们将在第10.7节中对PPCA进行更详细的研究。

◆

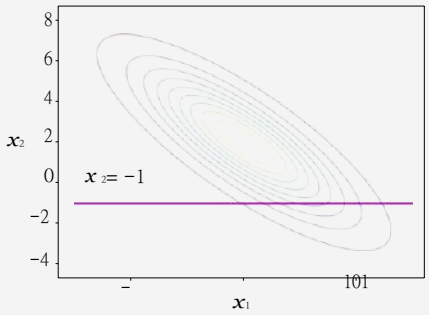
联合高斯分布$p(\boldsymbol{x}, \boldsymbol{y})$的边际分布$p(\boldsymbol{x})$（见(6.64)）本身就是高斯的，通过应用求和法则计算出来的(6.20)，并由

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{y})\mathrm{d}\boldsymbol{y} = \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx}\right). \tag{6.68}$$

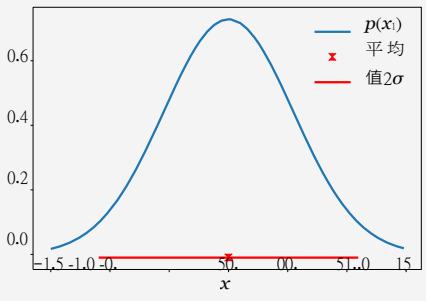
对于$p(\boldsymbol{y})$来说，相应的结果是成立的，它是通过对$\boldsymbol{x}$进行边际化处理而得到的。6.64)，我们忽略(即整合掉)所有我们不感兴趣的东西。这在图6.9(b)中得到了说明。

例子 6.6

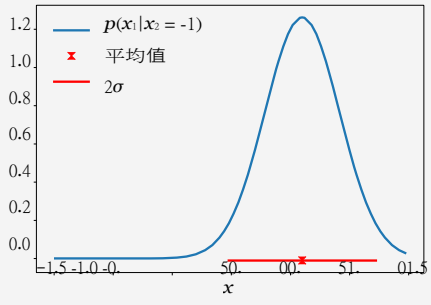
图（6.9a）双变量高斯；（b）联合高斯分布的边际是高斯；（c）高斯的条件分布也是高斯。



(a) Bivariate Gaussian。



(b) 边际分布。



(c) 条件性分布。

考虑双变量高斯分布（如图所示6.9):

$$p(x, x_2) = \mathcal{N}\left(\begin{bmatrix}0\\2\end{bmatrix}, \begin{bmatrix}0.3 & -\\ - & 15\end{bmatrix}\right) \tag{6.69}$$

我们可以计算单变量高斯的参数，条件是$x_2=1$，通过应用(6.66)和(6.67)来分别得到平均数和变异数。在数值上，这就是

$$\mu_{x_1|x_2=-1} = +0(-1)\cdot 0.2-(1--2) = 0.6 \tag{6.70}$$

和

$$\sigma^2_{x_1|x=_2-1} = 0.3 - (-1)\cdot 0.2 - (-1) = 0.1 \tag{6.71}$$

因此，条件高斯由以下公式给出

$$p(x_1 | x_2 = -1) = \mathcal{N}(0.6, 0.1) \tag{6.72}$$

相反，边际分布$p(x_1)$可以通过应用(6.68)，这基本上是在使用随机变量x的平均数和方差$_1$，我们可以得到

$$p(x_1) = \mathcal{N}(0, 0.3) \text{。} \tag{6.73}$$

6.5.2 高斯密度的乘积

对于线性回归（第9章），我们需要计算一个高斯相似度。此外，我们可能希望假设一个高斯先验（第9.3节）。我们应用贝叶斯定理来计算后验，其结果是似然和先验的相乘，也就是说，两个相乘的结果。高斯密度。两个高斯的*乘积*$\mathcal{N}(\boldsymbol{x}|\boldsymbol{a},\boldsymbol{A})\mathcal{N}(\boldsymbol{x}|\boldsymbol{b},\boldsymbol{B})$的推导是一个本章末尾的练习。是一个以$c\in\mathbb{R}$为尺度的高斯分布，由$c\,\mathcal{N}(\boldsymbol{x}|\boldsymbol{c},\boldsymbol{C})$给出。

$$\boldsymbol{C}=(\boldsymbol{A}^{-1}+\boldsymbol{B}^{-1})^{-1} \tag{6.74}$$

$$\boldsymbol{c}=\boldsymbol{C}(\boldsymbol{A}^{-1}\boldsymbol{a}+\boldsymbol{B}^{-1}\boldsymbol{b}) \tag{6.75}$$

$$c=(2\pi)^{-\frac{D}{2}}|\boldsymbol{A}+\boldsymbol{B}|^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}(\boldsymbol{a}-\boldsymbol{b})^{\top}(\boldsymbol{A}+\boldsymbol{B})^{-1}(\boldsymbol{a}-\boldsymbol{b})\right)。 \tag{6.76}$$

缩放常数c本身可以写成高斯密度的形式，在$\boldsymbol{a}$或$\boldsymbol{b}$中都有一个"膨胀的"协方差矩阵$\boldsymbol{A}+\boldsymbol{B}$。即，$c=\mathcal{N}_{\boldsymbol{a}}(\boldsymbol{b},\boldsymbol{A}+\boldsymbol{B})=\mathcal{N}_{\boldsymbol{b}}(\boldsymbol{a},\boldsymbol{A}+\boldsymbol{B})$。

备注。为了符号的方便，我们有时会用$\mathcal{N}(\boldsymbol{x}|\boldsymbol{m},\boldsymbol{S})$来描述高斯密度的函数形式，即使$\boldsymbol{X}$不是一个随机变量。在前面的演示中，我们刚刚做了这个工作，当时我们写道

$$c=\mathcal{N}_{\boldsymbol{a}}(\boldsymbol{b},\boldsymbol{A}+\boldsymbol{B})=\mathcal{N}_{\boldsymbol{b}}(\boldsymbol{a},\boldsymbol{A}+\boldsymbol{B})。 \tag{6.77}$$

这里，$\boldsymbol{a}$和$\boldsymbol{b}$都不是随机变量。然而，用这种方式写c，比(6.76)。 ◆

6.5.3 求和与线性变换

如果X、Y是独立的高斯随机变量（即联合分布为$p(\boldsymbol{x},\boldsymbol{y})=p(\boldsymbol{x})p(\boldsymbol{y})$），$p(\boldsymbol{x})=\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x)$和$p(\boldsymbol{y})=\mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_y)$，那么$\boldsymbol{x}+\boldsymbol{y}$也是高斯分布，并给出$\mathcal{N}(|$由

$$p(\boldsymbol{x}+\boldsymbol{y})=\mathcal{N}(\boldsymbol{\mu}_x+\boldsymbol{\mu}_y,\ \boldsymbol{\Sigma}_x+\boldsymbol{\Sigma}_y)。 \tag{6.78}$$

知道$p(\boldsymbol{x}+\boldsymbol{y})$是高斯的，平均数和协方差矩阵可以立即用(6.46)到(6.49)。当我们考虑到作用于随机变量的即日高斯噪声时，这一属性将非常重要，正如线性回归的情况一样（第9章）。

例子 6.7

由于期望是线性操作，我们可以得到独立高斯随机变量的加权和

$$p(a\boldsymbol{x}+b\boldsymbol{y})=\mathcal{N}(a\boldsymbol{\mu}_x+b\boldsymbol{\mu}_y,\ a^2\boldsymbol{\Sigma}_x+b^2\boldsymbol{\Sigma}_y) \tag{6.79}$$

*备注。*在第11章中有用的一个情况是高斯密 度 的加权和。这与高斯随机变量的加权和不同。 ◆

在该定理中6.12,随机变量x来自一个密度，该密度是两个密度$p_1(x)$和$p_2(x)$的混合物，由α加权。该定理可以推广到多变量随机变量的情况，因为线性期望也适用于多变量随机变量。然而，平方随机变量的概念需要用$\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}$来代替。

定理 *考虑***6.12.***两个单变量高斯密度的混合物*

$$p(x) = \alpha_1 p(x) + (1 - \alpha)p_2(x)\,, \tag{6.80}$$

其中标量$0<\alpha<1$*是混合权重，*$p_1(x)$*和*$p_2(x)$*是单变量的高斯密度（公式*(6.62)*），具有不同的参数。*
即$(\mu_1, \sigma_1)^2$ $(\mu_2, \sigma)^2$。

那么混合密度$p(x)$*的平均值由每个随机变量的平均值的加权和给出。*

$$\mathsf{E}[x] = \alpha_1\mu + (1 - \alpha)\mu_2。 \tag{6.81}$$

混合密度$p(x)$*的方差由以下公式给出*

$$\mathsf{V}[x] = \alpha\sigma^2 + (1 - \alpha)\sigma^2 。 + \alpha\mu^2 + (1 - \alpha)\mu^2 。 - [\alpha\mu + (1 - \alpha)\mu]^2 。 \tag{6.82}$$

证明 混合密度$p(x)$的平均值是由每个随机变量的平均值的加权和给出的。我们应用平均数的定义（定义6.4），然后插入我们的混合物(6.80)，得出的结果是

$$\mathsf{E}[x] = \int_{-\infty}^{\infty} xp(x)dx \tag{6.83a}$$

$$= \int_{-\infty}^{\infty} \alpha xp(x) + (1-\alpha)\, xp(x)\, dx \tag{6.83b}$$

$$= \alpha \int_{-\infty}^{\infty} xp(x)dx + (1-\alpha) \int_{-\infty}^{\infty} xp(x)dx \tag{6.83c}$$

$$= \alpha\mu + (1 - \alpha)\mu_2。 \tag{6.83d}$$

为了计算方差，我们可以使用原始分数版本的方差，即(6.44)，这需要对平方随机变量的期望值进行表达。这里我们使用随机变量的函数（平方）的期望值的定义（定义6.3),

$$\mathsf{E}[x^2] = \int_{-\infty}^{\infty} x^2 p(x)dx \tag{6.84a}$$

$$= \int_{-\infty}^{\infty} \left(\alpha^2 xp\,(x)_1 + (1-\alpha)^2 xp\,(x)_2 \right) dx \tag{6.84b}$$

$$= \alpha \int_{-\infty}^{\infty} x^2 p_1(x)\,dx + (1-\alpha) \int_{-\infty}^{\infty} x^2 p_2(x)\,dx \tag{6.84c}$$

$$= \alpha(\mu_1^2 + \sigma_1^2) + (1-\alpha)(\mu_2^2 + \sigma_2^2), \tag{6.84d}$$

在最后一个等式中，我们再次使用了原始分数版本的方差 (6.44)，得到$\sigma^2 = \mathbb{E}[x^2] - \mu^2$。这可以重新排列，即平方随机变量的期望值是平方平均数与方差之和。

因此，用(6.84d)减去(6.83d)，就可以得到方差。

$$\mathbb{V}[x] = \mathbb{E}[x^2] - (\mathbb{E}[x])^2 \tag{6.85a}$$

$$= \alpha(\mu_1^2 + \sigma_1^2) + (1-\alpha)(\mu_2^2 + \sigma_2^2) - (\alpha\mu_1 + (1-\alpha)\mu_2)^2 \tag{6.85b}$$

$$= \alpha\sigma_1^2 + (1-\alpha)\sigma_2^2 + \alpha\mu_1^2 + (1-\alpha)\mu_2^2 - [\alpha\mu_1 + (1-\alpha)\mu_2]^2 . \tag{6.85c}$$

□

*备注。*前面的推导对任何密度都成立，但由于高斯完全由均值和方差决定，因此可以用闭合形式确定混合物的密度。 ◆

对于混合密度来说，各个成分可以被认为是条件分布（以成分特性为条件）。

公式（6.85c）是条件方差公式的一个例子，也是被称为*总方差法则*，它一般指出，对于两个运行中的总方差法则支配变量X和Y认为，$\mathbb{V}_X[x] = \mathbb{E}_Y[\mathbb{V}_X[x\,y]] + \mathbb{V}_Y[\mathbb{E}_X[x\,y]]$，即$X$的（总）方差是预期条件方差加上条件平均数的方差。

我们在例6.17中考虑了一个双变量的标准高斯随机变量X，并对其进行了线性变换$\boldsymbol{Ax}$。结果是一个高斯随机变量，其均值为零，协方差为$\boldsymbol{AA}^\top$。显然，添加一个常数矢量将改变分布的平均数，而不影响其方差，也就是说，随机变量$\boldsymbol{x}+\boldsymbol{\mu}$是具有平均数$\boldsymbol{\mu}$和相同协方差的高斯。因此，任何线性/非线性

高斯随机变量的变换是高斯分布。 （ 任何线性/affine

考虑一个高斯分布的随机变量X $\boldsymbol{\mu}, \boldsymbol{\Sigma})$。对于一个给定的适当形状的矩阵$\boldsymbol{A}$，让$Y$是一个随机变量，如$\mathcal{N}$ $\mathbb{V}[\boldsymbol{y}] = \mathbb{V}[\boldsymbol{A}$ $\boldsymbol{y}=\boldsymbol{Ax}$是$\boldsymbol{x}$的一个转换版本。

$\boldsymbol{y}$，利用期望值是一个线性算子(6.50)，如下所示。

$$\mathbb{E}[\boldsymbol{y}] = \mathbb{E}[\boldsymbol{Ax}] = \mathbb{A}\mathbb{E}[\boldsymbol{x}] = \boldsymbol{A\mu} \text{ 。} \tag{6.86}$$

同样地，$\boldsymbol{y}$的方差可以通过使用(6.51):

$\boldsymbol{x}] = \mathsf{AV}[\boldsymbol{x}]\boldsymbol{A}^{\mathsf{T}} = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathsf{T}}$。 (6.87)

高斯随机变量的变换也是高斯分布。

这意味着，随机变量$\boldsymbol{y}$的分布是根据

$$p(\boldsymbol{y}) = \mathsf{N}\left(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}^{\mathsf{T}}\boldsymbol{A}\right)。 \tag{6.88}$$

现在让我们考虑一下反向转换：当我们知道一个随机变量的平均值是另一个随机变量的线性转换时
随机变量。对于一个给定的全等级矩阵$\boldsymbol{A} \in \mathbb{R}^{M \times N}$，其中$M$；?$N$，让$\boldsymbol{y} \in \mathbb{R}^{M}$是一个高斯随机变量，其平均值为$\boldsymbol{Ax}$，即。

$$p(\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{Ax}, \boldsymbol{\Sigma}\right) 。 \tag{6.89}$$

相应的概率分布$p(\boldsymbol{x})$是什么？如果$\boldsymbol{A}$是可逆的，那么我们可以写成$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{y}$，并应用上一段的变换。然而，在一般情况下，$\boldsymbol{A}$是不可反转的，我们使用类似于伪逆（3.57）的方法。也就是说，我们预先将两边都乘以$\boldsymbol{A}^{\mathrm{T}}$，然后反转$^{\mathrm{T}}\boldsymbol{AA}$，它是对称的和正定的，给我们的关系是

$$\boldsymbol{y} = \boldsymbol{Ax} \iff (^{\mathrm{T}}\boldsymbol{AA})^{-1}{}^{\mathrm{T}}\boldsymbol{Ay} = \boldsymbol{x} . \tag{6.90}$$

因此，$\boldsymbol{x}$是$\boldsymbol{y}$的一个线性变换，我们得到

$$p(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x} \mid (^{\mathrm{T}}\boldsymbol{AA})^{-1}{}^{\mathrm{T}}\boldsymbol{Ay}, (^{\mathrm{T}}\boldsymbol{AA})\boldsymbol{A}^{-1}{}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{A}(^{\mathrm{T}}\boldsymbol{AA})^{-1}\right) 。 \tag{6.91}$$

6.5.4 从多变量高斯分布中取样

我们不会解释计算机上随机抽样的微妙之处，感兴趣的读者可以参考Gentle(2004)。在多变量高斯的情况下，这个过程包括三个阶段：首先，我们需要一个伪随机数的来源，在区间[0,1]内提供一个均匀的样本；其次，我们使用非线性变换，如Box-Muller变换（Devroye，1986），从单变量高斯中获得一个样本；第三，我们整理这些样本的矢量，得到一个多变量标准正态$\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$的样本，$\boldsymbol{I}$。

对于一般的多变量高斯，也就是说，在平均值为非零，且协方差不是同一矩阵，我们使用高斯随机变量的线性变换的适当联系。假设我们对生成样本$\boldsymbol{x}_i, i = 1, \ldots, n$，从一个多变量的高斯分布，其平均值为$\boldsymbol{\mu}$，协方差矩阵为$\boldsymbol{\Sigma}$。我们将喜欢从一个提供多变量标准正态$\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$样本的采样器中构建样本。

为了计算矩阵的Cholesky因式分解，要求矩阵是对称和正定的（第3.2.3节）。协方差矩阵拥有这一特性。

为了获得多变量正态$\boldsymbol{\mu}$的样本，$\boldsymbol{\Sigma}$，我们可以使用高斯随机变量的线性变换的特性。如果$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$，那么$\boldsymbol{y} = \boldsymbol{Ax} + \boldsymbol{\mu}$，其中$\boldsymbol{AA}^{\mathrm{T}} = \boldsymbol{\Sigma}$是高斯分布，均值为$\boldsymbol{\mu}$，协方差矩阵为$\boldsymbol{\Sigma}$。对$\boldsymbol{A}$的一个方便的选择是使用协方差的Cholesky分解（第4.3节）。矩阵$\boldsymbol{\Sigma} = \boldsymbol{AA}^{\mathrm{T}}$。Cholesky分解法的好处是$\boldsymbol{A}$是三角形的，导致高效的计算。

6.6 共轭关系和指数族

我们在统计学教科书中发现的许多 "有名字 "的概率分布是为了模拟特定类型的现象。例如，我们在第1节中6.5.看到了高斯分布，这些分布也以复杂的方式相互关联（Leemis和McQueston，2008）。对于这个领域的初学者来说，要弄清楚该用哪种分布可能会让人不知所措。此外，这些分布中的很多
在统计和计算被用于 "计算机 "的时候，发现了这些数据。
通过铅笔和纸张。我们很自然地会问，在计算时代，什么是有意义的概念（Efron and Hastie,2016）。在上一节中，我们看到，当分布是高斯的时候，推理所需的许多操作都可以通过对等的方式计算。在这一点上，值得回顾一下在机器学习背景下操作概率分布的理想条件。

是一种工作描述。

1. 在应用概率规则时有一些 "封闭属性"，例如贝叶斯定理。通过封闭性，我们的意思是，应用一个特定的操作会返回一个相同类型的对象。
2. 随着我们收集更多的数据，我们不需要更多的参数来描述这个分布。
3. 由于我们对从数据中学习感兴趣，所以我们希望参数计算能有良好的表现。

事实证明，有一类分布被称为*指数族*
提供了适当的通用性平衡，同时保留了有利的编译和推理特性。在介绍指数系列之前，让我们再看看 "命名的 "概率分布的三个成员：伯努利（例6.8）、二项分布（例6.9）和贝塔分布（例6.10）。

--指数族

例子 6.8

伯努利分布

*伯努利分布*是一个单一的二元随机的分布
变量X的状态$x \in \{0,1\}$。它是由一个连续的pa------来支配的。
伯努利参数$\mu \in [0, 1]$，表示X=的1概率。 伯努利
分布Ber(μ)被定义为

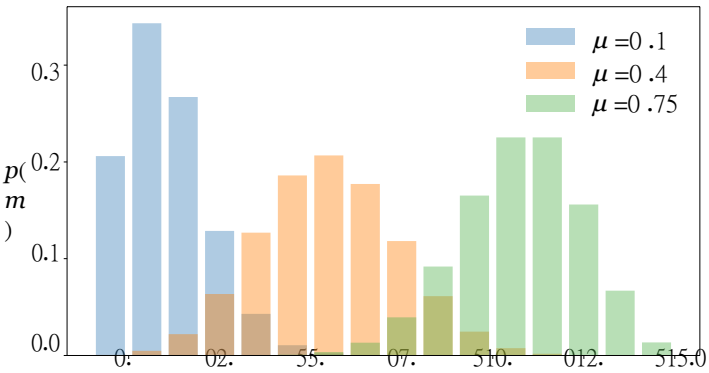
$$p(x \mid \mu) = \mu^{x}(1-\mu)^{1-x} \qquad x \in \{0,1\}, \tag{6.92}$$

$$\mathsf{E}[x] = \mu \tag{6.93}$$

$$\mathbb{V}[x] = \mu(1-\mu) \tag{6.94}$$

其中$\mathsf{E}[x]$和$\mathbb{V}[x]$是二元随机变量X的平均值和方差。

可以使用伯努利分布的一个例子是，当我们对翻转硬币时出现“人头”的概率建模感兴趣。



图二项分布的例子**6.10** μ ∈ {0.1,0 .4,0 .75}和N =15 .

备注。上面对伯努利分布的改写，即我们用布尔变量作为数字0或1，用指数表示，是机器学习教科书中经常使用的一个技巧。这方面的另一个例子是在表达多叉分布时。 ◆

例子（6.9二项分布）

Binomial distribution

*二项分布*是伯努利分布对整数分布的概括（如图所示6.10).特别是，二项分布可以用来描述在1伯努利分布的N个样本集合中观察到X=的m次方的概率
其中$p(X=1)=\mu\in[0,1]$。二项式分布Bin(N,μ)为定义为

$$p(m\mid N,\mu)=\binom{N}{m}\mu^{m}(1-\mu)^{N-m} \tag{6.95}$$

$$\mathsf{E}[m]=N\mu \text{ 。。} \tag{6.96}$$

$$\mathsf{V}[m]=N\mu(1-\mu)\text{ 。} \tag{6.97}$$

其中E$[m]$和V$[m]$分别是m的平均值和方差。

一个可以使用二项式的例子是，如果我们想描述在N个*掷硬币*的实验中观察到m个"头"的概率，如果在一个实验中观察到头的概率是μ。

例子（6.10贝塔分布）

贝塔分布

我们可能希望对有限区间上的连续随机变量进行建模。*贝塔分布*是连续随机变量[0,1]上通常用来表示某些二进制的概率。
事件（例如，支配伯努利分布的参数）。贝塔分布

分布Beta(α, β)（如图6.11所示）本身受两个参数$\alpha > 0$，$\beta > 0$的制约，定义为

$$p(\mu \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1} (1 - \mu)^{\beta-1} \tag{6.98}$$

$$\mathsf{E}[\mu] = \frac{\alpha}{\alpha + \beta}, \qquad \mathsf{V}[\mu] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{6.99}$$

其中$\Gamma(\cdot)$是伽马函数，定义为

$$\Gamma(t) := \int_0^\infty x^{t-1} \exp(-x) dx_\circ \qquad t > .0 \tag{6.100}$$

$$\Gamma(t + 1) = t\Gamma(t) \tag{6.101}$$

请注意，$(\circ)$中的Gamma函数的分数使Beta分布正常化。6.98)使Beta分布正常化。

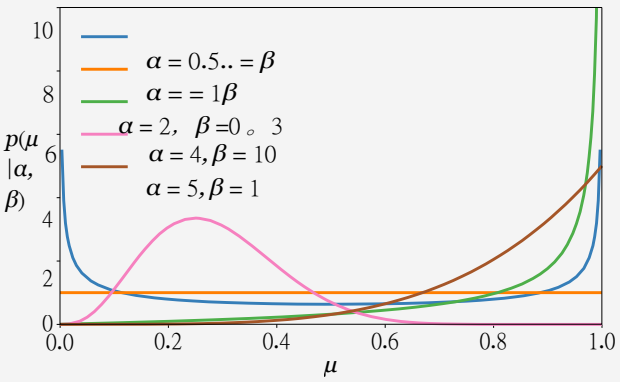


Figure 6.11 Examples of the Beta distribution for different values of α and β.

直观地说，α使概率质量向1，而β使概率能力质量向0。有一些特殊情况（Murphy，2012）。

- 对于$\alpha = = 1\beta$，我们得到均匀分布U[0, 1]。
- 对于$\alpha, \beta < 1$，我们得到一个双峰分布，在和01处有尖峰。对于$\alpha, \beta > 1$
- ，分布是单峰的。
- 对于α，$\beta > 1$和$\alpha = \beta$，分布是单模的，对称的，中心在区间[0，1]，也就是说，模式/平均值在$\frac{1}{2}$。

*备注。有一大堆*有名字的分布，它们以不同的方式相互关联（Leemis和McQueston,2008）。值得注意的是，每个命名的分布都是为了一个特定的原因而创建的，但也可能有其他的应用。了解创建一个特定分布

背后的原因，往往可以洞察到如何最好地使用它。我们介绍了前面的三个分布是

能够说明共轭性（第6.6.1节）和指数族（第6.6.2节）的概念。6.6.3). ◆

6.6.1 结合性

根据贝叶斯定理(6.23)，后验与先验和似然的乘积成正比。由于两个原因，先验的说明可能很棘手。首先，在我们看到任何数据之前，先验应该囊括我们对问题的了解。这通常很难描述。第二，通常不可能以分析方式计算后验分布。然而，有一些先验在计算上是方便的：*共轭先验*。

共轭之前的共轭

定义6.13（共轭先验）。如果后验与先验具有相同的形式/类型，则先验对于似然函数是*共轭的*。

共轭性特别方便，因为我们可以通过更新先验分布的参数来代数计算我们的后验分布。

备注。当考虑到概率分布的几何学时，共轭先验保留了与似然相同的距离结构（Agarwal和Daumé III， 2010）。 ◆

为了介绍共轭先验的一个具体例子，我们在前文中描述了6.11二项分布（定义于离散随机变量）和贝塔分布（定义于连续随机变量）。

例子（6.11贝塔-二项式共轭关系）。
考虑一个二项式随机变量$x \sim \text{Bin}(N, \mu)$，其中

$$p(x \mid N, \mu) = \binom{N}{x} \mu^x (1-\mu)^{N-x}, \quad x = 0, 1, \ldots, N \tag{6.102}$$

是指在*N次掷硬币*中发现*x*次结果为 "头 "的概率，其中μ是 "头 "的概率。我们在参数μ上放置一个Beta先验，即$\mu \sim \text{Beta}(\alpha, \beta)$，其中

$$p(\mu \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1} (1-\mu)^{\beta-1} \tag{6.103}$$

如果我们现在观察到一些结果$x=h$，也就是说，我们在*N次掷*硬币中看到*h*个头，我们计算出μ的后验分布为

$$p(\mu \mid x = h, N, \alpha, \beta) \propto p(x \mid N, \mu)\, p(\mu \mid \alpha, \beta) \tag{6.104a}$$
$$\propto \mu^h (1-\mu)^{(N-h)} \mu^{\alpha-1} (1-\mu)^{\beta-1} \tag{6.104b}$$
$$= \mu^{h+\alpha-1} (1-\mu)^{(N-h)+\beta-1} \tag{6.104c}$$

概率共轭	先验 后验
伯努利贝塔贝塔	
二项式BetaBeta	
高斯	/反伽马伽马 高斯/反
维萨特高斯/反维萨特 多项式	DirichletDirichlet

表 常见似然函数的共轭先验的例子**6.2**。

$$\propto \text{Beta}(h+\alpha,\ N-h+\beta) \tag{6.104d}$$

即后验分布是以Beta分布为先验，即Beta先验对于二项式似然函数中的参数μ是共轭的。

在下面的例子中，我们将推导出一个与Beta-Binomial共轭结果相似的结果。这里我们将证明Beta分布是Bernoulli分布的共轭先验。

例子（Beta6.12-Bernoulli Conjugacy）。
让$x\in\{0,\ 1\}$根据伯努利分布，其分布情况为
参数$\theta\in[0,\ 1]$，即$p(x=|1\ \theta)=\theta$。这也可以表示为
因为$p(x\mid\theta)=\theta^x(1-\theta)^{1-x}$。让$\theta$按照Beta分布法进行分布。
但有参数$\alpha,\ \beta$，即$p(\theta\mid\alpha,\beta)\propto\theta^{(\alpha-1}1-\theta)^{\beta-1}$。

将贝塔分布和伯努利分布相乘，我们可以得到

$$p(\theta\mid x,\alpha,\beta)=p(x\mid\theta)p(\theta\mid\alpha,\beta) \tag{6.105a}$$
$$\propto\theta^x(1-\theta)^{1-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{6.105b}$$
$$=\theta^{\alpha+x1-}(1-\theta)^{\beta+(1-x)} \tag{6.105c}$$
$$\propto p(\theta\mid\alpha+x,\beta+(1-x)) \tag{6.105d}$$

最后一行是参数为（$\alpha+x$；$\beta+(1-x)$）的β分布。

表格6.2中列出了一些参数的共轭先验参数的例子。

概率建模中使用的标准似然。分布，如

Gamma先验是分布。

多项式、反Gamma、反Wishart和Dirichlet可以在任何统计学文本中找到，例如Bishop(2006)中就有描述。

Beta分布是二项式和伯努利似然中参数μ的共轭先验。对于高斯似然函数，我们可以在平均数上放置一个共轭高斯先验。高斯似然在表中出现两次的原因是，我们需要区分单变量和多变量的情况。在单变量（标量）情况下，反Gamma是方差的共轭先验。在多变量情况下，我们使用共轭反Wishart分布作为协方差矩阵的先验。Dirichlet分布是共轭反Wishart

在单变量高斯似然中，精度（反方差）是共轭的，而Wishart先验在多变量高斯似然中，精度矩阵（反协方差矩阵）是共轭的。

多项式似然函数的门先验。关于进一步的细节，我们参考Bishop(2006)。

6.6.2 足够的统计数据

回顾一下，一个随机变量的统计量是该随机变量的一个确定性的函数。例如，如果$\boldsymbol{x} = [x_1, \ldots, x_N]^\top$是一个单变量高斯随机变量的向量，即$x_n \sim \mathcal{N}(\mu, \sigma^2)$，那么样本平均值$\hat{\mu} = \frac{1}{N}(x_1 + \ldots + x_N)$就是一个统计量。罗纳德-费舍尔爵士（Sir Ronald Fisher）提出了*充分统计量的概念*：有一些统计量将包含所有可用的信息，可以从所考虑的分布对应的数据中推断出来。换句话说，充分统计量包含了对人口进行推断所需的所有信息，也就是说，它们是足以代表分布的统计量。

充分的统计数据

对于一组以θ为参数的分布，让X是一个随机变量，其分布$p(x \mid \theta_0)$给定于一个未知的θ_0。如果统计量的向量$\varphi(x)$包含关于θ_0的所有可能的信息，则称为θ的充分统计量。为了更正式地说明 "包含所有可能的信息"，这意味着给定θ的x的概率可以被分解为不依赖于θ的部分，以及仅通过$\varphi(x)$依赖于θ的部分。Fisher-Neyman因式分解定理将这一概念正式化，我们在定理中说明了这一概念6.14中说明，无需证明。

Fisher-Neyman

定理（6.14 Fisher-Neyman）。*[Theorem 6.5 inLehmann and Casella (1998)] 设X有概率密度函数$p(x \mid \theta)$。那么统计量*
$\varphi(x)$对θ来说是充分的，当且仅当$p(x \mid \theta)$可以被写成以下形式时

$$p(x \mid \theta) = h(x)g_\theta(\varphi(x)), \tag{6.106}$$

其中$h(x)$是独立于θ的分布，g_θ通过充分统计量$\varphi(x)$捕捉到对θ的所有依赖性。

定理

如果$p(x \mid \theta)$不依赖于θ，那么$\varphi(x)$对任何函数φ来说都是充分统计量。更有趣的情况是$p(x \mid \theta)$只依赖于$\varphi(x)$而不依赖于x本身。在这种情况下，$\varphi(x)$是θ的一个充分统计量。

在机器学习中，我们考虑的是一个分布的有限数量的样本。我们可以想象，对于简单的分布（如例6.8中的伯努利分布），我们只需要少量的样本来估计分布的参数。我们也可以考虑相反的问题：如果我们有一组数据（来自一个未知分布的样本），哪个分布能给出最佳拟合？一个自然的问题是，随着我们观察到更多的数据，我们是否需要更多的参数θ来

描述分布？事实证明，一般来说答案是肯定的，这一点在非参数统计中有所研究（Wasserman,2007）。一个相反的问题是，考虑哪一类分布具有有限维度的

足够的统计量，也就是描述它们所需的参数数量不会任意增加。答案是指数族分布，在下一节中描述。

6.6.3 指数家族

在讨论分布（离散或连续随机变量的分布）时，我们可以有三个可能的抽象层次。在第一层次（最具体的一端），我们有一个固定参数的特殊命名的分布，例如单变量高斯分布 $\mathcal{N}(0, 1)$，其平均值为1零，方差为单位。在机器学习中，我们经常使用第二层次的抽象，也就是说，我们固定参数形式（单变量高斯）并从数据中推断参数。例如，我们假设单变量高斯 $\mathcal{N}(\mu, \sigma^2)$具有未知的平均值$\mu$和未知的方差$\sigma^2$，并使用最大似然拟合来确定最佳参数（$\mu, \sigma^2$）。我们将在第九章考虑线性回归时看到这样一个例子。第三个抽象层次是考虑分布族，在本书中，我们考虑前指数族。单变量高斯是指数家族成员的一个例子。许多广泛使用的统计模型，包括表格6.2,中所有 "命名 "的模型，都是指数家族的成员。它们都可以被统一为一个概念（Brown,1986）。

*备注。*一个简短的历史轶事。像数学和科学中的许多概念一样，指数族是由不同的研究者在同一时间独立发现的。1935-1936年，塔斯马尼亚的埃德温-皮特曼、巴黎的乔治-达尔莫瓦和纽约的伯纳德-库普曼分别表明，指数族是唯一在重复条件下享有有限维充分统计的族。

独立采样（Lehmann和Casella，1998）。 ◆

*指数族*是一个概率分布的家族，参数-$\boldsymbol{\theta} \in \mathbb{R}^D$，其形式为 指数族

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\left(\langle \boldsymbol{\theta}, \boldsymbol{\varphi}(\boldsymbol{x}) \rangle - A(\boldsymbol{\theta}) \right), \tag{6.107}$$

其中$\boldsymbol{\varphi}(\boldsymbol{x})$是充分统计量的向量。一般来说，任何内部程序（第3.2节）都可以用在(6.107)，为了具体化，我们在这里将使用标准点积($\langle \boldsymbol{\theta}, \boldsymbol{\varphi}(\boldsymbol{x}) \rangle = \boldsymbol{\theta}^\top \boldsymbol{\varphi}(\boldsymbol{x})$)。请注意，指数族的形式本质上是Fisher-Neyman定理中$g_\theta(\varphi(x))$的一个特殊表达(定理6.14).

通过在充分统计量向量$\boldsymbol{\varphi}(\boldsymbol{x})$中增加另一个条目（$\log h(\boldsymbol{x})$），并约束相应的参数$\theta_0=1$，可以将因子$h(\boldsymbol{x})$吸收到点乘项中。术语$A(\boldsymbol{\theta})$是归一化常数，确保分布相加或不相加。

栅格为1，被称为*对数分割函数*。一个好的直观的无对数分区

忽略这两个项，可以得到指数族的概念 功能

并将指数族视为分布形式的

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{\varphi}(\boldsymbol{x})\right) \text{。} \tag{6.108}$$

自然参数

对于这种形式的参数化，参数$\boldsymbol{\theta}$被称为*自然参数*。乍一看，指数族似乎是通过将指数函数添加到点乘的结果中而进行的一种穆丹式的转换。然而，有许多含义允许凸-------。

基于我们可以在$\boldsymbol{\varphi}(\boldsymbol{x})$中捕获数据信息的事实，我们可以进行便捷的建模和高效的计算。

例子 (6.13高斯为指数族)

考虑单变量高斯分布$\mathcal{N}\left(\mu, \sigma^2\right)$。让$\boldsymbol{\varphi}(x)=\begin{bmatrix} x \\ x^2 \end{bmatrix}$

然后通过使用指数族的定义。

$$p(x \mid \boldsymbol{\theta}) \propto \exp(\theta_1 x + \theta_2 x^2) \tag{6.109}$$

设置

$$\theta = \begin{bmatrix} \frac{\mu}{\sigma^2}, & -\frac{1}{2\sigma^2} \end{bmatrix}^{\mathrm{T}} \tag{6.110}$$

并将其代入(6.109)，我们得到

$$p(x \mid \boldsymbol{\theta}) \propto \exp\left[\frac{\mu}{\sigma^2} x - \frac{x^2}{2\sigma^2}\right] \propto \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] \tag{6.111}$$

因此，单变量高斯分布是指数分布的一个成员。

足够的统计量$\boldsymbol{\varphi}(x)=\begin{bmatrix} x \\ x^2 \end{bmatrix}$，以及自然参数-

中的$\boldsymbol{\theta}$所给出的指针。6.110).

例子（6.14伯努利为指数族）

回顾例题中的伯努利分布6.8

$$p(x \mid \mu) = \mu^x (1-\mu)^{1-x} \quad x \in \{0, 1\}. \tag{6.112}$$

这可以写成指数族的形式

$$\begin{aligned} p(x \mid \mu) &= \exp\left[\log \mu^x (1-\mu)^{1-x}\right] && \text{(6.113a)} \\ &= \exp\left[x \log \mu + (1-x) \log(1-\mu)\right] && \text{(6.113b)} \\ &= \exp\left[x \log \mu - x \log(1-\mu) + \log(1-\mu)\right] && \text{(6.113c)} \\ &= \exp\left[x \log \frac{\mu}{1-\mu} + \log(1-\mu)\right] && \text{(6.113d)} \end{aligned}$$

最后一行(6.113d)可以被认定为是指数族形式(6.107)，方法是观察

$$h(x) = 1 \tag{6.114}$$

$$\theta = \text{对-数}_{\mu} \tag{6.115}$$

$$\varphi(x) = x \tag{6.116}$$

$$A(\theta) = -\log(1-\mu) = \log(1+\exp(\theta)) \tag{6.117}$$

。

θ和μ之间的关系是可逆的，所以 $\mu = \dfrac{1}{1+\exp(-\theta)}$. (6.118)

关系(6.118)被用来获得()的右边相等。6.117).

*备注。*原始伯努利参数μ与

的自然参数θ被称为*sigmoid*或logistic函数。Ob -

sigmoid服务于$\mu \in (0, 1)$但θ R，因此sigmoid函数将一个实值挤压到(0, 1)的范围。这一特性在数学中是很有用的。

例如，它被用于逻辑回归（Bishop，2006，第4.3.2节），以及作为神经网络的非线性激活函数（Goodfellow等，2016，第6章）。 ◆

如何找到某个特定分布的共轭分布的参数形式往往并不明显（例如，表中的6.2).指数族提供了一种方便的方法来寻找分布的共轭对。考虑到随机变量X是指数族的成员(6.107):

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = h(\boldsymbol{x}) \exp((\boldsymbol{\theta}, \boldsymbol{\varphi}(\boldsymbol{x})) - A(\boldsymbol{\theta}))。 \tag{6.119}$$

指数家族的每个成员都有一个共轭先验（Brown，1986）。

$$p(\boldsymbol{\theta} \mid \boldsymbol{\gamma}) = h_c(\boldsymbol{\theta}) \exp(\boldsymbol{\gamma}_1 \cdot \boldsymbol{\theta} / \gamma_2 - A(\boldsymbol{\gamma})_c)。 \tag{6.120}$$

其中$\boldsymbol{\gamma} = \begin{matrix}\boldsymbol{\gamma}_1 \\ \gamma_2\end{matrix}$ 的维度为dim($\boldsymbol{\theta}$)+1。的充分统计量。

的共轭先验为 $\boldsymbol{\theta}$ $-A(\boldsymbol{\theta})$.通过使用一般的知识

根据指数族的共轭先验的形式，我们可以推导出对应于特定分布的共轭先验的函数形式。

例子 6.15

回顾伯努利分布的指数族形式(6.113d)

$$p(x \mid \mu) = \exp\left(x \log \frac{\mu}{1-\mu} + \log(1-\mu)\right). \tag{6.121}$$

典型的共轭先验的形式是

$$p(\mu \mid \alpha, \beta) = \frac{\mu}{1-\mu} \exp \alpha \log \frac{\mu}{1-\mu} + (\beta + \alpha) \log(1 - \mu) - A(\boldsymbol{\gamma}) 。 \tag{6.122}$$

其中我们定义$\boldsymbol{\gamma} := [\alpha, \beta + \alpha]^\top$，$h_c(\mu) := \mu/(1 - \mu)$。 等式符号 (6.122) 然后简化为

$$p(\mu \mid \alpha, \beta) = \exp [(\alpha - 1) \log \mu + (\beta - 1) \log(1 - \mu) - A_c(\alpha, \beta)] 。 \tag{6.123}$$

用非指数族的形式来表示，可以得到

$$p(\mu \mid \alpha, \beta) \propto \mu^{\alpha-1} (1-\mu)^{\beta-1} \tag{6.124}$$

我们将其确定为贝塔分布(6.98).在例子中6.12,中，我们假设Beta分布是Bernoulli分布的共轭先验，并证明它确实是共轭先验。在这个例子中，我们通过观察指数族形式的伯努利分布的典型共轭先验，得出了Beta分布的形式。

如上一节所述，指数族的主要动机是它们具有有限维的充分统计量。此外，共轭分布很容易写下来，而且共轭分布也来自指数族。从推断的角度来看，最大似然估计的表现很好，因为充分统计量的经验估计是充分统计量的群体值的最佳估计（回顾高斯的均值和协方差）。从优化的角度来看，对数似然函数是凹的，允许应用有效的优化方法（第七章）。

6.7 变量的变化/反变换

看起来，已知的分布非常多，但实际上，我们有名字的分布集是相当有限的。因此，了解转换后的随机变量是如何分布的往往是有用的。例如，假设X是一个按照单变量正态分布$\mathcal{N}(0, 1)$的随机变量，X^2的分布是什么？另一个在数学学习中很常见的例子是，给定X_1和X_2是单变量标准正态分布，X_1+X_2）的分布是什么？

计算X_1+X_2)分布的一个方案是计算X_1和X_2的平均数和方差，然后将它们结合起来。正如我们所看到的
在第6.4.4节中，当我们考虑随机变量的仿生变换时，我们可以计算出所产生的运行变量的平均值和方差。

ables。然而，我们可能无法获得变换下的分布的函数形式。此外，我们可能对随机变量的非线性变换感兴趣，而这些变换的闭合形式并不容易获得。

备注（记号）。在本节中，我们将对随机变量和它们的取值进行明确说明。因此，回顾一下，我们用大写字母X、Y来表示随机变量，用小写字母x、y来表示随机变量在目标空间的取值。我们将明确地把离散随机变量X的pmfs写成$P(X=x)$。对于连续随机变量X（第6.2.2），pdf被写成$f(x)$，cdf被写成$F_X(x)$。 ◆

我们将研究两种获得随机变量变换分布的方法：一种是使用累积分布函数定义的直接方法，另一种是改变变量的方法。

使用微积分的连锁规则（第5.2.2节）。变量变化的应用生成的时刻

该方法被广泛使用，因为它提供了一个"秘方"，可以尝试计算由于变换而产生的分布。我们将对单变量随机变量的技术进行说明，而对多变量随机变量的一般情况只简要地提供结果。离散随机变量的变换可以被直截了当地理解。假设有一个离散随机变量X，其pmf $P(X=x)$（第6.2.1），以及一个可逆函数$U(x)$。考虑一下变换形成的随机变量$Y := U(X)$，pmf $P(Y=y)$。那么

函数也可以用来研究随机的变换。

变量（Casella和Berger，2002，第2章）。

$$P(Y=y) = P(U(X)=y) \quad \text{利益转化} \quad (6.125a)$$
$$= P(X=U^{-1}(y)) \quad \text{逆转} \quad (6.125b)$$

其中我们可以看到，$x=U^{-1}(y)$。因此，对于离散的随机变量，变换直接改变了各个事件（概率经过适当的变换）。

6.7.1 分布函数技术

分布函数技术回到了第一原理，并使用了cdf $F_X(x)=P(X\leqslant x)$的定义，以及它的微分是pdf $f(x)$的事实（Wasserman,2004,第二章）。对于一个随机变量X和一个函数U，我们通过以下方法找到随机变量$Y := U(X)$的pdf

1.找到cdf。

$$F_Y(y) = P(Y \leqslant y) \quad (6.126)$$

2.对cdf $F_Y(y)$进行微分，得到pdf $f(y)$。

$$\frac{\mathrm{d}}{}$$

$$f(y) = \frac{F(y)}{\text{染料}}. \tag{6.127}$$

我们还需要记住，由于U的转换，随机变量的域可能已经改变。

例子 6.16

让X是一个连续的随机变量，其概率密度函数在0$\leqslant x \leqslant$上。 1

$$f(x) = 3x^2. \tag{6.128}$$

我们感兴趣的是找到$Y=X$的^{2}pdf。

函数f是x的增函数，因此得到的y的值位于区间[0，1]内。我们得到

$$\begin{aligned}
F_Y(y) &= P(Y \leqslant y) && \text{CDF的定义} && (6.129a)\\
&= P(X^2 \leqslant y) && \text{利益转换} && \\
&= P(X \leqslant y^{\frac{1}{2}}) && && (6.129b)\\
&= F_X(y^{\frac{1}{2}}) && \text{倒数} && (6.129c)\\
&= \int_0^{y^{\frac{1}{2}}} 3t^2 dt && \text{CDF的定义} && (6.129d)\\
&= t^3\big]_{t=0}^{t=y^{\frac{1}{2}}} && \text{cdf是一个定积分} && (6.129e)\\
&= y^{\frac{3}{2}}, \quad 0 \leqslant y \leqslant 1. && \text{整合的结果} && (6.129f)\\
& && && (6.129g)
\end{aligned}$$

因此，Y的cdf是

$$F_Y(y) = y^{\frac{3}{2}} \tag{6.130}$$

为0...y...1。为了得到pdf，我们对cdf进行微分

$$f_Y(y) = \frac{\mathrm{d}}{\text{染料}} F_Y(y) = \frac{3}{2} y^{\frac{1}{2}} \tag{6.131}$$

对于0 $\leqslant y \leqslant$ 1。

有反相的函数被称为双射函数（第2.7节）。

在例6.16中，我们考虑了一个严格单调递增的函数$f(x) = 3x^2$，这意味着我们可以计算一个反函数。一般来说，我们要求感兴趣的函数$y = U(x)$有一个in-verse $x = U^{-1}(y)$。通过考虑随机变量X的累积分布函数$F_X(x)$，可以得到一个有用的结果，并使用
它作为变换$U(x)$。这导致了下面的定理。

定理6.15。 *[Theorem 2.1.10 inCasella and Berger(2002)] 设X是一个具有严格单调的累积分布函数$F_X(x)$的连续随机变量。那么，随机变量Y定义为*

$$y := f_X(x) \tag{6.132}$$

其有均匀分布。

该定理6.15被称为*概率积分变换*，它是

概率积分的一种。

通过对均匀随机变量的抽样结果进行转换，用来推导出从分布中抽样的算法（Bishop，2006）。该算法的工作原理是：首先从一个均匀分布中产生一个样本，然后通过反cdf（假设有此功能）对其进行转换，以获得一个来自所需分布的样本。概率积分变换也被用于假设测试样本是否来自一个特定的分布（Lehmann和Romano，2005）。cdf的输出给出了一个均匀分布，这一想法也构成了copu-las的基础（Nelsen,2006）。

转变

6.7.2 变量的变化

本节中的分布函数技术是在第一原理的基础上，利用中值的定义和微分及积分的特性，从第一原理中得出的。6.7.1中的分布函数技术是从第一原理中推导出来的，它基于cdfs的定义，并使用了in- verses、微分和积分的特性。这种来自第一原理的论证依赖于两个事实。

1.我们可以将Y的cdf转化为X的cdf的表达式。 2.我们可以对cdf进行微分，以获得pdf。

Let us break down the reasoning step by step, with the goal of understand-
在定理中采用了更普遍的变量变化方法。6.16.

变量的改变

*备注："改变变量 "的名称来自于在面临困难的积分时改变积分的变量。*变量变化 "这一名称来自于当面临困难的积分时改变积分的变量的想法。对于单变量函数，我们使用积分的替换规则。

$$f(g(x))g'(x)dx = f(u)du\,, \quad \text{其中} \quad u = g(x)。 \tag{6.133}$$

这一规则的推导是基于微积分的连锁规则（5.32），并通过两次微积分基本定理的应用。微积分基本定理将积分和微分在某种程度上是彼此的 "逆 "这一事实正式化。通过对方程$u = g(x)$*的*微小变化（微分）的思考，可以获得对该规则的直观理解，即把$\Delta u = g'(x)\Delta x$看作$u = g(x)$的微分。通过替代$u = g(x)$，参数里面的
的右边的积分(6.133)的右侧成为$f(g(x))$。假设项du可以用$du \approx \Delta u = g'(x)\Delta x$*来*近似，并且$dx \approx \Delta x$，我们得到(6.133). ◆

考虑一个单变量的随机变量X，和一个*可反转的*函数U，它给我们另一个随机变量$Y = U(X)$。我们假设随机变量X有状态$x \in [a, b]$。根据cdf的定义，我们有

$$F(xy) = P(Y \leqslant y) 。 \tag{6.134}$$

我们对随机变量的一个函数U感兴趣

$$P(Y \leqslant y) = P(U(X) \leqslant y), \tag{6.135}$$

其中我们假设函数U是可逆的。一个区间上的可逆函数要么严格增加，要么严格减少。在U是严格增加的情况下，那么它的逆U^{-1}也是严格增加的。通过对$P(U(X) \leqslant y)$的参数应用逆U，我们可以得到

$$P(U(X) \leqslant y) = P(U^{-1}(X) \leqslant U^{-1}(y)) = P(X \leqslant U(^{-1}y)) \text{。} \tag{6.136}$$

( )中最右边的项是X的cdf的表达式。6.136)中最右边的一项是X的cdf的表达。

$$P(X \leqslant U^{-1}(y)) = \int_{a}^{U(^{-1}y)} f(x)dx \text{。} \tag{6.137}$$

现在我们有一个关于x的Y的cdf的表达。

$$F_Y(y) = \int_{a}^{U(^{-1}y)} f(x)dx \text{。} \tag{6.138}$$

为了得到pdf，我们对(6.138)与y有关。

$$f(y) = \frac{\mathrm{d}}{dy} F(y) = \frac{\mathrm{d}}{dy} \int_{a}^{U(^{-1}y)} f(x)dx \text{。} \tag{6.139}$$

请注意，右手边的积分是关于x的，但我们需要关于y的积分，因为我们是关于y的微分。6.133)来得到替代的结果

$$f(U^{-1}(y))U(^{-1}y)dy = f(x)dx \quad \text{，其中 } ex = U(^{-1}y) \text{。} \tag{6.140}$$

(6.140) 使用(6.140)在(6.139)的右边，我们可以得到

$$f(y) = \frac{\mathrm{d}}{\text{染料}} \int_{a}^{U(^{-1}y)} f_x(U^{-1}(y))U(^{-1}y)dy \text{。} \tag{6.141}$$

然后我们回顾一下，微分是一个线性算子，我们用下标x来提醒自己，$f_x(U(^{-1}y))$是x的函数，而不是y.再次引用微积分的基本定理，我们可以得到

$$f(y) = f_x(U^{-1}(y)) - \frac{\mathrm{d}}{\text{染料}} U^{-1}(y) \text{。} \tag{6.142}$$

回顾一下，我们假设U是一个严格增加的函数。对于递减函数，事实证明，当我们遵循同样的推导时，我们有一个负号。我们引入微分的绝对值，以便对增加和减少的U有相同的表达。

$$f(y) = f(U_{\cdot}(y))\left|\frac{\mathrm{d}}{xdy}U^{-1}(y)\right| 。 \tag{6.143}$$

这被称为*变量变化技术*。变量变化中的 术语$U^{-1}(y)$
(6.143)测量单位体积的变化程度，当应用U
(也见第5.3节中雅各布的定义）。

备注。与（6.125b）中的离散情况相比，我们有一个额外的因素$\det U^{-1}$（y）。连续情况下需要更多的注意，因为

$P(Y=y)$ = 对于0所有的y，概率密度函数$f(y)$并没有作为涉及y*的*事件的概率的描述。 ◆

到目前为止，我们在本节中一直在研究单变量的变化。多变量随机变量的情况是类似的，但由于绝对值不能用于多变量函数而变得复杂。相反，我们使用雅各布矩阵的行列式。回顾（5.58），雅各布矩阵是一个偏导数矩阵，非零行列式的存在表明我们可以反转雅各布矩阵。回顾第4.1节的讨论，行列式的产生是因为我们的微分（体积的立方体）被Jacobian转化为平行的lepipeds。让我们在下面的定理中总结一下前面的讨论，它给了我们一个多变量变化的秘诀。

定理6.16。 *[Billingsley(1995)中的定理17.2] 让$f(\boldsymbol{x})$为多变量连续随机变量X的概率密度值。如果矢量值函数$\boldsymbol{y}=U(\boldsymbol{x})$对于$\boldsymbol{x}$域内的所有数值都是可微和可逆的，那么对于$\boldsymbol{y}$的相应数值，$Y=U(X)$的概率密度由以下公式给出*

$$f(\boldsymbol{y}) = f_{\boldsymbol{x}}(U^{-1}(\boldsymbol{y}))\left|\det \frac{\partial}{\partial \boldsymbol{y}} U^{-1}(\boldsymbol{y})\right|. \tag{6.144}$$

这个定理乍一看很吓人，但关键是多变量随机变量的变化是按照单变量变化的程序进行的。首先，我们需要计算出反变换，并将其代入$\boldsymbol{x}$的密度中，然后计算出雅各布的行列式，并将结果相乘。下面的例子说明了双变量随机变量的情况。

例子 6.17

考虑一个双变量的随机变量X，其状态$\boldsymbol{x} = \begin{matrix} x_1 \\ x_2 \end{matrix}$ 和亲

弹性密度函数

$$f\begin{matrix} x_1 \\ x_2 \end{matrix} = \frac{1}{2\pi}\exp\left(-\frac{1}{2}\begin{matrix} x_1 \\ x_2 \end{matrix}^{\mathrm{T}}\begin{matrix} x_1 \\ x_2 \end{matrix}\right). \tag{6.145}$$

我们利用定理中的变量变化技术6.16来推导出

effect of a linear transformation (Section2.7) of the random variable. Consider a matrix $\boldsymbol{A} \in \mathrm{R}^{2\times 2}$ defined as

$$\boldsymbol{A} = \begin{matrix} ab \\ 唱片。 \end{matrix} \tag{6.146}$$

我们感兴趣的是找到状态为$\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}$的反式二元随机变量$Y$的概率密度函数。

回顾一下，对于变量的改变，我们需要$\boldsymbol{x}$的反变换作为$\boldsymbol{y}$的函数。因为我们考虑的是线性变换，所以反变换是由矩阵的反值给出的（见第2.2.2节）。对于×2矩阵2，我们可以明确地写出这个公式，具体如下

$$\begin{matrix} x_1 \\ XYAD_2 \end{matrix} = \boldsymbol{A}^{-1}\begin{matrix} y_1 \\ {}_2 \end{matrix} = \frac{1}{-BC-}\begin{matrix} d & -b\,y_1 \\ CAY_2 \end{matrix}。 \tag{6.147}$$

请注意，$-adbc$是$\boldsymbol{A}$的行列式（第4.1节），相关的概率密度函数由以下公式给出

$$f(\boldsymbol{x}) = f(\boldsymbol{A}^{-1}\boldsymbol{y}) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}^{-\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{y}\right)。 \tag{6.148}$$

矩阵乘以矢量相对于矢量的偏导是矩阵本身（第5.5节），因此

$$\frac{\partial}{\partial \boldsymbol{y}}\boldsymbol{A}^{-1}\boldsymbol{y} = \boldsymbol{A}^{-1}. \tag{6.149}$$

Recall from Section4.1that the determinant of the inverse is the inverse of the determinant so that the determinant of the Jacobian matrix is

$$\frac{\partial}{\partial \boldsymbol{y}}\boldsymbol{A}^{-1}\boldsymbol{y} = \frac{1}{广告-BC}. \tag{6.150}$$

现在我们能够应用西奥-雷姆的变量变化公式6.16将(6.148)与(6.150)相乘，可得

$$f(\boldsymbol{y}) = f(\boldsymbol{x})\det\frac{\partial}{\partial \boldsymbol{y}}\boldsymbol{A}^{-1}\boldsymbol{y} \tag{6.151a}$$

$$= \frac{1}{2\pi}\,曝光\,\boldsymbol{v}^{\mathrm{T}}\boldsymbol{A}^{-1\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{v}\,()\,|ad-bc|^{-1}。 \tag{6.151b}$$

虽然例子6.17是基于一个双变量的随机变量，这使得我们可以很容易地计算矩阵的逆值，而前面的关系对于更高的维度也是成立的。

*备注。*我们在第二节中看到6.5中的密度$f(\boldsymbol{x})$*实际上是标准的高斯分布。*6.148)实际上是标准的高斯分布，而转换后的密度$f(\boldsymbol{y})$是一个双变量高斯，协方差$\boldsymbol{\Sigma}=\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$。 ◆

我们将使用本章的观点来描述概率建模

在第8.4节中，我们将介绍一个图形化语言，以及在第8.5节中介绍一个图形化语言。我们将在第9章和第11章看到这些想法的直接机器学习应用。

6.8进一步阅读

Grinstead和Snell(1997)以及Walpole等人(2011)提供了更轻松的介绍，适合于自学。对概率的哲学方面感兴趣的读者应该考虑Hacking(2001)，而Downey(2014)则提出了一种与软件工程更相关的方法。关于指数族的概述可以在Barndorff-Nielsen（2014）中找到。我们将在第八章看到更多关于如何使用概率分布来为机器学习任务建模。具有讽刺意味的是，最近对神经网络兴趣的激增导致了对概率模型更广泛的赞赏。例如，归一化流量的想法（Jimenez Rezende and Mohamed,2015）依赖于变量的变化来转换随机变量。Goodfellow等人(2016)的书16中20的各章对应用于神经网络的变异推理方法进行了概述。

我们通过回避度量理论问题（Billingsley,1995;Pollard,2002），以及假设我们有实数和定义实数上的集合的方法以及它们适当的出现频率，来回避连续随机变量中的大部分困难。这些细节确实很重要，例如，在连续随机变量x、y的条件概率$p(y\,x)$的指定中（Proschan和Presnell，1998）。懒惰的符号隐藏了一个事实，即我们想指定$X=x$（这是一个度量为零的集合）。此外，我们对y的概率密度函数感兴趣。一个更精确的符号应该是$\mathrm{E}_y[f(y)\,\sigma(x)]$， 在这里我们取一个测试函数$f$在$y$上的期望，条件是$\sigma$代数的
x.对概率的细节感兴趣的更多技术性的听众，可以通过以下方式来了解------。或理论有许多选择（Jaynes，2003；MacKay，2003；Jacod and Protter，2004；Grimmett and Welsh，2014），包括一些非常技术性的讨论（Shiryayev，1984；Lehmann and Casella，1998；Dudley，2002；Bickel and Doksum， 2006；Çinlar， 2011）。另一种接近概率的方法是从期望的概念开始，然后 "向后 "推导出概率空间的必要属性（Whittle,2000）。由于机器学习允许我们在更复杂的数据类型上建立更复杂的分布模型，概率机器学习模型的开发者必须了解这些更多的技术方面。以概率建模为重点的机器学习文本包括MacKay（2003）；Bishop（2006）；

Rasmussen和Williams（2006）；Bar-ber（2012）；Murphy（2012）的书。

练习

6.1 考虑以下两个离散随机变量X和Y的双变量分布$p(x, y)$。

y_1	0.01	0.02	0.03	0.1	0.1
y_2	0.05	0.1	0.05	0.07	0.2
y_3	0.1	0.05	0.03	0.05	0.04
	x_1	x_2	x_3	x_4	x_5

X

计算。

a.边际分布$p(x)$和$p(y)$。

b.条件分布$p(x|Y=y_1)$和$p(y|X=x_3)$。

6.2 考虑两个高斯分布的混合物（如图6.4所示）。

$$0.4\mathcal{N}\left(\begin{bmatrix}10\\2\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right) + 0.6\mathcal{N}\left(\begin{bmatrix}8\\0\end{bmatrix}, \begin{bmatrix}0.4 & 2.0\\2.0 & 7\end{bmatrix}\right)$$

a.计算每个维度的边际分布。

b.计算每个边际分布的平均值、模式和中位数。 c.计算二维分布的平均值和模式。

6.3 你写了一个计算机程序，有时能编译，有时不能编译（代码不改变）。你决定用参数为μ的伯努利分布来模拟编译器的明显随机性（成功与不成功）x。

$$p(x \mid \mu) = \mu^x(1-\mu)^{1-x}, \quad x \in \{0,1\}$$

为Bernoulli可能性选择一个共轭先验，并计算正面的后验分布$p(\mu \mid x_1, \ldots, x_N)$。

6.4 有两个袋子。第一个袋子里有四个芒果和两个苹果；第二个袋子里有四个芒果和四个苹果。

我们还有一枚有偏见的硬币，它显示 "正面 "的概率为0.6，显示 "反面 "的概率为0.6 如果0.4.硬币显示 "正面"，我们从袋子1中随机挑选一个水果；否则我们从袋子中随机挑选一个水果 2.

你的朋友抛出硬币（你看不到结果），从相应的袋子里随机挑选一个水果，然后送给你一个芒果。

芒果是从2号袋中采摘的，其概率是多少？

提示：使用贝叶斯定理。

6.5 考虑一下时间序列模型

$$\begin{aligned} x_{t+1} &= Ax_t + \boldsymbol{w}, & \boldsymbol{w} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}) \\ y_t &= Cx_t + \boldsymbol{v}, & \boldsymbol{v} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{R}) \end{aligned}$$

其中，$\boldsymbol{w}$，$\boldsymbol{v}$是i.i.d.高斯噪声变量。此外，假设$p(\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$。

a.p(x0, $\boldsymbol{x}_1$. .) 的形式是什么？ ..... , x_T）的形式是什么？说明你的答案（你不

b.假设$p(x_t \mid \boldsymbol{y}_1, \ldots, y_t) = \mathrm{N}\,\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t$。（

1. 计算p($x_{t+1} \mid \boldsymbol{y}_1$,..... , y_t）。
2. 计算p($x_{t+1}, y_{t+1} \mid \boldsymbol{y}_1$,..... , y_t）。
3. 在时间$t+1$， 我们观察到值$y_{t+1} = \hat{\boldsymbol{y}}$。 计算条件分布$p(x_{t+1} \mid \boldsymbol{y}_1$。 ..... , y_{t+1}）。

6.6 证明(6.44)中的关系，它将方差的标准定义与方差的原始分数表达式联系起来。

6.7 证明(6.45)中的关系，该关系将数据集中的实例之间的成对差异与方差的原始分数表达联系起来。

6.8 用前指数族的自然参数形式表示伯努利分布，见(6.107).

6.9 将二项分布表示为指数族分布。也表示贝塔分布是一个指数族分布。证明贝塔分布和二项分布的乘积也是指数族的成员。

6.10 以两种方式推导出第6.5.2节中的关系。

a.通过完形填空

b.通过将高斯以其指数族的形式表达出来

两个高斯的*乘积*$\mathrm{N}(\boldsymbol{x} \mid \boldsymbol{a}, \boldsymbol{A})\,\mathrm{N}(\boldsymbol{x} \mid \boldsymbol{b}, \boldsymbol{B})$是一个非标准化的高斯分布c $\mathrm{N}\,\boldsymbol{x} \mid \boldsymbol{c}, \boldsymbol{C}$与$\boldsymbol{C} = (\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1})^{-1}$

$$\boldsymbol{c} = \boldsymbol{C}(\boldsymbol{A}^{-1}\boldsymbol{a} + \boldsymbol{B}^{-1}\boldsymbol{b})$$

$$c = (2\pi)^{-\frac{D}{2}} |\boldsymbol{A} + \boldsymbol{B}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{a} - \boldsymbol{b})^{\mathrm{T}}(\boldsymbol{A} + \boldsymbol{B})^{-1}(\boldsymbol{a} - \boldsymbol{b})\right)。$$

请注意，归一化常数c本身可以被认为是一个（归一化）高斯分布，在$\boldsymbol{a}$或$\boldsymbol{b}$中都有一个 "膨胀的 "协方差矩阵$\boldsymbol{A} + \boldsymbol{B}$，即$c = \mathrm{N}\,\boldsymbol{a} \mid \boldsymbol{b}, \boldsymbol{A} + \boldsymbol{B} = \mathrm{N}\,\boldsymbol{b} \mid \boldsymbol{a}, \boldsymbol{A} + \boldsymbol{B}$。

6.11 **迭代的期望。**

考虑两个随机变量x，y的联合分布$p(x, y)$。说明

$$\mathrm{E}_{Y}[xx] = \mathrm{E}_{Y}\mathrm{E}_{X}[x \mid y]。$$

这里，$\mathrm{E}_{X}[x \mid y]$表示在条件分布$p(x \mid y)$下x的期望值。

6.12 **操纵高斯随机变量。**

考虑一个高斯随机变量$\boldsymbol{x}\ \boldsymbol{x} \sim \mathrm{N}\,\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x$，其中$\boldsymbol{x} \in \mathrm{R}^D$.

此外，我们有

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} + \boldsymbol{w}。$$

其中$\boldsymbol{y} \in \mathrm{R}^E$，$\boldsymbol{A} \in \mathrm{R}^{E \times D}$，$\boldsymbol{b} \in \mathrm{R}^E$，和$\boldsymbol{w}\ \boldsymbol{w} \sim \mathrm{N}(0, \boldsymbol{Q})$是独立的高斯噪声。"独立 "意味着$\boldsymbol{x}$和$\boldsymbol{w}$是独立的随机变量，$\boldsymbol{Q}$是对角线的。

a.写下这种 $p(\boldsymbol{y} \mid \boldsymbol{x})$ 可能性。

b.分布p($\boldsymbol{y}$)= $\int p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$是高斯的。计算平均数$\boldsymbol{\mu}_y$和协方差$\boldsymbol{\Sigma}_y$。详细推导出你的结果。

c.随机变量$\boldsymbol{y}$正根据测量映射进行转换。

$$\boldsymbol{z} = \boldsymbol{C}\boldsymbol{y} + \boldsymbol{v},$$

其中$\boldsymbol{z} \in \mathbb{R}^F$，$\boldsymbol{C} \in \mathbb{R}^{F \times E}$，$\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R})$是独立的高斯噪声（测量）。

- 写下$p(\boldsymbol{z} \mid \boldsymbol{y})$。
- 计算$p(\boldsymbol{z})$，即平均值$\boldsymbol{\mu}_z$和协方差$\boldsymbol{\Sigma}_z$。详细推导出你的结果。

d.现在，一个值$\hat{\boldsymbol{y}}$被测量了。计算后验分布$p(\boldsymbol{x} \mid \hat{\boldsymbol{y}})$。

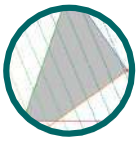
解决办法的提示。这个后验也是高斯的，也就是说，我们只需要确定其平均值和协方差矩阵。首先明确地计算出联合高斯$p(\boldsymbol{x}, \boldsymbol{y})$。这也要求我们计算交叉协方差$\boldsymbol{Cov}_{\boldsymbol{x},\boldsymbol{y}}[\boldsymbol{x}, \boldsymbol{y}]$和$\boldsymbol{Cov}_{\boldsymbol{y},\boldsymbol{x}}[\boldsymbol{y}, \boldsymbol{x}]$。然后应用高斯调节的规则。

6.13 概率积分转换

给定一个连续随机变量X，其cdf为$F_X(x)$，表明运行中的变量$Y := F_X(X)$是均匀分布的（定理6.15).

