

# 深度学习

2017年9月4日



# 目录

致谢	xvi
网站	xxii
数学符号	xxiii
<b>第一章 引言</b>	<b>1</b>
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
<b>第一部分 应用数学与机器学习基础</b>	<b>25</b>
<b>第二章 线性代数</b>	<b>27</b>
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
<b>第三章</b>	<b>概率与信息论</b>	<b>47</b>
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
<b>第四章</b>	<b>数值计算</b>	<b>72</b>
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
<b>第五章</b>	<b>机器学习基础</b>	<b>87</b>
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
<b>第二部分 深度网络：现代实践</b>		<b>143</b>
<b>第六章 深度前馈网络</b>		<b>145</b>
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
<b>第七章</b>	<b>深度学习中的正则化</b>	<b>197</b>
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
<b>第八章</b>	<b>深度模型中的优化</b>	<b>235</b>
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

<b>第九章</b>	<b>卷积网络</b>	<b>281</b>
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
<b>第十章</b>	<b>序列建模：循环和递归网络</b>	<b>319</b>
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
<b>第十一章</b>	<b>实践方法论</b>	<b>359</b>
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
<b>第十二章</b>	<b>应用</b>	<b>377</b>
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
<b>第三部分</b>	<b>深度学习研究</b>	<b>414</b>
<b>第十三章</b>	<b>线性因子模型</b>	<b>417</b>
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
<b>第十四章</b>	<b>自编码器</b>	<b>429</b>
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
<b>第十五章</b>	<b>表示学习</b>	<b>449</b>
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
<b>第十六章</b>	<b>深度学习中的结构化概率模型</b>	<b>475</b>
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

<b>第十七章 蒙特卡罗方法</b>	<b>502</b>
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
<b>第十八章 直面配分函数</b>	<b>516</b>
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
<b>第十九章 近似推断</b>	<b>538</b>
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
<b>第二十章</b>	<b>深度生成模型</b>	<b>559</b>
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

20.11 从自编码器采样 . . . . .	606
20.11.1 与任意去噪自编码器相关的马尔可夫链 . . . . .	607
20.11.2 夹合与条件采样 . . . . .	607
20.11.3 回退训练过程 . . . . .	608
20.12 生成随机网络 . . . . .	609
20.12.1 判别性 GSN . . . . .	610
20.13 其他生成方案 . . . . .	610
20.14 评估生成模型 . . . . .	611
20.15 结论 . . . . .	613
<b>参考文献</b>	<b>615</b>
<b>术语</b>	<b>679</b>

# 中文版致谢

首先，我们要感谢原作者在本书翻译时给予我们的大力帮助。特别是，原作者和我们分享了书中的原图和参考文献库，这极大节省了我们的时间和精力。

本书涉及的内容博大且思想深刻，如果没有众多同学和网友的帮助，我们不可能顺利完成翻译。

我们才疏学浅而受此重任，深知自身水平难以将本书翻译得很准确。因此我们完成草稿后，将书稿公开于 Github，及早接受网友的批评和建议。以下网友为本书的翻译草稿提供了很多及时的反馈和宝贵的修改意见：@tttwwy @tankeco @fairmiracle @GageGao @huangpingchun @MaHongP @acgtyrant @yanhuibin315 @Buttonwood @titicacafz @weijy026a @RuiZhang1993 @zymiboxpay @xingkongliang @oisc @tielei @yuduowu @Qingmu @HC-2016 @xiaomingabc @bengordai @Bojian @JoyFYan @minoriwww @khty2000 @gump88 @zdx3578 @PassStory @imwebson @wlbksy @roachsina @Elvinczp @endymecy @9578577 @linzhp @cncscottzheng @germany-zhu @zhangyafeikimi @showgood163 @kangqf @NeutronT @badpoem @kkpoker @Seaball @wheaio @angrymidiao @ZhiweiYang @corenel @zhaoyu611 @SiriusXDJ @dfcv24 @EmisXXY @FlyingFire @vsooda @friskit-china @poerin @ninesunqian @JiaqiYao @Sofring @wenlei @wizyoung @imageslr @indam @XuLYC @zhouqingping @freedomRen @runPenguin @piantou

在此期间，我们四位译者再次进行了校对并且相互之间也校对了一遍。然而仅仅通过我们的校对，实在难以发现翻译中存在的问题。因此，我们邀请一些同学和网友帮助我们校对。经过他们的校对，本书的翻译质量得到了极大的提升。在此我们一一列出，以表示我们由衷的感谢！

