

前 言

第 1 章

CHAPTER.1

强化学习基础 / 1

- 1.1 从监督学习到强化学习 / 1
- 1.2 强化学习的发展历史 / 4
- 1.3 强化学习的研究范畴 / 5
- 1.4 强化学习的应用领域 / 8

第 2 章

CHAPTER.2

强化学习研究范畴与应用领域 / 10

- 2.1 强化学习的核心概念 / 10
- 2.2 马尔可夫性和决策过程 / 11
- 2.3 值函数和策略学习 / 13

第 3 章

CHAPTER.3

学习值函数的强化学习算法 / 17

- 3.1 深度 Q 学习的基本理论 / 17
 - 3.1.1 深度 Q 网络 / 18
 - 3.1.2 经验池 / 19
 - 3.1.3 目标网络 / 21
- 3.2 深度 Q 学习的过估计 / 22
 - 3.2.1 过估计的产生原因 / 22
 - 3.2.2 Double Q -学习 / 23
- 3.3 深度 Q 学习的网络改进和高效采样 / 24
 - 3.3.1 竞争网络 / 24
 - 3.3.2 高效采样 / 26


- 3.4 周期后序迭代 Q 学习 / 27
- 3.5 Q 学习用于连续动作空间 / 29
 - 3.5.1 基于并行结构的 Q 学习 / 30
 - 3.5.2 基于顺序结构的 Q 学习 / 31
- 3.6 实例：使用值函数学习的雅塔力游戏 / 32
 - 3.6.1 环境预处理 / 33
 - 3.6.2 Q 网络的实现 / 37
 - 3.6.3 Q 学习的核心步骤 / 38

第 4 章
CHAPTER 4


策略梯度迭代的强化学习算法 / 41

- 4.1 REINFORCE 策略梯度 / 41
 - 4.1.1 策略梯度的基本形式 / 42
 - 4.1.2 降低策略梯度的方差 / 44
- 4.2 异步策略梯度法 / 45
 - 4.2.1 引入优势函数 / 46
 - 4.2.2 异步策略梯度 / 47
- 4.3 近端策略优化法 / 48
 - 4.3.1 裁剪的优化目标 / 49
 - 4.3.2 自适应的优化目标 / 49
- 4.4 确定性策略梯度 / 50
 - 4.4.1 Critic 学习 / 51
 - 4.4.2 Actor 学习 / 51
 - 4.4.3 拓展 1：探索噪声 / 52
 - 4.4.4 拓展 2：孪生 DDPG / 52
- 4.5 最大熵策略梯度 / 54
 - 4.5.1 熵约束的基本原理 / 54
 - 4.5.2 2SAC 算法 / 55
- 4.6 实例：使用策略梯度的 mujoco 任务 / 56
 - 4.6.1 Actor-Critic 网络实现 / 56
 - 4.6.2 核心算法实现 / 58


第 5 章
CHAPTER.5**基于模型的强化学习方法 / 61**

- 
- 5.1 如何使用模型来进行强化学习 / 61
 - 5.2 基于模型预测的规划 / 63
 - 5.2.1 随机打靶法 / 63
 - 5.2.2 集成概率轨迹采样法 / 64
 - 5.2.3 基于模型和无模型的混合算法 / 65
 - 5.2.4 基于想象力的隐式规划方法 / 66
 - 5.3 黑盒模型的理论框架 / 67
 - 5.3.1 随机下界优化算法 / 68
 - 5.3.2 基于模型的策略优化算法 / 70
 - 5.4 白盒模型的使用 / 71
 - 5.4.1 随机值梯度算法 / 71
 - 5.4.2 模型增强的演员-评论家算法 / 72
 - 5.5 实例：AlphaGo 围棋智能体 / 74
 - 5.5.1 网络结构介绍 / 74
 - 5.5.2 蒙特卡洛树搜索 / 75
 - 5.5.3 总体训练流程 / 79


