Reinforcement Learning China Summer School

RLChina 2021

习题课2
基于表的强化学习算法

林舒

中国科学院自动化研究所

2021年8月17日

*** 课程内容参考《动手学强化学习》http://hrl.boyuai.com/**

RLChina 2021暑期课习题课代码仓库

- 仓库链接
  - https://gitee.com/jidiai/summercourse2021
  - https://github.com/jidiai/SummerCourse2021

发布每天习题课作业所需资料

- 环境：提交链接、本地训练代码
- 算法：算法代码、训练框架、样例
- 说明：训练方法说明、补充说明等

注：
仓库会随着习题课进度逐步更新完善
请及时查看仓库链接获取最新版本

习题课第二天

任务：环境悬崖漫步 - 算法Q-learning & SARSA - 提交到Jidi平台，成绩优于随机10%

提交链接：http://www.jidiai.cn/cliffwalking

Env 👉**请看** env/cliffwalking.py

Q-learning 👉**请看** examples/algo/tabularq.py

Sarsa 👉**请看** examples/algo/sarsa.py

How to train your rl_agent:

have a go~

```
python main.py --scenario cliffwalking --algo sarsa --reload_config
```

```
python main.py --scenario cliffwalking --algo tabularq --reload_config
```

说明：

1. 算法需要在本地训练，及第平台提供了经典算法实现、训练框架和提交样例。
2. 在config文件夹里，已经保存了算法库对接多个环境和多个算法的训练参数。支持一键复现，只需要加 --reload_config这个参数（So cool...
3. 训练开始后，会生成models文件夹，在models/config_training里面保存了训练过程中的参数。可以试着不加reload_config，就在👈里调参，主run会自动上传这里的参数：例如python main.py --scenario cliffwalking --algo sarsa

Bonus

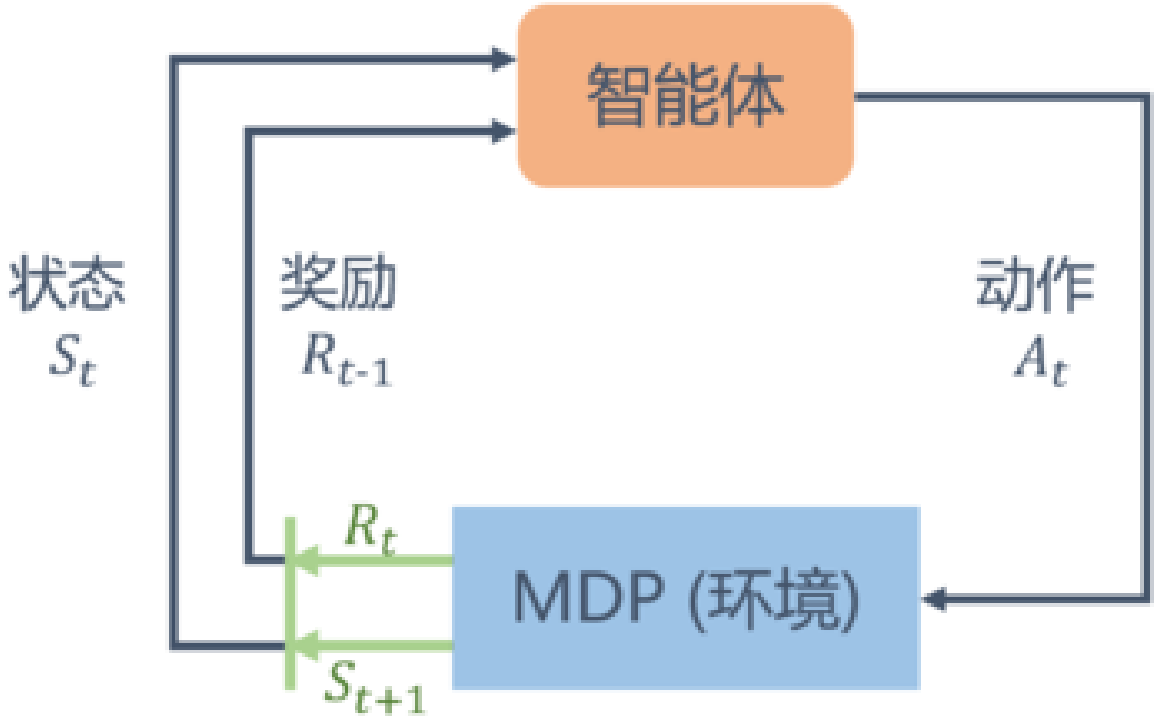
gridworld和cliffwarking都是网格环境，智能体tabularq依然是"冒险家"，sarsa还是"保险主义"。运行试试吧^0^

