Reinforcement Learning Task2

姚冠宇

2020211260070

先展示最终结果（程序可以直接运行）：

根据课本叙述，智能体执行完动作后到达的位置会受到风的影响，所以先考虑风的影响再执行完动作。（换句话说到达了有风的位置再执行动作才会被风吹）

总共迭代了60000轮

最优路线：

最后一行显示风力

```
['^^', 'vv', '<<', '>>','<^','v>','<v','^>']
```

分别对应上、下、左、右、左上、右下、左下、右上

```
Sarsa算法最终收敛得到的路径为：
oo oo oo oo oo oo oo oo oo oo
oo oo oo oo oo oo oo oo oo oo
oo oo oo oo oo oo oo oo oo oo
v> oo oo oo oo oo oo EE oo oo
oo v> oo oo oo oo oo oo oo oo
oo oo v> oo oo oo oo oo oo oo
oo oo oo v> v> v> ^> oo oo oo
== == == 11 11 11 22 22 11 ==
```

最优路线的步数：

```
Sarsa算法最终收敛得到的步数为： 7
```

最终得到每个状态的动作表格（一个位置可能存在多个策略故采用这种方式打印）：

如果看不清可以放大看或者运行程序也会打印出来这个表格

```
Sarsa算法最终收敛得到的策略为：
ooooo>>ooooooooo ooooooooooo<voo ooVVooooooooooo ooooooooov>oooo ooooooooooooo^> ooooooooooooo^> ooooooooov>oooo ooooooooooooo^> ooooooooooooo^> ooVVooooooooooo
ooooooooov>oooo ooooooooov>oooo ooooooooooo<voo ooooooooooo<voo ooooo>>ooooooooo ooooo>>ooooooooo ooooooooov>oooo ooooo>>ooooooooo ooooooooov>oooo ooooooooov>oooo
ooooooooov>oooo ooooooooov>oooo ooVVooooooooooo ooooooooooo<voo ooooooooov>oooo ooooooooov>oooo ooooooooov>oooo ooooooooov>oooo ooooooooov>oooo ooooooooov>oooo
ooooooooov>oooo ooooooooov>oooo ooooooooov>oooo ooooooooooo<voo ooooooooooo<voo ooooooooov>oooo ooooooooov>oooo EE ooooooooooo<voo ooooooooooo<voo
ooooooooov>oooo ooooooooov>oooo ooooooooov>oooo ooooooooov>oooo ooooooooov>oooo ooooooooov>oooo ooooooooov>oooo ooVVooooooooooo oooo<<ooooooooo ooooooooooo<voo
ooooo>>ooooooooo ooooooooov>oooo ooooooooov>oooo ooooooooov>oooo ooooooooov>oooo ooooooooov>oooo ooooo>>ooooooooo ooooooooov>oooo oooooooo<^ooooooo ooooooooov>oooo
ooooooooooooo^> ooooooooooooo^> ooooooo>>ooooooooo ooooooooov>oooo ooooooooov>oooo ooooooooov>oooo ooooooooooooo^> ^^vv<<>><^v><v^> ooooooooooo<voo oooo<<ooooooooo
=============== =============== =============== 1111111111111111 1111111111111111 1111111111111111 2222222222222222 2222222222222222 1111111111111111 ===============
```

大致讲解一下代码结构，代码里有详细的注释

Wind函数主要实现在特定位置的坐标转移（x在3，4，5，8时y上移一格，在6，7时y上移两格）

CliffWalkingEnv主要是环境，输入智能体的动作给出下一个状态，reward和是否到达终态的done

Sarsa类是智能体，主要维护一个nrow*ncol行，n_action列的qtable
我们初始化学习率alpha=0.1，折扣因子gamma=1，epsilon-贪婪中的epsilon=0.1。每一步选择动作价值最大的动作或随机选择动作。同时定义best action函数用来打印策略。定义update方法实现sarsa的q_table更新。
主函数进行了num_episodes=60000轮次训练，每10条序列打印一下这10条序列的平均回报。

Sarsa算法的更新公式为

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha[r_t + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)]$$

程序中也根据这个公式来更新q_table.

最后定义了两个打印函数print_agent()和print_route()。第一个用于打印智能体在每个位置会选择的动作（即打印策略），第二个用于打印智能体从开始位置出发会走的路线（即打印路径）。