Reinforcement Learning China Summer School

RLChina 2021

# 习题课3
# 基于神经网络的强化学习算法

林舒

中国科学院自动化研究所

2021年8月18日

* 课程内容参考《动手学强化学习》http://hrl.boyuai.com/
* 习题课代码仓库 https://gitee.com/jidiai/summercourse2021

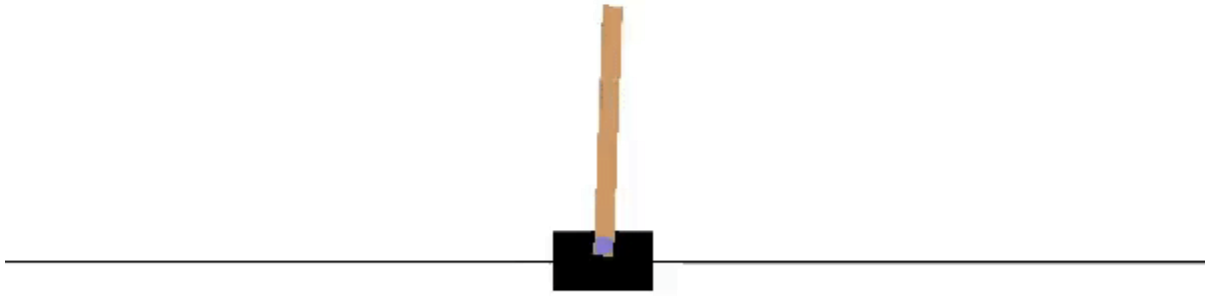
# 回顾：基于表的强化学习算法

- SARSA / Q-Learning
- 特点：需要计算、存储动作价值表$Q(s,a)$
- 缺点：
  - 无法存储大量数据
  - 无法将经验泛化到未见过的状态

- 适用问题：动作、状态空间有限且比较小

# 车杆游戏

- 及第科目http://www.jidiai.cn/cartpole



- 连续状态<车位置，车速度，杆角度，杆角速度>
- 动作{-1向左，+1向右}

- 无法使用表存储$Q(s,a)$!

# 用参数化函数拟合动作价值函数

- $Q_\theta(s,a) \approx Q\ (s,a)$
  - $Q_\theta$是关于$\theta$的可微函数，即偏导$\frac{\partial Q_\theta(s,a)}{\partial \theta}$存在
  - 一般采用特征线性组合函数或神经网络来拟合

# DQN——使用神经网络拟合$Q(s,a)$

- 及第秘籍http://www.jidiai.cn/dqn
- 价值更新：最小化均方误差MSE

$$\theta \leftarrow \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \left[ Q_\theta(s_i, a_i) - \left( r_i + \gamma \max_{a'} Q_\theta(s_i', a') \right) \right]^2$$

- 优化1：经验回放Experiment Replay
  - 将环境采样数据$\langle s, a, r, s' \rangle$存放在回放池
  - 每次训练时从回放池中随机采样
  - 作用：增强样本独立性；提高样本利用率
- 优化2：目标网络Target Network
  - 增加一套目标网络，与原训练网络结构相同但使用较旧参数$\theta'$
  - $\theta \leftarrow \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \left[ Q_\theta(s_i, a_i) - \left( r_i + \gamma \max_{a'} Q_{\theta'}(s_i', a') \right) \right]^2$
  - 每隔若干步，同步两个网络参数$\theta' \leftarrow \theta$
  - 作用：使训练相对稳定

# DQN

用任意初始参数$\theta$初始化训练网络$Q_\theta$

使用相同参数$\theta' \leftarrow \theta$初始化目标网络$Q_{\theta'}$

初始化回放池$R \leftarrow \emptyset$

REPEAT max_episodes次:

    $s \leftarrow S_0$

    WHILE $s$不是终止状态:

        $a \leftarrow \epsilon$-greedy策略根据$s$和$Q_\theta$选取动作

        $r, s' \leftarrow$ 采用动作$a$后，环境反馈的奖励和下一个状态

        将环境采样数据$\langle s, a, r, s' \rangle$存入$R$

        $s \leftarrow s'$

        IF $|R| \geq M$:

            从$R$中随机采样$N$个样本$\{\langle s_i, a_i, r_i, s_i' \rangle\}$

            对每个样本，用目标网络$Q_{\theta'}$计算$y_i = r_i + \gamma \max_{a'} Q_{\theta'}(s_i', a')$

            更新$Q_\theta$以最小化损失$L \leftarrow \frac{1}{N}\sum_{i=1}^{N}[Q_\theta(s_i, a_i) - y_i]^2$

            IF $Q_{\theta'}$滞后步数 = target_replace:

                同步$\theta' \leftarrow \theta$

# 习题课代码仓库里的相关文件说明

- course3\examples\algo\dqn\dqn.py
  - DQN算法核心实现
- course3\examples\networs\critic.py
  - 用于拟合动作价值函数的神经网络
- course3\examples\models\config_training\dqn_classic_CartPole-v0.yaml
  - 各模块的参数配置：包括算法、网络、环境、训练
- course3\examples\runner.py
  - 训练框架
- course3\examples\algo\homework*.py
  - 作业需要实现和提交的submission.py和critic.py

# 第三次作业：车杆游戏

- 及第科目介绍及提交入口
  - http://www.jidiai.cn/cartpole
- 作业本地训练环境、算法代码、训练说明等
  - https://gitee.com/jidiai/summercourse2021/tree/main/course3
  - https://github.com/jidiai/SummerCourse2021/tree/main/course3
- 作业要求
  - 训练车杆游戏的DQN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？


成功！

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“车杆”一行：

积分>80

即成功完成第三次作业