数据科学与工程数学基础

(第 1 版)

黄定江　编著

华东师范大学
上海

内 容 简 介

本书介绍了数据科学、人工智能和机器学习领域所需的核心数学基础知识，涉及矩阵计算、概率和信息论基础、优化基础。内容按照从模式分析到数据分析再到数学基础的思路来组织，围绕数据分析系统的核心构成：表示、模型和学习形成数据线和数学线两条线。数据线按照数据分析的处理流程、通过大量翔实的案例作为导引，引出所需数学；数学线紧扣数据线，按照知识内容发生的内在自然逻辑顺序展开。两者相辅相成，构成从具体到抽象、从抽象到具体的闭环。本书在数据科学的定位类似于《离散数学》在计算机科学的定位，配有相当数量的习题，可作为数据科学与大数据技术、人工智能、计算机科学和软件工程等相关专业的本科生或研究生的数学基础课程教材或参考书，也可作为学术和工业界科技人员了解和应用数据科学与大数据技术数学基础的参考手册。

序言

数据科学与工程核心课程的系列教材终于要面试了，这是一件鼓舞人心的事。作为华东师范大学数据学院的发起者和见证人，核心课程和系列教材一直是我心心念念的事情。值此教材出版发行之际，我很高兴能被邀请写几句话，做个回顾，分享一些感悟，也展望一下未来。

借着大数据热的东风，依托何积丰院士在 2007 年倡导成立的华东师范大学海量计算研究所，2012 年 6 月在时任 SAP 公司 CTO 史维学博士（Dr. Vishal Sikka）的支持下，我们成立了华东师范大学云计算与大数据研究中心。2013 年 9 月，学校发起成立作为二级独立实体的数据科学与工程研究院，开始软件工程一级学科下自设的数据科学与工程二级学科的博士和硕士研究生培养。在进行研究生培养的探索过程中，我们深切感受到我们的本科生的培养需要反思和改革。因此，到了 2016 年 9 月，研究院改制成数据科学与工程学院，随后就招收了数据科学与工程专业的本科生，第一届本科生已于 2020 年毕业。这是我们学院和专业的简单历史。经过这么几年的实践和思考，我们越发坚信当年对“数据科学与工程”这一名称的选择，“数据学院”和“数据专业”已经得到越来越多的认可，学院的师生也逐渐接受“数据人”这一称呼。

这里我想分享以下几方面的感悟：为什么要办数据专业？怎么办数据专业？教材为什么很重要？对人才培养有什么贡献？

为什么要办数据专业？数据是新能源，这是大家耳熟能详的一句话。说到能源，我们首先想到的是石油，所以大家就习惯把数据比喻成石油。但是，在我们看来，“新能源”对应的英文说法应该是“New Power”。“Data is Power”，这是我们的基本信念，也是我们为什么要办数据学院的根本动机。数据是人类文明史上的第三个重要的 Power，蒸汽能（Steam Power）和电能（Electric Power）引发了第一次和第二次工业革命。如果说蒸汽能和电能造就了从西方开始的两百多年的工业文明，数据能（Data Power）将把人们带入数字文明时代。数据是数字经济发展的重要的生产要素，这个生产要素不同于土地、劳动力，也不同于资本、技术。如果要给数据找一个恰当的比拟，也许只有十九世纪末伟大的发明家尼古拉 • 特斯拉发明的交流电。数据是新时代的交流电，就像上个世纪，交流电给世界带来的深刻变化一样，因为人们对数据能（Data Power）认识的提高，我们将进入一个“未来已来，一切重构”的时代。数据学院就像一百多年前的电力学院或电气学院。

怎么办数据专业？我们数据学院脱胎于软件工程学院，在此以前还有计算机科学与工程学院，数据相关的研究和偏向管理和图书情报方向的信息系统学科和专业也密切相关，应用数学、概率统计更是数据分析和处理的理论基础，不可或缺。到底什么样的专业才算是数据专业？起初的时候，这对我们来说基本上可以说是一个“灵魂拷问”。为此，我们发起成立了国内十五所高校三十多位知名教授组成立“高校数据科学与工程专业建设协作组”。我们相信，有了先进的理念，再加上集体的力量，数据专业建设的探索之路就能走通。协作组已经召开了四次研讨会，

