

目录

致谢	xvi
网站	xxii
数学符号	xxiii
第一章 引言	1
1.1 本书面向的读者	10
1.2 深度学习的历史趋势	11
1.2.1 神经网络的众多名称和命运变迁	12
1.2.2 与日俱增的数据量	17
1.2.3 与日俱增的模型规模	19
1.2.4 与日俱增的精度、复杂度和对现实世界的冲击	22
第一部分 应用数学与机器学习基础	25
第二章 线性代数	27
2.1 标量、向量、矩阵和张量	27
2.2 矩阵和向量相乘	29
2.3 单位矩阵和逆矩阵	31
2.4 线性相关和生成子空间	32
2.5 范数	34
2.6 特殊类型的矩阵和向量	36
2.7 特征分解	37

2.8	奇异值分解	39
2.9	Moore-Penrose 伪逆	40
2.10	迹运算	41
2.11	行列式	42
2.12	实例：主成分分析	42
第三章	概率与信息论	47
3.1	为什么要使用概率?	47
3.2	随机变量	49
3.3	概率分布	50
3.3.1	离散型变量和概率质量函数	50
3.3.2	连续型变量和概率密度函数	51
3.4	边缘概率	52
3.5	条件概率	52
3.6	条件概率的链式法则	53
3.7	独立性和条件独立性	53
3.8	期望、方差和协方差	54
3.9	常用概率分布	55
3.9.1	Bernoulli 分布	56
3.9.2	Multinoulli 分布	56
3.9.3	高斯分布	57
3.9.4	指数分布和 Laplace 分布	58
3.9.5	Dirac 分布和经验分布	59
3.9.6	分布的混合	59
3.10	常用函数的有用性质	61
3.11	贝叶斯规则	63
3.12	连续型变量的技术细节	64
3.13	信息论	65
3.14	结构化概率模型	69
第四章	数值计算	72
4.1	上溢和下溢	72
4.2	病态条件	73

4.3	基于梯度的优化方法	74
4.3.1	梯度之上: Jacobian 和 Hessian 矩阵	77
4.4	约束优化	82
4.5	实例: 线性最小二乘	85
第五章	机器学习基础	87
5.1	学习算法	87
5.1.1	任务 T	88
5.1.2	性能度量 P	91
5.1.3	经验 E	92
5.1.4	示例: 线性回归	94
5.2	容量、过拟合和欠拟合	97
5.2.1	没有免费午餐定理	102
5.2.2	正则化	104
5.3	超参数和验证集	105
5.3.1	交叉验证	106
5.4	估计、偏差和方差	108
5.4.1	点估计	108
5.4.2	偏差	109
5.4.3	方差和标准差	111
5.4.4	权衡偏差和方差以最小化均方误差	113
5.4.5	一致性	114
5.5	最大似然估计	115
5.5.1	条件对数似然和均方误差	116
5.5.2	最大似然的性质	117
5.6	贝叶斯统计	118
5.6.1	最大后验 (MAP) 估计	121
5.7	监督学习算法	122
5.7.1	概率监督学习	122
5.7.2	支持向量机	123
5.7.3	其他简单的监督学习算法	125
5.8	无监督学习算法	128
5.8.1	主成分分析	128

5.8.2	k -均值聚类	131
5.9	随机梯度下降	132
5.10	构建机器学习算法	133
5.11	促使深度学习发展的挑战	134
5.11.1	维数灾难	135
5.11.2	局部不变性和平滑正则化	135
5.11.3	流形学习	139
第二部分 深度网络：现代实践		143
第六章	深度前馈网络	145
6.1	实例：学习 XOR	148
6.2	基于梯度的学习	152
6.2.1	代价函数	153
6.2.1.1	使用最大似然学习条件分布	154
6.2.1.2	学习条件统计量	155
6.2.2	输出单元	156
6.2.2.1	用于高斯输出分布的线性单元	156
6.2.2.2	用于 Bernoulli 输出分布的 sigmoid 单元	157
6.2.2.3	用于 Multinoulli 输出分布的 softmax 单元	159
6.2.2.4	其他的输出类型	162
6.3	隐藏单元	165
6.3.1	整流线性单元及其扩展	166
6.3.2	logistic sigmoid 与双曲正切函数	168
6.3.3	其他隐藏单元	169
6.4	架构设计	170
6.4.1	万能近似性质和深度	171
6.4.2	其他架构上的考虑	174
6.5	反向传播和其他的微分算法	175
6.5.1	计算图	176
6.5.2	微积分中的链式法则	178
6.5.3	递归地使用链式法则来实现反向传播	179