7

持续优化



由于机器学习算法是在计算机上实现的，数学公式被表述为数值优化方法。本章介绍了训练机器学习模型的基本数值方法。训练一个机器学习模型往往可以归结为找到一组好的参数。良好 "的概念是由目标函数或概率模型决定的，我们将在本书的第二部分看到这些例子。给定一个目标函数，寻找最佳值是通过优化算法

完成的。 由于我们考虑

在R中D的数据和模型，我们面临的优化问题是*连续*优化问题，而不是离散变量的*组合*优化问题。

本章包括连续优化的两个主要分支（图7.1）：无约束的和有约束的优化。在本章中，我们将假设我们的目标函数是可分的（见第五章），因此我们可以在空间的每个位置获得梯度，以帮助我们找到最佳值。按照惯例，机器学习中的大多数目标函数都是要最小化的，也就是说，最佳值就是最小值。直观地讲，寻找最佳值就像寻找目标函数的梯度，而梯度指向我们上坡。我们的想法是向下坡移动（与梯度相反），并希望找到最深的点。对于无约束优化，这是我们唯一需要的概念，但有几个设计选择，我们将在第二节7.1.讨论。7.2).我们还将介绍一类特殊的问题（第7.3节中的凸优化问题），在这里我们可以对达到全局最优进行陈述。

考虑图中7.2.的函数 该函数有一个*全局最小值* 全局最小值

围绕$x=4.5$，其函数值约47为 。由于该函数是 "平滑的"，梯度可以用来帮助找到最小值，表明我们应该向右或向左走一步。

的最小值

这假定我们是在正确的碗里，因为存在另一个*局部*

围绕$x=0.7$*的最小值*。回顾一下，我们可以解决所有的静止的

静止的点是导数的实根，即梯度为零的点。

通过计算一个函数的导数并将其设为零，来确定该函数的点。对于

$$f(x) = x^4 + 7x^3 + 5x^2 - 17x + , \quad (37.1)$$

我们得到相应的梯度为

$$\frac{df(x)}{dx} = 4x^3 + 21x^2 + 10x - 17. \quad (7.2)$$

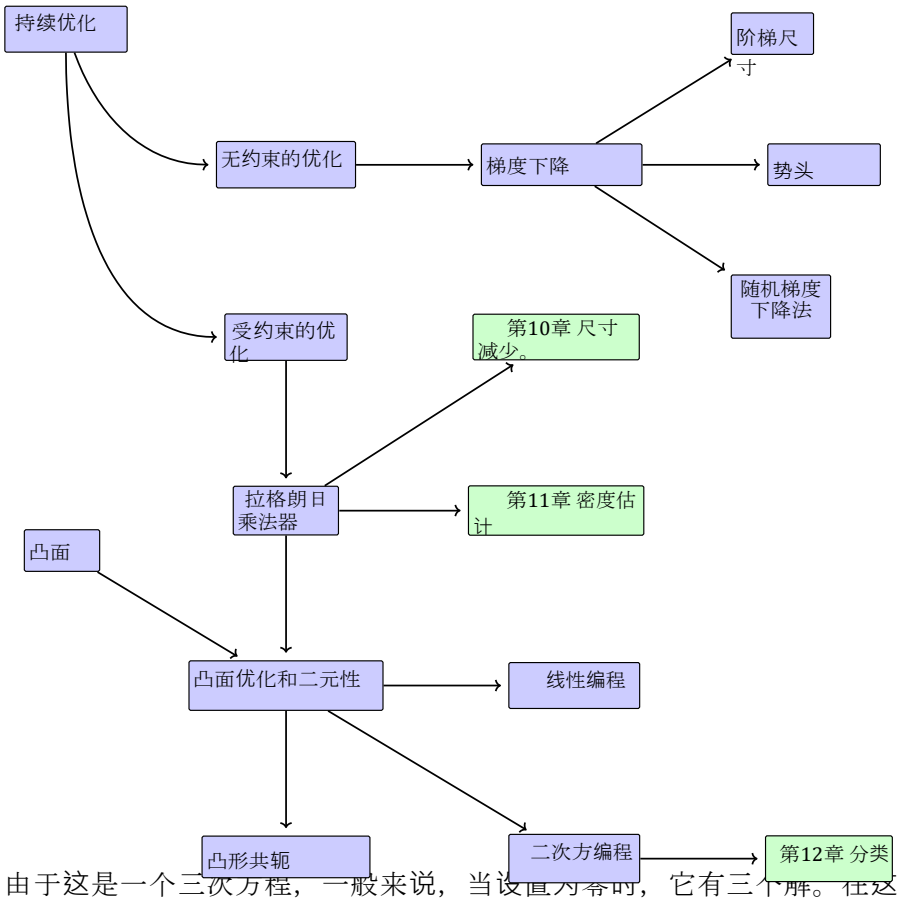


图 本章所介绍的与优化有关的概念的**7.1**思维导图。有两个主要概念：梯度下降和凸式优化。

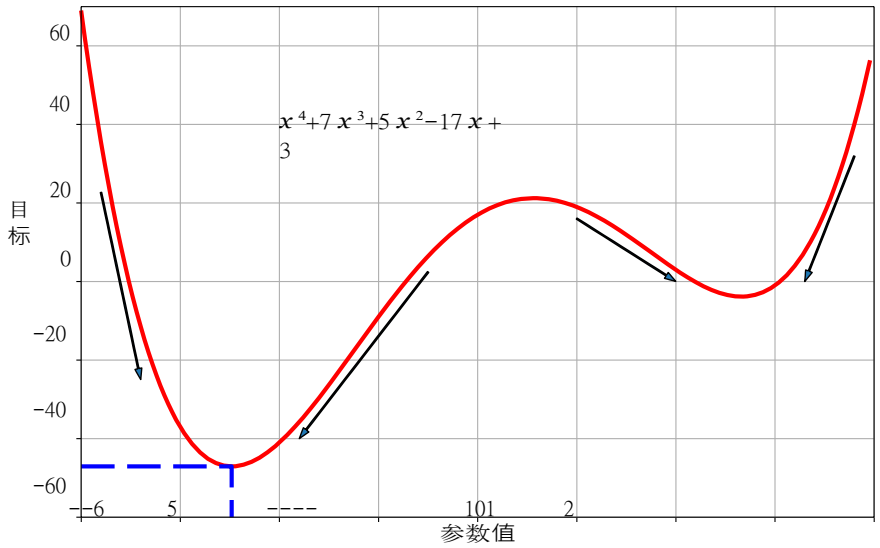
由于这是一个三次方程，一般来说，当设置为零时，它有三个解。在这个例子中，其中两个是最小值，一个是最大值（在x=1.附近4）。为了检查一个静止点是否是最小值或最大值，我们需要二次求导，并检查静止点的二次导数是正还是负。在我们的例子中，第二导数是

$$\frac{^{2}df(x)}{dx^{2}} = 12x^{2} + 42\,x + . \qquad 10(7.3)$$

通过替换我们目测的x=-4的值。5, -(1.4 , 0. 7, we $\frac{d^{2}f(x)}{dx^{2}} < 0$ 我们可以看到，正如预期的那样，中间点是一个最大值。

而另外两个静止点是最小值。

请注意，在前面的讨论中，我们避免了对x值进行分析求解，尽管对于低阶多项式，如前面提到的，我们可以这样做。一般来说，我们无法找到解析解，因此我们需要从某个值开始，比如x_0=6 ，然后沿着负梯度前进。负梯度表明我们应该去



图例7.2目标函数。负的梯度用箭头表示，而全球最小值是所表示的蓝色虚线。

准确，但不是多远（这被称为步长）。此外，如果我们如果从右边开始（例如$x_0=0$），负的梯度就会把我们带到错误的最小值。图7.2说明了这样一个事实：对于$x > 1$，负的梯度指向图中右边的最小值，它的目标值较大。

根据阿贝尔-鲁菲尼定理，一般来说，对于度数5以上的多项式没有代数解（阿贝尔，1826）。

在第7.3,我们将了解一类被称为凸函数的函数，它们对优化算法的起点不表现出这种棘手的依赖性。对于凸函数，所有的局部最小值是全局最小值。事实证明，许多机器学习的目标职能的设计使它们是凸的，我们将在第12章看到一个例子。

为凸函数所有局部最小值都是全局最小值。

本章到目前为止的讨论是关于一维函数的，在这里我们能够直观地看到梯度、下降方向和最优值的想法。在本章的其余部分，我们将在高维上发展同样的想法。不幸的是，我们只能在一个维度上对这些概念进行可视化，但有些概念并不能直接推广到更高的维度上，因此在阅读时需要注意一些。

7.1 使用梯度下降法进行优化

我们现在考虑求解一个实值函数的最小值问题

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}), \tag{7.4}$$

其中f :R^d R是一个目标函数，可以捕捉到手头的机器学习问题。我们假设我们的函数$\vec{f}$是可微的，而且我们无法通过分析找到一个封闭形式的解决方案。

我们使用行向量的惯例来表示梯度。

梯度下降是一种一阶优化算法。为了用梯度下降法找到一个函数的局部最小值，需要采取与当前点的函数梯度的负数成比例的步骤。回顾第5.1节，梯度指向最陡峭的上升方向。另一个有用的直觉是考虑函数处于某个值（$f(\boldsymbol{x}) = c$ 对于某个值cR）的线条集合，这被称为等高线。梯度∈指向的方向与我们希望优化的函数的轮廓线正交。

让我们考虑多变量函数。想象一个表面（由函数$f(\boldsymbol{x})$描述），有一个球从一个特定的位置$\boldsymbol{x}$开始$_0$，当球被释放时，它将沿着最陡峭的d-scent方向下山。梯度下降法利用了这样一个事实：如果从$\boldsymbol{x}$*开始*$_0$沿着f在$\boldsymbol{x}$处$_0$的负梯度（(f) $(\boldsymbol{x}_0)$）$^{\mathrm{T}}$的方向移动，$f(\boldsymbol{x}_0)$ −∇下降得最快。

$$\boldsymbol{x}_1 = \boldsymbol{x}_0 - \gamma((\nabla f)(\boldsymbol{x}_0))^{\mathrm{T}} \tag{7.5}$$

对于一个小的*步长*γ;?0那么，$f(\boldsymbol{x}_1) \leqslant f(\boldsymbol{x}_0)$。请注意，我们使用梯度的转置，否则尺寸将无法计算。

这一观察使我们可以定义一个简单的梯度下降算法。如果我们想找到一个函数$f(\boldsymbol{x}_*)$的局部最优值。R^n →R，$\boldsymbol{x}$l$f(\boldsymbol{x})$，我们从我们希望优化的参数的初始猜测$\boldsymbol{x}$开始$_0$，然后根据以下方法进行迭代

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \gamma_i((\nabla f)(\boldsymbol{x}_i))^{\mathrm{T}}。 \tag{7.6}$$

对于合适的步长γ_i，序列$f(\boldsymbol{x}_0)$;?$f(\boldsymbol{x}_1)$;? . ....收敛到一个局部最小值。

例子 7.1

考虑一个二维的二次函数

$$f\begin{bmatrix}x_1\\x_2\end{bmatrix} = \frac{1}{2}\begin{bmatrix}x_1\\x_2\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}2 & 1\\1 & 20\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix} - \begin{bmatrix}5\\3\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}x_1\\x_2\end{bmatrix} \tag{7.7}$$

有梯度的

$$\nabla f\begin{bmatrix}x_1\\x_2\end{bmatrix} = \begin{bmatrix}x_1\\x_2\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}2 & 1\\1 & 20\end{bmatrix} - \begin{bmatrix}5\\3\end{bmatrix}^{\mathrm{T}}. \tag{7.8}$$

从初始位置$\boldsymbol{x}_0 = [-3, -1]$开始$^{\mathrm{T}}$，我们迭代地应用(7.6)
以获得一连串收敛于最小值的估计值

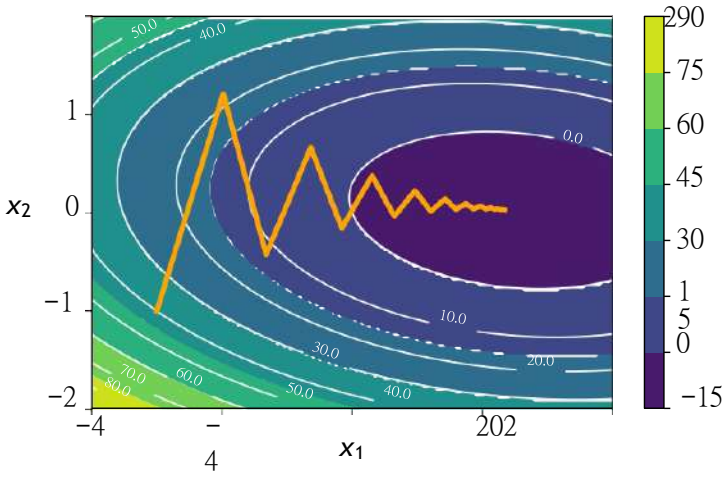


图 7.3梯度下降在一个二维二次元表面（以热图形式显示）。请看例子7.1的描述。

(图中说明了7.3).我们可以看到（从图中以及通过将$\boldsymbol{x}_0$插入(7.8)，在$\boldsymbol{x}_0$点的负梯度为0..085北方和东方，导致$\boldsymbol{x}_1 = [-1.98, 1.21]^\top$。重复这个论点得到$\boldsymbol{x}_2 = [-1.32, -0.42]^\top$，以此类推。

备注。梯度下降法在接近最小值时可能相对缓慢。它的渐进收敛率比许多其他方法要差。根据球滚下山的比喻，当表面是一个长而细的山谷时，问题的条件很差（Trefethen和Bau III，1997）。对于条件较差的凸问题，梯度下降法的效果越来越好。

"之字形"，因为梯度几乎正交于到最小点的最短直线；见图7.3. ◆

7.1.1 台阶大小

如前所述，选择一个好的步长在梯度上是很重要的。

梯度下降。如果步长太小，梯度下降会很慢。如果步长也是

- 如果步长选择得过大，梯度下降就会过冲，无法达成共识，甚至出现发散。我们将在下一节讨论动量的使用。这是一种可以平滑梯度上升的不稳定行为并抑制振荡的方法。

当函数值下降时，步长可以更大。试着增加步长。

自适应梯度方法在每次迭代时重新调整步长，这取决于函数的局部属性。有两种简单的方法（Toussaint，2012）。

- 当函数值在一个梯度步骤后增加时，步骤大小太大。撤销该步骤并减小步长。

称为学习率。

虽然 "撤销 "步骤似乎是一种资源浪费，但使用这种启发式方法可以保证单调收敛。

例子（7.2解决一个线性方程组）。
当我们解决$\boldsymbol{Ax}=\boldsymbol{b}$形式的线性方程时，在实践中我们会解决
通过找到最小平方误差的$\boldsymbol{x}_*$，$\boldsymbol{Ax}-\boldsymbol{b}$=近似值$\mathbf{0}$

$$\|\boldsymbol{Ax}-\boldsymbol{b}\|^2=(\boldsymbol{Ax}-\boldsymbol{b})^{\mathrm{T}}(\boldsymbol{Ax}-\boldsymbol{b}) \tag{7.9}$$

如果我们使用欧几里得准则。(7.9)相对于$\boldsymbol{x}$的梯度是

$$\nabla_{\boldsymbol{x}}=2(\boldsymbol{Ax}-\boldsymbol{b})^{\mathrm{T}}\boldsymbol{A}\ 。 \tag{7.10}$$

我们可以在梯度下降算法中直接使用这个梯度。然而，对于这个特殊的情况，事实证明有一个分析解，可以通过设置梯度为零来找到。我们将在第九章看到更多关于解决平方误差问题的内容。

*备注。*当应用于线性方程组$\boldsymbol{Ax}=\boldsymbol{b}$的求解时，梯度下降可能收敛缓慢。梯度下降法的收敛速度取决于*条件数*$\kappa=\frac{\sigma(\boldsymbol{A})_{\max}}{\sigma(\boldsymbol{A})_{\min}}$，其中是最大奇异值与最小奇异值的比率（第4.5节）。

条件编号

条件数本质上衡量的是最弯曲的方向与最不弯曲的方向的比率，这与我们的想象相一致，即条件差的问题是长而细的山谷。它们在一个方向上非常弯曲，但在另一个方向上非常平坦。而不是二-
直接求解$\boldsymbol{Ax}=\boldsymbol{b}$，可以改成求解$\boldsymbol{P}^{-1}(\boldsymbol{Ax}-\boldsymbol{b})=\mathbf{0}$，其中 -
$\boldsymbol{P}$被称为*预处理程序*。我们的目标是设计$\boldsymbol{P}$，使$\boldsymbol{P}^{-1}\boldsymbol{A}$有一个更好的条件数，但同时$\boldsymbol{P}^{-1}$也很容易计算。关于梯度下降、预处理和收敛的进一步信息，我们参考Boyd和Vandenberghe（2004，第9章）。

先决条件

◆

7.1.2 带动量的梯度下降

正如图中所示7.3,如图所示，如果优化表面的曲率是这样的，即有一些区域的比例很低，那么梯度下降的收敛就会非常慢。曲率是这样的，梯度下降的步骤在谷壁之间跳动，并以小的步骤接近最佳状态。为改善收敛性而提出的调整是给梯度下降一些内存。

Goh(2017)写了一篇关于梯度下降与动力的直观博文。

带动量的梯度下降法（Rumelhart等人，1986）是一种引入附加项来记忆前一次迭代中发生的情况的方法。这种记忆可以抑制震荡，使梯度更新更加平滑。继续用球来比喻，动量项模仿了一个不愿意改变方向的重

球的现象。我们的想法是让梯度更新具有记忆，以实现

移动平均线。基于动量的方法记住了更新
$\Delta_i \boldsymbol{x}$在每个迭代i，并将下一次更新确定为当前和之前梯度的线性组合

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \gamma_i((\nabla f)(\boldsymbol{x}_i))^{\mathrm{T}} + \alpha\Delta\boldsymbol{x}_i \tag{7.11}$$

$$\Delta\boldsymbol{x}_i = \boldsymbol{x}_i - \boldsymbol{x}_{i-1} = \alpha_{i-1}\Delta\boldsymbol{x} - \gamma_{i-1}((\nabla f)(\boldsymbol{x}_{i-1}))^{\mathrm{T}}, \tag{7.12}$$

其中$\alpha[0, \infty]$。有时我们只知道近似的梯度。在这种情况下，动量项是很有用的，因为它可以平均不同的梯度的噪声估计。获得近似梯度的一个特别有用的方法是使用随机近似，我们接下来讨论这个方法。

7.1.3 随机梯度下降法

计算梯度可能是非常耗时的。然而，通常可以找到一个 "廉价 "的梯度近似值。只要梯度的方向大致相同，近似的梯度仍然是有用的。
作为真实梯度。

随机梯度

血统

随机梯度下降法（通常简称为SGD）是梯度下降法的随机应用，用于最小化写成可微分函数之和的目标函数。这里的随机一词指的是，我们承认我们并不确切知道梯度，而只是知道它的一个嘈杂的近似值。通过限制近似梯度的概率分布，我们仍然可以在理论上保证SGD会收敛。

在机器学习中，给定$n = 1, \ldots, N$个数据点，我们经常考虑的目标函数是每个例子n所产生的损失L_n的总和。 在数学符号中，我们有如下形式

$$L(\boldsymbol{\theta}) = \sum_{n=1}^{N} L_n(\boldsymbol{\theta}), \tag{7.13}$$

其中$\boldsymbol{\theta}$是感兴趣的参数向量，也就是说，我们要找到最小化L的$\boldsymbol{\theta}$。回归（第9章）的一个例子是负对数可能性，它被表示为单个例子的对数可能性之和，所以

$$L(\boldsymbol{\theta}) = -\sum_{n=1}^{N} \log p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}), \tag{7.14}$$

其中$\mathbf{x} \in \mathbb{R}^D$是训练输入，$y_n$是训练目标，而$\boldsymbol{\theta}$
是回归模型的参数。

如前所述，标准梯度下降法是一种 "批量 "优化方法，即使用全部训练集进行优化。

通过更新参数向量，根据

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \gamma_i (\nabla \mathrm{L}(\boldsymbol{\theta}_i))^{\mathrm{T}} = \boldsymbol{\theta}_i - \gamma_i \sum_{n=1}^{N} (\nabla \mathrm{L}_n(\boldsymbol{\theta}_i))^{\mathrm{T}} \tag{7.15}$$

评估梯度总和可能需要对所有单个函数L的梯度进行昂贵的评估$_{ni}$。当训练集很大和/或没有简单的公式存在时，评估梯度总和变得非常昂贵。

考虑到$\sum_{n=}^{LN}(\nabla L_{倪}(\boldsymbol{\theta}))$项。7.15).我们可以减少与批量梯度下降不同的是，批量梯度下降使用的是n=1，.....的所有L_{nn}，我们随机选择L的一个子集来进行小批量梯度下降。，N，我们随机选择L的一个子集$_n$进行小批量梯度下降。在极端情况下，我们只随机选择一个L_n来估计梯度。关于为什么取一个数据子集是明智的，关键的见解是要认识到
对于梯度下降的收敛，我们只要求梯度是一个
对真实梯度的无偏估计。事实上，术语$\sum_{n=}^{LN}(\nabla L_{倪}(\boldsymbol{\theta}))$
中的(7.15)是对梯度预期值的经验估计（第6.4.1节）。因此，任何其他无偏的经验估计值，例如使用数据的任何子样本，都足以保证梯度下降的收敛性。

*备注。*当学习率以适当的速度下降时，在相对温和的假设下，随机梯度下降几乎肯定会收敛到局部最小值（Bottou,1998）。 ◆

为什么要考虑使用近似梯度？一个主要原因是实际执行的限制，比如中央处理器（CPU）/图形处理器（GPU）的内存大小或计算时间的限制。我们可以用估计经验平均数时的样本大小的方式来考虑用于估计梯度的子集的大小（第1节）。6.4.1).大批量的小样本将提供准确的梯度估计，减少参数更新的方差。此外，在成本和梯度的矢量实现中，大批量的小批量利用了高度优化的矩阵操作。方差的减少导致了更稳定的收敛，但每次梯度计算都会更昂贵。

相比之下，小型的迷你批次的估计是很快的。如果我们保持小批量，我们梯度估计中的噪声会让我们走出一些不好的局部优化，否则我们可能会陷入其中。在机器学习中，优化方法是通过在训练数据上最小化一个目标函数来进行训练的，但总体目标是提高泛化性能（第8章）。由于机器学习的目标不一定需要精确估计目标函数的最小值，所以使用小批量方法的近似梯度已经被广泛使用。随机梯度下降在大规模机器学习问题中非常有效（Bottou等人，2018）。

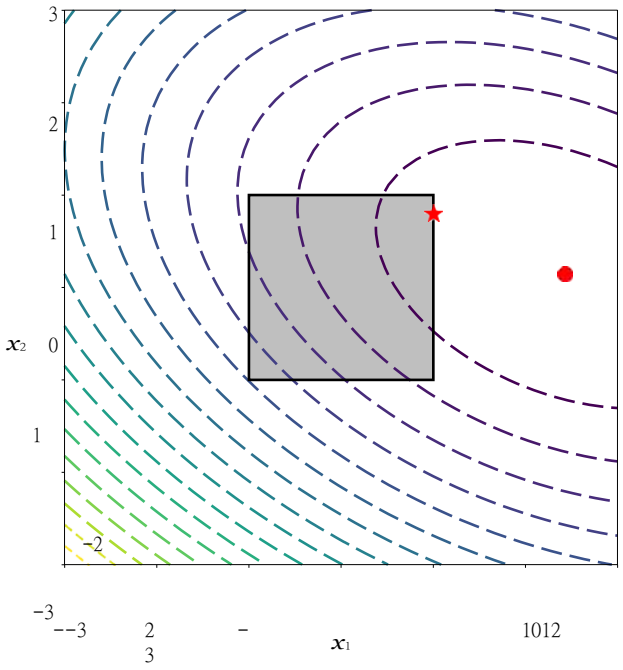


图 7.4 说明受约束的优化。实验结果无约束的问题（用等高线表示线）有一个右侧的最小值（表示由圆圈决定）。的。箱体限制 (-1 … *x* … 和1 - 1 $\leq y \leq$ 1) 需要-即最佳解决方案在盒子里，结果是一个最佳值星星表示的。

如在数百万张图片上训练深度神经网络（Dean等人，2012）、主题模型（Hoffman等人，2013）、强化学习（Mnih等人，2015），或训练大规模高斯过程模型（Hensman等人，2013；Gal等人，2014）。

7.2 有约束的优化和拉格朗日乘法器

在上一节中，我们考虑了求解一个函数的最小值的问题

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}), \tag{7.16}$$

其中f:R $^{D}\rightarrow$ R。

在本节中，我们有额外的约束。就是说，对于实值函数g_i :R^D R为i =1, … ., m，我们考虑有约束的优化问题（见图7.4为例）。

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \tag{7.17}$$

0 对于所有的 i =1, … …, m来说，受制于$og_i(\boldsymbol{x})$。，m.

值得指出的是，函数f和g_i在一般情况下可能是非凸的，我们将在下一节考虑凸的情况。

将有约束的问题（）转换成无约束的问题的一个明显但不太实用的方法是使用指标函数。7.17)转换为无约束的问题，就是使用一个指标函数

$$J(\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \mathbf{1}(g_i(\boldsymbol{x})), \tag{7.18}$$

where $\mathbf{1}(z)$ is an infinite step function

$$\mathbf{1}_0(z) = \begin{cases} 0 & \text{如果} z\,. \\ \infty & \text{否则} \end{cases} \tag{7.19}$$

如果不满足约束条件，这就会得到无限的惩罚，因此会提供相同的解决方案。然而，这种无限步函数同样难以优化。我们可以通过引入*拉格朗日乘法器*来克服这个困难。拉格朗日乘法器的概念是用一个线性函数来代替步骤函数。

拉格朗日乘法器 拉格朗日理论

我们对问题(7.17)的*拉格朗日*，引入拉格朗日乘数$\lambda_i \geq 0$对应于每个不等式约束的再光谱（Boyd和Vandenberghe，2004，第4章），以便

$$L(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{x}) \tag{7.20a}$$

$$= f(\boldsymbol{x}) + \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x})\,, \tag{7.20b}$$

在最后一行中，我们将所有的约束条件$g_i(\boldsymbol{x})$串联成一个向量$\boldsymbol{g}(\boldsymbol{x})$，将所有的拉格朗日乘数串联成一个向量$\boldsymbol{\lambda} \in \mathbb{R}^m$。

我们现在介绍拉格朗日对偶性的概念。一般来说，优化中的二重性是指将一组变量$\boldsymbol{x}$（称为原始变量）中的优化问题转换为另一组不同变量$\boldsymbol{\lambda}$（称为二重变量）中的优化问题。我们介绍两种不同的对偶性方法。在这一节中，我们讨论拉格朗日对偶性；在第二节中7.3.3,我们讨论Legendre-Fenchel对偶性。

定义 中的**7.1.**问题是(7.17)

$$\min_{\boldsymbol{x}} \; f(\boldsymbol{x}) \tag{7.21}$$
$$0 \qquad \text{对于所有的} \qquad i = 1, \ldots, m$$

首要问题 拉格朗日对偶问题

被称为*原始问题*，与原始变量x相对应，相关的*拉格朗日对偶问题*由以下公式给出

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}^m} \; D(\boldsymbol{\lambda}) \tag{7.22}$$

最小不等式

受制于 $\boldsymbol{\lambda} \geq \mathbf{0}$，在关于定义的讨论中，7.1,我们使用了两个同样具有*鞍点*意义的概念（Boyd和Vandenberghe，2004）。

其中$\boldsymbol{\lambda}$是对偶变量，$D(\boldsymbol{\lambda}) = \min_{\boldsymbol{x} \in \mathbb{R}^d} L(\boldsymbol{x}, \boldsymbol{\lambda})$。

备注。

首先是*最小不等式*，它说对于任何有两个参数$j(\boldsymbol{x}, \boldsymbol{y})$的函数，最大化小于最小值，即。

$$\max_{\boldsymbol{y}} \min_{\boldsymbol{x}} \phi(\boldsymbol{x}, \boldsymbol{y}) \leq \min_{\boldsymbol{x}} \max_{\boldsymbol{y}} \phi(\boldsymbol{x}, \boldsymbol{y}) \text{。} \tag{7.23}$$

这个不等式可以通过考虑不等式来证明

$$\text{对于所有的} \boldsymbol{x},\ \text{ymin}\, j(\boldsymbol{x}, \boldsymbol{y}) \leqslant \max_{\boldsymbol{xy}} j(\boldsymbol{x}, \boldsymbol{y}) \text{。} \tag{7.24}$$

请注意，取()左手边的$\boldsymbol{y}$上的最大值可以保持不等式，因为对所有的y都是真实的。7.24)的左手边保持不等式，因为不等式对所有的$\boldsymbol{y}$都是真的。同样，我们可以取(7.24)的右边的最小值，得到(7.23).

第二个概念是*弱二元性*，它用(7.23)来说明弱对偶性

原始值总是大于或等于对偶值。这一点将在(7.27). ◆

回顾一下，在(7.18)中的J(x)与(7.20b)中的拉格朗日的区别在于，我们已经将指标函数放宽为线性函数。因此，当λ;?0时，拉格朗日L（$\boldsymbol{x}$, $\boldsymbol{\lambda}$）是J（$\boldsymbol{x}$）的下限。因此，L($\boldsymbol{x}$, $\boldsymbol{\lambda}$)相对于$\boldsymbol{\lambda}$的最大值是

$$J(\boldsymbol{x}) = \max_{O} \mathsf{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \text{。} \tag{7.25}$$

回顾一下，最初的问题是最小化$J(\boldsymbol{x})$。

$$\min_{\boldsymbol{x}\in R^d} \max_{\boldsymbol{\lambda}\ O} \mathsf{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \text{。} \tag{7.26}$$

根据最小极大不等式(7.23)，可以看出，交换最小值和最大值的顺序会得到一个较小的值，即。

$$\min_{\boldsymbol{x}\in R^d} \max_{\boldsymbol{\lambda}\, O} \mathsf{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \text{ ; ? } \max_{\boldsymbol{\lambda}\, O} \min_{\boldsymbol{x}\in R^d} \mathsf{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \text{。} \tag{7.27}$$

这也被称为*弱二象性*。请注意，右弱对偶性的内部部分

手侧是对偶目标函数D($\boldsymbol{\lambda}$)，定义如下。与原来的优化问题相比，它有约束条件。

$\boldsymbol{x}\in \mathrm{R}^{Mind\ L}(\boldsymbol{x}, \boldsymbol{\lambda})$是一个针对给定$\boldsymbol{\lambda}$值的无约束优化问题。 如果解决$\boldsymbol{x}\in \mathrm{R}^{Mind\ L}(\boldsymbol{x}, \boldsymbol{\lambda})$很容易， 那么整个问题也很容易解决。我们可以从(7.20b)中看到，L($\boldsymbol{x}$, $\boldsymbol{\lambda}$)相对于λ是仿射的。 因此$\boldsymbol{x}\in \mathrm{R}^{mind\ L}(\boldsymbol{x}, \boldsymbol{\lambda})$是$\boldsymbol{\lambda}$的仿射函数的点状最小值，因此D($\boldsymbol{\lambda}$)是凹的，尽管$f(\ )$和$g_i(\ )$可能是非凸的。外部问题，即对$\boldsymbol{\lambda}$的最大化，是一个凹函数的最大值，可以有效地计算。假设$f(\ )$和$g_i(\ )$是可微的，我们通过对拉格朗日*对偶*问题进行微分，将微分设为零，然后求出最优值。我们将讨论两个第7.3.2节中的具体例子。7.3.1和7.3.2节中的具体例子，其中 f（ ）和g_i（ ）是凸的。

备注（平等约束）。考虑(7.17)有额外的平等约束

$$\text{闽}_{\boldsymbol{x}} \quad f(\boldsymbol{x})$$

$$0 \qquad \text{对于所有的} \qquad i = 1, \qquad \cdots\cdots$$

, m $h(\boldsymbol{x}_i)$ = 对于所有的j =, (7.28)
.....。 , $m\ h_j(\boldsymbol{x}) = 0$ for $all j = 1, \ldots, n$.

我们可以用两个不等式约束来代替平等约束的模型。也就是说，对于每个平等约束$h_j(\boldsymbol{x})=0$，我们可以用两个约束$h_j(\boldsymbol{x})\leq 0$和$h_j(\boldsymbol{x})\geq 0$代替。事实证明，所得到的拉格朗日乘数是没有约束的。

因此，我们将与(7.27)中的不平等约束相对应的拉格朗日乘数约束为非负值，而将与平等约束相对应的拉格朗日乘数不约束。(7.28)中的不平等约束所对应的拉格朗日乘数为非负值，而让平等约束所对应的拉格朗日乘数不受约束。

◆

7.3 凸面优化

我们的注意力集中在一类特别有用的优化问题上，我们可以保证全局最优。当$f(\cdot)$是一个凸函数，并且当涉及$g(\cdot)$和$h(\cdot)$的约束是凸集时，这被称为*凸优化问题*。在这种情况下，我们*有强对偶性*。对偶问题的最优解与原始问题的最优解是一样的。凸函数和凸集之间的区别在机器学习文献中往往没有严格的表述，但人们往往可以从上下文中推断出隐含的意义。

凸式优化问题

强二元性

凸集

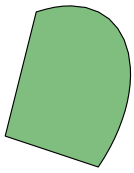
定义 一个**7.2.**集合是一个*凸集合*，如果对于任何$x, y \in \mathcal{C}$和任何标量的θ0的情况下1，我们有

$$\theta x + (1-\theta)y \in \mathcal{C} \text{。} \tag{7.29}$$

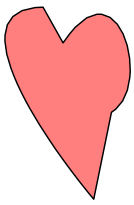
图例**7.5**

凸集

的。 凸集是指这样的集合，即连接集合中任何两个元素的直线位于该集合内。图7.5和7.6分别显示了凸集和非凸集。



图为非凸集的例子**7.6**。



凸函数 凹函数

凸函数是指这样的函数，即函数的任何两点之间的直线位于函数的上方。图中7.2显示了一个非凸函数，而图7.3显示了一个凸函数。另一个凸形

题记

函数如图所示 240 $+(1-\theta)f(\boldsymbol{y})$。 *持续优化* (7.30)
7.7.

*备注。凹形函数*是凸形函数

定义 7.3。设函数 $f:\mathbb{R}^D \to$ 的负数。 ◆

是一个域为凸集的函数。如果对于 f 涉及 $g(\ )$ 和 $h(\ )$ 的约束条件在(7.28)中涉及g( )和h( )的约束在一个标量值处截断了函数，从而产生了集合。凸函数和凸集合之间的另一个关系是考虑通过 "填入 "凸函数而得到的集合。凸函数是一个类似于碗的物体，我们想象一下把水倒进碗里，把它填满。这个被填满的集合，被称为 "凸函数 "的*表征*。

的域中的所有 $\boldsymbol{x}, \boldsymbol{y}$，以及对于任何标量 θ，且 $0 \ldots \theta \ldots 1$，我们有 凸函数，是一个凸集。

如果一个函数 $f:\mathbb{R}^n \to \mathbb{R}$ 是可微的，我们可以在以下方面指定凸性

$$f(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) \leqslant \theta f(\boldsymbol{x})$$

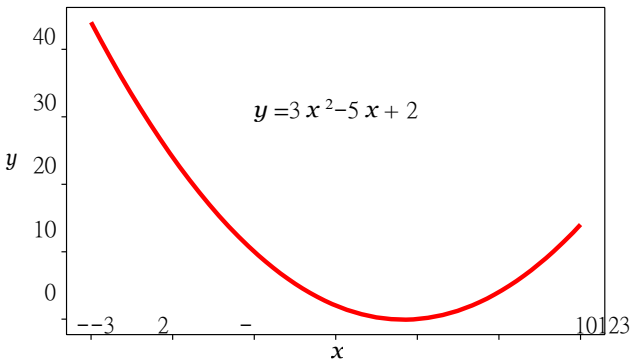


图 凸形的例子7.7
功能。

它的梯度∇x f（x）（第5.2节）。一个函数f (x)是凸的，当且仅当对任何两点x，y而言，其成立为

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} f(\boldsymbol{x})^{\top}(\boldsymbol{y} - \boldsymbol{x}) \text{。} \tag{7.31}$$

如果我们进一步知道一个函数f (x)是两次可微的，即Hessian (5.147)对x域中的所有值都存在，那么当且仅当$\nabla_{\boldsymbol{x}}^2 f(\boldsymbol{x})$是正半无限的时候，函数f (x)是凸的（Boyd 和 Vandenberghe,2004）。

例子 7.3

负熵f (x) = x log2 x对于x >来说0是凸的。图中是该函数的可视化图7.8,，我们可以看到该函数是凸的。为了说明前面的凸性定义，让我们检查一下两点x=和2x=的计算结果4。注意，为了证明凸性我们需要检查所有点x∈R的f（x）。

回顾定义 7.3.考虑两点之间的中间位置的一个点(即θ=0.)5；那么左手边是f (0. 5-2 + 0.5 - 4) = log3 2 3≈。4.75.右手边是0 . 5(2 log2 2) +0 . 5(4 log2 4) = +1 =45 . 并且因此，定义得到满足。

由于f (x)是可微的，我们可以选择使用(7.31).计算f（x）的导数，我们得到

$$\nabla(x \log_2 x) = 1 - \log x + x - \frac{1}{x \log_e 2} = \text{对数} x + \frac{1}{\text{原本} 2}. \tag{7.32}$$

使用相同的两个测试点x =和2x =4，( )的左手边由f (4)= 。7.31)的左手边是f (4) = 8。右手边是

$$f(\boldsymbol{x}) + \nabla_{\boldsymbol{x}}(\boldsymbol{y} - \boldsymbol{x}) = f(2) + \nabla f(2) \cdot (4 - 2) \tag{7.33a}$$

$$= +2\,(1 + \frac{1}{\text{原本} 2}) - 2 \approx 6.9\,. \tag{7.33b}$$

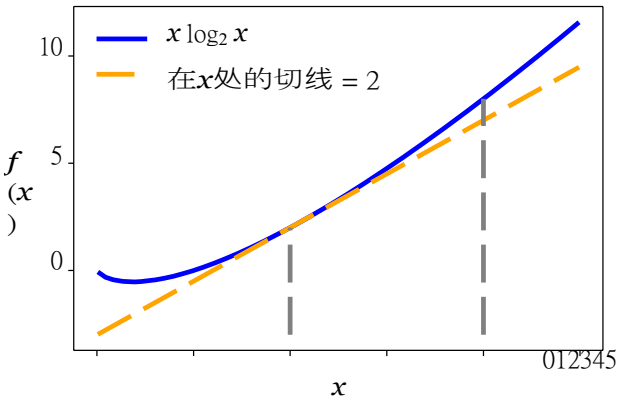


图 7.8负熵功能（这是凸）和其在x =2 的切线。

我们可以通过回顾定义，从第一原理上检查出一个函数或集合是凸的。在实践中，我们经常依靠预先服务于凸性的操作来检查一个特定的函数或集合是凸的。虽然细节上有很大的不同，但这又是我们在第二章中为向量空间介绍的闭合思想。

例子 7.4

凸函数的非负加权和是凸的。请注意，如果f是一个凸函数，而α;? 是0一个非负的标量，那么函数αf是凸的。我们可以通过将α乘以定义中方程的两边来了解这一点。7.3,并回顾一下，乘以一个非负数并不改变不等式。

如果f_1和f_2是凸函数，那么根据定义我们有

$$f(_1\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) \leqslant \theta f_1(\boldsymbol{x}) + (1-\theta)f_1(\boldsymbol{y}) \tag{7.34}$$

$$f(_2\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) \leqslant \theta f_2(\boldsymbol{x}) + (1-\theta)f_2(\boldsymbol{y}) \tag{7.35}$$

。

将两边相加，我们可以得到

$$f(_1\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) + f_2(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) \tag{7.36}$$

$$\theta f(_1\boldsymbol{x}) + (1-\theta)f_1(\boldsymbol{y}) + \theta f_2(\boldsymbol{x}) + (1-\theta)f_2(\boldsymbol{y})\ 。$$

其中右手边可以重新排列为

$$\theta(f_1(\boldsymbol{x}) + f_2(\boldsymbol{x})) + (1-\theta)(f_1(\boldsymbol{y}) + f_2(\boldsymbol{y})) \tag{7.37}$$

完成了凸函数之和是凸的证明。

结合前面两个事实，我们可以看到$\alpha f(_1\boldsymbol{x}) + \beta f_2(\boldsymbol{x})$对于$\alpha, \beta$;?0 是凸的。这个闭合属性可以用类似的论证来扩展，用于两个以上凸函数的非负加权和。

备注。( )中的不等式有时被称为詹森不等式。7.30)中的不等式有时被称为*詹森不等式*。 詹森不等式

事实上，一整类用于取凸函数的非负加权和的不等式都被称为詹森不等式。 ◆

综上所述，一个有约束的优化问题被称为*凸凸*凸优化。

的问题， 如果 问题

$$\begin{aligned} &\min_{\boldsymbol{x}} f(\boldsymbol{x}) \\ &\text{遵循 } g_i(\boldsymbol{x}) \leq 0 \quad \text{对于所有} \quad i = 1, \\ &\quad \ldots, \; m \; h_j(\boldsymbol{x}) = 0 \quad \text{对于所有} \\ &\quad j = 1, \ldots, n, \end{aligned} \tag{7.38}$$

其中所有函数f（$\boldsymbol{x}$）和g_i（$\boldsymbol{x}$）都是凸函数，所有h_j（$\boldsymbol{x}$）=0都是凸集。在下文中，我们将描述两类被广泛使用和充分理解的凸优化问题。

7.3.1 线性编程

考虑当所有前面的函数都是线性时的特殊情况，即：。

$$\begin{aligned} &\text{闽} \quad \boldsymbol{cx} \qquad \text{T} \\ &\boldsymbol{x} \in R^d \\ &\text{在} \boldsymbol{Ax} \leq \boldsymbol{b} \text{的条件下} \quad 。 \end{aligned} \tag{7.39}$$

其中$\boldsymbol{A} \in \mathbb{R}^{m \times d}$，$\boldsymbol{b} \in \mathbb{R}^m$，这就是所谓的*线性程序*。它有$d$个变量和$m$个线性约束。拉格朗日由以下公式给出

线性程序

线性程序是工业中最广泛使用的方法之一。

$$L(\boldsymbol{x}, \boldsymbol{\lambda}) = \boldsymbol{cx}^{\mathrm{T}} + \boldsymbol{\lambda}(^{\mathrm{T}}\boldsymbol{Ax} - \boldsymbol{b}), \tag{7.40}$$

其中$\boldsymbol{\lambda} \in \mathbb{R}^m$是非负的拉格朗日乘数的矢量。对应于$\boldsymbol{x}$的条款进行重新测算，可以得到

$$L(\boldsymbol{x}, \boldsymbol{\lambda}) = (\boldsymbol{c} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda})^{\mathrm{T}}\boldsymbol{x} - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{b}。 \tag{7.41}$$

取$L(\boldsymbol{x}, \boldsymbol{\lambda})$相对于$\boldsymbol{x}$的导数，并将其设为零，就可以得到

$$\boldsymbol{c} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} = . \qquad \boldsymbol{0} \tag{7.42}$$

因此，对偶拉格朗日是$D(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{b}$。 回顾一下，我们想使$D(\boldsymbol{\lambda})$最大化。除了由于$L(\boldsymbol{x}, \boldsymbol{\lambda})$的导数为零的约束外，我们还有一个事实，即$\boldsymbol{\lambda}$;?$\boldsymbol{0}$，导致

下面是一个双重优化问题 。 最小化原始的并最大限度地提高双。

$$\begin{aligned} &\text{最大} \quad -\boldsymbol{b}^{\mathrm{T}}\boldsymbol{\lambda} \\ &\boldsymbol{\lambda} \in R^m \\ &\text{主题 } \boldsymbol{toc} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} = \boldsymbol{0} \end{aligned} \tag{7.43}$$

这也是一个线性程序，但有m个变量。我们可以选择解决原始程序 (7.39)
或对偶(7.43)程序，这取决于

不管是m还是d大。回顾一下，d是变量的数量，m是原始线性程序中约束的数量。

例子 (7.5线性程序)

考虑线性程序

$$\min_{\boldsymbol{x}\in\mathbb{R}^2} \begin{bmatrix} -5 \\ -3 \end{bmatrix}^\top \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\text{受制于} \quad \begin{bmatrix} 2 & -4 \\ -1 & -2 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leq \begin{bmatrix} 33 \\ 8 \\ 5 \\ -1 \\ 8 \end{bmatrix} \tag{7.44}$$

有两个变量。这个程序也显示在图7.9中。目标函数是线性的，结果是线性轮廓线。标准形式的约束集被转化为图例。最佳值必须位于阴影（可行）区域内，并由星形表示。

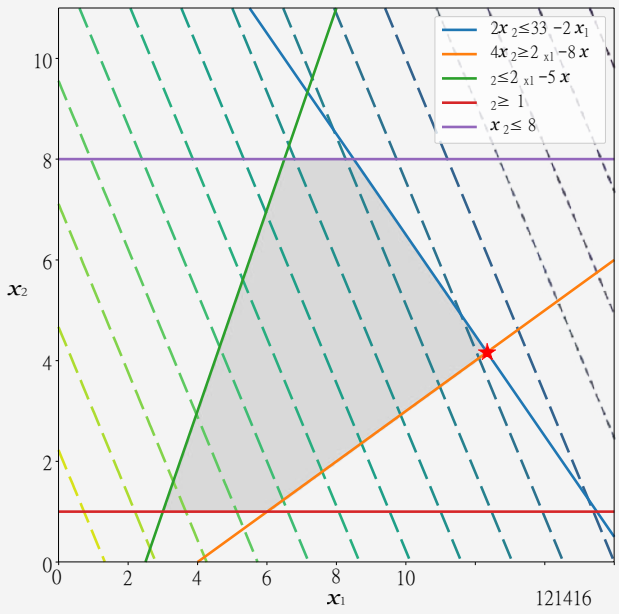


Figure 7.9 Illustration of a linear program. The unconstrained problem (indicated by the contour lines) has a minimum on the right side. The optimal value given the constraints are shown by the star.

7.3.2 二次方编程

考虑到凸二次元目标函数的情况，其中的约束是仿生的，即：。

$$\min_{\boldsymbol{x}\in R^d} \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \tag{7.45}$$

在$\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$的条件下 。

其中$\boldsymbol{A} \in \mathbb{R}^{m\times d}$、$\boldsymbol{b} \in \mathbb{R}^m$和$\boldsymbol{c} \in \mathbb{R}^d$。方形对称矩阵$\boldsymbol{Q} \in \mathbb{R}^{d\times d}$是正定的，因此，目标函数是凸的。这就是所谓的*二次方程序*。请注意，它有d个变量和m个线性约束。

例子（7.6二次方程序）。

考虑到二次方程序

$$\min_{\boldsymbol{x}\in\mathbb{R}^2} \frac{1}{2}\begin{bmatrix}x_1\\x_2\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}2 & 0\\0 & 4\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix} + \begin{bmatrix}5\\3\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}x_1\\x_2\end{bmatrix} \tag{7.46}$$

受制于

$$\begin{bmatrix}-1 & 0\\0 & 1\\0 & -1\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix} \leq \begin{bmatrix}1\\1\\1\end{bmatrix} \tag{7.47}$$

的两个变量。图7.4也说明了该程序。目标函数是二次函数，有一个正的半无限矩阵$\boldsymbol{Q}$，结果是椭圆的轮廓线。最佳值必须位于阴影（可行）区域内，用星形表示。

拉格朗日由以下公式给出

$$\mathrm{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) \tag{7.48a}$$

$$= \frac{1}{2}\begin{bmatrix}\boldsymbol{x}\\\boldsymbol{\lambda}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{Q} & \boldsymbol{A}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}\\\boldsymbol{\lambda}\end{bmatrix} + (\boldsymbol{c} + \boldsymbol{\lambda})^{\mathrm{T}}\boldsymbol{x} - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{b}, \tag{7.48b}$$

其中我们再次重新排列了这些条款。取$\mathrm{L}(\boldsymbol{x}, \boldsymbol{\lambda})$相对于$\boldsymbol{x}$的导数并将其设为零，可以得到

$$\boldsymbol{Q}\boldsymbol{x} + (\boldsymbol{c} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda}) = \boldsymbol{0}. \tag{7.49}$$

假设$\boldsymbol{Q}$是可逆的，我们得到

$$\boldsymbol{x} = -\boldsymbol{Q}^{-1}(\boldsymbol{c} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda}) 。 \tag{7.50}$$

将(7.50)到原始拉格朗日$\mathrm{L}(\boldsymbol{x}, \boldsymbol{\lambda})$中，我们得到对等拉格朗日

$$\mathrm{D}(\boldsymbol{\lambda}) = -\frac{1}{2}(\boldsymbol{c} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda})^{\mathrm{T}}\boldsymbol{Q}^{-1}(\boldsymbol{c} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda}) - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{b} 。 \tag{7.51}$$

因此，双重优化问题是由

$$\begin{aligned} &\max_{\boldsymbol{\lambda}\in\mathbb{R}^m} && -\frac{1}{2}(\boldsymbol{c}+\boldsymbol{A}^\top\boldsymbol{\lambda})^\top\boldsymbol{Q}^{-1}(\boldsymbol{c}+\boldsymbol{A}^\top\boldsymbol{\lambda})-\boldsymbol{\lambda}^\top\boldsymbol{b} \\ &\text{受制于} && \boldsymbol{\lambda}\geqslant\boldsymbol{0} \end{aligned} \tag{7.52}$$

我们将在第12章看到二次编程在机器学习中的应用。

7.3.3 Legendre-Fenchel变换和凸面共轭

让我们重新审视一下第二节中的对偶性概念。7.2,中的对偶思想，而不考虑约束条件。关于凸集的一个有用的事实是，它可以被其支持的超平面等效地描述。如果一个超平面与凸集相交，它就被称为凸集的*支持超平面*，并且

支持

超平面

的凸集只包含在它的一个侧面。回顾一下，我们可以把一个凸函数填满，得到外延，这是一个凸集。因此，我们也可以用它们的支持超平面来描述凸函数。此外，观察一下，支持超平面正好接触到凸函数，实际上是函数在该点的切线。并回顾一下，函数$f(\boldsymbol{x})$在某一点$\boldsymbol{x}$处$_0$的切线是在这一点上对该函数的梯度进行评估$\frac{\mathrm{d}f(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}}\Big|_{\boldsymbol{x}=\boldsymbol{x}_0}$。在综上所述，由于凸集可以等效地由其上位超平面描述，凸函数可以等效地由其梯度的函数描述。*Legendre变换*将这一概念正式化。

Legendre变换

物理学专业的学生经常会被介绍到Legendre变换，因为它与经典力学中的拉格朗日和哈密尔顿有关。

Legendre-Fenchel变换

凸共轭

我们从最一般的定义开始，不幸的是，这个定义有一个反直觉的形式，然后看一下特殊情况，把这个定义与前一段描述的直觉联系起来。*Legendre-Fenchel变换*是一个从凸可微函数$f(\boldsymbol{x})$到取决于切线$s(\boldsymbol{x})=\nabla_x f(\boldsymbol{x})$的函数的变换（在傅里叶变换的意义上）。值得强调的是，这是函数$f()$的变换，而不是变量$\boldsymbol{x}$或在$\boldsymbol{x}$处求值的函数。Legendre-Fenchel变换也被称为*凸共轭*（原因我们很快就会看到），并与二重性密切相关（Hiriart-Uruty和Lemaréchal， 2001, 第五章）。

定义7.4。一个函数f的*凸共轭*。$\mathbb{R}^D\to\mathbb{R}$是一个函数$f$，定义*如下

凸共轭

$$f^*(\boldsymbol{s})=\sup_{\boldsymbol{x}\in\mathbb{R}^D}\{\langle\boldsymbol{s},\boldsymbol{x}\rangle-f(\boldsymbol{x})\}。 \tag{7.53}$$

请注意，前面的凸共轭定义不需要函数f是凸的，也不需要是可微的。

在定义中，7.4，我们使用了一般的内积（第3.2节），但在本节的其余部分，我们

将考虑有限维向量之间的标准点积($\boldsymbol{s}, \boldsymbol{x} = \boldsymbol{s}\boldsymbol{x}^{\mathrm{T}}$)，以避免太多的技术细节。(

)为了理解定义7.4以几何学的方式来理解，考虑一个很好的推导。最容易理解的是在推理过程中画出推理。

请注意，由于我们看的是一个一维的问题，超平面可以简化为一条线[2]。考虑一条直线$y=sx+c$。回顾一下，我们能够通过支持超平面来描述凸函数，所以让我们尝试通过支持线来描述这个函数$f(x)$。固定直线$s \in \mathbb{R}$的梯度，对于f图形上的每一点$(x_0, f(x_0))$，找出c的最小值，使直线仍然与$(x_0, f(x_0))$相交。请注意，c的最小值是斜率为s的直线 "刚刚接触 "函数$f(x)=x^2$的地方。

$$y - f(x_0) = s(x - x_0)。 \tag{7.54}$$

这条线的y截点是$sx_0+f(x_0)$。- c的最小值，对于它

因此，$y=sx+c$与f的图形相交是

$$\inf_{x_0} sx_0 + f(x_0)。 \tag{7.55}$$

按照惯例，前面的凸共轭被定义为这个的负数。本段的推理并不依赖于以下事实

我们选择了一个一维的凸和可微的函数，而对于$f: \mathbb{R}^D \to \mathbb{R}$，它们是非凸的和不可微分的。

经典的Legendre变换是在$\mathbb{R}^D$中的凸可微函数上定义的。

*备注。*凸可微函数，如例子$f(x) = x^2$是一个很好的特例，这里不需要上位数，而且函数和它的Legendre反式之间有一个一一对应的关系。让我们从第一原理中推导出这一点。对于一个凸可微函数，我们知道在x_0处，切线接触到$f(x_0)$，所以

$$f(x_0) = sx_0 + c。 \tag{7.56}$$

回顾一下，我们想用它的梯度$\nabla_x f(x)$来描述凸函数$f(x)$，并且$s = \nabla_x f(x_0)$。我们重新排列以得到$-c$的表达式，从而得到

$$-c = sx_0 - f(x_0)。 \tag{7.57}$$

请注意，$-c$随x_0的变化而变化，因此也随s的变化而变化，这就是为什么我们可以把它看作是s的一个函数，我们称之为

$$f^*(s) := sx_0 - f(x_0)。 \tag{7.58}$$

对比(7.58)与定义7.4,相比较，我们看到(7.58)是一个特例（没有上位数）。 ◆

共轭函数有很好的特性；例如，对于凸函数，再次应用Legendre变换

会让我们回到原始函数。就像$f(x)$的斜率是s一样，$f^*(s)$的斜率是 *持续优化*

以下两个例子显示了凸共轭在机器学习中的常见用途。

例子（7.7凸形共轭物）

为了说明凸共轭的应用，考虑二次函数

$$f(\boldsymbol{y}) = \frac{\lambda}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{y}^{-1} \tag{7.59}$$

基于正定矩阵$\boldsymbol{K} \in \mathbb{R}^{n\times n}$，我们表示原始的变量为$\boldsymbol{y} \in \mathbb{R}^n$，对偶变量为$\boldsymbol{a} \in \mathbb{R}^n$。

应用定义，7.4,我们得到函数

$$f^*(\boldsymbol{a}) = \sup_{\boldsymbol{y}\in\mathbb{R}^n} (\boldsymbol{y}, \boldsymbol{a}) - \frac{\lambda}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{y}^{-1} \tag{7.60}$$

由于该函数是可微的，我们可以通过取导数和关于$\boldsymbol{y}$的导数来找到最大值，并将其设为零。

$$\frac{\partial\,(\boldsymbol{y},\ \boldsymbol{a}) - \frac{\lambda}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{y}^{-1}\ ]}{\partial \boldsymbol{y}} = (\boldsymbol{a} - \lambda\boldsymbol{K}\boldsymbol{y})^{\mathrm{T}} \tag{7.61}$$

因此，当梯度为零时，我们有$\boldsymbol{y} = \boldsymbol{K}^{-1}\boldsymbol{a}$。代入进入(7.60)得出的结果是

$$f^*(\boldsymbol{a}) = \frac{1}{\lambda}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{a} - \frac{\lambda}{2}\frac{1}{\lambda}\boldsymbol{a}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{K}^{-1}\frac{1}{\lambda}\boldsymbol{K}\boldsymbol{a} = \frac{1}{2\lambda}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{a} \tag{7.62}$$

例子 7.8

在机器学习中，我们经常使用函数之和；例如，训练集的目标函数包括训练集中每个ex- ample的损失之和。在下文中，我们推导出凸共轭的的损失之和$f(t)$，其中f:这也说明了应用的问题。凸共轭物对矢量情况的描述。让$\mathrm{L}(\boldsymbol{t}) = \sum_{i=1}^{n} f_i(t_i)$然后

$$\mathrm{L}^*(\boldsymbol{z}) = \sup_{\boldsymbol{t}\in\mathbb{R}^n} (\boldsymbol{z}, \boldsymbol{t}) - \sum_{i=1}^{n} f_i(t_i) \tag{7.63a}$$

$$= \sup_{\boldsymbol{t}\in\mathbb{R}^n} \sum_{i=1}^{n} z_i t_i - f_i(t_i) \quad \text{点积的定义} \tag{7.63b}$$

$$= \sum_{i=1}^{n} \sup_{t_i\in\mathbb{R}} z_i t_i - f_i(t_i) \tag{7.63c}$$

$= \sum_{i=1}^{n} f_i(z_i)$ 共轭的 (7.63d) 定义

回顾一下，在第7.2中，我们使用拉格朗日乘法器得出了一个对偶优化问题。此外，对于凸优化问题，我们有很强的对偶性，即原始问题和对偶问题的解相匹配。这里描述的Legendre-Fenchel变换也可以用来推导出一个对偶优化问题。此外，当函数是凸的和可微的时候，最高值是唯一的。为了进一步研究这两种方法之间的关系，让我们考虑一个线性平等约束的凸优化问题。

例子 7.9

设$f(\boldsymbol{y})$和$g(\boldsymbol{x})$为凸函数，$\boldsymbol{A}$为适当维度的实矩阵，使$\boldsymbol{Ax}=\boldsymbol{y}$。

$$\min_{\boldsymbol{x}\boldsymbol{A}\boldsymbol{x}=\boldsymbol{y}} f(\boldsymbol{A}\boldsymbol{x}) + g(\boldsymbol{x}) = \min f(\boldsymbol{y}) + g(\boldsymbol{x}) \text{。} \tag{7.64}$$

通过引入约束条件$\boldsymbol{Ax}=\boldsymbol{y}$的拉格朗日乘数$\boldsymbol{u}$。

$$\min_{\boldsymbol{A}\boldsymbol{x}=\boldsymbol{y}} f(\boldsymbol{y}) + g(\boldsymbol{x}) = \min_{\boldsymbol{x},\boldsymbol{y}} \max_{\boldsymbol{u}} f(\boldsymbol{y}) + g(\boldsymbol{x}) + (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y})^{\mathrm{T}}\boldsymbol{u} \tag{7.65a}$$

$$= \max_{\boldsymbol{u}} \min_{\boldsymbol{x},\boldsymbol{y}} f(\boldsymbol{y}) + g(\boldsymbol{x}) + (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y})^{\mathrm{T}}\boldsymbol{u} \tag{7.65b}$$

其中最后一步交换最大和最小是由于f（$\boldsymbol{y}$）和g（$\boldsymbol{x}$）是凸函数。通过拆分点积项，收集$\boldsymbol{x}$和$\boldsymbol{y}$。

$$\max_{\boldsymbol{u}} \min_{\boldsymbol{x},\boldsymbol{y}} f(\boldsymbol{y}) + g(\boldsymbol{x}) + (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y})^{\mathrm{T}}\boldsymbol{u} \tag{7.66a}$$

$$= \max_{\boldsymbol{u}} \min_{\boldsymbol{y}} -\boldsymbol{y}^{\mathrm{T}}\boldsymbol{u} + f(\boldsymbol{y}) + \min_{\boldsymbol{x}} (\boldsymbol{A}\boldsymbol{x})^{\mathrm{T}}\boldsymbol{u} + g(\boldsymbol{x}) \tag{7.66b}$$

$$= \max_{\boldsymbol{u}} \min_{\boldsymbol{y}} -\boldsymbol{y}^{\mathrm{T}}\boldsymbol{u} + f(\boldsymbol{y}) + \min_{\boldsymbol{x}} \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{u} + g(\boldsymbol{x}) \tag{7.66c}$$

回顾一下凸共轭（定义7.4）以及点状突起的事实

For general inner products, we have to replace $\boldsymbol{A}^{\mathrm{T}}$ with the adjoint $\boldsymbol{A}^{*}$.

控件是对称的。

$$\max_{\boldsymbol{u}} \min_{\boldsymbol{y}} -\boldsymbol{y}^{\mathrm{T}}\boldsymbol{u} + f(\boldsymbol{y}) + \min_{\boldsymbol{x}} \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{u} + g(\boldsymbol{x}) \tag{7.67a}$$

$$= \max_{\boldsymbol{u}} -f^{*}(\boldsymbol{u}) - g^{*}(-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{u}) \tag{7.67b}$$

因此，我们已经证明，

$$\min_{\boldsymbol{x}} f(\boldsymbol{A}\boldsymbol{x}) + g(\boldsymbol{x}) = \max_{\boldsymbol{u}} -f^{*}(\boldsymbol{u}) - g^{*}(-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{u}) \tag{7.68}$$

。

Legendre-Fenchel共轭对可以表示为凸优化问题的数学学习问题非常有用。特别是，对于独立适用于每个例子的凸损失函数，共轭损失是推导对偶问题的一种方便方法。

7.4 进一步阅读

连续优化是一个活跃的研究领域，我们并不试图对最近的进展提供一个全面的说明。

从梯度下降的角度来看，有两个主要的弱点，这两个弱点都有自己的一套文献。第一个挑战是，梯度下降是一个一阶算法，不使用关于表面曲率的信息。当有长谷的时候，梯度会垂直于感兴趣的方向。动量的概念可以被概括为一类通用的加速方法（Nesterov,2018）。共轭梯度方法通过考虑以前的方向来避免梯度下降所面临的问题（Shewchuk,1994）。二阶方法，如牛顿方法，使用Hessian来提供曲率信息。通过考虑目标函数的曲率，产生了许多选择步长和动量等想法（Goh，2017；Bottou等人，2018）。L-BFGS等准牛顿方法试图使用更便宜的计算方法来接近Hessian（Nocedal和Wright，2006）。最近，人们对计算下降方向的其他度量产生了兴趣，产生了诸如镜像下降（Beck和Teboulle,2003）和自然梯度（Toussaint,2012）等方法。

第二个挑战是如何处理非微分的函数。当函数中存在结点时，梯度方法就不能很好地定义。在这些情况下，可以使用*亚梯度方法*（Shor,1985）。关于优化非微分函数的更多信息和算法，我们可以参考Bertsekas(1999)的书。有大量关于数值解决连续优化问题的不同方法的文献，包括约束性优化问题的算法。Luenberger(1969)和Bonnans等人(2006)的书是欣赏这些文献的良好起点。Bubeck(2015)提供了一份关于连续优化的最新调查。

Hugo Gonçalves的博客也是一个很好的资源，可以更容易地介绍Legendre-Fenchel变换：https://tinyurl.com/ydaal7hj

现代机器学习的应用往往意味着数据集的大小禁止使用批量梯度下降，因此，随机梯度下降是目前大规模机器学习方法的主力。最近的文献调查包括Hazan（2015）和Bot- tou等人（2018）。

对于对偶性和凸优化，Boyd和Vanden- berghe(2004)的书包括在线讲座和幻灯片。Bertsekas(2009)提供了更多的数学处理方法，而最近的一本书是由Bertsekas的一位朋友所写的。

优化领域的主要研究人员是Nesterov（2018）。凸优化以凸分析为基础，对凸函数的更多基础性结果感兴趣的读者可以参考Rock- afellar(1970)、Hiriart-Uruty and Lemar échal(2001)以及Borwein and Lewis(2006)。Legendre-Fenchel变换在上述关于凸分析的书籍中也有涉及，但在Zia等人(2009)的书中有一个更适合初学者的预处理。Polyak(2016)对Legendre-Fenchel变换在凸优化算法分析中的作用进行了调查。

练习

7.1 考虑单变量函数

$$f(x) = x^3 + 6x^2 - 3x - 5.$$

找到它的静止点，并指出它们是最大值、最小值还是鞍点。

7.2 考虑随机梯度下降的更新方程（公式（7.15））。

写下当我们使用一个迷你批次大小的更新。

7.3 请考虑以下说法是真的还是假的。

a.任何两个凸集的交点都是凸的。 b.任何两个凸集的联合都是凸的。

c.一个凸集A与另一个凸集B的差是凸的。

7.4 请考虑以下说法是真的还是假的。

a.任何两个凸函数的和是凸的。 b.任何两个凸函数的差是凸的。 c.任何两个凸函数的积是凸的。 d.任何两个凸函数的最大值是凸的。

7.5 用矩阵符号将以下优化问题表示为标准线性程序

$$\max_{x \in \mathbb{R}^2,\ \xi \in \mathbb{R}} \ p^T x + \xi$$

受制于$\xi \geq 0$, $x_0 \leq 3$和$x_1 \leq 3$。

7.6 考虑图7.9所示的线性程序。

$$\min_{x \in \mathbb{R}^2} \ -5x_1 - 3x_2$$

受

$$\begin{bmatrix} 2 & 2 \\ 1 & 2 \\ 0 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leq \begin{bmatrix} 8 \\ 5 \\ -1 \\ 8 \end{bmatrix}$$

利用拉格朗日对偶性推导出对偶线性程序。

7.7 考虑图7.4所示的二次方程序。

$$\min_{x\in R^2} \frac{1}{2} x^T \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix} x + \begin{bmatrix} -1 \\ 0 \end{bmatrix}^T x + 5$$

$$\text{受制于} \quad \begin{bmatrix} 0 & -1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} x \leq \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

利用拉格朗日对偶性推导出对偶二次方程序。

7.8 考虑以下凸式优化问题

$$\min_{w\in R^D} \frac{1}{2} w^T w$$

主体是$y_n w^T x_n \geq 1$。

通过引入拉格朗日乘数λ，推导出拉格朗日对偶。

7.9 考虑$x\in\mathbb{R}^D$的负熵。

$$f(x) = \sum_{d=1}^{D} x_d \log x_d .$$

通过假设标准点积，推导出凸共轭函数$f^*(s)$。

提示：取一个适当的函数的梯度，并将梯度设为零。

7.10 考虑一下函数

$$f(x) = \frac{1}{2} x^T A x + b^T x + c \text{。}$$

其中A是严格正定的，这意味着它是可逆的。推导出$f(x)$的凸共轭。

提示：取一个适当的函数的梯度，并将梯度设为零。

7.11 铰链损失（这是支持向量机使用的损失）由以下公式给出

$$L(\alpha) = \max\{0, 1 - \alpha\}.,$$

如果我们对应用梯度方法（如L-BFGS）感兴趣，而不想求助于子梯度方法，我们需要平滑铰链损失中的扭结。计算铰链损失的凸共轭$L^*(\beta)$，β是对偶变量。增加一个*英镑2的*近似项，并计算所得函数的共轭值

$$L^*(\beta) + \frac{\gamma}{2}\beta^2 \text{。}$$

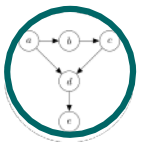
其中γ是一个给定的超参数。

第二部分

中央机器学习问题

8

当模型遇到数据时



在本书的第一部分，我们介绍了构成许多机器学习方法基础的数学。希望读者能够从第一部分中学习到数学语言的初级形式，我们现在将用它来描述和讨论机器学习。本书的第二部分介绍了机器学习的四个支柱。

- 回归（第九章）
- 降维（第十章） 密度估计（第十一章） 分
- 类（第十二章）

本书这一部分的主要目的是说明如何利用本书第一部分介绍的数学概念来设计机器学习算法，以解决四大支柱范围内的任务。我们不打算介绍先进的机器学习概念，而是提供一套实用的方法，使读者能够应用他们从本书第一部分中获得的知识。它还为已经熟悉数学的读者提供了进入更广泛的机器学习文献的途径。

8.1 数据、模型和学习

在这一点上，值得暂停一下，考虑一下机器学习算法所要解决的问题。正如第一章所讨论的，机器学习系统有三个主要组成部分：数据、模型和学习。机器学习的主要问题是 "什么是
我们所说的好模型是指什么？"。*模型*这个词有很多微妙之处，我们 认为模型
在本章中，我们将多次重提这个问题。如何客观地定义 "好 "这个词也不完全明显。机器学习的指导原则之一是，好的模型应该在未见过的数据上表现良好。这就要求我们定义一些性能指标，如准确度或与地面真相的距离，以及找出在这些性能指标下表现良好的方法。本章涵盖了一些必要的数学和统计语言的片段，这些语言通常是

表例**8.1** 数据来自一个虚构的人资源数据库不在一个数字格式。

命名	性别	学位	邮政编号	年龄	年薪
阿迪雅	M	硕士	W21BG	36	89563
鲍勃	M	博士学位	EC1A1BA	47	123543
ÄÄÄ	F	BEcon	SW1A1BH	26	23989
大辅	M	理学士	SE207AT	68	138769
尹志平	F	MBA	SE10AA	33	113888

谈到机器学习模型的时候，我们会简要地介绍一下目前训练模型的最佳做法。通过这样做，我们简要地介绍了目前训练模型的最佳做法，使所产生的预测器在我们尚未看到的数据上表现良好。

如第一章所述，我们在使用 "机器学习算法 "这一短语时有两种不同的含义：训练和预测。我们将在本章中描述这些想法，以及在不同模型中进行选择的想法。我们将在第8.4节中介绍经验风险最小化的框架8.2,、最大似然的原则8.3,和概率模型的思想。我们在第8.4节中简要介绍了用于指定概率模型的图形语言。8.5最后在第8.6节讨论模型选择。本节的其余部分阐述了机器学习的三个主要组成部分：数据、模型和学习。

8.1.1 作为载体的数据

我们假设我们的数据可以被计算机读取，并以数字格式精确地表示。数据被假定为表格形式（图8.1)，我们认为表格的每一行都代表一个特定的立场或例子，每一列都是一个特定的特征。近年来，机器学习已经被应用于许多类型的数据，这些数据显然不是以表格的数字格式出现的，例如基因组序列、网页的文本和图像内容以及社交媒体图。我们不讨论识别好的特征的重要和挑战方面。这些方面很多都取决于领域的专业知识，并且需要仔细的工程设计，近年来，它们被归入数据科学的范畴（Stray,2016;Adhikari and DeNero,2018）。

假设数据是在一个整洁的格式（Wickham, 2014；Codd, 1990）。

即使我们有表格式的数据，也要做出选择以获得数字表示。例如，在表8.1中，性别一栏（一个分类变量）可以转换为数字0代表 "男性 "和代表1"女性"。或者说，基因-可分别用数字 -1、$+1$表示（如表所示8.2).此外，使用领域知识往往是很重要的在构建表征时，例如知道大学学位从学士到硕士再到博士，或者意识到

所提供的邮编不仅仅是一串字符，实际上是对伦敦的一个地区的编码。在表8.2,我们将表8.1中的数据转换为数字格式，每个邮编被表示为两个数字。

性别标识	学位	纬度（单位：度）	经度（单位：度）	年龄	年薪（以千计）
-1	2	51.5073	0.1290	36	89.563
-1	3	51.5074	0.1275	47	123.543
+1	1	51.5071	0.1278	26	23.989
-1	1	51.5075	0.1281	68	138.769
+1	2	51.5074	0.1278	33	113.888

表**8.2** 来自于虚构的人资源数据库(见表8.1),转换为数字格式。

一个经纬度。即使是有可能被直接读入机器学习算法的数字数据，也应该仔细考虑单位、比例和约束。在没有额外信息的情况下，我们应该对数据集的所有列进行移位和缩放，使它们的经验平均值和0经验方差分别为 1。 在本书中，我们假设领域专家已经对数据进行了适当的转换，即每个输入$\boldsymbol{x}_n$是一个D维实数矢量。

这被称为*特征*、*属性*或*协变量*。我们认为一个数据集的特征是

的形式，如表8.2.所示。 请注意，在新的数字表示法中，我们去掉了表的"姓名 "一栏。8.1在新的数字表示法中，我们去掉了表的名称一栏。这样做有两个主要原因。(1)我们不希望识别符（姓名）对机器学习任务来说是有意义的；(2)我们可能希望对数据进行匿名处理，以帮助保护雇员的隐私。

属性协变量

在本书的这一部分，我们将用N来表示数据集中的试题数量，并用小写的n =1，... ... *来索*引这些例子。，N .我们假设给我们一组数字数据，以向量数组的形式表示（表8.2).每一行都是一个特定的个体$\boldsymbol{x}_n$，通常在机器学习中被称为一个*例子*或*数据点*。子

例证

数据点

n指的是这是数据集中N个例子中的第n个例子。每一列代表一个关于这个例子的特定特征，我们将这些特征索引为d =1，. . .回顾一下，数据被表示为矢量，这意味着每个例子（每个数据点）是一个D维矢量。表格的方向源于数据库界，但对于一些机器学习算法（例如第10章），将例子表示为列向量更为方便。

让我们考虑根据表的数据，从年龄预测年薪的问题。8.2.这被称为监督学习问题

其中，我们有一个与每个例子xlabel相关的n

标签y_n（工资）。

(年龄）。标签y_n有各种其他名称，包括目标、再反应变量和注释。一个数据集被写成一组例子-标签对（$\boldsymbol{x}_1$，y_1），... ., ($\boldsymbol{x}_n$, y)$_n$, . . ., ($\boldsymbol{x}_N$, y_N)。例子的表格$\boldsymbol{x}_1, . . ., \boldsymbol{x}_N$ 通常被串联起来，并被写成X$\in \mathbb{R}^{N\times D}$。图8.1说明了由表的最右边两列组成的数据集，8.2,其中x=年龄，y=工资。

{}

我们使用本书第一部分中介绍的概念来正式说明

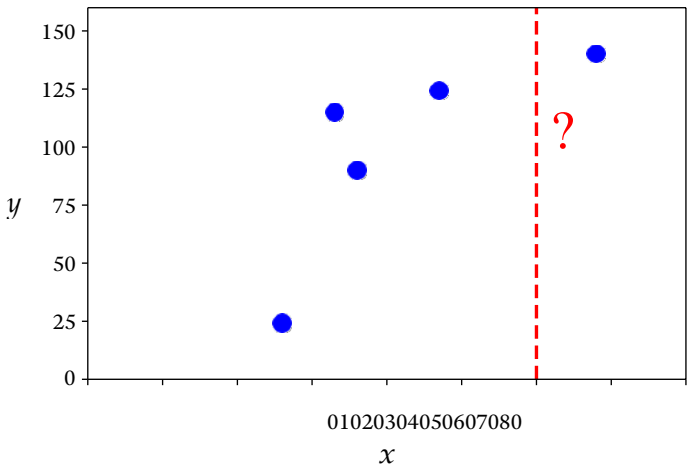


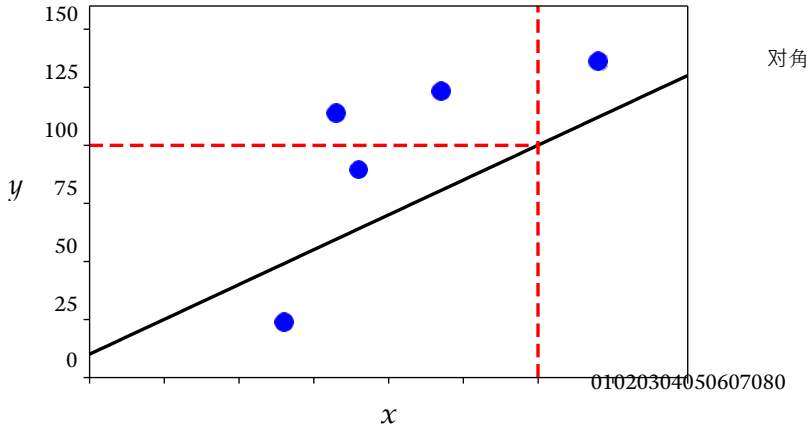
图 为线性回归的Toy**8.1**数据。训练数据（x_n, y_n）对从最右边的两个表8.2的各列。我们感兴趣的是一个60岁的人的工资（x=60）。图示为纵向的红色虚线，这不是训练数据的一部分。

诸如上一段中的机器学习问题。将数据表示为向量$\boldsymbol{x}_n$，可以让我们使用线性al-gebra的概念（在第二章介绍）。在许多机器学习算法中，我们需要额外地能够比较两个向量。正如我们在第9章和第12章中所看到的，计算两个例子之间的相似性或距离使我们能够正式确定具有相似特征的例子应该具有相似的标签这一直觉。两个向量的比较需要我们构建一个几何体（在第三章中解释），并允许我们使用第七章中的技术来优化所产生的学习问题。

既然我们有数据的矢量表示，我们就可以通过操作数据来找到潜在的更好的表示。我们将以两种方式讨论寻找好的表征：寻找原始特征向量的低维近似值，以及使用原始特征向量的非线性高维组合。在第十章中，我们将看到一个通过寻找主成分找到原始数据空间的低维近似值的例子。寻找主成分与第四章中介绍的特征值和奇异值分解的概念密切相关。对于高维表示，我们将看到一个明确的*特征图*φ（），它允许我们$_n$使用高维表示φ（$\boldsymbol{x}_n$）来表示in-put$\boldsymbol{x}$。高维表征的主要原因是我们可以将新的特征构建为原始特征的非线性组合，这反过来可以使学习问题变得更容易。我们将在本节中讨论特征图，并说明这个特征图是如何使我们的学习更容易。9.2并在第12.4节中说明这个特征图是如何导致*阿卡诺的*。近年来，深度学习方法（Goodfellow等人，2016）在使用数据本身来学习新的好的特征方面显示出前景，并且在计算机视觉、语音识别和自然语言处理等领域取得了很大的成功。在本书的这一部分，我们将不涉及神经网络，但读者可参阅

特征图

内核



图：8.2函数实例（黑色实心对角线）和其预测值在 $x = 60$，即。$f(60) = 100$.