确定了称为 CST 的专业建设路线图，C 代表 Curriculum（培养计划），S 代表 Syllabus（课程大纲），T 代表 Textbook（教材建设）。在得知我们的工作后，ACM/IEEE 计算机工程学科规范主席 John Impagliazzo 教授邀请我们参与了 ACM/IEEE 数据科学学科规范的制定。协作组达成共识：专业课程分为基础课、核心课、方向课三类，核心课是体现专业区分度的一组课程。与数据专业（DSE）最相近的专业就是计算机科学与工程（CSE）与软件工程（SE）两个专业，我们确定的第一批 DSE 区别于 CSE 和 SE 的 8 门核心课程是：数据科学与工程导论，数据科学与工程数学基础、数据科学与工程算法基础，应用统计与机器学习、当代人工智能，云计算系统、分布式计算系统、当代数据管理系统。随后我们又确定两门课纳入这个系列，分别是：区块链导论——原理、技术与应用，数据中台初阶教程。数据专业作为一个新专业，三类课程的边界还不清晰，我们重点关注核心课程上面，核心课有遗漏的知识点可以纳入基础课或方向课。这样可以保证知识体系的完整性，简单起步，快速迭代。随着实践和认识的深入，逐渐明晰三类课程的边界，形成完善的培养计划。

教材为什么很重要？建设好一个专业，培养计划和课程体系固然很重要，但落实在根本上是教材。一套好的教材是建成一个好的专业的前提，放眼看去，无论是国内还是国外，无论是具体高校还是国家区域层面，这都是不争的事实，好的专业都有成体系的好的教材。当然，现在的教材已经不仅仅指的是单纯的一本教科书，还有深层次的内容，比如说具体的教学内容和教学方式。我们都知道，教材是知识的结晶，是站到巨人肩膀上去的台阶。在自然科学领域，确实如此，一百年前我们民族的仁人志士呼唤“赛先生”，在中华大地上科学的传播带来了翻天覆地的变化。在更广泛的领域，教材也还是技术、工艺和文化的传承，是产业发展的助推器。拿信息技术来举例，技术的源头和产业的发祥地都在美国和欧洲，像 IBM、Lucent、Oracle 等跨国企业在我国商业上取得的巨大成功无一不与他们重视教材开发密切相关。试想一下，我们的学生在课堂上学的都是他们的研究和研发的东西，等走上工作岗位，自然会对熟悉的技术和系统有亲近感，这应该是产业或产品生态最重要的一个环节。本世纪以来，随着互联网的蓬勃发展，人们已经深刻认识到，互联网改变世界。在人类的文明史上，没有任何一项科研成果像互联网这样深刻地改变人，改变世界。互联网之所以能改变世界，是因为它真正发挥了数据的威力。互联网实现了信息技术发展从“以计算为中心”到“以数据为中心”的路径转变。用“昔日王谢堂前燕，飞入寻常百姓家”来形容很多我们以前甚至当前教材上的一些内容，可能毫不为过。以互联网为代表的新型产业的发展，极大地推动了技术的进步，我们已经到了可以编写自己的教材，形成自己的技术体系和科学理论体系的时候了。我们是现代科学的后来者，已经习惯了从科学到技术再到应用的路径，现在有了成功的应用，企业也发展出了领先的技术，学界可以在此基础上发展出技术体系和科学理论体系，应用、技术和科学的联动才是真正的创新之路。

对人才培养有什么贡献？在信息技术领域，迄今为止我们更多的是参考或沿袭了西方发达国家的培养计划和教材体系，在改革开放以来的四十年，这种“拿来主义”的做法很有效，培养了大量的人才，推动了我国的社会经济发展。但总的来说，我们的高校在这一领域更像是在培养“驾驶员”，培养开车的人，现在到了需要我们来培养自己的造车人的时候了。技术发展趋

