

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. 第五章介绍最优运输, 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, 如李天佑在最优运输, 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

第一章 数学基础	1
1.1 线性代数基础	1
1.2 概率基础	12
1.2.1 概率空间	13
1.2.2 随机变量	13
1.2.3 概率不等式	21
1.3 优化基础	25
1.3.1 凸集与凸函数	25
1.3.2 拉格朗日函数与对偶问题	31
1.3.3 带约束优化问题的最优性理论	36
1.4 习题	40
第二章 随机数值代数	43
2.1 简介	43
2.2 矩阵乘法的随机算法	44
2.2.1 随机抽样算法	44
2.2.2 基于列抽样的随机矩阵乘法	46
2.2.3 误差分析	47
2.3 特征值分解与奇异值分解	51
2.3.1 部分特征值问题的变分情形	51
2.3.2 瑞利-里茨过程	53
2.3.3 奇异值分解	54
2.3.4 特征值分解与奇异值分解的关系	56
2.4 矩阵奇异值分解随机算法	58

2.4.1	两阶段计算框架	59
2.4.2	基于列采样的奇异值分解随机算法	60
2.4.3	基于幂法的奇异值分解随机算法	65
2.4.4	应用举例	70
2.5	超定最小二乘问题随机算法	72
2.5.1	基于行采样的随机算法	73
2.5.2	基于哈达玛矩阵采样的随机算法	75
2.5.3	应用举例	76
2.6	习题	77
第三章	压缩感知	81
3.1	简介	81
3.1.1	背景介绍	81
3.1.2	压缩感知工业界应用	84
3.1.3	稀疏优化需要解决的基本问题	85
3.2	稀疏性相关条件	86
3.2.1	Spark 及其性质	86
3.2.2	零空间定义及其相关性质	89
3.3	RIP 条件下 l_1 题与 l_0 题的等价性	92
3.3.1	RIP 定义及其相关性质	93
3.3.2	l_1 化问题的最优性条件	95
3.3.3	l_1 化问题解的唯一性	97
3.4	鲁棒压缩感知	103
3.4.1	无噪声情形	103
3.4.2	噪声情形	104
3.5	稀疏优化算法	104
3.5.1	贪婪算法	104
3.5.2	近似点梯度算法	106
3.5.3	加速近似点梯度算法	108
3.5.4	交替乘子方向算法	109
3.6	应用举例	114
3.6.1	压缩感知理论验证	114
3.6.2	压缩感知	115
3.7	习题	117

第四章 低秩矩阵优化	121
4.1 简介	121
4.2 协同过滤模型	123
4.3 矩阵分解模型	124
4.3.1 模型背景	125
4.3.2 算法介绍	127
4.4 核范数优化模型	129
4.4.1 基于核范数的低秩最小化模型	129
4.4.2 与压缩感知的关系	130
4.4.3 矩阵稀疏性相关条件	131
4.4.4 核范数的性质	133
4.4.5 带约束情形下的核范数最小化问题	136
4.4.6 核范数最小化问题的二次惩罚框架	140
4.5 其他矩阵优化问题	142
4.5.1 鲁棒主成分分析	142
4.5.2 低秩因子分解	143
4.5.3 非负矩阵补全	144
4.5.4 稀疏协方差矩阵估计	144
4.6 应用举例	146
4.6.1 低秩矩阵补全	146
4.6.2 鲁棒主成分分析	147
4.7 习题	149
第五章 最优运输	151
5.1 简介	151
5.1.1 距离概念的延拓	151
5.1.2 最优运输的发展	152
5.2 最优运输问题	154
5.2.1 离散情形下的最优运输问题	154
5.2.2 一般情形下的最优运输问题 *	158
5.2.3 最优运输距离	161
5.3 算法简介	166
5.3.1 网络流算法	167
5.3.2 Sinkhorn 算法	175

