

目录

基础部分	1
第 1 章 深度学习入门	2
1.1 简介	2
1.2 感知器	3
1.3 多层感知器	7
1.4 激活函数	9
1.5 损失函数	11
1.6 优化	13
1.6.1 梯度下降和误差的反向传播	13
1.6.2 随机梯度下降和自适应学习率	15
1.6.3 超参数筛选	17
1.7 正则化	18
1.7.1 过拟合	18
1.7.2 权重衰减	18
1.7.3 Dropout	20
1.7.4 批标准化	20
1.7.5 其他缓和过拟合的方法	21
1.8 卷积神经网络	22
1.9 循环神经网络	25
1.10 深度学习的实现样例	28
1.10.1 张量和梯度	28
1.10.2 定义模型	29
1.10.3 自定义层	31
1.10.4 多层感知器：MNIST 数据集上的图像分类	33

1.10.5	卷积神经网络：CIFAR-10 数据集上的图像分类	35
1.10.6	序列到序列模型：聊天机器人	36
第 2 章	强化学习入门	43
2.1	简介	43
2.2	在线预测和在线学习	46
2.2.1	简介	46
2.2.2	随机多臂赌博机	48
2.2.3	对抗多臂赌博机	50
2.2.4	上下文赌博机	51
2.3	马尔可夫过程	52
2.3.1	简介	52
2.3.2	马尔可夫奖励过程	54
2.3.3	马尔可夫决策过程	57
2.3.4	贝尔曼方程和最优性	61
2.3.5	其他重要概念	64
2.4	动态规划	64
2.4.1	策略迭代	65
2.4.2	价值迭代	67
2.4.3	其他 DP 的：异步 DP、近似 DP 和实时 DP	68
2.5	蒙特卡罗	70
2.5.1	蒙特卡罗预测	70
2.5.2	蒙特卡罗控制	71
2.5.3	增量蒙特卡罗	72
2.6	时间差分学习	73
2.6.1	时间差分预测	73
2.6.2	Sarsa：在线策略 TD 控制	77
2.6.3	Q-Learning：离线策略 TD 控制	80
2.7	策略优化	80
2.7.1	简介	80
2.7.2	基于价值的优化	84
2.7.3	基于策略的优化	89
2.7.4	结合基于策略和基于价值的方法	105

第 3 章 强化学习算法分类	110
3.1 基于模型的方法和无模型的方法	111
3.2 基于价值的方法和基于策略的方法	113
3.3 蒙特卡罗方法和时间差分方法	114
3.4 在线策略方法和离线策略方法	115
第 4 章 深度 Q 网络	119
4.1 Sarsa 和 Q-Learning	121
4.2 为什么使用深度学习: 价值函数逼近	121
4.3 DQN	123
4.4 Double DQN	124
4.5 Dueling DQN	125
4.6 优先经验回放	127
4.7 其他改进内容: 多步学习、噪声网络和值分布强化学习	128
4.8 DQN 代码实例	131
第 5 章 策略梯度	146
5.1 简介	146
5.2 REINFORCE: 初版策略梯度	147
5.3 Actor-Critic	149
5.4 生成对抗网络和 Actor-Critic	150
5.5 同步优势 Actor-Critic	152
5.6 异步优势 Actor-Critic	153
5.7 信赖域策略优化	154
5.8 近端策略优化	157
5.9 使用 Kronecker 因子化信赖域的 Actor-Critic	159
5.10 策略梯度代码例子	162
5.10.1 相关的 Gym 环境	162
5.10.2 REINFORCE: Atari Pong 和 CartPole-V0	165
5.10.3 AC: CartPole-V0	173
5.10.4 A3C: BipedalWalker-v2	176
5.10.5 TRPO: Pendulum-V0	181
5.10.6 PPO: Pendulum-V0	192

第 6 章 深度 Q 网络和 Actor-Critic 的结合	200
6.1 简介	200
6.2 深度确定性策略梯度算法	201
6.3 孪生延迟 DDPG 算法	203
6.4 柔性 Actor-Critic 算法	206
6.4.1 柔性策略迭代	206
6.4.2 SAC	207
6.5 代码例子	209
6.5.1 相关的 Gym 环境	209
6.5.2 DDPG: Pendulum-V0	209
6.5.3 TD3: Pendulum-V0	215
6.5.4 SAC: Pendulum-v0	225
研究部分	236
第 7 章 深度强化学习的挑战	237
7.1 样本效率	237
7.2 学习稳定性	240
7.3 灾难性遗忘	242
7.4 探索	243
7.5 元学习和表征学习	245
7.6 多智能体强化学习	246
7.7 模拟到现实	247
7.8 大规模强化学习	251
7.9 其他挑战	252
第 8 章 模仿学习	258
8.1 简介	258
8.2 行为克隆方法	260
8.2.1 行为克隆方法的挑战	260
8.2.2 数据集聚合	261
8.2.3 Variational Dropout	262
8.2.4 行为克隆的其他方法	262
8.3 逆向强化学习方法	263
8.3.1 简介	263
8.3.2 逆向强化学习方法的挑战	264