- 第一章（引言）：刘畅、许丁杰、潘雨粟和 NeutronT 对本章进行了阅读，并对

很多语句提出了不少修改建议。林中鹏进行了校对，他提出了很多独到的修改建议。

- 第二章（线性代数）：许丁杰和骆徐圣阅读本章，并修改语句。李若愚进行了校对，提出了很多细心的建议。
- 第三章（概率与信息论）：许丁杰阅读本章，并修改语句。李培炎和何翊卓进行了校对，并修改了很多中文用词，使翻译更加准确。
- 第四章（数值计算）：张亚霏阅读本章，并对其他章节也有提出了一些修改建议。张源源进行了校对，并指出了原文可能存在的问题，非常仔细。
- 第五章（机器学习基础）：郭浩和黄平春阅读本章，并修改语句。李东和林中鹏进行了校对。本章篇幅较长，能够有现在的翻译质量离不开这四位的贡献。
- 第六章（深度前馈网络）：周卫林、林中鹏和张远航阅读本章，并提出修改意见。
- 第七章（深度学习中的正则化）：周柏村进行了非常细心的校对，指出了大量问题，令翻译更加准确。
- 第八章（深度模型中的优化）：房晓宇和吴翔阅读本章。黄平春进行了校对，他提出的很多建议让行文更加流畅易懂。
- 第九章（卷积网络）：赵雨和潘雨粟阅读本章，并润色语句。丁志铭进行了非常仔细的校对，并指出很多翻译问题。
- 第十章（序列建模：循环和递归网络）：刘畅阅读本章。赵雨提供了详细的校对建议，尹瑞清根据他的翻译版本，给我们的版本提出了很多建议。虽然仍存在一些分歧，但我们两个版本的整合，让翻译质量提升很多。
- 第十二章（应用）：潘雨粟进行了校对，在他的校对之前，本章阅读起来比较困难。他提供的修改建议，不仅提高了行文流畅度，还提升了译文的准确度。
- 第十三章（线性因子模型）：贺天行阅读本章，修改语句。杨志伟校对本章，润色大量语句。
- 第十四章（自编码器）：李雨慧和黄平春进行了校对。李雨慧提升了语言的流畅度，黄平春纠正了不少错误，提高了准确性。

- 第十五章（表示学习）：cnscozzheng 阅读本章，并修改语句。
- 第十七章（蒙特卡罗方法）：张远航提供了非常细致的校对，后续还校对了一遍，使译文质量大大提升。
- 第十八章（直面配分函数）：吴家楠进行了校对，提升了译文准确性和可读性。
- 第十九章（近似推断）：黄浩军、张远航和张源源进行了校对。这章虽篇幅不大，但内容有深度，译文在三位的帮助下提高了准确度。

所有校对的修改建议都保存在 Github 上，再次感谢以上同学和网友的付出。经过这五个多月的修改，草稿慢慢变成了初稿。尽管还有很多问题，但大部分内容是可读的，并且是准确的。当然目前的翻译仍存在一些没有及时发现的问题，因此翻译也将持续更新，不断修改。我们非常希望读者能到 Github 提建议，并且非常欢迎，无论多么小的修改建议，都是非常宝贵的。

此外，我们还要感谢魏太云学长，他帮助我们与出版社沟通交流，并给予了我们很多排版上的指导。

最后，感谢我们的导师张志华教授，没有老师的支持，我们难以完成翻译。

# 原书致谢

如果没有他人的贡献，这本书将不可能完成。我们感谢为本书提出建议和帮助组织内容结构的人：Guillaume Alain, Kyunghyun Cho, Çağlar Gülçehre, David Krueger, Hugo Larochelle, Razvan Pascanu and Thomas Rohée。