第 6 章
CHAPTER.6**值分布式强化学习算法 / 80**

- 
- 6.1 离散分布投影的值分布式算法 / 80
 - 6.2 分位数回归的值分布式算法 / 83
 - 6.2.1 分位数回归 / 83
 - 6.2.2 Wasserstein 距离 / 86
 - 6.2.3 QR-DQN 算法 / 88
 - 6.2.4 单调的分位数学习算法 / 90
 - 6.3 隐式的值分布网络 / 91
 - 6.4 基于值分布的代价敏感学习 / 93
 - 6.4.1 IQN 中的代价敏感学习 / 94
 - 6.4.2 基于 IQN 的 Actor-Critic 模型的代价敏感学习 / 95
 - 6.5 实例：基于值分布的 Q 网络实现 / 95

• 6.5.1 IQN 模型构建 / 96

• 6.5.2 IQN 损失函数 / 97

第 7 章 强化学习中的探索算法 / 100

CHAPTER.7

• 7.1 探索算法的分类 / 100

• 7.2 基于不确定性估计的探索 / 102

• 7.2.1 参数化后验的算法思路 / 103

• 7.2.2 重采样 DQN (Bootstrapped DQN) / 104

• 7.3 进行虚拟计数的探索 / 110

• 7.3.1 基于图像生成模型的虚拟计数 / 110

• 7.3.2 基于哈希的虚拟计数 / 113

• 7.4 根据环境模型的探索 / 114

• 7.4.1 特征表示的学习 / 115

• 7.4.2 随机网络蒸馏 / 117

• 7.4.3 Never-Give-Up 算法 / 117

• 7.5 实例：蒙特祖玛复仇任务的探索 / 119

• 7.5.1 RND 网络结构 / 119

• 7.5.2 RND 的训练 / 120

• 7.5.3 RND 用于探索 / 121

第 8 章 多目标强化学习算法 / 122

CHAPTER.8

• 8.1 以目标为条件的价值函数 / 122

• 8.1.1 最大熵 HER / 125

• 8.1.2 动态目标 HER / 125

• 8.2 监督式的多目标学习 / 126

• 8.2.1 Hindsight 模仿学习 / 127

• 8.2.2 加权 Hindsight 模仿学习 / 128

• 8.3 推广的多目标学习 / 129

• 8.4 实例：仿真机械臂的多目标抓取 / 131

• 8.4.1 多目标实验环境 / 131

• 8.4.2 HER 的实现方法 / 132



第 9 章

CHAPTER.9

8.4.3 MEP 的算法实现 / 134

层次化强化学习算法 / 135

9.1 层次化学习的重要性 / 135

9.2 基于子目标的层次化学习 / 136

9.2.1 封建网络的层次化学习 / 137

9.2.2 离策略修正的层次化学习 / 139

9.2.3 虚拟子目标的强化学习方法 / 140

9.3 基于技能的层次化学习 / 141

9.3.1 使用随机网络的层次化学习 / 142

9.3.2 共享分层的元学习方法 / 143

9.4 基于选项的层次化学习 / 145

9.4.1 option 与半马尔可夫决策过程 / 145

9.4.2 Option-Critic 结构 / 147

9.5 实例：层次化学习 Ant 走迷宫任务 / 148

第 10 章

CHAPTER.10

基于技能的强化学习算法 / 156

10.1 技能学习的定义 / 156

10.2 互信息最大化的技能学习算法 / 157

10.2.1 多样性最大化技能学习算法 / 157

10.2.2 其他基于互信息的技能学习方法 / 160

10.3 融合环境模型的技能学习算法 / 162

10.4 最大化状态覆盖的技能学习算法 / 164

10.5 实例：人形机器人的技能学习 / 166

第 11 章

CHAPTER.11

离线强化学习算法 / 171

11.1 离线强化学习中面临的困难 / 171

11.2 策略约束的离线学习 / 172

11.2.1 BCQ 算法 / 175

11.2.2 BRAC 算法 / 177

- 11.2.3 TD3-BC 算法 / 178
- 11.3 使用保守估计的离线学习 / 179
 - 11.3.1 分布式 CQL / 181
- 11.4 基于不确定性的离线学习 / 182
 - 11.4.1 UWAC 算法 / 182
 - 11.4.2 MOPO 算法 / 183
 - 11.4.3 PBRL 算法 / 185
- 11.5 监督式的离线学习 / 188
 - 11.5.1 DT 算法 / 188
 - 11.5.2 RVS 算法 / 189
- 11.6 实例：使用离线学习的 D4RL 任务集 / 190
 - 11.6.1 D4RL 数据集的使用 / 193
 - 11.6.2 CQL 算法实现 / 193
 - 11.6.3 TD3-BC 算法实现 / 195