势如此，国际形势也对我们提出了这样的要求。我们处在一个大变局的时代，世界充满不确定性，开放和创新是应对不确定性的不二之选。创新成为人才培养的第一性原理，更新观念，变革教育，卓越育人是我们华东师范大学新时期人才培养的基本理念。人才培养是大学的第一要务，科学研究、社会服务和文化传承是大学的另外三大职能，大学通过这三大职能的实现可以更好地服务于人才培养。这也是数据专业核心课程系列教材的建设的指导思想，我们计划久久为功的这一件事是我们践行这一理解的一个小小的行动。

最后，要特别表示感谢，感谢华东师范大学和高等教育出版社的支持和鼓励，感谢数据专业建设协助组的各位老师们的通力协作和辛勤劳动，也要感谢数据学院师生的信任和付出。心有所信，方能行远；因为相信，所以看见。希望作为探路者的艰辛能成为大家学术和事业生涯中的一笔重要财富。

“The best way to predict the future is to invent it”——Alan Kay

周傲英
华东师范大学
2020 年 11 月于上海

前言

本书主要介绍数据科学、机器学习和人工智能所依赖的数学基础，包括：线性代数、概率与信息论和优化理论。我们知道数据的表示需要向量，机器学习中函数模型的权重可以用矩阵来表示；数据中的不确定性或随机性描述通常由概率来刻画，大数定律为统计机器学习模型的成功提供了理论基础；而优化为最终训练出一套可靠的模型参数提供了强大的数值计算支撑。

尽管线性代数、概率与信息论和优化理论的很多内容研究已经持续了一个世纪以上，但是直到近二十多年来人们才发现它们已然成为数据科学建模求解的核心数学基础，比如，奇异值分解的广泛应用、最大似然和最大后验的成功运用、凸优化方法可靠和迅速的求解等等，使得这些理论和方法足以嵌入到基于计算机程序运行的数据分析和人工智能算法设计之中。

但是就像很多其它学科利用线性代数、概率统计和优化作为基础工具一样，现实世界的数据问题如何转换为一个线性代数计算或概率估计或优化求解问题是不容易的，特别是数据科学领域的问题与其它领域的不同之处在于它对这三部分知识的需求是如此的交错复杂和浑然一体，比如，矩阵既可以用于表示数据，但它也是函数模型变换的一部分；协方差矩阵巧妙的融合了概率和线性代数，把方差和矩阵捏合在一起，从而能用作主成分分析的建模对象；数据科学大部分优化问题是非凸的，判断它是不是凸的或者将某个问题表述为凸优化的形式是比较困难的，这可以部分地借助对称正半定矩阵的概念等来实现。

本书目的

因此，本书的主要目的是帮助读者快速理清和掌握数据科学、机器学习和人工智能领域所需的相关数学知识，即表示数据所需的向量和矩阵的概念与运算，以及数值线性代数的四大核心议题；构建数据概率模型所需的概率基础和相关的统计和信息论准则；判断、描述以及求解凸优化问题的方法和背景知识等等。全书包括四个部分，共 12 章内容。

第一部分：绪论。也即第 1 章，主要介绍数据科学与工程数学基础在数据科学与大数据技术专业中的定位、应用背景、服务学科领域和主要数学内容的构成以及相关的数学基础简史，使读者对本书有初步的了解。这一章，我们会对从图像感知到自然语言处理再到数据分析与机器学习做一个简要的概览，让读者能够从“应用驱动”的角度来了解数据科学所涉及的和所需的数学基础，为全书的数据案例和数学内容展开做好铺垫。

第二部分：数据的低维表示——矩阵分析。涵盖了从第 2 章到第 6 章的主要内容。

第 2 章主要按数据的向量和矩阵表示、数据的向量和矩阵空间、数据空间的关系以及数据空间上代数结构建立的过程来具体介绍数据科学与工程所涉及的向量和矩阵的计算所需的基本知识，包括向量和矩阵基本概念和运算、向量空间、线性映射和线性变换、矩阵的基本特征和矩阵的特征分解等。