我们感谢为本书内容提供反馈的人。其中一些人对许多章都给出了建议：Martín Abadi, Guillaume Alain, Ion Androutsopoulos, Fred Bertsch, Olexa Bilaniuk, Ufuk Can Biçici, Matko Bošnjak, John Boersma, Greg Brockman, Alexandre de Brébisson, Pierre Luc Carrier, Sarath Chandar, Pawel Chilinski, Mark Daoust, Oleg Dashevskii, Laurent Dinh, Stephan Dreseitl, Jim Fan, Miao Fan, Meire Fortunato, Frédéric Francis, Nando de Freitas, Çağlar Gülçehre, Jurgen Van Gael, Javier Alonso García, Jonathan Hunt, Gopi Jeyaram, Chingiz Kabytayev, Lukasz Kaiser, Varun Kanade, Asifullah Khan, Akiel Khan, John King, Diederik P. Kingma, Yann LeCun, Rudolf Mathey, Matías Mattamala, Abhinav Maurya, Kevin Murphy, Oleg Mürk, Roman Novak, Augustus Q. Odena, Simon Pavlik, Karl Pichotta, Eddie Pierce, Kari Pulli, Roussel Rahman, Tapani Raiko, Anurag Ranjan, Johannes Roith, Mihaela Rosca, Halis Sak, César Salgado, Grigory Sapunov, Yoshinori Sasaki, Mike Schuster, Julian Serban, Nir Shabat, Ken Shirriff, Andre Simpelo, Scott Stanley, David Sussillo, Ilya Sutskever, Carles Gelada Sáez, Graham Taylor, Valentin Tolmer, Massimiliano Tomassoli, An Tran, Shubhendu Trivedi, Alexey Umnov, Vincent Vanhoucke, Marco Visentini-Scarzanella, Martin Vita, David Warde-Farley, Dustin Webb, Kelvin Xu, Wei Xue, Ke Yang, Li Yao, Zygmunt Zajac and Ozan Çağlayan.

我们也要感谢对单个章节提供有效反馈的人：

- 数学符号：Zhang Yuanhang.

- 第一章 (引言): Yusuf Akgul, Sebastien Bratieres, Samira Ebrahimi, Charlie Gorichanaz, Brendan Loudermilk, Eric Morris, Cosmin Pârvulescu and Alfredo Solano.
- 第二章 (线性代数): Amjad Almahairi, Nikola Banić, Kevin Bennett, Philippe Castonguay, Oscar Chang, Eric Fosler-Lussier, Andrey Khalyavin, Sergey Oreshkov, István Petrás, Dennis Prangle, Thomas Rohée, Gitanjali Gulve Sehgal, Colby Toland, Alessandro Vitale and Bob Welland.
- 第三章 (概率与信息论): John Philip Anderson, Kai Arulkumaran, Vincent Dumoulin, Rui Fa, Stephan Gouws, Artem Oboturov, Antti Rasmus, Alexey Surkov and Volker Tresp.
- 第四章 (数值计算): Tran Lam An Ian Fischer and Hu Yuhuang.
- 第五章 (机器学习基础): Dzmitry Bahdanau, Justin Domingue, Nikhil Garg, Makoto Otsuka, Bob Pepin, Philip Popien, Emmanuel Rayner, Peter Shepard, Kee-Bong Song, Zheng Sun and Andy Wu.
- 第六章 (深度前馈网络): Uriel Berdugo, Fabrizio Bottarel, Elizabeth Burl, Ishan Durugkar, Jeff Hlywa, Jong Wook Kim, David Krueger and Aditya Kumar Praharaaj.
- 第七章 (深度学习中的正则化): Morten Kolbæk, Kshitij Lauria, Inkyu Lee, Sunil Mohan, Hai Phong Phan and Joshua Salisbury.
- 第八章 (深度模型中的优化): Marcel Ackermann, Peter Armitage, Rowel Atienza, Andrew Brock, Tegan Maharaj, James Martens, Kashif Rasul, Klaus Strobl and Nicholas Turner.
- 第九章 (卷积网络): Martín Arjovsky, Eugene Brevdo, Konstantin Divilov, Eric Jensen, Mehdi Mirza, Alex Paino, Marjorie Sayer, Ryan Stout and Wentao Wu.
- 第十章 (序列建模: 循环和递归网络): Gökçen Eraslan, Steven Hickson, Razvan Pascanu, Lorenzo von Ritter, Rui Rodrigues, Dmitriy Serdyuk, Dongyu Shi and Kaiyu Yang.
- 第十一章 (实践方法论): Daniel Beckstein.

- 第十二章 (应用): George Dahl, Vladimir Nekrasov and Ribana Roscher.
- 第十三章 (线性因子模型): Jayanth Koushik.
- 第十五章 (表示学习): Kunal Ghosh.
- 第十六章 (深度学习中的结构化概率模型): Minh Lê and Anton Varfolom.
- 第十八章 (直面配分函数): Sam Bowman.
- 第十九章 (近似推断): Yujia Bao.
- 第二十章 (深度生成模型): Nicolas Chapados, Daniel Galvez, Wenming Ma, Fady Medhat, Shakir Mohamed and Grégoire Montavon.
- 参考文献: Lukas Michelbacher and Leslie N. Smith.