第5.6节对反向传播的数学描述，这是训练神经网络的一个关键概念。

8.1.2 作为功能的模型

一旦我们有了适当的矢量表示法中的数据，我们就可以进入到构建一个预测函数（称为*预测器*）的业务。 预测器

在第一章中，我们还没有语言来精确描述模型。利用本书第一部分的概念，我们现在可以介绍 "模型 "的含义。我们在本书中提出了两种主要方法：作为函数的预测器和作为概率模型的预测器。我们在此对前者进行描述，后者在下一小节进行描述。

预测器是一个函数，当给定一个特定的输入例子（在我们的例子中是一个特征向量）时，产生一个输出。现在，考虑输出是一个单一的数字，即一个实值标量输出。这可以被写成

$$f : \mathbb{R}^D \to \mathbb{R}, \tag{8.1}$$

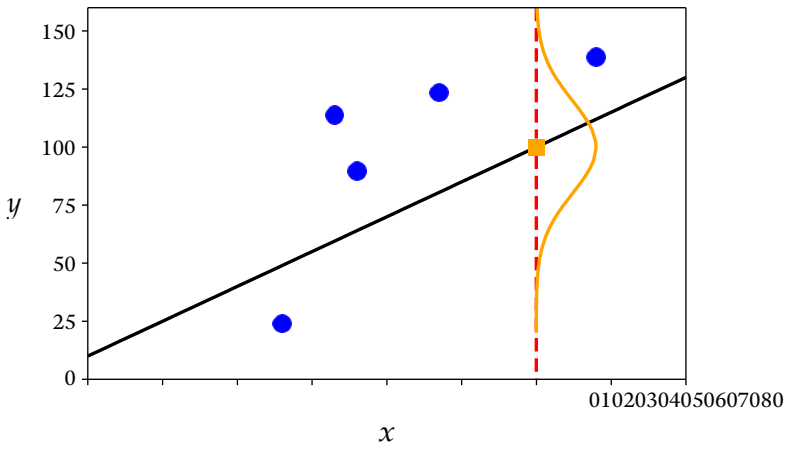
其中，输入向量$\boldsymbol{x}$是D维的（有D个特征），然后应用于它的函数f（写成$f(\boldsymbol{x})$）返回一个实数。图8.2说明了一个可能的函数，可以用来计算输入值x的预测值。

在本书中，我们不考虑所有函数的一般情况，这将涉及到对函数分析的需求。相反，我们考虑的是线性函数的特殊情况

$$f(\boldsymbol{x}) = \boldsymbol{\theta}^{\top}\boldsymbol{x} + \theta_0 \tag{8.2}$$

这一限制意味着第2章和第3章的内容有助于精确地描述一个预测器的概

非概率性（与接下来描述的概率性观点相反）。



图例**8.3**函数（黑色实心对角线）及其预测的不确定性在*x*=（60画成高斯）。

机器学习的观点。线性函数在可以解决的问题的普遍性和需要的基础数学的数量之间取得了良好的平衡。

8.1.3 作为概率分布的模型

我们经常认为数据是对一些真正的潜在效应的嘈杂观察，并希望通过应用机器学习，我们可以从噪声中识别出信号。这就要求我们有一种语言来量化噪声的影响。我们通常也希望有表达某种不确定性的预测器，例如，量化我们对某个特定测试数据点的预测值的信心。正如我们在第六章所看到的，概率论提供了一种量化不确定性的语言。图8.3说明了函数的预测不确定性是一个高斯分布。

我们可以把预测器看作是一个单一的函数，而不是把预测器看作是概率模型，即描述可能函数的分布的模型。在本书中，我们将自己限制在具有有限维参数的分布的特殊情况下，这使得我们能够描述概率模型而不需要随机过程和随机度量。对于这种特殊情况，我们可以把概率模型看作是多变量概率分布，它已经允许有一类丰富的模型。

我们将在第8.4节中介绍如何使用概率的概念（第6章）来定义机器学习模型，并在第8.5节中介绍一种以紧凑方式描述概率模型的图形语言。8.5.

8.1.4 学习就是寻找参数

学习的目标是找到一个模型及其相应的参数，从而使产生的预测器在未见过的数据上表现良好。在讨论机器学习算法时，在概念上有三个不同的算法阶段。

1.预测或推理 2.训练或参数估计
3.超参数调整或模型选择

预测阶段是指我们在以前未见过的测试数据上使用训练好的预测器。换句话说，参数和模型选择已经固定，预测器被应用于代表新输入数据点的新向量。正如第一章和上一小节所概述的，在本书中我们将考虑机器学习的两个流派，对应于预测器是一个函数还是一个概率模型。当我们有一个概率模型（在第8.4节进一步讨论），预测阶段被称为推理。

备注。 遗憾的是，对于不同的算法阶段，没有一致的命名。"推理 "一词有时也被用来表示
概率模型的参数估计，较少用于指非概率模型
的预测。 ◆

训练或参数估计阶段是我们根据训练数据来调整我们的预测模型。我们希望在给定的训练数据中找到好的预测器，这样做有两种主要策略：根据某种质量衡量标准找到最佳预测器（有时称为找到点估计），或使用贝叶斯推理。寻找点估计可以适用于这两种类型的预测器，但贝叶斯推理需要概率模型。

对于非概率模型，我们遵循*经验风险*经验风险的原则。
经验风险*最小化*，我们在章节8.2.中描述了经验风险最小化直接提供了一个优化问题来寻找好的参数。 最小化
帐目。对于统计模型，*最大似然*原则被用于最大似然。
以找到一套好的参数（第8.3).我们还可以使用概率模型对参数的不确定性进行建模，我们将在第二节中更详细地讨论这个问题。8.4.

我们使用数值方法来寻找 "适合 "数据的好参数，大多数训练方法可以被认为是寻找目标的最大值的爬坡方法，例如，相似度的最大值。
罩。为了应用爬坡方法，我们使用了在《公约》中描述的梯度。

第5章，并实施第7章的数值优化方法。

如第一章所述，我们感兴趣的是根据数据学习一个模型，使其在未来的数据上表现良好。这对我们来说是不够的

当模型遇到数据时

最小化目标。因此，在机器学习的目标中经常会有一个额外的减号。

交叉验证

为了使模型能够很好地适应训练数据，预测器需要在未见过的数据上很好地表现出来。我们使用*交叉验证法*模拟我们的预测器在未来的未见数据上的行为（第8.2.4).正如我们在本章中所看到的，为了实现在未见过的数据上表现良好的目标，我们需要在训练数据上的拟合和寻找现象的 "简单 "解释之间取得平衡。这种权衡是通过正则化（第8.2.3节）或增加先验（第8.3.2节）来实现的。在哲学中，这被认为既不是归纳也不是演绎，而是被称为*归纳*。根据《*斯坦福哲学百科全书*》，归纳法是推理到最佳解释的过程（Douven, 2017）。

诱拐

一个好的电影标题是 "人工智能诱导"。

我们经常需要对预测器的结构做出高层次的建模决定，例如使用的组件数量或考虑的概率分布类别。对组件数量的选择元素是一个*超参数*的例子，这个选择可以显著影响模型的性能。在不同的模型中进行选择的问题被称为*模型选择*，我们在第1节中描述了这个问题。8.6.对于非概率模型，模型选择通常是通过*嵌套交叉验证*来完成的，这一点我们将在第8.6.1.1节中描述。

超参数模型选择

嵌套

交叉验证

使用模型选择来选择我们模型的超参数。

*备注。*参数和超参数之间的区别在某种程度上是任意的，主要是由可以进行数值优化的内容和需要使用搜索技术的内容之间的区别驱动的。另一种考虑区别的方式是将参数视为概率模型的明确参数，而将超参数视为eters（更高级别的参数）作为控制这些明确参数

分布的参数。 ◆

在接下来的章节中，我们将研究机器学习的三种类型：经验风险最小化（Section8.2）、最大似然原则（第8.3）和概率建模（第8.4节）。

8.2 经验性的风险最小化

在掌握了所有的数学知识后，我们现在可以介绍一下学习的含义。机器学习的 "学习 "部分可以归结为根据训练数据来估计参数。

在这一节中，我们考虑预测器是一个函数的情况，并在第12节中考虑概率模型的情况。 8.3.我们描述了经验风险最小化的想法，它最初是由支持向量机的提议推广的（在第12章中描述）。然而，它的一般原则是广泛适用的，并允许我们在不明确构建概率模型的情况下提出什么是学习

的问题。有四个主要的设计选择，我们将在下面的小节中详细介绍。

节8.2.1我们允许预测器采取的函数集合是什么？

科目8.2.2我们如何衡量预测器在训练数据上的表现如何？

科目8.2.3我们如何从仅有的训练数据中构建在未见过的测试数据上表现良好的预测器？

科目8.2.4在模子的空间上进行搜索的程序是什么？

8.2.1 假设的函数类

假设我们给出了N个例子$\boldsymbol{x}_n \in \mathbb{R}^D$和相应的标量la- bels $y_n \in \mathbb{R}$。，$(\boldsymbol{x}_N, y_N)$。考虑到这些数据，我们想估计一个预测器$f(\,,\boldsymbol{\theta}) : \mathbb{R}^D \to \mathbb{R}$，用$\boldsymbol{\theta}$*作为参数*。

$$f(\boldsymbol{x}_n, \boldsymbol{\theta}^*) \approx y_n \quad \text{，对于所有} \quad n=1, \ldots, N. \tag{8.3}$$

在本节中，我们使用符号$\hat{y}_n = f(\boldsymbol{x}_n, \boldsymbol{\theta}^*)$来表示预测器的输出。

*备注。*为了便于表述，我们将以监督学习（我们有标签）的方式描述经验风险最小化。这简化了假设类和损失函数的定义。它在机器学习中，选择一个参数化的函数类也是很常见的，例如仿生函数。 ◆

例子 8.1

我们介绍了普通最小二乘法的回归问题，以说明经验性风险最小化。第九章将对回归进行更全面的阐述。当标签y_n是实值时，预测器的一个流行的函数类别选择是仿生函数集。我们选择一个

Affine functions are often referred to as linear functions in 机器学习。

通过串联一个加号1为一个仿生函数提供更紧凑的记号。该预测器接收代表单个例子的特征向量$\boldsymbol{x}_n$作为输入，并产生一个实值的输出，即$f : \mathbb{R}^{D+1} \to \mathbb{R}$。

领土单位特征$\boldsymbol{x}_n = [1, x_n^{(1)}, x_n^{(2)}, \ldots, x_n^{(D)}]^\mathrm{T}$。即$\boldsymbol{x}_n = [1, x_n^{(1)}, x_n^{(2)}, \ldots, x_n^{(D)}]^\mathrm{T}$，该参数向量相应地是$\boldsymbol{\theta} = [\theta_0, \theta_1, \theta_2, \ldots, \theta_D]^\mathrm{T}$，允许我们把预测器写成一个线性函数

$$f(\boldsymbol{x}_n, \boldsymbol{\theta}) = \boldsymbol{\theta}^\mathrm{T} \boldsymbol{x}_n . \tag{8.4}$$

这个线性预测器等同于仿生模型

$$f(\boldsymbol{x}_n, \boldsymbol{\theta}) = \theta_0 + \sum_{d=1}^{D} \theta_d x_n^{(d)} \tag{8.5}$$

本章以前的数字有一条直线作为预测器，这意味着我们假设了一个仿射函数。

我们可能希望将非线性函数作为预测器，而不是线性函数。神经网络的最新进展允许对更复杂的非线性函数类进行有效计算。

考虑到函数的类别，我们要寻找一个好的预测器。现在我们开始讨论经验风险最小化的第二个要素：如何衡量预测器对训练数据的拟合程度。

8.2.2 训练的损失函数

考虑到某一特定例子的标签y_n；以及我们根据x做出的n相应的预测$\hat{y}_n$。为了定义适合数据的含义，我们需要指定一个*损失函数*（y_n，$\hat{y}_n$），该*函数*将地面真实标签和预测作为输入，并产生一个非负数（称为损失），代表我们在这个特定预测上的错误程度。我们寻找一个好的参数向量$\boldsymbol{\theta}^*$的目标是使N个训练例子的平均损失最小。

损失函数

"错误 "这一表达方式经常被用来表示损失。

在机器学习中通常有一个假设，即实例集（$\boldsymbol{x}_1$，y_1），...，($\boldsymbol{x}_N$, y_N)是*独立和相同的分布*。独立这个词（第6.4.5节）意味着两个数据点（$\boldsymbol{x}_i$，y_i）和（$\boldsymbol{x}_j$，y_j）在统计上不相互依赖，意味着经验均值是对群体均值的良好估计（第6.4.1节）。这意味着我们可以在训练数据上使用损失的经验平均值。对于一个给定的*训练集*（$\boldsymbol{x}_1$，y_1），...，($\boldsymbol{x}_N$, y_N)，我们引入实例矩阵$\boldsymbol{X} := [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top \in \mathbb{R}^{N\times D}$和标签向量$\boldsymbol{y} := [y_1, \ldots, y_N]^\top \in \{\}^N$的符号。使用这个矩阵符号，平均损失为

独立和相同的分布

训练集

$$\text{雷姆普}(f, \boldsymbol{X}, \boldsymbol{y}) = \frac{1}{N}\sum_{n=1}^{N} (y_n, \hat{y}_n) 。 \tag{8.6}$$

其中$\hat{y}_n = f(\boldsymbol{x}_n, \boldsymbol{\theta})$。 方程(8.6)被称为*经验风险*，取决于三个参数，即预测器f和数据$\boldsymbol{X}$、$\boldsymbol{y}$。这种学习的一般策略被称为*经验风险最小化*。

经验风险

经验性风险最小化

示例（8.2最小二乘法损失）。

继续以最小二乘回归为例，我们明确指出，我们用平方损失（y_n，$\hat{y}_n$）$= (y_n - \hat{y}_n)^2$来衡量训练期间的错误成本。 我们希望最小化经验风险(8.6)，

这是在数据上损失的平均数

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^D} \frac{1}{N}\sum_{n=1}^{N}(y_n - f(\boldsymbol{x}_n,\boldsymbol{\theta}))^2 \tag{8.7}$$

其中我们替换了预测器$\hat{y}_n = f(\boldsymbol{x}_n, \boldsymbol{\theta})$。通过使用我们选择的线性预测器$f(\boldsymbol{x}_n, \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \boldsymbol{x}_n$，我们得到优化问题

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^D} \frac{1}{N}\sum_{n=1}^{N}(y_n - \boldsymbol{\theta}^\top \boldsymbol{x}_n)^2 \tag{8.8}$$

这个方程可以用矩阵形式等价表示

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^D} \frac{1}{N}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2 \tag{8.9}$$

这就是所谓的*最小二乘法问题*。通过求解正常方程，存在一个闭合形式的解，我们将在第二节讨论这个问题。9.2.

least-squares problem

我们对一个只在训练数据上表现良好的预测器不感兴趣。相反，我们寻求的是一个在未见过的测试数据上表现良好（风险低）的预测器。更正式地说，我们感兴趣的是找到一个
预测器f（参数固定），使预期风险最小。

$$R_{\text{true}}(f) = \mathbb{E}_{\boldsymbol{x},y}[\ell(y, f(\boldsymbol{x}))], \tag{8.10}$$

其中y是标签，$f(\boldsymbol{x})$是基于例子$\boldsymbol{x}$的预测。
符号$R_{\text{true}}(f)$表示，这是真正的风险，如果我们能获得
无限的数据量。期望值是在（无限的）所有
上。
可能的数据和标签。有两个实际问题是由我们希望最小化预期风险而产生的，我们在下面两个小节中讨论。

Another phrase的集合

通常用于预期风险的是 "人口风险"。

- 我们应该如何改变我们的训练程序以实现良好的泛化？我们如何
- 从（有限的）数据中估计预期风险？

*备注。*许多机器学习任务都规定了相关的性能指标，例如预测的准确性或均方根误差。性能度量可以更复杂，对成本敏感，并捕捉特定应用的细节。原则上，经验风险最小化的损失函数的符号应该直接对应于机器学习任务所指定的性能指标。在实践中，损失函数的设计与机器学习任

务所指定的性能指标往往是不匹配的。 *当模型遇到数据时*

职能和性能测量。这可能是由于易于实施或优化效率等问题造成的。 ◆

8.2.3 减少过拟合的正则化

本节描述了对经验风险最小化的补充，使其能够很好地泛化（近似于最小化预期风险）。我们认为训练机器学习预测器的目的是为了让我们在未见过的数据上表现良好，也就是说，预测器的泛化能力很强。我们通过保留整个数据集的一部分来模拟这种未见过的数据。这个保留的数据集被称为*测试集*。鉴于预测器的函数种类足够丰富，我们基本上可以记忆训练数据，以获得零经验风险。虽然这对于最小化训练数据上的损失（也就是风险）是很好的，但我们并不期望预测器能很好地推广到未见过的数据。在实践中，我们只有一个有限的数据集，因此我们把数据分成训练集和测试集。训练集用于拟合模型，而测试集（机器学习算法在训练期间没有看到）则用于评估泛化性能。重要的是，用户在观察完测试集后，不要再循环到新一轮的训练。我们用下标和$_{\text{train}}$来$_{\text{test}}$分别表示训练集和测试集。我们将在本节中重新审视这个使用有限数据集来评估预期风险的想法。8.2.4.

测试组

即使只知道预测器在测试集上的表现，也会泄露信息（Blum and Hardt,2015）。

事实证明，经验风险最小化会导致*过度拟合*，也就是说，预测器与训练数据的拟合过于紧密，不能很好地概括新数据（Mitchell,1997）。这种在训练集上有非常小的平均损失，但在测试集上有很大的平均损失的普遍现象往往发生在我们有很少的数据和一个复杂的混合类。对于一个特定的预测器f（参数固定），当来自训练集的风险估计在测试集上出现时，就会出现过拟合现象。

过度拟合

训练数据$\mathbf{R}_{\text{emp}}(f, \boldsymbol{X}_{\text{train}}, y_{\text{train}})$低估了预期风险$\mathbf{R}_{\text{true}}(f)$。因为我们通过使用测试集$\mathrm{R}_{\text{emp}}(f, \boldsymbol{X}_{\text{test}}, y_{\text{test}})$上的经验风险来估计预期风险$\mathbf{R}_{\text{true}}(f)$，如果测试风险比训练风险大得多，这就是过拟合的表现。我们重新审视一下以下的想法

第1节中的过度拟合8.3.3.

因此，我们需要通过引入一个惩罚项，使优化器更难返回一个过于灵活的预测器，从而在某种程度上偏向于寻找经验风险的最小化。在机器学习中，惩罚项被称为*正则化*。正则化是在经验风险最小化的精确解决方案和解决方案的大小或复杂性之间进行妥协的一种方式。

正规化

例子（正则8.3化最小二乘法）。

正则化是一种不鼓励对优化问题进行复杂或极端解决的方法。最简单的正则化策略是

来代替最小二乘法问题

$$\min_{\boldsymbol{\theta}} \frac{1}{N}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|^2 \tag{8.11}$$

在前面的例子中，通过增加一个只涉及$\boldsymbol{\theta}$的惩罚项，使问题"正规化"。

$$\min_{\boldsymbol{\theta}} \frac{1}{N}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|^2+\lambda\|\boldsymbol{\theta}\|^2 \tag{8.12}$$

附加项$\|\boldsymbol{\theta}\|^2$被称为*正则器*，而参数$\lambda$是*正则化参数*。正则化参数交易

正则器

正则化参数

如果我们遇到过度拟合，参数值的大小往往会变得相对较大（Bishop，2006）。

正则化项有时被称为*惩罚项*，其中双使得向量$\boldsymbol{\theta}$更接近于原点。正则化的思想也作为参数的先验概率出现在概率模型中。回顾第6.6节，为了使后验分布与先验分布具有相同的形式，先验和似然必须是一致的。我们将在第12章中重温这一思想，8.3.2.我们将看到正则器的思想等同于大边际的思想。

惩罚项

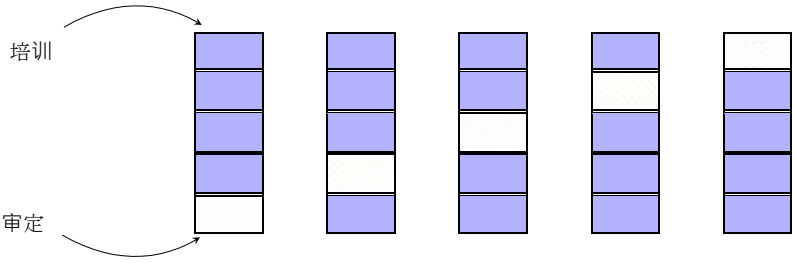
8.2.4 交叉验证来评估泛化性能

我们在上一节提到，我们通过在测试数据上应用预测器来估计泛化误差。这个数据是有时也被称为*验证集*。验证集是一个子我们把可用的训练数据的集合放在一边。这种方法的一个实际问题是，数据量是有限的，理想情况下，我们会使用尽可能多的可用数据来训练模型。这就要求我们保持小的验证集，这将导致预测性能的嘈杂估计（具有高方差）。解决这些相互矛盾的目标（大的训练集，大的验证集）的一个办法是集）是使用*交叉验证法*。K-fold交叉验证法有效地分割了将数据分成K个块，其中$K-1$个*块*构成训练集，最后一个块作为验证集（类似于前面概述的想法）。交叉验证通过（理想情况下）所有的块的分配组合来迭代；见图8.4.对验证集的所有K个选择重复这一程序，并对K个运行中的模型性能进行平均。

验证集

交叉验证法

我们将数据集划分为两组$\mathsf{D}=\mathsf{R}\cup\mathsf{V}$，使其不重叠（$\mathsf{R}\cap\mathsf{V}=\emptyset$），其中$\mathsf{V}$是验证集，并在$\mathsf{R}$上训练我们的模型。



图：*K*8.4-折交叉验证。数据集被分为*K*=5块，其中*K* 1*块* - 作为训练集（蓝色），1块作为验证集（橙色填充）。

验证集（例如V 通过计算验证集上训练模型的均方根误差（RMSE））。更确切地说，对于每个分区*k*，训练数据产生(*k*)一个预测器*f*(*k*)，然后将其应用于验证集以(*k*)计算经验风险*R*(*f* (*k*),(*k*))。我们通过验证集和训练集的所有可能分区进行循环，并计算出预测器的平均概括误差。交叉验证法近似于预期的泛化误差

$$\mathsf{V}\quad \mathsf{E}_{\mathsf{V}}[R(f,\mathsf{V})] \approx \frac{1}{K}\sum_{k=1}^{K} R(f^{(k)},\mathsf{V}^{(k)}), \tag{8.13}$$

其中*R*(*f* (*k*),(*k*))是(*k*)预测器在验证集上的风险（例如RMSE）(*k*)。近似有两个来源：第一，由于有限的训练集，导致不是最好的*f*(*k*)；第二，由于有限的验证集，导致对风险*R*(*f*(*k*) ,)的估计不准确(*k*)。*K*-折交叉验证的一个潜在缺点是训练模型*K*次的计算成本，如果训练成本很高的话，会造成很大的负担。在实践中，仅看直接参数往往是不够的。例如，我们需要探索多个复杂度参数（如多个重组参数），这些参数可能不是模型的直接参数。根据这些超参数来评估模型的质量，可能会导致训练运行的数量与模型参数的数量成指数关系。我们可以使用嵌套交叉验证法（第8.6.1节）来寻找好的超参数。

然而，交叉验证是一个*令人尴尬的平行*问题，即，Lit-

令人尴尬的是需要并行

的努力，将问题分成若干并行任务。如果有足够的计算资源（如云计算、服务器群），交叉验证不需要比单一的性能评估更长的时间。

在这一节中，我们看到经验风险最小化是基于以下概念：假设类函数、损失函数和正则化。在第8.3节中，我们将看到使用概率分布来代替损失函数和正则化概念的效果。

8.2.5 进一步阅读

由于经验性风险最小化的最初发展（Vapnik,1998）是用大量的理论语言表述的，因此许多后续发展都是理论性的。研究的领域被称为*统计学习理论*（Vapnik，1999；Evgeniou等人，2000；Hastie等人，2001年；Von Luxburg和Sch ölkopf,2011)。最近的一个机器建立在理论基础上并开发出高效学习算法的学习教科书是Shalev-Shwartz和Ben-David（2014）。

统计学习理论

正则化的概念起源于不理想问题的解决（Neumaier，1998）。这里提出的方法被称为*Tikhonov正则化*，还有一个密切相关的约束性版本称为Ivanov正则化。Tikhonov正则化与偏差-变异权衡和特征选择有很深的关系（Bühlmann和Van De Geer，2011）。交叉验证的另一种方法是自举法和杰克尼夫法（Efron和Tibshirani，1993；Davidson和Hinkley，1997；Hall，1992）。

Tikhonov正规化

将经验风险最小化（第8.2节）视为 "无概率 "是不正确的。有一个潜在的未知的概率分布$p(\boldsymbol{x}, y)$来控制数据的生成。然而，经验风险最小化的方法对分布的选择是不可知的。这与明确要求知道$p(\boldsymbol{x}, y)$的标准统计方法形成对比。此外，由于分布是一个在例子$\boldsymbol{x}$和标签y上的联合分布，标签可以是非确定性的。与标准统计学相比，我们不需要指定标签y的噪声分布。

8.3 参数估计

在本节中，8.2,我们没有明确地使用概率分布对我们的问题进行建模。在这一节中，我们将看到如何使用概率分布来模拟观察过程中的不确定性和预测器参数的不确定性。在第8.3.1,我们将介绍似然，它类似于经验风险最小化中的损失函数的概念（第8.2.2节）。先验的概念（第8.3.2节）类似于正则化的概念（第8.2.3节）。

8.3.1 最大似然估计

最大似然估计（MLE）背后的想法是定义一个最大似然函数。

准确计算参数，使我们能够找到一个适合数据的模型。

好。估计问题主要集中在*似然*函数上，即 likelihood

更确切地说，是它的负对数。对于由随机变量$\boldsymbol{x}$表示的数据和由概率密度$p(\boldsymbol{x} \mid \boldsymbol{\theta})$参数化的系列来说$\boldsymbol{\theta}$，*负的对数可能性*是由 负的对数可能性

$$L(\boldsymbol{\theta}) = -\log p(\boldsymbol{x} \mid \boldsymbol{\theta}) . \tag{8.14}$$

符号（$\theta_{\boldsymbol{x}}$）强调了参数$\boldsymbol{\theta}$是变化的，而数据$\boldsymbol{x}$是固定的。在写负对数似然时，我们经常放弃对$\boldsymbol{x}$的提及，因为它实际上是$\boldsymbol{\theta}$的函数，当代表数据中不确定性的随机变量从上下文中明确时，我们将它写成$L(\boldsymbol{\theta})$。

让我们来解释一下概率密度$p(\boldsymbol{x} \mid \boldsymbol{\theta})$对于$\boldsymbol{\theta}$的固定值是什么建模。它是一个分布，对给定参数设置的数据的不确定性进行建模。对于一个给定的数据集$\boldsymbol{x}$，可能性允许我们表达对参数$\boldsymbol{\theta}$的不同设置的偏好，我们可以选择更 "可能 "产生数据的设置。

在一个补充的观点中，如果我们认为数据是固定的（因为它已经被观察到了），而我们改变参数$\boldsymbol{\theta}$，$L(\boldsymbol{\theta})$告诉我们什么？它告诉我们，对于观察结果$\boldsymbol{x}$来说，$\boldsymbol{\theta}$的特定设置有多大可能。基于这第二种观点，最大似然估计器为我们提供了一组数据的最可能参数$\boldsymbol{\theta}$。

我们考虑有监督的学习环境，在那里我们得到了对$(\boldsymbol{x}_1, y_1), \ldots$我们感兴趣的是构建一个预测器，将特征向量$\boldsymbol{x}_n$作为输入并产生预测$y_n$（或与之相近的东西），也就是说，给定一个向量$\boldsymbol{x}_n$，我们希望得到标签$y_n$的概率分布。换句话说，我们指定标签的条件概率分布，在特定的参数设置$\boldsymbol{\theta}$下，给定例子。

例子 8.4

第一个经常使用的例子是规定给定例子的标签的条件概率是一个高斯分布。换句话说，我们假设可以用均值为零的独立高斯噪声（参考第6.5节）来解释我们的观察不确定性。$\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$。我们进一步假设，线性模型$\boldsymbol{x}_n^\top \boldsymbol{\theta}$被用于预测。这意味着我们为每个例子指定一个高斯似然值标签对$(\boldsymbol{x}_n, y_n)$。

$$p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}) = \mathcal{N}(y_n \mid \boldsymbol{x}_n^\top \boldsymbol{\theta}, \sigma^2) . \tag{8.15}$$

给定参数$\boldsymbol{\theta}$的高斯似然的图示如下图中8.3.我们将在第9.2节中看到如何用高斯分布来明确地扩展前面的表达。

独立和相同的分布

我们假设一组例子$(x_1, y_1), \ldots, (x_N, y_N)$是*独立和相同的分布*（i.i.d.）。独立 "一词（第6.4.5节）意味着涉及整个数据集（$\mathcal{Y} = \{y_1, \ldots, y_N\}$和$\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$）的似然性被分解为以下似然性的乘积

各个例子

$$p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\theta}) = \prod_{n=1}^{N} p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}), \tag{8.16}$$

其中$p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})$是一个特定的分布(在Ex- ample中是高斯分布)8.4)."相同分布 "这一表述意味着，乘积中的每项(8.16)的分布是相同的，而且所有这些项都有相同的参数。从优化的角度来看，计算可以分解为更简单的函数之和的函数往往更容易。
因此，在机器学习中，我们经常考虑负的对数可能性 Recall $\log(ab) = \log(a) + \log(b)$

$$\mathrm{L}(\boldsymbol{\theta}) = -\log p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\theta}) = -\sum_{n=1}^{N} \log p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}) \text{。} \tag{8.17}$$

虽然解释$\boldsymbol{\theta}$在$p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})$中的条件的右边这一事实是很有诱惑力的 (8.15)，因此应该被解释为观察到的和固定的，但这种解释是不正确的。负对数可能性（θ）是$\boldsymbol{\theta}$的函数。因此，要找到一个好的参数向量$\boldsymbol{\theta}$来解释数据（$\boldsymbol{x}_1, y_1$），....,（$\boldsymbol{x}_N, y_N$），要使关于$\boldsymbol{\theta}$的负对数似然$\mathrm{L}(\boldsymbol{\theta})$最小。

备注。(8.17)中的负号是一个历史遗留问题，是由于我们希望最大化似然的惯例，但数值优化文献倾向于研究函数的最小化。 ◆

例子 8.5

继续我们的高斯似然的例子(8.15)，负对数似然可以改写为

$$\mathrm{L}(\boldsymbol{\theta}) = -\sum_{n=1}^{N} \log p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}) = -\sum_{n=1}^{N} \log \mathrm{N}\left(y_n \mid \boldsymbol{x}_n^{\mathrm{T}} \boldsymbol{\theta}, \sigma^2\right) \tag{8.18a}$$

$$= -\sum_{n=1}^{N} \log \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{(y_n - \boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{\theta})^2}{2\sigma^2}\right) \tag{8.18b}$$

$$= -\sum_{n=1}^{N} \log \exp\left(-\frac{(y_n - \boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{\theta})^2}{2\sigma^2}\right) - \sum_{n=1}^{N} \log \sqrt{\frac{1}{2\pi\sigma^2}} \tag{8.18c}$$

$$= \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{\theta})^2 - \sum_{n=1}^{N} \log \sqrt{\frac{1}{2\pi\sigma^2}}. \tag{8.18d}$$

由于σ是给定的，（8.18d）中的第二项是常数，并且最小化$\mathrm{L}(\boldsymbol{\theta})$。
对应于解决最小二乘法问题（与(8.8)表示的第一项。

事实证明，对于高斯似然，所产生的优化

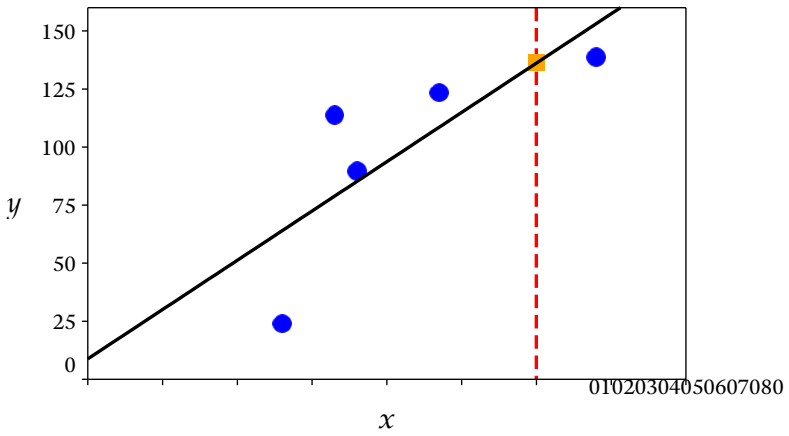
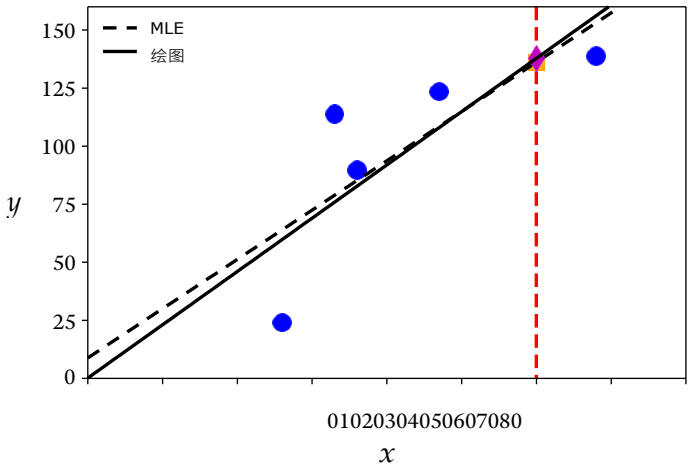


图 对于**8.5**给定的数据，参数的最大似然估计结果为黑色的对角线。橙色方块显示了最大似然的值
预测在
$x=60$.



图：**8.6**比较预测与最大似然估计和MAP估计在
$x=60$ .....先验偏向于斜率为
在这个例子中，使截距更接近于零的偏差实际上是增加了斜率。在这个例子中，使截距接近于零的偏差实际上增加了斜率。

与最大似然估计相对应的问题有一个封闭形式的解决方案。我们将在第九章看到更多的细节。图8.5显示了一个回归数据集和由最大似然参数引起的函数。最大似然估计可能会出现过拟合（Section8.3.3），类似于无规则化经验风险最小化(第8.2.3节)。对于其他似然函数，即如果我们用非高斯分布来模拟我们的噪声，最大似然估计可能没有一个封闭式的分析解。在这种情况下，我们要采用第七章中讨论的数值优化方法。

8.3.2 最大A后验估计

如果我们有关于参数$\boldsymbol{\theta}$分布的先验知识，我们可以在似然中乘以一个附加项。这个附加项是关于参数$p(\boldsymbol{\theta})$的一个先验概率分布。对于一个给定的

先验，在

观察到一些数据$\boldsymbol{x}$，我们应该如何更新$\boldsymbol{\theta}$的分布？换句话说，我们应该如何表示我们在观察了数据$\boldsymbol{x}$之后对$\boldsymbol{\theta}$有了更具体的了解这一事实？正如第6.3节所讨论的，贝叶斯定理为我们提供了一个原则性的工具来更新我们的概率分布。

准确地说是随机变量的分布。它允许我们计算出一个*后验*分布 posterior
$p(\boldsymbol{\theta}|\boldsymbol{x})$（更具体的知识）对参数$\boldsymbol{\theta}$从一般
*先验*声明（先验分布）$p(\boldsymbol{\theta})$和函数$p(\boldsymbol{x}|\boldsymbol{\theta})$，即 先验的
将参数$\boldsymbol{\theta}$和观察到的数据$\boldsymbol{x}$联系起来（称为*似然*）： 似然

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{x})}。 \tag{8.19}$$

回顾一下，我们感兴趣的是找到使后验最大化的参数$\boldsymbol{\theta}$。由于分布$p(\boldsymbol{x})$不依赖于$\boldsymbol{\theta}$，我们可以忽略优化的分母的值，得到

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})。 \tag{8.20}$$

前面的比例关系隐藏了数据的密度$p(\boldsymbol{x})$，这可能很难估计。我们现在不是估计负对数可能性的最小值，而是估计负对数可能性的最小值。

这就是所谓的*最大后验数*--最大后验数。

判断（*MAP估计*）。图中显示了添加零均值高斯先验的效果。8.6. 后验估计
MAP估计

例子 8.6

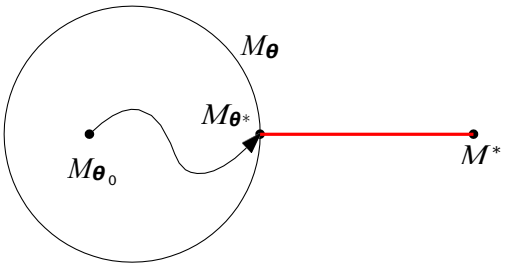
除了前面的高斯似然假设外，我们还假设参数向量分布为多变量的

平均值为零的高斯，即$p(\boldsymbol{\theta})=\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$，其中$\boldsymbol{\Sigma}$是协方差---。
的矩阵（第6.5节）。请注意，高斯的共轭先验
也是高斯的（第6.6.1节），因此我们期望后验分布也是高斯的。我们将在第九章看到最大后验估计的细节。

在机器学习中，包含关于好的参数所在的先验知识的想法是很普遍的。另一种观点是正则化，它引入了一个附加项，使得到的参数偏离原点。8.2.3,是正则化的想法，它引入了一个附加项，使得到的参数偏向于接近原点。最大后验估计可以被认为是连接非概率和概率世界的桥梁，因为它明确承认需要一个先验分布，但它仍然只产生一个参数的点估计。

*备注。*最大似然估计$\boldsymbol{\theta}_{\mathrm{ML}}$具有以下特性（Lehmann和Casella，1998；Efron和Hastie，2016）。

- 渐进的一致性。MLE收敛于真实值，在



图为模型**8.7**拟合。在一个参数化的类中模型的M，我们优化模型参数以最小化与真实（未知）模型M的距离$\boldsymbol{\theta}^*$。

无限多的观察结果的极限，加上一个近似于正常的随机误差。

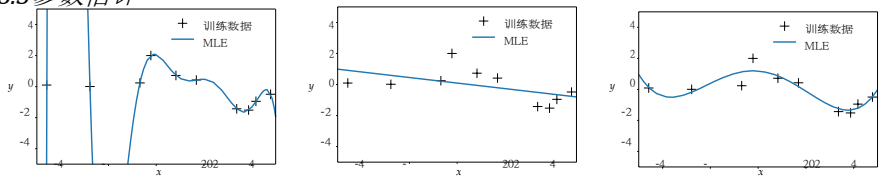
- 实现这些特性所需的样本大小可能相当大。
- 误差的方差以$1/N$衰减，其中N是数据点的数量。
- 特别是，在 "小 "数据体系中，最大似然估计会导致*过度拟合*。

最大似然估计（和最大先验估计）的原理是利用概率模型来推理数据和模型参数的不确定性。然而，我们还没有把概率建模发挥到极致。在本节中，所产生的训练过程仍然产生预测器的点估计，即训练返回一组代表最佳预测器的参数值。在本节中，8.4,我们将认为参数值也应被视为随机变量，而且我们将在以下情况下使用完整的参数分布，而不是估计该分布的 "最佳 "值。做出预测。

8.3.3 模型拟合

考虑一下这样的情况：我们给了一个数据集，我们对将一个参数化的模型拟合到数据上感兴趣。当我们谈论 "拟合 "时，我们通常是指优化/学习模型参数，使其最小化一些损失函数，例如，负对数可能性。关于最大似然（第8.3.1节）和最大后验估计（第8.3.2），我们已经讨论了两种常用的模型拟合算法。

模型的参数化定义了一个模型类$M_{\boldsymbol{\theta}}$，用它可以我们可以操作。例如，在线性回归环境中，我们可以将输入x和（无噪声）观测值y之间的关系定义为$y = ax + b$，其中$\boldsymbol{\theta} := a, b$是模型参数。在这种情况下，模型参数$\boldsymbol{\theta}_{\{\}}$描述了仿生函数系列，即斜率为$a$的直线，其偏移量为0$b$。 假设数据来自



(a) 过度拟合 (b) 适应性不足。 (c) 适应性强。

图为不同模型类别与回归数据集的**8.8**拟合（通过最大似然法）。

从一个我们不知道的模型M^*开始。对于一个给定的训练数据集，我们优化$\boldsymbol{\theta}$，使$M_{\boldsymbol{\theta}}$尽可能地接近M^*，其中 "接近性 "由我们优化的目标函数定义（例如，训练数据上的损失平方）。图8.7说明了这样一个情况：我们有一个小的模型类别（用圆圈M表示$_{\boldsymbol{\theta}}$），而数据生成模型M^*位于所考虑的模型集之外。我们在$M_{\boldsymbol{\theta}}$处开始我们的参数搜索。在优化之后，即当我们获得最佳的可行参数$\boldsymbol{\theta}^*$时，我们将区分三种不同的情况。(i) 过度拟合。

(ii) 拟合不足，和(iii) 拟合良好。我们将对这三个概念的含义给出一个高层次的直觉。

粗略地说，*过拟合*是指准元化的模型类太丰富了，无法对M^*所生成的数据集进行建模，也就是说，$M_{\boldsymbol{\theta}}$可以对更复杂的数据集进行建模。例如，如果数据集是由一个线性函数生成的，而我们将$M_{\boldsymbol{\theta}}$定义为七阶多项式类，那么我们不仅可以为线性函数建模，还可以为二阶、三阶等多项式建模。过度的模型拟合通常有大量的参数。一个我们经常观察到的一个方法是检测使得过于灵活的模型类$M_{\boldsymbol{\theta}}$使用其所有的建模能力来减少训练误差。如果训练数据是有噪声的，那么它就会在噪声本身中找到一些有用的信号。当我们远离训练数据进行预测时，这将引起巨大的问题。图8.8(a)给出了一个回归中过拟合的例子，其中模型参数是通过最大似然法学习的。8.3.1).我们将在第8.8节中更多地讨论回归中的过拟合问题。9.2.2.

过拟合的情况。

在实践中，观察到模型的训练风险较低，但在交叉验证过程中测试风险较高，这就是过拟合。8.2.4).

当我们遇到*欠拟合时*，我们会遇到相反的问题其中模型类$M_{\boldsymbol{\theta}}$不够丰富。例如，如果我们的数据集是由一个正弦函数产生的，但是$\boldsymbol{\theta}$只对直线进行了参数化，那么最佳优化程序将不会让我们接近真实的模型。然而，我们仍然优化参数，并找到模拟数据集的最佳直线。图8.8(b)显示了一个因为不够灵活而不适合的模型的例子。拟合不足的模型一般都有很少的参数。

--欠拟合。

第三种情况是当参数化的模型类别基本正确时。那么，我们的模型拟合得很好，也就是说，它既不过度拟合，也不不足拟合。这意味着我们的模型类别足够丰富，可以描述我们所给的数据集。图8.8(c)显示了一个相当适合给定数据集的模型。理想情况下。

这是我们想要使用的模型类，因为它具有良好的泛化特性。

在实践中，我们经常定义非常丰富的模型类$M_{\boldsymbol{\theta}}$，其中有许多派别。诸如深度神经网络等。为了缓解过度拟合的问题，我们可以使用正则化（第8.2.3节）或先验指标（第8.3.2节）。我们将在第8.2节讨论如何选择模型类别。8.6.

8.3.4 进一步阅读

当考虑概率模型时，最大似然估计原则概括了线性模型的最小二乘回归思想，我们将在第九章详细讨论。当限制预测器具有线性形式，并有一个额外的非线性函数j应用于输出时，即。

$$p(y_n|\boldsymbol{x}_n, \boldsymbol{\theta}) = \phi(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}_n)\,, \tag{8.21}$$

我们可以考虑其他预测任务的模型，如二元分类或计数数据建模（McCullagh and Nelder,1989）。一个
另一种观点是，考虑来自前者的可能性。
的模型(第6.6节)。这类模型在参数和数据之间具有线性依赖性，并具有潜在的非线性变换j（称为*链接函数*），被称为*广义线性模型*（Agresti，2002，第4章）。

链接函数 广义线性模型

最大似然估计有着丰富的历史，最初是由Ronald Fisher爵士在20世纪30年代提出的。我们将在本节中对概率模型的概念进行扩展。8.4.在使用概率模型的研究人员中，有一个争论是贝叶斯统计学和自由统计学之间的讨论。正如第6.1.1节所提到的，这可以归结为概率的定义。回顾第6.1节，我们可以认为概率是对逻辑推理的概括（通过允许不确定性）（Cheeseman，1985；Jaynes，2003）。最大似然性估计的方法在本质上是频繁主义的，感兴趣的读者可以参考Efron和Hastie(2016)对贝叶斯和频繁主义统计的平衡看法。

在一些概率模型中，最大似然法可能是不可能的。读者可以参考更高级的统计学教科书，如Casella和Berger(2002)，了解一些方法，如矩量法、M-估计和估计方程。

8.4 概率建模和推理

在机器学习中，我们经常关注对数据的解释和分析，例如对未来事件的

生成过程

预测和决策。为了使这一任务更具有可操作性，我们经常建立模型来描述产生观察数据的*生成过程*。

例如，我们可以通过两个步骤来描述一个抛硬币实验的结果（"头 "或 "尾"）。首先，我们定义一个参数μ，作为伯努利分布的参数来描述 "正面 "的概率（第6章）；其次，我们可以从伯努利分布$p(x|\mu) = \text{Ber}(\mu)$中抽出一个结果xhead，tail。参数$\mu$产生了一个特定的 $\in \{$ 数据集，并取决于所使用的硬币。由于μ是事先不知道的，也不可能直接观察到，所以我们需要一些机制来了解关于μ的情况，给定抛硬币实验的观察结果 $\mathcal{X}$ 。在下文中，我们将讨论如何将概率模型用于这一目的。

8.4.1 概率模型

一个概率的模型是由所有随机变量的联合分布指定的。

概率模型将实验的不确定方面表示为概率分布。使用概率模型的好处是，它们为建模、推理、预测和模型选择提供了一套来自概率理论（第6章）的统一和一致的工具。

在概率建模中，观察变量$\boldsymbol{x}$和隐藏参数$\boldsymbol{\theta}$的联合分布$p(\boldsymbol{x}, \boldsymbol{\theta})$是最重要的。它使
囊括了来自以下方面的信息。

- 先验和似然（乘积规则，第6.3节）。
- 边际似然$p(\boldsymbol{x})$将在模型选择中发挥重要作用（第6节）。8.6)，可以通过取联合分布和整合参数来计算(总和规则，第6.3节)。
- 后验，可以通过将接头除以边际似然得到。

只有联合分布具有这种特性。因此，一个概率模型是由其所有随机变量的联合分布指定的。

参数估计可以被表述为一个优化问题。

8.4.2 贝叶斯推理

机器学习的一个关键任务是利用模型和数据来发现模型的隐藏变量$\boldsymbol{\theta}$的值，给定观察到的变量
$\boldsymbol{x}$. 在第8.3.1节中，我们已经讨论了使用最大似然法或最大后验法来估计模型参数$\boldsymbol{\theta}$的两种方法。在这两种情况下，我们都能得到一个最佳的$\boldsymbol{\theta}$值，因此参数估计的关键算法问题是解决一个优化问题。一旦这些点估计值$\boldsymbol{\theta}^*$是已知的，我们就用它们来进行预测。更具体地说，预测分布将是$p(\boldsymbol{x}\,|\,\boldsymbol{\theta}^*)$，我们在似然函数中使用$\boldsymbol{\theta}^*$。

正如第6.3节所讨论的那样，仅仅关注后向分布的某些统计量（如使后向分布最大化的参数$\boldsymbol{\theta}^*$）会导致信息的损失，这在一个系统中可能是至关重要的。

使用预测值$p(\boldsymbol{x}\mid\boldsymbol{\theta}^*)$来做决策。这些决策系统通常有不同于似然的目标函数，即平方误差损失或错误分类误差。因此，拥有完整的后验分布会非常有用，并导致更稳健的决策。*贝叶斯推理*就是要找到这个后验分布（Gelman等人，2004）。对于一个数据集$\mathcal{X}$、一个参数先验$p(\boldsymbol{\theta})$和一个似然函数，后验

贝叶斯推理是关于学习随机变量的分布。贝叶斯推理

$$p(\boldsymbol{\theta}\mid\mathcal{X})=\frac{p(\mathcal{X}\mid\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{X})},\qquad p(\mathcal{X})=\int p(\mathcal{X}\mid\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta},\tag{8.22}$$

是通过应用贝叶斯定理得到的。关键的想法是利用贝叶斯定理来反转参数$\boldsymbol{\theta}$和数据之间的关系（由似然给出），以获得后验分布$p(\boldsymbol{\theta}\mid\mathcal{X})$。

贝叶斯推理倒置了参数和数据之间的关系。

$\mathcal{X}$有一个关于参数的后验分布，其含义是它可以被用来将不确定性从参数传播到数据。更具体地说，在参数的分布$p(\boldsymbol{\theta})$下，我们的预测将是

$$p(\boldsymbol{x})=\int p(\boldsymbol{x}\mid\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}=\mathbb{E}_{\boldsymbol{\theta}}[p(\boldsymbol{x}\mid\boldsymbol{\theta})],\tag{8.23}$$

而且它们不再依赖于模型参数$\boldsymbol{\theta}$，这些参数已经被边缘化/积分化了。方程式(8.23)显示，预测是所有可信的参数值$\boldsymbol{\theta}$的平均值，其中可信度由参数分布$p(\boldsymbol{\theta})$来概括。

在第1节中讨论了参数估计和贝叶斯法则后，在第2节中讨论了贝叶斯法则。8.3和Bayesian in-

在这里，让我们比较一下这两种学习方法。通过最大似然或MAP估计产生一个一致的参数点估计$\boldsymbol{\theta}^*$，要解决的关键计算问题是优化。与此相反，贝叶斯推理产生一个（正）分布，需要解决的关键计算问题是整合。用点估计进行预测是直接的，而贝叶斯框架中的预测需要解决另一个整合问题；见(8.23).然而，贝叶斯推理给我们提供了一种原则性的方法来纳入先验知识，说明侧面信息，并纳入结构知识，所有这些在参数估计的背景下都不容易做到。此外，在数据高效学习的背景下，参数不确定性对预测的传播在风险评估和探索的决策系统中是有价值的（Deisenroth等人，2015；Kamthe和Deisenroth，2018）。

虽然贝叶斯推理是一个学习参数和进行预测的数学原则性框架，但由于我们需要解决的整合问题，它也有一些实践上的挑战；见(8.22)和(8.23).更具体地说，如果我们不对参数选择共轭先验（第6.6.1节），那么（8.22）和（）中的积分就不能用分析法解决。8.23)中的积分是不可分析的，而且我们无法计算出正负值。

在封闭的形式下，我们需要对后验、预测或边际似然进行分析。在这些情况下，我们需要求助于近似方法。在这里，我们可以使用随机性近似，如马尔科夫链蒙特卡洛（MCMC）（Gilks等，1996），或者确定性近似，如拉普拉斯近似（Bishop，2006；Barber，2012；Murphy，2012），变分法（Jordan等，1999；Blei等，2017），或者期望传播法（Minka，2001a）。

尽管有这些挑战，贝叶斯推理已经成功地应用于各种问题，包括大规模的主题建模（Hoff-man等人，2013）、点击率预测（Graepel等人，2010）、控制系统中的数据高效强化学习（Deisenroth等人，2015）、在线排名系统（Herbrich等人，2007）和大规模的识别系统。有一些通用工具，如贝叶斯优化（Brochu等人，2009；Snoek等人，2012；Shahriari等人，2016），对于有效搜索模型或算法的元参数是非常有用的成分。

备注。在机器学习文献中，（随机）"变量 "和 "参数 "之间可能有某种程度的分离。当参数被估计时（例如，通过最大似然法），变量通常被边缘化掉。在本书中，我们对这种分离并不严格，因为原则上，我们可以在任何参数上放置一个先验，并将其整合出来，这将使参数变成一个随机变量。

能够根据上述的分离。 ◆

8.4.3 潜在变量模型

在实践中，有时需要有额外的*潜变量*zlatent variable
(除了模型参数$\boldsymbol{\theta}$之外）作为模型的一部分（Moustaki等人，2015）。这些潜变量与模型参数$\boldsymbol{\theta}$不同，因为它们没有明确地对模型进行参数化。潜变量可以描述数据产生的过程，从而有助于提高模型的可预测性。它们也经常简化模型的结构，使我们能够定义更简单和更丰富的模型结构。模型结构的简化往往与较少的模型参数相伴而行（Paquet，2008；Murphy，2012）。潜变量模型的学习（至少通过最大似然）可以通过期望最大化（EM）算法（Demp-ster等，1977；Bishop，2006）以一种原则性的方式完成。例如，在这种潜变量

有帮助的是用于降维的主成分分析（第10章）、用于密度估计的高斯混合模型（第11章）、用于时间序列建模的隐马尔科夫模型（Maybeck,

1979）或动力系统（Ghahramani和Roweis，1999；Ljung，1999）以及元学习和任务泛化（Hausman等人，2018；Sæ- mundsson等人，2018）。虽然这些潜变量的引入

可能会使模型结构和生成过程更容易，但在潜在变量模型中的学习通常是困难的，我们将在第11章中看到这一点。

由于潜变量模型也允许我们定义从参数生成数据的过程，让我们来看看这个生成过程。用$\boldsymbol{x}$表示数据，用$\boldsymbol{\theta}$表示模型参数，用$\boldsymbol{z}$表示潜变量，我们得到条件分布

$$p(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{\theta}) \tag{8.24}$$

这使得我们能够为任何模型参数和潜在变量生成数据。鉴于$\boldsymbol{z}$是潜伏变量，我们对其放置一个先验$p(\boldsymbol{z})$。

正如我们之前讨论的模型一样，带有潜变量的模型可以在我们在第8.3节和第8.4.2节讨论的框架内用于参数学习和推理。为了促进学习（例如，通过最大似然估计或贝叶斯推断），我们采用两步程序。首先，我们计算模型的似然$p(\boldsymbol{x} \mid \boldsymbol{\theta})$，它不依赖于潜变量。其次，我们使用这个似然来进行参数估计或贝叶斯推断，其中我们使用的表达方式与第8.4.2节中的完全相同。8.3和8.4.2节中的表达式完全相同。

由于似然函数$p(\boldsymbol{x} \mid \boldsymbol{\theta})$是给定模型参数的数据预测分布，我们需要将潜在变量边缘化，以便

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \int p(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{\theta}) p(\boldsymbol{z}) d\boldsymbol{z}, \tag{8.25}$$

其中$p(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{\theta})$在(8.24)，$p(\boldsymbol{z})$是潜变量的先验。请注意，似然不能依赖于潜变量$\boldsymbol{z}$，它只是数据$\boldsymbol{x}$和模型参数$\boldsymbol{\theta}$的一个函数。

似然是数据和模型参数的函数，但与潜变量无关。

中的似然值(8.25) 直接允许通过最大似然进行参数估计。对于模型参数$\boldsymbol{\theta}$的条件先验，MAP估计也是直接的，这一点在第二节中讨论过。8.3.2. 此外，利用似然(8.25)的贝叶斯推断(第8.4.2节)在一个潜在变量模型中以通常的方式工作。我们在模型参数上放置一个先验$p(\boldsymbol{\theta})$，并使用贝叶斯定理得到一个后验分布

$$p(\boldsymbol{\theta} \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{X})} \tag{8.26}$$

在给定的数据集上的 $\mathbf{X}$ 模型参数。(8.26)可用于贝叶斯推理框架下的预测；见(8.23).在这个潜在变量模型中，我们面临的一个挑战是，相似条件$p(\mathbf{X} \mid \boldsymbol{\theta})$需要对潜在变量ac-的边际化。

根据(8.25).除了当我们选择共轭先验$p(\boldsymbol{z})$为$p(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{\theta})$，()中的边际化是不可分析的，我们需要借助于近似值(Bishop,2006;Paquet,2008;Mur -8.25)中的边际化在分析上是不可行的，我们需要求助于近似的方法(Bishop,2006;Paquet,2008;Murphy,2012;Moustaki et al.,2015）。

类似于参数后验 (8.26)，我们可以根据以下公式计算出潜变量的后验

$$p(\boldsymbol{z} \mid \mathrm{X}) = \frac{p(\mathrm{X} \mid \boldsymbol{z})p(\boldsymbol{z})}{p(\mathrm{X})}, \quad p(\mathrm{X} \mid \boldsymbol{z}) = \mathrm{p}(\mathrm{X} \mid \boldsymbol{z}, \boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}, \quad (8.27)$$

其中$p(\boldsymbol{z})$是潜变量的先验，$p(\mathrm{X} \mid \boldsymbol{z})$要求我们把模型参数$\boldsymbol{\theta}$整合出来。

考虑到分析求解积分的困难，显然，在一般情况下，同时确定潜变量和模型参数是不可能的（Bishop，2006；Murphy，2012）。一个比较容易计算的数量是潜变量的后验分布，但以模型参数为条件，即。

$$p(\boldsymbol{z} \mid \mathrm{X}, \boldsymbol{\theta}) = \frac{p(\mathrm{X} \mid \boldsymbol{z}, \boldsymbol{\theta})p(\boldsymbol{z})}{p(\mathrm{X} \mid \boldsymbol{\theta})}, \quad (8.28)$$

其中$p(\boldsymbol{z})$是潜在变量的先验，$p(\mathrm{X} \mid \boldsymbol{z}, \boldsymbol{\theta})$在(8.24).

在第10章和第11章中，我们分别推导了PCA和高斯混合模型的似然函数。此外，我们还计算了PCA和高斯混合模型的潜变量的后向分布(8.28)在PCA和高斯混合模型的潜变量上。

*备注。*在接下来的章节中，我们可能不会对潜伏变量$\boldsymbol{z}$和不确定的模型参数$\boldsymbol{\theta}$进行如此明确的区分，并将模型参数也称为 "潜伏 "或 "隐藏"，因为它们是未观察到的。在第10章和第11章中，当我们使用潜变量$\boldsymbol{z}$时，我们将注意其区别，因为我们将有两种不同类型的隐藏变量：模型参数$\boldsymbol{\theta}$和潜变量$\boldsymbol{z}$。 ◆

我们可以利用概率模型的所有元素都是随机变量这一事实，定义一种统一的语言来表示它们。在第8.5节中，我们将看到一种简明的图形语言来表示概率模型的结构。我们将在随后的章节中使用这种图形语言来描述概率模型。

8.4.4 进一步阅读

机器学习中的概率模型（Bishop，2006；Barber，2012；Murphy，2012）为用户提供了一种方法，以一种原则性的方式捕捉数据和预测模型的不确定性。Ghahramani（2015）对机器学习中的概率模型进行了简短的回顾。给定一个概率模型，我们可能足够幸运，能够通过分析计算出感兴趣的参数。然而，一般来说，分析解决方案是罕见的，计算方法

，如抽样（Gilks等人，1996；Brooks等人，2011）和变异推理（Jordan等人，1999；Blei等人）。

Moustaki等人（2015）和Paquet（2008）对潜变量模型中的贝叶斯推断进行了很好的概述。

近年来，人们提出了几种编程语言，旨在将软件中定义的变量视为与概率分布相对应的随机变量。其目的是能够编写复杂的概率分布函数，而在引擎盖下，编译器自动处理贝叶斯推理规则。这个快速变化的领域被称为*概率编程*。

概率编程

8.5 有向图形模型

在这一节中，我们介绍一种用于指定概率模型的图形语言，称为*有向图形模型*。它提供了一个紧凑的是指定概率模型的简洁方式，并允许读者直观地解析随机变量之间的依赖关系。图形模型直观地捕捉了所有随机变量的联合分布可以被分解为只取决于这些变量子集的因素的乘积的方式。在第8.4,我们将概率模型的联合分布确定为感兴趣的关键数量，因为它包括关于先验、似然和后验的信息。

有向图形化模型

模型

有向图形模型也被称为贝叶斯网络。

叶斯网络

图形模型

有向图形模型/贝

然而，联合分布本身可能相当复杂，它并没有告诉我们任何关于概率模型的结构属性。例如，联合分布$p(a, b, c)$并没有告诉我们任何关于独立关系的信息。这就是图形模型发挥作用的地方。本节依靠独立和条件独立的概念，如第6.4.5节所述。

在一个*图形模型*中，节点是随机变量。在图8.9(a)中，节点代表随机变量a、b、c，边代表变量之间的概率关系，例如，条件概率。

*备注。*并非每个分布都可以用特定的图形模型来表示。这方面的讨论可以在Bishop（2006）中找到。◆

概率图式模型有一些方便的特性。

- 它们是可视化概率模型结构的一种简单方法。它们可以用来设计或激励新型的统计模型。仅仅检查图就可以让我们了解到一些特性，例如条件独立性。
- 统计模型中推理和学习的复杂计算可以用图形操作来表达。

8.5.1 图形语义学

*有向图形模型/贝叶斯网络*是一种在概率模型中表示条件依赖关系的方法。它们提供了一种可视化的

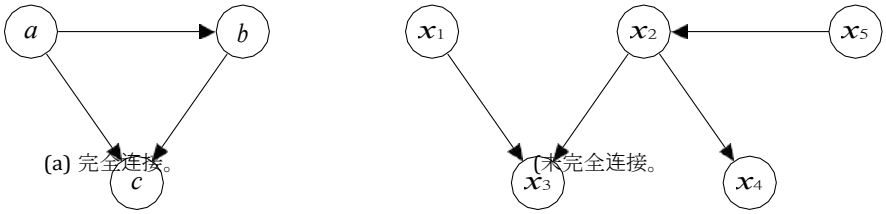
描写条件概率，因此，提供了一个简单的语言。

描述复杂的相互依存关系的准则。模块化的描述

也带来了计算上的简化。两个节点（随机变量）之间的定向链接（箭头）表示条件概率。例如，图8.9(a)中a和b之间的箭头给出了给定a的b的条件概率$p(b \mid a)$。

假设，箭头可以用来表示因果关系（Pearl，2009）。

图为有向图形模型的例子**8.9**。



(a) 完全连接。 (b) 未完全连接。

如果我们对联合分布的因子化有所了解，就可以从联合分布中推导出定向图形模型。

例子 8.7

考虑联合分布

$$p(a, b, c) = p(c \mid a, b)p(b \mid a)p(a) \tag{8.29}$$

三个随机变量a、b、c的联合分布的因式分解在(8.29)告诉我们一些关于随机变量之间的关系。

- c直接取决于a和b。b直接取
- 决于a。
- a既不取决于b也不取决于c。

对于(8.29)，我们得到图8.9(a)中的有向图形模型。

一般来说，我们可以从因子化的联合分布中构建相应的有向图形模型，如下所示。

1. 为所有随机变量创建一个节点。
2. 对于每个条件分布，我们从对应于分布条件的变量的节点向图中添加一个定向链接（箭头）。

图形布局取决于联合分布的因子化。

图形布局取决于联合分布的因式分解的选择。

我们讨论了如何从一个已知的联合分布的因式分解中得到相应的有向图式模型。现在，我们将做

恰恰相反，描述如何从一个给定的图形模型中提取一组随机变量的联合分布。

例子 8.8

看一下图8.9(b)中的图形模型，我们利用了两个适当的联系。

- 我们寻求的联合分布$p(x_1, \ldots, x_5)$是一组条件的乘积，图中每个节点都有一个条件。在这个特定的例子中，我们将需要五个条件式。
- 每个条件只取决于图中相应节点的父节点。例如，x_4将以x_2为条件。

这两个特性产生了所需的联合分布的因式分解

$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2)p(x_3 | x_1)p(x_4 | x_2) \text{。} \quad (8.30)$$

一般来说，联合分布$p(\boldsymbol{x})=p(x_1, \ldots, x_K)$是由以下公式给出的

$$p(\boldsymbol{x}) = \prod_{k=1}^{K} p(x_k \mid \mathrm{Pa}_k) , \quad (8.31)$$

其中Pa_k表示 "x_k的父节点"。x_k的父节点是有箭头指向x_k的节点。

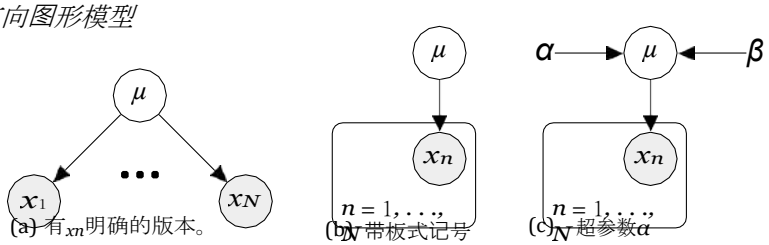
在本小节的最后，我们举一个抛硬币实验的具体例子。考虑一个伯努利实验（例6.8），这个实验的结果x是 "人头 "的概率是

$$p(x \mid \mu) = \mathrm{Ber}(\mu) \text{。} \quad (8.32)$$

我们现在重复这个实验N次，观察结果$x_1, \ldots, x_N$
因此，我们可以得到联合分布

$$p(x_1, \ldots, x_N | \mu) = \prod_{n=1}^{N} p(x_n | \mu) \text{。} \quad (8.33)$$

右手边的表达式是每个单独结果的伯努利分布的乘积，因为实验是独立的。回顾第6.4.5节，统计上的独立性意味着分布的因子化。为了写出这组图形模型，我们要区分未观察的/潜在的变量和观察变量。从图形上看，观察变量用阴影节点表示，这样我们就得到了图8.10（a）中的图形模型。我们看到单参数μ对所有的x_n都是一样的，$n=1, \ldots, N$，因为结果x_n是相同分布的。图8.10(b)给出了这种情况下的一个更紧凑但等效的图形模型，在这里我们用



(a) 有x_n明确的版本。 (b) 带板式记号的版本。 (c) 超参数α和β对潜在的μ。

图为重复伯努利实验的图形**8.10**模型。

*板块*记号。板块（盒子）重复里面的一切（在这种情况下，的观测值x_n)N次。因此，两种图形模型都是等效的，但板块符号更紧凑。图形模型可以立即

板块

*超*先验是对第一层先验参数的第二层布的超先验。图8.10(c)将Beta(α, β)先验置于潜变量μ上。α和β是确定性的参数，即不是随机变量，我们省略了围绕它的圆圈。

先验分

8.5.2 有条件的独立和d-分离

有向图形模型允许我们只通过寻找联合分布的条件独立性（第6.4.5节）关系属性。

在图中。一个叫做d-separation（Pearl，1988）的概念是这方面

的关键。

考虑一个一般的有向图，其中 、 、 是任意不相交的节点集（其联合可能小于图中的完整节点集）。我们希望确定一个特定的结条件独立性声明，"A有条件地独立于B给定C"，表示为

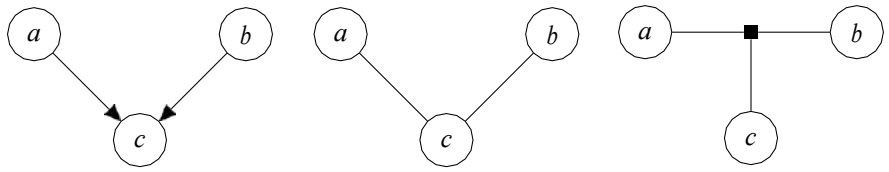
$$a \perp b \mid c\,, \tag{8.34}$$

是由一个给定的有向无环图所暗示的。为此，我们考虑所有可能的路径（忽略箭头方向的路径），从任何节点到任何节点。如果任何这样的路径包括任何节点，并且以下任何一种情况为真，则称其为阻断。

- 路径上的箭头在节点处头尾相接或尾尾相接，而节点在集合C中。
- 箭头在节点处首尾相接，而且该节点和它的任何子孙都不在集合C中。

如果所有的路径都被阻断了，那么A和B可以说是*d分离*

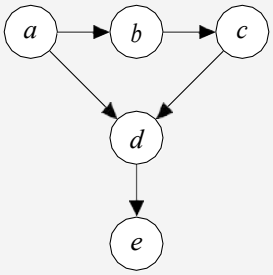
，图中所有变量的联合分布将满足 $A \perp B \mid C$。



(a) 定向图形模型 (b) 无定向图形模型 (c) 因素图

图三类**8.12**图形模型：(a)有向图形模型（贝叶斯网络）；(b)无向图形模型（马尔科夫随机场）；(c)因子图。

例子（8.9有条件的独立）。



图D**8.11**-分离的例子。

考虑到图中的图形模型 8.11.目视检查给我们提供了

$$b \perp d \mid a, c \tag{8.35}$$
$$a \perp c \mid b \tag{8.36}$$
$$b \not\perp \perp d \mid c \tag{8.37}$$
$$a \not\perp \perp c \mid b, e \tag{8.38}$$

有向图形模型允许紧凑地表示原生模型，我们将在第11章中看到有向图形模型的例子。9,10,和11章中看到有向图形模型的例子。这种表示方法，加上条件独立性的概念，使我们能够将各自的概率模型分解成更容易优化的表达式。

概率模型的图形表示使我们能够直观地看到我们所做的设计选择对模型结构的影响。我们经常需要对模型的结构进行高级假设。这些建模假设（超参数）会影响预测性能，但不能用我们目前看到的方法直接选择。我们将在第二节讨论选择结构的不同方法。8.6.