5.3.3	沃瑟斯坦重心算法	183
5.4	应用举例	187
5.4.1	一维最优运输问题	187
5.4.2	图像色彩自适应	188
5.4.3	形状插值	189
5.5	习题	191
第六章	次模优化	193
6.1	简介	193
6.2	次模函数和次模性	194
6.2.1	常用符号与定义	194
6.2.2	次模函数的定义	196
6.2.3	次模函数的基本性质	202
6.3	次模最大化问题	206
6.3.1	经典贪心算法	207
6.3.2	延迟贪心算法	209
6.4	次模最小化问题	211
6.4.1	连续扩展和洛瓦兹扩展	211
6.4.2	次模多面体和基多面体	215
6.4.3	次模最小化问题的等价形式	216
6.4.4	投影次梯度算法	218
6.5	数值实验	222
6.5.1	最优选址	222
6.5.2	图的最大流问题	225
6.5.3	小结	228
6.6	习题	229
第七章	相位恢复	231
7.1	简介	231
7.1.1	应用举例：图像的相位信息	232
7.1.2	应用举例：X 射线晶体学	233
7.1.3	相位恢复的发展	234
7.2	经典相位恢复模型	235
7.2.1	投影算法	236

7.2.2	ADMM 算法与投影算法的关系	239
7.3	半定松弛	241
7.3.1	相位提升	242
7.3.2	相位切割	246
7.4	非凸优化模型	248
7.4.1	目标函数的非凸性	248
7.4.2	复变函数的可微性与 Wirtinger 导数	248
7.4.3	Wirtinger 梯度流算法	251
7.5	应用实例	254
7.5.1	超短激光脉冲测量	254
7.5.2	相干衍射成像	255
7.6	习题	256
第八章	深度学习	259
8.1	背景介绍	259
8.1.1	简介	259
8.1.2	深度学习的历史以及意义	260
8.1.3	常见数据集	260
8.2	神经网络构架	265
8.2.1	前馈神经网络	265
8.2.2	卷积神经网络	268
8.2.3	循环神经网络	276
8.2.4	Transformer 模型	279
8.2.5	图神经网络	284
8.2.6	AI for science	290
8.2.7	总结	291
8.3	深度学习模型	291
8.3.1	损失函数	292
8.3.2	优化模型	293
8.3.3	反向传播	294
8.3.4	初始化	296
8.3.5	梯度消失与梯度爆炸	297
8.4	深度学习优化算法	299
8.4.1	随机梯度下降算法	300

8.4.2	自然梯度法	304
*8.5	机器学习理论基础	313
8.5.1	误差分解	313
8.5.2	泛化界	316
8.5.3	VC 维数	317
8.5.4	拉德马赫复杂度	319
8.6	应用举例	327
8.6.1	步长调整策略	327
8.6.2	训练与测试	328
8.6.3	数值实验	329
8.7	习题	330
第九章	强化学习	333
9.1	简介	333
9.2	动态规划简介	335
9.2.1	最短路径问题的动态规划	335
9.2.2	随机性最短路径问题的动态规划算法	336
9.3	马尔科夫决策过程	338
9.3.1	马尔科夫决策过程的基本要素	338
9.3.2	价值函数与增益	343
9.3.3	贝尔曼方程	347
9.3.4	价值迭代与策略迭代	354
9.4	基于价值的强化学习	362
9.4.1	马尔科夫决策过程中的采样	362
9.4.2	无模型控制问题的建模与随机优化算法	363
9.4.3	时序差分算法	367
9.4.4	Q 学习与深度 Q 学习	372
9.5	基于策略的强化学习	376
9.5.1	策略梯度定理	377
9.5.2	原始策略梯度算法	381
9.5.3	方差缩减方法	383
9.5.4	Actor-Critic 算法	386
9.5.5	信赖域策略优化算法	387
9.5.6	近似点策略优化算法	393

9.6 应用举例	396
9.6.1 悬崖漫步	397
9.6.2 推车立杆	398
9.7 习题	402
附录 A 符号表	405
参考文献	409
索引	429