第 12 章
CHAPTER.12

元强化学习算法 / 197

- 12.1 元强化学习的定义 / 197
- 12.2 基于网络模型的元强化学习方法 / 198
 - 12.2.1 使用循环神经网络的元强化学习方法 / 199
 - 12.2.2 基于注意力机制和时序卷积的方法 / 200
- 12.3 元梯度学习 / 201
- 12.4 元强化学习中探索方法 / 205
 - 12.4.1 结构化噪声探索法 / 205
 - 12.4.2 利用后验采样进行探索 / 206
- 12.5 实例：元学习训练多任务猎豹智能体 / 209

第 13 章
CHAPTER.13

高效的强化学习表示算法 / 216

- 13.1 为什么要进行表示学习 / 216
- 13.2 对比学习的特征表示 / 217
 - 13.2.1 基本原理和 SimCLR 算法 / 218
 - 13.2.2 MoCo 算法 / 220



13.2.3 基于对比学习的 CURL 算法 / 222

13.2.4 基于对比学习的 ATC 算法 / 223

13.2.5 基于对比学习的 DIM 算法 / 224

13.2.6 对比学习和互信息理论 / 225

13.2.7 完全基于图像增广的方法 / 228

13.3 鲁棒的特征表示学习 / 229

13.3.1 互模拟特征 / 229

13.3.2 信息瓶颈特征 / 232

13.4 使用模型预测的表示学习 / 235

13.5 实例：鲁棒的仿真自动驾驶 / 237

第 14 章

CHAPTER.14

强化学习在智能控制中的
应用 / 240

14.1 机器人控制 / 240

14.1.1 机械臂操作任务的控制 / 240

14.1.2 足式机器人的运动控制 / 242

14.1.3 多任务机器人控制 / 244

14.1.4 面临的挑战 / 246

14.2 电力优化控制 / 248

14.2.1 电力管理任务 / 248

14.2.2 需求响应 / 249

14.3 交通指挥优化控制 / 251

14.3.1 多信号灯合作控制 / 251

14.3.2 大规模信号灯控制方法 / 252

14.3.3 元强化学习信号灯控制 / 254

第 15 章

CHAPTER.15

强化学习在机器视觉中的应用 / 256

15.1 神经网络结构搜索 / 256

15.1.1 利用强化学习解决 NAS / 256

15.1.2 其它前沿方法 / 259

15.2 目标检测和跟踪中的优化 / 262

- 15.2.1 强化学习与目标检测 / 263
- 15.2.2 强化学习与实时目标跟踪 / 264
- 15.3 视频分析 / 266

第 16 章

CHAPTER.16

强化学习在语言处理中的应用 / 269

- 16.1 知识图谱系统 / 269
- 16.2 智能问答系统 / 271
 - 16.2.1 事后目标回放法 / 273
 - 16.2.2 多任务对话系统 / 273
- 16.3 机器翻译系统 / 275
 - 16.3.1 NMT 中奖励的计算 / 276
 - 16.3.2 策略梯度方差处理 / 277

第 17 章

CHAPTER.17

强化学习在其他领域中的应用 / 278

- 17.1 医疗健康系统 / 278
 - 17.1.1 动态治疗方案 / 278
 - 17.1.2 重症监护 / 280
 - 17.1.3 自动医疗诊断 / 281
- 17.2 个性化推荐系统 / 282
 - 17.2.1 策略优化方法 / 283
 - 17.2.2 基于图的对话推荐 / 284
- 17.3 股票交易系统 / 285
 - 17.3.1 FinRL 强化学习框架 / 286
 - 17.3.2 FinRL 训练示例 / 287

附录

相关学习资料 / 290