8.3.3	生成对抗模仿学习	265
8.3.4	生成对抗网络指导性代价学习	266
8.3.5	对抗性逆向强化学习	268
8.4	从观察量进行模仿学习	269
8.4.1	基于模型方法	269
8.4.2	无模型方法	272
8.4.3	从观察量模仿学习的挑战	277
8.5	概率性方法	277
8.6	模仿学习作为强化学习的初始化	279
8.7	强化学习中利用示范数据的其他方法	280
8.7.1	将示范数据导入经验回放缓存	280
8.7.2	标准化 Actor-Critic	281
8.7.3	用示范数据进行奖励塑形	282
8.8	总结	282
第 9 章	集成学习与规划	289
9.1	简介	289
9.2	基于模型的方法	290
9.3	集成模式架构	292
9.4	基于模拟的搜索	293
9.4.1	朴素蒙特卡罗搜索	294
9.4.2	蒙特卡罗树搜索	294
9.4.3	时间差分搜索	295
第 10 章	分层强化学习	298
10.1	简介	298
10.2	选项框架	299
10.2.1	战略专注作家	300
10.2.2	选项-批判者结构	303
10.3	封建制强化学习	305
10.3.1	封建制网络	305
10.3.2	离线策略修正	307
10.4	其他工作	309

第 11 章 多智能体强化学习	315
11.1 简介	315
11.2 优化和均衡	316
11.2.1 纳什均衡	317
11.2.2 关联性均衡	318
11.2.3 斯塔克尔伯格博弈	320
11.3 竞争与合作	321
11.3.1 合作	321
11.3.2 零和博弈	321
11.3.3 同时决策下的竞争	322
11.3.4 顺序决策下的竞争	323
11.4 博弈分析架构	324
第 12 章 并行计算	326
12.1 简介	326
12.2 同步和异步	327
12.3 并行计算网络	329
12.4 分布式强化学习算法	330
12.4.1 异步优势 Actor-Critic	330
12.4.2 GPU/CPU 混合式异步优势 Actor-Critic	332
12.4.3 分布式近端策略优化	333
12.4.4 重要性加权的行动者-学习者结构和可扩展高效深度强化学习	336
12.4.5 Ape-X、回溯-行动者和分布式深度循环回放 Q 网络	338
12.4.6 Gorila	340
12.5 分布式计算架构	340
应用部分	343
第 13 章 Learning to Run	344
13.1 NeurIPS 2017 挑战: Learning to Run	344
13.1.1 环境介绍	344
13.1.2 安装	346
13.2 训练智能体	347
13.2.1 并行训练	348
13.2.2 小技巧	351
13.2.3 学习结果	352

第 14 章 鲁棒的图像增强	354
14.1 图像增强	354
14.2 用于鲁棒处理的强化学习	356
第 15 章 AlphaZero	366
15.1 简介	366
15.2 组合博弈	367
15.3 蒙特卡罗树搜索	370
15.4 AlphaZero: 棋类游戏的通用算法	376
第 16 章 模拟环境中机器人学习	388
16.1 机器人模拟	389
16.2 强化学习用于机器人任务	405
16.2.1 并行训练	407
16.2.2 学习效果	407
16.2.3 域随机化	408
16.2.4 机器人学习基准	409
16.2.5 其他模拟器	409
第 17 章 Arena: 多智能体强化学习平台	412
17.1 安装	413
17.2 用 Arena 开发游戏	413
17.2.1 简单的单玩家游戏	414
17.2.2 简单的使用奖励机制的双玩家游戏	416
17.2.3 高级设置	420
17.2.4 导出二进制游戏	424
17.3 MARL 训练	427
17.3.1 设置 X-Server	427
17.3.2 进行训练	429
17.3.3 可视化	431
17.3.4 致谢	431
第 18 章 深度强化学习应用实践技巧	433
18.1 概览: 如何应用深度强化学习	433
18.2 实现阶段	434
18.3 训练和调试阶段	440

总结部分	445
附录 A 算法总结表	446
附录 B 算法速查表	451
B.1 深度学习	451
B.1.1 随机梯度下降	451
B.1.2 Adam 优化器	452
B.2 强化学习	452
B.2.1 赌博机	452
B.2.2 动态规划	453
B.2.3 蒙特卡罗	454
B.3 深度强化学习	458
B.4 高等深度强化学习	467
B.4.1 模仿学习	467
B.4.2 基于模型的强化学习	468
B.4.3 分层强化学习	470
B.4.4 多智能体强化学习	471
B.4.5 并行计算	472
附录 C 中英文对照表	476

