大数据分析中的算法

文再文课题组编著

前言

随着科技的发展, 新的实际问题和新的技术不断出现. 从海量数据中提取有效信息, 发现潜在规律, 支撑科学决策, 已成为科学研究与工程实践中的重要课题. 要把这些问题解决好, 通常需要建立合适的数学模型, 设计可实现且高效的算法, 研究模型与算法的理论性质, 并在实际任务中检验其效果. 相关研究在很大程度上拓展了传统数学的应用边界, 持续为应用数学与运筹学提供新的问题来源, 也推动了计算机、数学、概率统计、运筹学等学科的交叉与融合.

本书旨在介绍大数据分析中一些核心的数学方法与算法工具. 全书内容可分为四个部分.

第一部分是数学基础. 第一章回顾线性代数、概率论与优化理论中的基本概念, 包括向量与矩阵运算、概率不等式、凸分析基础以及拉格朗日对偶等内容, 为理解后续各章的模型与算法提供必要的预备知识.

第二部分介绍面向大规模数据的随机算法与稀疏/低秩建模方法. 第二章讨论随机数值代数, 重点介绍矩阵乘法、奇异值分解以及超定最小二乘问题的随机化算法, 说明如何在计算资源受限的情况下处理规模很大的数据. 第三章与第四章分别介绍压缩感知与低秩矩阵优化: 压缩感知关注如何从少量测量数据中恢复稀疏信号; 低秩矩阵优化则围绕矩阵补全、鲁棒主成分分析等问题展开. 两类方法都利用了问题的结构性假设 (稀疏或低秩), 在信号处理与推荐系统等任务中具有代表性.

第三部分选取近年来发展较快的若干专题. 第五章介绍最优输运, 作为比较概率分布差异的一种数学框架, 同时也给出网络流、Sinkhorn 等典型算法及应用例子. 第六章讨论次模优化, 强调次模性所体现的 ‘‘边际效应递减’’ 结构, 并介绍次模最大化与最小化的基本算法思路及实验. 第七章以相位恢复为例, 展示在信息不完整 (仅有幅度测量) 的情形下如何重建原始信号, 并讨论半定松弛与非凸优化等不同处理路线.

第四部分聚焦人工智能领域的核心算法. 第八章介绍深度学习, 从基本神经网络架构出发, 讨论卷积神经网络、Transformer 等主流模型, 并进一步介绍训练中的优化算法与相关理论基础. 第九章讨论强化学习, 系统介绍马尔科夫决策过程、基于价值的方法 (如 Q 学习) 与基于策略的方法 (如策略梯度、Actor-Critic 等), 并结合典型例子说明这些方法在序列决策任务中的作用.

本书可作为数学、统计学、计算机科学、数据科学与大数据技术等专业高年级本科生、研究生及相关领域研究人员的教材或参考书. 希望读者通过本书的学习, 能够了解大数据分析中一些基本问题的数学模型、典型算法及其理论背景, 逐步培养建模、算法设计与问题分析的能力, 为进一步学习与研究打下基础.

诚挚感谢北京大学北京国际数学研究中心和数学科学学院的长期资助和支持. 本书部分内容参考了高教出版社出版的《最优化: 建模、算法与理论》教材, 感谢刘浩洋、户将和李勇锋的付出与努力. 本书内容在北京大学数学科学学院多次开设的 ‘‘大数据分析中的算法’’ 课程中使用, 感谢课题组同学在初稿整理方面, 如李天佑在最优输运, 袁浩然在相位恢复, 丁思哲在次模优化和强化学习, 崔敏在次模优化等等章节的帮助和支持.

限于作者的知识水平, 书中恐有不妥之处, 恳请读者不吝批评和指正.

文再文

北京，2026 年 2 月

目录