8.5.3 进一步阅读

关于概率图形模型的介绍可以在Bishop（2006，第8章）中找到，关于不同的应用和相应的算法含义的广泛描述可以在Koller和Friedman（2009）的书中找到。有三种主要的概率图形模型。

- *有向图形模型*（*贝叶斯网络*）；见图8.12（a）
- *无向图形模型*（*马尔科夫随机场*）；见图8.12（b）
- *因子图*；见图8.12（c）

有向图形模型
贝叶斯网络
无定向图形模型
马尔科夫随机场
因素图

图形模型允许基于图形的推理和学习算法，例如，通过本地消息传递。应用范围包括网络游戏中的排名（Herbrich等人，2007）和计算机视觉（例如，图像分割、语义标签、图像去噪、图像修复（Kittler和Eglein，1984；Sucar和Gillies，1994；Shotton等人，2006；Szeliski等人，2008）到编码理论（McEliece等，1998），解决线性方程组（Shental等，2008），以及信号处理中的迭代贝叶斯状态估计（Bickson等，2007；Deisenroth和Mohamed，2012）。

有一个在实际应用中特别重要的话题，我们在本书中没有讨论，那就是结构化预测的理念（Bakir等人，2007；Nowozin等人，2014），它允许机器学习模型处理结构化的预测，例如序列、树和图。神经网络模型的普及允许使用更灵活的概率模型，从而产生了许多有用的结构化模型的应用（Goodfellow等人，2016，第16章）。近年来，由于图形模型在因果推断中的应用，人们对其重新产生了兴趣（Pearl,2009；Imbens和Rubin,2015；Peters等人,2017；Rosenbaum,2017）。

8.6 模型选择

在机器学习中，我们经常需要做出高层次的建模决定，这些决定对模型的性能有着至关重要的影响。我们所做的选择（例如，似然的函数形式）影响了模型中自由参数的数量和类型，从而也影响了模型的灵活性。和模型的可表达性。更复杂的模型在A多项式中更加灵活
在这个意义上，它们可以被用来描述更多的数据集。例如，1度的多项式（一条直线$y=a_0+a_1x$）只能用来描述输入x和观测值y之间的线性关系，2度的多项式可以额外描述输入和观测值之间的二次方关系。

的表达能力更强。一个普遍的问题

现在人们会认为，非常灵活的模型通常比简单的模型要好，因为它们

$y = a_0 + a_1 x_2^2 + a_2$也可以通过设置$a_2 = 0$来描述线性函数，也就是说，严格来说，它比*当模型遇到数据时*一阶多项式更具表达力。

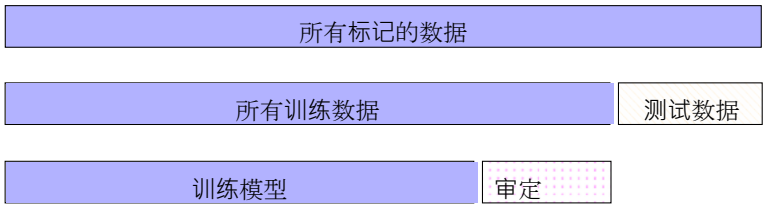


图 8.13嵌套交叉验证。我们进行两级的K-折交叉验证。

在训练时，我们只能使用训练集来评估模型的性能并学习其参数。然而，训练集上的性能并不是我们真正感兴趣的东西。在第8.3,我们已经看到，最大似然估计会导致过度拟合，特别是当训练数据集很小的时候。理想情况下，我们的模型（也）在测试集上工作得很好（测试集在训练时是不可用的）。因此，我们需要一些机制来评估一个模型对未见过的测试数据的*泛化情况*。*模型选择*正是关注这个问题的。

8.6.1 嵌套交叉验证法

我们已经看到了一种可以用于模型选择的方法（第8.2.4节中的交叉验证法）。回顾一下，交叉验证法通过反复将数据集分割成训练集和验证集来提供泛化误差的估计。我们可以再一次应用这个想法，也就是说，对于每一次分割，我们可以再进行一轮交叉验证。这有时被称为*嵌套式交叉验证*；见图8.13. 内层的交叉验证是指对数据集的训练和验证。

嵌套

交叉验证

级是用来估计一个特定的模型或超参数选择在内部验证集上的性能。外层用于估计内循环所选择的最佳模型的泛化性能。我们可以在内循环中测试不同的模型和超参数选择。为了区分这两个层次，用于估计的集合是

测试组 验证组

归纳性能的*测试集*通常被称为*测试集*，用于选择最佳模型的集子被称为*验证集*。内循环估计一个给定模型的泛化误差的预期值(8.39)，通过使用验证集上的经验误差对其进行近似，即。

$$\mathbb{E}_{\mathcal{V}_k}[\mathbf{r}(\mathsf{v}\mid m)] \approx \frac{1}{K}\sum_{k=1}^{K}\mathbf{r}(\mathsf{v}^{(k)}\mid m)\,, \tag{8.39}$$

标准误差定义为$\sigma/\sqrt{K}$。其中K是实验的数量，σ是每个实验的风险的标准偏差。

其中$\mathbf{R}(M_j)$是模型M在验证集上的经验风险（如均方根误差）。我们对所有模型重复这个程序，并选择表现最好的模型。请注意，交叉验证不仅给我们提供了预期的泛化误差，而且我们还可以获得高阶统计数据，例如，标准误差，对模型的不确定性的估计。

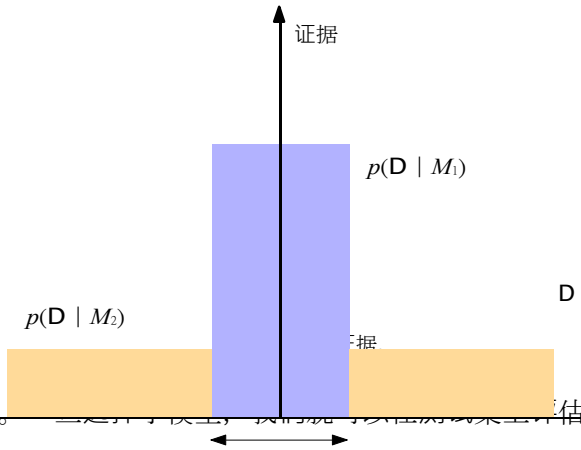


图8.14 贝叶斯推理体现了奥卡姆剃刀。横轴描述了所有可能的数据集的空间。证据（纵轴）评估了一个模型对现有数据的预测程度。 $\boldsymbol{D}$ 由于 $p(\boldsymbol{D} \mid M_i)$需要整合到1，我们应该选择模型与

改编自来自MacKay（2003）。

平均估计值是。 是选择模型，我们就可以在测试集上评估最终的性能。

8.6.2 贝叶斯模型选择

有很多选择模型的方法，本节将介绍其中一些。一般来说，它们都试图在模型的复杂性和数据拟合之间进行权衡。我们假设较简单的模型比复杂的模型更不容易过度拟合，因此，模型选择的目的是找到能合理解释数据的最简单的模型。这个概念是

又称*奥卡姆剃刀*。

奥卡姆剃刀

*备注。*如果我们把模型选择当作一个假设检验问题，我们要寻找与数据一致的最简单的假设（Mur-phy，2012）。 ◆

人们可以考虑在模型上放置一个倾向于更简单模型的先验。然而，没有必要这样做。一个 "自动的奥卡姆剃刀 "在贝叶斯概率的应用中得到了定量的体现（Smith and Spiegelhalter, 1980; Jefferys and Berger, 1992; MacKay, 1992）。图8.14,图，给了我们一个基本的直觉，为什么复杂和非常有表现力的模型可能会变成一个不太可能的模型？

这些预测是对给定数据集D进行建模的选择。

代表所有可能的数据集的空间。如果我们对数据下模型M的$_i$后验概率 $p(M_i \mid \mathsf{D})$感兴趣，我们可以采用贝叶斯定理。假设在所有模型上有一个统一的先验$p(M)$，贝叶斯定理对模型的奖励程度与它们的前验概率成正比。

是由一个归一化的概率分布来量化的，也就是说，它需要积分/求和到 1。

预测了所发生的数据。这种对数据的预测给定了模型

一个简单的模型M_1只能 对少量的数据集进行

证据预测，这由$p($ $M_1)$显示；一个更强大的模型M_2，例如，比M_1有更多的自由参数，能够

来预测更多的数据集。然而，这意味着M_2对区域C中的数据集的预测不如M好$_1$。假设为两个模型分配了相等的先验概率。那么，如果数据集落入区域C，功能较弱的模型M_1是更有可能的模型。

在本章的前面，我们认为模型需要能够解释数据，也就是说，应该有办法从一个给定的模型中生成数据。此外，如果模型已经从数据中适当地学习了，那么我们期望生成的数据应该与经验数据相似。为此，将模型选择表述为一个分层推理问题是很有帮助的，这使我们能够计算模型的后验分布。

让我们考虑一个有限数量的模型$M = \{M_1, \ldots, M_K\}$其中每个模型M_k拥有参数$\boldsymbol{\theta}_k$。

贝叶斯模型

在模型集上选择一个先验$p(M)$。相应的*生成性*让我们从这个模型中产生数据的生成*过程*是

$$M_k \sim p(M) \tag{8.40}$$

$$\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta} \mid M_k) \tag{8.41}$$

$$\mathbf{D} \sim p(\mathbf{D} \mid \boldsymbol{\theta}_k) \tag{8.42}$$

图 8.15 贝叶斯模型选择中分层生成过程的说明。我们在模型的集合上放置一个先验$p(M)$。对于每个模型，都有一个分布$p(\boldsymbol{\theta} \mid M)$对相应的模型参数的影响，这被用于产生数据$\mathbf{D}$.

并在图中说明8.15.给定一个训练 D集，我们应用贝叶斯定理，计算模型的后验分布为

$$p(M_k \mid \mathbf{D}) \propto p(M_k)p(\mathbf{D} \mid M_k) \text{。} \tag{8.43}$$

注意，这个后验不再依赖于模型参数$\boldsymbol{\theta}_k$，因为它们在贝叶斯设定中已经被整合掉了，因为

$$p(\mathbf{D} \mid M_k) = \int p(\mathbf{D} \mid \boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k \mid M_k)d\boldsymbol{\theta}_k \text{，} \tag{8.44}$$

模型证据边际似然

其中，$p(\boldsymbol{\theta}_k \mid M_k)$是模型参数$\boldsymbol{\theta}_k$的先验分布。
模型M_k.的术语(8.44)被称为*模型证据*或*边际似然*。从( )中的后验，我们确定MAP估计。8.43)，我们确定MAP估计

$$M = \arg\max_{M_k} p(M_k \mid \mathbf{D}) \text{。} \tag{8.45}$$

在统一先验$p(M_k)=\frac{1}{K}$，即给予每个模型相同的（先验）概率的情况下，确定模型的MAP估计相当于挑选出使模型证据最大化的模型(8.44).

备注（可能性和边际可能性）。似然和边际似然（证据）之间有一些重要区别。虽然似然容易过拟合，但边际似然通常不会，因为模型参数已经被边际化了（即我们不再需要拟合参数）。此外，边际似然

罩子自动体现了模型复杂性和数据拟合之间的权衡（奥卡姆剃刀）。

◆

8.6.3 用于模型比较的贝叶斯系数

考虑比较两个概率模型M_1、M_2的问题，给定一个数据集。如果我们计算出后验$p(M_1|\mathcal{D})$和$p(M_2|\mathcal{D})$，我们可以计算出后验的比率

$$\frac{p(M_1|\mathcal{D})}{p(M_2|\mathcal{D})} = \frac{\frac{p(\mathcal{D}|M_1)p(M_1)}{p(\mathcal{D})}}{\frac{p(\mathcal{D}|M_2)p(M_2)}{p(\mathcal{D})}} = \frac{p(M_1)}{p(M_2)}\frac{p(\mathcal{D}|M_1)}{p(\mathcal{D}|M_2)}. \tag{8.46}$$

后期赔率　先前的赔率　贝叶斯系数

后验数的比率也被称为*后验几率*。(8.46)右边的第一个分数 --后验赔率

，即先验赔率，衡量我们的先验（初始）信念对M的青睐程度。8.46)右边的第一项

，即*先验几率*，衡量的是我们的先验（初始）信念对M_1比对M_2有多少 先验几率

。

lihoods（右手边的第二个分数）被称为*Bayes系数* Bayes系数

并衡量与M_2相比，数据$\mathcal{D}$被M预测的程度。

*备注。杰弗里斯-林德利悖论*指出，"贝叶斯系数总是 杰弗里斯-林德利的。

倾向于更简单的模型，因为在一个复杂的模型下，具有扩散性先验的数 悖论

据的概率将非常小"（Murphy,2012）。这里，扩散性先验指的是不倾向于特定模型的先验，也就是说。

许多模型在这个先验条件

下是先验可信的。◆

如果我们对模型选择一个统一的先验，那么(8.46)中的先验几率项是1，即后验几率是边际可能性（贝叶斯系数）的比率。8.46)是1，即后验几率是边际可能性（贝叶斯系数）的比率

$$\frac{p(\mathcal{D}|M_1)}{p(\mathcal{D}|M_2)}. \tag{8.47}$$

如果贝叶斯系数大于1，我们就选择模型M_1，否则就选择模型M_2。与频繁主义统计类似，在结果的"显著性"之前，人们应该考虑比率的大小，有一些准则（Jeffreys，1961）。

备注（计算边际似然）。边际似然在模型选择中起着重要作用。我们需要计算贝叶斯系数(8.46)和模型的后验分布(8.43).

不幸的是，计算边际似然需要我们解决一个积分(8.44).这种积分通常是难以分析的，我们将不得不求助于近似技术，例如，数字积分（Stoer和Burlirsch，2002），使用蒙特卡罗的随机近似（Murphy，2012），或贝叶斯蒙特卡罗技术（O'Hagan，1991；Rasmussen和Ghahramani，2003）

。

然而，在一些特殊情况下，我们可以解决这个问题。在第6.6.1节，我们讨论了共轭模型。如果我们选择一个共轭参数先验$p(\boldsymbol{\theta})$，我们可以以封闭形式计算边际似然。在第6章中

之三，我们将在线性回归 的背景下做的正是这个

。 ◆

在本章中，我们已经看到了对机器学习基本概念的简要介绍。在本书的其余部分中，我们将看到

在第8.4节中的三种不同的学习方式是如何应用于机器学习的四大支柱（回归、降维和分类）的。8.2,8.3和8.4节中的三种不同的学习方式如何应用于机器学习的四大支柱（回归、降维、密度估计和分类）。

8.6.4 进一步阅读

我们在本节开始时提到，有一些高级别的建模选择会影响模型的性能。例子包括以下几点。

- 回归中的多项式的度数 混合模型中的成分数
- 深度神经网络的网络结构 支持向量机的内核类型
- PCA中潜在空间的维度
- 优化算法中的学习率（时间表）。

在参数化模型中，参数的数量往往与模型类的复杂性有关。

Rasmussen和Ghahramani(2001)表明，自动奥卡姆剃刀不一定会惩罚模型中的参数数量，但它在函数的复杂性方面是活跃的。他们还表明，自动奥卡姆剃刀也适用于具有许多参数的贝叶斯非参数模型，例如高斯过程。

如果我们专注于最大似然估计，存在一些启发式的模型选择方法，以阻止过度拟合。它们被称为信息准则，我们选择具有最大值的模型。

Akaike信息准则

Akaike信息准则（AIC）（Akaike，1974年）

$$\log p(\boldsymbol{x} \mid \boldsymbol{\theta}) - M \tag{8.48}$$

纠正了最大似然估计器的偏差，增加了一个惩罚项，以补偿具有大量参数的更复杂模型的过度拟合。这里，M是模型参数的数量。AIC估计了一个给定模型所损失的相对信息。

Bayesian

贝叶斯信息准则（BIC）（Schwarz，1978）。

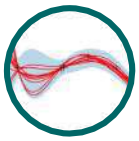
信息标准

$$\log p(\boldsymbol{x}) = \log \int p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \approx \log p(\boldsymbol{x} \mid \boldsymbol{\theta}) - \frac{1}{2} M \text{对数} N \tag{8.49}$$

可用于指数族分布。这里，N是数据点的数量，M是参数的数量。BIC对模型复杂性的惩罚比AIC更严重。

9

线性回归

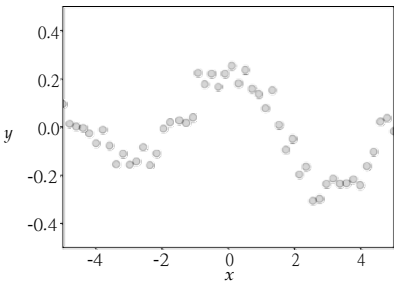


下面，我们将应用第2、5、6和7章的数学概念来解决线性回归（曲线拟合）问题。在

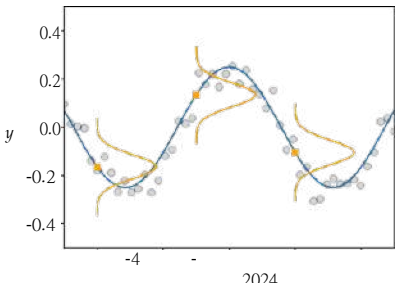
回归，我们的目标是找到一个函数f，将输入$\boldsymbol{x} \in \mathbb{R}^D$到相关对应的函数值$f(\boldsymbol{x}) \in \mathbb{R}$。我们假设给定一组训练输入$\boldsymbol{x}_n$和相应的噪声观测值$y_n = f(\boldsymbol{x}_n) + E$，其中$E$是一个描述测量/观测噪声和潜在的未建模过程（我们在本章中不会进一步考虑）的i.i.d. 随机变量。在本章中，我们假设零均值的高斯噪声。我们的任务是找到一个函数，这个函数不仅可以模拟训练数据，而且可以很好地概括预测不属于训练数据的输入位置的函数值（见第八章）。图中给出了这样一个回归问题的说明。9.1.图9.1(a)中给出了一个典型的回归设置。对于一些输入值x_n，我们观察到（有噪声的）函数值$y_n = f(x_n) + E$。任务是推断产生数据的函数f，并对新输入位置的函数值有良好的泛化作用。图9.1(b)给出了一个可能的解决方案，其中我们还显示了以函数值$f(x)$为中心的三个分布，代表了数据中的噪声。

回归中去

回归是机器学习的一个基本问题，回归问题出现在不同的研究领域和应用中。



(a) 回归问题：观察到的噪声函数值，我们希望从中推断出产生数据的基本函数。



(b) 回归方案：可能产生数据的函数（蓝色），并标明函数值在相应的投入（橙色分布）的测量噪声。

图（9.1a）数据集；（b）回归问题的可能解决方案。

包括时间序列分析（如系统识别）、控制和机器人学（如强化学习、正向/反向模型学习）、优化（如直线搜索、全局优化）和深度学习应用（如计算机游戏、语音到文本翻译、图像识别、自动视频注释）。回归也是分类算法的一个关键成分。找到一个回归函数需要解决各种问题，包括以下问题。

通常情况下，噪声的类型也可以是一种 "模型选择"，但在本章中我们将噪声固定为高斯。

- **模型（类型）的选择和**回归函数的**参数**化。给定一个数据集，哪些函数类（如多项式）是为数据建模的良好候选者，以及我们应该选择什么特定的参数化（如多项式的程度）？正如本节所讨论的，模型选择8.6,允许我们比较各种模型，以找到能合理解释训练数据的最简单模型。
- **寻找好的参数。**在选择了回归函数的模型后，我们如何找到好的模型参数？在这里，我们需要研究不同的损失/目标函数（它们决定了什么是 "好的 "拟合）和优化算法，使我们能够最小化这种损失。**过度拟合和模型选择。**过度拟合是一个问题，当回归函数与训练数据拟合得 "太好"，但对未见过的测试数据并不适用。过度拟合通常发生在基础模型（或其参数化）过于灵活和富有表现力的情况下；见第8.6. 我们将研究其根本原因并讨论如何在线性回归的背景下减轻过度拟合的影响。
- **损失函数和参数预设之间的关系。**损失函数（优化目标）通常是由概率分析模型激发和引起的。我们将研究损失函数和引起这些损失的基本先验假设之间的联系。
- **不确定性建模。**在任何实际环境中，我们只能获得有限的、可能是大量的（训练）数据来选择模型类别和相应的参数。鉴于这个有限的训练数据并不涵盖所有可能的情况，我们可能希望描述剩余的参数不确定性，以获得测试时对模型预测的信心；训练集越小，不确定性建模就越重要。对不确定性的一致建模使模型的预测具有信心界限。

在下文中，我们将使用第三章、第五章、第六章和第七章的数学工具来解决线性回归问题。我们将讨论最大似然法和最大后验法（MAP）估计，以找到最佳模型参数。利用这些参数估计，我们将对泛化误差和过拟合进行简单的研究。在本章的最后，我们将讨论贝叶斯线性回归，它允许我们在更高层次上推理模型参数，从而消除最大似然和MAP估计中遇到的一些问题。

9.1 问题的提出

Because of the presence of observation noise, we will adopt a probabilistic approach and explicitly model the noise using a likelihood function. More specifically, throughout this chapter, we consider a regression problem with the likelihood function

$$p(y \mid \boldsymbol{x}) = \mathcal{N}\, y \mid f(\boldsymbol{x}),\ \sigma^2 。 \tag{9.1}$$

这里，$\boldsymbol{x} \in \mathbb{R}^D$是输入，$y \in \mathbb{R}$是噪声函数值（目标）。
通过(9.1)，$\boldsymbol{x}$和y之间的函数关系为

$$y = f(\boldsymbol{x}) + \varepsilon , \tag{9.2}$$

其中，$\varepsilon \sim \mathcal{N}(0, \sigma^2)$是独立、相同分布（i.i.d.）的高斯测量噪声，平均值为0，方差为σ^2。 我们的目标是找到一个与产生数据的未知函数f相近（类似）的函数，并且能很好地概括。

在本章中，我们重点讨论参数模型，也就是说，我们选择一个参数化的函数，并找到对数据建模 "效果好 "的参数$\boldsymbol{\theta}$。在线性回归中，我们考虑的特殊情况是参数$\boldsymbol{\theta}$在我们的模型中线性出现。线性回归的一个例子是这样给出的

$$p(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{N}\, y \mid \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta},\ \sigma^2 \tag{9.3}$$

$$\Leftrightarrow \quad y = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{9.4}$$

其中$\boldsymbol{\theta} \in \mathbb{R}^D$是我们寻求的参数。用(9.4)描述的那类函数是通过原点的直线。9.4)描述的一类函数是通过原点的直线。在(9.4),
我们选择了一个参数化$f(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}$。

一个狄拉克三角（delta函数）在任何地方都是零，除了一个点，它的积分是1。可以认为是$\sigma^2 \to 0$的极限中的高斯。 可能性

(9.3)是在$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}$处评估的$y$的概率密度函数。请注意，唯一的不确定性来源来自于观测噪声（因为在（）中假定$\boldsymbol{x}$和$\boldsymbol{\theta}$是已知的。9.3)).如果没有观测噪声，$\boldsymbol{x}$和y之间的关系将是确定性的，(9.3)将是一个狄拉克三角。

例子 9.1
对于$x, \theta \in \mathbb{R}$，线性回归模型在(9.4)描述了直线
(线性函数），而参数θ是直线的斜率。图9.2(a)显示了一些不同θ值的函数实例。

线性回归指的是在参数上是线性的模型。

中的线性回归模型。9.3)-(9.4图9.2(a)显示了此类函数的例子。我们将在后面看到，对于非线性变换$\boldsymbol{\varphi}$，$y = \boldsymbol{\varphi}(\mathrm{T}\boldsymbol{x})\boldsymbol{\theta}$也是一个线性回归模型，因为 "线性回归"

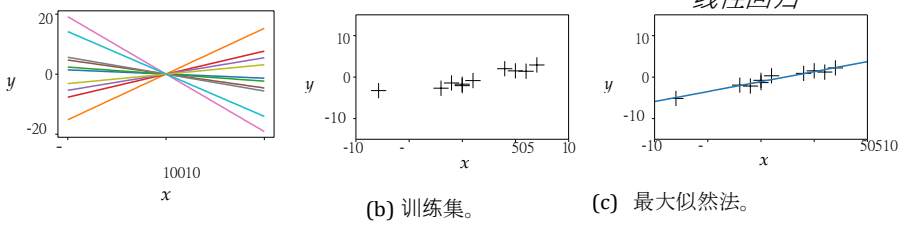


图 线性**9.2**回归实例。(a)属于这一类的函数实例；(b)训练集；(c)最大似然估计。

(a) 可以用( )中的线性模型描述的函数（直线）实例。9.4).

(b) 训练集。

(c) 最大似然法。

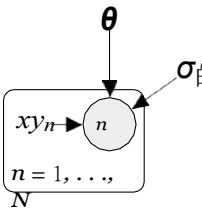
指的是 "参数线性 "的模型，即通过输入特征的线性组合来描述一个函数的模型。这里，"特征 "是输入$\boldsymbol{x}$的表示方法$\boldsymbol{\varphi}(\boldsymbol{x})$。

在下文中，我们将更详细地讨论如何找到好的爸爸。参数$\boldsymbol{\theta}$，以及如何评价一个参数集是否 "效果好"。目前，我们假设噪声方差σ^2是已知的。

9.2 参数估计

训练集

图为线性回归的概率**9.3**图形模型。观察到的随机变量有阴影。确定性的已知的数值是没有圆圈的。



考虑到线性回归的设置(9.4)，并假设我们得到一个*训练集* $:= (\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$由$N$个输入$\boldsymbol{x}_n \in \mathbb{R}^D$和相应的观察/目标$y_n \in \mathbb{R}$组成，$n = 1, \ldots, N$。相应的图形模型在图中给出 9.3.请注意，$y_i$和$y_j$是有条件独立的，给定它们各自的输入$\boldsymbol{x}_i$，$\boldsymbol{x}_j$，所以似然因子根据以下情况进行分解

$$p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\theta}) = p(y_1, \ldots, y_N \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{\theta}) \tag{9.5a}$$

$$= \prod_{n=1}^{N} p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{\theta}, \sigma^2\right) \tag{9.5b}$$

其中我们定义了$\mathbf{X} := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$和$\mathbf{Y} := \{y_1, \ldots, y_N\}$作为集合的训练输入和相应的目标，分别。由于噪声分布，似然和因子$p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})$是高斯的；见(9.3).

下面，我们将讨论如何找到线性回归模型的最佳参数$\boldsymbol{\theta}^* \in \mathbb{R}^D$ (9.4).一旦找到了参数$\boldsymbol{\theta}^*$，我们就可以通过使用这个参数估计来预测函数值，在(9.4)，这样，在一个任意的测试输入$\boldsymbol{x}_*$，相关的目标y_*的分布是

$$p(y_* \mid \boldsymbol{x}_*, \boldsymbol{\theta}^*) = \mathcal{N}\left(y_* \mid \boldsymbol{x}_*^{\mathrm{T}}\boldsymbol{\theta}^*, \sigma^2\right) \tag{9.6}$$

在下文中，我们将看一下通过最大似然来估计参数，这个话题我们已经在第二节中在某种程度上涵盖了。8.3.

9.2.1 最大似然估计

一个广泛使用的寻找理想参数$\boldsymbol{\theta}_{\mathrm{ML}}$的方法是*最大概率估计*，在这里我们找到参数$\boldsymbol{\theta}_{\mathrm{ML}}$，使其最大化。可能性（9.5b）。直观地讲，最大化可能性意味着最大化训练数据的预测分布是由模型参数决定的。我们得到最大似然参数为

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\theta}) \text{。} \tag{9.7}$$

最大似然法。估算-最大化似然是指在给定参数的情况下使（训练）数据的预测分布最大化。似然不是参数中的概率分布。

备注。似然$p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$不是$\boldsymbol{\theta}$中的概率分布。它只是参数$\boldsymbol{\theta}$的一个函数，但没有积分到（1即没有归一化），甚至可能无法对$\boldsymbol{\theta}$*进行*整数化。9.7)是一个归一化的概率分布在$\boldsymbol{y}$。◆

为了找到理想的参数$\boldsymbol{\theta}_{\mathrm{ML}}$，使可能性最大化，我们通常会进行梯度上升（或梯度下降，对负的似然）。然而，在我们这里考虑的线性回归的情况下，由于对数是一个（严格）单调递增的函数，一个函数f的最优与$\log f$的最优是相同的。存在一个闭合形式的解决方案，这使得迭代梯度下降没有必要。在实践中，我们不是直接最大化似然，而是对似然函数进行对数变换，并最小化负对数似然。

备注（对数变换）。由于可能性（9.5b）是N个高斯分布的乘积，对数变换是有用的，因为（a）它不会出现数值下溢，（b）微分规则会变得更简单。更具体地说，当我们将N个概率相乘（其中N是数据点的数量）时，数字下溢将是一个问题，因为我们不能表示非常小的数字，如10^{-256}.A。此外，对数转换将把乘积变成对数概率的总和，这样相应的梯度就是单个梯度的总和，而不是重复应用乘积规则（5.46）到计算N个项的乘积的梯度。◆

为了找到我们线性回归问题的最佳参数$\boldsymbol{\theta}_{\mathrm{ML}}$，我们最小化负对数可能性

$$-\log p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\theta}) = -\log \mathrm{p}(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}) \text{，} \tag{9.8}$$

由于我们对训练集的独立假设，我们利用了可能性（9.5b）对数据点数量的因数化。

在线性回归模型(9.4)，可能性是高斯的(由于高斯加性噪声项)，这样我

$$\log p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}) = -\frac{1}{2\sigma^2}(y_n - \boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{\theta})^2 + \text{常数}, \tag{9.9}$$

其中常数包括独立于$\boldsymbol{\theta}$的所有项。使用(9.9)在

负的对数可能性(9.8)，我们得到(忽略常数项)

$$\frac{1}{2\sigma^2}(\boldsymbol{\theta}) := \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \boldsymbol{x}_n^\top\boldsymbol{\theta})^2 \tag{9.10a}$$

$$= \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) = \frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2。 \tag{9.10b}$$

负对数似然函数也是称为*错误函数*。

设计矩阵

误差的平方经常被用作衡量距离的标准。回顾一下第3.1节的内容，即$\boldsymbol{x} = \boldsymbol{x}^\top\boldsymbol{x}$，如果我们选择点的乘积为内积。

其中我们定义*设计矩阵*$\boldsymbol{X} := [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top \in \mathbb{R}^{N\times D}$为训练输入的集合，$\boldsymbol{y} := [y_1, \ldots, y_N]^\top \in \mathbb{R}^N$为集合所有训练目标的向量。注意，设计矩阵$\boldsymbol{X}$中的第$n$行对应于训练输入$\boldsymbol{x}_n$。

观测值y_n和相应模型预测值$\boldsymbol{x}_n^\top\boldsymbol{\theta}$之间的平方误差之和等于$\boldsymbol{y}$和$\boldsymbol{X}\boldsymbol{\theta}$之间的平方距离。

有了(9.10b)，我们现在有了一个负对数可能性的具体形式我们需要优化的函数。我们立即看到(9.10b)对$\boldsymbol{\theta}$是二次的，这意味着我们可以找到一个唯一的全局解决方案$\boldsymbol{\theta}_{\mathrm{ML}}$来最小化负对数似然。我们可以通过以下方法找到全局最优的梯度，将其设为$\mathcal{L}$的梯度，将其设为$\mathcal{L}$g并o求解$\boldsymbol{\theta}$。

利用第五章的结果，我们计算出相对于参数的梯度为$\mathcal{L}$

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\boldsymbol{\theta}} = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\left(\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})\right) \tag{9.11a}$$

$$= \frac{1}{2\sigma^2}\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\left(\boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\theta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\theta}\right) \tag{9.11b}$$

$$= \frac{1}{\sigma^2}(-\boldsymbol{y}^\top\boldsymbol{X} + \boldsymbol{\theta}^\top\boldsymbol{X}^\top\boldsymbol{X}) \in \mathbb{R}^{1\times D}。 \tag{9.11c}$$

最大似然估计器$\boldsymbol{\theta}_{\mathrm{ML}}$解决了$= (\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\boldsymbol{\theta}}\mathbf{o}^\top$必要的操作性）。

忽略重复数据点的可能性，$\mathrm{rk}(\boldsymbol{X}) = D$

$\boldsymbol{x}$

如果$N < D$，也就是说，我们没有更多的参数比数据点多。

质量条件），我们得到

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\boldsymbol{\theta}} = \mathbf{o}^\top \overset{(9.11c)}{\iff} \boldsymbol{\theta}^\top\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{y}^\top\boldsymbol{X} \tag{9.12a}$$

$$\overset{\mathrm{M}}{\mathrm{L}} \quad \boldsymbol{y}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \tag{9.12b}$$

$$\iff \boldsymbol{\theta}_{\mathrm{ML}}^\top)$$

$$\iff \boldsymbol{\theta}_{\mathrm{ML}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}。 \tag{9.12c}$$

我们可以用$(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$对第一个方程进行右乘，因为如果$\mathrm{rk}(\boldsymbol{X}) = D$，$\boldsymbol{X}^\top\boldsymbol{X}$是正定的，其中$\mathrm{rk}(\boldsymbol{X})$表示$\boldsymbol{X}$的等级。

*备注。*设置梯度为是$\mathbf{o}^\top$一个必要且充分的条件，我们得到一个全局最小值，因为Hessian $\nabla^2_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) = \boldsymbol{X}^\top\boldsymbol{X} \in \mathbb{R}^{D\times D}$是正定的。 ◆

备注。(9.12c)中的最大似然解要求我们解决一个形式为$\boldsymbol{A}\boldsymbol{\theta} = \boldsymbol{b}$*的*线性方

程组 $\boldsymbol{A}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})$，$\boldsymbol{b}=\boldsymbol{X}\boldsymbol{y}^{\mathrm{T}}$。 ◆

例子（9.2拟合线）。
让我们看一下图9.2,，我们的目标是用最大似然估计法将一条直线$f(x) = \theta x$，其中θ是一个未知的斜率，拟合到数据集上。图9.2(a)中显示了该模型类（直线）的函数实例。对于图9.2(b)所示的数据集，我们用(9.12c)找到斜率参数θ的最大似然估计，得到图9.2(c)中的最大似然线性函数。

有特征的最大似然估计

到目前为止，我们考虑了（）中描述的线性回归设置。9.4)中描述的线性回归设置，它允许我们使用最大似然法对数据进行直线拟合。

估计。然而，当它

表现力是不够的。

来拟合更多有趣的数据。幸运的是，线性回归为我们提供了一种在线性回归框架内拟合非线性函数的方法。由于 "线性回归 "仅指 "参数的线性"，我们可以对输入$\boldsymbol{x}$进行任意的非线性变换$\boldsymbol{\varphi}(\boldsymbol{x})$，然后线性地组合这个变换的组成部分。相应的线性回归模型是

$$p(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = \mathsf{N}\, y \mid \boldsymbol{\varphi}(^{\mathsf{T}}\boldsymbol{x})\boldsymbol{\theta},\ \sigma^2$$
$$\Leftrightarrow y = \boldsymbol{\varphi}(^{\mathsf{T}}\boldsymbol{x})\boldsymbol{\theta} + E = \sum_{k=0}^{K-1} \theta_k \varphi_k(\boldsymbol{x}) + E \tag{9.13}$$

。

其中$\boldsymbol{\varphi}: \mathcal{D} \to \mathsf{RR}^K$是输入$\boldsymbol{x}$和RR的（非线性）变换。
$\varphi_k: \mathcal{D} \to \mathsf{RR}$是*特征向量*$\boldsymbol{\varphi}$的第$k$*个*分量。 注意， 特征向量
模型参数$\boldsymbol{\theta}$仍然只以线性方式出现。

线性回归时，直线的

指的是 "参数中的线性 "回归模型，但输入可以经历任何非线性转换。

示例（9.3多项式回归）。
我们关注的是一个回归问题$y = \boldsymbol{\varphi}(^{\mathsf{T}}x)\boldsymbol{\theta} + E$， 其中$x \in \mathsf{R}$和$\boldsymbol{\theta} \in \mathsf{R}^K$.在这种情况下经常使用的一个变换是

$$\boldsymbol{\varphi}(x) = \begin{bmatrix} \varphi_0(x) \\ \varphi_1(x) \\ \vdots \\ \varphi_{K-1}(x) \end{bmatrix} = \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \\ \vdots \\ x^{K-1} \end{bmatrix} \in \mathsf{R}^K. \tag{9.14}$$

这意味着我们将原来的一维输入空间 "提升 "为一个由所有单项式x组成^{k}的K维特征空间，$K = 0, \ldots, K-1$.有了这些特征，我们可以为度数为0.5的多项式建立模型在线性回归的框架内的$K-1$。度的多项式

$K-1$是

$$f(x) = \sum_{k=0}^{K-1} \theta_k x^k = \boldsymbol{\varphi}(x)^\top \boldsymbol{\theta} \tag{9.15}$$

其中$\boldsymbol{\varphi}$定义在(9.14)，$\boldsymbol{\theta} = [\theta_0, \ldots, \theta_{K-1}]^\top \in \mathbb{R}^K$包含了(线性)参数$\theta_k$。

现在让我们来看看线性回归模型中的参数$\boldsymbol{\theta}$的最大似然估计(9.13).我们考虑训练输入$\boldsymbol{x}_n \in \mathbb{R}^D$和目标$y_n \in \mathbb{R}$，$n = 1, \ldots, N$，并定义*特征矩阵（设计矩阵）*为

特征矩阵设计矩阵

$$\boldsymbol{\Phi} := \begin{bmatrix} \boldsymbol{\varphi}(\boldsymbol{x}_1)^\top \\ \vdots \\ \boldsymbol{\varphi}(\boldsymbol{x}_N)^\top \end{bmatrix} = \begin{bmatrix} \varphi_0(\boldsymbol{x}_1) & \cdots & \varphi_{K-1}(\boldsymbol{x}_1) \\ \varphi_0(\boldsymbol{x}_2) & \cdots & \varphi_{K-1}(\boldsymbol{x}_2) \\ \vdots & & \vdots \\ \varphi_0(\boldsymbol{x}_N) & \cdots & \varphi_{K-1}(\boldsymbol{x}_N) \end{bmatrix} \in \mathbb{R}^{N \times K}, \tag{9.16}$$

其中，$\Phi_{ij} = \varphi_j(\boldsymbol{x}_i)$，$\Phi_j : \mathbb{R}^D \to \mathbb{R}$。

示例（9.4二阶多项式的特征矩阵）。
对于一个二阶多项式和N个训练点$x_n \in \mathbb{R}$, $n = 1, \ldots, N$，特征矩阵为

$$\boldsymbol{\Phi} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 \end{bmatrix}. \tag{9.17}$$

有了定义在(9.16)，线性回归模型的负对数可能性(9.13)可以写成

$$-\log p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\theta}) = \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}) + \text{const}。 \tag{9.18}$$

对比(9.18)与(9.10b)中"无结构"模型的负对数可能性相比较，我们立即发现我们只需要用$\boldsymbol{\Phi}$代替$\boldsymbol{X}$。由于$\boldsymbol{X}$和$\boldsymbol{\Phi}$都与我们希望优化的参数$\boldsymbol{\theta}$无关，我们立即得到了*最大可能性估计*

最大似然估计

$$\boldsymbol{\theta}_{\text{ML}} = (\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{y} \tag{9.19}$$。

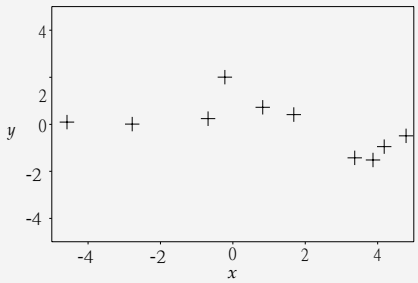
中定义的具有非线性特征的线性回归问题的(9.13).

*备注。*当我们在没有特征的情况下工作时，我们要求$\boldsymbol{X}^\top\boldsymbol{X}$是可逆的，

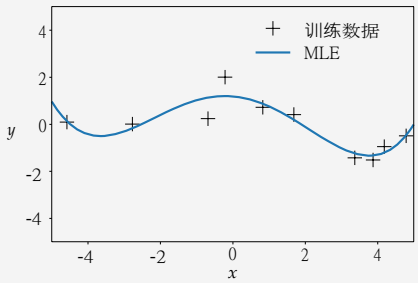
当 *9.2 参数估计*
rk(***X***)=
D 时就是这种情况，也就是说，***X*** 的列

是线性独立的。在(9.19)中，我们因此要求$\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\in\mathrm{R}^{K\times K}$是可逆的。当且仅当$\mathrm{rk}(\mathbf{\Phi})=K$时，情况就是如此。

例子（9.5最大似然多项式拟合）。



(a) 回归



数据集。(b)4由max-大数似然估计。

Figure 9.4 Polynomial regression: (a)dataset consisting of (x_n, y_n) pairs, $n = 1, \ldots, 10$; (b)度数的最大似然多项式 4.

考虑图9.4（a）中的数据集。该数据集由$N = 10$对(x_n, y_n)，其中$x_n \sim \mathrm{U}[-5, 5]$，$y_n = -\sin(x/5) + \cos(x) + \varepsilon$其中$\varepsilon \sim \mathrm{N}(0, 0.2^2)$。

我们用4最大似然估计法拟合度数的多项式。即参数$\boldsymbol{\theta}_{\mathrm{ML}}$是在(9.19).最大似然估计在任何测试位置$x_*$，都能得到函数值$\boldsymbol{\phi}(x_*)^{\mathrm{T}}\boldsymbol{\theta}_{\mathrm{ML}}$，其结果是如图9.4（b）所示。

估算噪声方差

到目前为止，我们假设噪声方差σ^2是已知的。然而，我们也可以利用最大似然估计的原理来获得噪声方差的最大似然估计器σ^2。为了做到这一点，我们遵循标准程序。我们写下对数可能性，计算其相对于$\sigma^2>0$的导数，将其设为0，然后求解。对数可能性由以下公式给出

$$\log p(\mathbf{y}\mid \mathbf{X}, \boldsymbol{\theta}, \sigma^2) = \sum_{n=1}^{N} \log \mathcal{N}(y_n \mid \boldsymbol{\phi}(\boldsymbol{x}_n)^{\mathrm{T}}\boldsymbol{\theta}, \sigma^2) \tag{9.20a}$$

$$= \sum_{n=1}^{N} \left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_n - \boldsymbol{\phi}(\boldsymbol{x}_n)^{\mathrm{T}}\boldsymbol{\theta})^2\right) \tag{9.20b}$$

$$= -\frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \boldsymbol{\phi}(\boldsymbol{x}_n)^{\mathrm{T}}\boldsymbol{\theta})^2 + \text{const.} \tag{9.20c}$$

那么，关于σ的对数可能性的偏导[2]就是

$$\frac{\partial \text{对数} p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}, \sigma)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} s = 0 \tag{9.21a}$$

$$\Leftrightarrow \sigma^2 = \frac{s}{N} \tag{9.21b}$$

这样我们就可以确定

$$\sigma^2_{\text{劳动部}} = \frac{1}{N} \sum_{n=1}^{N} (y_n - \boldsymbol{\phi}(\mathbf{x}_n)^\top \boldsymbol{\theta})^2. \tag{9.22}$$

因此，噪声方差的最大似然估计是无噪声函数值$\boldsymbol{\phi}(\mathbf{x}_n)^\top \boldsymbol{\theta}$与输入locations $\mathbf{x}$处n相应的噪声观测值y之间的平方距离的经验平均值n。

9.2.2 线性回归中的过度拟合

我们刚刚讨论了如何使用最大似然估计来拟合数据的线耳模型（例如，多项式）。我们可以通过计算产生的误差/损失来评估模型的质量。一种方法是计算负对数似然（9.10b），我们将其最小化以确定最大似然估计。另外，鉴于噪声参数σ^2不是一个自由的模型参数，我们可以忽略$1/\sigma^2$的比例，因此我们最终得到一个平方误差-损失函数$\|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\|^2$。我们通常不使用这个平方损失，而是使用*均值根*

均方根
误差平方误差（*RMSE*）
RMSE

$$\sqrt{\frac{1}{N}\|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\|^2} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (y_n - \boldsymbol{\phi}^\top(\mathbf{x}_n)\boldsymbol{\theta})^2}, \tag{9.23}$$

它(a)允许我们比较不同大小的数据集的误差，(b)具有与观察到的函数相同的比例和单位。

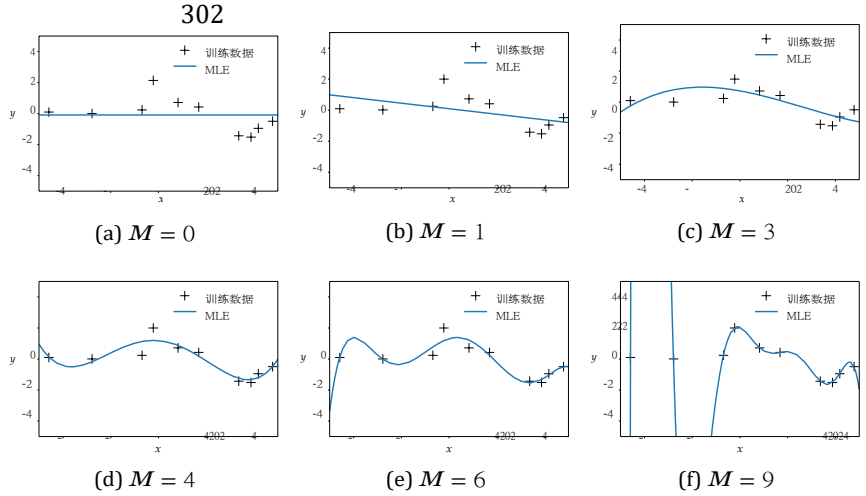
RMSE为

例如，如果我们拟合一个模型，将邮政编码（$\mathbf{x}$是以纬度、经度为单位）与房价（Y值为欧元）的关系，那么RMSE也是以欧元为单位，而平方误差则是以欧元为单位。

负对数可能性是无单位的。

在欧元[2]。如果我们选择包括原始负对数可能性（9.10b）中的因子σ^2，那么我们最终会得到一个无单位的目标，也就是说，在前面的例子中，我们的目标将不再是以欧元为单位或欧元[2]。

对于模型的选择（见第8.6），我们可以使用RMSE（或负对数可能性），通过找到使目标最小化的多项式度数M来确定多项式的最佳度数。鉴于多项式的度数是一个自然数，我们可以进行粗暴的搜索，列举出所有（合理的）M的值。对于一个大小为N的训练集，只需测试$0 \ldots M \leq N-1$。对于$M < N$，最大似然估计值是唯一的。对于$M \geq N$，我们有更多的参数



(a) $M = 0$ (b) $M = 1$ (c) $M = 3$

(d) $M = 4$ (e) $M = 6$ (f) $M = 9$

图为不同多项式度数M的最大**9.5**似然拟合。

比数据点多，需要解决一个欠确定的线性方程组（$\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$在(9.19)也将不再是可逆的)，因此有无限多可能的最大似然估计。

图9.5显示了一些由图9.4(a)的数据集的最大似然法确定的多项式拟合，其中 N=10个观测值。我们注意到，低度的多项式（例如，常数（M =0）或线性（M = 1））对数据的拟合效果很差，因此，对真正的基本函数的表示很差。对于度数M = 3, . . ., 6其拟合结果看起来貌似合理，并顺利地对数据进行插值。当我们到了更高的度数多项式，我们注意到它们对数据的拟合效果越来越好。在$M = N 1 = 9$ 的典型情况下，该函数将通过每一个数据点。然而，这些高阶多项式会疯狂地摆动，对产生数据的基本函数的表现力很差，因此我们会受到*过度拟合的*影响。

时 M = N1是极端的，因为否则的话就会出现空的空间。相应的线性方程组将是非线性的，而且我们将有无限多的线性回归问题的最优解。

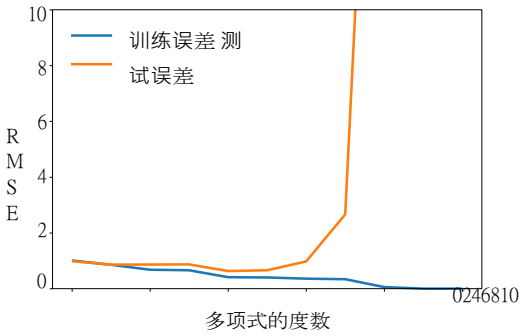
过度拟合

请记住，我们的目标是通过对新的（未见过的）数据做出准确的预测来实现良好的泛化。我们通过考虑一个单独的测试集，其中的数据200点是用生成训练集的相同程序生成的，从而对泛化性能对度数为M的多项式的依赖性有了一些定量的了解。作为测试输入，我们在[5, 5]的区间内选择了200个点的线性网格。对于每个M的选择，我们评估训练数据和测试数据的RMSE (9.23)的训练数据和测试数据。

请注意，噪声方差 $\sigma^2>0$。

现在看一下测试误差，它是对相应多项式的基因化特性的定性衡量，我们注意到，最初的测试误差是下降的；见图9.6(橙色)。对于四阶多项式，测试误差相对较低，并且在度数5之前保持相对稳定。然而，从度数6开始，测试误差明显增加，高阶多项式的泛化能力非常差。

系。在这个特殊的例子中，这也可以从相应的



图为培训**9.6**
和测试错误。

训练误差

测试错误

图9.5中的最大似然拟合。请注意，当多项式的度数增加时，*训练误差*（图9.6中的蓝色曲线）从未增加。在我们的例子中，最好的概括（*测试误差最小的*点）是在度数为M=4的多项式中得到的。

9.2.3 最大A后验估计

我们刚刚看到，最大似然估计很容易出现过拟合。我们经常观察到，如果遇到过拟合，参数值的大小会变得相对较大（Bishop，2006）。

To mitigate the effect of huge parameter values, we can place a prior distribution $p(\boldsymbol{\theta})$ on the parameters. The prior distribution explicitly encodes what parameter values are plausible (before having seen any data). For example, a Gaussian prior $p(\theta) = \mathcal{N}(0, 1)$ on a single parameter θ encodes that parameter values are expected lie in the interval $[-2, 2]$ (two standard deviations around the mean value). Once a dataset $(\boldsymbol{X}, \boldsymbol{Y})$ 的情况下，我们不寻求最大化似然，而是寻求使后验分布$p(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Y})$最大化的参数。这个过程被称为*最大后验*（*MAP*）估计。

最大的后验性MAP

通过应用贝叶斯定理（第6.3节），可以得到参数$\boldsymbol{\theta}$的后验，$\boldsymbol{X}$给定的训练数据、、、的后验为

$$p(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Y}) = \frac{p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\boldsymbol{Y} \mid \boldsymbol{X})}。 \tag{9.24}$$

由于后验明确地取决于参数先验$p(\boldsymbol{\theta})$，先验将对我们找到的作为后验最大化的参数向量产生影响。我们将在下文中更明确地看到这一点。使后验最大化的参数向量$\boldsymbol{\theta}_{\text{MAP}}$ (9.24)就是MAP估计。

为了找到MAP估计，我们遵循与最大似然估计相似的步骤。我们从对数转换开始，计算对数后验，即

$$\log p(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Y}) = \log p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) + \text{const}, \tag{9.25}$$

其中常数包括独立于$\boldsymbol{\theta}$的条款。9.25)是对数似然$p(\ ,\boldsymbol{\theta})$和对数先验$p(\boldsymbol{\theta})$的总和，因此MAP估计将是先验（我们在观察数据之前对可信参数值的建议）和与数据有关的似然之间的 "折衷"。

为了找到MAP估计值$\boldsymbol{\theta}_{\text{MAP}}$，我们使关于$\boldsymbol{\theta}$的负对数后验分布最小化，也就是说，我们解决

$$\boldsymbol{\theta}_{\text{MAP}} \in \arg\min_{\boldsymbol{\theta}} \{-\log p(\mathsf{Y} \mid \mathsf{X}, \boldsymbol{\theta}) - \log p(\boldsymbol{\theta})\}。 \tag{9.26}$$

负对数后验相对于$\boldsymbol{\theta}$的梯度为

$$-\frac{\mathrm{d}\log p(\boldsymbol{\theta} \mid \mathsf{X}, \mathsf{Y})}{\mathrm{d}\boldsymbol{\theta}} = -\frac{\mathrm{d}\log p(\mathsf{Y} \mid \mathsf{X}, \boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} - \frac{\mathrm{d}\log p(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}, \tag{9.27}$$

其中，我们将右手边的第一项确定为$\boldsymbol{\theta}$(9.11c）中负对数似然的梯度。

在参数上有一个（共轭的）高斯先验$p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{m}_0, \boldsymbol{S}_0)$，特别是$\mathbf{m}_0 = \mathbf{0}$，$\boldsymbol{S}_0 = b^2\boldsymbol{I}$。$\boldsymbol{\theta}$，线性回归设置的负对数后验(9.13)，我们获得负对数后验

$$-\log p(\boldsymbol{\theta} \mid \mathsf{X}, \mathsf{Y}) = \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}) + \frac{1}{2b^2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta} + \text{const}。 \tag{9.28}$$

这里，第一项对应于来自对数似然的贡献，第二项则来自对数后验。那么，关于参数$\boldsymbol{\theta}$的对数后验的梯度是

$$-\frac{\mathrm{d}\log p(\boldsymbol{\theta} \mid \mathsf{X}, \mathsf{Y})}{\mathrm{d}\boldsymbol{\theta}} = \frac{1}{\sigma^2}(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Phi}) + \frac{1}{b^2}\boldsymbol{\theta}^{\mathrm{T}}。 \tag{9.29}$$

我们将$_{\text{MAP}}$通过设定这个梯度来找到MAP估计值$\boldsymbol{\theta}$，并$\mathbf{0}^{\mathrm{T}}$求解$\boldsymbol{\theta}_{\text{MAP}}$。我们得到

$$\begin{aligned}
&\frac{1}{\sigma^2}(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Phi}) + \frac{1}{b^2}\boldsymbol{\theta}^{\mathrm{T}} = \mathbf{0}^{\mathrm{T}} && (9.30a)\\
\Leftrightarrow\ &\boldsymbol{\theta}^{\mathrm{T}}\left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \boldsymbol{I}\frac{\sigma^2}{b^2}\right) - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Phi} = \mathbf{0}^{\mathrm{T}} && (9.30b)\\
\Leftrightarrow\ &\boldsymbol{\theta}^{\mathrm{T}}\left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \frac{\sigma^2}{b^2}\boldsymbol{I}\right) = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Phi} && (9.30c)\\
\Leftrightarrow\ &\boldsymbol{\theta}^{\mathrm{T}} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \frac{\sigma^2}{b^2}\boldsymbol{I}\right)^{-1} && (9.30d)
\end{aligned}$$

因此，MAP估计是（通过转置最后一个等式的两边）$\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$是对称的。

$$\boldsymbol{\theta}_{\text{MAP}} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \frac{\sigma^2}{b^2}\boldsymbol{I}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{y}\ . \tag{9.31}$$

积极的半确切地说。额外的术语在(9.31)是严格的

将( )中的MAP估计值与最大似然估计值进行比较。9.31)中的MAP估计值与最大似然估计值

在(9.19)，我们看到两个解决方案之间的唯一区别

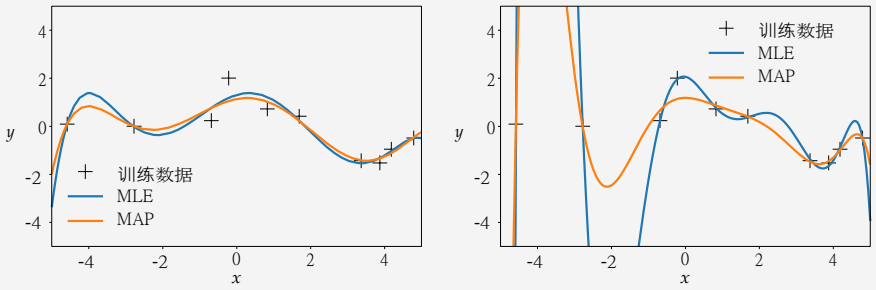
是正定的，因此存在逆。

的附加项$\boldsymbol{I}\sigma^2$。 这个项保证了

$\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}+\boldsymbol{I}\frac{\sigma^2}{b^2}$是对称的和严格正定的（即它的逆存在，MAP估计是线性方程组的唯一解）。此外，它还反映了正则器的影响。

例子（多项式回归的9.6MAP估计）。

在本节9.2.1的多项式回归例子中，我们把一个高斯在参数$\boldsymbol{\theta}$上的先验$p(\boldsymbol{\theta})=\mathcal{N}(\mathbf{0},\boldsymbol{I})$并确定MAP根据(9.31).在图中，9.7,我们同时显示了最大6度（左）和度（8右）多项式的似然和MAP估计。先验（正则器）对低度多项式没有发挥重要作用，但保持了函数的相对平滑性为高阶多项式。虽然MAP估计可以突破过拟合的界限，但它并不是这个问题的一般解决方案，所以我们需要一个更有原则的方法来解决过拟合问题。



(a) 度6的多项式 .(b)度8的多项式.

Figure 9.7 Polynomial regression: maximum likelihood and MAP estimates. (a)Polynomials of degree 6; (b)polynomials of degree 8.

9.2.4 作为规范化的MAP估计

不在参数$\boldsymbol{\theta}$上设置先验分布，而是通过*正则化*惩罚参数的振幅来减轻过拟合的影响也是可能的。在*正则化最小二乘法*中，我们考虑损失函数

正规化

正则化最小二乘法

$$\|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{\theta}\|^2+\lambda\|\boldsymbol{\theta}\|_2^2 \tag{9.32}$$。

我们使其相对于$\boldsymbol{\theta}$最小化（见第8.2.3).这里，第一个项是*数据拟合项*（也叫*错位项*），它与下列各项成正比

数据拟合术语

拟合项即负对数似然；见（9.10b）。第二个项被称为

正则器

正则化，*正则化参数*λ控制0正则化的 "严格程度"。

正则化参数

*备注。*我们可以选择任何p-norm $\|\cdot\|_p$来代替Euclidean norm $\|\cdot\|_2$，在(9.32).在实践中，较小的p值会导致较稀疏的解决方案。这里，"稀疏 "意味着许多参数值$\theta_d=0$，这也是

对变量选择很有用。对于$p=1$，正则器被称为LASSOLASSO（最小绝对收缩和选择算子），是由Tib- shirani（1996）提出。

正则器$\lambda\|\boldsymbol{\theta}\|_2^2$在(9.32)中的正则器可以解释为负对数高斯先验，我们在MAP估计中使用它；见(9.26).更具体地说，用高斯先验$p(\boldsymbol{\theta})=\mathcal{N}(\mathbf{0}, b^2\boldsymbol{I})$，我们得到负对数高斯先验

$$-\log p(\boldsymbol{\theta}) = \frac{1}{2b^2}\|\boldsymbol{\theta}\|_2^2 + \text{const} \tag{9.33}$$

因此，对于$\lambda=\frac{1}{2b^2}$正则化项和负对数高斯先验是相同的。

鉴于( )中的正则化最小二乘法损失函数由与负对数可能性密切相关的项和负对数优先权组成。9.32)中的正则化最小二乘损失函数由与负对数似然和负对数先验密切相关的项组成，因此，当我们最小化这一损失时，我们得到的解决方案与()中的MAP估计非常相似，这并不奇怪。9.31).更具体地说，最小化正则化的最小二乘损失函数可以得到

$$\boldsymbol{\theta}_{\text{RLS}} = (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{y}\,, \tag{9.34}$$

中的MAP估计完全相同。9.31)，σ^2其中σ^2是噪声方差和(各向同性的)高斯先验的b^2方差。$p(\boldsymbol{\theta}) = \mathcal{N}\mathbf{o}\,, b\boldsymbol{I}^2$。

一个点估计是一个单一的具体参数值，而不像合理的参数设置的分布。

到目前为止，我们已经涵盖了使用最大似然和MAP估计的参数估计，其中我们找到了点估计值$\boldsymbol{\theta}^*$，以实现一个目标函数（似然或后验）。我们看到，最大似然法和MAP估计都可能导致过拟合。在下一节中，我们将讨论贝叶斯线性回归，我们使用贝叶斯推理（第8.4节）来寻找未知参数的后验分布，随后我们用它来进行预测。更具体地说，对于预测，我们将对所有可信的参数集进行平均，而不是专注于一个点估计。

9.3 贝叶斯线性回归

之前，我们研究了线性回归模型，在这些模型中，我们通过最大似然法或MAP法来估计模型参数$\boldsymbol{\theta}$。我们发现MLE可能会导致严重的过拟合，特别是在小数据的情况下。MAP解决了这一问题，它将一个先验的在参数上起到正则器的作用。贝叶斯线性

*贝叶斯线性回归将*参数先验的概念推进一步，甚至不试图计算参数的点估计，而是在进行预测时考虑参数的全部后验分布。这意味着我们不拟合任何参数，而是计算所有合

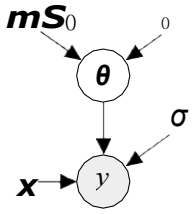
理的参数设置的平均值（根据后验）。

9.3.1 模型

In Bayesian linear regression, we consider the model

$$\begin{aligned} prior p(\boldsymbol{\theta}) &= \mathsf{N}\left(\boldsymbol{m}_0, \boldsymbol{S}_0\right), \\ likelihood p(y \mid \boldsymbol{x}, \boldsymbol{\theta}) &= \mathsf{N}\left(y \mid \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\theta}, \sigma^2\right)。 \end{aligned} \tag{9.35}$$

图为贝叶斯线性回归的图形9.8模型。



我们现在明确地将高斯先验$p(\boldsymbol{\theta})=\boldsymbol{m}_0, \boldsymbol{S}_0$放在$\boldsymbol{\theta}$上，这将参数向量变成了一个随机变量。这使我们能够写出图9.8中相应的图形模型，其中我们明确了$\boldsymbol{\theta}$上高斯先验的参数。完整的proba-bilistic模型，即观察到的和未观察到的ran-dom变量y和$\boldsymbol{\theta}$的联合分布，分别为

$$p(y, \boldsymbol{\theta} \mid \boldsymbol{x}) = p(y \mid \boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) 。 \tag{9.36}$$

9.3.2 先前的预测

在实践中，我们通常对参数值$\boldsymbol{\theta}$本身不那么感兴趣。相反，我们的关注点往往在于我们用这些参数值进行的预测。在贝叶斯设置中，当我们进行预测时，我们采取参数分布，并在所有合理的参数设置上取平均值。更具体地说，为了对输入的$\boldsymbol{x}$进行预测$_*$，我们将$\boldsymbol{\theta}$积分出来，得到

$$p(y_* \mid \boldsymbol{x}_*) = \mathrm{p}(y_* \mid \boldsymbol{x}_*, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = \mathsf{E}_{\boldsymbol{\theta}}[p(y_* \mid \boldsymbol{x}_*, \boldsymbol{\theta})] , \tag{9.37}$$

我们可以把它解释为根据先验分布$p(\boldsymbol{\theta})$，对所有可能的参数$\boldsymbol{\theta}$进行的$y_*\boldsymbol{x}_*, \boldsymbol{\theta}$的平均预测。请注意，使用先验分布的预测只需要我们指定输入$\boldsymbol{x}_*$，但不需要训练数据。

In our model (9.35), we chose a conjugate (Gaussian) prior on $\boldsymbol{\theta}$ so that the predictive distribution is Gaussian as well (and can be computed in closed form): With the prior distribution $p(\boldsymbol{\theta}) = \boldsymbol{m}_0, \boldsymbol{S}_0$ we obtain the predictive distribution as

$$p(y_* \mid \boldsymbol{x}_*) = \mathsf{N}\left(\boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}_*)\boldsymbol{m}_0, \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}_*)\boldsymbol{S}_0\boldsymbol{\phi}(\boldsymbol{x}_*) + \sigma^2\right), \tag{9.38}$$

其中我们利用了（i）由于共轭性（见第6.6节）和高斯的边际化属性（见第6.5节），预测是高斯的，（ii）高斯噪声是独立的，所以

$$\mathsf{V}[y_*] = \mathsf{V}_{\boldsymbol{\theta}}[\boldsymbol{\phi}(^{\mathrm{T}}\boldsymbol{x}_*)\boldsymbol{\theta}] + \mathsf{V}_E[E] , \tag{9.39}$$

和(iii)y_*是$\boldsymbol{\theta}$的线性变换，这样我们就可以通过使用(6.50)和(6.51)分别应用分析计算预测的平均值和协方差的规则。在(9.38)中，预测方差中的项$\boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}_*)\boldsymbol{S}_0\boldsymbol{\phi}(\boldsymbol{x}_*)$明确地说明了与不确定性有关的

与参数$\boldsymbol{\theta}$，而σ^2是由于测量噪声造成的不确定性贡献。

如果我们对预测无噪声的函数值$f(\boldsymbol{x}_*)$=感兴趣，那么
$\boldsymbol{\varphi}^{\mathrm{T}}(\boldsymbol{x}_*)\boldsymbol{\theta}$ instead of the noise-corrupted targets y_* we obtain

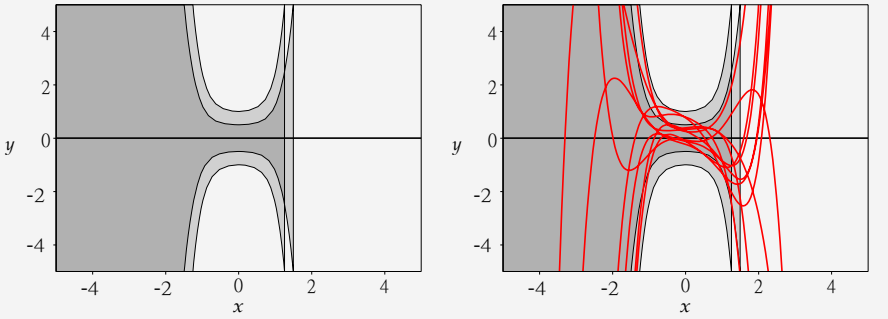
$$p(f(\boldsymbol{x}_*)) = \mathrm{N}\ \boldsymbol{\varphi}^{\mathrm{T}}(\boldsymbol{x}_*)\boldsymbol{m}_0,\ \boldsymbol{\varphi}^{\mathrm{T}}(\boldsymbol{x}_*)\boldsymbol{S}_0\boldsymbol{\varphi}(\boldsymbol{x}_*)\ , \quad (9.40)$$

它与(9.38)的区别在于预测方差中省略了噪声方差σ^2。

备注（函数上的分布）。由于我们可以将参数的分布表示为分布$p(\boldsymbol{\theta})$使用一组样本$\boldsymbol{\theta}_i$*的参数分布*$p(\boldsymbol{\theta})$，每一个样本$\boldsymbol{\theta}_i$都会产生一个函数$f_i(\cdot)=\boldsymbol{\theta}_i^{T}\boldsymbol{\varphi}(\cdot)$，由此可见，参数分布$p(\boldsymbol{\theta})$会在函数上诱导一个分布$p(f(\cdot))$。这里我们用符号$(\cdot)$来明确表示函数关系。 ◆

诱导出一个关于函数的分布。

例子（9.7先验大于函数）。



(a)函数的先验分布。(b)来自先验分布的样本

Figure 9.9 Prior over functions. (a)Distribution over functions represented by the mean function (black line) and the marginal uncertainties (shaded), representing the 67% and 95% confidence bounds, respectively; (b)samples from the prior over 职能，它们是由参数先验的样本所引起的。

职能。

我们选择一个参数先验$p(\boldsymbol{\theta})=\mathrm{N}(\boldsymbol{0},\ \frac{1}{2}\boldsymbol{I})$，图9.9可视化了函数的诱导先验分布（阴影区域：暗色的灰色。67%的置信度；浅灰色。95%的置信度）由该参数先验引起的，包括该先验的一些函数样本。

让我们考虑一个贝叶斯线性回归问题，其度数为多项式的5

$\boldsymbol{\theta}_i \sim p(\boldsymbol{\theta})$然后计算对应的函数$f_i(\cdot)=\boldsymbol{\theta}_i^{\mathrm{T}}\boldsymbol{\varphi}(\cdot)$。我们最终得到的函数样本是通过首先对一个参数向量进行采样而得到的

阳离子$x_*\in[-5,\ 5]$，我们对其应用特征函数$\varphi(\cdot)$。 我们对其应用特征图数的不确定性（用阴影区域表示）完全是由于参数的不确定性造成的

。9.9完全是由于参数的不确定性，因为我们考虑的是无噪声的预测分布(9.40).

到目前为止，我们研究了使用参数先验$p(\boldsymbol{\theta})$计算预测。然而，当我们有一个参数后验（给定一些训练数据，），预测和推理的原则与(9.37)--我们只需要用后验代替先验$p(\boldsymbol{\theta})$。

XY

$p(\boldsymbol{\theta} \mid X, Y)$。在下文中，我们将详细推导出后验分布，然后再利用它进行预测。

9.3.3 后期分布

给定一个输入$\boldsymbol{x}_n \in \mathbb{R}^D$和相应的观测值$y_n \in \mathbb{R}$的训练集，$n = 1, \ldots, N$，我们使用贝叶斯定理计算参数的后验，即

$$p(\boldsymbol{\theta} \mid X, Y) = \frac{p(Y \mid X, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(Y \mid X)}, \tag{9.41}$$

其中是对训练输入的集合和对应的训练目标的集合。此外，$p(Y \mid X, \boldsymbol{\theta})$是似然，$p(\boldsymbol{\theta})$是参数先验，而

$$p(Y \mid X) = \int p(Y \mid X, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = \mathbb{E}_{\boldsymbol{\theta}}[p(Y \mid X, \boldsymbol{\theta})] \tag{9.42}$$

边际似然/证据，它与参数无关。

边际可能性
证据
边际似然是参数先验下的期望似然。

$\boldsymbol{\theta}$，并确保后验是归一化的，也就是说，它整合到1。我们可以把边际似然看作是所有可能的参数设置（关于先验分布$p(\boldsymbol{\theta})$）的平均似然。

Theorem 9.1 (Parameter Posterior)**.** *In our model* (9.35)*, the parameter posterior* (9.41) *can be computed in closed form as*

$$p(\boldsymbol{\theta} \mid X, Y) = \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{m}_N, \boldsymbol{S}_N), \tag{9.43a}$$

$$\boldsymbol{S}_N = (\boldsymbol{S}_0^{-1} + \sigma^{-2} \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}, \tag{9.43b}$$

$$\boldsymbol{m}_N = \boldsymbol{S}_N (\boldsymbol{S}_0^{-1} \boldsymbol{m}_0 + \sigma^{-2} \boldsymbol{\Phi}^\top \boldsymbol{y}), \tag{9.43c}$$

其中，下标N表示训练集的大小。

证明 贝叶斯定理告诉我们，后验$p(\boldsymbol{\theta} \mid X, Y)$与似然$p(Y \mid X, \boldsymbol{\theta})$和先验$p(\boldsymbol{\theta})$的乘积成正比。

$$\text{后验} \quad p(\boldsymbol{\theta} \mid X, Y) = \frac{p(Y \mid X, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(Y \mid X)} \tag{9.44a}$$

$$\text{Likelihood} \quad p(Y \mid X, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{\Phi}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}) \tag{9.44b}$$

$$\text{Prior} \quad p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{m}_0, \boldsymbol{S}_0) \tag{9.44c}$$

我们可以将问题转化为对数空间，并通过完成平方求解后验的平均值和协方差，而不是看先验和似然的乘积。

对数优先权和对数可能性之和为

$$\log \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{\Phi}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}) + \log \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{m}_0, \boldsymbol{S}_0) \tag{9.45a}$$

$$= -\frac{1}{2}(\boldsymbol{y}^{-2} - \boldsymbol{\Phi\theta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\Phi\theta}) + (\boldsymbol{\theta} - \boldsymbol{m}_0)^{\mathrm{T}}\boldsymbol{S}_0^{-1}(\boldsymbol{\theta} - \boldsymbol{m}_0) + \text{const} \tag{9.45b}$$

其中常数包含与$\boldsymbol{\theta}$无关的项。我们将在下文中忽略常数。现在我们对（9.45b）进行因式分解，得到的结果是

$$-\frac{1}{2}\left(\sigma^{-2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}-\sigma\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Phi}2^{-2}\boldsymbol{\theta}+\boldsymbol{\theta}\sigma\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}^{-2}\boldsymbol{\theta}+\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{\theta}\right)$$

$$2-\boldsymbol{m}\boldsymbol{S}^{\mathrm{T}-1}\boldsymbol{\theta}+\boldsymbol{m}\boldsymbol{S}\boldsymbol{m}^{\mathrm{T}-1} \tag{9.46a}$$

$$=-\frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\sigma^{-2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}+\boldsymbol{S}^{-1}\boldsymbol{\theta}-2(\sigma\boldsymbol{\Phi}^{-2}\boldsymbol{y}+\boldsymbol{S}^{-1}\boldsymbol{m})^{\mathrm{T}}\boldsymbol{\theta}+\text{const}, \tag{9.46b}$$

其中常数包含（9.46a）中的黑色项，它们与$\boldsymbol{\theta}$无关。橙色项是在$\boldsymbol{\theta}$中呈线性的项，而蓝色项是在$\boldsymbol{\theta}$中呈二次的项。

$$p(\boldsymbol{\theta}\mid\mathrm{X},\mathrm{Y})=\exp(\log p(\boldsymbol{\theta}\mid\mathrm{X},\mathrm{Y}))\propto\exp(\log p(\mathrm{Y}\mid\mathrm{X},\boldsymbol{\theta})+\log p(\boldsymbol{\theta})) \tag{9.47a}$$

$$\propto\exp\left(-\frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\sigma^{-2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}+\boldsymbol{S}^{-1}\right)\boldsymbol{\theta}-2\left(\sigma\boldsymbol{\Phi}^{-2}\boldsymbol{y}+\boldsymbol{S}^{-1}\boldsymbol{m}\right)^{\mathrm{T}}\boldsymbol{\theta}. \tag{9.47b}$$

其中我们在最后一个表达式中使用了（9.46b）。

剩下的任务就是把这个（未归一化的）高斯变成与$\boldsymbol{\theta}$ $\boldsymbol{m}_N$，$\boldsymbol{S}$成正比的形式$_N$，也就是说，我们需要确定平均值$\boldsymbol{m}_N$和协方差矩阵$\boldsymbol{S}_N$。为了做到这一点，我们使用以下概念

的*完成方程*。所需的对数后验是

完成方格的

$$\log\mathrm{N}(\boldsymbol{\theta}\mid\boldsymbol{m}_N,\boldsymbol{S}_N)=-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{m}_N)^{\mathrm{T}}\boldsymbol{S}_N^{-1}(\boldsymbol{\theta}-\boldsymbol{m}_N)+\text{const} \tag{9.48a}$$

$$=-\frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{S}_N^{-1}\boldsymbol{\theta}-2\boldsymbol{m}_N^{\mathrm{T}}\boldsymbol{S}_N^{-1}\boldsymbol{\theta}+\boldsymbol{m}_N^{\mathrm{T}-1}\boldsymbol{S}_N\boldsymbol{m}_N. \tag{9.48b}$$

在这里，我们把二次形式$(\boldsymbol{\theta}-\boldsymbol{m}_N)^{\mathrm{T}}\boldsymbol{S}_N^{-1}(\boldsymbol{\theta}-\boldsymbol{m}_N)$分解为一个$\boldsymbol{\theta}$的二次项（蓝色），$\boldsymbol{\theta}$的线性项（橙色），以及常数项（黑色）。这使得我们现在可以$_N$通过以下方式找到$\boldsymbol{S}_N$和$\boldsymbol{m}_N$

由于$p(\boldsymbol{\theta}\mid X,Y)=\mathrm{N}(\boldsymbol{m}_N,\boldsymbol{S}_N)$，$_N$认为$\boldsymbol{\theta}_{\mathrm{MAP}}=m_N$。

与（9.46b）和（9.48b）中的彩色表达式相匹配，从而得到

$$\boldsymbol{S}_N^{-1}=\boldsymbol{\Phi}\sigma^{\mathrm{T}-2}\boldsymbol{I}\boldsymbol{\Phi}+\boldsymbol{S}_0^{-1} \tag{9.49a}$$

$$\Leftrightarrow\boldsymbol{S}_N=(\sigma^{-2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}+\boldsymbol{S}_0^{-1})^{-1} \tag{9.49b}$$

和

$$\boldsymbol{m}_N^{\mathrm{T}}\boldsymbol{S}_N^{-1}=(\sigma\boldsymbol{\Phi}^{-2}\boldsymbol{y}+\boldsymbol{S}_0^{-1}\boldsymbol{m}_0)^{\mathrm{T}} \tag{9.50a}$$

$$\Leftrightarrow\boldsymbol{m}_N=\boldsymbol{S}_N(\sigma\boldsymbol{\Phi}^{-2}\boldsymbol{y}+\boldsymbol{S}_0^{-1}\boldsymbol{m}_0)。 \tag{9.50b}$$

□

备注（补全方程的一般方法）。如果我们得到了一个方程

$$\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} - 2\mathbf{a}^{\mathrm{T}}\mathbf{x} + \mathrm{const}_1, \tag{9.51}$$

其中$\mathbf{A}$是对称和正定的，我们希望将其转化为以下形式

$$(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}(\mathbf{x} - \boldsymbol{\mu}) + \mathrm{const}_2, \tag{9.52}$$

我们可以通过设置来做到这一点

$$\boldsymbol{\Sigma} := \mathbf{A}, \tag{9.53}$$

$$\boldsymbol{\mu} := \boldsymbol{\Sigma}^{-1}\mathbf{a} \tag{9.54}$$

和$\mathrm{const}_2 = \mathrm{const}_1 - \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\mu}$。 ◆

我们可以看到(9.47b)中指数内部的项的形式是(9.51)，其中有

$$\mathbf{A} := \sigma^{-2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \mathbf{S}_0^{-1}, \tag{9.55}$$

$$\mathbf{a} := \sigma^{-2}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{y} + \mathbf{S}_0^{-1}\mathbf{m}_0. \tag{9.56}$$

由于(9.46a)这样的方程中，A、$\mathbf{a}$可能难以确定，因此将这些方程转化为(9.51)的形式，将二次项、线性项和常数解耦，从而简化了寻找所需的解决方案。

9.3.4 后期预测

在(9.37)中，我们使用参数先验$p(\boldsymbol{\theta})$计算了测试输入$\mathbf{x}_*$时y_*的预测分布。原则上，鉴于在我们的共轭模型中，先验和后验都是高斯（具有不同的参数），用参数后验$p(\boldsymbol{\theta}|X,Y)$进行预测没有根本的不同。因此，按照章节中的相同推理，9.3.2,我们可以得到（后验）预测分布

$$p(y_* \mid X, Y, \mathbf{x}_*) = \int p(y_* \mid \mathbf{x}_*, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid X, Y) \mathrm{d}\boldsymbol{\theta} \tag{9.57a}$$

。

$$= \int \mathcal{N}\left(y_* \mid \boldsymbol{\phi}(\mathbf{x}_*)^{\mathrm{T}}\boldsymbol{\theta}, \sigma^2\right) \mathcal{N}\left(\boldsymbol{\theta} \mid \mathbf{m}_N, \mathbf{S}_N\right) \mathrm{d}\boldsymbol{\theta} \tag{9.57b}$$

$$= \mathcal{N}\left(y_* \mid \boldsymbol{\phi}(\mathbf{x}_*)^{\mathrm{T}}\mathbf{m}_N, \boldsymbol{\phi}(\mathbf{x}_*)^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_*) + \sigma^2\right)。 \tag{9.57c}$$

$\mathbb{E}[y_* \mid X, Y, \mathbf{x}_*] = \boldsymbol{\phi}(\mathbf{x}_*)^{\mathrm{T}}\mathbf{m}_N = \boldsymbol{\phi}(\mathbf{x}_*)^{\mathrm{T}}\boldsymbol{\theta}_{\mathrm{MAP}}$.