我们还要感谢那些允许我们从他们的出版物中复制图片、数据的人。我们在图片标题的文字中注明了他们的贡献。

我们还要感谢 Lu Wang 为我们写了 pdf2htmlEX, 我们用它来制作这本书的网页版本, Lu Wang 还帮助我们改进了生成的 HTML 的质量。

我们还要感谢 Ian 的妻子 Daniela Flori Goodfellow, 在 Ian 的写作过程中的耐心支持和检查。

我们还要感谢 Google Brain 团队提供了学术环境, 从而使得 Ian 能够花费大量时间写作此书并接受同行的反馈和指导。我们特别感谢 Ian 的前任经理 Greg Corrado 和他的现任经理 Samy Bengio 对这个项目的支持。最后我们还要感谢 Geoffrey Hinton 在写作困难时的鼓励。

# 网站

[www.deeplearningbook.org](http://www.deeplearningbook.org)

这本书伴随有上述网站。网站提供了各种补充材料，包括练习、讲义幻灯片、错误更正以及其他应该对读者和讲师有用的资源。

# 数学符号

本节简要介绍本书所使用的数学符号。我们在第二章至第四章中描述大多数数学概念，如果你不熟悉任何相应的数学概念，可以参考对应的章节。

## 数和数组

$a$	标量 (整数或实数)
$\mathbf{a}$	向量
$\mathbf{A}$	矩阵
$\mathbf{A}$	张量
$\mathbf{I}_n$	$n$ 行 $n$ 列的单位矩阵
$\mathbf{I}$	维度蕴含于上下文的单位矩阵
$\mathbf{e}^{(i)}$	标准基向量 $[0, \dots, 0, 1, 0, \dots, 0]$ ，其中索引 $i$ 处值为 1
$\text{diag}(\mathbf{a})$	对角方阵，其中对角元素由 $\mathbf{a}$ 给定
$a$	标量随机变量
$\mathbf{a}$	向量随机变量
$\mathbf{A}$	矩阵随机变量

## 集合和图

$\mathbb{A}$	集合
$\mathbb{R}$	实数集
$\{0, 1\}$	包含 0 和 1 的集合
$\{0, 1, \dots, n\}$	包含 0 和 $n$ 之间所有整数的集合
$[a, b]$	包含 $a$ 和 $b$ 的实数区间
$(a, b]$	不包含 $a$ 但包含 $b$ 的实数区间
$\mathbb{A} \setminus \mathbb{B}$	差集, 即其元素包含于 $\mathbb{A}$ 但不包含于 $\mathbb{B}$
$\mathcal{G}$	图
$Pa_{\mathcal{G}}(x_i)$	图 $\mathcal{G}$ 中 $x_i$ 的父节点

## 索引

$a_i$	向量 $\mathbf{a}$ 的第 $i$ 个元素, 其中索引从 1 开始
$a_{-i}$	除了第 $i$ 个元素, $\mathbf{a}$ 的所有元素
$A_{i,j}$	矩阵 $\mathbf{A}$ 的 $i, j$ 元素
$\mathbf{A}_{i,:}$	矩阵 $\mathbf{A}$ 的第 $i$ 行
$\mathbf{A}_{:,i}$	矩阵 $\mathbf{A}$ 的第 $i$ 列
$A_{i,j,k}$	3 维张量 $\mathbf{A}$ 的 $(i, j, k)$ 元素
$\mathbf{A}_{::,i}$	3 维张量的 2 维切片
$\mathfrak{a}_i$	随机向量 $\mathbf{a}$ 的第 $i$ 个元素

## 线性代数中的操作

- $\mathbf{A}^\top$  矩阵  $\mathbf{A}$  的转置
- $\mathbf{A}^+$   $\mathbf{A}$  的Moore-Penrose 伪逆
- $\mathbf{A} \odot \mathbf{B}$   $\mathbf{A}$  和  $\mathbf{B}$  的逐元素乘积 (Hadamard 乘积)
- $\det(\mathbf{A})$   $\mathbf{A}$  的行列式