第 3 章介绍了线性代数的几何：度量和投影，包括向量的范数和内积、矩阵的范数和内积、

矩阵的四个基本子空间、投影以及特殊的正交矩阵等。这些概念有助于我们从几何的角度来理解线性代数的基本概念以及在数据科学中的应用。如范数和内积将被用作定义数据的各种相似性度量，以及防止数据模型过拟合的正则化手段；投影既是一个几何量，也是一个变换，在数据科学的降维任务中具有本质的作用。

第 4 章介绍了五种常用的矩阵分解方法，包括 LU（三角）分解、QR（正交）三角分解、谱（特征）分解、Cholosky 分解和奇异值分解等。线性代数包含很多有趣的矩阵，如：对角阵、三角矩阵、正交矩阵、对称矩阵、置换矩阵、投影矩阵和关联矩阵等等。在这些矩阵当中对称正（半）定矩阵是核心，因为数据科学与机器学习中大部分矩阵都是非方阵，而非方阵总是可以通过与其自身的转置相乘得到对称正（半）定矩阵。对称正（半）定矩阵有正（非负）的特征值，并且有正交的特征向量，它也可以表示成一些秩 1 矩阵的线性组合，因此可以方便的用于做低秩近似计算。在机器学习中，我们主要处理的是这些大规模的对称正定矩阵或复杂的非方阵矩阵，需要借助矩阵分解的技术，特别是奇异值分解，把它表示为对角阵、三角阵和正交矩阵的乘积等等，然后利用这些特殊且简单的矩阵实现复杂矩阵的特征值等矩阵基本特征的快速计算，并用于数据压缩、数据降维、稀疏优化以及低秩矩阵恢复问题的求解等等，这对帮助理解原本复杂的高维数据矩阵的结构和性质具有重要的作用。

第 5 章介绍了数值线性代数三大核心主题内容，包括线性方程组的求解、最小二乘问题和特征值的求解。数据科学中的很多问题最终都归结为上述三类问题的求解，因此这一章主要介绍线性方程组的类型和解的结构，引入基于矩阵分解的线性方程组和最小二乘问题的求解方法，并讨论解的敏感性，这些内容将与后续优化问题求解、数据科学中的线性回归问题相联系。此外，还介绍了大规模矩阵求解特征值的一些计算方法，包括幂迭代法，这已被广泛应用于数据科学中的搜索技术 pagerank 的矩阵特征值计算。

第 6 章主要介绍向量和矩阵微分。包括向量和矩阵函数，以及数据科学和统计机器学习中常见的各种函数（包括模型函数、损失函数和目标函数等）、深度神经网络中函数的构造（包括模型函数和激活函数等），梯度和高阶导数的定义和性质、向量值函数和矩阵函数的梯度和求解方法以及用迹微分法求梯度的方法，并引入深度网络中的反向传播和自动微分求解方法。这一章介绍的函数模型是数据科学中两大类型的模型之一。这些内容将在优化方法和数据科学中的各种优化问题求解中反复使用。

第三部分：数据的随机表示——概率和信息论。涵盖了从第 7 章至第 9 章的内容。

第 7 章回顾概率论的基本概念，建立用随机变量和分布来描述数据中的不确定性的思想。包括概率论的基本概念、随机变量及其分布、随机变量的数字特征、概率不等式、大数定律和中心极限定理、随机过程初步等。其中，概率不等式在机器学习的理论分析，通常也称为计算学习理论，如 PAC 可学习性以及算法的泛化界和收敛性分析等方面具有重要的应用。此外，大数据定律将被推广用于统计学习理论中经验风险最小化准则的建立，而随机过程则在深度强化学习中具有广泛的应用。

第 8 章介绍香农熵、信息熵、KL 散度和微分熵等信息论基本概念和性质，并引入基于熵概

念的信息度量准则和数据科学建模原理。信息论与机器学习有着紧密的联系，学习某种意义上就是一个熵减的过程，学习的过程也就是使信息的不确定度下降的过程，因此这些内容可以用于创造和改进学习算法（主要是分类问题），甚至衍生出了一个新方向——信息理论学习。特别可用于数据科学中基于概率和熵的相似性度量，这与第 3 章中非概率的相似性度量形成对应。

