

目录

第一章 简介	1
第二章 马尔科夫决策过程	3
2.1 马尔科夫过程	3
2.2 马尔科夫奖励过程	5
2.3 马尔科夫决策过程	10
2.4 编程实践——学生马尔科夫决策示例	18
2.4.1 收获和价值的计算	18
2.4.2 验证贝尔曼方程	21
第三章 动态规划寻找最优策略	31
3.1 策略评估	32
3.2 策略迭代	34
3.3 价值迭代	36
3.4 异步动态规划算法	39
3.5 编程实践——动态规划求解小型方格世界最优策略	40
3.5.1 小型方格世界 MDP 建模	40
3.5.2 策略评估	45
3.5.3 策略迭代	45
3.5.4 价值迭代	46
第四章 不基于模型的预测	49
4.1 蒙特卡罗强化学习	49
4.2 时序差分强化学习	51
4.3 n 步时序差分学习简介	58
4.4 编程实践：蒙特卡罗学习评估 21 点游戏的玩家策略	62

4.4.1	二十一点游戏规则	62
4.4.2	将二十一点游戏建模为强化学习问题	63
4.4.3	游戏场景的搭建	64
4.4.4	生成对局数据	74
4.4.5	策略评估	75
第五章	不基于模型的控制	79
5.1	行为价值函数的重要性	80
5.2	ϵ -贪婪策略	81
5.3	现时策略蒙特卡罗控制	82
5.4	现时策略时序差分控制	83
5.4.1	Sarsa 算法	83
5.4.2	Sarsa(λ) 算法	86
5.4.3	比较 Sarsa 和 Sarsa(λ)	88
5.5	借鉴策略 Q 学习算法	91
5.6	编程实践：蒙特卡罗学习求二十一点游戏最优策略	93
5.7	编程实践：构建基于 gym 的有风格子世界及个体	96
5.7.1	gym 库简介	97
5.7.2	状态序列的管理	98
5.7.3	个体基类的编写	100
5.8	编程实践：各类学习算法的实现及与有风格子世界的交互	105
5.8.1	Sarsa 算法	106
5.8.2	Sarsa(λ) 算法	107
5.8.3	Q 学习算法	108
第六章	价值函数的近似表示	111
6.1	价值近似的意义	111
6.2	目标函数与梯度下降	114
6.2.1	目标函数	114
6.2.2	梯度和梯度下降	116
6.3	常用的近似价值函数	120
6.3.1	线性近似	120
6.3.2	神经网络	121
6.3.3	卷积神经网络近似	124

6.4	DQN 算法	129
6.5	编程实践：基于 PyTorch 实现 DQN 求解 PuckWorld 问题	130
6.5.1	基于神经网络的近似价值函数	131
6.5.2	实现 DQN 求解 PuckWorld 问题	134
第七章	基于策略梯度的深度强化学习	141
7.1	基于策略学习的意义	141
7.2	策略目标函数	143
7.3	Actor-Critic 算法	146
7.4	深度确定性策略梯度 (DDPG) 算法	149
7.5	编程实践：DDPG 算法实现	151
7.5.1	连续行为空间的 PuckWorld 环境	151
7.5.2	Actor-Critic 网络的实现	152
7.5.3	确定性策略下探索的实现	156
7.5.4	DDPG 算法的实现	157
7.5.5	DDPG 算法在 PuckWorld 环境中的表现	163
第八章	基于模型的学习和规划	165
8.1	环境的模型	165
8.2	整合学习与规划——Dyna 算法	167
8.3	基于模拟的搜索	167
8.3.1	简单蒙特卡罗搜索	169
8.3.2	蒙特卡罗树搜索	170
第九章	探索与利用	171
9.1	多臂赌博机	171
9.2	常用的探索方法	174
9.2.1	衰减的 ϵ -贪婪探索	174
9.2.2	不确定行为优先探索	174
9.2.3	基于信息价值的探索	178
第十章	Alpha Zero 算法浅析	181
10.1	Alpha Zero 算法核心思想	181
10.1.1	蒙特卡罗树搜索	184

10.2 编程实践:AlphaZero 代码赏析	186
10.2.1 蒙特卡罗树搜索	186

Author: 叶强 qqiangye@gmail.com