术语$\boldsymbol{\phi}(\mathbf{x}_*)^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_*)$反映了与之相关的后验不确定性。

注意，$\mathbf{S}_N$通过$\boldsymbol{\Phi}$取决于训练输入；见（9.43b）。预测均值$\boldsymbol{\phi}(\mathbf{x}_*)^{\mathrm{T}}\mathbf{m}_N$与用MAP估计值$\boldsymbol{\theta}_{\mathrm{MAP}}$做出的预测相吻合。

备注（边际似然和后验预测分布）。通过替换(9.57a)中的积分，预测分布可以等效地写成期望值$\mathbb{E}_{\boldsymbol{\theta}|p_{X,Y}}[y_* \mid \boldsymbol{x}_*, \boldsymbol{\theta}]$，这里的期望值是相对于参数后验$p(\boldsymbol{\theta} \mid X, Y)$而言的。

用这种方式来写后验预测分布，突出了与边际似然的密切相似性(9.42). 边际似然和后验预测分布的关键区别在于：（i）边际似然可以被认为是预测训练目标$\boldsymbol{y}$，而不是测试目标y_*；（ii）边际似然是关于参数先验而不是参数后验的。

原告。 ◆

备注（无噪声函数值的平均值和方差）。在许多情况下，我们对（有噪声的）观测值y_*的预测分布$p(y_* \mid X, Y, \boldsymbol{x}_*)$不感兴趣，而是希望得到（无噪声的）函数值$f(\boldsymbol{x}_*) = \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}_*)\boldsymbol{\theta}$的分布。我们通过利用均值的特性来确定相应的矩。

差异，从而得出了

$$\begin{aligned}\mathbb{E}[f(\boldsymbol{x}_*) \mid X, Y] &= \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\phi}(^{\mathrm{T}}\boldsymbol{x}_*)\boldsymbol{\theta} \mid X, Y] = \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x})_*\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta} \mid X, Y] \\ &= \boldsymbol{\phi}(^{\mathrm{T}}\boldsymbol{x}_*)\boldsymbol{m}_N = {}^{\mathrm{T}}_{m_N}\boldsymbol{\phi}(\boldsymbol{x}_*) \text{。}\end{aligned} \tag{9.58}$$

$$\begin{aligned}\mathbb{V}[\boldsymbol{\theta} f(\boldsymbol{x}_*) \mid X, Y] &= \mathbb{V}_{\boldsymbol{\theta}}[\boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}_*)\boldsymbol{\theta} \mid X, Y] \\ &= \boldsymbol{\phi}(^{\mathrm{T}}\boldsymbol{x})_*\mathbb{V}_{\boldsymbol{\theta}}[\boldsymbol{\theta} \mid X, Y]\boldsymbol{\phi}(\boldsymbol{x}_*) \\ &= \boldsymbol{\phi}(^{\mathrm{T}}\boldsymbol{x}_*)\boldsymbol{S}_N\,\boldsymbol{\phi}(\boldsymbol{x}_*) \text{。}\end{aligned} \tag{9.59}$$

我们看到，预测均值与噪声观测的预测均值相同，因为噪声的均值()为 ，而预测方差只相差σ^2，也就是测量噪声的方差。当我们预测有噪声的函数值时，我们需要把σ^2作为不确定性的来源，但对于无噪声的预测，不需要这个项。这里。

唯一剩下的不确定性来自于参数后验。

归纳出参数诱导出一个关于函数的分布。

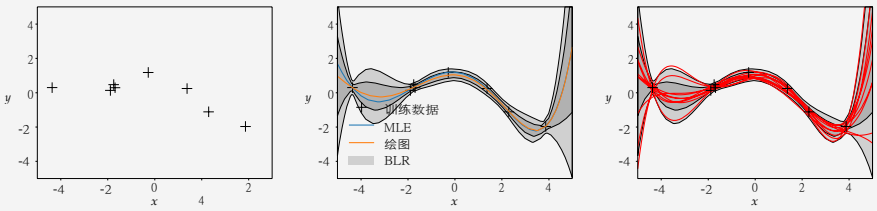
备注（函数上的分布）。我们对参数$\boldsymbol{\theta}$*进行*积分的事实诱导了一个函数分布。如果我们从参数后验中抽出$\boldsymbol{\theta}_i \sim p(\boldsymbol{\theta} \mid X, Y)$，我们得到一个单一的函数re
化 $\boldsymbol{\theta}_i^{\mathrm{T}}\boldsymbol{\phi}(\cdot)$。 *平均值函数*， 即所有预期函数平均值函数的集合
值$\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{f}(\cdot) \mid \boldsymbol{\theta}, X, Y]$， 这个分布的函数是${}^{\mathrm{T}}_{m_N}\boldsymbol{\phi}(\cdot)$。 边际）方差，即函数$\boldsymbol{f}(\cdot)$ 的方差，由$\boldsymbol{\phi}^{\mathrm{T}}(\cdot)\ \boldsymbol{S}_N\boldsymbol{\phi}(\cdot)$ 给出。 ◆

例子（9.8函数的后验）

让我们重新审视贝叶斯线性回归问题，其度数为多项式的5 .我们选择一个参数先验$p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \frac{1}{4}\boldsymbol{I})$。图9.9可视化由参数先验引起的函数的先验和
从这个先验中得到的样本函数。

图为9.10显示了我们通过贝叶斯线性回归得到的函数的后验。训练数据集显示在面板（a）；面板（b）显示了函数的后验分布，包括我们通过最大似然和MAP估计获得的函数。我们使用MAP估计得到的函数也对应于贝叶斯线性回归环境中的后验平均函数。面板(c)显示了在该函数的后验分布下的一些可信的函数实现（样本）。

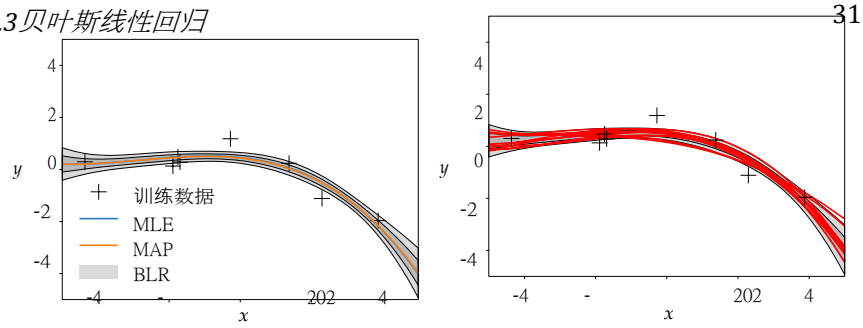
图 9.10贝叶斯线性回归和函数的后验。(a)训练数据；(b)函数的后验分布；(c)函数的后验样本。



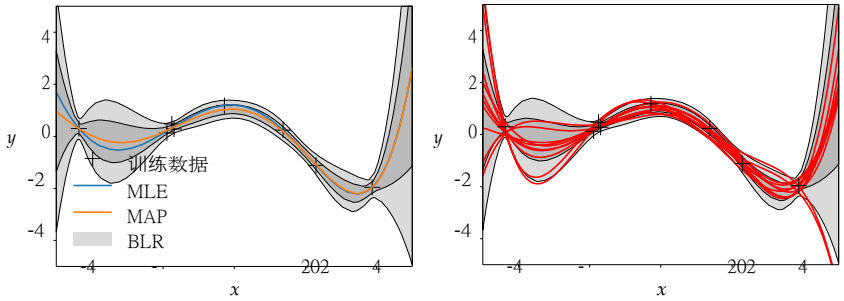
(a) 训练数据。

(b)后置函数代表(c)由边际不确定的后置函数的样本，这些样本是由67%和95%的预测参数后置的样本组成的(阴影)。信心界限，最大似然估计（MLE）和MAP估计（MAP），后者与后验平均函数相同。

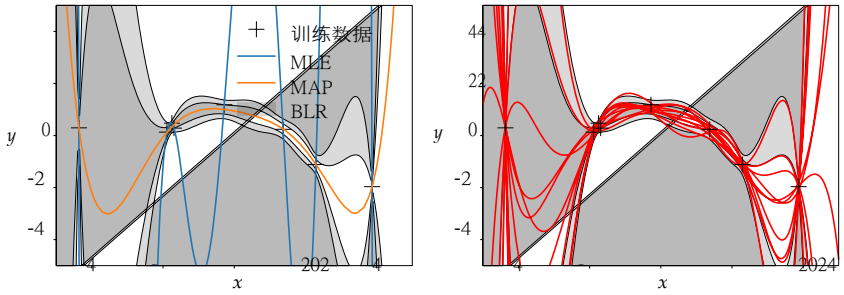
"机器学习的数学"草案（2022-01-11）。反馈：https://mml-book.com。



(a) 度数为M=的多项式的后验分布(3左)和来自正-的样本。职能上的劣势（右）。



(b) 度数为M=的多项式的后验分布（5左）和函数上的后验样本（右）。



(c) 度数为M=的多项式的后验分布（7左）和来自函数的正后验样本（右）。

图 贝叶斯**9.11**线性回归。左侧面板。阴影区域表示67%（深灰色）的和95%（浅灰色）的预测置信界线。贝叶斯线性回归模型的平均值恰好与MAP估计。预测的不确定性是指噪声项和后验参数不确定性之和，这取决于测试输入的位置。右图：来自后验分布的抽样函数。

后验的不确定性是巨大的。当我们在决策系统中使用这些预测时，这些信息可能是至关重要的，因为错误的决定会产生重大的后果（例如，在强化学习或机器人技术中）。

9.3.5 计算边际可能性

在第8.6.2节中，我们强调了边际似然对贝叶斯模型选择的重要性。在下文中，我们将计算参数为共轭高斯先验的贝叶斯线性回归的边际似然，也就是说，正是我们在本章中所讨论的设定。

只是为了回顾一下，我们考虑以下生成过程。

$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{m}_0, \boldsymbol{S}_0) \tag{9.60a}$$

$$y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{x}_n^\top \boldsymbol{\theta}, \sigma^2), \tag{9.60b}$$

$n = 1, \ldots, N$.边际似然由以下公式给出

边际似然可以是被解释为预期可能性在先验的情况下，即。$\mathbb{E}_{\boldsymbol{\theta}}[p(Y \mid X, \boldsymbol{\theta})]$。

$$p(Y \mid X) = \int p(Y \mid X, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \tag{9.61a}$$

$$= \int \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}) \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{m}_0, \boldsymbol{S}_0) \mathrm{d}\boldsymbol{\theta}, \tag{9.61b}$$

我们分两步来计算边际似然。首先，我们表明边际似然是高斯的（作为$\boldsymbol{y}$的分布）；其次，我们计算这个高斯的平均值和共同方差。

1. 边际似然是高斯的。从第6.5.2节，我们知道
   (i) 两个高斯随机变量的乘积是一个（未归一化的）高斯分布，以及（ii）一个高斯随机变量的线性变换是高斯分布。在（9.61b）中，我们需要一个线性变换来使$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I})$变成$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$的形式，对于某些$\boldsymbol{\mu}, \boldsymbol{\Sigma}$。一旦这样做了，就可以用闭合形式解决这个积分。其结果是两个高斯数的乘积的归一化常数。归一化常数本身具有高斯形状；见（6.76）。
2. 均值和协方差。我们通过利用随机变量仿射变换的均值和协方差的标准结果来计算边际似然的均值和协方差；见第6.4.4节。边际似然的平均值被计算为

$$\mathbb{E}[Y \mid X] = \mathbb{E}_{\boldsymbol{\theta}, \boldsymbol{\varepsilon}}[\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}] = X\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}] = \boldsymbol{X}\boldsymbol{m}_0 。 \tag{9.62}$$

请注意，$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$是一个i.i.d.随机变量的向量。协方差矩阵给定为

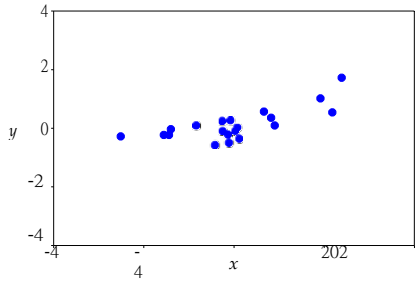
$$\mathrm{Cov}[Y \mid X] = \mathrm{Cov}_{\boldsymbol{\theta}, \boldsymbol{\varepsilon}}[\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}] = \mathrm{Cov}_{\boldsymbol{\theta}}[\boldsymbol{X}\boldsymbol{\theta}] + \sigma^2 \boldsymbol{I} \tag{9.63a}$$

$$= \boldsymbol{X}\,\mathrm{Cov}[\boldsymbol{\theta}]\boldsymbol{X}^\top + \sigma^2 \boldsymbol{I} = \boldsymbol{X}\boldsymbol{S}_0\boldsymbol{X}^\top + \sigma^2 \boldsymbol{I} 。 \tag{9.63b}$$

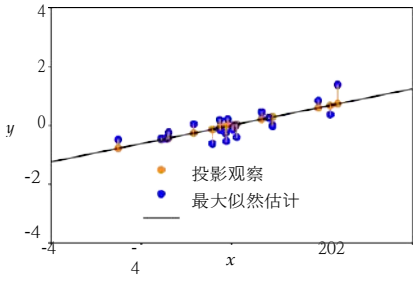
因此，边际可能性为

$$p(Y \mid X) = (2\pi)^{-\frac{N}{2}} \det(\boldsymbol{X}\boldsymbol{S}_0\boldsymbol{X}^\top + \sigma^2 \boldsymbol{I})^{-\frac{1}{2}} \tag{9.64a}$$

$$\exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{m}_0)^\top (\boldsymbol{X}\boldsymbol{S}_0\boldsymbol{X}^\top + \sigma^2 \boldsymbol{I})^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{m}_0)\right)$$



(a) 回归数据集由输入位置x_n的函数值$f(x_n)$的嘈杂ob-servations y_n（蓝色）组成。



(b) 橙色的点是噪声观测值（蓝点）对线$\theta_{ML}x$的投影。线性回归问题的最大似然解找到了一个子空间（线），观测值的整体投影误差（橙线）被最小化了。

图：最小二乘法的几何**9.12**解释。(a)数据集；(b)最大似然解被解释为一个投影。

$$= \mathsf{N}\left(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{m}_0,\ \boldsymbol{X}\boldsymbol{S}_0\boldsymbol{X}^{\mathrm{T}} + \sigma^2\boldsymbol{I}\right) . \tag{9.64b}$$

鉴于与后验预测分布的密切联系（见本节前面关于边际似然和后验预测分布的备注），边际似然的功能形式不应该太令人惊讶。

9.4 作为正交投影的最大似然法

在通过许多代数推导出最大似然和MAP估计之后，我们现在将提供最大似然估计的几何解释。让我们考虑一个简单的线性回归环境

$$y = x\theta + E, \quad E \sim \mathsf{N}\left(0, \sigma^2\right) , \tag{9.65}$$

其中，我们考虑线性函数$f : \mathbb{R} \to \mathbb{R}$通过原点（为了清楚起见，我们在这里省略了特征）。参数θ决定了直线的斜率。图9.12（a）显示了一个一维数据集。

有了训练数据集$(x_1, y_1), \ldots, (x_N, y_N)$，我们回顾一下本节的结果。9.2.1的结果，得到斜率参数的最大似然估计值为

$$\theta_{\mathrm{ML}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}}{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}} \in \mathbb{R} , \tag{9.66}$$

其中，$\boldsymbol{X} = [x_1, \ldots, x_N]^{\mathrm{T}} \in \mathbb{R}^N$，$\boldsymbol{y} = [y_1, \ldots, y_N]^{\mathrm{T}} \in \mathbb{R}^N$。

这意味着对于训练输入$\boldsymbol{X}$，我们获得了最佳（最大可能性）重建的训练目标为

$$\boldsymbol{X}\theta_{\mathrm{ML}} = \boldsymbol{X}\frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}}{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}} = \frac{\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}}{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}\boldsymbol{y} , \tag{9.67}$$

也就是说，我们得到的是$\boldsymbol{y}$和$\boldsymbol{X}\theta$之间误差最小的近似值。

由于我们正在寻找$\boldsymbol{y}=\boldsymbol{X}\theta$的解，我们可以把线性回归看作是一个解决线性方程组的问题。因此，我们可以把我们在第二章和第三章中讨论的线性代数和解析几何的概念联系起来。特别是，仔细观察在(9.67)，我们看到最大似然估计器θ_{ML}在我们的前

线性回归可以被认为是一种解决线性方程组的方法。

从(9.65)有效地做了一个$\boldsymbol{y}$的正交投影到$\boldsymbol{X}$所跨越的一维子空间。回顾关于或-的结果。我们将$\boldsymbol{X}\boldsymbol{X}^{+}$确定为来自科3.8的投射。

$\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$

矩阵，θ_{ML}作为投射到一维的坐标

最大似然线性回归执行一个正交的投影。

$\boldsymbol{X}$和$\boldsymbol{X}\theta_{\mathrm{ML}}$所横跨的R的$N$子空间是X的正交投影。$\boldsymbol{y}$到这个子空间上。

因此，最大似然解也提供了一个几何上的最优解，即在$\boldsymbol{X}$所跨越的子空间中找到与相应观测值$\boldsymbol{y}$ "最接近 "的向量，其中 "最接近 "意味着函数值y_n与$x_n\theta$的最小（平方）距离。这是通过正交投影实现的。图9.12(b)显示了噪声观测值在子空间上的投影，它使原始数据集和其投影之间的平方距离最小（注意x坐标是固定的），这相当于最大似然解。

In the general linear regression case where

$$y = \boldsymbol{\varphi}(^{\mathrm{T}}\boldsymbol{x})\boldsymbol{\theta} + \quad E, E \sim \mathsf{N}\ 0,\ \sigma(^{2} \qquad 9.68)$$

与矢量值特征$\boldsymbol{\varphi}(\boldsymbol{x})$ $\in \mathsf{R}^{K}$，我们可以再次解释最大似然结果

$$\boldsymbol{y} \approx \boldsymbol{\Phi}\boldsymbol{\theta}_{\mathrm{ML}}, \qquad (9.69)$$

$$\boldsymbol{\theta}_{\mathrm{ML}} = (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})\ \boldsymbol{\Phi}^{-1\mathrm{T}}\boldsymbol{y} \qquad (9.70)。$$

作为对R的$N$$K$维子空间的投影，它被特征矩阵$\boldsymbol{\Phi}$的列所覆盖；见第3.8.2节。

如果我们用来构建特征矩阵$\boldsymbol{\Phi}_k$的特征函数是正交的（见第3.7节），我们会得到一个特殊的情况，即$\boldsymbol{\Phi}$的列形成一个正交基（见第3.5节），这样$\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}=\boldsymbol{I}$，这将导致投影

$$\boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathrm{T}})\boldsymbol{\Phi}^{-1\mathrm{T}}\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{y} = \sum_{k=1}^{K} \boldsymbol{\varphi}_k \boldsymbol{\varphi}_k^{\mathrm{T}}\ \boldsymbol{y} \qquad (9.71)$$

因此，最大似然投影只是$\boldsymbol{y}$在各个基向量$\boldsymbol{\varphi}_k$上的投影之和，也就是$\boldsymbol{\Phi}$的列。此外，由于基的正交性，不同特征之间的耦合已经消失了。信号处理中许多流行的基函数，如小波和傅里叶基，都是正交的基函数。

当基数不是正交的时候，我们可以通过使用Gram-Schmidt过程将一组线性独立的基函数转换为正交基数；见第3.8.3节和(Strang,2003)。

9.5进一步阅读

在这一章中，我们讨论了高斯似然和模型参数的共轭高斯预设的线性回归。这为闭合形式的贝叶斯推断提供了条件。然而，在某些应用中，我们可能希望选择不同的似然函数。例如，在

在二元*分类*环境中，我们只观察到两种可能的（分类）结果，而高斯似然在这种情况下是不合适的。相反，我们可以选择一个伯努利似然，它将返回预测标签的概率为（1或0）。我们可以参考Barber(2012)、Bishop(2006)和Murphy(2012)的书中对分类问题的深入介绍。一个非高斯可能性很重要的不同例子是计数数据。计数是非负整数，在这种情况下，二项式或泊松似然是比高斯式更好的选择。

分类

所有这些例子都属于*广义线性模型*的范畴，是一种柔性允许响应变量的线性回归的可行概括具有高斯分布以外的误差分布。GLMG广义线性通过允许线性模型通过一个平滑的、可反转的函数$\sigma(\ )$与观测值相关，从而使线性回归得到了概括，该函数可能是非线性的，因此$y = \sigma(f(\boldsymbol{x}))$，其中$f(\boldsymbol{x}) = \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x})$是线性回归模型的(9.13).因此，我们可以用函数构成来考虑广义线性模型$y = \sigma f$，其中f是线性回归模型，σ是激活函数。请注意，虽然我们谈论的是"广义线性模型"，但输出y并不是在*逻辑回归*中，我们选择logistic回归。*logistic sigmoid* $\sigma(f) = \frac{1}{1+\exp(-f)} \in [0, 1]$，这可以解释为logistic sigmoid观察到y=的1概率是伯努利随机变量$y \in \{0, 1\}$。函数$\sigma(\cdot)$被称为*传递函数*或*激活函数*，其的倒数被称为*典范链接函数*。从这个角度来看，广义线性模型是（深度）前馈神经网络的构建模块，这一点也很明显。如果我们考虑一个广义线性模型$\boldsymbol{y}=\sigma(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})$，其中$\boldsymbol{A}$是一个权重矩阵，$\boldsymbol{b}$是一个偏置向量，我们把这个广义线性模型确定为一个具有激活函数$\sigma()$的单层神经网络。现在我们可以通过以下方式递归地组成这些函数

广义线性的模型

模型是深度神经网络的构建模块。

转移函数

激活函数 典型链接函数

对于普通的线性回归来说，激活函数只是一个身份。

一个伟大的帖子，关于GLMs和深度网络之间的关系是可在https://tinyurl.com/glm-dnn。

$$\begin{aligned}\boldsymbol{x}_{k+1} &= \boldsymbol{f}_k(\boldsymbol{x}_k)\\ \boldsymbol{f}_k(\boldsymbol{x}_k) &= \sigma_k(\boldsymbol{A}_k\boldsymbol{x}_k + \boldsymbol{b}_k)\end{aligned} \tag{9.72}$$

对于$k = 0, \ldots, K-1$，其中$\boldsymbol{x}_0$是输入特征，$\boldsymbol{x}_K = \boldsymbol{y}$是观察到的输出，这样，$\boldsymbol{f}_{K-1} \circ \cdots \circ \boldsymbol{f}_0$是一个$K$层的深度神经网络。因此，这个深度神经网络的构建模块是

中定义的广义线性模型(9.72).神经网络（Bishop, 1995;Goodfellow等,2016）比线性回归模型更有表现力和灵活性。然而，最大似然参数估计是一个非凸的优化问题，在完全贝叶斯设置中参数的边际化在分析上是难以解决的。

高斯过程

我们简要地暗示了一个事实，即参数上的分布产生了回归函数的分布。*高斯过程*（Ras-mussen和Williams，2006）是回归模型，其中函数上的分布概念是核心。高斯过程不是把分布放在参数上，而是直接把分布放在函数空间上，而不通过参数 "绕道"。为此，高斯过程利用了*内核技巧*（

内核技巧

Scholkopf和Smola，2002），它允许我们计算两个函数值之间的内积 $k(x_i, x_j)$。高斯过程与贝叶斯线性回归和上口向量回归密切相关，但也可以解释为贝叶斯神经系统。

单一隐藏层的网络，单元数趋于无穷大（Neal，1996；Williams，1997）。关于高斯过程的出色介绍可以在MacKay（1998）和Rasmussen和Williams（2006）中找到。

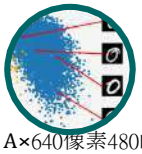
在本章的讨论中，我们重点讨论了高斯参数先验，因为它们允许在线性回归模型中进行闭式推理。然而，即使在具有高斯似然的回归环境中，我们也可以选择一个非高斯先验。考虑一种情况，即输入为 $\boldsymbol{x} \in \mathbb{R}^D$，我们的训练集很小，大小为 N D。在这种情况下，我们可以选择一个强制执行稀疏性的参数先验，也就是说，一个试图设置为尽可能多的参数（*变量选择*）。这种先验提供了一个比高斯先验更强的正则器，而高斯先验往往会导致一个在提高预测精度和模型的可解释性。拉普拉斯先验就是一个经常被用于此目的的例子。对参数采用拉普拉斯先验的线性再回归模型等同于L1正则化的线性回归（*LASSO*）（Tibshirani，1996）。拉普拉斯分布在零点处有一个尖锐的峰值（其一阶导数是不连续的），它的概率质量比高斯分布更接近零，高斯分布鼓励参数为0零。因此，非零参数与回归问题有关，这也是我们谈论 "变量选择 "的原因。

变量选择

LASSO

10

用主成分分析降低维度

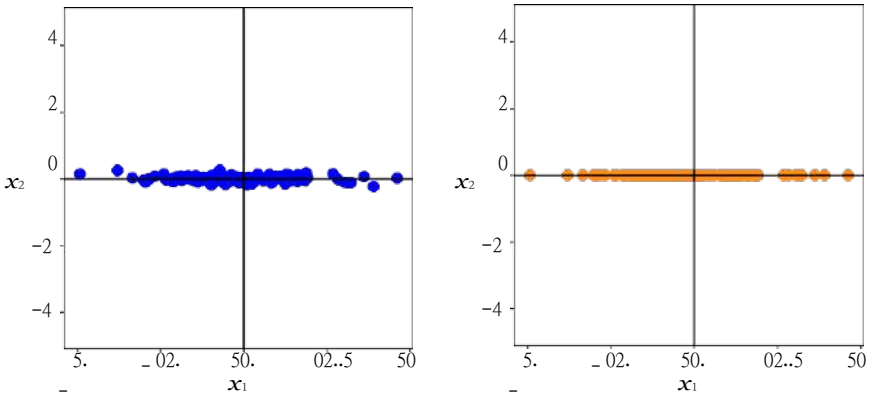


A×640像素480的情况彩色图像是一个数据点，在百万维空间，每个像素都对三个维度做出反应，每个颜色通道（红、绿、蓝）都有一个。

直接处理高维数据，如图像，会出现一些困难。它很难分析，解释很困难，可视化几乎是不可能的，而且（从实际角度来看）数据向量的存储可能很昂贵。然而，高维数据往往具有我们可以利用的特性。例如，高维数据往往是超完整的，也就是说，许多维度是多余的，可以通过其他维度的组合来表达。此外，高维数据中的维度往往是相关的，所以数据拥有内在的低维结构。降维利用了结构和相关性，使我们能够在不损失信息的前提下，用更紧凑的数据表达方式来工作。我们可以把降维看作是一种压缩技术，类似于jpeg或mp3，它们是图像和音乐的压缩算法。

在本章中，我们将讨论*主成分分析*（*PCA*），这是一种线性*降维*的算法。PCA由Pearson(1901)和Hotelling(1933)提出，已经存在了100多年，现在仍然是数据压缩和数据可视化的最常用技术之一。它也被用于识别简单模式、潜在因素和高维数据的结构。在

主成分分析。
分析
PCA
维度
减少



(a) 数据集的x_1和x_2坐标。

(b) 压缩的数据集，其中只有x_1的协整是相关的。

图 10.1 说明：维度减少。(a)该原始数据集差异不大沿着x_2方向。(b)该(a)中的数据可以单独用x_1坐标表示，几乎没有损失。

卡胡宁-罗夫变换

在信号处理界，PCA也被称为*Karhunen-Loève变换*。在本章中，我们根据对基和基的变化（第2.6.1和2.7.2节）、投影（第3.8节）、特征值（第4.2节）、高斯分布（第6.5节）和约束优化（第7.2节）的理解，从第一原理推导出PCA。

降维通常是利用高维数据（如图像）的一个特性，即它通常位于一个低维的子空间上。图10.1给出了一个二维的说明性例子。尽管图10.1(a)中的数据并不完全位于一条线上，但数据在x_2方向上的变化不大，因此我们可以把它当作一条线来表达--几乎没有损失；见图10.1(b)。为了描述数据，在
图10.1(b)，只需要x_1坐标，数据位于$\mathbb{R}^2$的一维子空间中。

10.1问题设置

在PCA中，我们感兴趣的是找到数据点$\boldsymbol{x}_n$的投影$\tilde{\boldsymbol{x}}_n$，n这些投影与原始数据点尽可能相似，但其内在维度明显降低。图10.1给出了一个关于这种情况的说明。

更具体地说，我们考虑一个独立的数据集 $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$，$\boldsymbol{x}_n \in \mathbb{R}^D$，其平均值$\mathbf{0}$拥有*数据协方差矩阵*（6.42）。

数据协方差矩阵

$$\boldsymbol{S} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top. \tag{10.1}$$

此外，我们假设存在一个低维的压缩表达（代码）。

$$\boldsymbol{z}_n = \boldsymbol{B}^\top \boldsymbol{x}_n \in \mathbb{R}^M \tag{10.2}$$

的$\boldsymbol{x}_n$，其中我们定义了投影矩阵

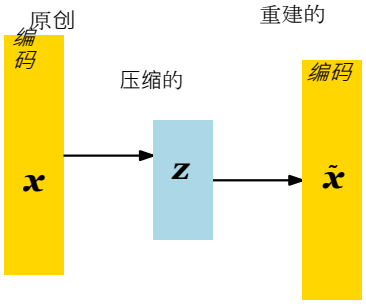
$$\boldsymbol{B} := [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M] \in \mathbb{R}^{D \times M}. \tag{10.3}$$

我们假设$\boldsymbol{B}$的列是正交的（定义3.7），所以
$\boldsymbol{b}_i^\top \boldsymbol{b}_j = 0$ if and only if $i = j$ and $\boldsymbol{b}_i^\top \boldsymbol{b}_i = 1$. We seek an M-dimensional 子空间$U \subseteq$ *种族投影*，$\dim(U) = M < D$，我们将数据投射到该空间上。我们用$\tilde{\boldsymbol{x}}_n \in U$表示投影数据，用$\boldsymbol{z}_n$表示它们的坐标（相对于$U$的基向量$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M$）。我们的目的是找到投影$\tilde{\boldsymbol{x}}_n \in \mathbb{R}^D$（或者等同于代码$\boldsymbol{z}_n$和基向量$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M$）

栏目的$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M$形成一个$M$维子空间的基础，在这个子空间中，投影数据$\tilde{\boldsymbol{x}}_n = \boldsymbol{B}\boldsymbol{z}_n \in \mathbb{R}^D$是活动。

以使它们与原始数据$\boldsymbol{x}_n$尽可能相似，并使损失最小化。
由于压缩的原因。

例子（10.1坐标表示法/代码）
考虑$\mathbb{R}^2$的典范基础$\boldsymbol{e}_1=[1, 0]$，$\boldsymbol{e}_2=[0, 1]$。从



图为PCA的图形10.2说明。在PCA中，我们找到了一个压缩的版本 的原始数据。 $\boldsymbol{x}$ 压缩的数据可以是 ，重构为 ~ ，它生活在original data space, but has an intrinsic lower-dimensional representation than $\boldsymbol{x}$.

第二章，我们知道$\boldsymbol{x}\in\mathsf{R}^2$可以表示为一个线性组合---。这些基向量的划分，例如：。

$$\begin{matrix}5\\3\end{matrix} = 5\boldsymbol{e}_1+3\,\boldsymbol{e}_2 \tag{10.4}$$

然而，当我们考虑以下形式的向量时

$$\tilde{\boldsymbol{x}} = \begin{matrix}0\\z\end{matrix} \in \mathsf{R}^2, \quad z\in\mathsf{R} \tag{10.5}$$

它们总是可以写成$0\boldsymbol{e}_1+ z\boldsymbol{e}_2$。为了表示这些向量，只需记住/存储$\tilde{\boldsymbol{x}}$相对于$\boldsymbol{e}_2$向量的*坐标/代码*$z$。

更确切地说， $\tilde{\boldsymbol{x}}$矢量的集合（有标准的矢量加法和标量乘法）形成一个矢量子空间$\boldsymbol{U}$（见第2.4节），$\dim(\boldsymbol{U})$=因为$1\boldsymbol{U}$=span[$\boldsymbol{e}_2$]。

矢量空间的维度对应于其基向量的数量（见第2.6.1节）。

在这一节中10.2,我们将找到低维表征，以重新保留尽可能多的信息并使压缩损失最小化。第二节给出了PCA的另一种推导。10.3,其中我们将是指最大限度地减少重建误差的平方值$\boldsymbol{x}_n$ $\tilde{\boldsymbol{x}}_n$我是...原始数据$\boldsymbol{x}_n$和其投影$\tilde{\boldsymbol{x}}_n$之间的关系。

图中10.2说明了我们在PCA中考虑的设置，其中$\boldsymbol{z}$代表了压缩数据$\tilde{\boldsymbol{x}}$的低维表示，并扮演了瓶颈的角色，它控制了多少信息可以在$\boldsymbol{x}$和$\tilde{\boldsymbol{x}}$之间流动。 在PCA中， 我们考虑原始数据$\boldsymbol{x}$和其低维代码$\boldsymbol{z}$之间的线性关系，因此$\boldsymbol{z}=\boldsymbol{B}\boldsymbol{x}$ 和对于一个合适的矩阵$\boldsymbol{B}$， $\tilde{\boldsymbol{x}} = \boldsymbol{B}\boldsymbol{z}$。 基于思考的动机作为数据压缩技术的PCA，我们可以将图中的箭头解释为代表编码器和解码器的一对操作。10.2中的箭头解释为代表编码器和解码器的一对操作。由$\boldsymbol{B}$代表的线性映射可以被认为是一个解码器，它将低维代码$\mathsf{z}\in\mathsf{R}^M$映射回原始数据空间$\mathsf{R}^D$。同样，$\boldsymbol{B}$可以被认为是一个编码器，它将原始数据$\boldsymbol{x}$编

码为低维（压缩）代码$\boldsymbol{z}$。

在这一章中，我们将使用MNIST数字数据集作为一个重新设计的数据集。

图为MNIST数据集中的手写数字实例**10.3**。 http://yann.lecun.com/exdb/mnist/。

发生的例子，其中包含了 60，000手写数字0到9 。每个数字都是大小为2828的灰度图像，即包含784个像素，因此我们可以把这个数据集中的每个图像解释为一个向量$\boldsymbol{x}\in\mathbb{R}^{784}$， 这些数字的例子见图10.3.

10.2 最大差异视角

图10.1给出了一个例子，说明如何用一个坐标来表示一个二维数据集。在图10.1(b)中，我们选择了忽略数据的x_2坐标，因为它没有增加太多的内涵，所以压缩后的数据与图10.1(a)中的原始数据相似。我们可以选择忽略x_1坐标，但那样的话，压缩后的数据就与原始数据非常不一样了，数据中的许多信息就会丢失。

如果我们把数据中的信息含量解释为数据集的 "空间填充 "程度，那么我们可以通过观察数据的扩散来描述数据中所包含的信息。从第6.4.1节中，我们知道方差是数据分布的一个指标，我们可以推导出PCA是一种降维算法，它在数据的低维表示中最大化方差以保留尽可能多的信息。图10.4说明了这一点。

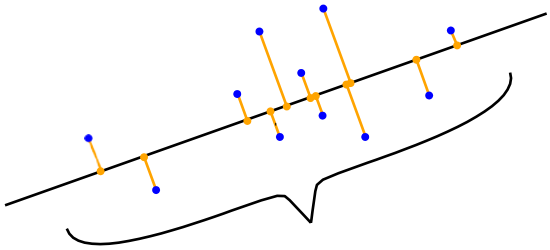
考虑到本节中讨论的设定，我们的目标是找到一个矩阵B（见（））。10.1,我们的目标是找到一个矩阵$\boldsymbol{B}$（见(10.3)），通过将数据投射到B的列$\boldsymbol{b}_1, \ldots, b$所跨越的子空间上，在压缩数据时尽可能多地保留信息。在数据压缩后保留最多的信息，相当于在数据中捕获最大的方差。
低维代码（Hotelling，1933）。

备注。(居中的数据)对于()中的数据协方差矩阵，我们假设居中的数据。10.1)中，我们假设了居中数据。我们可以在不损失基因的情况下做出这个假设。让我们假设$\boldsymbol{\mu}$是数据的平均值。利用我们在第6.4.4节中讨论的方差的特性，我们可以得到

$$\mathbb{V}_{\boldsymbol{z}}[\boldsymbol{z}] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}(\boldsymbol{x}-\boldsymbol{\mu})] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{B}\boldsymbol{x} - \boldsymbol{B}\boldsymbol{\mu}] = \mathbb{V}[_{\boldsymbol{x}}\boldsymbol{B}\boldsymbol{x}], \tag{10.6}$$

即，低维代码的方差不取决于

数据的平均数。因此，在本节的其余部分中，**0**我们不失时机地假设数据具有均值。在这个假设下，低维代码的平均数也是，因为$\mathbf{0}\mathbb{E}_{\boldsymbol{z}}[\boldsymbol{z}]=\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{B}\boldsymbol{x}]=\mathbf{B}\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}]=\mathbf{0}$. ◆



图中10.4PCA找到了一个当数据（蓝色）被投射到这个子空间（橙色）上时，能保持尽可能多的方差（数据的扩散）的低维子空间（线）。

10.2.1 差异最大的方向

我们使用连续的方式使低维代码的方差最大化方法。我们首先寻求一个单一的矢量$\boldsymbol{b}_1 \in \mathbb{R}^D$，使其达到最大。的方差，也就是说，我们的目标是使预测数据的方差最大化。$_1$的第一个坐标$\boldsymbol{z} \in \mathbb{R}$，因此，

$$V_1 := \mathbb{V}[z_1] = \frac{1}{N}\sum_{n=1}^{N} z_{1n}^2 \tag{10.7}$$

该矢量$\boldsymbol{b}_1$将是第一列的$\boldsymbol{B}$矩阵和因此是M个正态基向量中的第一个，即跨越低维的子空间。

最大化，其中我们利用了数据的i.i.d.假设，将z_{1n}定义为$\boldsymbol{x}_n \in \mathbb{R}^M$的低维表示$\boldsymbol{z}_n \in \mathbb{R}^D$的第一个坐标。

$$z_{1n} = \boldsymbol{b}_1^\top \boldsymbol{x}_n, \tag{10.8}$$

即，它是$\boldsymbol{x}_n$的正交投影到$\boldsymbol{b}_1$所跨越的一维子空间的坐标（第3.8节）。我们将(10.8)代入(10.7)，可以得到

$$V_1 = \frac{1}{N}\sum_{n=1}^{N} (\boldsymbol{b}_1^\top \boldsymbol{x}_n)^2 = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{b}_1^\top \boldsymbol{x}_n \boldsymbol{x}_n^\top \boldsymbol{b}_1 \tag{10.9a}$$

$$= \boldsymbol{b}_1^\top \left(\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top\right) \boldsymbol{b}_1 = \boldsymbol{b}_1^\top \boldsymbol{S} \boldsymbol{b}_1, \tag{10.9b}$$

其中$\boldsymbol{S}$是数据协方差矩阵，定义在(10.11)。在(10.9a)中，我们利用了两个向量的点积相对于其参数是对称的这一事实，即$\boldsymbol{b}_1^\top \boldsymbol{x}_n = \boldsymbol{x}_n^\top \boldsymbol{b}_1$。

请注意，任意增加矢量$\boldsymbol{b}_1$in-的幅值。增加V_1，也就是说，一个长度为2倍的向量$\boldsymbol{b}_1$可以导致V_1这可能是四倍大。因此，我们将所有解决方案限制在$\|\boldsymbol{b}_1\|^2 = 1$，这导致了一个受限的优化问题，在这个问题中，我们寻求数据变化最大的方向。

可以通过以下方法找到指向最大方差方向的

随着解空间被限制为单位向量，向量$\boldsymbol{b}_1$

$\iff \boldsymbol{b}_1 = 1$

受约束的优化问题

$$\max_{\boldsymbol{b}_1} \boldsymbol{b}_1^\top \boldsymbol{S}\boldsymbol{b}_1 \quad \text{受益于} \|\boldsymbol{b}_1\|_2^2 = 1 \tag{10.10}$$

根据第7.2节，我们得到拉格朗日

$$L(\boldsymbol{b}_1, \lambda_1) = \boldsymbol{b}_1^\top \boldsymbol{S}\boldsymbol{b}_1 + \lambda_1(1 - \boldsymbol{b}_1^\top \boldsymbol{b}_1) \tag{10.11}$$

来解决这个受限的优化问题。的偏导数
关于$\boldsymbol{b}_1$和λ_1的L_1是

$$\frac{\partial L}{\partial \boldsymbol{b}_1} = 2\boldsymbol{b}_1^\top \boldsymbol{S} - 2\lambda_1 \boldsymbol{b}_1^\top, \quad \frac{\partial L}{\partial \lambda_1} = 1 - \boldsymbol{b}_1^\top \boldsymbol{b}_1, \tag{10.12}$$

分别。将这些偏导数设为，$\mathbf{0}$我们可以得到以下关系

$$\boldsymbol{S}\boldsymbol{b}_1 = \lambda_1 \boldsymbol{b}_1, \tag{10.13}$$

$$\boldsymbol{b}_1^\top \boldsymbol{b}_1 = 1. \tag{10.14}$$

通过与特征值分解的定义相比较（第4.4节），我们看到$\boldsymbol{b}_1$是数据协方差矩阵$\boldsymbol{S}$的一个特征向量，而拉格朗日乘数λ_1起到了对应特征值的作用。这种特征向量的特性(10.13)使我们可以将方差目标(10.10)改写为

$$V_1 = \boldsymbol{b}_1^\top \boldsymbol{S}\boldsymbol{b}_1 = \lambda_1 \boldsymbol{b}_1^\top \boldsymbol{b}_1 = \lambda_1, \tag{10.15}$$

。

这个量$\sqrt{\lambda_1}$也被称为单位的*装载量*向量$\boldsymbol{b}_1$代表主子空间跨度所占数据的标准差$[\boldsymbol{b}_1]$。

主成分 .

即数据投射到一个一维子空间上的方差等于与横跨这个子空间的基向量$\boldsymbol{b}_1$相关的特征值λ_1。因此，为了使低维代码的方差最大化，我们选择与数据协方差矩阵的最大特征值相关的基向。这个特征向量被称为第一*主成分*。我们可以通过将坐标z_n映射回数据空间来确定主成分$\boldsymbol{b}_1$在原始数据空间中的影响/贡献，这样我们就得到了投影数据点

$$\tilde{\boldsymbol{x}}_n = b z_n = \boldsymbol{b}_1 \boldsymbol{b}_1^\top \boldsymbol{x}_n \in \mathbb{R}^D \tag{10.16}$$

在原始数据空间中。

备注。 虽然$\tilde{\boldsymbol{x}}_n$是一个D维向量，但它只需要一个坐标z_n来表示它与基向量$\boldsymbol{b}_1 \in \mathbb{R}^D$的关系。

10.2.2 具有最大方差的M维子空间

假设我们已经找到了前$m-1$个主成分，即$\boldsymbol{S}$的$m-1$个特征向量，它们与最大的$m-1$个特征值相关。由于$\boldsymbol{S}$是对称的，光谱定理（Theorem4.15）指出，我们可以用这些特征向量来构造一个正交的特征基。

(一般来说，第m个主成分可以通过减去前$m-1$个主成分$b_1, \ldots, b$的影响来找到D。$\boldsymbol{b}_{m-1}$的影响，从而试图找到能够压缩剩余信息的主成分。然后我们就可以得出

新的数据矩阵

$$\hat{\boldsymbol{X}} := \boldsymbol{X} - \sum_{i=1}^{m-1} b_i \boldsymbol{b}_i \boldsymbol{X} = \boldsymbol{X} - \boldsymbol{B}\boldsymbol{X}_{m-1} 。 \tag{10.17}$$

其中，$\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{D\times N}$包含数据点作为列向量和$\boldsymbol{B} := \sum_{i=1}^{m-1} \boldsymbol{b}_i \boldsymbol{b}_i^\top$是一个投影矩阵，投射到$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{m-1}$所跨越的子空间。[1] 矩阵$\hat{\boldsymbol{X}} := [\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_N] \in \mathbb{R}^{D\times N}$在(10.17)包含数据中尚未被压缩的信息。

备注（符号）。*在*本章中，我们不遵循收集数据$\boldsymbol{x}_1, \ldots$这意味着我们的数据矩阵$\boldsymbol{X}$是一个D N矩阵，而不是传统的N D矩阵。我们这样选择的原因是，代数运算可以顺利进行，不需要对矩阵进行转置，也不需要将向量重新定义为左乘到矩阵的行向量。 ◆

为了找到第m个主成分，我们使方差最大化。

$$V_m = \mathbb{V}[z_m] = \frac{1}{N}\sum_{n=1}^{N} z_{nm}^2 = \frac{1}{N}\sum_{n=1}^{N} (\boldsymbol{b}_m^\top \hat{\boldsymbol{x}}_n)^2 = \boldsymbol{b}_m^\top \hat{\boldsymbol{S}} \boldsymbol{b}_m , \tag{10.18}$$

受制于$\|\boldsymbol{b}_m\|^2 = 1$，其中我们遵循与（10.9b）相同的步骤并将$\hat{\boldsymbol{S}}$定义为转换后数据集的数据协方差矩阵$\hat{\boldsymbol{X}} := [\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_N]$。 如前所述，当我们看了第一条主线单独的组件，我们解决了一个受限的优化问题，并发现最优解$\boldsymbol{b}_m$是与$\hat{\boldsymbol{S}}$的最大特征值相关的$\hat{\boldsymbol{S}}$的特征向量。

结果发现，$\boldsymbol{b}_m$也是$\boldsymbol{S}$的一个特征向量。更一般地说，集合$\boldsymbol{S}$和$\hat{\boldsymbol{S}}$的特征向量是相同的。 由于$\boldsymbol{S}$和$\hat{\boldsymbol{S}}$都是对称的度量，我们可以找到一个特征向量的ONB（光谱定理4.15），即$\boldsymbol{S}$和$\hat{\boldsymbol{S}}$都存在D个不同的特征向量。接下来，我们表明，$\boldsymbol{S}$的每个特征向量都是$\hat{\boldsymbol{S}}$的特征向量。 假设我们已经找到了$\hat{S}$的特征向量$\boldsymbol{b}_1, \ldots, \hat{\boldsymbol{S}}_{m-1}$的$\boldsymbol{b}$。考虑S的一个特征向量$\boldsymbol{b}_i$，即$\boldsymbol{S}\boldsymbol{b}_i = \lambda_i \boldsymbol{b}_i$。一般来说。

$$\hat{\boldsymbol{S}}\boldsymbol{b}_i = \frac{1}{N}\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}^\top \boldsymbol{b}_i = \frac{1}{N}(\boldsymbol{X} - \boldsymbol{B}\boldsymbol{X})(\boldsymbol{X} - \boldsymbol{B}\boldsymbol{X})^\top \boldsymbol{b}_i \tag{10.19a}$$

$$= (\boldsymbol{S} - \boldsymbol{S}\boldsymbol{B} - \boldsymbol{B}\boldsymbol{S} + \boldsymbol{B}\boldsymbol{S}\boldsymbol{B})\boldsymbol{b}_i 。 \tag{10.19b}$$

我们对两种情况进行区分。如果$i \geqslant m$，即$\boldsymbol{b}_i$是一个不在前$m-1$个主成分中的特征向量，那么$\boldsymbol{b}_i$与前$m-1$个主成分是正交的，$\boldsymbol{B}\boldsymbol{b}_i = \boldsymbol{0}$。 如果$i < m$，即$\boldsymbol{b}_i$在前$m-1$个主成分中，那么$\boldsymbol{b}_i$是一个基向量。

的主子空间，$\boldsymbol{B}_{m-1}$投射到该子空间。 由于$b_1, \ldots \boldsymbol{b}_{m-1}$是这个主子空间的一个ONB，我们得到$_{m-1}\boldsymbol{B}\boldsymbol{b}_i = \boldsymbol{b}_i$。

$$_{m-1}\boldsymbol{B}\boldsymbol{b}_i = \text{bif}_i \quad i < m, \qquad _{m-1}\boldsymbol{B}\boldsymbol{b}_i = \boldsymbol{0} \quad \text{如果 } i \not< m。 \tag{10.20}$$

在$i<m$的情况下， 通过使用(10.20) 在(10.19b)中， 我们得到$\hat{\boldsymbol{S}}_i\boldsymbol{b} = (\boldsymbol{S} - {}_{m-1}\boldsymbol{B}\boldsymbol{S})\ \boldsymbol{b}_i = \boldsymbol{S}\boldsymbol{b}_i = \lambda_i\boldsymbol{b}_i$， 即$\boldsymbol{b}_i$也是$\hat{\boldsymbol{S}}$的一个特征向量，其特征是值$\lambda_i$。 具体来说。

$$\hat{\boldsymbol{S}}_m\boldsymbol{b} = \boldsymbol{S}\boldsymbol{b}_m = \lambda_m\boldsymbol{b}_m。 \tag{10.21}$$

公式(10.21)显示，$\boldsymbol{b}_m$不仅是$\boldsymbol{S}$的一个特征向量，也是
。 具体来说， λ_m是$\hat{\boldsymbol{S}}$的最大特征值，λ_m是第m个
$\boldsymbol{S}$的最大特征值，并且都有相关的特征向量$\boldsymbol{b}_m$。

在$i<m$的情况下，通过使用(10.20) 在(10.19b)中，我们得到

$$\hat{\boldsymbol{S}}_i\boldsymbol{b} = (\boldsymbol{S} - \boldsymbol{S}\boldsymbol{B}_{m-1} - {}_{m-1}\boldsymbol{B}\boldsymbol{S} + {}_{m-1}\boldsymbol{B}\boldsymbol{S}\boldsymbol{B})_{m-1}\boldsymbol{b}_i = = \boldsymbol{0}0\boldsymbol{b}\,(_i 10.22)$$

这意味着，$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{m-1}$也是$\hat{\boldsymbol{S}}$的特征向量， 但它们与特征值相关，所以$0\boldsymbol{b}_1, \ldots \ldots, b$跨越$\hat{S}$的空空间。$\boldsymbol{b}_{m-1}$横跨$\hat{\boldsymbol{S}}$的无效空间。 总的来说，$\boldsymbol{S}$的每个特征向量也是$\hat{\boldsymbol{S}}$的一个特征向量。 然而。如果$\boldsymbol{S}$的特征向量是（m1）维主子空间的一部分，那么$\hat{\boldsymbol{S}}$的相关特征值是0.

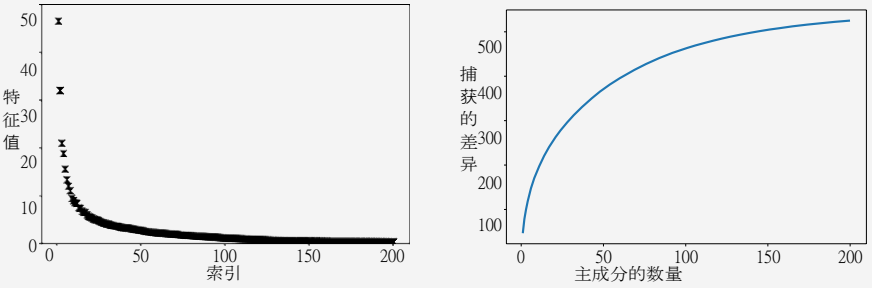
这一推导表明，在 "我 "与 "我 "之间存在着紧密的联系。的M维子空间的最大方差和特征值分解。我们将在第一节中重新审视这种联系。10.4.

通过关系(10.21)和$\boldsymbol{b}\boldsymbol{m}\boldsymbol{b}_m$=1， 数据的方差亲剖析到第$m$个主成分是

$$V_m = \boldsymbol{b}\boldsymbol{m}\boldsymbol{S}\boldsymbol{b}_m \overset{(10.21)}{=} \lambda_m\boldsymbol{b}\boldsymbol{m}\boldsymbol{b}_m = \lambda_m。 \tag{10.23}$$

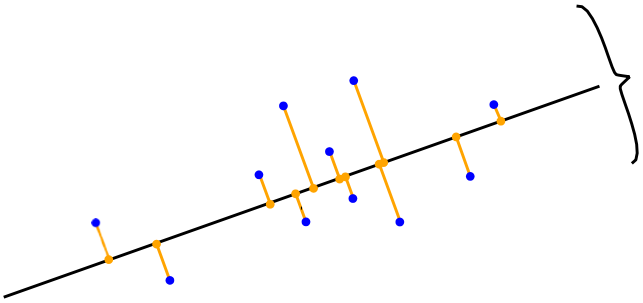
这意味着数据的方差在投射到一个M维的子空间时，等于与数据协方差矩阵的相应特征向量有关的特征值之和。

例子 （MNIST的10.2特征值 "8"）。



图：MNIST "8 "的训练数据的属性**10.5**。(a)按降序排列的特征值；(b)与最大特征值相关的主成分捕获的方差。

(a) (b)所有数字 "8 "的数据协方差矩阵的特征值（按降序排列）。训练集的MNIST。



图为投影方法的说明**10.6**。找到一个子空间（线），使投影（橙色）和原始（蓝色）数据之间的差异向量长度最小。

总的来说，为了找到一个能保留尽可能多信息的M维子空间，PCA告诉我们要选择矩阵$\boldsymbol{B}$中的列(10.3)作为数据协方差矩阵$\boldsymbol{S}$的M个特征向量，与M个最大的特征值相关。PCA能够用前M个主成分捕获的最大方差量是

$$V = \sum_{m=1}^{M} \lambda_m \,, \tag{10.24}$$

其中λ_m是数据协方差矩阵的M个最大特征值$\boldsymbol{S}$.因此，通过PCA进行数据压缩所损失的方差是

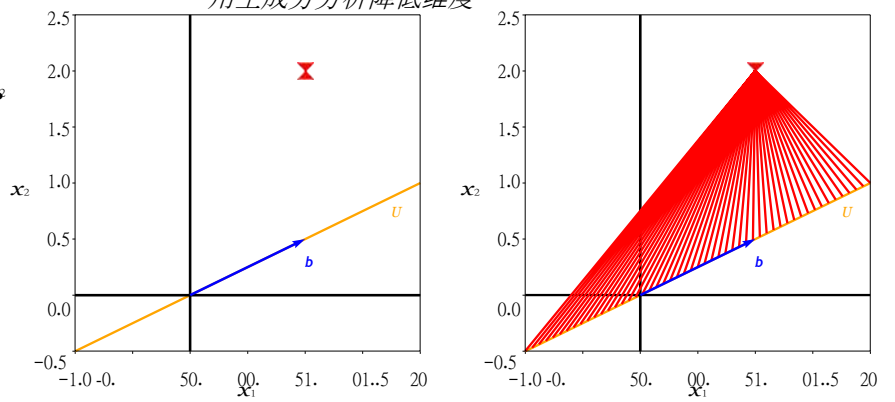
$$J_M := \sum_{j=M+1}^{D} \lambda_j = V_D - V_M \,. \tag{10.25}$$

取代这些绝对量，我们可以将捕获的相对方差定义为$\frac{V_M}{V_D}$，而将压缩后损失的相对方差定义为$1-\frac{V_M}{V_D}$。

10.3 投影透视

在下文中，我们将把PCA推导为一种直接最小化平均重建误差的算法。这一观点使我们能够将PCA解释为实现一个最佳的线性自动编码器。我们将大量借鉴第二章和第三章的内容。

在上一节中，我们通过最大化投影空间的方差来推导PCA，以保留尽可能多的信息。在



图：简化**10.7**的投影设置。(a) 一个二维向量$\in \mathcal{R}^2$(红叉)应投射到一个一维子空间$U \subseteq \mathcal{R}^2$上所跨越的范围$\boldsymbol{b}$。(b)显示了$\boldsymbol{x}$和一些候选者$\tilde{\boldsymbol{x}}$。

(a) 设置。 (b) 50个不同$\tilde{\boldsymbol{x}}_i$的差异$\boldsymbol{x} - \tilde{\boldsymbol{x}}_i$由红线表示。

下面，我们将查看原始数据$\boldsymbol{x}_n$和它们的重建$\tilde{\boldsymbol{x}}$之间的差异向量，$_n$并使这个距离最小化，以便$\boldsymbol{x}_n$和$\tilde{\boldsymbol{x}}_n$尽可能地接近。图10.6说明了这种设置。

10.3.1 设置和目标

假设R的^D（有序）正态基（ONB）B=（$\boldsymbol{b}_1$，...，$\boldsymbol{b}_D$），即$_{bi}\,\boldsymbol{b}_j$=当且1仅当i=j，否则0。

从第2.5节我们知道，对于R的^{D}一个基（b$_1$, ..., $\boldsymbol{b}_D$），任何$\boldsymbol{x}$ $\in$ R^{D}可以写成R的^{D}基向量的线性组合，即：。

$$\boldsymbol{x} = \sum_{d=1}^{D} \zeta_d \boldsymbol{b}_d = \sum_{m=1}^{M} \zeta_m \boldsymbol{b}_m + \sum_{j=M+1}^{D} \zeta_j \boldsymbol{b}_j \tag{10.26}$$

矢量$\in U$可以是$\mathcal{R}^3$中某一平面上的向量，该平面的维度为 2，但的向量仍有三座标相对于标准的基础$\mathcal{R}^3$.

对于合适的坐标ζ_d $\in$ R。

我们感兴趣的是要找到向量$\tilde{\boldsymbol{x}}$ $\in$R^D，它们生活在较低的维的子空间$U \subseteq$R^D，dim(U)=M，因此，

$$\tilde{\boldsymbol{x}} = \sum_{m=1}^{M} z_m \boldsymbol{b}_m \in U \subseteq \mathrm{R}^D \tag{10.27}$$

是尽可能地与$\boldsymbol{x}$相似。注意，此时我们需要假设$\tilde{\boldsymbol{x}}$的坐标z_m和$\boldsymbol{x}_m$的ζ不完全相同。

在下文中，我们正是用$\tilde{\boldsymbol{x}}$的这种表示法来寻找最佳坐标$\boldsymbol{z}$和基向量$\boldsymbol{b}_1$，. . .$\boldsymbol{b}$，$_M$使$\tilde{\boldsymbol{x}}$与原始数据点$\boldsymbol{x}$尽可能相似，也就是说，我们的目标是最小化（欧氏）距离$\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|$。图10.7说明了这种设置。

在不丧失一般性的情况下，我们假设数据集X={$\boldsymbol{x}_1$, . ., $\boldsymbol{x}_N$}，$\boldsymbol{x}_n \in$R^D，以 **0**，即E[X]=为中心**0**。如果没有零均值假设，则

我们将得出完全相同的解决方案，但符号将大大地更加杂乱。

我们感兴趣的是找到$\mathbb{R}^D$的低维子空间U的最佳线性投影$\dim(U)=M$和正态基向量$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M$。我们将这个子空间$U$称为*主子空间*。 主子空间

数据点的投影表示为

$$\tilde{\boldsymbol{x}}_n := \sum_{m=1}^{M} z_{mn}\boldsymbol{b}_m = \boldsymbol{B}\boldsymbol{z}_n \in U \subseteq \mathbb{R}^D。 \tag{10.28}$$

其中$\boldsymbol{z}_n := [z_{1n}, \ldots, z_{Mn}]^\top \in \mathbb{R}^M$是$\tilde{\boldsymbol{x}}_n$相对于基础$(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M)$的坐标向量。更具体地说，我们感兴趣的是让$\tilde{\boldsymbol{x}}_n$尽可能地与$\boldsymbol{x}_n$相似。

我们在下面使用的相似性度量是$\boldsymbol{x}_n$和$\tilde{\boldsymbol{x}}_n$之间的平方距离（欧氏规范）$\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n$。因此，我们定义我们的目标是最小化平均欧氏距离的平方（*重构误差*）（Pearson,1901）。

$$J_M := \frac{1}{N}\sum_{n=1}^{N} \|\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n\|^2, \tag{10.29}$$

其中我们明确指出，我们将数据投射到的子空间的维度是M。为了找到这个最佳的线性投影，我们需要找到主子空间的正态基以及相对于这个基的投影的坐标$\boldsymbol{z}_n \in \mathbb{R}^M$。

为了找到主子空间的坐标$\boldsymbol{z}_n$和ONB，我们采用了两步方法。首先，我们优化给定ONB（$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M$）的坐标$\boldsymbol{z}_n$；其次，我们找到最优的ONB。

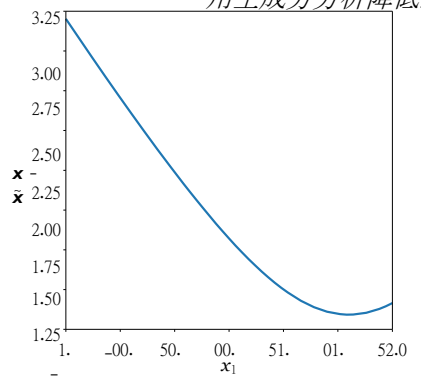
10.3.2 寻找最佳坐标

让我们从寻找最佳坐标$z_{1n}, \ldots, z_{Mn}$的projec- tions $\tilde{\boldsymbol{x}}_n$为$n = 1, \ldots, N$。考虑图10.7(b)，其中主子空间被一个单一的矢量$\boldsymbol{b}$所跨越。从几何学上讲，找到最佳坐标z相当于找到相对于$\boldsymbol{b}$的线性投影$\tilde{\boldsymbol{x}}$的表示，使$\tilde{\boldsymbol{x}}$的距离最小。$\boldsymbol{x}$. 从图10.7(b)可以看出，这将是正交投影，在下文中我们将确切地表明这一点。

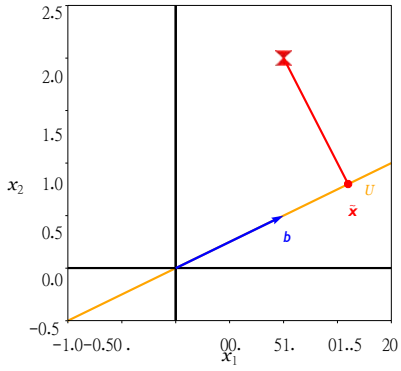
我们假设$U \subseteq \mathbb{R}^D$的ONB（$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M$）。为了找到相对于这个基础的最佳共轭坐标$\boldsymbol{z}_m$，我们需要偏导数

$$\frac{\partial J_M}{\partial z_{in}} = \frac{\partial J_M}{\partial \tilde{\boldsymbol{x}}_n}\frac{\partial \tilde{\boldsymbol{x}}_n}{\partial z_{in}}。 \tag{10.30a}$$

$$\frac{\partial J_M}{\partial \tilde{\boldsymbol{x}}_n} = \frac{2}{N}(\tilde{\boldsymbol{x}}_n - \boldsymbol{x}_n)^\top \in \mathbb{R}^{1\times D}, \tag{10.30b}$$



(a) 距离 $\boldsymbol{x} - \tilde{\boldsymbol{x}}$ 为一些 $\tilde{\boldsymbol{x}} = z_1\boldsymbol{b} \in \text{span}[\quad]$；设置见面板（b）。



(b) 在面板(a)中最小化距离的矢量 $\tilde{\boldsymbol{x}}$ 是它在 U 上的正交投影。 投影 $\tilde{\boldsymbol{x}}$ 相对于跨越 U 的基础矢量 $\boldsymbol{b}$ 的坐标是什么？

图 10.8 最优投影的向量到 $\boldsymbol{x} \in \mathcal{R}^2$ 一个的一维子空间（从图中延续出来的 10.7).(a)距离 $\boldsymbol{x}$ 对于一些 $\tilde{\boldsymbol{x}} \in U$. (b)正交投影和最佳坐标。

是我们需要扩展 $\boldsymbol{b}$ 的因素，以便 "达到" 。

$$\frac{\partial \tilde{\boldsymbol{x}}_n}{\partial z_{in}} \overset{(10.28)}{=} \frac{\partial}{\partial z_{in}} \sum_{m=1}^{M} z_{nm} \boldsymbol{b}_m = \boldsymbol{b}_i \tag{10.30c}$$

对于 $i = 1, \ldots, M$，这样，我们得到

$$\frac{\partial J_M}{\partial z_{in}} \overset{(10.30b)\ (10.30c)}{=} -\frac{2}{N}(\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n)\boldsymbol{b}_i \overset{(10.28)}{=} -\frac{2}{N}\left(\boldsymbol{x}_n - \sum_{m=1}^{M} z_{nm}\boldsymbol{b}_m\right)\boldsymbol{b}_i \tag{10.31a}$$

$$\overset{\text{ONB}}{=} -\frac{2}{N}(\boldsymbol{x}_n \boldsymbol{b}_i - z_{in}\boldsymbol{b}_i \boldsymbol{b}_i) = -\frac{2}{N}(\boldsymbol{x}_n \boldsymbol{b}_i - z_{in}) 。 \tag{10.31b}$$

n相对于基向量的最佳投影的坐标 $\boldsymbol{b}, \ldots, \boldsymbol{b}_M$ 是正交的坐标。 $\boldsymbol{x}_n$ 的预测到主子空间。

因为 $\boldsymbol{b}_i \boldsymbol{b}_i = 1$. 将此偏导设为立即0得到最佳坐标

$$z_{in} = \boldsymbol{x}_n \boldsymbol{b}_i = \boldsymbol{b}_i \boldsymbol{x}_n \tag{10.32}$$

对于 $i = 1, \ldots, M$ 和 $n = 1, \ldots, N$.这意味着投影 $\tilde{\boldsymbol{x}}_n$ 的最佳共轭坐标 z_{in} 是原始数据点 $\boldsymbol{x}$ 的 n 正交投影的坐标（见第3.8节）到一因此，$\boldsymbol{b}$ 所 i 跨越的维度子空间。

- $\boldsymbol{x}$ 的最优线性投影 $\tilde{\boldsymbol{x}}_{nn}$ 是一个正交投影。 $\tilde{\boldsymbol{x}}_n$ 相对于基 $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M)$ 的坐标
- 是 $\boldsymbol{x}_n$ 对主子空间的正交投影的坐标。
- 正交投影是给定目标的最佳线性映射(10.29).
- $\boldsymbol{x}$ 的坐标 ζ_m 在(10.26)和 $\tilde{\boldsymbol{x}}$ 的坐标 z_m 在(10.27)

对于$m=1,\ldots$因为$U=^{\perp}$ span$[\boldsymbol{b}_{M+1},\ldots,\boldsymbol{b}]_D$是$U=$ span$[\boldsymbol{b}_1,\ldots,\boldsymbol{b}]_M$的正交补充（见第3.6节）。

备注（正交投影与正交基矢）。让我们简单回顾一下第3.8节的正交投影。如果$(\boldsymbol{b}_1,\ldots,\boldsymbol{b}_D)$是$\mathbb{R}^D$一个正态基，那么

$\boldsymbol{b}_j^{\top}\boldsymbol{x}$ 的正交投影的坐标。

$$\tilde{\boldsymbol{x}}=\boldsymbol{b}_j(\boldsymbol{b}_j^{\top}\boldsymbol{b}_j)^{-1}\boldsymbol{b}_j^{\top}\boldsymbol{x}=\boldsymbol{b}_j\boldsymbol{b}_j^{\top}\boldsymbol{x}\in\mathbb{R}^D \tag{10.33}$$

是$\boldsymbol{x}$在第j*个基*向量所跨越的子空间上的正交投影，而$z_j=\boldsymbol{b}_j^{\top}\boldsymbol{x}$是这个投影相对于跨越该子空间的基向量$\boldsymbol{b}_j$的$j$坐标，因为$z\boldsymbol{b}_j=\tilde{\boldsymbol{x}}$。图10.8(b)说明了这种设置。

$\boldsymbol{x}$ 的子空间跨度为 $\boldsymbol{b}_j$。

更一般地说，如果我们的目标是投射到$\mathbb{R}^D$一个M维子空间，我们得到$\boldsymbol{x}$的正交投射到M维子空间的正交基向量$\boldsymbol{b}_1,\ldots,\boldsymbol{b}_M$为

$$\tilde{\boldsymbol{x}}=\boldsymbol{B}(\underbrace{\boldsymbol{B}^{\top}\boldsymbol{B}}_{=\boldsymbol{I}})^{-1}\boldsymbol{B}^{\top}\boldsymbol{x}=\boldsymbol{B}\boldsymbol{B}^{\top}\boldsymbol{x}, \tag{10.34}$$

其中我们定义$\boldsymbol{B}:=[\boldsymbol{b}_1,\ldots,\boldsymbol{b}_M]\in\mathbb{R}^{D\times M}$如第3.8节所述，这个投影相对于有序基$(\boldsymbol{b}_1,\ldots,\boldsymbol{b}_M)$的坐标为$\boldsymbol{z}:=\boldsymbol{B}^{\top}\boldsymbol{x}$。

我们可以把坐标看作是在一个由$(\boldsymbol{b}_1,\ldots,\boldsymbol{b}_M)$定义的新坐标系中的投影矢量的表示。请注意，al-
虽然$\tilde{\boldsymbol{x}}\in\mathbb{R}^D$，我们只需要$M$个坐标$z_1,\ldots,z_M$来表示
这个向量；其他$D-M$坐标相对于基向量而言
$(\boldsymbol{b}_{M+1},\ldots,\boldsymbol{b}_D)$总是0. ◆

到目前为止，我们已经表明，对于一个给定的ONB，我们可以通过正交投影到主子空间找到$\tilde{\boldsymbol{x}}$的最佳坐标。在下文中，我们将确定什么是最佳基础。

10.3.3 寻找主子空间的基点

为了确定主子空间的基向量$\boldsymbol{b}_1,\ldots,$主子空间$_M$的$\boldsymbol{b}$，我们用到目前为止的结果重新表述损失函数(10.29)，使用我们迄今为止的结果。这将使我们更容易找到基向量。为了重新表述损失函数，我们利用之前的结果，得到

$$\tilde{\boldsymbol{x}}_n=\sum_{m=1}^{M}z_m\boldsymbol{b}_m\overset{(10.32)}{=}\sum_{m=1}^{M}(\boldsymbol{b}_m^{\top}\boldsymbol{x}_n)\boldsymbol{b}_m=\boldsymbol{B}\boldsymbol{B}^{\top}\boldsymbol{x}_n. \tag{10.35}$$

我们现在利用点积的对称性，可以得到

$$\tilde{\boldsymbol{x}}_n=\sum_{m=1}^{M}\boldsymbol{b}_m\boldsymbol{b}_m^{\top}\boldsymbol{x}_n. \tag{10.36}$$

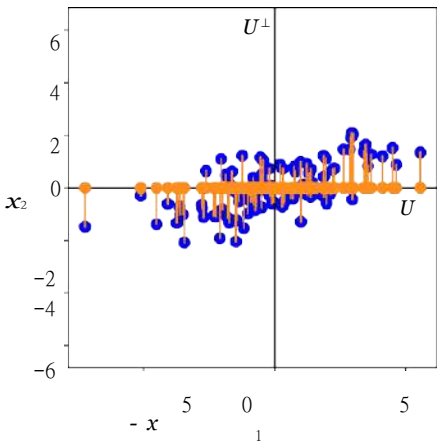


图 10.9 正交投影和流离失所向量。当预测数据点 x_n（蓝色）到子空间 U 上，我们得到 $\tilde{x}_n$（橙色）。的位移向量 $\tilde{x}_n - x_n$ 谎言完全在 U_2 的正交补数中。U_1.