6.5.4	全连接 MLP 中的反向传播计算	181
6.5.5	符号到符号的导数	182
6.5.6	一般化的反向传播	185
6.5.7	实例：用于 MLP 训练的反向传播	188
6.5.8	复杂化	190
6.5.9	深度学习界以外的微分	191
6.5.10	高阶微分	193
6.6	历史小记	193
第七章	深度学习中的正则化	197
7.1	参数范数惩罚	198
7.1.1	L^2 参数正则化	199
7.1.2	L^1 参数正则化	202
7.2	作为约束的范数惩罚	204
7.3	正则化和欠约束问题	206
7.4	数据集增强	207
7.5	噪声鲁棒性	208
7.5.1	向输出目标注入噪声	209
7.6	半监督学习	209
7.7	多任务学习	210
7.8	提前终止	211
7.9	参数绑定和参数共享	217
7.9.1	卷积神经网络	218
7.10	稀疏表示	218
7.11	Bagging 和其他集成方法	220
7.12	Dropout	222
7.13	对抗训练	230
7.14	切面距离、正切传播和流形正切分类器	232
第八章	深度模型中的优化	235
8.1	学习和纯优化有什么不同	235
8.1.1	经验风险最小化	236
8.1.2	代理损失函数和提前终止	237

8.1.3	批量算法和小批量算法	237
8.2	神经网络优化中的挑战	241
8.2.1	病态	242
8.2.2	局部极小值	243
8.2.3	高原、鞍点和其他平坦区域	244
8.2.4	悬崖和梯度爆炸	246
8.2.5	长期依赖	247
8.2.6	非精确梯度	248
8.2.7	局部和全局结构间的弱对应	248
8.2.8	优化的理论限制	250
8.3	基本算法	251
8.3.1	随机梯度下降	251
8.3.2	动量	253
8.3.3	Nesterov 动量	256
8.4	参数初始化策略	256
8.5	自适应学习率算法	261
8.5.1	AdaGrad	261
8.5.2	RMSProp	262
8.5.3	Adam	262
8.5.4	选择正确的优化算法	263
8.6	二阶近似方法	265
8.6.1	牛顿法	266
8.6.2	共轭梯度	267
8.6.3	BFGS	270
8.7	优化策略和元算法	271
8.7.1	批标准化	271
8.7.2	坐标下降	274
8.7.3	Polyak 平均	274
8.7.4	监督预训练	275
8.7.5	设计有助于优化的模型	277
8.7.6	延拓法和课程学习	278

第九章	卷积网络	281
9.1	卷积运算	282
9.2	动机	285
9.3	池化	290
9.4	卷积与池化作为一种无限强的先验	295
9.5	基本卷积函数的变体	296
9.6	结构化输出	306
9.7	数据类型	307
9.8	高效的卷积算法	309
9.9	随机或无监督的特征	310
9.10	卷积网络的神经科学基础	311
9.11	卷积网络与深度学习的历史	317
第十章	序列建模：循环和递归网络	319
10.1	展开计算图	320
10.2	循环神经网络	323
10.2.1	导师驱动过程和输出循环网络	326
10.2.2	计算循环神经网络的梯度	328
10.2.3	作为有向图模型的循环网络	330
10.2.4	基于上下文的 RNN 序列建模	334
10.3	双向 RNN	336
10.4	基于编码-解码的序列到序列架构	338
10.5	深度循环网络	340
10.6	递归神经网络	341
10.7	长期依赖的挑战	343
10.8	回声状态网络	345
10.9	渗漏单元和其他多时间尺度的策略	347
10.9.1	时间维度的跳跃连接	347
10.9.2	渗漏单元和一系列不同时间尺度	348
10.9.3	删除连接	348
10.10	长短期记忆和其他门控 RNN	349
10.10.1	LSTM	349
10.10.2	其他门控 RNN	351

10.11	优化长期依赖	352
10.11.1	截断梯度	353
10.11.2	引导信息流的正则化	355
10.12	外显记忆	355
第十一章	实践方法论	359
11.1	性能度量	360
11.2	默认的基准模型	362
11.3	决定是否收集更多数据	363
11.4	选择超参数	364
11.4.1	手动调整超参数	364
11.4.2	自动超参数优化算法	367
11.4.3	网格搜索	368
11.4.4	随机搜索	369
11.4.5	基于模型的超参数优化	370
11.5	调试策略	371
11.6	示例：多位数字识别	374
第十二章	应用	377
12.1	大规模深度学习	377
12.1.1	快速的 CPU 实现	378
12.1.2	GPU 实现	378
12.1.3	大规模的分布式实现	380
12.1.4	模型压缩	381
12.1.5	动态结构	382
12.1.6	深度网络的专用硬件实现	384
12.2	计算机视觉	385
12.2.1	预处理	385
12.2.1.1	对比度归一化	386
12.2.2	数据集增强	389
12.3	语音识别	390
12.4	自然语言处理	392
12.4.1	n -gram	392