关于算法训练

- 习题课算法的训练，可以选择以下两种方式之一
  - 本地训练
    - 从习题课代码仓库下载环境、算法、训练框架代码等
    - 根据训练说明，在自己的机器上完成本地训练
  - 在线训练
    - 根据和鲸平台的手册，在平台上完成在线训练
    - 手册链接： https://jidi-images.oss-cn-beijing.aliyuncs.com/rlchina2021/%5BRLChina%E4%B9%A0%E9%A2%98%E8%AF%BE%5D%E5%92%8C%E9%B2%B8%E5%B9%B3%E5%8F%B0.pdf

智能体与环境交互

智能体策略 $\pi(a|s) = \Pr(\mathrm{A_t} = \mathrm{a}|\mathrm{S_t} = \mathrm{s})$



基于动态规划的强化学习算法

- 状态价值函数

$$V^{\pi}(s) = \mathbb{E}[R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \cdots | s_0 = s, \pi]$$
$$= R(s) + \gamma \sum_{s'} p(s'|s, \pi(s)) V^{\pi}(s')$$

- 最优状态价值函数

$$V^*(s) = \max_{\pi} V^{\pi}(s)$$
$$= R(s) + \max_{a} \gamma \sum_{s'} p(s'|s, a) V^*(s')$$

- 最优策略

$$\pi^*(s) = \arg\max_{a} \sum_{s'} p(s'|s, a) V^*(s')$$
$$V^*(s) = V^{\pi^*}(s) \geq V^{\pi}(s)$$

策略迭代

- 适用范围
  - 模型已知
  - 动作空间和状态空间有限
  - 规模较小的问题，收敛相对较快
- 策略迭代过程

$\pi \leftarrow$ 随机初始化策略

`while` 未收敛:

策略评估 $V^{\pi}(s) \leftarrow \sum_a (r(s,\pi(s)) + \gamma \sum_{s'} p(s'|s,a) V^{\pi}(s'))$

策略提升 $\pi'(s) \leftarrow \arg\max_a \{r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^{\pi}(s')\}$

$\pi \leftarrow \pi'$

`return` π

价值迭代

- 适用范围
  - 模型已知
  - 动作空间和状态空间有限
  - 规模较大的问题，计算效率比策略迭代更高
- 价值迭代过程

初始化状态价值函数 $V(s) \leftarrow 0$

`while` 未收敛:

更新价值函数$V'(s) \leftarrow \max_a\{r(s,a) + \gamma \sum_{s'} p(s'|s,a)V(s)\}$

$V \leftarrow V'$

计算最优策略$\pi(s) \leftarrow \arg\max_a\{r(s,a) + \gamma \sum_{s'} p(s'|s,a)V(s')\}$

`return` π

基于时序差分的强化学习算法

- 动作价值函数

$$Q^{\pi}(s,a) = \mathbb{E}[R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \cdots | s_0 = s, a_0 = a, \pi]$$
$$= \mathbb{E}[R(s) + \gamma Q(s_1, a_1) | s_0 = s, a_0 = a, \pi]$$

- 时序差分的动作价值函数更新

$$Q(s,a) = Q(s,a) + \alpha[R(s) + \gamma Q(s',a') - Q(s,a)]$$

- 根据动作价值函数进行策略提升

$$\pi'(s) = \arg\max_a \sum_{s'} p(s'|s,a)V(s')$$
$$= \arg\max_a Q(s,a)$$

SARSA

- 适用范围
  - 模型未知，在线策略
  - 动作空间和状态空间有限
  - 策略相对保守
- SARSA过程

$Q(s,a) \leftarrow$ 随机初始化动作价值函数，终止状态为0

重复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

$a' \leftarrow \epsilon$-greedy策略根据s'和Q选取动作

更新$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)]$

$s \leftarrow s'$

计算最优策略$\pi(s) \leftarrow \arg\max_a Q(s,a)$

return π

Q-Learning

- 适用范围
  - 模型未知，离线策略
  - 动作空间和状态空间有限
  - 策略相对激进，训练需要的样本数量更少
- Q-Learning过程

$Q(s,a) \leftarrow$ 随机初始化动作价值函数，终止状态为0

重复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

更新$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s,a)]$

$s \leftarrow s'$

计算最优策略$\pi(s) \leftarrow \arg\max_a Q(s,a)$

return π

第二次作业：悬崖行走

- 及第科目介绍及提交入口
  - http://www.jidiai.cn/cliffwalking
- 作业本地训练环境、算法代码、训练说明等
  - https://gitee.com/jidiai/summercourse2021/tree/main/course2
  - https://github.com/jidiai/SummerCourse2021/tree/main/course2

作业提示

- 下载习题课仓库summercourse2021-main
- 本地训练SARSA或Q-Learning，产生model文件*.pth
  - 在course2/examples下启动命令行
  - python main.py --scenario cliffwalking --algo sarsa
  - course2/examples/models/cliffwalking/sarsa/run1/trained_model
- 将*.pth复制到course2/examples/algo/homework文件夹
- 实现submission.py
- 在及第上提交submission.py和*.pth

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“悬崖行走”一行：

积分>-90

即成功完成第二次作业