由于我们通常可以将原始数据点 x_n 写成所有基向量的线性组合，因此可以认为

$$x_n = \sum_{d=1}^{D} z_{dn} b_d \overset{(10.32)}{=} \sum_{d=1}^{D} (x_n^\top b_d) b_d = \sum_{d=1}^{D} b_d b_d^\top x_n \tag{10.37a}$$

$$= \sum_{m=1}^{M} b_m b_m^\top x_n + \sum_{j=M+1}^{D} b_j b_j^\top x_n, \tag{10.37b}$$

其中，我们将 D 项之和分成 M 项之和和 $D-M$ 项之和。有了这个结果，我们发现位移矢量 $x_n - \tilde{x}_n$，即原始数据点和其投影之间的差异矢量，是

$$x_n - \tilde{x}_n = \sum_{j=M+1}^{D} b_j b_j^\top x_n \tag{10.38a}$$

$$= \sum_{j=M+1}^{D} (x_n^\top b_j) b_j . \tag{10.38b}$$

这意味着差值正是数据点对主子空间的正交补数的投影。我们确定主子空间的

(10.38a) 中的Trix $b_j b_j^\top$ 作为投影矩阵来执行这个任务。投影。$j=M+1$ 因此，位移矢量 $x_n - \tilde{x}_n$ 位于子空间中如图所示，它与主子空间正交。10.9.

备注（低秩近似）。在(10.38a)中，我们看到，将 x_n 投射到 $\tilde{x}_n$ 上的projection矩阵由以下公式给出

$$\sum_{m=1}^{M} b_m b_m^\top = BB^\top . \tag{10.39}$$

根据构造，作为秩一矩阵 $b_m b_m^\top$ 的总和，我们看到 $BB^\top$ 是

对称的，并且有M级。因此，平均重建误差的平方也可以写为

$$\frac{1}{N}\sum_{n=1}^{N}\|\boldsymbol{x}_n-\tilde{\boldsymbol{x}}_n\|^2=\frac{1}{N}\sum_{n=1}^{N}\|\boldsymbol{x}_n-\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{x}_n\|^2 \tag{10.40a}$$

$$=\frac{1}{N}\sum_{n=1}^{N}\|(\boldsymbol{I}-\boldsymbol{B}\boldsymbol{B}^\top)\boldsymbol{x}_n\|^2. \tag{10.40b}$$

寻找正态基向量$\boldsymbol{b}_1,\ldots,\boldsymbol{b}_M$，使差异最小化--PCA找到最佳的原始数据$\boldsymbol{x}_n$和它们的投影$\tilde{\boldsymbol{x}}_n$之间的关系，相当于找到身份矩阵$\boldsymbol{I}$的最佳等级-$M$近似$\boldsymbol{B}\boldsymbol{B}^\top$（见第4.6节）。◆

身份矩阵的等级-M近似。

现在我们有了所有的工具来重新表述损失函数(10.29).

$$J_M=\frac{1}{N}\sum_{n=1}^{N}\|\boldsymbol{x}_n-\tilde{\boldsymbol{x}}_n\|^2\overset{(10.38b)}{=}\frac{1}{N}\sum_{n=1}^{N}\sum_{j=M+1}^{D}(\boldsymbol{b}_j^\top\boldsymbol{x}_n)^2\boldsymbol{b}_j^\top\boldsymbol{b}_j. \tag{10.41}$$

我们现在明确地计算平方法则，并利用这一事实，即$\boldsymbol{b}_j$形成一个ONB，由此可得

$$J_M=\frac{1}{N}\sum_{n=1}^{N}\sum_{j=M+1}^{D}(\boldsymbol{b}_j^\top\boldsymbol{x}_n)^2=\frac{1}{N}\sum_{n=1}^{N}\sum_{j=M+1}^{D}\boldsymbol{b}_j^\top\boldsymbol{x}_n\boldsymbol{x}_n^\top\boldsymbol{b}_j \tag{10.42a}$$

$$=\frac{1}{N}\sum_{n=1}^{N}\sum_{j=M+1}^{D}\boldsymbol{b}_j^\top\boldsymbol{x}_n\boldsymbol{x}_n^\top\boldsymbol{b}_j, \tag{10.42b}$$

其中我们利用了上一步中点积的对称性，写成$\boldsymbol{b}_j^\top\boldsymbol{x}_n=\boldsymbol{x}_n^\top\boldsymbol{b}_j$.我们现在交换和，得到

$$J_M=\sum_{j=M+1}^{D}\boldsymbol{b}_j^\top\left[\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top\right]\boldsymbol{b}_j=\sum_{j=M+1}^{D}\boldsymbol{b}_j^\top\boldsymbol{S}\boldsymbol{b}_j \tag{10.43a}$$

$$=\sum_{j=M+1}^{D}\operatorname{tr}(\boldsymbol{b}_j^\top\boldsymbol{S}\boldsymbol{b}_j)=\sum_{j=M+1}^{D}\operatorname{tr}(\boldsymbol{S}\boldsymbol{b}_j\boldsymbol{b}_j^\top)=\operatorname{tr}\Big(\sum_{j=M+1}^{D}\boldsymbol{S}\boldsymbol{b}_j\boldsymbol{b}_j^\top\Big), \tag{10.43b}$$

投影矩阵

其中，我们利用了跟踪算子tr()(见(4.18))是线性的且对其参数的循环排列不变的特性。由于我们假设我们的数据集是居中的，即$\mathbb{E}[\boldsymbol{X}]=\boldsymbol{0}$，我们将$\boldsymbol{S}$确定为数据协方差矩阵。由于(10.43b)中的投影矩阵是CON.

结构为秩一矩阵$\boldsymbol{b}_j\boldsymbol{b}_j^\top$之和，其本身的秩为$D$。

方程(10.43a)意味着我们可以把平均重建误差的平方值等效为数据的协方差矩阵。

最小化平均平方重建误差等同于最小化数据协方差矩阵在主子空间正交补数上的投影。

投射到主子空间的正交补充上。因此，最小化平均重建误差平方相当于最小化数据投射到我们所忽略的子空间，即主子空间的正交补数上时的方差。等价地，我们使保留在主子空间中的投影方差最大化，这使投影损失立即与本节中讨论的PCA的最大方差表述相联系，10.2.但这也意味着我们将获得与最大方差观点相同的解决方案。因此，我们省略了与第1节中提出的相同的推导。10.2的推导，并从投影的角度对前面的结果进行总结。

最小化平均平方重建误差等同于最大化

当投射到M维主子空间时，平均重建误差的平方为

$$J_M = \sum_{j=M+1}^{D} \lambda_j, \quad (10.44)$$

预测数据的方差。

其中λ_j是数据协方差矩阵的特征值。因此，为了最小化(10.44)，我们需要选择最小的D M特征值，这就意味着它们相应的特征向量是主子空间的正交补数的基础。因此，这意味着主子空间的基础包括与最大的特征值相关的特征向量$\boldsymbol{b}_1, \ldots \boldsymbol{b}_M$，与数据协方差矩阵的最大$M$个特征值相关。

示例（10.3MNIST数字嵌入）。

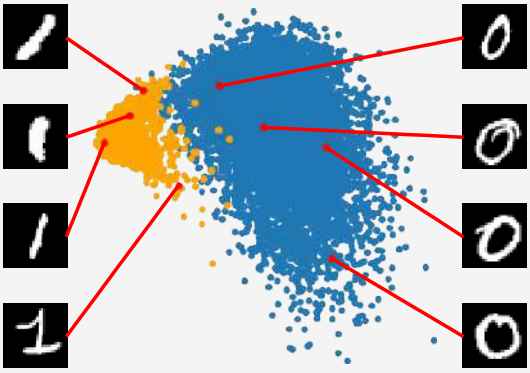


图 10.10 MNIST数字的嵌入0(蓝色)和1(橙色）在一个二维的使用PCA的主子空间。数字"0"和"1"在主子空间中的四个嵌入与它们相应的原始数字以红色标示。

图中10.10将MMIST数字"0"和"1"的训练数据嵌入到前两个主成分所跨越的矢量子空间中。我们观察到"0"（蓝点）和"1"（橙点）之间有一个相对清晰的分离，我们看到每个个体内部的变化。

簇。数字 "0 "和 "1 "在主子空间中的四个嵌入与它们相应的原始数字以红色和蓝色显示。请注意，"0 "的集合内的变化明显大于 "1 "的集合内的变化。

10.4 特征向量计算和低秩近似法

在前面的章节中，我们得到了主子空间的基础，即与数据协方差矩阵的最大特征值相关的特征向量。

$$\boldsymbol{S} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^\top, \tag{10.45}$$

$$\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{D\times N}。 \tag{10.46}$$

请注意，$\boldsymbol{X}$是一个 $D \times N$矩阵，即它是 "典型 "数据矩阵的转置（Bishop，2006；Murphy，2012）。为了得到特征值（和的相应特征向量），我们可以遵循两种方法：

使用严格分解或SVD来计算特征向量。

- 我们进行一个特征分解（见第4.2节），直接计算$\boldsymbol{S}$的特征值和特征向量。
- 我们使用奇异值分解法（见第4.5节）。由于$\boldsymbol{S}$是对称的，并被分解为$\boldsymbol{X}\boldsymbol{X}^\top$（忽略因子$\frac{1}{N}$），所以特征-$\boldsymbol{S}$的值是$\boldsymbol{X}$的平方奇异值。

更具体地说，$\boldsymbol{X}$的SVD由以下公式给出

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top, \tag{10.47}$$

其中$\boldsymbol{U} \in \mathbb{R}^{D\times D}$和$\boldsymbol{V} \in \mathbb{R}^{N\times N}$是正交矩阵，$\boldsymbol{\Sigma} \in \mathbb{R}^{D\times N}$是一个矩阵，其唯一的非零条目是奇异值$\sigma_{ii} \geqslant 0$。由此可见，

$$\boldsymbol{S} = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^\top = \frac{1}{N}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top = \boldsymbol{U}\frac{1}{N}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top。 \tag{10.48}$$

根据第4.5节的结果，我们可以得到，$\boldsymbol{U}$的列是$\boldsymbol{S}$的特征向量。

$\boldsymbol{X}\boldsymbol{X}^\top$（因此也是$\boldsymbol{S}$）的特征向量。 此外，特征值$\boldsymbol{S}_d$的$\lambda$与$\boldsymbol{X}$的奇异值的关系是通过

$$\lambda_d = \frac{\sigma_d^2}{N}. \tag{10.49}$$

$\boldsymbol{S}$的特征值和$\boldsymbol{X}$的奇异值之间的这种关系提供了最大方差观点（第10.2节）和奇异值分解之间的联系。

10.4.1 使用低秩矩阵近似法的PCA

为了使投影数据的方差最大化（或使重建的平均平方误差最小化），PCA选择$\boldsymbol{U}$中的列作为与数据协方差矩阵S的M个最大特征值相关的特征向量。10.48)中的列是与数据协方差矩阵$\boldsymbol{S}$的M个最大特征值相关的特征向量，这样我们就把$\boldsymbol{U}$确定为(10.3)，它将原始数据投射到一个维度为M的低维子空间。*Eckart-Young定理*（Theorem4.25 in 第4.6节）提供了一种估计低维代表的直接方法。考虑最好的等级-M近似

埃卡特-杨（Eckart-Young

定理

$$\tilde{\boldsymbol{X}}_M := \operatorname{argmin}_{\operatorname{rk}(\boldsymbol{A})\leq M} \|\boldsymbol{X} - \boldsymbol{A}\|_2 \in \mathbb{R}^{D\times N} \tag{10.50}$$

的，其中$\|\cdot\|_2$是（4.93）中定义的谱准则。Eckart-Young定理指出，$\tilde{\boldsymbol{X}}_M$是通过在顶部截断SVD-M而得到的。

奇异值。换言之，我们得到

$$\tilde{\boldsymbol{X}}_M = \underset{D\times M}{\boldsymbol{U}_M} \underset{M\times M}{\boldsymbol{\Sigma}_M} \underset{M\times N}{\boldsymbol{V}_M^\top} \in \mathbb{R}^{D\times N} \tag{10.51}$$

正交矩阵$\boldsymbol{U}_{:M} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M] \in \mathbb{R}^{D\times M}$和$\boldsymbol{V}_{:M} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_M] \in \mathbb{R}^{N\times M}$和一个对角矩阵$\boldsymbol{\Sigma}_M \in \mathbb{R}^{M\times M}$其对角条目是$\boldsymbol{X}$的$M$个最大奇异值。

10.4.2 实用方面

寻找特征值和特征向量对于其他需要矩阵分解的基础机器学习方法也很重要。在理论上，正如我们在第4.2节所讨论的，我们可以把特征值作为特征多项式的根来解决。然而，对于大于4的N，这是不可能的，因为我们需要找到一个度数或更高的5多义词的根。然而，*阿贝尔-鲁菲尼定理*（Ruffini, 1799；Abel,1826）指出，这个问题不存在代数解。因此，在实践中，我们使用迭代方法求解特征值或奇异值，这些方法在所有现代线性代数软件包中都能实现。

Abel-Ruffini

定理

辽宁

或

np.linalg.svd

在许多应用中（如本章介绍的PCA），我们只需要几个特征向量。如果计算完整的去构成，然后丢弃所有具有超出前几个特征值的特征向量，那就太浪费了。事实证明，如果我们只对前几个特征向量（具有最大的特征值）感兴趣，那么直接优化这些特征向量的迭代过程在计算上要比完整的eigendecomposition（或SVD）更有效率。在只需要第一个特

权力迭代

征向量的极端情况下，一种叫做*幂迭代*的简单方法非常有效。功率迭代选择一个随机的向量$\boldsymbol{x}_0$，它不在

$\boldsymbol{S}$的无效空间，并遵循迭代原则

$$x_{k+1} = \frac{\boldsymbol{S}\boldsymbol{x}_k}{\|\boldsymbol{S}\boldsymbol{x}_k\|}, \quad k = 0, 1, \ldots. \tag{10.52}$$

这意味着向量$\boldsymbol{x}_k$在每次迭代中都乘以$\boldsymbol{S}$，然后这个向量序列指向与$\boldsymbol{S}$的最大特征值相关的特征向量1。最初的Google PageRank算法（Page等人，1999）使用这样的算法，根据网页的超链接进行排名。

If 是可逆的，它是充分的，以确保 $\boldsymbol{x}_0$ I=O .

10.5 高维度的PCA

为了进行PCA，我们需要计算数据协方差矩阵。在D维中，数据协方差矩阵是一个D D矩阵。计算这个矩阵的特征值和特征向量在计算上是很昂贵的，因为它在D中是立体扩展的。因此，正如我们前面所讨论的，PCA在非常高的维度上是不可行的。例如，如果我们的$\boldsymbol{x}_n$是有10IV个000像素的图像（如像素100100图像），我们将需要计算10IV00010个协方差矩阵的特征分解。在下文中，我们提供了一个解决这个问题的方法，即我们的数据点大大少于维数，即N D。

假设我们有一个居中的数据集$\boldsymbol{x}_1, \ldots$那么数据的协方差矩阵为

$$\boldsymbol{S} = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^\top \in \mathbb{R}^{D\times D}, \tag{10.53}$$

其中$\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$是一个$D$ N矩阵，其列是数据点。

现在我们假设N D，即数据点的数量小于数据的维度。如果没有重复的数据点，协方差矩阵$\boldsymbol{S}$的等级是N，所以它有D N + 1许多特征值，这些特征值是0。直观地说，这意味着存在一些冗余。在下文中，我们将利用这一点，把D D协方差矩阵变成N N协方差矩阵，其特征值都是正的。

在PCA中，我们最后得到的是特征向量方程

$$\boldsymbol{S}\boldsymbol{b}_m = \lambda_m \boldsymbol{b}_m, \quad m = 1, \ldots, M, \tag{10.54}$$

其中$\boldsymbol{b}_m$是主子空间的一个基向量。让我们稍微重写一下这个方程。随着$\boldsymbol{S}$在(10.53)，我们得到

$$\boldsymbol{S}\boldsymbol{b}_m = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{b}_m = \lambda_m \boldsymbol{b}_m. \tag{10.55}$$

我们现在从左手边乘以$\boldsymbol{X}^\top \in \mathbb{R}^{N\times D}$，得到的是

$$\frac{1}{N}\underbrace{\boldsymbol{X}^\top\boldsymbol{X}}_{N\times N}\boldsymbol{X}^\top\boldsymbol{b}_m = \lambda_m \boldsymbol{X}^\top\boldsymbol{b}_m \iff \frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{c}_m = \lambda_m \boldsymbol{c}_m, \tag{10.56}$$

我们得到一个新的特征向量/特征值方程：λ_m仍然是特征值，这证实了我们在第4.5.3节的结果，即非零的

$\boldsymbol{XX}^\top$的特征值等于$\boldsymbol{X}^\top\boldsymbol{X}$的非零特征值。我们得到与$\lambda$相关的矩阵$\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X} \in \mathbb{R}$的$N\times N$特征向量$\boldsymbol{c}_m$为$\boldsymbol{X}\boldsymbol{b}_m$。假设我们没有重复的数据点，这个矩阵有$N$级，并且是可倒置的。这也意味着$\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X}$具有与数据协方差矩阵$\boldsymbol{S}$相同的（非零）特征值，但这现在是一个$N\times N$矩阵，因此我们可以比对原始$D\times D$数据协方差矩阵更有效地计算特征值和特征向量。现在我们有了$\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X}$的特征向量，我们要重新覆盖原始特征向量，我们仍然需要PCA。目前，我们知道了$\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X}$的特征向量，如果我们把我们的特征值/左乘以用$\boldsymbol{X}$表示的特征向量方程，我们得到

$$\frac{1}{N}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{c}_m = \lambda_m\boldsymbol{X}\boldsymbol{c}_m \tag{10.57}$$

而我们又恢复了数据协方差矩阵。这现在也意味着我们恢复了$\boldsymbol{X}\boldsymbol{c}_m$作为$\boldsymbol{S}$的一个特征向量。

备注。 如果我们想应用第10.6节中讨论的PCA算法，我们需要将$\boldsymbol{S}$的特征向量$\boldsymbol{X}\boldsymbol{c}$归一化$_m$，使其具有规范1。 ◆

10.6 实践中PCA的关键步骤

在下文中，我们将通过一个运行中的例子来了解PCA的各个步骤，其摘要见图10.11.我们得到了一个二维数据集（图10.11(a)），我们想用PCA将其投影到一个一维子空间。

1. **平均值减法** 我们首先通过计算数据的中心化来进行计算。

   数据集的平均数$\boldsymbol{\mu}$，并从每一个数据点中减去它。这可以确保数据集的平均值为0（图10.11（b））。严格来说，均值减去并不是必须的，但可以减少数字概率的风险。

   问题。

2. **标准化** 将数据点除以每个维度$d=1, \ldots$的数据集的标准偏差σ_d。现在数据是无单位的，它在1每个轴上都有方差，这在图10.11(c)中用两个箭头表示。这一步完成了数据的*标准化*。

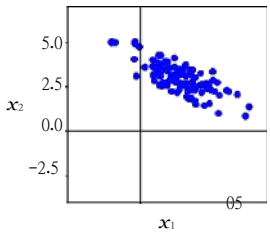
   标准化

3. **协方差矩阵的特征分解**计算

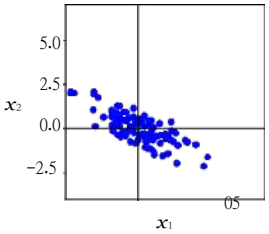
   数据协方差矩阵及其特征值和相应的特征向量。由于协方差矩阵是对

d)中，特征向量的比例是由共变矩阵的大小决定的。

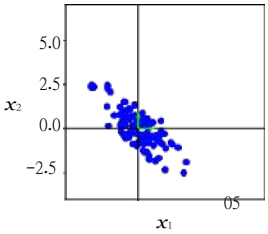
称的，光谱定理（Theorem 4.15）指出，我们可以找到一个特征向量的ONB。在图10.11(



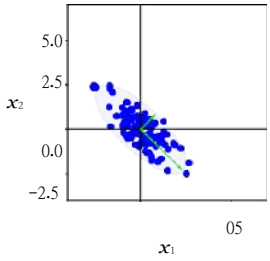
(a) 原始数据集。



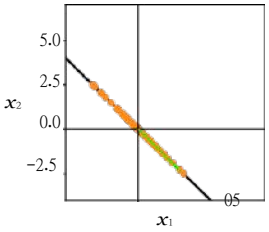
(b) 第1步：通过对每个数据点的平均数进行分折来进行居中。



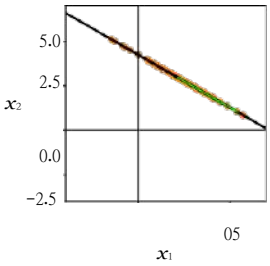
(c) 第二步：除以标准差，使数据无单位。数据沿1每个轴有差异。



(d) 第三步：计算数据协方差矩阵（椭圆）的特征值和特征向量（箭头）。



(e) 第四步：将数据投射到主子空间上。



(f) 撤销标准化，将投影数据移回(a)中的原始数据空间。

图表步骤10.11 的PCA。(a) 原始数据集。(b) 定心。(c) 除以标准差。(d) eigendecomposition; (e) projection; (f)映射回原始数据空间。

响应的特征值。较长的矢量横跨主子空间，我们用U表示。 数据协方差矩阵用椭圆表示。

4. **投影** 我们可以将任何数据点$\boldsymbol{x}_*\in$R投影^{D}到主子空间上。为了做到这一点，我们需要$_*$用第$\boldsymbol{d}$维的训练数据$_d$的平均值μ_d和标准差σ对$\boldsymbol{x}$进行标准化。

分别为，所以

$$x_*^{(d)} \leftarrow \frac{x_*^{(d)} - \mu_d}{\sigma_d} \quad ,d= 1, \ldots, D \,, \tag{10.58}$$

其中$x^{(d)}$是$\boldsymbol{x}$的第$\boldsymbol{d}$个分量。 我们得到投影为

$$\tilde{\boldsymbol{x}}_* = \boldsymbol{B}\boldsymbol{B}\boldsymbol{x}(_* \tag{10.59}$$

座标为

$$\boldsymbol{z}_* = \boldsymbol{B}\boldsymbol{x} \quad _* \tag{10.60}$$

关于主子空间的基础。这里，$\boldsymbol{B}$是包含与数据协方差矩阵的最大特征值相关的特征向量作为列的ma- trix。PCA返回的是坐标(10.60)，而不

是投影$\boldsymbol{x}_*$。

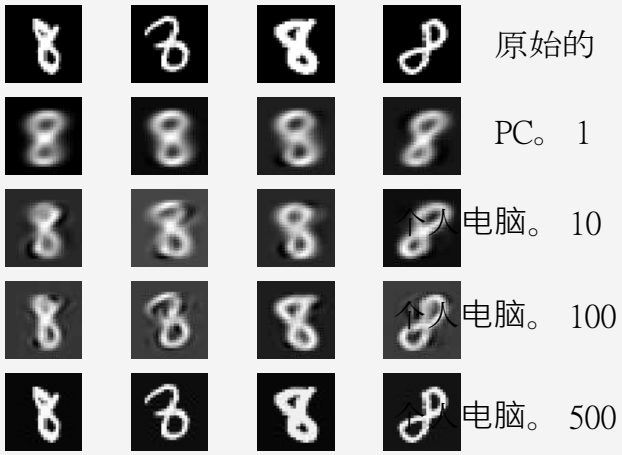
在对我们的数据集进行标准化后，(10.59)只能得到标准化数据集背景下的投影。为了得到我们在原始数据空间（即标准化之前）的投影，我们需要撤销标准化(10.58)，并在加入平均值之前乘以标准差，这样我们就可以得到

$$\tilde{x}_*^{(d)} \leftarrow \tilde{x}_*^{(d)} \sigma_d + \mu_d \quad d = 1, \ldots, D . \tag{10.61}$$

图10.11(f)说明了在原始数据空间的投影。

例子（10.4MNIST数字：重构）

在下文中，我们将把PCA应用于MNIST数字数据集，该数据集包含60*WWW*手写数字的000例子09。每个数字都是一个大小28为 ，28即包含784个像素的图像，因此我们可以将这个数据集中的每个图像解释为一个矢量$\boldsymbol{x} \in \mathbb{R}^{784}$。 10.3.



图：增加主成分的数量对重建的影响**10.12**。

为了说明问题，我们将PCA应用于MNIST数字的一个子集，并将重点放在数字 "8 "上。我们使用了数字 "8 "的训练5,389图像，并确定了本章所详述的主子空间。然后我们用学到的投影矩阵来重建一组测试图像，如图所示。10.12.图10.12的第一行显示了一组来自测试集的四个原始数字。下面几行显示的正是这些数字在使用尺寸1为 、10 、100 和500 的主子空间时的重建情况。我们看到，即使使用单维的主子空间，我们也能得到一个半成品的原始数字的重构，然而，它是模糊的和通用的。随着主成分（PC）数量的增加，重新构建的数字变得更加清晰，更多的细节被考虑在内。随着主500成分数量的增加

我们可以有效地获得一个接近完美的重建。如果我们选择784台电脑，我们将恢复准确的数字，而没有任何压缩损失。

图中10.13显示了平均重建误差的平方，它是

$$\frac{1}{N}\sum_{n=1}^{N}\|\boldsymbol{x}_n-\tilde{\boldsymbol{x}}_n\|^2=\sum_{i=M+1}^{D}\lambda_i\,, \tag{10.62}$$

作为主成分数量M的函数。我们可以看到，主成分的重要性迅速下降，增加更多的主成分只能获得微弱的收益。这与我们在图10.5中的观察完全吻合，我们在图中发现，预测数据的大部分方差只被几个主成分所占据。有了这些主成分550，我们基本上就可以完全重建训练数据了。
含有数字 "8 "的数据（边界周围的一些像素显示为
在整个数据集中没有变化，因为它们总是黑色的）。

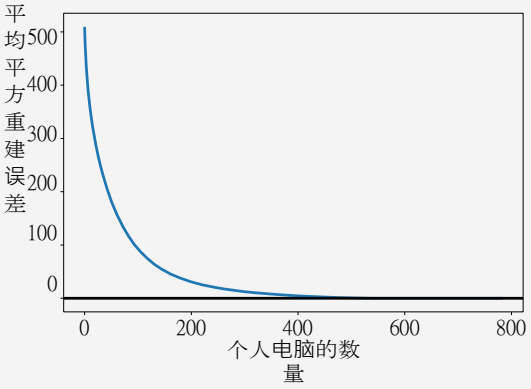


Figure 10.13 Average squared reconstruction error as a function of the number of principal components. The average squared reconstruction error is the sum of the eigenvalues in the orthogonal complement of the principal subspace.

10.7 潜在变量的视角

在前面的章节中，我们使用最大方差和投射观点，在没有任何概率模型概念的情况下推导出PCA。一方面，这种方法可能很吸引人，因为它允许我们避开所有与概率论有关的数学困难，但另一方面，概率论模型将为我们提供更多的灵活性和有用的见解。更具体地说，一个概率模型将

- 来一个似然函数，我们可以明确地处理噪声观测（我们之前甚至没有讨论过）。
- 允许我们通过边际似然进行贝叶斯模型比较，如第8.6节所述
- 将PCA看作是一个生成模型，它允许我们模拟新的数据

- 通过应用贝叶斯定理，允许我们与相关算法建立直接的联系 处理随机
- 缺失的数据维度
- 给我们一个新数据点的新颖性的概念
- 给我们一个原则性的方法来扩展模型，例如，扩展到PCA模型的混合物
- 让我们在前面的章节中得出的PCA作为一个特例
- 通过对模型参数进行边际化处理，允许进行完全的贝叶斯式处理

通过引入一个连续值的潜变量$\boldsymbol{z} \in \mathbb{R}^M$，有可能将PCA表述为一个概率潜变量模型。Tipping和Bishop（1999）提出了这种潜变量模型，即*概率PCA*（*PPCA*）。

概率PCA

PPCAPPCA解决了上述大部分问题，我们通过最大化投影空间的方差或最小化重建误差得到的PCA解决方案，是在无噪声环境下最大似然估计的特例。

10.7.1 生成过程和概率模型

在PPCA中，我们明确地写出了用于线性二重性降低的概率模型。为此，我们假设一个连续的潜在变量$\boldsymbol{z} \in \mathbb{R}^M$与标准正态先验$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$和线性关系的关系。潜在变量和观察到的$\boldsymbol{X}$数据之间的关系，其中

$$\boldsymbol{x} = \boldsymbol{B}\boldsymbol{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon} \in \mathbb{R}^D, \tag{10.63}$$

其中$\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$是高斯观测噪声，$\boldsymbol{B} \in \mathbb{R}^{D\times M}$和$\boldsymbol{\mu} \in \mathbb{R}^D$描述了从潜伏到观察的线性/阿芬映射。变量。因此，PPCA通过以下方式将潜在变量和观测变量联系起来

$$p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{B}, \boldsymbol{\mu}, \sigma^2) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{B}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2\boldsymbol{I}) \text{。} \tag{10.64}$$

总的来说，PPCA诱导了以下生成过程。

$$\boldsymbol{z}_n \sim \mathcal{N}(\boldsymbol{z} \mid \boldsymbol{0}, \boldsymbol{I}) \tag{10.65}$$

$$\boldsymbol{x}_n | \boldsymbol{z}_n \sim \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{B}\boldsymbol{z}_n + \boldsymbol{\mu}, \sigma^2\boldsymbol{I}) \tag{10.66}$$

为了产生一个给定模型参数的典型的数据点，我们遵循一个*祖先的抽样*方案。我们首先从$p(\boldsymbol{z})$中抽取一个潜变量$\boldsymbol{z}_n$。然后，我们用$\boldsymbol{z}_n$在(10.64)来抽取一个以被抽取的$\boldsymbol{z}_n$为条件的数据点，即$\boldsymbol{x}_n | p(\boldsymbol{x} | \boldsymbol{z}_n, \boldsymbol{B}, \boldsymbol{\mu}, \sigma^2)$。

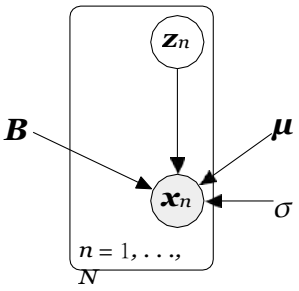
祖传采样

这个生成过程使我们能够将概率模型（即所有随机变量的联合分布；见第8.4节）写为

$$p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{B}, \boldsymbol{\mu}, \sigma^2) = p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{B}, \boldsymbol{\mu}, \sigma^2)\, p(\boldsymbol{z}) \text{，} \tag{10.67}$$

这 10.7 概 *潜在变量的视角*

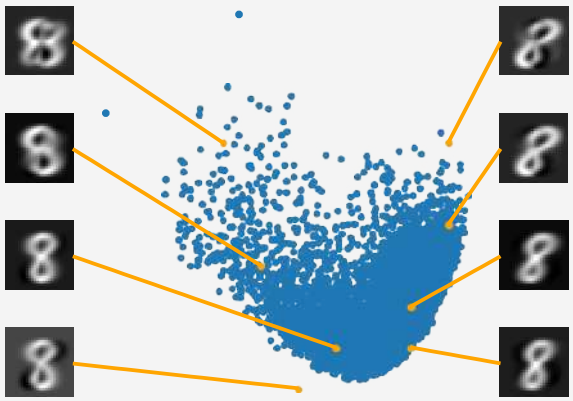
立 即 产 生 了 图 中 的 图 形 模 型 10.14 图 中 的 图 形 模 型 ， 使 用 第 8.5 节 的 结 果 。



图为概率PCA的图形**10.14**模型。观察结果 x_n 明确地取决于相应的潜在变量 $z_n \sim \mathcal{N}(0, I)$。σn模型参数 B,μ 和似然参数σ在整个数据集中是共享的。

备注。注意连接潜在变量$\boldsymbol{z}$和观察数据$\boldsymbol{x}$*的*箭头方向。箭头从$\boldsymbol{z}$指向$\boldsymbol{x}$，这意味着PPCA模型为高维观测数据$\boldsymbol{x}$*假设*了一个低维的潜在原因$\boldsymbol{z}$。最后，我们显然对在一些观测数据下找到关于$\boldsymbol{z}$*的*东西感兴趣。为了达到这个目的，我们将应用贝叶斯推理来 "反转 "这个箭头，从观察值到潜变量。 ◆

例子（10.5使用潜变量生成新数据）



图：生成**10.15**新的MNIST数字。潜在变量 z 可用于产生新的数据 $\tilde{x} \sim p(x|z)$。越是接近我们在训练数据上停留的时间越长，生成的数据就越真实。

图中10.15显示了使用二维主子空间时，通过PCA找到的MNIST数字 "8 "的潜在坐标（蓝点）。我们可以查询这个潜在空间中的任何向量$\boldsymbol{z}_*$，并生成一个图像$\boldsymbol{x}_* = \boldsymbol{B}\boldsymbol{z}_*$。我们展示了八个这样的生成图像与它们相应的潜在空间表示。根据我们查询潜在空间的位置，生成的图像看起来有所不同（形状、旋转、大小等）。如果我们远离训练数据进行查询，我们会看到越来越多的伪影，例如，左上角和右上角的数字。请注意，这些生成的图像的内在维度只有两个。

可能性不取决于潜伏变量 z.

10.7.2 概率和联合分布

利用第六章的结果，我们通过整合潜在变量$\boldsymbol{z}$（见第8.4.3节）来获得这个proba-bilistic模型的似然，因此

$$p(\boldsymbol{x} \mid \boldsymbol{B}, \boldsymbol{\mu}, \sigma) = \int p(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{B}, \boldsymbol{\mu}, \sigma) p(\boldsymbol{z}) d\boldsymbol{z} \quad (10.68a)$$

$$= \int \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{B}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) \mathcal{N}(\boldsymbol{z} \mid \boldsymbol{0}, \boldsymbol{I}) d\boldsymbol{z} 。 \quad (10.68b)$$

从第6.5节，我们知道这个积分的解是一个高斯分布，其平均值为

$$\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] = \mathbb{E}_{\boldsymbol{z}}[\boldsymbol{B}\boldsymbol{z} + \boldsymbol{\mu}] + \mathbb{E}_{\boldsymbol{\varepsilon}}[\boldsymbol{\varepsilon}] = \boldsymbol{\mu} \quad (10.69)$$

并具有协方差矩阵

$$\mathbb{V}[\boldsymbol{x}] = \mathbb{V}_{\boldsymbol{z}}[\boldsymbol{B}\boldsymbol{z} + \boldsymbol{\mu}] + \mathbb{V}_{\boldsymbol{\varepsilon}}[\boldsymbol{\varepsilon}] = \mathbb{V}_{\boldsymbol{z}}[\boldsymbol{B}\boldsymbol{z}] + \sigma^2 \boldsymbol{I} \quad (10.70a)。$$

$$= \mathbb{B}\mathbb{V}[\boldsymbol{z}\boldsymbol{z}]\boldsymbol{B}^\top + \sigma^2 \boldsymbol{I} = \boldsymbol{B}\boldsymbol{B}^\top + \sigma^2 \boldsymbol{I} 。 \quad (10.70b)$$

(10.68b)中的似然可以用于模型参数的最大似然或MAP估计。

备注。 我们不能使用(10.64)进行最大似然估计，因为它仍然取决于潜变量。我们对最大似然（或MAP）估计所要求的似然函数应该只是数据$\boldsymbol{x}$和模型参数的函数。

但必须不依赖于潜在的变量。 ◆

从第6.5节中，我们知道高斯随机变量$\boldsymbol{z}$和它的线性/affine变换$\boldsymbol{x}=\boldsymbol{B}\boldsymbol{z}$是共同的高斯离散。

分布的。我们已经知道边际$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z} \mid \boldsymbol{0}, \boldsymbol{I})$和$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{B}^\top + \sigma^2 \boldsymbol{I})$。缺失的交叉协方差给定为

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{z}] = \mathrm{Cov}_{\boldsymbol{z}}[\boldsymbol{B}\boldsymbol{z} + \boldsymbol{\mu}] = \boldsymbol{B}\,\mathrm{Cov}_{\boldsymbol{z}}[\boldsymbol{z}, \boldsymbol{z}] = \boldsymbol{B} 。 \quad (10.71)$$

因此，PPCA的概率模型，即潜伏随机变量和观察随机变量的联合分布明确地给出为

$$p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{B}, \boldsymbol{\mu}, \sigma) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{z}\end{bmatrix} \mid \begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{B}\boldsymbol{B}^\top + \sigma^2\boldsymbol{I} & \boldsymbol{B}\\ \boldsymbol{B}^\top & \boldsymbol{I}\end{bmatrix}\right), \quad (10.72)$$

平均值向量的长度为$D+M$，协方差矩阵的大小为$(d+m)\times(d+m)$。

10.7.3 后期分布

中的联合高斯分布$p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{B}, \boldsymbol{\mu}, \sigma^2)$10.72)中的联合高斯分布，我们可以通过应用以下公式立即确定后验分布$p(\boldsymbol{z} \mid \boldsymbol{x})$。

6.5.1节中的高斯条件规则。那么，给定一个观察值$\boldsymbol{x}$的潜变量的后验分布是

$$p(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{z} \mid \boldsymbol{m}, \boldsymbol{C}), \tag{10.73}$$

$$\boldsymbol{m} = \boldsymbol{B}^\top(\boldsymbol{B}\boldsymbol{B}^\top + \sigma^2\boldsymbol{I})^{-1}(\boldsymbol{x} - \boldsymbol{\mu}), \tag{10.74}$$

$$\boldsymbol{C} = \boldsymbol{I} - \boldsymbol{B}^\top(\boldsymbol{B}\boldsymbol{B}^\top + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{B}。 \tag{10.75}$$

请注意，后验协方差并不取决于观察数据。

$\boldsymbol{x}_*$.对于数据空间中的一个新观察点$\boldsymbol{x}$，我们用(10.73)来确定相应的潜伏变量$\boldsymbol{z}_*$的后验分布$p(\boldsymbol{z}_* \mid \boldsymbol{x}_*)$。共变矩阵$\boldsymbol{C}$使我们能够评估嵌入的信心如何。一个具有小行列式的协方差矩阵$\boldsymbol{C}$（衡量体积）告诉我们，潜伏嵌入$\boldsymbol{z}_*$是相当确定的。如果我们得到一个具有很大方差的后验分布$p(\boldsymbol{z}_* \mid \boldsymbol{x}_*)$，我们可能会面临一个异常点。然而，我们可以探索这个后验分布，以了解在这个后验下还有哪些数据点$\boldsymbol{x}$是可信的。为了做到这一点，我们利用了PPCA的生成过程，它允许我们通过生成在这个后验下合理的新数据来探索潜在变量的后验分布。

1.从潜在变量的后验分布中抽出一个潜在变量$\boldsymbol{z}_* \sim p(\boldsymbol{z} \mid \boldsymbol{x}_*)$(10.73).

2.对重建的向量进行采样 $\tilde{\boldsymbol{x}}_* \sim p(\boldsymbol{x} \mid \boldsymbol{z}_*, \boldsymbol{B}, \boldsymbol{\mu}, \sigma^2)$ 来自(10.64).

如果我们多次重复这个过程，我们可以探索潜变量z的后验分布(10.73)上的潜在变量$\boldsymbol{z}_*$及其对观察数据的影响。抽样过程有效地假设了数据，这在后验分布下是可信的。

10.8 进一步阅读

我们从两个角度得出PCA。(a) 最大化投影空间的方差；(b) 最小化平均重建误差。然而，PCA也可以从不同的角度进行解释。让我们回顾一下我们所做的事情。我们把高维数据$\boldsymbol{x} \in \mathbb{R}^D$，用一个矩阵$\boldsymbol{B}$来找到低维的表示$\boldsymbol{z} \in \mathbb{R}^M$。$\boldsymbol{B}$的列是数据协方差矩阵$\boldsymbol{S}$的特征向量，与最大的特征值有关。一旦我们有了低维表征$\boldsymbol{z}$，我们就可以得到它的高维版本（在原始的的数据空间）为$\boldsymbol{x}$ $\tilde{\boldsymbol{x}} = \boldsymbol{B}\boldsymbol{z} = \boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{x} \in \mathbb{R}^D$，其中$\boldsymbol{B}\boldsymbol{B}^\top$是一个投影矩阵。

$\approx$

我们也可以把PCA看作是一个线性*自动编码器*，如图-自动编码器所示自动编码器将数据$\boldsymbol{x} \in \mathbb{R}^D$编码为一个*代码*$\boldsymbol{z} \in \mathbb{R}^M$。$n$并将其解码为与$\boldsymbol{x}$相似的$\tilde{\boldsymbol{x}}_n \in \mathbb{R}^D$。从数据映射到编码被称为*编码器*，而从编码到原编码的映射则被称为编码器。

在最终的数据空间中，被称为*解码器*。如果我们考虑线性映射，其中 解码器

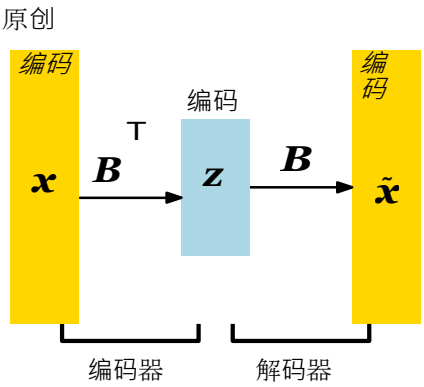


图10.16 PCA可以被看作是一个线性自动编码器。它编码了高维的数据：变成一个低维表示法(代码) $\boldsymbol{z} \in \mathbb{R}^M$和使用解码器进行解码。解码后的向量$\tilde{\boldsymbol{x}}$是正交投影的原始数据到的$M$维主子空间。

编码由$\boldsymbol{z}_n = \boldsymbol{B}\boldsymbol{x}_n$给出，我们感兴趣的是最小化数据$\boldsymbol{x}_n$和其重建$\tilde{\boldsymbol{x}}_n = \boldsymbol{B}\boldsymbol{z}$之间的平均平方误差，$n = 1, \ldots, N$，我们得到

$$\frac{1}{N}\sum_{n=1}^{N}\|\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n\|^2 = \frac{1}{N}\sum_{n=1}^{N}\|\boldsymbol{x}_n - \boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{x}_n\|^2 . \tag{10.76}$$

这意味着我们最终得到的目标函数与(10.29)中的目标函数，我们在章节中讨论过10.3因此，当我们最小化自动编码损失的平方时，我们会得到PCA解决方案。如果我们用非线性映射取代PCA的线性映射，我们会得到一个非线性自动编码器。一个突出的例子是深度自动编码器，其中的线性函数被深度神经网络取代。在这种情况下，编码器也被称为*识别网络*或*推理网络*，而解码器也被称为*生成器*。

识别网络

推理网络
生成器

PCA的另一种解释与信息论有关。我们可以把代码看作是原始数据点的一个较小或压缩的版本。当我们使用代码重建我们的原始数据时，我们并没有得到准确的数据点，而是得到一个稍微扭曲或有噪音的版本的。这意味着，我们的压缩是 "有损的"。直观地说，我们希望最大限度地提高原始数据和低维代码之间的相关性。更正式地说，这与相互信息有关。然后，我们将得到与我们在第1节中讨论的PCA相同的解决方案。10.3通过最大化相互信息，即信息理论中的一个核心概念（MacKay,2003），我们将得到与我们在第二节中讨论的PCA解决方案相同的解决方案。

该代码是原始数据的压缩版本。

在我们关于PPCA的讨论中，我们假设模型的参数，即$\boldsymbol{B}$、$\boldsymbol{\mu}$和似然参数σ^2是已知的。Tipping和Bishop(1999)描述了如何在PPCA设置中得出这

些参数的最大似然估计（注意我们在本章中使用了不同的符号）。最大似然参数，当pro-

将D维的数据探测到M维的子空间，是

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n, \tag{10.77}$$

$$\boldsymbol{B}_{\mathrm{ML}} = \boldsymbol{T}(\boldsymbol{\Lambda} - \sigma^2\boldsymbol{I})^{\frac{1}{2}}\boldsymbol{R} \tag{10.78}$$

。

$$\sigma^2_{\mathrm{ML}} = \frac{1}{D-M}\sum_{j=M+1}^{D} \lambda_j, \tag{10.79}$$

其中$\boldsymbol{T} \in \mathbb{R}^{D\times M}$包含数据协方差矩阵的$M$个特征向量，$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_M) \in \mathbb{R}^{M\times M}$是一个对角线矩阵，其对角线上有与主轴相关的特征值，$\boldsymbol{R} \in \mathbb{R}^{M\times M}$是一个任意的正交矩阵。最大似然解$\boldsymbol{B}_{\mathrm{ML}}$在任意正交变换下是唯一的，例如，我们可以将$\boldsymbol{B}_{\mathrm{ML}}$与任何旋转矩阵$\boldsymbol{R}$进行右乘，这样(10.78)本质上是一个奇异值分解（见第4.5节）。Tipping和Bishop(1999)给出了一个证明的概要。

在(10.78)被保证为正半无限，因为数据协方差矩阵的最小特征值被噪声方差σ自下而上约束[2]。

中给出的$\boldsymbol{\mu}$的最大似然估计值(10.77)是数据的样本平均值。( ) 中给出的观测噪声方差σ^2的最大似然估计值是主子空间正交补数的平均方差。10.79)是主子空间的正交补数中的平均方差，即平均遗留方差。

我们无法用前M个主成分捕捉到的信息被视为观察噪声。

在σ0的无噪声限制下，PPCA和PCA提供了相同的解决方案。由于数据协方差矩阵$\boldsymbol{S}$是对称的，它可以被二元化（见第4.4节），也就是说，存在一个$\boldsymbol{S}$的特征向量矩阵$\boldsymbol{T}$，以便

$$\boldsymbol{S} = \boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}^{-1}。 \tag{10.80}$$

在PPCA模型中，数据协方差矩阵是高斯似然$p(\boldsymbol{x}\,|\,\boldsymbol{B}, \boldsymbol{\mu}, \sigma^2)$的协方差矩阵，它是$\boldsymbol{B}\boldsymbol{B}^\top + \sigma^2\boldsymbol{I}$，见（10.70b）。对于$\sigma$0，我们得到$\boldsymbol{B}\boldsymbol{B}^\top$，所以这个数据协方差必须等于PCA数据协方差（以及它的因子化，在(10.80)，所以

$$\mathrm{Cov}[\boldsymbol{X}] = \boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}^{-1} = \boldsymbol{B}\boldsymbol{B}^\top \Leftrightarrow \boldsymbol{B} = \boldsymbol{T}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{R}, \tag{10.81}$$

即，我们在( )中得到最大似然估计。10.78)为σ=0。 从(10.78)和(10.80)，很明显，(P)PCA对数据协方差矩阵进行了一次去构成。

在流媒体环境中，数据按顺序到达，建议使用迭代期望最大化（EM）算法进行最大似然估计（Roweis，1998）。

为了确定潜在变量的维度（代码的长度，我们将数据投射到的低维子空间的维度），Gavish和Donoho（2014）提出了一个启发式方法，即如果

我们可以指定数据的噪声方差σ^2，我们应该

抛弃所有小于$\sqrt{\frac{4\sigma^2}{D}}$的特征值。另外，我们可以使用(嵌套）交叉验证（第8.6.1节）或贝叶斯模型选择方法（在第8.6.2节中讨论）来确定对数据内在维度的良好估计（Minka，2001b）。

Bayesian PCA

与我们在第九章中关于线性回归的讨论类似，我们可以在模型的参数上放置一个先验分布，并将它们整合出来。通过这样做，我们(a)避免了参数的点估计和这些点估计带来的问题（见第8.6节），(b)自动选择潜空间的适当维度M。在Bishop(1999)提出的*贝叶斯PCA*中，模型参数被置于先验$p(\boldsymbol{\mu}, \boldsymbol{B}, \sigma^2)$上。生成过程允许我们将模型参数整合出来，而不是对它们进行调节，这就解决了过拟合问题。由于这种整合在分析上是难以实现的，Bishop(1999)建议使用近似推理方法，如MCMC或变量推理。关于这些近似推理技术的更多细节，我们参考Gilks等人（1996）和Blei等人（2017）的工作。

因素分析

在PPCA中，我们考虑了线性模型$p(\boldsymbol{x}_n | \boldsymbol{z}_n) = \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{B}\boldsymbol{z}_n + \boldsymbol{\mu}, \sigma^2\boldsymbol{I})$，先验$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$，其中所有观测维受相同数量的噪声影响。如果我们允许每个观察维度d有不同的方差σ_d^2，我们就会得到*因子分析*（FA）（Spearman,1904;Bartholomew等人，2011）。这意味着FA比PPCA给了似然更多的灵活性，但仍然迫使数据被模型参数$\boldsymbol{B}$、$\boldsymbol{\mu}$所解释。然而，FA不再允许有封闭式的最大似然解，因此我们需要使用迭代方案，如期望最大化算法，来估计模型参数。虽然在PPCA中，所有的站位点都是全局最优，但这在FA中不再成立。与PPCA相比，如果我们缩放数据，FA不会改变，但如果我们旋转数据，它就会返回不同的解决方案。

一个过于灵活的可能性将能够解释更多的问题，而不仅仅是噪音。

一种与PCA密切相关的算法是*独立的com---*。

独立的*分量分析*（*ICA*）（Hyvarinen等人，2001）。再次从ICAlatent-variable perspective $p(\boldsymbol{x}_n|\boldsymbol{z}_n) = \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{B}\boldsymbol{z}_n + \boldsymbol{\mu}, \sigma^2\boldsymbol{I})$我们现在把$\boldsymbol{z}_n$的先验改为非高斯分布。ICA可以被用来

盲源隔离

用于*盲源分离*。想象一下，你在一个繁忙的火车站，有许多人说话。你的耳朵扮演着麦克风的角色，他们在火车站线性地混合不同的语音信号。盲源分离的目标是识别混合信号的组成部分。正如之前在PPCA的最大似然估计中所讨论的，原始PCA解决方案对任何旋转都是不变的。因此，PCA可以识别信号所在的最佳低维子空间，但不能识别

10.8进一步阅读 信号本身（Murphy，2012）。ICA通过修改潜源的先验分布$p(\boldsymbol{z})$来解决这个问题

来要求非高斯预设$p(\boldsymbol{z})$。关于ICA的更多细节，我们可以参考Hyvarinen等人（2001）和Murphy（2012）的书。

PCA、因子分析和ICA是线性模型降维的三个例子。Cunningham和Ghahramani（2015）对线性降维进行了更广泛的调查。

我们在这里讨论的(P)PCA模型允许几个重要的例外情况。在第10.5,我们解释了当输入维度D明显大于数据点的数量N时如何进行PCA。通过利用PCA可以通过计算(许多)内积来进行的见解，这个想法可以被推到极端，考虑到
形成无穷大的特征。*内核技巧*是kernelkernel技巧的基础
*PCA*允许我们隐含地计算无限维度特征之间的内积（Sch ölkopf等人， 1998；

内核PCA

Sch ölkopf和Smola， 2002）。有一些非线性降维技术是可以去掉的。从PCA的角度来看（Burges（2010）提供了一个很好的概述）。我们在本节之前讨论的PCA的自动编码器观点
可用于将PCA作为*深度自动编码器*的一个特例。在深度自动编码器中
深度自动编码器，编码器和解码器都由多层前馈神经网络表示，它们本身就是非线性映射。如果我们将这些神经网络中的激活函数设置为身份，那么该模型就等同于PCA。一种不同的方法来
非线性降维是*高斯过程*的*潜变*高斯过程
Lawrence(2005)提出的*GP-LVM模型*（*GP-LVM*）。GP-LVM从我们用来推导PPCA的潜变量角度出发，用高斯过程（GP）取代了潜变量$\boldsymbol{z}$和观测值$\boldsymbol{x}$之间的线性关系。GP-LVM不是估计映射的参数（就像我们在PPCA中做的那样），而是将模型参数边际化，对潜变量$\boldsymbol{z}$进行点估计。

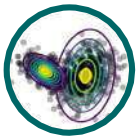
潜在变量模型
GP-LVM

对贝叶斯PCA，Titsias和Lawrence
*LVM*贝叶斯GP-LVM。
(2010)在潜变量$\boldsymbol{z}$上保持一个分布，并使用近似推理将它们也整合出来。

提出的*贝叶斯GP-*

11

用高斯混合模型进行密度估计



在前面几章中，我们已经介绍了机器学习中的两个基本问题：回归（第9章）和降维（第8章）。10).在本章中，我们将看看机器学习的第三个支柱：密度估计。在我们的旅程中，我们将介绍一些重要的概念，如期望最大化（EM）算法和混合模型的密度估计的潜在变量观点。

当我们将机器学习应用于数据时，我们通常旨在以某种方式表示数据。一个直接的方法是将数据点本身作为数据的代表；见图11.1为例。然而，如果数据集很庞大，或者我们对表示数据的特征感兴趣，这种方法可能是无益的。在密度计算中，我们用一个参数族的密度来紧凑地表示数据，例如高斯或贝塔分布。例如，我们可能正在寻找一个数据集的平均数和方差，以便用高斯分布来紧凑地表示数据。平均数和方差可以使用我们在第8.3节中讨论的工具找到：最大似然或最大后验估计。然后，我们可以用这个高斯的均值和方差来表示数据的基础分布，也就是说，我们认为数据集是这个分布的一个典型实现，如果我们从这个分布中取样的话。

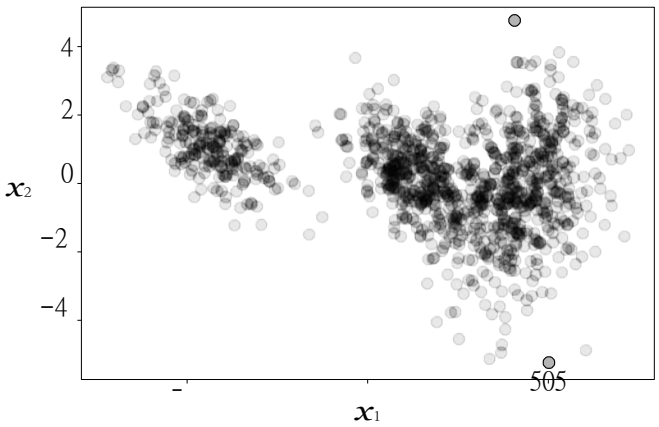


图 11.1
无法用高斯表示的二维数据集。

在实践中，高斯分布（或类似于我们迄今为止所认识的所有其他分布）的建模能力是有限的。例如，高斯对产生图中数据的密度的近似值11.1将是一个糟糕的近似值。在下文中，我们将研究一个更有代表性的分布系列，我们可以用它来进行密度估计。

混合模型。 混合模型

混合模型可用于通过K个简单（基础）分布的凸组合来描述一个分布$p(\boldsymbol{x})$。

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k p_k(\boldsymbol{x}) \tag{11.1}$$

$$0 \leqslant \pi_k \leqslant 1, \quad \sum_{k=1}^{K} \pi_k = 1, \tag{11.2}$$

其中成分p_k是一个基本分布系列的成员。

例如，高斯、伯努利或伽马，而π_k是*混合物权重*。 混合物权重

混合模型比相应的基数分布更具表现力，因为它们允许多模态的数据表示，即它们可以描述具有多个 "集群 "的数据集，如图中的例子。11.1.

我们将专注于高斯混合模型（GMM），其中基本分布是高斯的。对于一个给定的数据集，我们的目标是使模型参数的似然性最大化，以训练GMM。为此，我们将使用第五章、第六章和第7.2节的结果。然而，与我们之前讨论的其他应用（线性回归或PCA）不同，我们不会找到一个封闭形式的最大似然解。相反，我们将得出一组依赖性的同步方程，我们只能通过迭代来解决这些问题。

11.1 高斯混合模型

*高斯混合模型*是一种密度模型，我们将有限合模型相结合。 高斯混合模型与高斯混合模型

数量为K的高斯分布$\mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$所以

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \tag{11.3}$$

$$0 \leqslant \pi_k \leqslant 1, \quad \sum_{k=1}^{K} \pi_k = 1, \tag{11.4}$$

其中我们定义$\boldsymbol{\theta} := \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k : k = 1, \ldots, K\}$作为模型的所有参数的集合。这个高斯分布的凸形组合是由高斯分布和高斯分布组成的。

与简单的高斯分布（我们从（）中恢复）相比，该分布为我们提供了更

大的灵活性，使我们能够对复杂的densi- ties进行建模。11.3)1。图中给出了一个说明11.2,显示了加权的

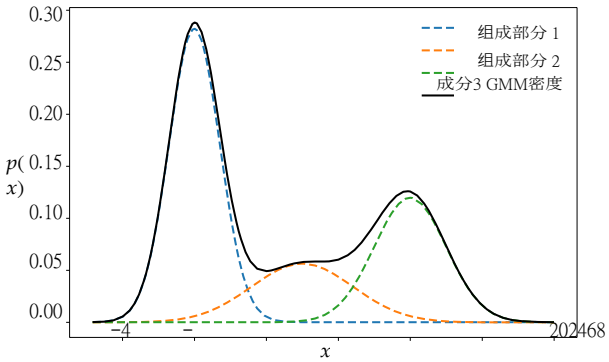


图 高斯11.2混合模型。高斯混合分布（黑色）是由高斯分布的凸组合组成的，比任何单独的成分都更有表现力。虚线代表加权的高斯成分。

成分和混合物的密度，它被认为是

$$p(x \mid \boldsymbol{\theta}) = 0.5\,\mathcal{N}\left(x \mid -2, \tfrac{1}{2}\right) + 0.2\,\mathcal{N}\left(x \mid 1, 2\right) + 0.3\,\mathcal{N}\left(x \mid 4, 1\right) \quad (11.5)$$

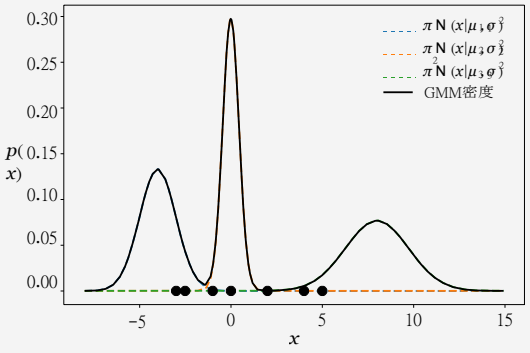
11.2 通过最大似然法学习参数

假设我们得到一个数据集 $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$，其中$\boldsymbol{x}_n, n = 1, \ldots, N$，是从一个未知的分布$p(\boldsymbol{x})$中提取的i.i.d。我们的目标是为这个未知的分布找到一个很好的近似/代表。

通过一个具有K个混合成分的GMM，我们可以得到分布$p(\boldsymbol{x})$。GMM的参数是K个均值$\boldsymbol{\mu}_k$，协方差$\boldsymbol{\Sigma}_k$，以及混合物权重π_k。

$$\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k : k = 1, \ldots, K\}.$$

示例（11.1初始设置）



图初始11.3设置。GMM（黑色）有三个混合成分（虚线）和七个数据点（圆盘）。

在本章中，我们将有一个简单的运行实例，帮助我们说明和形象化重要的概念。

我们考虑一个一维数据集$X = \{-3, -2.5, -1, 0, 2, 4, 5\}$由7个数据点组成，希望找到一个具有$K=3$个分量的GMM，对数据的密度进行建模。我们将混合成分初始化为

$$p_1(x) = \mathcal{N}(x \mid -4, 1) \tag{11.6}$$

$$p_2(x) = \mathcal{N}(x \mid 0, 0.2) \tag{11.7}$$

$$p_3(x) = \mathcal{N}(x \mid 8, 3) \tag{11.8}$$

并给它们分配相等的权重$\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$。相应的模型（和数据点）见图11.3.

在下文中，我们将详细介绍如何获得模型参数$\boldsymbol{\theta}_{\mathrm{ML}}$的最大似然值。我们首先写下似然值，即给定参数的训练数据的预测分布。我们利用我们的i.i.d.假设，这导致了因子化的似然。

$$p(X \mid \boldsymbol{\theta}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n \mid \boldsymbol{\theta}) \quad , p(\boldsymbol{x}_n \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{11.9}$$

其中每个单独的似然项$p(\boldsymbol{x}_n \mid \boldsymbol{\theta})$是高斯混合密度。然后我们得到对数似然为

$$\log p(X \mid \boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n \mid \boldsymbol{\theta}) = \sum_{n=1}^{N} \log \underbrace{\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{=L} . \tag{11.10}$$

我们的目标是找到参数$\boldsymbol{\theta}^*_{\mathrm{ML}}$，使定义在(11.10).我们的 "正常 "程序是计算L相对于模型参数$\boldsymbol{\theta}$的对数似然的梯度$\mathrm{d}L/\mathrm{d}\boldsymbol{\theta}$，设置为$0$，对于最大的然而，与我们之前的例子不同的是，对于$\boldsymbol{\theta}$的估计（例如，当我们讨论线性回归的时候，我们就已经讨论过了）。第9.2节），我们无法得到一个封闭式的解决方案。然而，我们可以利用一个迭代方案来找到好的模型参数$\boldsymbol{\theta}_{\mathrm{ML}}$，这将是GMMs的EM算法。其关键思想是每次更新一个模型参数，同时保持其他参数固定。

*备注。*如果我们把单一的高斯看作是所需的密度，那么在(11.10)中的和消失，对数可以直接应用于高斯分量，这样我们就可以得到

$$\log \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log \det(\boldsymbol{\Sigma}) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^{\top} \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) 。 \tag{11.11}$$

这种简单的形式使我们能够找到$\boldsymbol{\mu}$和$\boldsymbol{\Sigma}$的闭合形式的最大似然估计，如第八章所讨论的。在(11.10)中，我们不能移动

的对数，所以我们无法得到一个简单的封闭式的最大似然解。 ◆

一个函数的任何局部最优都表现出这样的特性：它相对于参数的梯度必须消失（必要条件）；见第七章。在我们的例子中，我们得到以下必要条件，当

我们对( )中的对数似然进行优化。11.10)中的对数似然值与GMM参数$\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, π_k有关。

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_k} = \mathbf{0} \Leftrightarrow \sum_{n=1}^{N} \frac{\partial \log p(\boldsymbol{x}_n | \boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} = , \quad (\mathbf{0}11.12)$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}_k} = \mathbf{0} \Leftrightarrow \sum_{n=1}^{N} \frac{\partial \log p(\boldsymbol{x}_n | \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} = , \quad (\mathbf{0}11.13)$$

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = 0 \Leftrightarrow \sum_{n=1}^{N} \frac{\partial \log p(\boldsymbol{x}_n | \boldsymbol{\theta})}{\partial \pi_k} = . \quad 0(11.14)$$

对于所有三个必要条件，通过应用链式规则（见第5.2.2节），我们需要形式为的偏导数

$$\frac{\partial \text{的} \log p(\boldsymbol{x}_n | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{p(\boldsymbol{x}_n | \boldsymbol{\theta})} \frac{\partial p(_{xn} | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \quad (11.15)$$

其中$\boldsymbol{\theta}=\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k, k=1, \ldots, K\}$是模型参数，而

$$\frac{1}{p(\boldsymbol{x}_n | \boldsymbol{\theta})} = \frac{1}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \quad (11.16)$$

在下文中，我们将计算偏导数(11.12)到(11.14).但在这之前，我们要介绍一个将在本章剩余部分发挥核心作用的量：责任。

11.2.1 职责

我们定义数量

$$r_{nk} := \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \quad (11.17)$$

责任

作为第n个数据点的第k个混合成分的*责任*。第k个混合成分对数据点$\boldsymbol{x}_n$的责任r_{nk}与似然率成正比。

$$p(\boldsymbol{x}_n | \pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (11.18)$$

给定数据点的混合成分。因此，混合组份

$\boldsymbol{r}^n$ 遵循一个玻尔兹曼/吉布斯分布。

当数据点可能是来自该混合成分的合理样本时，该成分对数据点有很

高的责任。请注意，$\boldsymbol{r}_n := [r_{n1}, \ldots, r_{nK}] \in \mathbb{R}^K$ 是一个（归一化）概率向量，即。

$\sum_k r_{nk}=1$，$r_{nk}\geq 0$。这个概率向量在K个混合成分中分配概率质量，我们可以把$\boldsymbol{r}_n$看作是一个

将$\boldsymbol{x}_n$"软分配"给K个混合成分。因此，再

响应性r_{nk}来自(11.17)表示$\boldsymbol{x}_n$是由第k个混合成分产生的概率。

--责任 r_{nk}是第k个混合成分产生第n个数据点的概率。

例子(11.2责任)

对于我们图中11.3,的例子，我们计算了责任r_{nk}

$$\begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 \\ 0.057 & 0.943 & 0.0 \\ 0.001 & 0.999 & 0.0 \\ 0.0 & 0.066 & 0.934 \\ 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix} \in^{RN\times K} \tag{11.19}$$

这里第n行告诉我们所有混合成分对x的$_n$责任。一个数据点的所有K责任的总和（每一行的总和）是 1。第k列给我们提供了第k个混合成分的责任概况。我们可以看到，第三个混合成分（第三列）对前四个数据点都不负责，但

承担了其余数据点的大部分责任。一列的所有条目之和为我们提供了数值N_k，即第k个混合成分的总责任。在我们的例子中，我们得到$N_1 = 2.058$, $N_2 = 2.008$, $N_3 = 2.934$

将看到，更新方程都取决于责任，这使得一个闭合形式的解

因此，对最大似然估计问题的处理是不可能的。然而，对于给定的责任，我们将一次更新一个模型参数，同时保持其他参数固定。在这之后，我们将重新计算责任。这两个步骤的迭代最终会收敛到一个最优化，是EM算法的一个具体实例。我们将在第二节中更详细地讨论这个问题。11.3.

11.2.2 更新手段

定理（11.1GMM均值的更新）。**GMM***的平均参数*$\boldsymbol{\mu}_k$, $k = 1, \ldots, K$, *GMM的均值参数的更新由以下公式给出*

$$\boldsymbol{\mu}_k^{新} = \frac{\sum_{n=1}^{N} r_{nk}\boldsymbol{x}_n}{\sum_{n=1}^{N} r_{nk}}, \tag{11.20}$$

其中责任r_{nk}*定义在*(11.17).

备注。()中各个混合成分的均值$\boldsymbol{\mu}_k$的更新取决于所有的均值、协方差矩阵Σ和混合权重π。11.20)中$_{nk}$的各个均值、协方差矩阵$\boldsymbol{\Sigma}_k$和混合体权重π_k，都取决于(11.17). 因此，我们无法一次获得所有$\boldsymbol{\mu}_k$的闭合式解决方案。 ◆

证明 从(11.15)，我们可以看到，关于平均参数$\boldsymbol{\mu}_k$的对数似然的梯度，$k = 1, ..., K$，需要我们计算部分导数

$$\frac{\partial p(\boldsymbol{x}_n \mid \boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} = \sum_{j=1}^{K} \pi_j \frac{\partial \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\partial \boldsymbol{\mu}_k} = \pi_k \frac{\partial \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}_k} \tag{11.21a}$$

$$= \pi_k (\boldsymbol{x}_n - \boldsymbol{\mu}_k) \boldsymbol{\Sigma}_k^{-1} \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{11.21b}$$

其中，我们利用了只有第k个混合成分取决于$\boldsymbol{\mu}_k$的情况。

我们用(11.21b)中的结果(11.15)中的结果，并把所有的东西放在一起，这样，所需的L相对于$\boldsymbol{\mu}_k$的偏导就得到了

$$\frac{\partial L}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} \frac{\partial \log p(\boldsymbol{x}_n \mid \boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} \frac{1}{p(\boldsymbol{x}_n \mid \boldsymbol{\theta})} \frac{\partial p(\boldsymbol{x}_n \mid \boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} \tag{11.22a}$$

$$= \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu}_k) \boldsymbol{\Sigma}_k^{-1} \underbrace{\left(\frac{\pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \right)}_{= r_{nk}} \tag{11.22b}$$

$$= \sum_{n=1}^{N} r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k) \boldsymbol{\Sigma}_k^{-1} . \tag{11.22c}$$

这里我们用(11.16)和(11.21b)中的部分推导结果来得到(11.22b)。数值r_{nk}是我们在（）中定义的责任。11.17).

我们现在求解(11.22c)中的$\boldsymbol{\mu}_k^{\text{new}}$，使$\frac{\partial L(\boldsymbol{\mu}_k^{\text{new}})}{\partial \boldsymbol{\mu}_k} = \mathbf{0}$得到

$$\sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n = \sum_{n=1}^{N} r_{nk} \boldsymbol{\mu}_k^{\text{新}} \iff \boldsymbol{\mu}_k^{\text{new}} = \frac{\sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n}{\sum_{n=1}^{N} r_{nk}} = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n , \tag{11.23}$$

其中我们定义了

$$N_k := \sum_{n=1}^{N} r_{nk} \tag{11.24}$$

作为整个数据集的第k个混合成分的总责任。定理的证明到此结束11.1. □

直观地说，(11.20)可以解释为对平均值的重要性加权的蒙特卡洛估计

356 ，其中数据点$\boldsymbol{x}$的重要性权重高斯混合模型分类的责任r_{nk}，$k=1,\ldots,K$.