12.4.2	神经语言模型	394
12.4.3	高维输出	396
12.4.3.1	使用短列表	396
12.4.3.2	分层 Softmax	397
12.4.3.3	重要采样	399
12.4.3.4	噪声对比估计和排名损失	401
12.4.4	结合 n -gram 和神经语言模型	401
12.4.5	神经机器翻译	402
12.4.5.1	使用注意力机制并对齐数据片段	403
12.4.6	历史展望	406
12.5	其他应用	407
12.5.1	推荐系统	407
12.5.1.1	探索与利用	409
12.5.2	知识表示、推理和回答	410
12.5.2.1	知识、联系和回答	410
第三部分	深度学习研究	414
第十三章	线性因子模型	417
13.1	概率 PCA 和因子分析	418
13.2	独立成分分析	419
13.3	慢特征分析	421
13.4	稀疏编码	423
13.5	PCA 的流形解释	426
第十四章	自编码器	429
14.1	欠完备自编码器	430
14.2	正则自编码器	431
14.2.1	稀疏自编码器	431
14.2.2	去噪自编码器	433
14.2.3	惩罚导数作为正则	434
14.3	表示能力、层的大小和深度	434
14.4	随机编码器和解码器	435

14.5	去噪自编码器	436
14.5.1	得分估计	437
14.5.2	历史展望	440
14.6	使用自编码器学习流形	440
14.7	收缩自编码器	445
14.8	预测稀疏分解	447
14.9	自编码器的应用	448
第十五章	表示学习	449
15.1	贪心逐层无监督预训练	450
15.1.1	何时以及为何无监督预训练有效?	452
15.2	迁移学习和领域自适应	457
15.3	半监督解释因果关系	461
15.4	分布式表示	466
15.5	得益于深度的指数增益	471
15.6	提供发现潜在原因的线索	472
第十六章	深度学习中的结构化概率模型	475
16.1	非结构化建模的挑战	476
16.2	使用图描述模型结构	479
16.2.1	有向模型	480
16.2.2	无向模型	482
16.2.3	配分函数	484
16.2.4	基于能量的模型	485
16.2.5	分离和 d-分离	487
16.2.6	在有向模型和无向模型中转换	490
16.2.7	因子图	493
16.3	从图模型中采样	494
16.4	结构化建模的优势	495
16.5	学习依赖关系	496
16.6	推断和近似推断	497
16.7	结构化概率模型的深度学习方法	498
16.7.1	实例:受限玻尔兹曼机	499

第十七章 蒙特卡罗方法	502
17.1 采样和蒙特卡罗方法	502
17.1.1 为什么需要采样?	502
17.1.2 蒙特卡罗采样的基础	503
17.2 重要采样	504
17.3 马尔可夫链蒙特卡罗方法	506
17.4 Gibbs 采样	510
17.5 不同的峰值之间的混合挑战	511
17.5.1 不同峰值之间通过回火来混合	513
17.5.2 深度也许会有助于混合	514
第十八章 直面配分函数	516
18.1 对数似然梯度	516
18.2 随机最大似然和对比散度	518
18.3 伪似然	524
18.4 得分匹配和比率匹配	526
18.5 去噪得分匹配	528
18.6 噪声对比估计	529
18.7 估计配分函数	531
18.7.1 退火重要采样	533
18.7.2 桥式采样	536
第十九章 近似推断	538
19.1 把推断视作优化问题	539
19.2 期望最大化	541
19.3 最大后验推断和稀疏编码	542
19.4 变分推断和变分学习	544
19.4.1 离散型潜变量	545
19.4.2 变分法	551
19.4.3 连续型潜变量	554
19.4.4 学习和推断之间的相互作用	556
19.5 学成近似推断	556
19.5.1 醒眠算法	557

19.5.2	学成推断的其他形式	557
第二十章	深度生成模型	559
20.1	玻尔兹曼机	559
20.2	受限玻尔兹曼机	561
20.2.1	条件分布	562
20.2.2	训练受限玻尔兹曼机	563
20.3	深度信念网络	564
20.4	深度玻尔兹曼机	566
20.4.1	有趣的性质	568
20.4.2	DBM 均匀场推断	569
20.4.3	DBM 的参数学习	571
20.4.4	逐层预训练	572
20.4.5	联合训练深度玻尔兹曼机	574
20.5	实值数据上的玻尔兹曼机	578
20.5.1	Gaussian-Bernoulli RBM	578
20.5.2	条件协方差的无向模型	579
20.6	卷积玻尔兹曼机	583
20.7	用于结构化或序列输出的玻尔兹曼机	585
20.8	其他玻尔兹曼机	586
20.9	通过随机操作的反向传播	587
20.9.1	通过离散随机操作的反向传播	588
20.10	有向生成网络	591
20.10.1	sigmoid 信念网络	591
20.10.2	可微生成器网络	592
20.10.3	变分自编码器	594
20.10.4	生成式对抗网络	597
20.10.5	生成矩匹配网络	600
20.10.6	卷积生成网络	601
20.10.7	自回归网络	602
20.10.8	线性自回归网络	602
20.10.9	神经自回归网络	603
20.10.10	NADE	604