第 9 章介绍概率模型。包括数据建模的概率思想、模型的参数估计和非参数估计、概率模型的图语言描述和统计决策理论。其中数据建模的概率思想将引出数据科学和机器学习中模型的概率表示和类型等；模型的参数估计和非参数估计重点介绍极大似然、极大后验、直方图估计、核密度估计和非参回顾估计等；概率模型的图语言描述将给出条件独立性、有向非循环图、无向图、团和势等，这为以后学习朴素贝叶斯、隐马尔科夫等概率图模型内容奠定基础；统计决策理论主要涉及模型参数估计的好坏判断，这与机器学习中建立模型的策略密切相关。这一章介绍的概率模型是数据科学中另一大类型的模型之一，与第 6 章中的非概率模型，也即函数模型形成对应。

第四部分：数据的数值优化——凸优化。也即第 10 章至第 12 章的内容。

第 10 章介绍优化的基础理论。包括优化问题的分类，凸集和凸函数的定义和判别方法以及保凸运算，引入凸优化问题的定义和标准形式，并介绍数据科学和机器学习中常见的典型优化问题。事实上，机器学习中通过经验风险最小化准则建立的很多问题都可以建模为凸优化问题。

第 11 章介绍拉格朗日对偶函数和拉格朗日对偶问题，把标准形式（可能是非凸）的优化问题转化为对偶问题进行求解；并引出优化问题的最优性条件；介绍数据科学中各种常见的优化问题的对偶性问题。数据科学和机器学习中的很多问题是非凸的，比如最大割问题，可以通过转换为对偶问题进行有效求解。

第 12 章介绍无约束优化问题的性质和求解方法，包括直线搜索、梯度下降、最速下降、随机梯度下降方法等零阶和一阶方法；约束优化问题的求解方法，包括可行函数法和罚函数法；凸优化问题求解的高阶算法，包括牛顿法、内点法和拟牛顿法等二阶方法以及深度学习中一些常见的优化技术。这些方法将用于数据科学与机器学习中各种优化问题的求解。

读者范围

本书主要面向“数据科学与大数据技术”、“人工智能”、“计算机科学”等专业的本科生或低年级研究生。对于在工作中需要用到数值线性代数、概率估计和数学优化，或者更一般地说，用到计算数学的科研人员、科学家以及工程师，本书也较为合适。这些人群包括直接从事数据分析、机器学习和人工智能算法的科技工作者，亦包括一些工作在其他科学和工程领域但是需要借助数据科学数学基础的科技工作者，这些领域包括计算科学、经济学、金融、统计学、数据挖掘等。在阅读本书之前，读者只需要掌握现代微积分的基础知识即可。如果读者对一些基本的线性代数和基本的概率论有一定的了解，应能较好地理解本书的所有论证和讨论。当然，我们希望即使没有学过线性代数和概率论的读者也能够理解本书所有的基本思想和内容要点。

使用本书作为教材

我们希望本书能够在不同的课程中作为基本教材或者是参考教材来发挥它的作用，这些课

程包括数据科学的数学基础、人工智能的数学基础、机器学习的数学基础和计算机科学的数学基础（偏应用）等。从 2018 年开始，我们即在华东师范大学数据学院的本科生和低年级研究生的同名课程中使用本书的初稿。我们的经验表明，用 3 个学分，也即 48 学时到 54 学时，可以粗略讲授本书的大部分内容。如果用一个 4 学分的课程时间，也即 64 学时到 72 学时，讲课进度就可以比较从容，也可以增加更多的例子，并且可以更加详尽地讨论有关理论。若能用 5 个学分的课程时间，就可以对奇异值分解、最小二乘问题、特征值的计算、线性规划和二次规划（对于以应用为目的的学生极为重要）这些基本内容进行较广泛的细致讨论，或者加强这些内容对应算法方面的介绍或对学生布置更多的习题训练。本书可以作为线性代数、概率统计、线性优化和非线性优化等基础的参考读物。此外，对于像数学系更关注理论的课程，本书可以作为辅助教材，它提供了一些简单的实际例子。

