Reinforcement Learning China Summer School

RLChina 2021

习题课4
多智能体合作

林舒

中国科学院自动化研究所

2021年8月19日

* 课程内容参考 汪军教授《Multi-agent AI》（公众号回复MAAI）
* 习题课代码仓库 https://gitee.com/jidiai/summercourse2021

回顾：马尔可夫决策过程

- $MDP = \langle S, A, p, r, \gamma \rangle$
  - 状态集 $S = \{s_1, s_2, \dots, s_n\}$
  - 动作集 $A = \{a_1, a_2, \dots, a_m\}$
  - 状态转移函数
    $p(s' \mid s, a) = \Pr(S_{t+1} = s' \mid S_t = s, A_t = a)$
  - 奖励函数 $r(s, a)$
  - 折扣因子 $\gamma \in [0,1)$

多智能体随机博弈

- $SG = \langle n, S, A, p, r, \gamma \rangle$
  - 智能体数量 n
  - 状态集 $S = \{s_1, s_2, \ldots, s_n\}$
  - 动作集 $\boldsymbol{A} = A_1 \times A_2 \times \cdots \times A_n$
  - 状态转移函数
    $p(s' \mid s, \boldsymbol{a}) = \Pr(S_{t+1} = s' \mid S_t = s, A_t = \boldsymbol{a})$
    其中联合动作$\boldsymbol{a} = [a_1, a_2, \cdots, a_n] \in \boldsymbol{A}$
  - 奖励函数 $\boldsymbol{r}(s, \boldsymbol{a}) = [r_1(s, \boldsymbol{a}), r_2(s, \boldsymbol{a}), \cdots, r_n(s, \boldsymbol{a})]$
  - 折扣因子 $\gamma \in [0,1)$

随机博弈分类

- 根据 $\boldsymbol{r}(s,\boldsymbol{a}) = [r_1(s,\boldsymbol{a}), r_2(s,\boldsymbol{a}), \cdots, r_n(s,\boldsymbol{a})]$ 的特点分类：

1. 纯合作
   - 每个状态和联合动作对$(s,\boldsymbol{a})$下，各智能体获得的奖励相同
   - $r_1(s,\boldsymbol{a}) = \mathrm{r}_2(s,\boldsymbol{a}) = \cdots = r_n(s,\boldsymbol{a})$
2. 纯竞争
   - 每个状态和联合动作对$(s,\boldsymbol{a})$下，各智能体获得的奖励和为0
   - $r_1(s,\boldsymbol{a}) + \mathrm{r}_2(s,\boldsymbol{a}) + \cdots + r_n(s,\boldsymbol{a}) = 0$
3. 混合（其他情况）

IQL——Independent Q-Learning

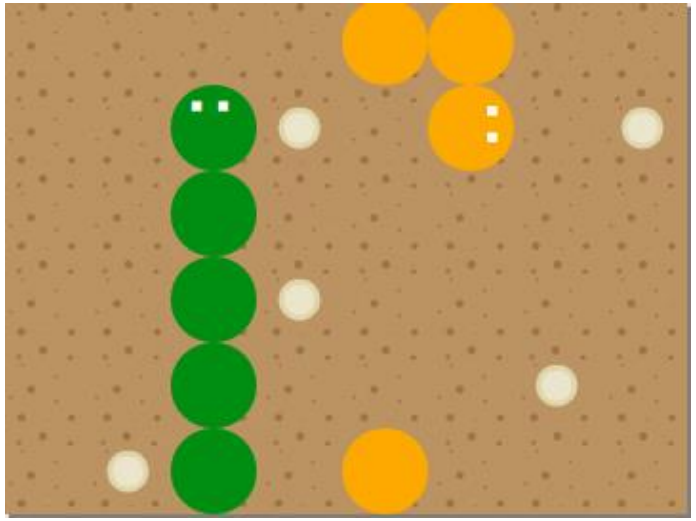
- 及第秘籍http://www.jidiai.cn/iql
- 核心思想：
  - 每个智能体独立使用DQN训练
  - 训练智能体i时，将其他智能体直接看作环境的一部分
- 优点：
  - 直接沿用单智能体算法和训练框架
- 缺点：
  - 智能体间缺乏合作和沟通
  - 环境不稳定，收敛比较慢
- 在工程实践上，具有不错的效果

多智能体合作环境下的价值网络共享

- 同构多智能体
  - 各智能体有完全相同的能力（观测能力、行动能力）
  - 各智能体可相互替换
  - 各智能体可以采用完全相同的策略π
- 共享价值网络
  - 在IQL算法中，不同智能体共享价值网络（Q_θ，$Q_{\theta'}$）
  - 优点：
    - 减少训练参数
  - 缺点：
    - 可能陷入局部最优，或者出现内部竞争

贪吃蛇2P——同构多智能体纯合作环境

- 及第科目http://www.jidiai.cn/snake2p



- 控制两条蛇，在规定步数（30）内通过吃豆子增加长度
- 若一条蛇头撞上自己或另一条蛇的蛇身会死亡，并随机以长度3重生
- 最终在第50步时，积分=(蛇A长度-3)+(蛇B长度-3)

第四次作业：贪吃蛇(2P)游戏

- 及第科目→单方多智能体
  - http://www.jidiai.cn/snake2p
- 作业本地训练环境、算法代码、训练说明等
  - https://gitee.com/jidiai/summercourse2021/tree/main/course4
  - https://github.com/jidiai/SummerCourse2021/tree/main/course4
- 作业要求
  - 训练贪吃蛇(2P)游戏的多智体合作算法
  - 将homework里的submission.py填写完整
  - 将submission.py，critic.py，critic_*.pth提交到及第平台

如何判断是否成功完成作业？

查看成绩：

登录 及第Jidi →

点击右上角个人头像，点击个人中心 →

在“贪吃蛇(2P)”一行：

积分>8

即成功完成第四次作业

更多多智能体算法

- MADDPG
  - http://www.jidiai.cn/maddpg
  - 集中式训练，分布式执行
  - 既能用于合作环境，也能用于竞争环境
- Bidirectionally-Coordinated Network (BiCNet)
  - http://www.jidiai.cn/bicnet
  - 建立智能体间沟通协调机制
  - 主要用于同构或异构智能体间合作