因此，平均数$\boldsymbol{\mu}_k$被拉向一个数据点$\boldsymbol{x}_n$，其强度如图
平均值被拉向相应的混合成分具有高责任的数据点，即高可能性r_{nk}。图11.4说明了这一点。我们还可以把(11.20)中的平均值解释为所有数据点在分布图下的期望值。
命名为

$$\boldsymbol{r}_k := [r_{1k}, \ldots, r_{Nk}] / N_k, \tag{11.25}$$

这是个归一化的概率向量，即。

$$\boldsymbol{\mu}_k \leftarrow \sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n \tag{11.26}$$

更新**11.4**。
在GMM中混合成分的平均参数。
均值是指
拔向个别数据点，用所给的权重。
相应的责任。

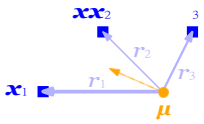
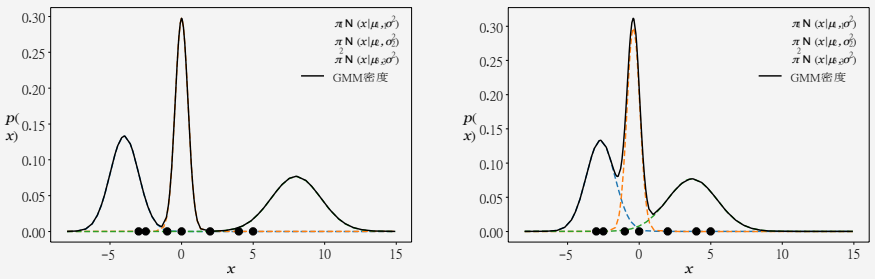


图11.5 在GMM中更新均值的效果。
(a)GMM
(b)更新均值μ后的GMM，$_k$同时保留方差和混合权重。

例子（11.3平均更新）



(a)GMM密度和各个组成部分 (b)更新平均值之前的GMM密度和各个组成部分 。更新平均值 之后。

在我们图中11.3,的例子中，均值被更新为以下几点。

$$\mu_1 :- 4 \rightarrow -2.7 \tag{11.27}$$
$$\mu_2 : \rightarrow 0 \ -0.4 \tag{11.28}$$
$$\mu_3 : \rightarrow 83 \ .7 \tag{11.29}$$

在这里我们看到，第一和第三种混合成分的平均值向数据的体系移动，而第二种成分的平均值则没有如此大的变化。图11.5说明了这种变化，图11.5(a)显示了更新平均值之前的GMM密度，图11.5(b)显示了更新平均值μ_k之后的GMM密度。

$\pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j$为所有$j = 1, \ldots, K$，这样，( )中的更新取决于11.20)中的更新取决于在GMM的所有参数上，有一个闭合形式的解决方案，我们发现
第9.2节中的线性回归或第10章中的PCA所保留的数据，无法获得。

11.2.3 更新协方差

定理（11.2GMM协方差的更新）。 *共同方差参数$\boldsymbol{\Sigma}_k$的更新，$k=1,\ldots,$ GMM的K的更新由以下公式给出*

$$\boldsymbol{\Sigma}_k^{new}=\frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top \tag{11.30}$$

其中r_{nk}和N_k的定义见(11.17)和(11.24)，分别定义。

证明 为了证明定理11.2,我们的方法是计算对数似然相对于协方差$\boldsymbol{\Sigma}_k$的偏导，将其设为$\mathbf{0}$，并求解$\boldsymbol{\Sigma}_k$。我们从我们的一般方法开始

$$\frac{\partial L}{\partial \boldsymbol{\Sigma}_k}=\sum_{n=1}^{N}\frac{\partial \log p(\boldsymbol{x}_n|\boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k}=\sum_{n=1}^{N}\frac{1}{p(\boldsymbol{x}_n|\boldsymbol{\theta})}\frac{\partial p(\boldsymbol{x}_n|\boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k}。 \tag{11.31}$$

我们已经知道$p(\boldsymbol{x}_n|\boldsymbol{\theta})$来自(11.16).为了得到剩余的部分导数$\partial p(\boldsymbol{x}_n|\boldsymbol{\theta})/\partial \boldsymbol{\Sigma}_k$，我们写下高斯分布$p(\boldsymbol{x}_n|\boldsymbol{\theta})$的定义（见(11.9)），并放弃除第$k$项之外的所有项。我们
然后得到

$$\frac{\partial p(\boldsymbol{x}_n|\boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} \tag{11.32a}$$

$$=\frac{\partial}{\partial \boldsymbol{\Sigma}_k}\pi_k(2\pi)^{-\frac{D}{2}}\det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\right) \tag{11.32b}$$

$$=\pi_k(2\pi)^{-\frac{D}{2}}\frac{\partial}{\partial \boldsymbol{\Sigma}_k}\det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\right)$$
$$+\det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}}\frac{\partial}{\partial \boldsymbol{\Sigma}_k}\exp\left(-\frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\right). \tag{11.32c}$$

我们现在使用的身份是

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_k}\det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}}=-\frac{1}{2}\det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}}\boldsymbol{\Sigma}_k^{-1}, \tag{11.33}$$

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_k}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)=-\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1} \tag{11.34}$$

并得到(经过一些重排)所需的偏导，在(11.31)中要求的偏导为

$$\frac{\partial p(\boldsymbol{x}_n|\boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k}=\pi_k\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$$

$$-\left(\frac{1}{2}\boldsymbol{\Sigma}_k^{-1}\frac{1}{N_k}\sum_{n=1}^{N}(\boldsymbol{x}_{kn}-\boldsymbol{\mu}_k)(\boldsymbol{x}_{kn}-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}\right)\pi_k \tag{11.35}$$

把所有东西放在一起，对数可能性的偏导数

关于$\boldsymbol{\Sigma}_k$，由以下公式给出

$$\frac{\partial L}{\partial \boldsymbol{\Sigma}_k} = \frac{\partial \log p(\boldsymbol{x}_n|\boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} = \sum_{n=1}^{N} \frac{1}{p(\boldsymbol{x}_n|\boldsymbol{\theta})} \frac{\partial p(\boldsymbol{x}_n|\boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} \tag{11.36a}$$

$$= \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \log \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{11.36b}$$

$$= \sum_{n=1}^{N} r_{nk} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \log \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{11.36c}$$

$$= \sum_{n=1}^{N} r_{nk} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \left[\log \pi_k - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| - \frac{1}{2} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k) \right] \tag{11.36d}$$

我们看到，责任r_{nk}也出现在这个偏导中。将这个偏导设为$\mathbf{0}$，我们得到必要的最优性条件

$$N_k \boldsymbol{\Sigma}_k^{-1} = \boldsymbol{\Sigma}_k^{-1} \sum_{n=1}^{N} r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} \tag{11.37a}$$

$$\iff \boldsymbol{I} = \sum_{n=1}^{N} r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}. \tag{11.37b}$$

通过对$\boldsymbol{\Sigma}_k$的求解，我们得到

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \tag{11.38}$$

其中$\boldsymbol{r}_k$是定义在（）的概率向量。11.25).这给了我们一个简单的$\boldsymbol{\Sigma}_k$的更新规则，对于$k=1, \ldots, K$，并证明了定理11.2. □

与()中$\boldsymbol{\mu}_k$的更新类似，我们可以解释()中协方差的更新。11.20)中对协方差的更新，我们可以把(11.30)中协方差的更新可以解释为重要性加权的中心数据$\tilde{\boldsymbol{X}}_k := \{\boldsymbol{x}_1 - \boldsymbol{\mu}_k, \ldots, \boldsymbol{x}_N - \boldsymbol{\mu}_k\}$。

例子 (11.4差异更新)

在我们图中11.3,的例子中，差值被更新如下。

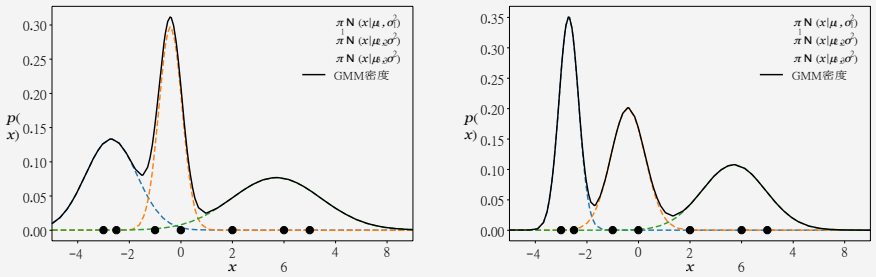
$$\sigma_1^2: 10.14 \to \tag{11.39}$$

$$\sigma_2^2: 0.2 \to 0.44 \tag{11.40}$$

$$\sigma_3^2: 31.53 \to \tag{11.41}$$

这里我们看到，第一和第三部分的方差明显缩小，而第二部分的方差略有增加。

图11.6说明了这种设置。图11.6(a)与图11.5(b)相同(但是放大了)，显示了更新方差之前的GMM密度及其各个组成部分。图11.6(b)显示了更新方差后的GMM密度。



图：更新GMM中方差的效果**11.6**。(a)更新方差前的GMM；(b)更新方差后的GMM，同时保留平均值和混合权重。

(a)GMM密度和个别成分 (b)更新方差前的GMM密度和个别成分 。更新方差后。

的数据点$\boldsymbol{x}$的加权协方差的蒙特卡洛估计$_n$，其中的权重是

与平均参数的更新一样，这个更新取决于所有的π_j、$\boldsymbol{\mu}_j$、$\boldsymbol{\Sigma}_j$，$j = 1, \ldots, K$，通过责任r_{nk}，这禁止了一个封闭式的解决方案。

11.2.4 更新混合物的权重

定理11.3（GMM混合物权重的更新）。*GMM的混合权重被更新为*

$$\pi^{new} = \frac{Nk}{kN} \quad ,k=1, \ldots, K, \tag{11.42}$$

其中N是数据点的数量，N_k的定义在(11.24).

为了找到对数似然相对于权重参数π_k的偏导，$k=1$，。，K，我们通过1使用拉格朗日乘法器（见第7.2节）来考虑con-$_{straintk}\pi_k$=。Lagrangian是

$$\mathsf{L} = \mathsf{L} + \lambda \sum_{k=1}^{K} \pi_k - 1 \tag{11.43a}$$

$$= \sum_{n=1}^{N} \text{原} \sum_{k=1}^{K} \pi_k \mathsf{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k + \lambda \sum_{k=1}^{K} \pi_k - 1 \quad , \tag{11.43b}$$

L 其中的对数似然是来自(11.10)，第二项是平等约束，即所有的混合权重的总和需要达到

1.我们得到关于π_k的部分导数为

$$\frac{\partial \mathsf{L}}{\partial \pi_k} = \sum_{n=}^{N} \frac{\mathsf{N}(x_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k}{LK} + \lambda \tag{11.44a}$$

$$1 \frac{j=1}{\sum_{N}^{1} \pi_k \mathsf{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k} + \lambda = \frac{Nk}{\pi_k} + \lambda \,, \tag{11.44b}$$

$$\pi \mathsf{k} \mathsf{L}_K \sum_{n=1}^{} \sum_{j=1}^{} \pi \mathsf{N}(\boldsymbol{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}$$

$$=N$$

和相对于拉格朗日乘数λ的偏导为

$$\frac{\partial \mathsf{L}}{\partial \lambda} = \sum_{k=1}^{K} \pi_k - 1. \tag{11.45}$$

将两个偏导数设为（0最佳的必要条件），可得到方程组

$$\pi_k = -\frac{Nk}{\lambda}, \tag{11.46}$$

$$1 = \sum_{k=1}^{K} \pi_k \,. \tag{11.47}$$

使用(11.46)中的(11.47)并求解π_k，我们得到

$$\sum_{k=1}^{K} \pi = \Leftarrow 1 \Rightarrow \quad \sum_{k=1}^{K} Nk = \Leftarrow 1 \Rightarrow -\frac{N}{\lambda} \quad = \Leftarrow 1 \Rightarrow \ \lambda = -N \text{。} \tag{11.48}$$

这使得我们可以用$-N$代替λ，在(11.46)，得到

$$\pi^{\text{new}} = \frac{Nk}{kN}, \tag{11.49}$$

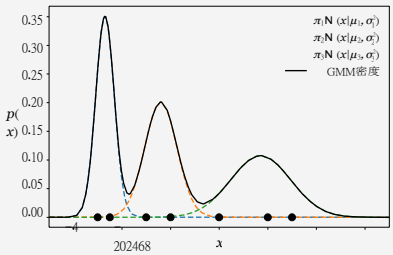
这给了我们权重参数π_k的更新，并证明了Theo- rem11.3.

□

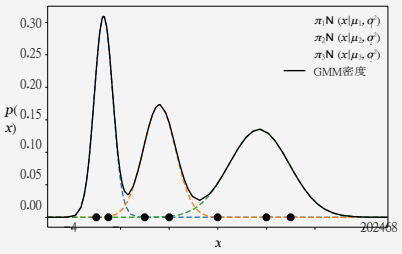
我们可以把(11.42)中的混合权重是第k个聚类的责任与数据点数量的比率。由于$N = \sum_k N_k$，数据点的数量也可以解释为所有混合成分的总责任，这样π_k就是第k个混合成分对数据集的相对重要性。

备注。由于$N_k = \sum_{i=1}^{N} r_{nk}$的更新方程(11.42)为混合-结构权重$\pi_k$也取决于$\mathsf{L}$所有的$\pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, j = 1, \ldots, K$通过再责任$r_{nk}$。 ◆

例子（11.5重量参数更新）



(a) 在更新混合物权重之前，GMM密度和个别成分。



(b) 更新混合物权重后的GMM密度和各个组成部分。

图：在GMM中更新混合物权重的效果**11.7**。(a)GMM (b)更新混合权重后的GMM，同时保留了平均值和方差。注意不同尺度的纵轴。

在我们图中11.3,的运行例子中，混合物的权重被更新如下。

$$\pi_1: \frac{1}{3} \rightarrow 0.29 \tag{11.50}$$
$$\pi_2: \frac{1}{3} \rightarrow 0.29 \tag{11.51}$$
$$\pi_3: \frac{1}{3} \rightarrow 0.42 \tag{11.52}$$

在这里，我们看到第三个组成部分得到了更多的权重/重要性，而其他组成部分则变得稍微不那么重要。图11.7说明了更新混合权重的效果。图11.7(a)与图11.6(b)相同，显示了更新混合权重之前的GMM密度及其各个成分。图11.7(b)显示了更新混合权重后的GMM密度。

总之，在更新了一次均值、方差和权重之后，我们得到了图11.7(b)所示的GMM。与图中所示的初始化相比，11.3,我们可以看到，参数更新导致GMM密度向数据点转移了一些质量。

在更新一次均值、方差和权重后，图11.7(b)中的GMM拟合已经明显好于图中11.3.的初始化。这也可以从对数似然值中得到证明，在一个完整的更新周期后，对数似然值从28 。3(在一个完整的更新周期4后，对数似然值从.（初始化）上升到14.。

11.3 EM算法

复杂的方式来估计参数。然而，结果表明有一个简单的*迭代方案*，通过最大似然来寻找参数估计问题的解决方案。期望最大化算法(*EM算法*)

不幸的是，在(11.20)，(11.30)和(11.42)中的更新并不构成混合物模型的参数$\boldsymbol{\mu}_k$、$\boldsymbol{\Sigma}_k$、π_k更新的闭式解决方案，因为责任r_{nk}取

rithm）是由Dempster等人（1977年）提出的，是一个通用的迭代方案，用于学习混合模型中的参数（最大似然或MAP），更广泛地说，用于学习潜在变量模型。

在我们的高斯混合模型的例子中，我们选择了$\boldsymbol{\mu}_k$、$\boldsymbol{\Sigma}_k$、π的初始值，k并交替使用，直到收敛于

- *E步。*评估责任r_{nk}（数据点n属于混合成分k的后验概率）。
- *M-步骤。*使用更新的责任来重新估计参数

$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$.

EM算法的每一步都会增加对数似然函数（Neal和Hinton，1999）。对于收敛性，我们可以直接检查对数似然或参数。一个用于估计GMM参数的EM算法的具体实例如下。

1. 初始化$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$。
2. *E-步骤。*对每个数据点$\boldsymbol{x}_n$的责任r_{nk}进行评估，使用current-租用参数$\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$。

$$r_{nk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{11.53}$$

3. *M-步骤。*重新估计参数$\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$，使用当前负责的方法。在更新了"E-步骤"我们可以看到

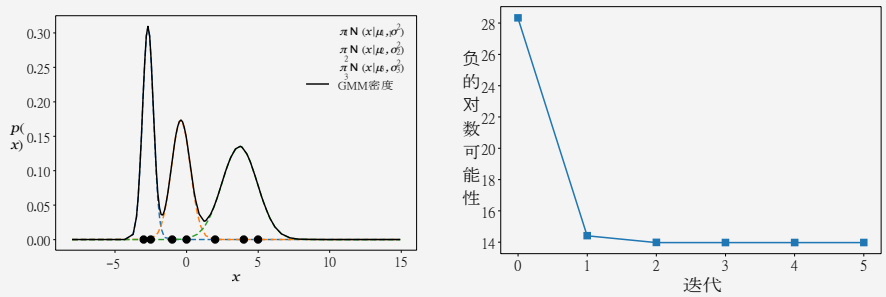
$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n, \tag{11.54}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^{\top}, \tag{11.55}$$

$$\pi_k = \frac{N_k}{N}。 \tag{11.56}$$

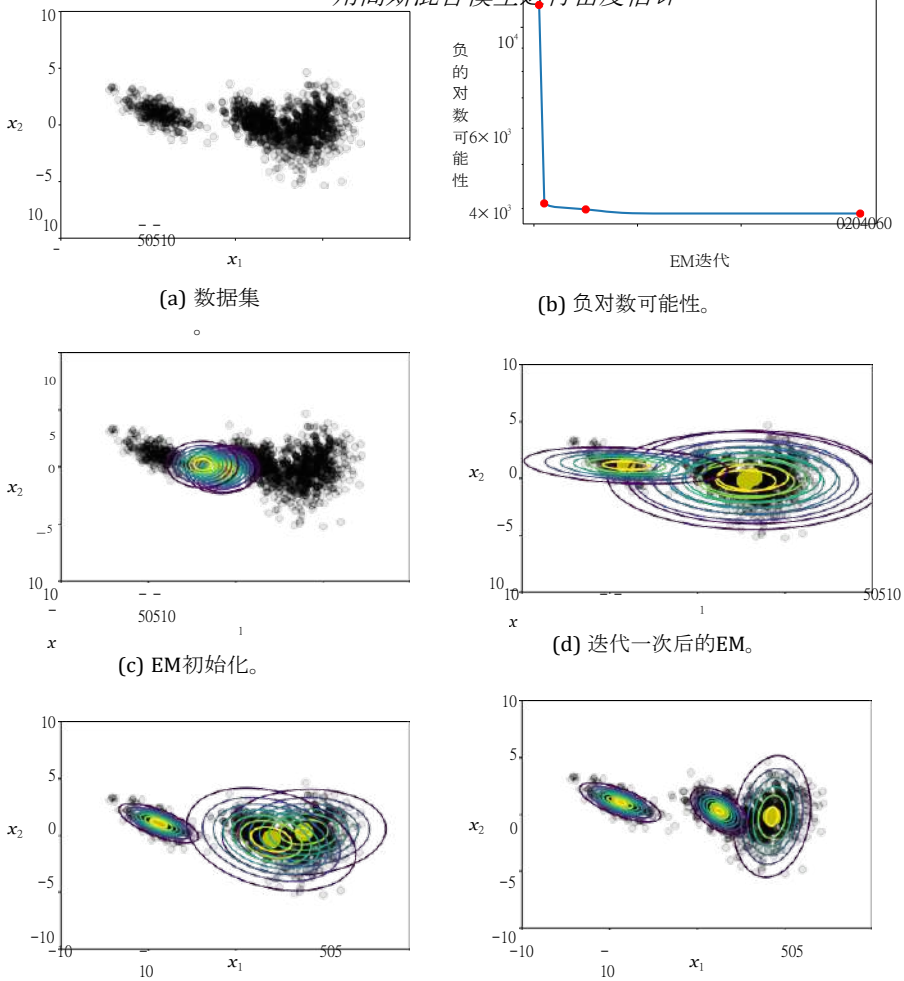
中的"能力R_{nk}"后，途径$\boldsymbol{\mu}_k$在(11.54)，它们随后被用于在(11.55)来更新相应的协方差。

示例（11.6GMM拟合）。



(a) 最终的GMM拟合。经过五次迭代，EM（b）负对数似然值作为算法的函数收敛并返回这个GMM。EM迭代。

图11.8 应用于图11.2.中GMM的EM算法 (a) 最终GMM拟合。(b) 负面作为EM迭代的函数的对数似然。



(a) 数据集。

(b) 负对数可能性。

(c) EM初始化。

(d) 迭代一次后的EM。

(e) 迭代10后的EM。

(f) 迭代62后的EM。

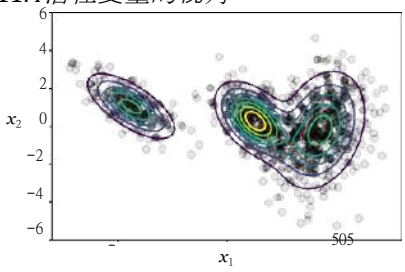
图：用EM算法对二维数据集拟合有三个成分的高斯混合模型的说明**11.9**。(a)数据集；(b)负数对数可能性(越低越好)是EM迭代的一个函数。红色的圆点表示的迭代，其中混合物相应的GMM的分量拟合结果见(c)。通过(f)。该黄色圆盘表示高斯混合成分的平均值。

图 11.10 （a）显示了最终的GMM拟合。

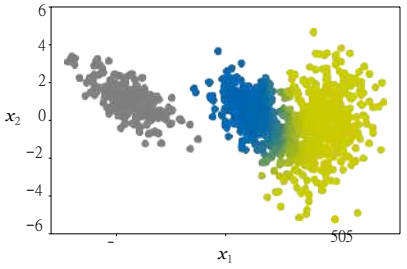
当我们在图11.3,中的例子上运行EM时，经过五次迭代后得到了图11.8(a)所示的最终结果，图11.8(b)显示了负对数可能性是如何作为EM迭代的函数而演变的。最终的GMM是这样给出的

$$p(x) = 0.29\,\mathcal{N}\left(x \mid -2.75, 0.06\right) + 0.28\,\mathcal{N}\left(x \mid -0.50, 0.25\right) + 0.43\,\mathcal{N}\left(x \mid 3.64, 1.63\right) \tag{11.57}$$

我们将EM算法应用于图中所示的二维数据集。11.1图中所示的二维数据集，其中 $K =$ 混合3成分。图中11.9说明了EM算法的一些步骤，并显示了作为EM迭代函数的负对数似然（图11.9(b)）。图11.10(a)显示



(a) 迭代62后的GMM拟合。



(b) 数据集根据混合物成分的反应能力而着色。

图为EM收敛时的**11.10**GMM拟合和责任。(a)EM收敛时的GMM拟合；(b)每个数据点根据混合成分的责任而着色。

对应的最终GMM拟合。图11.10(b)直观地显示了数据点的混合成分的最终责任。当EM收敛时，数据集根据混合成分的责任被着色。虽然单一的混合成分显然对左边的数据负责，但右边的两个数据簇的重叠可能是由两个混合成分产生的。很明显，有一些数据点不能唯一地分配给单一成分（蓝色或黄色），这样，这两个集群对这些点的责任大约0是.5.....。

11.4 潜在变量的视角

我们可以从离散潜变量模型的角度来看待GMM，也就是说，潜变量$\boldsymbol{z}$只能达到一组有限的值。这与PCA不同，PCA中的潜变量是R中的M连续值数字。

概率论观点的优点是：（i）它将对我们在前面几节中做出的一些临时决定进行说明，（ii）它允许将责任具体解释为后验概率，以及（iii）更新模型参数的迭代算法可以以一种原则性的方式推导出潜变模型中最大似然参数估计的EM算法。

11.4.1 生成过程和概率模型

为了推导出GMMs的概率模型，思考生成过程是很有用的，也就是让我们使用概率模型生成数据的过程。

我们假设一个有K个成分的混合模型，一个数据点$\boldsymbol{x}$正好可以由一个混合成分产生。我们引入一个二元指标变量$z_k \in \{0,1\}$，有两个状态（见第6.2节），表示第k个混合成分是否产生了该数据点

以致于

$$p(\boldsymbol{x} \mid z_k = 1) = \mathsf{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right). \tag{11.58}$$

我们定义$\boldsymbol{z} := [z_1, \ldots, z_K] \in \mathbb{R}^K$为概率向量，由$K$个很多0的s和恰好一个1.例如，对于$K = 3$，一个有效的$\boldsymbol{z}$将是$\boldsymbol{z} = [z_1, z_2, z_3] = [0, 1, 0]$，这将选择第二个混合成分，因为$z_2 = 1$。

*备注。*有时这种概率分布被称为"多努利"，是伯努利分布对两个以上数值的概括（Murphy，2012）。 ◆

单热编码

1-的-K

LT的属性意味着，$\boldsymbol{z}$的属性是$\sum_{k=1}^{K} z_k = 1$因此，$\boldsymbol{z}$是一个*单热的编码*（还有：*1-of-K表示法*）。

到目前为止，我们假设指标变量z_k是已知的。然而，在实践中，情况并非如此，因此我们将一个先验分布

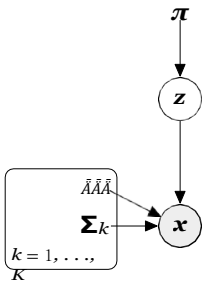
$$p(\boldsymbol{z}) = \boldsymbol{\pi} = [\pi_1, \ldots, \pi_K], \quad \sum_{k=1}^{K} \pi_k = 1, \tag{11.59}$$

在潜变量$\boldsymbol{z}$上，那么第k个条目

$$\pi_k = p(z_k = 1) \tag{11.60}$$

的概率向量描述了第k个混合成分产生数据点$\boldsymbol{x}$的概率。

图为有单个数据点的GMM的图形**11.11**模型。



备注（从GMM中抽样）。这个潜在变量模型的构建（见图11.11中相应的图形模型）适合用非常简单的抽样程序（生成过程）来生成数据。

1. 采样$z^{(i)} \sim p(\boldsymbol{z})$。
2. 采样$\boldsymbol{x}^{(i)} \sim p(\boldsymbol{x} \mid z^{(i)} = 1)$。

在第一步中，我们根据$p(\boldsymbol{z}) = \boldsymbol{\pi}$随机选择一个混合成分$i$（通过一热编码-ing $\boldsymbol{z}$）；在第二步中，我们从相应的混合成分中抽取一个样本。当我们放弃潜变量的样本，只剩下$\boldsymbol{x}^{(i)}$时，我们就有了GMM的有效样本。这种抽样，即随机变量的样本取决于图形模型中变量的父代的样本，被称为*祖先抽样*。 ◆

祖传采样

一般来说，概率模型是由数据和潜变量的联合分布来定义的（见第8.4节）。由于先验$p(\boldsymbol{z})$定义在(11.59)和(11.60)中定义的先验p(z)和条件$p(\boldsymbol{x} \mid \boldsymbol{z})$来自(11.58)，我们通过以下方式得到这个联合分布的所有K个分量

$$p(\boldsymbol{x}, z_k = 1) = p(\boldsymbol{x} \mid z_k = 1)p(z_k = 1) = \pi_k \mathsf{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \tag{11.61}$$

对于$k=1,\ldots,K$，因此，

$$p(\boldsymbol{x}, \boldsymbol{z}) = \begin{bmatrix} p(\boldsymbol{x}, z_1 = 1) \\ \vdots \\ p(\boldsymbol{x}, z_K = 1) \end{bmatrix} = \begin{bmatrix} \pi_1 \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \\ \vdots \\ \pi_K \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K) \end{bmatrix}, \tag{11.62}$$

这充分说明了概率模型。

11.4.2 可能性

为了得到潜在变量模型中的似然数$p(\boldsymbol{x} \mid \boldsymbol{\theta})$，我们需要将潜在变量边缘化（见第8.4.3节）。在我们的例子中，这可以通过从联合$p(\boldsymbol{x}, \boldsymbol{z})$中求和出所有的潜变量来完成。11.62)中的所有潜变量相加，这样就可以

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \sum_{\boldsymbol{z}} p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{z}) p(\boldsymbol{z} \mid \boldsymbol{\theta}), \qquad \boldsymbol{\theta} := \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k : k = 1, \ldots, K\}. \tag{11.63}$$

我们现在明确地对概率模型的参数$\boldsymbol{\theta}$提出条件，我们以前省略了这些参数。在(11.63)中，我们对$\boldsymbol{z}$的所有K个可能的单热编码进行加总，用$\sum_{\boldsymbol{z}}$表示。由于每个$\boldsymbol{z}$中只有一个非零单项，所以$\boldsymbol{z}$只有K种可能的配置/设置。

$$\begin{matrix} 1 \\ 0 \\ \vdots \\ 0 \end{matrix}, \begin{matrix} 0 \\ 1 \\ \vdots \\ 0 \end{matrix}, \ldots, \begin{matrix} 0 \\ 0 \\ \vdots \\ 1 \end{matrix}. \tag{11.64}$$

在( )中对$\boldsymbol{Z}$的所有可能配置进行求和，相当于查看Z向量的非零项，并写上11.63)中所有可能的配置求和，相当于看一下$\boldsymbol{Z}$向量的非零项，然后写出

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \sum_{\boldsymbol{z}} p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{z}) p(\boldsymbol{z} \mid \boldsymbol{\theta}) \tag{11.65a}$$

$$= \sum_{k=1}^{K} p(\boldsymbol{x} \mid \boldsymbol{\theta}, z_k = 1) p(z_k = 1 \mid \boldsymbol{\theta}) \tag{11.65b}$$

因此，所需的边际分布被赋予为

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) \overset{(11.65b)}{=} \sum_{k=1}^{K} p(\boldsymbol{x} \mid \boldsymbol{\theta}, z = 1) p(z = 1 \mid \boldsymbol{\theta}) \tag{11.66a}$$

$$= \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{11.66b}$$

我们将其确定为GMM模型，从(11.3).给定一个 X数据集，我们立即得到似然值

$$p(\mathbf{X} \mid \boldsymbol{\theta}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n \mid \boldsymbol{\theta}) \overset{(11.66b)}{=} \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{11.67}$$

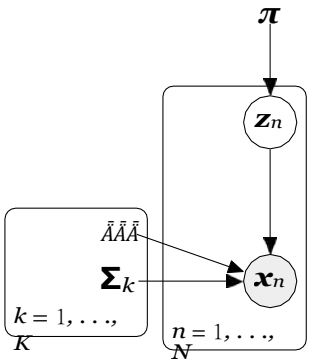


图11.12 有N个数据点的GMM的图形模型。

这正是GMM的可能性，从(11.9).因此，带有潜指标z的潜变量模型$_k$是对高斯混合模型的一种等效思考方式。

11.4.3 后期分布

让我们简单看一下潜在变量的后验分布

$\boldsymbol{z}$.根据贝叶斯定理，产生了数据点$\boldsymbol{x}$*的*第k个组件的后验

$$p(z_k = 1 \mid \boldsymbol{x}) = \frac{p(z_k = 1)p(\boldsymbol{x} \mid z_k = 1)}{p(\boldsymbol{x})}, \tag{11.68}$$

其中边际p（$\boldsymbol{x}$）在（11.66b）中给出。这就得到了第k个指标变量z的后验分布$_k$

$$p(z_k = 1 \mid \boldsymbol{x}) = \frac{p(z_k = 1)p(\boldsymbol{x} \mid z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\boldsymbol{x} \mid z_j = 1)} = \frac{\pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}, \tag{11.69}$$

注意，我们省略了对GMM的明确调节。

参数$\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$其中$k = 1, \ldots, K$.

11.4.4 扩展到完整的数据集

到目前为止，我们只讨论了数据集只包括一个数据点$\boldsymbol{x}$的情况。然而，先验和后验的概念可以直接扩展到N个数据点 的情况：$= \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$.在GMM的概率解释中，每一个数据点$\boldsymbol{x}$的$_n$位置是评估自己的潜在变量

$$\boldsymbol{z}_n = [z_{n1}, \ldots, z_{nK}] \in \mathbb{R}^K。 \tag{11.70}$$

以前（当我们只考虑单一数据点$\boldsymbol{x}$时），我们省略了索引n，但现在这变得很重要。

我们在所有潜在变量$\boldsymbol{z}$中n共享相同的先验分布$\boldsymbol{\pi}$。相应的图形模型如图所示11.12,其中我们使用板块符号。

条件分布$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N|\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N)$对数据点进行因子化，并给出了

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N|\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{z}_n) \text{。} \tag{11.71}$$

为了得到后验分布$p(z_{nk}=1|\boldsymbol{x}_n)$，我们遵循与第11节相同的推理，并应用贝叶斯定理得到后验分布。11.4.3并应用贝叶斯定理，得到

$$p(z_{nk}=1|\boldsymbol{x}_n) = \frac{p(\boldsymbol{x}_n|z_{nk}=1)p(z_{nk}=1)}{\sum_{j=1}^{K} p(\boldsymbol{x}_n|z_{nj}=1)p(z_{nj}=1)} \tag{11.72a}$$

$$= \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} = r_{nk} \text{。} \tag{11.72b}$$

这意味着$p(z_{nk}=1|\boldsymbol{x}_n)$是第$k$个混合物成分产生数据点$\boldsymbol{x}_n$的（后验）概率，对应于$r_{nk}$我们在(11.17).现在，责任也不仅有一个直观的，而且有一个数学上合理的解释，即后验概率。

11.4.5 重新审视EM算法

我们介绍的EM算法是一种最大似然估计的迭代方案，可以从潜在变量的角度以一种原则性的方式导出。考虑到模型参数的当前设置$\boldsymbol{\theta}^{(t)}$，E步骤计算预期对数可能性

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta}^{(t)}}[\log p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})] \text{。} \tag{11.73a}$$

$$= \int \log p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}^{(t)}) \mathrm{d}\boldsymbol{z} \text{，} \tag{11.73b}$$

其中，$\log p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})$的期望值是相对于潜在变量的后向$p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}^{(t)})$而言的。M步骤通过最大化（11.73b）来选择一套最新的模型参数$\boldsymbol{\theta}^{(t+1)}$。

尽管EM迭代确实增加了对数似然，但并不能保证EM收敛到最大似然解。EM算法有可能收敛到对数可能性的局部最大值。在多次EM运行中，可以使用不同的参数$\boldsymbol{\theta}$的初始化，以减少在一个坏的局部最优中结束的风险。我们在此不做进一步的详述，而是参考Rogers和Girolami（2016）以及Bishop（2006）的精彩阐述。

11.5 进一步阅读

GMM可以被认为是一个生成模型，因为它是
使用祖先抽样直接生成新数据（Bishop，2006）。对于给定的GMM参数 $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, k = 1, \ldots, K$，我们从概率向量$[\pi_1, \ldots, \pi]$中$_K$抽取一个索引$k$，然后抽取一个数据点$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$。如果我们这样重复$N$次，我们就会得到一个由GMM生成的数据集。图11.1是用这个方法生成的
程序。

在本章中，我们假设组件的数量K是已知的。在实践中，情况往往不是这样的。然而，我们可以使用嵌套交叉验证，如第8.6.1节所讨论的，来找到好的模型。高斯混合模型与K-means聚类算法密切相关。K-means也使用EM算法将数据点分配给聚类。如果我们把GMM中的平均值当作聚类中心，而忽略协方差（或者把它们设为$\boldsymbol{I}$），我们就得到了K-means。正如MacKay(2003)所描述的那样，K-means将数据点 "硬 "分配到聚类中心$\boldsymbol{\mu}_k$，而GMM则是 "软 "分配。通过责任。

我们只触及了GMMs和EM算法的潜变量角度。请注意，EM可以用于一般的潜变量模型的参数学习，例如非线性状态空间模型（Ghahramani和Roweis，1999；Roweis和Ghahramani，1999）以及Barber（2012）讨论的强化学习。因此，GMM的latent-variable per- spective有助于以一种原则性的方式推导出相应的EM算法（Bishop，2006；Barber，2012；Murphy，2012）。

我们只讨论了寻找GMM参数的最大似然估计（通过EM算法）。对最大似然的标准批评也适用于此。

- 与线性回归一样，最大似然法也会出现严重的过拟合现象。在GMM的情况下，当混合成分的平均值与数据点相同而协方差趋向于$\mathbf{0}$.然后，可能性接近无限大。Bishop（2006）和Barber (2012)详细讨论了这个问题。
- 我们只得到了$k = 1, \ldots, K$的参数$\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$的点估计，这并没有显示参数值的不确定性。，$K$，这并不能说明参数值的不确定性。贝叶斯方法会在参数上设置一个先验。

  等值，可以用来获得参数的后验分布。这个后验允许我们计算模型证据（边际似然），它可以用于模型比较，这给了我们一个确定混合成分数量的原则性方法。不幸的是，在这种情况下，封闭式推理是不可能的，因为这个模型没有共轭先验。然而，近似方法，如变异推理，

可以用来获得一个近似的后验（Bishop,2006）。

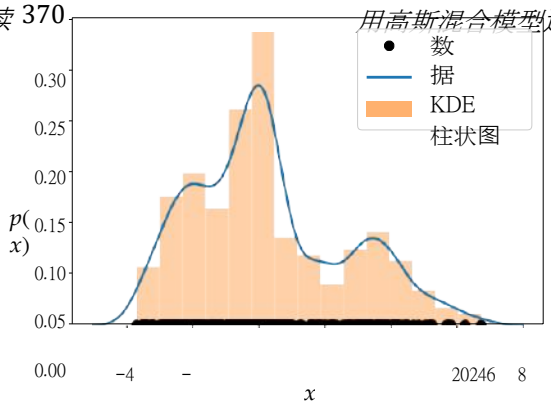


图 11.13直方图（橙色条）和内核密度估计（蓝线）。核心密度估计器产生了对基本密度的平滑估计，而直方图是对有多少数据点（黑色）落入某一区域的不平滑的计数测量。

在这一章中，我们讨论了密度估计的混合模型。有大量的密度估计技术可用。在实践中。我们经常使用直方图和内核密度估计。

单仓。

直方图

*直方图*提供了一种非参数的方式来表示连续的数据量，是由Pearson（1895）提出的。直方图是通过对数据空间进行 "分档 "并计算每个档位上有多少数据点来构成的。然后在每个仓的中心画一个条形图，条形图的高度与该仓内的数据点的数量成正比。仓的大小是一个关键的超参数，一个不好的选择会导致过拟合和欠拟合。交叉验证，如第8.2.4节所述，可以
用来确定一个好的bin大小。

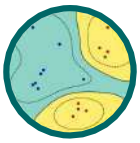
核心密度估算

*核心密度估计*是由Rosenblatt(1956)和Parzen(1962)独立提出的，是一种非参数的密度估计方法。给定N个独立样本，核密度估计器表示基础分布为

$$p(\boldsymbol{x}) = \frac{1}{Nh}\sum_{n=1}^{N} k\left(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\right), \tag{11.74}$$

其中k是一个核函数，即一个非负函数，它的积分为，1h>0是一个平滑/带宽参数，它的作用类似于直方图中的bin大小。请注意，我们在数据集中的每一个数据点$\boldsymbol{x}_n$上放置一个核。常用的核函数是均匀分布和高斯分布。核密度估计与直方图密切相关，但通过选择一个合适的核，我们可以保证密度估计的平滑性。图11.13说明了对于一个给定的数据集的数据250点，直方图和核密度估计器（高斯形核）之间的区别。

12



用支持向量机进行分类

在许多情况下，我们希望我们的机器学习算法能够预测一些（离散）结果中的一个。例如，一个电子邮件客户端将邮件分为个人邮件和垃圾邮件，这有两个结果。另一个例子是一个望远镜，它可以识别夜空中的物体是星系、恒星还是行星。通常有少量的结果，更重要的是这些结果上通常没有额外的结构。在本章中，我们考虑输出二元值的预测器，即只有两种可能的结果。这种机器学习任务被称为*二进制分类*。这与第九章不同，第九章中我们考虑的是具有连续值输出的预测问题。

结构的一个例子是，如果结果是有序的，比如小号、中号和大号的T恤衫。

二元分类

对于二元分类，标签/输出可以达到的可能值的集合是二元的，在本章中我们用+1、1表示。换句话说，我们考虑的预测器形式为 $\{-\}$

$$f : \mathbb{R}^D \to \{+1, -1\} . \tag{12.1}$$

回顾第八章，我们把每个例子（数据点） $\boldsymbol{x}$ 表示$_n$为一个D实数的特征向量。这些标签通常被分别称为*正类*和*负类*。我们应该注意不要推断出+1类的直观属性的积极性。例如，在癌症检测任务中，一个患癌症的病人往往被标记为+1。原则上，可以使用任何两个不同的值，例如，真，假，0，或1红，蓝。二元分类的问题已经得到了很好的研究，我们将对其他方法的调查推迟到第二节 $\{\{\}$ 12.6$\}$。

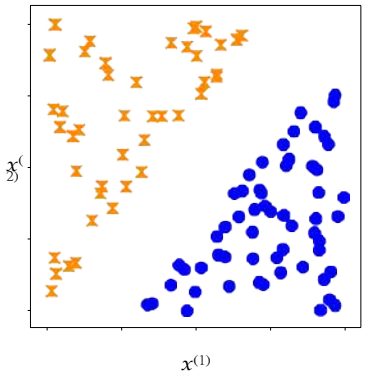
输入示例 x_n 也可以被称为输入、数据点、特征或实例。

类

对于概率模型，使用0，作为1二元表示法在数学上是很方便的；见例6.12 $\{\}$ 的注释。

我们提出了一种被称为支持向量机（SVM）的方法，它解决了二元分类任务。和回归一样，我们有一个有监督的学习任务，我们有一组例子 $\boldsymbol{x}_n \in \mathbb{R}^D$ 沿着与它们相应的（二进制）标签y_n+1，1。给定 $\{$ 一个由 $\}$ 例子-标签对（$\boldsymbol{x}_1, y_1$），... ...组成的训练数据集。，$(\boldsymbol{x}_N, y_N)$，$\{\}$ 我们希望估计模型的参数，以获得最小的分类误差。与第九章类似，我们考虑一个线性模型，并将非线性隐藏在例子的变换φ中（9.13）。我们将在第9章中重新审视φ。12.4.

SVM在许多应用中提供了最先进的结果，并有良好的理论保证（Steinwart和Christmann，2008）。我们选择用二进制分类法来说明有两个主要原因



图例12.1二维数据，说明了数据的直观性，我们可以找到一个线性分类器，将橙色的十字架和蓝色的圆盘分开。

SVMs。首先，SVM允许用一种几何方法来思考监督机器学习的问题。在第九章中，我们从概率模型的角度考虑了机器学习问题，并使用最大似然估计和贝叶斯推理对其进行了攻击，而在这里，我们将考虑另一种方法，对机器学习任务进行几何推理。它在很大程度上依赖于一些概念，如内积和投影，我们在第三章中讨论过这些概念。我们发现SVM具有指导意义的第二个原因是，与第九章相比，SVM的优化问题不允许有分析性的解决方案，因此我们需要借助第七章中介绍的各种优化工具。

机器学习的SVM观点与第九章的最大似然观点有细微的不同。最大似然观点是基于数据分布的概率观点建立模型，并从中得出优化问题。相比之下，SVM的观点是基于几何学的直觉，从设计一个在训练中要优化的特定函数开始。我们在第二章中已经看到了类似的情况10,中已经看到了类似的情况，我们从几何原理中导出了PCA。在SVM的情况下，我们从设计一个损失函数开始，按照经验风险最小化的原则（第8.2节），在训练数据上使其最小化。

让我们推导出对应于在例子-标签对上训练SVM的优化问题。直观地说，我们设想二元分类数据，它们可以被一个超平面分开，如图所示12.1.在这里，每个例子$\boldsymbol{x}_n$（维度2为 ）是一个二维位置 （$\boldsymbol{x}^{(1)}$和$\boldsymbol{x}^{(2)}$），而相应的二元标签$\boldsymbol{y}_n$是其中之一
两个不同的符号（橙色十字或蓝色圆盘）。"超平面 "是机器学习中常用的一个词，我们在第2.8节已经遇到过超平面。超平面是一个二度空间D1的仿射子空间（如果相应的矢量空间的维数是D）。例子由两个类组成（有两个可能的标签），它们的特征（代表例子的向量的分量）为
摆放的方式使我们能够通过画一条直线将它们分开/分类。

在下文中，我们将寻找两类的线性分离器的想法正式化。我们介绍了余量的概念，然后扩展了线性分离器，以允许例子落在 "错误 "的一边，产生分类错误。我们提出了两种形式化SVM的同等方法：几何观点（第12.2.4节）和损失函数观点（第12.2.5节）。12.2.5).我们使用拉格朗日乘法器推导出SVM的对偶版本（第7.2节）。对偶SVM使我们能够观察到SVM的第三种形式化方式：从每个类的例子的凸壳来看（第12.3.2).最后，我们简要地描述了核以及如何用数值解决非线性核-SVM优化问题。

12.1分离超平面

给出两个以向量$\boldsymbol{x}_i$和$\boldsymbol{x}$表示的$_j$例子，计算它们之间的相似性的一种方法是使用内积$\boldsymbol{x}_i$，$\boldsymbol{x}_j$。回顾第3.2节，内积与两个向量之间的角度密切相关。两个向量之间的内积值取决于每个向量的长度（规范）。此外，内积允许我们严格地定义几何概念，如正交性和pro-jections。

许多分类算法的主要思想是用R表示^{D}数据，然后对这个空间进行划分，最好是让具有相同标签的例子（而没有其他例子）处于同一分区。在二元分类的情况下，空间将被划分为两部分，分别对应于正类和负类。我们考虑一个特别方便的分区，就是用超平面将空间（线性）分成两半。让例子$\boldsymbol{x}$ R^{D}是数据空间的一个元素。考虑一个函数

$$f:\mathrm{R}^D \to \mathrm{R} \tag{12.2a}$$

$$\boldsymbol{x} \to f(\boldsymbol{x}) := (\boldsymbol{w}, \boldsymbol{x}) + b\,, \tag{12.2b}$$

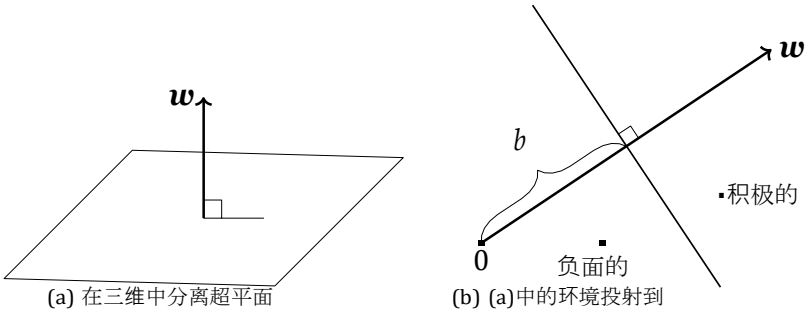
的参数化的wR^{D}和bR。回顾第2.8节，hy-perplane是仿射子空间。因此，我们将二元分类问题中分隔两类的超平面定义为

$$\{\boldsymbol{x} \in \mathrm{R}^D : f(\boldsymbol{x}) = 0\}\,. \tag{12.3}$$

超平面的图示如图所示12.2,其中矢量$\boldsymbol{w}$是超平面的法向矢量，b是截距。我们可以推导出$\boldsymbol{w}$是超平面的法向量，在(12.3)中，选择超平面上的任意两个例子$\boldsymbol{x}_a$和$\boldsymbol{x}_b$，并证明它们之间的矢量与$\boldsymbol{w}$正交。

$$f(\boldsymbol{x}_a) - f(\boldsymbol{x}_b) = (\boldsymbol{w}, \boldsymbol{x}_a) + b - ((\boldsymbol{w}, \boldsymbol{x}_b) + b) \tag{12.4a}$$

$$= (\boldsymbol{w}, \boldsymbol{x}_a - \boldsymbol{x}_b)\,, \tag{12.4b}$$



(a) 在三维中分离超平面 (b) (a)中的环境投射到飞机

图：分离超平面的方程式**12.2** (12.3).(a)在三维中表示该方程的标准方法。(b)为了便于绘制，我们把超平面的边缘放在上面。

其中第二行是由内积的线性得到的（第3.2节）。由于我们选择$\boldsymbol{x}_a$和$\boldsymbol{x}_b$都在超平面上，这意味着$f(\boldsymbol{x}_a) = 0$和$f(\boldsymbol{x}_b) = 0$，因此$\langle \boldsymbol{w}, \boldsymbol{x}_a - \boldsymbol{x}_b \rangle = 0$。回顾一下，当两个向量的内积为零时，它们是正交的。因此，我们得到$\boldsymbol{w}$与超平面上的任何矢量都是正交的。

*备注：从第二章可以看出，我们可以用不同的方式来考虑向量。*回顾第二章，我们可以用不同的方式思考向量。在本章中，我们把参数向量$\boldsymbol{w}$看作是一个指示方向的箭头，也就是说，我们认为$\boldsymbol{w}$是一个几何向量。相比之下，我们把示例向量$\boldsymbol{x}$看作是一个数据点（由其坐标表示），也就是说，我们把$\boldsymbol{x}$看作是一个向量的坐标相对于标准基础。 ◆

当遇到一个测试例子时，我们根据该例子出现在超平面的哪一边，将其分为正向或负向。请注意，(12.3)不仅定义了一个超平面；它还定义了一个方向。换句话说，它定义了超平面的正反两面。因此，为了对一个测试例子$\boldsymbol{x}_{\text{test}}$进行分类，我们计算函数$f(\boldsymbol{x}_{\text{test}})$的值，如果$f(\boldsymbol{x}_{\text{test}}) > 0$，则将该例子分类为+1，否则为−1。从几何学的角度考虑，正面的例子位于超平面的 "上方"，负面的例子位于超平面的 "下方"。

训练分类器时，我们要确保有正标签的例子在超平面的正侧，即。

$$\langle \boldsymbol{w}, \boldsymbol{x}_n \rangle + b \geqslant 0 \quad \text{当 } y_n = +1 \tag{12.5}$$

，有负数标签的例子是在负数一边，即：。

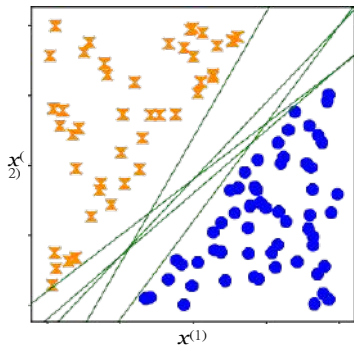
$$\langle \boldsymbol{w}, \boldsymbol{x}_n \rangle + b < 0 \quad \text{当 } y_n = -1 \text{。} \tag{12.6}$$

请参考图12.2以了解正负例子的几何直观。这两个条件经常以一个方程的形式出现

$$y_n(\langle \boldsymbol{w}, \boldsymbol{x}_n \rangle + b) \geqslant 0 \text{。} \tag{12.7}$$

方程(12.7)等同于(12.5)和(12.6当我们将( )和( )两边相乘时，相当于( )和( )。12.5)和(12.6)两边分别乘以$y_n = +1$和$y_n = -1$。

是与超平面上的任何矢量正交的。



图中12.3可能的分离超平面。有许多线性分类器（绿线）可以将橙色的十字架和蓝色的圆盘分开。

12.2 原始支持向量机

基于点到超平面的距离的概念，我们现在可以讨论支持向量机了。对于一个数据集$(\boldsymbol{x}_1, y_1), \ldots$对于一个线性可分离的数据集（x，y）$\{(\boldsymbol{x}_N, y_N)\}$，我们有无限多的候选超平面（参考图12.3），因此也有分类器，可以解决我们的分类问题而没有任何（训练）误差。为了找到一个唯一的解决方案，一个想法是选择一个分离超平面，使正面和负面的例子之间的边际最大化。换句话说，我们希望正反两方面的例子被一个大的余量分开（第12.2.1).在下文中，我们计算一个例子和一个超平面之间的距离，从而得出边际。回顾一下，超平面上离给定点（例子$\boldsymbol{x}_n$）最近的点是通过正交投影得到的（第3.8节）。

具有大余量的分类器被证明具有良好的概括性（Steinwart和Christmann，2008）。

12.2.1 保证金的概念

*边际的*概念在直觉上很简单。假设数据集是线性可分离的，那么它就是分离超平面到数据集中最接近的例子的距离。然而，当试图正式确定这个距离时，有一个技术上的问题可能会让人困惑。这个技术问题是，我们需要定义一个测量距离的尺度。一个潜在的尺度是考虑数据的尺度，即$\boldsymbol{x}$的n原始值。这有问题，因为我们可以改变$\boldsymbol{x}$的测量单位n，改变$\boldsymbol{x}$的n值，从而改变到超平面的距离。正如我们很快会看到的，我们根据超平面的方程来定义比例 (12.3)本身来定义比例。

边缘

一个超平面可能有两个或多个最接近的例子。

考虑一个超平面$\boldsymbol{w}, \boldsymbol{x}+b$，和一个例子$\boldsymbol{x}_a$，如图所示12.4.在不丧失一般性的情况下，我们可以认为例子$\boldsymbol{x}_a$在超平面的正侧，即$\boldsymbol{w}, \boldsymbol{x}_a + b > 0$。我们想计算$\boldsymbol{x}_a$与超平面的距离$r > 0$。我们通过考虑$\boldsymbol{x}_a$在超平面上的正交投影（第3.8节）来做到这一点，我们用x_a表示。由于$\boldsymbol{w}$是正交于

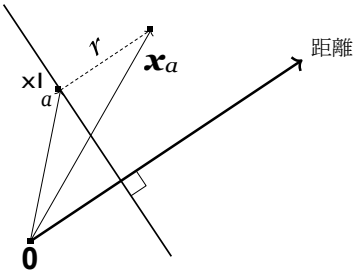


图12.4 向量加法来表达超平面。$\boldsymbol{x}' = \boldsymbol{x}_a + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$. a

如果$\boldsymbol{w}$的长度是已知的，那么我们就可以用这个缩放系数r系数来计算$\boldsymbol{x}_a$和$_{xa}$之间的t绝对距离。为方便起见，我们选择使用一个单位长度的向量（其规范为1），并得到这个通过将$\boldsymbol{w}$除以其规范，。$\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$使用矢量加法（第2.4节），我们获得

$$\boldsymbol{x} = \boldsymbol{x}_a + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \tag{12.8}$$

对r的另一种思考方式是，它是$\boldsymbol{x}_a$在$\boldsymbol{w}/\|\boldsymbol{w}\|$所跨越的子空间中的坐标。我们现在已经把$\boldsymbol{x}_a$与超平面的距离表示为$r$，如果我们选择$\boldsymbol{x}_a$是最接近超平面的点，这个距离$r$就是边际。

回顾一下，我们希望正面的例子离超平面的距离大于r，负面的例子离超平面的距离大于dis-tance r（在负面方向）。类比于(12.5)和(12.6)组合成(12.7)，我们把这个目标表述为

$$y_n(\langle \boldsymbol{w}, \boldsymbol{x}_n \rangle + b) \geq r \tag{12.9}$$

换句话说，我们结合了例子至少是r远离超平面（在正和负方向），变成一个单一的不等式。

由于我们只对方向感兴趣，所以我们增加一个假设，即我们的模型中，参数向量$\boldsymbol{w}$的长度为单位，即$\|\boldsymbol{w}\| = 1$。其中我们使用欧氏规范$\|\boldsymbol{w}\| = \sqrt{\boldsymbol{w}^\top \boldsymbol{w}}$（第3.1节）。这假设也允许更直观地解释距离$r$(12.8)，因为它是一个长度1为.....的向量的缩放系数。

我们将看到其他内在产品的选择（第3.2节），在第3.2节。12.4.

*备注。*熟悉边际的其他表述的读者会注意到，如果SVM是由Sch ölkopf和Smola（2002）提供的，例如，我们对$\|\boldsymbol{w}\| = 1$的定义与标准表述1不同。在第12.2.3,我们将展示这两种方法的等效性。

◆

将三个要求收集到一个单一的约束性优化中

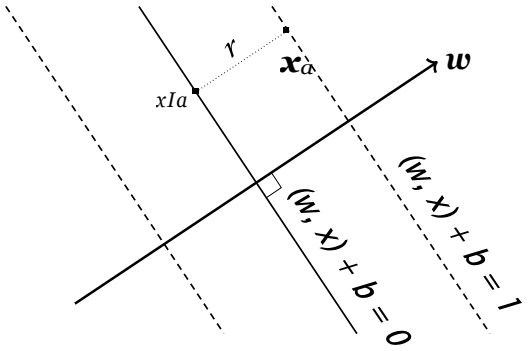


图12.5 边际的推导：$r = 。 \frac{1}{/\boldsymbol{w}/}$

问题，我们得到目标

$$\begin{array}{l} \text{最} \quad r \\ \text{大} \quad \text{ÃÃÃÃ} \\ \boldsymbol{w},b,r \end{array} \quad \text{主体是} \underbrace{\text{玩具}_{(n}(\boldsymbol{w}, \boldsymbol{x}_n) + b)}_{\text{数据处理}} \underbrace{\text{\textcircled{r}}}_{} \underbrace{r}_{} , \underbrace{\text{lwl}}_{} = 1 , \quad r > 0 , \tag{12.10}$$

统计数据

也就是说，我们要在确保数据位于超平面的正确一侧的同时，最大限度地提高余量r。

备注。 边际的概念在数学学习中是非常普遍的。Vladimir Vapnik和Alexey Chervonenkis用它来说明，当边际较大时，函数类的"复杂性"较低，因此学习是可能的（Vapnik，2000）。事实证明，这个概念对理论上分析泛化误差的各种不同方法都很有用（Steinwart和Christmann,2008; Shalev-Shwartz和Ben-David,2014）。 ◆

12.2.2 保证金的传统推导

在上一节中，我们通过观察我们只对w的方向感兴趣而不是它的长度来推导出(12.10)是通过观察我们只对$\boldsymbol{w}$的方向感兴趣，而不是它的长度，从而得出$\boldsymbol{w}$=的假设1。 在本节中，我们通过不同的假设来推导边际最大化问题。我们不选择参数向量是标准化的，而是为数据选择一个尺度。我们选择这个尺度，使预测器$\boldsymbol{w}$，$\boldsymbol{x}$+b的值在最接近的例子中为1。让我们也用$\boldsymbol{x}_a$来表示数据集中最接近超平面的例子。

回顾一下，我们目前考虑的是线性可分离的数据。

图12.5与图12.4相同，只是现在我们重新调整了轴的比例，使例子$\boldsymbol{x}_a$正好位于边缘上，即（$\boldsymbol{w}$，$\boldsymbol{x}_a$）+。
由于$^{t}_{xa}$是$\boldsymbol{x}_a$在超平面上的正交投影，它根据定义，必须位于超平面上，即。

$$(\boldsymbol{w}, {}^{\mathrm{t}}_{xa}) + b = . \quad 0(12.11)$$

通过将(12.8)代入(12.11)，我们得到

$$\left\langle \boldsymbol{w}, \boldsymbol{x}_a - r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \right\rangle + b = 0 \quad (12.12)$$

利用内积的双线性（见第3.2节），我们可以得到

$$\langle \boldsymbol{w}, \boldsymbol{x}_a \rangle + b - r\frac{\langle \boldsymbol{w}, \boldsymbol{w} \rangle}{\|\boldsymbol{w}\|} = 0 \quad (12.13)$$

请注意，第一项是由1我们的规模假设，即（$\boldsymbol{w}$，$\boldsymbol{x}_a$）+b=1。从第3.1节的（3.16），我们知道（$\boldsymbol{w}$，$\boldsymbol{w}$）=$\|\boldsymbol{w}\|^2$。因此，第二项减少为$r\|\boldsymbol{w}\|$。利用这些简化，我们得到

$$r = \frac{1}{\|\boldsymbol{w}\|}. \quad (12.14)$$

这意味着我们用法向量$\boldsymbol{w}$推导出了距离r的超平面。乍一看，这个方程是反直觉的，因为我们也可以认为似乎已经用向量$\boldsymbol{w}$的长度推导出了与超平面的距离，但我们还不知道这个向量。一种思考方式是把距离r看作是一个临时变量，我们只在这个推导中使用。因此，在本节的其余部分，我们将用.表示[1]到超平面的距离。在第12.2.3，我们将看到，选择保证金等于我们的前面的$\boldsymbol{w}$=的假设在1第1.2.1.

[1] 的距离是指 投 到超平面时产生的投射误差。

类似于得到(12.9)，我们希望正反两方面的例子离超平面至少有1的距离，这就得到了条件

$$y_n(\langle \boldsymbol{w}, x_n \rangle + b) \geq 1 \quad (12.15)$$

将边际最大化与实例需要在超平面的正确一侧（基于它们的标签）这一事实结合起来，我们可以得到

$$\max_{\boldsymbol{w},b} \frac{1}{\|\boldsymbol{w}\|} \quad (12.16)$$

受制于$y_n(\langle \boldsymbol{w}, \boldsymbol{x}_n \rangle + b) \geq 1$，对于所有 $n = 1, \ldots, N$. (12.17)

而不是像( )中那样最大化准则的倒数。12.16)，我们常常最小化平方准则。我们还经常包括一个常数，[1]这个常数可以使$\frac{1}{2}$不影响最佳的$\boldsymbol{w}$、b，但在我们计算梯度时，会产生一个更整洁的形式。然后，我们的目标变成

[1] 平方法则最小化。凸二次方程序设计的结果

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \quad (12.18)$$

受制于$y_n(\langle \boldsymbol{w}, \boldsymbol{x}_n \rangle + b) \geq 1$对于所有 $n = 1, \ldots, N$. (12.19)

的问题（第12.5).

方程(12.18)被称为*硬边际SVM*。采用

硬边际SVM的原因是

表达“困难”的原因是该表述不允许对边际条件进行任何修改。我们将在本节中看到12.2.4中看到，这

如果数据不是线性可分离的，"硬 "条件可以放宽，以适应违规行为。

*12.2.3 为什么我们可以将保证金设置为*1

在第12.2.1节中，我们认为我们希望最大化某个值r，它代表了最接近超平面的例子的距离。在这一节中12.2.2,我们对数据进行了缩放，使最接近的例子与1超平面的距离相同。在本节中，我们将这两个推导联系起来，并表明它们是等价的。

定理 *最大化***12.1.***保证金r，在这里我们考虑归一化权重，如*(12.10)*,*

$$\begin{aligned} &\text{最}\quad r \\ &\text{大}\quad \boldsymbol{w},b,r \\ &\text{主体是}\underbrace{\text{玩具}_n(\boldsymbol{w},\boldsymbol{x}_n)+b)}_{\text{数据处理}} \underbrace{\geq r}_{} , \quad \underbrace{\|\boldsymbol{w}\|=1}_{\text{统计数据}} r>0 , \end{aligned} \tag{12.20}$$

相当于对数据进行缩放，从而使差值达到统一。

$$\begin{aligned} &\text{闽}\quad \boldsymbol{w},b \quad \frac{1}{2}\underbrace{\text{岩石}^2}_{\text{硕生}} \\ &\text{主体是}\underbrace{\text{玩具}_n(\boldsymbol{w},\boldsymbol{x}_n)+b)}_{\text{数据传输}} \geq 1 。 \end{aligned} \tag{12.21}$$

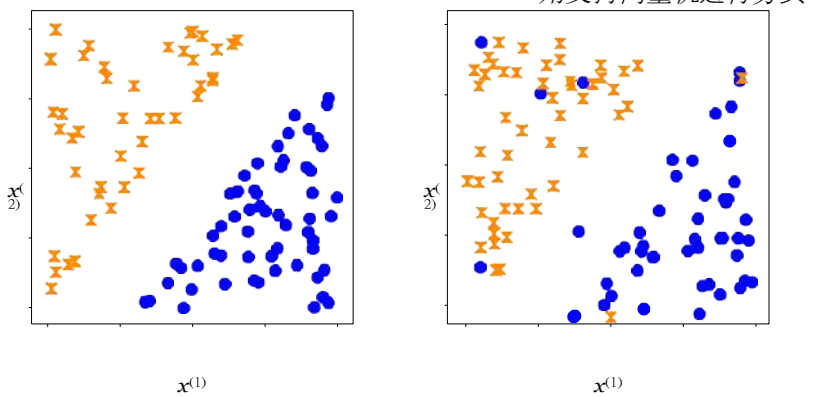
证明 考虑(12.20).由于平方是对非负参数的严格单调反演，如果我们在目标中考虑r^2，则最大值保持不变。由于$\|\boldsymbol{w}\|=1$，我们可以用一个新的权重向量$\boldsymbol{w}'$来重构该方程，该向量没有被明确地归一化使用 $\frac{\boldsymbol{w}'}{\|\boldsymbol{w}'\|}$.我们得到

$$\begin{aligned} &\text{最}\quad r^2 \\ &\text{大}\quad \boldsymbol{w}',b,r \\ &\text{主题}\text{玩具}_n \quad \frac{(\text{wt})}{\|\boldsymbol{w}'\|}, x_n + b \geq r, r>0 。 \end{aligned} \tag{12.22}$$

方程(12.22)明确指出，距离r是正的。因此，我们可以用第一个约束条件除以r，得到的结果是

请注意，$r>0$因为我们假定是线性可分离的，因此不存在除以r的问题。

$$\begin{aligned} &\boldsymbol{w}',b,r^2 \\ &\text{主题}\text{玩具}_n \quad \left\|\frac{\boldsymbol{w}'}{r}\right\|, x_n + \frac{b}{r} \geq 1 , \quad r>0 \end{aligned} \tag{12.23}$$



(a) 可线性分离的数据，有很大的余地　　(b) 非线性可分离的数据

图 12.6
(a) 可线性分离的和
(b) 非线性可分离的数据。

将参数重新命名为$\boldsymbol{w}^{tt}$和b^{tt}。 因为$\boldsymbol{w}^{tt}=$ $\frac{\boldsymbol{w}^{l}}{\|\boldsymbol{w}^{\top}\|r}$，重新排列为$r$给出了

$$\frac{\boldsymbol{w}^{t}}{\|\boldsymbol{w}^{t}\|}=\frac{1}{\|\boldsymbol{u}\|}\cdot\frac{\boldsymbol{w}^{t}}{\|\boldsymbol{w}^{t}\|}=\frac{1}{r}. \tag{12.24}$$

通过将这个结果代入(12.23)，我们得到

$$\max_{\boldsymbol{w},b}\ \frac{1}{\|\boldsymbol{w}\|_2^2} \tag{12.25}$$

主体是$(n(\boldsymbol{w}^{tt},\boldsymbol{x}_n)+b^{tt})\geq 1$。

最后一步是观察，最大化$\frac{1}{\|\boldsymbol{w}\|_2^2}$产生相同的解决方案作为$\|\boldsymbol{w}\|_2^2$的最小化，这就结束了定理的证明。12.1. □

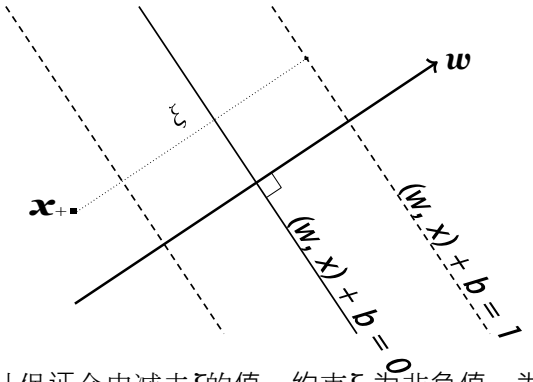
12.2.4 软边际SVM。几何学观点

在数据不是线性可分离的情况下，我们可能希望允许一些例子落在边缘区域，甚至在超平面的错误一侧，如图所示。12.6.

允许一些分类误差的模型被称为*软性*软边际SVM *边际SVM*。在这一节中，我们用几何论证来推导所产生的优化问题。在第12.2.5,我们将利用损失函数的思想推导出一个等效的优化问题。使用拉格朗日乘数（第7.2节），我们将在第7.2节中推导出SVM的双重优化问题。12.3.这个对偶优化问题让我们观察到SVM的第三个解释：作为一个超平面，它将对应于正面和反面数据例子的凸壳之间的线一分为二（第7节）。12.3.2).