致谢

本书写作参考了 Glbert Strange 教授的 Linear Algebra and Its Application 和 Linear Algebra and Learning from data,Larry Wasserman 教授的 All of Statistics, Thomaas Cover 教授的 Elements of Information Theory，Giuseppe Calafiorce 教授和 El Ghaoui Laurent 教授的 Optimization Models, Stephen Boyd 教授和 Lieven Vandenberghe 教授的 Convex Optimization 等经典数学教材以及 Vladimir Vapnik 教授的 Statistical learning theory, Hastie Trevor 教授, Tibshirani Robert 教授和 Friedman Jerome 教授的 The Elements of Statistical Learning, Ian Goodfellow, Yoshua Bengio, Aaron Courville 的 Deep Learning 等经典的机器学习教材。

本书是在华东师范大学周傲英副校长和数据科学与工程学院的大力支持下历时两年多完成的，虽然两年的时间不算短，并且主要内容也在华东师范大学数据学院本科生和低年级研究生的同名课程中使用过并取得了不错的讲授和学习效果，但是作为为“数据科学与大数据技术”这样一个崭新的硬专业提供一本适用的教材，这点时间显然是不够的，很多内容还没有得到很好的打磨以适应不同层次水平的学生或相关的科研人员。但我们还是希望能够快速出版以满足日益增长的专业需求和读者们对这一领域持续探索的热情。我们只能期待在使用的过程中不断获得反馈以便快速迭代，从而获得更广泛的使用普遍性。这正如数据科学、人工智能和计算机科学这一领域从业者的行事准则：上线、迭代更新、再迭代，...，直至打磨稳定。我们也计划采用这种方式，所以恳请读者们如果碰到任何书本有关的问题，能及时反馈给我们 djhuang@dase.ecnu.edu.cn，以便我们能够改进，我们将不吝感激。

本书的写作过程中得到了来自由华东师范大学、哈尔滨工业大学、中国人民大学、中山大学、东北大学、西北工业大学、河南大学以及桂林电子科技大学等 15 所高校组成的数据专业协作组以及华东师范大学出版社和高等教育出版社的专家们的反馈和建议，同时也获得了华东师范大学很多同事，我课题组的研究生们以及我课程上的学生们的反馈和建议。篇幅所限，我们无法一一表达我们的感谢，只能在此对大家一并表达诚挚的谢意。

最后要特别感谢我的课题组的研究生们，我的博士生郝珊锋、申弋斌、刘友超和硕士生唐赟喆、赖叶静、张洋、余若男、汤路民、杨康、周雪茗、杨礼孟、王明和李特等同学，他们花

费了很多时间来协助我一起修改、编辑书中的公式、表格和图片等，才使得本书能够快速面世。郝珊锋和唐赟喆也协助我一起制作了与本书配套的同名课程的 MOOC 视频（本课程在融优学堂和超星泛雅 http://mooc1.chaoxing.com/course/208843967.html 上线），感谢他们的努力付出。

限于作者的知识水平，书中难免有不妥和错误之处，恳请读者不吝批评和指正。

黄定江

2020 年 10 月

数学符号

下面简要介绍本书所使用的数学符号。如果你不熟悉数学符号所表示的数学概念，可以参考对应的章节。

数据集

$\mathbb{X}$	输入空间
$\mathbb{Y}$	输出空间
$\boldsymbol{x} \in \mathbb{X}$	输入，实例
$y \in \mathbb{Y}$	输出，标记
$T = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_N, y_N)\}$	训练数据集
N	样本容量
$(\boldsymbol{x}_i, y_i)$	第 i 个训练数据点
$\boldsymbol{x} = (\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, ..., \boldsymbol{x}^{(n)})^T$	输入向量，n 维实数向量
$\boldsymbol{x}_i^{(j)}$	输入向量 $\boldsymbol{x}_i$ 的第 j 分量