## 微积分

- $\frac{dy}{dx}$   $y$  关于  $x$  的导数
- $\frac{\partial y}{\partial x}$   $y$  关于  $x$  的偏导
- $\nabla_x y$   $y$  关于  $\mathbf{x}$  的梯度
- $\nabla_{\mathbf{X}} y$   $y$  关于  $\mathbf{X}$  的矩阵导数
- $\nabla_{\mathbf{X}} y$   $y$  关于  $\mathbf{X}$  求导后的张量
- $\frac{\partial f}{\partial \mathbf{x}}$   $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  的Jacobian矩阵  $\mathbf{J} \in \mathbb{R}^{m \times n}$
- $\nabla_{\mathbf{x}}^2 f(\mathbf{x})$  or  $\mathbf{H}(f)(\mathbf{x})$   $f$  在点  $\mathbf{x}$  处的Hessian矩阵
- $\int f(\mathbf{x}) d\mathbf{x}$   $\mathbf{x}$  整个域上的定积分
- $\int_{\mathbb{S}} f(\mathbf{x}) d\mathbf{x}$  集合  $\mathbb{S}$  上关于  $\mathbf{x}$  的定积分

### 概率和信息论

$a \perp b$	a 和 b 相互独立的随机变量
$a \perp b \mid c$	给定 c 后条件独立
$P(a)$	离散变量上的概率分布
$p(a)$	连续变量（或变量类型未指定时）上的概率分布
$a \sim P$	具有分布 $P$ 的随机变量 a
$\mathbb{E}_{x \sim P}[f(x)]$ or $\mathbb{E}f(x)$	$f(x)$ 关于 $P(x)$ 的期望
$\text{Var}(f(x))$	$f(x)$ 在分布 $P(x)$ 下的方差
$\text{Cov}(f(x), g(x))$	$f(x)$ 和 $g(x)$ 在分布 $P(x)$ 下的协方差
$H(x)$	随机变量 x 的香农熵
$D_{\text{KL}}(P \parallel Q)$	P 和 Q 的KL 散度
$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	均值为 $\boldsymbol{\mu}$ 协方差为 $\boldsymbol{\Sigma}$ , $\mathbf{x}$ 上的高斯分布

## 函数

$f: \mathbb{A} \rightarrow \mathbb{B}$	定义域为 $\mathbb{A}$ 值域为 $\mathbb{B}$ 的函数 $f$
$f \circ g$	$f$ 和 $g$ 的组合
$f(\mathbf{x}; \boldsymbol{\theta})$	由 $\boldsymbol{\theta}$ 参数化, 关于 $\mathbf{x}$ 的函数 (有时为简化表示, 我们忽略 $\boldsymbol{\theta}$ 记为 $f(\mathbf{x})$ )
$\log x$	$x$ 的自然对数
$\sigma(x)$	Logistic sigmoid, $\frac{1}{1 + \exp(-x)}$
$\zeta(x)$	Softplus, $\log(1 + \exp(x))$
$\ \mathbf{x}\ _p$	$\mathbf{x}$ 的 $L^p$ 范数
$\ \mathbf{x}\ $	$\mathbf{x}$ 的 $L^2$ 范数
$x^+$	$x$ 的正数部分, 即 $\max(0, x)$
$\mathbf{1}_{\text{condition}}$	如果条件为真则为 1, 否则为 0

有时候我们使用函数  $f$ , 它的参数是一个标量, 但应用到一个向量、矩阵或张量:  $f(\mathbf{x})$ ,  $f(\mathbf{X})$ , or  $f(\mathbf{X})$ 。这表示逐元素地将  $f$  应用于数组。例如,  $\mathbf{C} = \sigma(\mathbf{X})$ , 则对于所有合法的  $i$ 、 $j$  和  $k$ ,  $C_{i,j,k} = \sigma(X_{i,j,k})$ 。

## 数据集和分布

$p_{\text{data}}$	数据生成分布
$\hat{p}_{\text{train}}$	由训练集定义的经验分布
$\mathbb{X}$	训练样本的集合
$\mathbf{x}^{(i)}$	数据集的第 $i$ 个样本 (输入)
$\mathbf{y}^{(i)}$ or $\mathbf{y}^{(i)}$	监督学习中与 $\mathbf{x}^{(i)}$ 关联的目标
$\mathbf{X}$	$m \times n$ 的矩阵, 其中行 $\mathbf{X}_{i,:}$ 为输入样本 $\mathbf{x}^{(i)}$