松弛变量。

关键的技术思想是引入一个*松弛变量*ξ_n来对应对每个例子-标签对$(\boldsymbol{x}_n, y_n)$，允许一个特定的例子在超平面的边际范围内，甚至在超平面的错误一侧（参考



图中软12.7边际SVM允许例子在边际内或在超平面的错误一侧。松弛变量ξ衡量一个正的例子的距离$\boldsymbol{x}_+$对正边际超平面$\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = 1$ $\boldsymbol{x}$当是在错误的一面。

图12.7).我们$_n$从保证金中减去ξ的值，约束ξ$_n$为非负值。为了鼓励样本的正确分类，我们在目标中加入ξ$_n$。

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\xi_n \tag{12.26a}$$

$$_n(\langle \boldsymbol{w}, \boldsymbol{x}_n \rangle + b) - 1\xi(_n \tag{12.26b}$$

$$\xi_n 0 \tag{12.26c}$$

软边际SVM正则化

对于n=1, . . .与硬边际SVM的优化问题(12.18)的优化问题，这个问题被称为*软边际SVM*。参数C>0可以换取余量的大小和我们拥有的总的松弛量。这个参数被称为*正则化参数*，因为，如

我们将在下一节看到，目标函数（12.26a）中的边际项是一个正则化项。边际项$\|\boldsymbol{w}\|^2$被称为

正则器

正则器，在许多关于数值优化的书籍中，正则器是指ularization参数与该术语相乘（第8.2.3节）。这与我们在本节中的表述不同。在这里，大的C值意味着低的正则化，因为我们给了松弛变量更大的权重，因此给了那些不在边际正确一侧的例子更多的优先权。

这个正则化有其他的参数化，这就是为什么（12.26a）也经常被称为C-SVM。

备注。在软边际SVM的表述中（12.26a），$\boldsymbol{w}$被规范化了，但b没有被规范化。我们可以通过观察正则化项不包含b来了解这一点。

妨碍了理论分析（Steinwart和Christmann，2008，第1章）并降低了计算效率（Fan等人，2008）。 ◆

12.2.5 软边际SVM。损失函数视图

让我们考虑一种不同的方法来推导SVM，遵循经验风险最小化原则（第8.2节）。对于SVM，我们

选择超平面作为假说类，也就是

$$f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b. \qquad (12.27)$$

。

我们将在本节中看到，边际对应的是正则---。

损失项。剩下的问题是，什么是*损失函数*？在CON -Loss函数中

与第九章相比，我们考虑的是回归问题（预测器的输出是一个实数），

在本章中，我们考虑的是二元分类问题（预测器的输出是两个标签之一 $\{+1,1\}$.因此，每个单一例子-标签对的误差/损失函数需要适合于二元分类。例如，用于回归的平方损失（9.10b）不适合于二元分类。

初步分类。

*备注。*二元标签之间理想的损失函数是计算预测和标签之间不匹配的数量。这意味着，对于应用于实例$\boldsymbol{x}_n$的预测器f，我们将输出$f(\boldsymbol{x}_n)$与标签y_n进行比较，如果它们匹配，我们将损失定义为0，如果是1它们不匹配。这用$\mathbf{1}(f(\boldsymbol{x}_n) \neq y_n)$表示，被称为*零-一损失*。不幸的是，零-一损失导致了寻找最佳参数$\boldsymbol{w}$、b的组合 零-一损失优化问题。组合优化问题（与连续优化问题相反

第七章中讨论的）一般来说，解决 这些问题更具有挑战性。 ◆

与SVM相对应的损失函数是什么？考虑预测器$f(\boldsymbol{x}_n)$的输出与标签y之间的n误差。损失描述了训练数据上的误差。一个等价的方法是

推导(12.26a)的方法是使用*铰链损失* shinge loss

$$H(t) = \max\{0, 1 - t\} \text{ 其中} t = yf(\boldsymbol{x}) = y(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b) \text{。} \qquad (12.28)$$

If $f(\boldsymbol{x})$ is on the correct side (based on the corresponding label y) of the hyperplane, and further than distance 1, this means that $t \geq 1$ and the hinge loss returns a value of zero. If $f(\boldsymbol{x})$ is on the correct side but too close to the hyperplane ($0 < t < 1$), the example $\boldsymbol{x}$ is within the margin, and the hinge loss returns a positive value. When the example is on the wrong side of the hyperplane ($t < 0$), the hinge loss returns an even larger value, which increases linearly. In other words, we pay a penalty once we are closer than the margin to the hyperplane, even if the prediction is correct, and the penalty increases linearly. An alternative way to express the hinge loss is by considering it as two linear pieces

$$H(t) = \begin{cases} 0 & \text{如果 } t \geq 1, \\ 1 - t & \text{if } t < 1 \end{cases} \qquad (12.29)$$

as illustrated in Figure12.8. The loss corresponding to the hard margin SVM12.18is defined as

$$H(t) = \begin{cases} 0 & \text{如果 } t \geq 1 \\ \infty & \text{if } t < 1 \end{cases}. \tag{12.30}$$

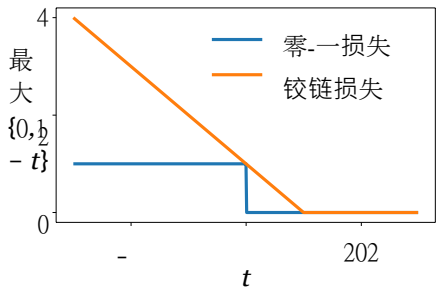


图 铰链损失是一个凸形的零-一损失的上界。

这种损失可以解释为永远不允许在边缘内有任何例子。

对于一个给定的训练集 $(x_1, y_1), \ldots, (x_N, y_N)$，我们寻求最小化总损失，同时用$H_2$-正则化来规范目标（见第8.2.3节）。使用铰链损失（12.28），我们可以得到无约束的优化问题

$$\min_{w,b} \underbrace{\frac{1}{2}\|w\|^2}_{\text{改革开放}} + C\sum_{n=1}^{N} \underbrace{\max\{0, 1 - y_n((w, x_n) + b)\}}_{\text{误区}}. \tag{12.31}$$

正则器损失项

()中的第一个项被称为正则化项或正则器。12.31)中的第一个项被称为正则化项或*正则器*(见第8.2.3节），第二个项被称为*损失项*或*误差*

误差*term*。回顾一下第12.2.4中提到，w项[1]直接[2]来源于边际。换句话说，利润率最大化可以被解释为

正规化 *正规化*。

原则上，( )中的无约束优化问题可以用第7.1节所述的(次)梯度下降法直接解决。12.31)中的无约束优化问题可以用第7.1节所述的(次)梯度下降方法直接求解。为了说明（12.31）和（12.26a）是等价的，请注意铰链损失(12.28) 基本上由两个线性部分组成，如(12.29).考虑一下单个例子-标签对的铰链损失(12.28).我们可以用带有两个约束条件的松弛变量ξ的最小化来代替对t的铰链损失的最小化。以方程的形式。

$$\min_t \max\{0, 1 - t\} \tag{12.32}$$

相当于

$$\min_{\xi,t} \text{闽南语} \tag{12.33}$$

受制于 $\xi \geq 0, \xi \geq 1 - t$。

通过将这个表达式代入(12.31)并重新排列其中的约束条件，我们正好得到软边际SVM(12.26a)。

*备注。*让我们把本节中损失函数的选择与第九章中线性回归的损失函数进行对比。回顾一下第9.2.1节，为了找到最大似然估计值，我们通常最小化

负对数似然。此外，由于高斯噪声的线性回归的似然项是高斯的，每个例子的负对数似然是一个平方误差函数。误差平方函数是在寻找最大似然时被最小化的损失函数
解决办法。 ◆

12.3 双支持向量机

前面几节对SVM的描述，在变量$\boldsymbol{w}$和b方面，被称为原始SVM。回顾一下，我们认为输入$\boldsymbol{x} \in \mathrm{R}^D$有$D$个特征。由于$\boldsymbol{w}$与$\boldsymbol{x}$的维度相同，这意味着操作化问题的参数数量（$\boldsymbol{w}$的维度）与特征数量呈线性增长。

在下文中，我们考虑一个等效的优化问题（所谓的对偶观点），它与特征的数量无关。相反，参数的数量随着训练集中的例子数量的增加而增加。我们在第十章中看到了一个类似的想法，在那里我们用一种不随特征数量增长的方式来表达学习问题。这对于我们拥有比训练数据集中的例子数量更多的特征的问题是很有用的。双重SVM还有一个额外的优势，就是它很容易允许应用内核，我们将在本章的最后看到。"对偶 "这个词经常出现在数学文献中，在这种特殊情况下，它指的是凸对偶性。下面几个小节基本上是凸对偶性的应用，我们在第7.2节中讨论过这个问题。

12.3.1 通过拉格朗日乘法器的凸显二重性

回顾一下原始软边际SVM（12.26a）。我们把变量$\boldsymbol{w}, b,$
和ξ对应的原始SVM的原始变量。在第七章中，我们使用α_n。
0为对应于约束条件（12.26b）的拉格朗日乘数，即例子被正确分类，γ_n.
0为对应于松弛变量的非负性约束的拉格朗日乘数。
见（12.26c）。然后，拉格朗日由以下公式给出

用λ作为拉格朗日乘法器。在本节中，我们遵循SVM文献中通常选择的符号，并使用α和γ。

$$\mathrm{L}(\boldsymbol{w}, b, \xi, \alpha, \gamma) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\alpha_n(\langle \boldsymbol{w}, \boldsymbol{x}_n\rangle + b - y_n + \xi_n) - \sum_{n=1}^{N}\gamma_n\xi_n \tag{12.34}$$

$$-\alpha_n(y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) - 1 + \xi_n) \qquad \gamma_n \xi_n$$

紧张（ 约束(12.26b) 紧张（C） 12.26） 紧张（C） 紧张（C）

通过对拉格朗日(12.34)分别与三个基本变量$\boldsymbol{w}$、b和ξ*进行微分*，我们可以得到

$$\frac{\partial \mathrm{L}}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{n=1}^{N} \alpha_n y_n \boldsymbol{x}_n \quad , \tag{12.35}$$

$$\frac{\partial \mathrm{L}}{\partial b} = -\sum_{n=1}^{N} \alpha_n y_n \, , \tag{12.36}$$

$$\frac{\partial \mathrm{L}}{\partial \xi_n} = C - \alpha_n - \gamma_n \, . \tag{12.37}$$

现在我们通过将这些偏导数中的每一个设为零来找到拉格朗日的最大值。通过设置(12.35)为零，我们发现

$$\boldsymbol{w} = \sum_{n=1}^{N} \alpha_n y_n \boldsymbol{x}_n \, , \tag{12.38}$$

代表人定理

代表者定理实际上是一系列定理的集合，说最小化经验风险的解决方案位于由实例定义的子空间（第2.4.3节）中。

这是*代表者定理*（Kimeldorf and Wahba,1970）的一个特殊实例。等式(12.38)指出，原始的最佳权重向量是例子$\boldsymbol{x}_n$的线性组合。回顾一下第2.6.1节，这意味着优化问题的解在训练数据的范围内。此外，通过设置(12.36)为零意味着最优权重向量是例子的仿射组合。代表者定理对于正则化经验风险最小化的非常普遍的设置是成立的（Hofmann等人，2008；Argyriou和Dinuzzo，2014）。该定理有更普遍的版本（Scholkopf等人，2001），其存在的必要和充分条件可以在Yu等人（2013）中找到。

*备注。*代表者定理(12.38)也提供了"支持向量机"这一名称的解释。例子$\boldsymbol{x}_n$，对其相应的参数$\alpha_n = 0$，对解$\boldsymbol{w}$没有贡献，在所有。其他例子，其中$\alpha_n > 0$，被称为*支持向量*，因为它们"支持"超平面。◆

支持向量

通过将$\boldsymbol{w}$的表达式代入拉格朗日(12.34)，我们得到对偶

$$\mathrm{D}(\xi, \alpha, \gamma) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j (\boldsymbol{x}_i, \boldsymbol{x}_j) - \sum_{i=1}^{N} y_i \alpha_i \sum_{j=1}^{N} y_j \alpha_j (\boldsymbol{x}_j, \boldsymbol{x}_i)$$

$$+ C \sum_{i=1}^{N} \xi_i - b \sum_{i=1}^{N} y_i \alpha_i + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i \xi_i - \sum_{i=1}^{N} \gamma_i \xi_i \, . \tag{12.39}$$

请注意，不再有任何涉及原始变量$\boldsymbol{w}$*的*条款。

通过设置(12.36)为零，我们得到$\sum_{n=1}^{N} y_n \alpha_n = 0$。因此，术语

涉及的b也消失了。回顾一下，内积是对称的，而且

双线性（见第3.2节）。因此，（12.39）中的前两个项是在相同的对象上。这些项（蓝色）可以被简化，我们得到拉格朗日

$$D(\xi, \alpha, \gamma) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j (\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{i=1}^{N} \alpha_i + \sum_{i=1}^{N} (C - \alpha_i - \gamma_i)\xi_i \tag{12.40}$$

这个方程的最后一项是包含松弛变量ξ_i的所有项的集合。通过设置(12.37)为零，我们可以看到(12.40)中的最后一项也是零。此外，通过使用相同的方程并回顾拉格朗日乘法器γ_i是非负的，我们得出结论：$\alpha_i \leqslant C$.我们现在得到SVM的对偶优化问题，它完全以拉格朗日乘法器α_i的形式出现。回顾一下拉格朗日对偶性（定义7.1），我们将对偶问题最大化。这等同于最小化负对偶问题，这样我们

最终得到*双*SVMdual SVM

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j (\boldsymbol{x}_i, \boldsymbol{x}_j) - \sum_{i=1}^{N} \alpha_i \\ &\text{受制于} \quad \sum_{i=1}^{N} y_i \alpha_i = 0 \\ &0 \leqslant \alpha_i \leqslant C \quad \text{对于所有} \quad i = 1, \ldots, N. \end{aligned} \tag{12.41}$$

中的平等约束是由(12.41)中的平等约束是通过设置(12.36)为零。不平等约束$\alpha_i \geqslant 0$是对不平等约束的拉格朗日乘数所施加的条件（第7.2节）。不平等约束$\alpha_i \leqslant C$在上一段已经讨论过了。

SVM中的一组不等式约束被称为 "盒子约束"，因为它们限制了拉格朗日分子的向量$\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_N] \in \mathbb{R}^N$在每个轴上由0和$C$定义的盒子内。这些轴对齐的盒子在数值计算中实现起来特别有效。

求解器（例如，第5章）。

事实证明，

一旦我们得到对偶参数$\boldsymbol{\alpha}$，我们就可以通过使用代表者定理来恢复原始参数$\boldsymbol{w}$ (12.38).让我们把最基本的参数称为$\boldsymbol{w}^*$。然而，仍然存在如何获得参数b^*的问题。考虑一个正好位于边际边界上的例子$\boldsymbol{x}_n$，即（$\boldsymbol{w}^*$，$\boldsymbol{x}_n$）$+b=y_n$。

应该计算所有支持向量的$y_n - \langle \boldsymbol{w}^*, \boldsymbol{x}_n \rangle$，并取这个绝对值差的中值为

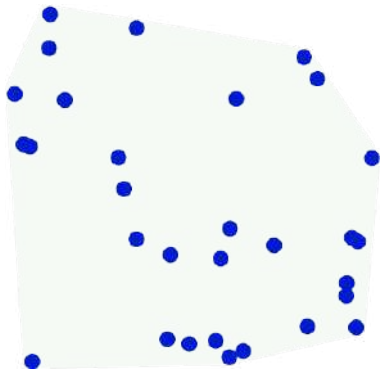
或-1。因此，唯一的未知数是b，它可以通过以下方式计算出来

$$b^* = y_n - \langle \boldsymbol{w}^*, \boldsymbol{x}_n \rangle \,. \tag{12.42}$$

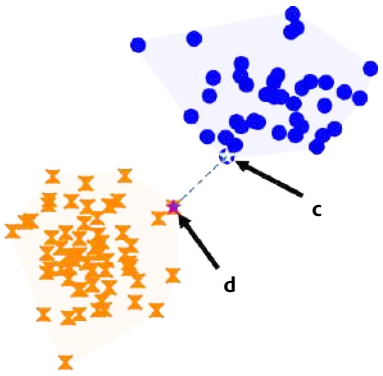
*备注。*原则上说，可能没有完全位于边缘的例子。在这种情况下，我们

恰好位于边际上的例子是指其对偶参数严格位于箱体约束内的例子*用支持向量机进行分类*
$0 < \alpha_i < C$. 这是用以下方法得出的
Karush Kuhn Tucker条件，例如Scholkopf和Smola(2002)。

图 12.9凸面体 (a) 凸面点的船体，其中一些位于边界内；(b) 正负例子周围的凸形船体。



(a) 凸面体。



(b) 正（蓝色）和负（橙色）例子周围的凸面罩。两个凸面集之间的距离是差值向量$\boldsymbol{c}$-$\boldsymbol{d}$的长度。

这方面的推导可以在http://fouryears. eu/2012/06/07/the-svm-bias-term-conspiracy/找到*。 ◆

12.3.2 双重SVM。凸面船体视图

获得对偶SVM的另一种方法是考虑另一种几何学论证。考虑具有相同标签的例子$\boldsymbol{x}_n$的集合。我们希望建立一个包含所有例子的凸集，使其成为最小的可能集合。这被称为 "凸壳"，如图所示12.9.

让我们首先建立一些关于点的凸组合的直觉。考虑两个点$\boldsymbol{x}_1$和$\boldsymbol{x}_2$以及相应的非负权重α_1，α_2，使得$0 \leqslant \alpha_1 + \alpha_2 = 1$。方程$\alpha\boldsymbol{x}_{11}+\alpha_{22}\boldsymbol{x}$描述了每个考虑当我们在$\boldsymbol{x}_1$和$\boldsymbol{x}$之间的直线上的点会发生什么。第三个点$\boldsymbol{x}_3$和一个权重$\alpha_3$，使得$0 \leqslant \alpha_n \leqslant 1$。$\sum_{n=1}^{3} \alpha_n = 1$。这三个点$\boldsymbol{x}_1$、$\boldsymbol{x}_2$、$\boldsymbol{x}_3$的凸组合跨越了一个二维区域。这个区域的*凸体*是由以下几个部分组成的三角形

凸壳

的每一对点所对应的边。当我们添加更多的点，并且点的数量变得大于尺寸的数量时，一些点将在凸壳内，正如我们在图12.9(a)中看到的那样。

一般来说，建立一个凸的凸壳可以通过引入与每个例子$\boldsymbol{x}_n$相对应的非负权重$\alpha_n \geqslant 0$来完成，那么凸壳可以被描述为一个集合

$$\operatorname{conv}(\boldsymbol{X}) = \sum_{n=1}^{N} \alpha_n \boldsymbol{x}_n \quad \text{与} \quad \sum_{n=1}^{N} \alpha_n = 1 \quad \text{和} \quad \alpha_n \geqslant 0, \quad (12.43)$$

。

对于所有n =1, . . .如果对应于正类和负类的两个点云是分开的，那么凸壳就不会重叠。给定训练数据（$\boldsymbol{x}_1$，y_1），. . ., ($\boldsymbol{x}_N$, y_N)，我们形成两个凸面罩，分别对应于正类和负类。我们选取一个点$\boldsymbol{c}$，它位于正面样本集的凸面壳中，并且最接近于负面类的分布。同样地，我们在负面例子集的凸壳中选取一个点$\boldsymbol{d}$，它最接近正面类的分布；见图12.9(b)。我们定义$\boldsymbol{d}$和$\boldsymbol{c}$之间的差异向量为

$$\boldsymbol{w} := \boldsymbol{c} - \boldsymbol{d} 。 \tag{12.44}$$

像前面的情况一样挑选点$\boldsymbol{c}$和$\boldsymbol{d}$，并要求它们彼此最接近，相当于最小化$\boldsymbol{w}$的长度/规范，因此我们最终得到相应的优化问题

$$\arg\min_{\boldsymbol{w}} \|\boldsymbol{w}\| = \arg\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2 。 \tag{12.45}$$

由于$\boldsymbol{c}$必须在正的凸壳中，它可以表示为正的例子的凸组合，即对于非负的系数
α_n^+

$$\boldsymbol{c} = \sum_{n:y_n=+1} \alpha_n^+ \boldsymbol{x}_n . \tag{12.46}$$

In (12.46), we use the notation $n : y_n = +1$ to indicate the set of indices n for which $y_n = +1$. Similarly, for the examples with negative labels, we obtain

$$\boldsymbol{d} = \sum_{n:y_n=-1} \alpha_n^- \boldsymbol{x}_n . \tag{12.47}$$

通过替代(12.44), (12.46)，以及(12.47)代入(12.45)，我们可以得到目标

$$闵_{\boldsymbol{\alpha}} \frac{1}{2} \left\| \sum_{n:y_n=+1} \alpha_n^+ \boldsymbol{x}_n - \sum_{n:y_n=-1} \alpha_n^- \boldsymbol{x}_n \right\|^2 . \tag{12.48}$$

设$\boldsymbol{\alpha}$是所有系数的集合，即$\boldsymbol{\alpha}^+$和$\boldsymbol{\alpha}^-$的连接。回顾一下，我们要求对于每个凸壳，它们的系数之和为1。

$$\sum_{n:y_n=+1} \alpha_n^+ = 1 \quad 和 \quad \sum_{n:y_n=-1} \alpha_n^- = 。 1 \tag{12.49}$$

这意味着约束条件

$$\sum_{n=1}^{N} y_n \alpha_n = . \quad 0 \tag{12.50}$$

这个结果可以通过乘出各个班级来看到

$$\sum_{n=1}^{N} y_n \alpha_n = \sum_{n:y_n=+1} (+1)\alpha_n^+ + \sum_{n:y_n=-1} (-1)\alpha_n^- \tag{12.51a}$$

$$= \sum_{n:y_n=+1} \alpha_n^+ - \sum_{n:y_n=-1} \alpha_n^- = 1 - 1 = 0 \tag{12.51b}$$

目标函数（12.48)和约束条件(12.50)，以及$\boldsymbol{\alpha}$的假设$\mathbf{0}$，给了我们一个有约束的（凸）优化问题。这个优化问题可以被证明是与以下问题相同的双重硬边缘SVM（Bennett和Bredensteiner，2000a）。

缩小的船体

*备注。*为了得到软边际对偶，我们考虑缩小的船体。*缩小的船体*与凸形船体相似，但对系数$\boldsymbol{\alpha}$的大小有一个上限。$\boldsymbol{\alpha}$元素的最大可能值限制了凸形船体的大小。换言之，

对$\boldsymbol{\alpha}$的约束将凸壳缩小到一个较小的体积（Bennett和Bredensteiner，2000b）。 ◆

12.4 果核

考虑到双重SVM的表述(12.41).注意，目标中的内积只发生在例子$\boldsymbol{x}_i$和$\boldsymbol{x}_j$之间，例子和参数之间没有内积。因此，如果我们考虑用一组特征$\boldsymbol{\varphi}(\boldsymbol{x}_i)$来表示$\boldsymbol{x}_i$，对偶SVM的唯一变化就是替换内积。这种模式，即分类方法（SVM）的选择和特征表示$\boldsymbol{\varphi}$（$\boldsymbol{x}$）的选择可以分开考虑，为我们独立探索这两个问题提供了灵活性。在这一节中，我们讨论了$\boldsymbol{\varphi}(\boldsymbol{x})$的表示，并简要介绍了核子的概念，但不涉及技术细节。

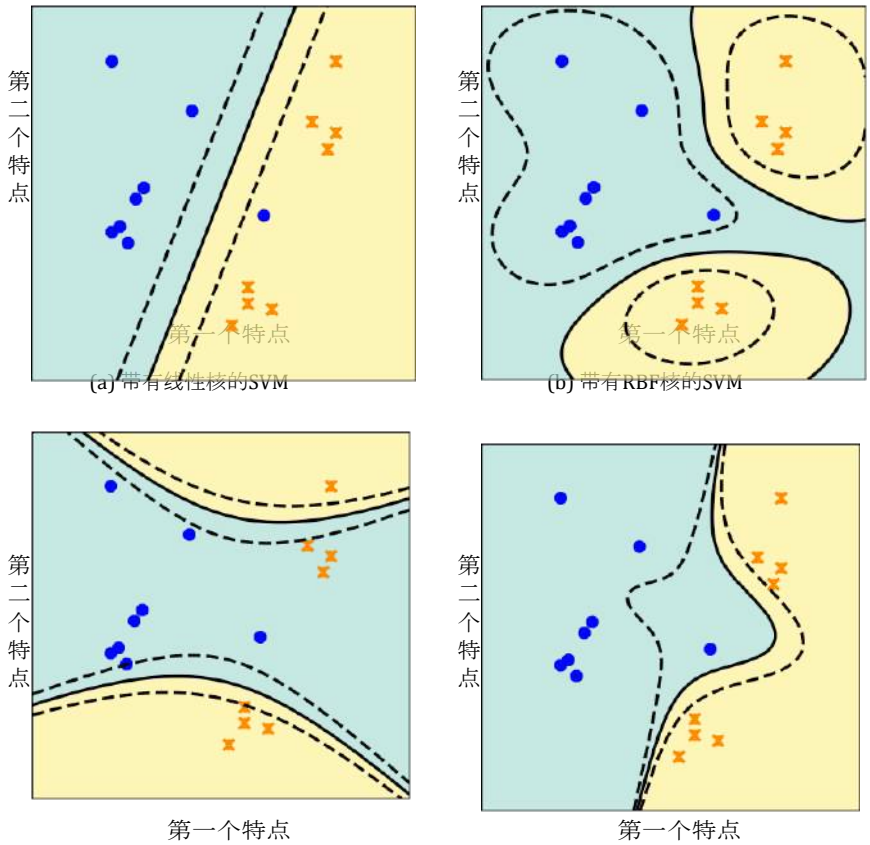
由于$\boldsymbol{\varphi}(\boldsymbol{x})$可能是一个非线性函数，我们可以使用SVM（它假定是一个线性分类器）来构建在例子$\boldsymbol{x}$中是非线性的分类器，这为用户处理非线性可分离的数据集提供了除了软边际之外的第二个途径。事实证明，有很多算法和统计方法都具有我们在对偶SVM中观察到的这种特性：唯一的内积是发生在实例之间的内积。我们没有*明确*定义非线性特征图$\boldsymbol{\varphi}(\ )$，也没有计算实例$\boldsymbol{x}_i$和$\boldsymbol{x}_j$之间产生的内积，而是定义了一个相似度函数$k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ b-

对于某类相似度函数，即*内核*，相似度函数*隐含地*定义了一个非线性特征图φ(-)。根据定义，核是指函数$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$，对于这些函数，存在一个希尔伯特空间$\mathcal{H}$，而$\boldsymbol{\varphi} : \mathcal{X} \to \mathcal{H}$是一个特征图，使得

内核

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \boldsymbol{\varphi}(\boldsymbol{x}_i), \boldsymbol{\varphi}(\boldsymbol{x}_j) \rangle_{\mathcal{H}}。 \tag{12.52}$$

核函数的输入可以非常普遍，不一定限于$\mathbb{R}^D$。



(a) 带有线性核的SVM (b) 带有RBF核的SVM

(c) 带有多项式（2度）内核的SVM (d) 带有多项式（3度）内核的SVM

图为具有不同核的SVM**12.10**。请注意，虽然决策边界是非线性的，但所解决的基本问题是一个线性分离超平面（尽管有一个非线性核）。

有一个唯一的再现核希尔伯特空间与每个核k相关联（Aronszajn,1950;Berlinet和Thomas-Agnan,2004）。在这个唯一的关联，$\varphi(\boldsymbol{x})=k(\cdot,\boldsymbol{x})$被称为*典范特征图*。从内积到核函数的泛化 (12.52)是被称为*内核诀窍*（Schölkopf and Smola， 2002; Shawe-Taylor and Cristianini,2004），因为它隐藏了明确的非线性特征图。

典范特征map

kernel trick

矩阵$\boldsymbol{K} \in \mathbb{R}^{N\times N}$，由内积或应用产生。$k(\boldsymbol{x}_i,\boldsymbol{x}_j)$对数据集的计算，被称为*格拉姆矩阵*，通常只是格拉姆矩阵被称为*内核矩阵*。核必须是对称的和正的核矩阵半定函数，因此每个核矩阵$\boldsymbol{K}$都是对称的和正半定理（第3.2.3节）。

$$\forall \boldsymbol{z}\in\mathbb{R}^N: \quad \boldsymbol{z}\boldsymbol{K}\boldsymbol{z}^\top \geqslant 0 \text{。} \tag{12.53}$$

一些流行的多变量实值数据的核的例子$\boldsymbol{x}_i$

$\mathbb{R}^D$是多项式核、高斯径向基函数核和有理二次核（Schölkopf和Smola，2002；Rasmussen

和Williams,2006）。图12.10说明了不同的核对分离超平面的效果，在一个例子数据集上。请注意，我们仍然在求解超平面，也就是说，假设类的函数仍然是线性的。非线性的表面是由于核函数的原因。

备注。不幸的是，对于刚刚起步的机器学习者来说，"内核 "这个词有多种含义。在本章中，"核 "这个词来自于再现核希尔伯特空间（RKHS）的概念（Aron-szajn,1950;Saitoh,1988）。我们已经讨论了线耳代数中的核的概念（第2.7.3节），其中核是空空间的另一个词。机器学习中 "核 "这个词的第三个常见用法是
内核密度估计中的平滑核（第11.5). ◆

由于显式表示$\boldsymbol{\varphi}(\boldsymbol{x})$在数学上等同于核表示$k(\boldsymbol{x}_i, \boldsymbol{x})_j$，实践者通常会设计核函数，使其能够比显式特征图之间的内积更有效地进行计算。例如，考虑多项式核（Schölkopf and Smola,2002），当输入维度很大时，显式扩展中的项数增长很快（即使是低度的多项式）。核函数只需要对每个输入维度进行一次乘法，这可以大大节省计算量。另一个例子是高斯基函数核(Schölkopf and Smola,2002; Rasmussen and Williams,2006)，其中相应的特征空间是无限维的。在这种情况下，我们不能明确地表示特征空间，但仍然可以使用核来计算一对例子之间的相似性。

内核技巧的另一个有用的方面是，不需要原始数据已经被表示为多变量实值数据。请注意，内积是在函数$\boldsymbol{\varphi}(\ )$的输出上定义的，但并不限制输入为实数。因此，该函数
$\boldsymbol{\varphi}(\cdot)$和内核函数$k(\ ,\ )$可以定义在任何物体上，例如。
集合、序列、字符串、图形和分布（Ben-Hur等人，2008。
Gärtner,2008; Shi et al.,2009; Sriperumbudur et al.,2010; Vishwanathan et al.,2010）。

核的选择以及核的参数，通常是通过嵌套交叉验证来选择的（第8.6.1节）。

12.5 数值解决方案

我们通过研究如何用第7章中提出的概念来表达本章中得出的问题来结束对SVM的讨论。我们考虑两种不同的方法来寻找SVM的最优解。首先，我们考虑SVM8.2.2的损失观点，并将其作为一个无约束的优化问题。然后，我们将原始和双重SVM的约束版本表达为标准形式的二次方程序7.3.2。

考虑到SVM的损失函数视图(12.31).这是一个凸的无约束的优化问题，但铰链损失(12.28)是没有区别的。

可变的。因此，我们采用子梯度方法来解决它。然而，铰链损失几乎在任何地方都是可微的，除了铰链处的一个点$t = 1$。 在这一点上，梯度是一组位于0和1.之间的可能值。因此，铰链损失的子梯度g由以下公式给出

$$g(t) = \begin{cases} -1 & t < 1 \\ 0 & t > 1 \\ [-1, 0] & t = 1 \end{cases} \tag{12.54}$$

利用这个子梯度，我们可以应用第7.1节中提出的优化方法。

原始和对偶SVM的结果都是一个凸的二次函数问题（约束性优化）。请注意，(12.26a)中的原始SVM有优化变量，其大小与输入实例的尺寸D相同。中的对偶SVM(12.41)中的优化变量的大小与实例的数量N相同。

为了用二次编程的标准形式（7.45）表达原始SVM，让我们假设我们使用点积（3.5）作为

内积。我们重新排列原始SVM的方程式（12.26a），回顾一下，从

这样，优化变量都在右边，约束的不等式符合标准形式。这就产生了优化

在第3.2节中，我们用点积这个词来表示内积在欧几里得向量空间。

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\xi_n \tag{12.55}$$

$$\text{受制于} \quad y_n(\boldsymbol{x}_n^\top \boldsymbol{w} + b) \geqslant 1 - \xi_n, \quad \xi_n \geqslant 0$$

$n = 1, \ldots, N$.通过将变量$\boldsymbol{w}$、b、$\boldsymbol{x}_n$串联成一个向量，并仔细收集各条款，我们得到以下软边际SVM的矩阵形式。

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}\begin{bmatrix}\boldsymbol{w}\\ b\\ \boldsymbol{\xi}\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{I}_D & \boldsymbol{0}_{D,N+1}\\ \boldsymbol{0}_{N+1,D} & \boldsymbol{0}_{N+1,N+1}\end{bmatrix}\begin{bmatrix}\boldsymbol{w}\\ b\\ \boldsymbol{\xi}\end{bmatrix} + \begin{bmatrix}\boldsymbol{0}_{D+1,1}\\ C\boldsymbol{1}_{N,1}\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{w}\\ b\\ \boldsymbol{\xi}\end{bmatrix}$$

$$\text{受制于} \quad -\begin{bmatrix}\boldsymbol{Y}\boldsymbol{X} & \boldsymbol{y} & \boldsymbol{I}_N\\ \boldsymbol{0}_{N,D+1} & & I_N\end{bmatrix}\begin{bmatrix}\boldsymbol{w}\\ b\\ \boldsymbol{\xi}\end{bmatrix} \leqslant -\begin{bmatrix}\boldsymbol{1}_{N,1}\\ \boldsymbol{0}_{N,1}\end{bmatrix}. \tag{12.56}$$

在前面的优化问题中，最小化是在帕-------。
参数$[\boldsymbol{w}, b, \boldsymbol{\xi}] \in \mathbb{R}^{D+1+N}$， 我们使用符号。$\boldsymbol{I}_m$代表大小为$m \times m$的身份矩阵，$\boldsymbol{0}_{m,n}$代表大小为$m \times n$的零矩阵，$\boldsymbol{1}_{m,n}$代表大小为$m \times n$的一矩阵。 此外，$\boldsymbol{y}$是标签$[y_1, \ldots, y_N]$的向量 ，$\boldsymbol{Y} = \operatorname{diag}(\boldsymbol{y})$

是一个N乘N的矩阵，其中对角线的元素来自$\boldsymbol{y}$，而

$\boldsymbol{X} \in \mathbb{R}^{N \times D}$是通过串联所有实例得到的矩阵。

我们同样可以对SVM的双重版本进行条件集合(12.41).为了以标准形式表达对偶SVM，我们首先要表达核矩阵$\boldsymbol{K}$，使每个条目为$K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$。如果我们有一个明确的特征表示$\boldsymbol{x}_i$，那么我们就定义$K_{ij} = \boldsymbol{x}_i^\top \boldsymbol{x}_j$。为方便记述，我们引入了一个除了对角线以外到处都是零的矩阵，在那里我们存储标签，也就是$\boldsymbol{Y} = \text{diag}(\boldsymbol{y})$。对偶SVM可以写成

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \frac{1}{2} \boldsymbol{\alpha}^\top \boldsymbol{Y} \boldsymbol{K} \boldsymbol{Y} \boldsymbol{\alpha} - \boldsymbol{1}_{N,1}^\top \boldsymbol{\alpha} \\ &\text{受制于} \begin{bmatrix} -\boldsymbol{y}^\top \\ \boldsymbol{I}_N \\ -\boldsymbol{I}_N \end{bmatrix} \boldsymbol{\alpha} \leq \begin{bmatrix} \boldsymbol{0}_{N+2,1} \\ C\boldsymbol{1}_{N,1} \end{bmatrix} \end{aligned} \tag{12.57}$$

备注。在第7.3.1和7.3.2节中，我们介绍了标准形式的

的约束为不等式约束。我们将把对偶SVM的平等约束表达为两个不平等约束，即：。

$$\boldsymbol{A}\boldsymbol{x} = \text{bis替换为} \quad \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \text{and} \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}。 \tag{12.58}$$

凸优化方法的特定软件实现可以提供表达平等约束

的能力。 ◆

由于对SVM有许多不同的看法，因此有许多方法来解决由此产生的优化问题。这里提出的方法，即用标准的凸优化形式表达SVM问题，在实践中并不经常使用。两个主要的SVM求解器的实现是Chang和Lin（2011）（开源）和Joachims（1999）。由于SVM有一个清晰明确的优化问题，许多基于数值优化技术的方法（Nocedal和Wright，2006）可以被应用（Shawe-Taylor和Sun，2011）。

12.6 进一步阅读

SVM是研究二元分类的众多方法之一。其他方法包括感知器、逻辑回归、Fisher dis- criminant、最近邻、天真贝叶斯和随机森林（Bishop，2006；Murphy，2012）。关于离散序列的SVMs和核的简短教程可以在Ben-Hur等人（2008）中找到。SVM的发展与第8.2节中讨论的经验风险最小化密切相关。因此，SVM具有很强的理论属性（Vapnik，2000；Stein-wart和Christmann，2008）。关于核方法的书（Sch ölkopf和

Smola,2002）包括支持向量机的许多细节和

如何优化它们。一本更广泛的关于核方法的书（Shawe- Taylor and Cristianini,2004）也包括了许多针对不同机器学习问题的线性代数方法。

使用Legendre-Fenchel变换的思想可以得到对偶SVM的另一种推导（第7.3.3节）。这个推导考虑了SVM的无约束表述中的每一个项(12.31)分别计算它们的凸共轭(Rifkin and Lippert, 2007)。对SVM的函数分析观点（也是规整方法观点）感兴趣的读者可以参考Wahba（1990）的工作。核子的理论阐述（Aronszajn，1950；Schwartz，1964；Saitoh，1988；Manton和Amblard，2015）需要有线性算子的基础（Akhiezer和Glazman，1993）。核子的概念已经被推广到Banach空间（Zhang等人，2009）和Kre˘ın空间（Ong等人，2004；Loosli等人，2016）。

请注意，铰链损失有三种等效的表现形式，如(12.28)和(12.29)，以及()中的约束性优化问题。12.33).表述(12.28)在比较SVM损失函数和其他损失函数时经常被使用（Steinwart,2007）。两件式公式(12.29)便于计算子参数，因为每一块都是线性的。第三种表述方式(12.33)，正如在第1节中所看到的12.5,可以使用凸二次方程序设计（第7.3.2节）工具。

由于二元分类是机器学习中一项被研究得很透彻的任务，所以有时也会使用其他的词，如歧视、分离和决定。此外，有三个数量可以作为二元分类器的输出。首先是线性函数本身的输出（通常称为分数），它可以取任何实际值。这个输出可以用来对例子进行排名，二元分类可以被认为是在排名的例子上选取一个阈值（Shawe-Taylor and Cristianini,2004）。第二个量通常被认为是二元分类器的输出，它是在通过非线性函数将其值限制在一个有界范围内（例如在区间[0, 1]）后确定的输出。一个常见的非线性函数是sigmoid函数（Bishop,2006）。当非线性的结果是经过良好校准的(Gneiting and Raftery,2007;Reid and Williamson,2011)，这被称为类别概率估计。二元分类器的第三个输出是最终的二元决策{+1, −1}，这是最被认为是分类器输出的一个。

SVM是一个二进制分类器，并不自然地适合于概率解释。有几种方法可以将线性函数的原始输出（分数）转换为校准的类概率估计值（$P(Y=1 \mid X=\boldsymbol{x})$），这些方法涉及额外的校准步骤（Platt，2000；Zadrozny和Elkan，2001；Lin等人，2007）。

从训练的角度来看，有许多相关的概率方法。我们在本节末尾提到12.2.5
我们在本节末尾提到，有一个再

损失函数和似然之间的关系（也可以比较第8.2和8.3节）。最大似然法对应于

在训练过程中，一个经过良好校准的转换被称为逻辑回归，它来自于一类称为广义线性模型的方法。从这个角度看逻辑回归的细节，可以在Agresti（2002，第5章）和McCullagh和Nelder（1989，第4章）中找到。

当然，我们可以通过使用贝叶斯逻辑回归估计后验分布，对分类器的输出采取更加贝叶斯的观点。贝叶斯观点还包括先验的规格，其中包括设计选择，如与似然的共轭（第6.6.1节）。传统上，我们可以把潜函数视为先验，这导致了高斯过程分类（Rasmussen和Williams,2006，第3章）。

参考文献

Abel, Niels H. *D*1826.*émonstration de l'Impossibilitée de la Résolution Algébrique des Équations Genérales qui Passent le Quatrième Degrè*. Grøndahl and Søn.

Adhikari, Ani, and DeNero, John. 2018. *计算和推理思维。数据科学的基础*.Gitbooks.

Agarwal, Arvind, and Daumé III, Hal. 2010. 共轭优先权的几何学观点。*机器学习*,**81** (1), 99-113.

Agresti, A. 2002.*Categorical Data Analysis*.Wiley.

Akaike, Hirotugu. 1974. 统计模型识别的新视角.*IEEE Transactions on Automatic Control*,**19** (6), 716-723.

Akhiezer, Naum I., and Glazman, Izrail M. *Theory*1993. *of Linear Operators in Hilbert Space*.Dover Publications.

Alpaydin, Ethem. 2010. *Introduction to Machine Learning*.麻省理工学院出版社。

Amari, Shun-ichi. 2016. *Information Geometry and Its Applications*.Springer.Argyriou, Andreas, and Dinuzzo, Francesco. 2014. A Unifying View of Representer 定理。In:*国际机器学习会议论文集*。

Aronszajn, Nachman. 1950. Reproducing Kernels的理论。*Transactions of the American Mathematical Society*,**68** , 337-404.

Axler, Sheldon. 2015. *Linear Algebra Done Right*.Springer.

Bakir, Gökhan, Hofmann, Thomas, Bölkopf, Bernhard, Smola, Alexander J., Taskar, Ben, and Vishwanathan, S. V. N. （编辑）。2007. *预测结构化数据*。麻省理工学院出版社。

Barber, David. 2012. *Bayesian Reasoning and Machine Learning*.剑桥大学新闻界。

Barndorff-Nielsen, Ole.2014.*信息和指数家族。In Statistical The- ory*.Wiley.

Bartholomew, David, Knott, Martin, and Moustaki, Irini.2011.*潜变量模型和因子分析。A Unified Approach*.Wiley.

Baydin, Atılım G., Pearlmutter, Barak A., Radul, Alexey A., and Siskind, Jeffrey M. 2018。机器学习中的自动差异化。A Survey.*Journal of Machine Learning Research*,**18** , 1-43.

Beck, Amir, and Teboulle, Marc.2003.Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization.*运筹学通讯》*, **31**(3), 167- 175.

Belabbas, Mohamed-Ali, and Wolfe, Patrick J. Spectral2009. Methods in Machine Learning and New Strategies for Very Large Datasets.*美国国家科学院院刊》*。0810600105.

Belkin, Mikhail, and Niyogi, Partha.2003.Laplacian Eigenmaps for Dimensionality Reduction and Data Representation.*Neural Computation*,**15** (6), 1373-1396.

Ben-Hur, Asa, Ong, Cheng Soon, Sonnenburg, Sören, Bölkopf, Bernhard, and Rätsch, Gunnar.2008.用于计算生物学的支持向量机和内核。*PLoS计算生物学*, **4**（10）, e1000173。

Bennett, Kristin P., and Bredensteiner, Erin J. 2000a.SVM分类器中的双重性和几何学。In:*国际机器学习会议论文集*。

Bennett, Kristin P., and Bredensteiner, Erin J. 2000b.学习中的几何学。第132-145页。*工作中的几何学》*。美国数学协会。

Berlinet, Alain, and Thomas-Agnan, Christine. 2004. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*.Springer.

Bertsekas, Dimitri P. *Nonlinear*1999. *Programming*.Athena Scientific.Bertsekas,

Dimitri P. *Convex*2009. *Optimization Theory*.Athena Scientific.

Bickel, Peter J., and Doksum, Kjell.2006.*Mathematical Statistics, Basic Ideas and Selected Topics*.Vol. 1.Prentice Hall.

Bickson, Danny, Dolev, Danny, Shental, Ori, Siegel, Paul H., and Wolf, Jack K. 2007. 通过信念传播的线性检测。In:*通信、控制和计算的年度Allerton会议论文集*。

Billingsley, Patrick. 1995. *Probability and Measure*.Wiley.

Bishop, Christopher M. 1995.*Neural Networks for Pattern Recognition*.Clarendon Press.

Bishop, Christopher M. 1999.Bayesian PCA.In:*Advances in Neural Information Processing Systems*.

Bishop, Christopher M. 2006.*Pattern Recognition and Machine Learning*.Springer.Blei, David M., Kucukelbir, Alp, and McAuliffe, Jon D. Variational2017. Inference:A

统计学家的评论.*美国统计学会杂志*，**112**（518），859-877。

Blum, Arvim, and Hardt, Moritz.2015.The Ladder:一个用于数学学习竞赛的可靠的排行榜。In:*International Conference on Machine Learning*.

Bonnans, J. Frédéric, Gilbert, J. Charles, Lemaréchal, Claude, and Sagastizábal, Clau- dia A. *Numerical*2006. *Optimization:Theoretical and Practical Aspects*.Springer.

Borwein, Jonathan M., and Lewis, Adrian S. 2006.凸*面分析与非线性优化*》。2nd edn.Canadian Mathematical Society.

Bottou, Léon. 1998. Online Algorithms and Stochastic Approximations. 第9-42页。*在线学习和神经网络*。剑桥大学出版社。

Bottou, Léon, Curtis, Frank E., and Nocedal, Jorge. 2018. 大规模机器学习的优化方法。*SIAM Review*, **60**(2), 223-311.

Boucheron, Stephane, Lugosi, Gabor, and Massart, Pascal. 2013. *Concentration Inequalities:A Nonasymptotic Theory of Independence.*牛津大学出版社。

Boyd, Stephen, and Vandenberghe, Lieven. 2004. *凸式优化*.剑桥大学出版社。

Boyd, Stephen, and Vandenberghe, Lieven.2018.*Introduction to Applied Linear Algebra*.剑桥大学出版社。

Brochu, Eric, Cora, Vlad M., and de Freitas, Nando. 2009. *昂贵成本函数的贝叶斯优化教程，并应用于主动用户建模和分层强化学习*。Tech. rept.TR-2009-023.英属哥伦比亚大学计算机科学系。

Brooks, Steve, Gelman, Andrew, Jones, Galin L., and Meng, Xiao-Li (eds). 2011. *马尔科夫链蒙特卡洛手册*.Chapman and Hall/CRC.

Brown, Lawrence D. 1986.*统计指数族的基本原理。With Ap- plications in Statistical Decision Theory*.Institute of Mathematical Statistics.

Bryson, Arthur E. A1961. Gradient Method for Optimizing Multi-Stage Allocation Processes.In:*哈佛大学数字计算机及其应用研讨会论文集*。

Bubeck, Sébastien. 2015. 凸面优化。算法和复杂性。 *Founda- tions and Trends in Machine Learning*,**8** (3-4), 231-357.

Bühlmann, Peter, and Van De Geer, Sara. 2011. *高维数据的统计*。斯普林格。

Burges, Christopher.2010.降维。A Guided Tour.*机器学习的基础和趋势*，**2**（4），275-365。

Carroll, J Douglas, and Chang, Jih-Jie.1970.通过 "Eckart-Young "分解位置的*N*次方泛化分析多维标度的个体差异。*Psychometrika*, **35**(3), 283-319.

Casella, George, and Berger, Roger L. 2002.*统计推理*》。Duxbury.Çinlar, Erhan. 2011. *Probability and Stochastics*. Springer.

Chang, Chih-Chung, and Lin, Chih-Jen.2011.LIBSVM:A Library for Support Vector Machines.*ACM Transactions on Intelligent Systems and Technology*,**2** , 27:1-27:27.

Cheeseman, Peter.1985.In Defense of Probability.In:*人工智能国际联合会议论文集*.

Chollet, Francois, and Allaire, J. J. *Deep*2018. *Learning with R*. Manning Publications.Codd, Edgar F. *The*1990. *Relational Model for Database Management*.Addison-Wesley
朗文出版社。

Cunningham, John P., and Ghahramani, Zoubin. 2015. Linear Dimensionality Reduc- tion:Survey, Insights, and Generalizations.*Journal of Machine Learning Research*,**16** , 2859-2900.

Datta, Biswa N. 2010.*Numerical Linear Algebra and Applications*.SIAM.

Davidson, Anthony C., and Hinkley, David V. 1997。*Bootstrap Methods and Their Application*.剑桥大学出版社。

Dean, Jeffrey, Corrado, Greg S., Monga, Rajat, and Chen, et al. 大型2012.分布式深度网络。In:*神经信息处理系统的进展*》。Deisenroth, Marc P., and Mohamed, Shakir. 2012. Expectation Propagation in Gaus- sian Process Dynamical Systems.第2618-2626页。*神经信息处理系统的进展信息处理系统*。

Deisenroth, Marc P., and Ohlsson, Henrik.2011.关于高斯滤波和平滑的一般观点：解释当前和派生新算法。In:*美国控制会议论文集*》。

Deisenroth, Marc P., Fox, Dieter, and Rasmussen, Carl E. Gaussian2015. Processes for Data-Efficient Learning in Robotics and Control.*IEEE Transactions on Pattern Analysis and Machine Intelligence*,**37** (2), 408-423.

Dempster, Arthur P., Laird, Nan M., and Rubin, Donald B. Maximum1977. Likelihood from Incomplete Data via the EM Algorithm.*皇家统计学会杂志*，**39**（1），1-38。

Deng, Li, Seltzer, Michael L., Yu, Dong, Acero, Alex, Mohamed, Abdel-rahman, and Hinton, Geoffrey E. Binary2010. Coding of Speech Spectrograms Using a Deep Auto-Encoder.In:*Interspeech会议论文集*.

Devroye, Luc. 1986. *Non-Uniform Random Variate Generation*.Springer.

Donoho, David L., and Grimes, Carrie. 2003. Hessian Eigenmaps:高维数据的局部线性嵌入技术。*美国国家科学院院刊*，**100**（10），5591-5596。

Fl, Zdeněk. 2009. *Optimal Quadratic Programming Algorithms:With Applications to Variational Inequalities*.Springer.

Douven, Igor.2017.绑架。In:*Stanford Encyclopedia of Philosophy*.斯坦福大学元物理学研究实验室。

Downey, Allen B. 2014。*Think Stats:探索性数据分析*。2nd edn.O'Reilly Media.

Dreyfus, Stuart.1962.变量问题的数值解法.*Journal of Mathematical Analysis and*

Applications,**5** (1), 30-45.
Drumm, Volker, and Weil, Wolfgang. 2001. *Lineare Algebra und Analytische Geometrie*.
讲义，卡尔斯鲁厄大学（TH）。
Dudley, Richard M. *Real* 2002.*Analysis and Probability*.剑桥大学出版社。

Eaton, Morris L. 2007.*Multivariate Statistics:A Vector Space Approach*.数学统计研究所讲义.

埃卡特，卡尔，和杨，盖尔。1936.一个矩阵对另一个低等级矩阵的逼近。*Psychometrika*, **1**(3), 211-218.

Efron, Bradley, and Hastie, Trevor. 2016. *Computer Age Statistical Inference:算法、证据和数据科学*。剑桥大学出版社。

Efron, Bradley, and Tibshirani, Robert J. 1993.*Bootstrap简介*.Chap- man and Hall/CRC.

Elliott, Conal.2009.美丽的差异化。In:*International Conference on Func- tional Programming*.

Evgeniou, Theodoros, Pontil, Massimiliano, and Poggio, Tomaso. 2000. Statistical Learning Theory:A Primer.*International Journal of Computer Vision*,**38** (1), 9-13.

Fan, Rong-En, Chang, Kai-Wei, Hsieh, Cho-Jui, Wang, Xiang-Rui, and Lin, Chih-Jen. 2008. LIBLINEAR: A Library for Large Linear Classification.*机器学习研究杂志*, **9**, 1871-1874.

Gal, Yarin, van der Wilk, Mark, and Rasmussen, Carl E. 2014.稀疏高斯过程回归和潜变量模型中的分布式变量推断。In:*神经信息处理系统的进展*。

Gärtner, Thomas. 2008. *Kernels for Structured Data*. World Scientific.

Gavish, Matan, and Donoho, David L. The 2014.Optimal Hard Threshold for Singular价值是3.4*IEEE Transactions on Information Theory*,**60** (8), 5040-5053.

Gelman, Andrew, Carlin, John B., Stern, Hal S., and Rubin, Donald B. 2004.*Bayesian Data Analysis*.Chapman and Hall/CRC.

Gentle, James E.2004. *Random Number Generation and Monte Carlo Methods*.斯普林格。

Ghahramani, Zoubin. 2015. Probabilistic Machine Learning and Artificial Intelligence.*自然》*, **521**, 452-459。

Ghahramani, Zoubin, and Roweis, Sam T. 1999.Learning Nonlinear Dynamical Systems Using an EM Algorithm.In:*Advances in Neural Information Processing Systems*.MIT Press.

Gilks, Walter R., Richardson, Sylvia, and Spiegelhalter, David J. 1996。*Markov Chain Monte Carlo in Practice*.Chapman and Hall/CRC.

Gneiting, Tilmann, and Raftery, Adrian E. 2007.严格正确的评分规则、预言和估计。*美国统计学会杂志*，**102**（477），359-378。

Goh, Gabriel. 2017.为什么动量真的有用。*Distill*.

Gohberg, Israel, Goldberg, Seymour, and Krupnik, Nahum.2012.*线性算子的踪迹和定子*。Birkhäuser.

Golan, Jonathan S. 2007.*The Linear Algebra a Beginning Graduate Student Ought to Know*.Springer.

Golub, Gene H., and Van Loan, Charles F. 2012.*矩阵计算*.JHU Press.Goodfellow, Ian, Bengio, Yoshua, and Courville, Aaron. 2016. *深度学习*.麻省理工学院新闻界。

Graepel, Thore, Candela, Joaquin Q., Borchert, Thomas, and Her-brich, Ralf.2010.微软Bing搜索引擎中赞助搜索广告的网络规模贝叶斯式点击率预测。In:*Interna- tional Conference on Machine Learning的论文集*。

Griewank, Andreas, and Walther, Andrea.2003.Automatic Differenti- ation简介。In:*应用数学和机械学论文集*。

Griewank, Andreas, and Walther, Andrea.2008.*Evaluating Derivatives, Principles and Techniques of Algorithmic Differentiation*.SIAM.

Grimmett, Geoffrey R., and Welsh, Dominic.2014.*概率。An Introduction*.牛津大学出版社。

Grinstead, Charles M., and Snell, J. Laurie. 1997. *概率论简介*.美国数学学会。

Hacking, Ian. 2001. *Probability and Inductive Logic*.Cambridge University Press.Hall, Peter. 1992. *The Bootstrap and Edgeworth Expansion*.Springer.

Hallin, Marc, Paindaveine, Davy, and Šiman, Miroslav. 2010. 多变量量表瓷砖和多输出回归量子。从*英镑1的*优化到半空间的深度。*Annals of Statistics*,**38** , 635-669.

Hasselblatt, Boris, and Katok, Anatole. 2003. *动力学第一课与最新发展全景》*。剑桥大学出版社。

Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome.2001.*统计学习的要素--数据挖掘、推理和预测*。Springer.

Hausman, Karol, Springenberg, Jost T., Wang, Ziyu, Heess, Nicolas, and Riedmiller, Martin. 2018. 为可转移的机器人技能学习一个嵌入空间。In:*学习表征国际会议论文集》（Proceedings of the International Conference on Learning Representations）*。

Hazan, Elad.2015.在线凸式优化简介.*优化的基础和趋势*, **2**(3-4), 157-325.

Hensman, James, Fusi, Nicolo, and Lawrence, Neil D. 用于大数据的高斯2013.过程。In:*人工智能中的不确定性会议记录*。

Herbrich, Ralf, Minka, Tom, and Graepel, Thore.2007.TrueSkill(TM):A Bayesian Skill Rating System.In:*神经信息处理系统的进展*。

Hiriart-Urruty, Jean-Baptiste, and Lemaréchal, Claude. 2001. *凸面分析的基本原理*.Springer.

Hoffman, Matthew D., Blei, David M., and Bach, Francis.2010.Latent Dirichlet Allocation的在线学习.*神经信息处理系统的进展》*。

Hoffman, Matthew D., Blei, David M., Wang, Chong, and Paisley, John. 2013. Stochas- tic Variational Inference.*机器学习研究杂志*, **14**（1）, 1303-1347。

Hofmann, Thomas, Schölkopf, Bernhard, and Smola, Alexander J. Kernel2008. Meth-. ods in Machine Learning.*Annals of Statistics*,**36** (3), 1171-1220.

Hogben, Leslie. 2013. *Handbook of Linear Algebra*.Chapman and Hall/CRC.

Horn, Roger A., and Johnson, Charles R. *Matrix* 2013.*Analysis*.剑桥大学出版社。

Hotelling, Harold. 1933. 对统计变量的复合体进行主成分分析.*教育心理学杂志*, **24**, 417-441.

Hyvarinen, Aapo, Oja, Erkki, and Karhunen, Juha. 2001. *独立成分分析*.Wiley.

Imbens, Guido W., and Rubin, Donald B. *Causal*2015. *Inference for Statistics, Social and Biomedical Sciences*.剑桥大学出版社。

Jacod, Jean, and Protter, Philip. 2004. *Probability Essentials*.Springer.

Jaynes, Edwin T. *Probability*2003. *Theory:The Logic of Science*.剑桥大学出版社。

Jefferys, William H., and Berger, James O. 1992.Ockham's Razor and Bayesian Analysis.*American Scientist*,**80** , 64-72.

Jeffreys, Harold.1961. The *Theory of Probability*. 牛津大学出版社。

Jimenez Rezende, Danilo, and Mohamed, Shakir. 2015. 用Nor-malizing Flows进行变异推理。In:*国际机器学习会议论文集*。Jimenez Rezende, Danilo, Mohamed, Shakir, and Wierstra, Daan. 2014. 深度生成模型中的随机反向传播和近似推理。In:*Pro-国际机器学习会议论文集》*。

Joachims, Thorsten.1999.*核方法的进展-支持向量学习*.麻省理工学院出版社。Chap.

Making Large-Scale SVM Learning Practical, pages 169-184.
Jordan, Michael I., Ghahramani, Zoubin, Jaakkola, Tommi S., and Saul, Lawrence K. 1999.图形模型的变异方法介绍。*Machine Learn- ing*,**37** , 183-233.

Julier, Simon J., and Uhlmann, Jeffrey K. 1997.卡尔曼滤波法对非线性系统的新扩展。In:*航空航天/国防传感、模拟和控制研讨会论文集*。

Kaiser, Marcus, and Hilgetag, Claus C. 2006.由于神经系统中的长距离投射，非最佳元件放置，但处理路径短。*PLoS Computational Biology*,**2** (7), e95.

Kalman, Dan.1996.A Singularly Valuable Decomposition:The SVD of a Matrix.*Col-lege Mathematics Journal*,**27** (1), 2-23.

Kalman, Rudolf E. A1960. New Approach to Linear Filtering and Prediction Problems. *美国机械工程师协会的交易-基础工程杂志*，**82**（D系列），35-45。

Kamthe, Sanket, and Deisenroth, Marc P. 2018.数据高效的强化学习与概率模型预测控制。In:*人工智能和统计学国际会议论文集*。

Katz, Victor J. *A*2004. *History of Mathematics*.Pearson/Addison-Wesley.

Kelley, Henry J. 1960.最佳飞行路径的梯度理论。*阿斯杂志*，**30**（10），947-954。

Kimeldorf, George S., and Wahba, Grace.1970.A Correspondence between Bayesian Estimation on Stochastic Processes and Smoothing by Splines.*Annals of Mathemat-ical Statistics*,**41** (2), 495-502.

Kingma, Diederik P., and Welling, Max. 2014. Auto-Encoding Variational Bayes. In: *学习表征国际会议论文集》（Proceedings of the International Conference on Learning Representations*）。

Kittler, Josef, and Föglein, Janos. 1984. 多光谱像素数据的情境分类。*图像和视觉计算*, **2**（1），13-29。

Kolda, Tamara G., and Bader, Brett W. Tensor2009. Decompositions and Applications. *SIAM Review*, **51**(3), 455-500.

Koller, Daphne, and Friedman, Nir.2009.*Probabilistic Graphical Models*.麻省理工学院出版社。Kong, Linglong, and Mizera, Ivan. 2012. Quantile Tomography:使用量值与多变量数据.*Statistica Sinica*,**22** , 1598-1610.

Lang, Serge. 1987. *Linear Algebra*.Springer.

Lawrence, Neil D. 2005.Probabilistic Non-Linear Principal Component Analysis with Gaussian Process Latent Variable Models.*Journal of Machine Learning Research*,**6** (Nove.), 1783-1816.

Leemis, Lawrence M., and McQueston, Jacquelyn T. 2008.单变量分布关系。*American Statistician*, **62**(1), 45-53.

Lehmann, Erich L., and Romano, Joseph P. 2005. *Testing Statistical Hypotheses*. 斯普林格。

Lehmann, Erich Leo, and Casella, George. 1998. *The Theory of Point Estimation*.Springer.Liesen, Jörg, and Mehrmann, Volker. 2015. *Linear Algebra*. 斯普林格。

Lin, Hsuan-Tien, Lin, Chih-Jen, and Weng, Ruby C. A2007. Note on Platt's Probabilistic Outputs for Support Vector Machines.*机器学习*, **68**, 267-276.

Ljung, Lennart. 1999. *系统识别:Theory for the User*.Prentice Hall.

Loosli, Gaëlle, Canu, Stéphane, and Ong, Cheng Soon. 2016. Learning SVM in Krein Spaces.*IEEE Transactions of Pattern Analysis and Machine Intelligence*,**38** (6), 1204-1216.

Luenberger, David G. *Optimization*1969. *by Vector Space Methods*.Wiley.

MacKay, David J. C. 1992.Bayesian Interpolation.*Neural Computation*,**4** , 415-

447.MacKay, David J. C. Introduction1998. to Gaussian Processes.第133-165页。
Bishop, C. M. (ed), *Neural Networks and Machine Learning*.Springer.

MacKay, David J. C. *Information*2003. *Theory, Inference, and Learning Algorithms*.
剑桥大学出版社。

Magnus, Jan R., and Neudecker, Heinz.2007.*Matrix Differential Calculus with Applications in Statistics and Econometrics*.Wiley.

Manton, Jonathan H., and Amblard, Pierre-Olivier. 2015. A Primer on Reproducing Kernel Hilbert Spaces.*Foundations and Trends in Signal Processing*,**8** (1-2), 1-126.Markovsky, Ivan. 2011. *Low Rank Approximation:Algorithms, Implementation, Appli-.*

cations.斯普林格。

Maybeck, Peter S. 1979.*Stochastic Models, Estimation, and Control*.Academic Press.McCullagh, Peter, and Nelder, John A. *Generalized*1989. *Linear Models*.CRC Press.

McEliece, Robert J., MacKay, David J. C., and Cheng, Jung-Fu.1998.涡轮解码作为Pearl "信念传播 "算法的一个实例。*IEEE通信领域精选期刊*，**16**（2），140-152。

Mika, Sebastian, Rätsch, Gunnar, Weston, Jason, Schölkopf, Bernhard, and Müller, Klaus-Robert.1999.带核的Fisher判别分析。第41-48页。*信号处理中的神经网络研讨会论文集*。

Minka, Thomas P. 2001a.*A Family of Algorithms for Approximate Bayesian Inference*. 麻省理工学院的博士论文。

Minka, Tom.2001b.Automatic Choice of Dimensionality of PCA.In:*神经信息处理系统的进展*。

Mitchell, Tom. 1997. *机器学习*.McGraw-Hill.

Mnih, Volodymyr, Kavukcuoglu, Koray, and Silver, David, et al. Human2015.-Level Control through Deep Reinforcement Learning.*自然*》，**518**529-533。

Moonen, Marc, and De Moor, Bart. 1995. *SVD和信号处理，III：算法、架构和应用*。Elsevier.

Moustaki, Irini, Knott, Martin, and Bartholomew, David J. *Latent*-2015.*Variable Modeling*.美国癌症协会。第1-10页。

Müller, Andreas C., and Guido, Sarah. 2016. *用Python进行机器学习的介绍。A Guide for Data Scientists*.O'Reilly出版社。

Murphy, Kevin P. *Machine* 2012.*Learning:A Probabilistic Perspective*.麻省理工学院出版社。

Neal, Radford M. 1996.*Bayesian Learning for Neural Networks*.博士论文，多伦多大学计算机科学部。

Neal, Radford M., and Hinton, Geoffrey E. 1999.A View of the EM Algorithm that justifies Incremental, Sparse, and Other Variants.第355-368页。*图形模型中的学习*。麻省理工学院出版社。

内尔森，罗杰。2006.*Copulas 简介*.Springer.Nesterov，Yuri.

2018. *Convex Optimization讲座*.Springer.

Neumaier, Arnold. 1998. Solving Ill-Conditioned and Singular Linear Systems:A Tutorial on Regularization.*SIAM Review*,**40** , 636-666.

Nocedal, Jorge, and Wright, Stephen J. 2006.*Numerical Optimization*.Springer.Nowozin, Sebastian, Gehler, Peter V., Jancsary, Jeremy, and Lampert, Christoph H..

(编辑). 2014. *Advanced Structured Prediction*.麻省理工学院出版社。

O'Hagan, Anthony.1991.Bayes-Hermite Quadrature.*Journal of Statistical Planning and Inference*,**29** , 245-260.

Ong, Cheng Soon, Mary, Xavier, Canu, Stéphane, and Smola, Alexander J. Learn2004.-ing with Non-Positive Kernels。在。*国际机器学习会议论文集*。

Ormoneit, Dirk, Sidenbladh, Hedvig, Black, Michael J., and Hastie, Trevor.2001.学习和跟踪循环的人类运动。In:神经*信息处理系统的进展*。

Page, Lawrence, Brin, Sergey, Motwani, Rajeev, and Winograd, Terry. 1999. *PageRank 引文排名。为网络带来秩序。* Tech. rept.Stanford Info- Lab.

Paquet, Ulrich. 2008. *Bayesian Inference for Latent Variable Models*.博士论文，剑桥大学。

Parzen, Emanuel. 1962. On Estimation of a Probability Density Function and Mode. *数学统计年鉴*, **33**(3), 1065-1076.

Pearl, Judea.1988.*The Probabilistic Reasoning in Intelligent Systems:合理推理的网络*.Morgan Kaufmann.

Pearl, Judea.2009.*因果关系:Model, Reasoning and Inference*.2nd edn.剑桥大学出版社。

皮尔逊，卡尔。1895.对进化的数学理论的贡献。II.均质材料中的斜度变化。*皇家学会哲学论文集A：数学、物理和工程科学*，**186**，343-414。

Pearson, Karl. 1901. 论与空间中的点系统最接近的线和平面. *哲学杂志*，**2**（11），559-572。

Peters, Jonas, Janzing, Dominik, and Scholkopf, Bernhard. 2017. *因果推理的要素。基础和学习算法*。麻省理工学院出版社。

Petersen, Kaare B., and Pedersen, Michael S. *The*2012. *Matrix Cookbook*.Tech. rept. 丹麦技术大学。

Platt, John C. 2000.支持向量机的概率输出和与正则化似然法的比较.In:In: *Advances in Large Margin Classifiers*.

Pollard, David. 2002. *A User's Guide to Measure Theoretic Probability*.剑桥大学出版社。

Polyak, Roman A. 2016.现代优化中的Legendre变换.第437-507页。Goldengorin, B. (ed), *Optimization and Its Applications in Control and Data Sciences*.Springer.

Press, William H., Teukolsky, Saul A., Vetterling, William T., and Flannery, Brian P. *Numerical*2007. *Recipes:科学计算的艺术*。剑桥大学出版社。

Proschan, Michael A., and Presnell, Brett.1998.Expect the Unexpected from Conditional Expectation.*American Statistician*, **52**(3), 248-252.

Raschka, Sebastian, and Mirjalili, Vahid. 2017. *Python Machine Learning:Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow*.Packt Publish- ing.

Rasmussen, Carl E., and Ghahramani, Zoubin.2001.奥卡姆的剃刀。In:*神经信息处理系统的进展*。

Rasmussen, Carl E., and Ghahramani, Zoubin.2003.Bayesian Monte Carlo.In:*神经信息处理系统的发展*。

Rasmussen, Carl E., and Williams, Christopher K. I. 2006.*Gaussian Processes for Machine Learning*.麻省理工学院出版社。

Reid, Mark, and Williamson, Robert C. 2011.二元实验的信息、分歧和风险。*Journal of Machine Learning Research*,**12** , 731-817.

Rifkin, Ryan M., and Lippert, Ross A. Value 2007.Regularization and Fenchel Duality. *机器学习研究杂志*, **8**, 441-479.

Rockafellar, Ralph T. *Convex*1970. *Analysis*.Princeton University Press.

Rogers, Simon, and Girolami, Mark. 2016. *A First Course in Machine Learning*.Chapman and Hall/CRC.

Rosenbaum, Paul R. *Observation*2017. *and Experiment:因果推理简介*》。哈佛大学出版社。

Rosenblatt, Murray. 1956. Remarks on Some Non-parametric Estimates of a Density Function.*Annals of Mathematical Statistics*,**27** (3), 832-837.

Roweis, Sam T. PCA 和 SPCA 的 EM1998. 算法。 第626-632页。*神经信息处理系统的进展*.

Roweis, Sam T., and Ghahramani, Zoubin. 1999. A Unifying Review of Linear Gaussian Models.*Neural Computation*,**11** (2), 305-345.

Roy, Anindya, and Banerjee, Sudipto. 2014. *Linear Algebra and Matrix Analysis for Statistics*.Chapman and Hall/CRC.

Rubinstein, Reuven Y., and Kroese, Dirk P. *Simulation*2016. *and the Monte Carlo Method*.Wiley.

Ruffini, Paolo.1799.*Teoria Generale delle Equazioni, in cui si Dimostra Impossibile la Soluzione Algebraica delle Equazioni Generali di Grado Superiore al Quarto*.Stamperia di S. Tommaso d'Aquino.

Rumelhart, David E., Hinton, Geoffrey E., and Williams, Ronald J. Learning 1986.Representations by Back-Propagating Error.*自然》*，**323**（6088），533-536。

Sæmundsson, **S**r, Hofmann, Katja, and Deisenroth, Marc P. Meta 2018.Rein-forcement Learning with Latent Variable Gaussian Processes.In:*人工智能中的不确定性会议记录*。

Saitoh, Saburou.1988.*复制核的理论及其应用*.Longman Scientific and Technical.

Sä**r**kkä, Simo. 2013. *Bayesian Filtering and Smoothing*. 剑桥大学出版社。**S**chölkopf, Bernhard, and Smola, Alexander **J**. *Learning* 2002.*with Kernels - Support Vector Machines, Regularization, Optimization, and Beyond*.麻省理工学院出版社。

Schölkopf, Bernhard, Smola, Alexander J., and **M**üller, Klaus-Robert. 1997. 核心主成分分析。载于。*人工神经网络国际会议论文集*。

Schölkopf, Bernhard, Smola, Alexander **J**., and **M**üller, Klaus-Robert. 1998. 非线性成分分析是一个核心特征值问题。*神经计算*，**10**（5），1299-1319。

Schölkopf, Bernhard, Herbrich, Ralf, and Smola, Alexander J. A 2001.Generalized Representer Theorem。In:*计算学习理论国际会议论文集*。

Schwartz, Laurent.1964.Sous Espaces Hilbertiens d'Espace Vectoriels Topologiques et Noyaux Associés. *Journal d'Analyse Mathématique*,**13** , 115-256.

Schwarz, Gideon E. Estimating1978. the Dimension of a Model.*Annals of Statistics*, **6**(2), 461-464.

Shahriari, Bobak, Swersky, Kevin, Wang, Ziyu, Adams, Ryan P., and De Freitas, Nando. 2016. Taking the Human out of the Loop:贝叶斯优化的回顾。*IEEE论文集*, **104**(1), 148-175.

Shalev-Shwartz, Shai, and Ben-David, Shai. 2014. *了解机器学习。From Theory to Algorithms*.剑桥大学出版社。

Shawe-Taylor, John, and Cristianini, Nello. 2004. *Kernel Methods for Pattern Analysis*. 剑桥大学出版社。

Shawe-Taylor, John, and Sun, Shiliang.2011. 支持向量机的优化方法回顾.*Neurocomputing*,**74** (17), 3609-3618.

Shental, Ori, Siegel, Paul H., Wolf, Jack K., Bickson, Danny, and Dolev, Danny.2008.线性方程系统的高斯信念传播求解器.第1863-1867的。*信息理论国际研讨会论文集》*。

Shewchuk, Jonathan R. 1994.*没有痛苦的共轭梯度法介绍*.

Shi, Jianbo, and Malik, Jitendra. 2000. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,**22** (8), 888-905.

Shi, Qinfeng, Petterson, James, Dror, Gideon, Langford, John, Smola, Alexander J., and Vishwanathan, S. V. N. 2009.结构化数据的哈希核。*机器学习研究杂志》*，2615-2637。

Shiryayev, Albert N. *Probability*1984..Springer.

Shor, Naum Z. *Minimization*1985. *Methods for Non-Differentiable Functions*.Springer.Shotton, Jamie, Winn, John, Rother, Carsten, and Criminisi, Antonio. 2006. Texton-Boost:Joint Appearance, Shape and Context Modeling for Multi-Class Object Recog- nition and Segmentation.In:*欧洲计算机会议论文集*

愿景。
Smith, Adrian F. M., and Spiegelhalter, David. 1980. Bayes Factors and Choice Criteria for Linear Models.*皇家统计学会杂志B*, **42**（2）, 213-220。

Snoek, Jasper, Larochelle, Hugo, and Adams, Ryan P. Practical2012. Bayesian Optimization of Machine Learning Algorithms.In: *神经信息处理系统的进展*。

Spearman, Charles. 1904. "一般智力，"客观地确定和测量。*美国心理学杂志*，**15**（2），201-292。

Sriperumbudur, Bharath K., Gretton, Arthur, Fukumizu, Kenji, Schölkopf, Bernhard, and Lanckriet, Gert R. G. 2010.Hilbert Space Embeddings and Metrics on Probability Measures.*Journal of Machine Learning Research*,**11** , 1517-1561.

Steinwart, Ingo. 2007. 如何比较不同的损失函数及其风 险 . *构造性逼近*, **26**, 225-287.

Steinwart, Ingo, and Christmann, Andreas. 2008. *支持向量机*.Springer.Stoer, Josef, and Burlirsch, Roland. 2002. *Introduction to Numerical Analysis*.Springer.Strang, Gilbert. 1993. The Fundamental Theorem of Linear Algebra.*The American 数学月刊*, **100**(9), 848-855.

Strang, Gilbert. 2003. *Introduction to Linear Algebra*.Wellesley-Cambridge Press.Stray, Jonathan. 2016. *The Curious Journalist's Guide to Data*.Tow Center for Digital 哥伦比亚大学新闻研究生院的新闻学。

Strogatz, Steven. 2014. 为困惑者和受创伤者写数学。*美国数学学会通知*，**61**（3），286-291。

Sucar, Luis E., and Gillies, Duncan F. 1994.Probabilistic Reasoning in High-Level Vision.*Image and Vision Computing*,**12** (1), 42-60.

Szeliski, Richard, Zabih, Ramin, and Scharstein, Daniel, et al. A2008. Comparative Study of Energy Minimization Methods for Markov Random Fields with Smoothness-Based Priors.*IEEE Transactions on Pattern Analysis and Machine In- telligence*,**30** (6), 1068-1080.

Tandra, Haryono.2014. 变量定理与微积分基本定理在勒贝斯格积分中的关系.*数学教学*，**17**（2），76-83。

Tenenbaum, Joshua B., De Silva, Vin, and Langford, John C. 2000.A Global Geometric Framework for Nonlinear Dimensionality Reduction.*科学*，**290**（5500），2319-2323.

Tibshirani, Robert. 1996. Regression Selection and Shrinkage via the Lasso.*Journal of the Royal Statistical Society B*,**58** (1), 267-288.

Tipping, Michael E., and Bishop, Christopher M. 1999.Probabilistic Principal Component Analysis.*皇家统计学会杂志。Series B*, **61**(3), 611-622.

Titsias, Michalis K., and Lawrence, Neil D. 2010.Bayesian Gaussian Process Latent Variable Model.In:*人工智能与统计国际会议论文集》*。

Toussaint, Marc.2012. *关于梯度下降的一些说明*。 https://ipvs.informatik.uni-stuttgart.de/mlr/marc/notes/gradientDescent.pdf。

Trefethen, Lloyd N., and Bau III, David. 1997. *Numerical Linear Algebra*.SIAM.Tucker, Ledyard R. Some1966. Mathematical Notes on Three-Mode Factor Analysis. *Psychometrika*, **31**(3), 279-311.

Vapnik, Vladimir N. *Statistical*1998. *Learning Theory*.Wiley.

Vapnik, Vladimir N. An1999. Overview of Statistical Learning Theory.*IEEE Transactions on Neural Networks*,**10** (5), 988-999.

Vapnik, Vladimir N. 2000.*统计学习理论的性质》*。Springer.Vishwanathan, S. V. N., Schraudolph, Nicol N., Kondor, Risi, and Borgwardt, Karsten M. Graph 2010.Kernels.*机器学习研究杂志*, **11**, 1201-1242.

von Luxburg, Ulrike, and Schölkopf, Bernhard. 2011. 统计学习理论。模型、概念和结果。第651-706页。D. M. Gabbay, S. Hartmann,

J.Woods (ed), *Handbook of the History of Logic*, vol. Elsevier10..

Wahba, Grace. 1990. *Spline Models for Observational Data*.工业与应用数学学会.

Walpole, Ronald E., Myers, Raymond H., Myers, Sharon L., and Ye, Keying. 2011. *Probability and Statistics for Engineers and Scientists*.Prentice Hall.

Wasserman, Larry. 2004. *All of Statistics*.Springer.

Wasserman, Larry.2007.*All of Nonparametric Statistics*.Springer.Whittle, Peter. 2000. *Probability via Expectation*.Springer.

Wickham, Hadley.2014.Tidy Data.*统计软件杂志*, **59**, 1-23.Williams, Christopher K. I. 1997. 无限网络的计算。In:进展*情况神经信息处理系统*。

Yu, Yaoliang, Cheng, Hao, Schuurmans, Dale, and Szepesvári, Csaba. 2013. 代表者定理的特征。在。*国际机器学习会议论文集*。

Zadrozny, Bianca, and Elkan, Charles. 2001. Obtaining Calibrated Probability Estimates from Decision Trees and Naive Bayesian Classifiers.In:*国际机器学习会议论文集》（Proceedings of the International Conference on Machine Learning*）。

Zhang, Haizhang, Xu, Yuesheng, and Zhang, Jun. 2009. 重现机器学习的核Banach空间 *Journal of Machine Learning Research*,**10** , 2741-2775.Zia, Royce K. P., Redish, Edward F., and McKay, Susan R. Making2009. Sense of the Legendre变换。*美国物理学杂志*, **77**（614）, 614-622。