向量和矩阵

a	标量 (整数或实数)
$\boldsymbol{a}$	向量
$\boldsymbol{A}$	矩阵
A	张量
$\boldsymbol{I}_n$	n 行 n 列的单位矩阵
$\boldsymbol{I}$	维度蕴含于上下文的单位矩阵
$\boldsymbol{e}^{(i)}$	索引 i 处值为 1 其它值为 0 的标准基向量
$\text{diag}(\boldsymbol{a})$	对角方阵，其中对角元素由 $\boldsymbol{a}$ 给定
a_i	向量 $\boldsymbol{a}$ 的第 i 个元素，其中索引从 1 开始
a_{-i}	除了第 i 个元素,$\boldsymbol{a}$ 的所有元素
$a_{i,j}$	矩阵 $\boldsymbol{A}$ 的 i 行 j 列元素
$\boldsymbol{A}_{i,:}$	矩阵 $\boldsymbol{A}$ 的第 i 行
$\boldsymbol{A}_{:,i}$	矩阵 $\boldsymbol{A}$ 的第 i 列
a_i	随机向量 $\boldsymbol{a}$ 的第 i 个元素
$\mathbb{A}$	集合
$\mathbb{R}$	实数集

$\mathbb{C}$	复数域集
$\mathbb{A}\backslash\mathbb{B}$	差集，即其元素包含于 $\mathbb{A}$ 但不包含于 $\mathbb{B}$
$\mathbb{R}^n$	n 维实向量空间
$\mathbb{C}^n$	n 维复向量空间
$\dim(\mathbb{V})$	空间 $\mathbb{V}$ 的维数
$\boldsymbol{A}^{-1}$	矩阵的逆
$\boldsymbol{A}^\top$	矩阵 $\boldsymbol{A}$ 的转置
$\boldsymbol{A}^\dagger$	$\boldsymbol{A}$ 的 Moore-Penrose 伪逆
$\boldsymbol{A}\odot\boldsymbol{B}$	$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的逐元素乘积 (Hadamard 乘积)
$\lvert\boldsymbol{A}\rvert$	$\boldsymbol{A}$ 的行列式
$\text{rank}(\boldsymbol{A})$	矩阵的秩
A_{ij}	元素 a_{ij} 的代数余子式
$\boldsymbol{A}^*$	$\boldsymbol{A}$ 的伴随矩阵
$\text{Tr}\,(\boldsymbol{A})$	矩阵的迹
λ	矩阵的特征值
	范数
$\lVert\cdot\rVert$	范数
$\lVert\boldsymbol{x}\rVert_2$	向量的 l_2 范数
$\lVert\boldsymbol{x}\rVert_1$	向量的 l_1 范数
$\lVert\boldsymbol{x}\rVert_\infty$	向量的 l_∞ 范数
$\lVert\boldsymbol{X}\rVert_2$	矩阵 $\boldsymbol{X}$ 的谱范数
$\text{sim}_{\cos}(\boldsymbol{x},\boldsymbol{y})$	余弦相似度
$\text{vec}(\boldsymbol{A})$	矩阵的向量化
$\lVert\boldsymbol{A}\rVert_F$	矩阵的 F 范数
$\lVert\boldsymbol{A}\rVert_*$	矩阵的核范数
$\text{Col}\,(\boldsymbol{A})$	$\boldsymbol{A}$ 的列空间
$\text{Row}\,(\boldsymbol{A})$	行空间
$\text{Null}\,(\boldsymbol{A})$	零空间
$\text{Null}\,(\boldsymbol{A}^{\text{T}})$	左零空间
	微分
$\frac{dy}{dx}$	y 关于 x 的导数
$\frac{\partial y}{\partial x}$	y 关于 x 的偏导
$\nabla_{\boldsymbol{x}}y$	y 关于 $\boldsymbol{x}$ 的梯度
$\nabla_{\boldsymbol{X}}y$	y 关于 $\boldsymbol{X}$ 的矩阵导数

$\nabla_{\mathsf{X}} y$	y 关于 X 求导后的张量
$\frac{\partial f}{\partial x}$	f对$\boldsymbol{x} \in \mathbb{R}^m$ 的 Jacobian 矩阵
$\nabla_{\boldsymbol{x}}^2 f(\boldsymbol{x})$或$\boldsymbol{H}_f(\boldsymbol{x})$	f 在点 $\boldsymbol{x}$ 处的 Hessian 矩阵

概率基础

Ω	样本空间
E	随机试验
A	事件
$P(A)$	事件A发生的概率
$P(B\|A)$	事件A发生的情况下，事件B发生的概率
$F_X(x)$	累积分布函数 CDF
$f_X(x)$	概率密度函数
x^+	从右边趋向于x
$E(X)$	随机变量X的期望
$D(X)$	随机变量X的方差
$\mathrm{Cov}(X,Y)$	随机变量 X、Y 的协方差
$N(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$	均值为 μ 协方差为 $\boldsymbol{\Sigma}$，$\boldsymbol{x}$ 的高斯分布
$\Phi_X(t)$	X的矩母函数，t为实数
$\mathbb{E}_{x\sim P}[f(x)]$	$f(x)$关于$P(x)$的期望
$X_n \xrightarrow{P} X$	X_n依概率收敛于X
$X_n \rightsquigarrow X$	X_n依分布收敛于X
$X_n \xrightarrow{qm} X$	X_n均方意义下收敛于X
$\Phi(z)$	标准正态分布的累积分布函数
$R_{\exp}(f)$	期望风险
$R_{\mathrm{emp}}(f)$	经验风险

信息论基础

$I(x_i)$	事件x_i的自信息
$I(x_i;y_i)$	事件x_i和事件y_i的互信息
$H(X)$	随机变量 X 的信息熵
$p(x_i)$	事件x_i的概率分布
$H(\boldsymbol{p})$	熵函数
$D_{KL}(P\|Q)$	P 和 Q 的 KL 散度
$H(P,Q)$	P 和 Q交叉熵
$h(X)$	连续随机变量X的微分熵

概率模型和参数估计

θ	待估参数
Θ	待估参数可能的取值
$L(\theta)$	样本在参数θ下的似然函数
$f(x;\theta)$	由 θ 参数化，关于 x 的函数
$p(\mathcal{D}\|\theta)$	由 θ 参数化，获得给定数据$\mathcal{D}$ 的概率
$l(\theta)$	对数似然函数
$\mathbf{1}_{condition}$	如果条件为真则为 1，否则为 0
$K(u)$	参数为u的核函数
$\Gamma(x)$	伽马函数
$\text{Beta}(a,b)$	Beta 函数
$Pa(x_i)$	随机变量x_i的父节点
$\Phi_C(x_C)$	团的势函数，变量x_C属于集合C
$NLL(\theta)$	模型参数为θ 的极小负 log 似然损失
$\boldsymbol{w}$	权重向量

优化

$\mathbf{aff}C$	集合$\mathbb{C}$的仿射包
$\mathbf{relint}C$	集合$\mathbb{C}$的相对内部
$\mathbf{conv}C$	集合$\mathbb{C}$的凸包
$\mathbf{int}C$	集合$\mathbb{C}$的内部
$\mathbf{cl}C$	集合$\mathbb{C}$的闭包
$\mathbf{bd}C$	集合$\mathbb{C}$的边界：$\mathbf{bd}C = \mathbf{cl}C \backslash \mathbf{int}C$
I_C	集合$\mathbb{C}$的示性函数
S_C	集合$\mathbb{C}$的支撑函数
$x \preceq y$	向量x和y之间的分量不等式
S^n	对称的$n \times n$矩阵
S^n_+, S^n_{++}	对称半正定、正定 $n \times n$矩阵
$\mathbb{R}_+, \mathbb{R}_{++}$	非负、正实数
$\mathbf{epi}f$	函数f的上镜图
$\mathbf{prob}S$	事件S的概率
$\mathbf{dom}f$	函数f的定义域
$\lambda_{max}(\boldsymbol{X}), \lambda_{min}(\boldsymbol{X})$	对称矩阵 X 的最大、最小特征值
$\text{dist}(A,B)$	集合（或点）A和B之间的距离
∇f	函数的导数
f^*	f的共轭函数

目录