**实验三**

2017202109 刘洋凡

1. 实验目的

利用机器学习中的Q-Learning算法实现CartPole游戏得分

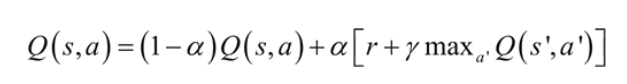
1. 算法思路

有一个由状态 s 描述的环境（s∈S，S 是所有可能状态的集合），一个能够执行动作 a 的 agent（a∈A，A 是所有可能动作的集合），智能体的动作致使智能体从一个状态转移到另外一个状态。智能体的行为会得到奖励，而智能体的目标就是最大化奖励。

在 Q learning 中，智能体计算能够最大化奖励 R 的状态-动作组合，以此学习要采取的动作（策略 π），在选择动作时，智能体不仅要考虑当前的奖励，还要尽量考虑未来的奖励：



智能体从任意初始状态 Q 开始，选择一个动作 a 并得到奖励 r，然后更新状态为 s'（主要受过去的状态 s 和动作 a 的影响），新的 Q 值为：

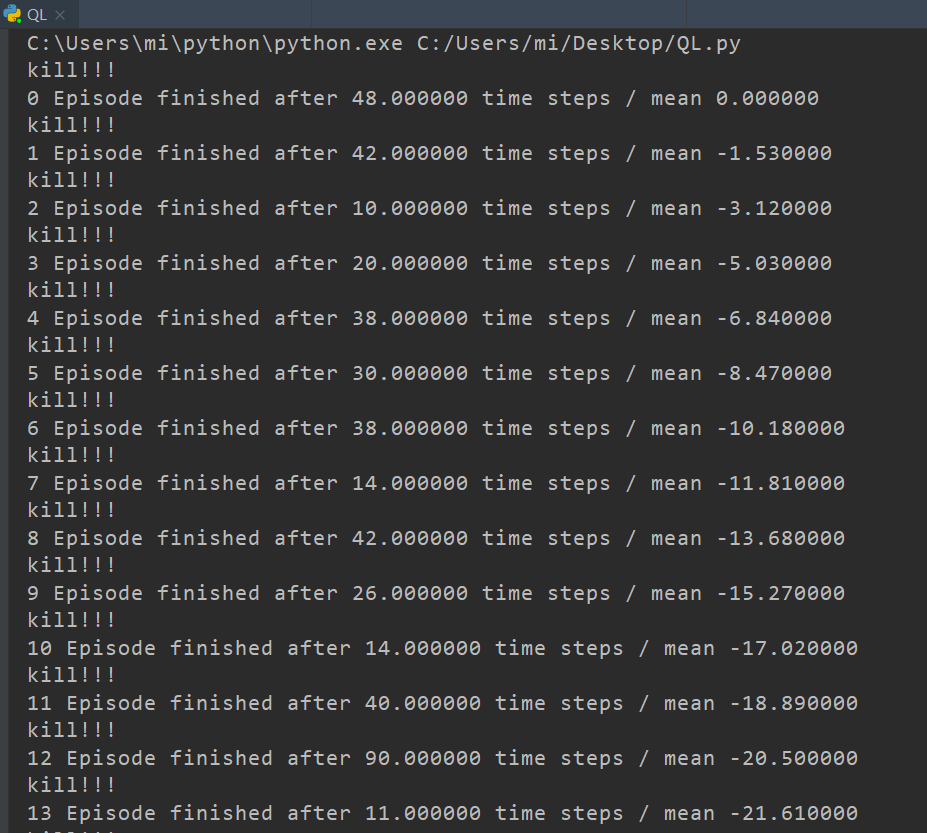


其中，α 是学习率，γ 是折扣因子。第一项保留 Q 的旧值，第二项对 Q 值进行更新估计（包括当前奖励和未来动作的折扣奖励），这会导致在结果状态不满意时降低 Q 值，从而确保智能体在下一次处于此状态时不会选择相同的动作。类似地，当对当前状态满意时，对应的 Q 值将增加。

Q learning 的最简单实现包括维护和更新一个状态-动作值的对应表，表格大小为 N×M，其中 N 是所有可能状态的数量，M 是所有可能动作的数量。对于大多数环境来说，这个表格会相当大，表格越大，搜索所需的时间越长，存储表格所需的内存越多，因此该方案并不可行。

1. 实验结果

程序刚开始运行时，得分比较低，基本都在50以下



上百次训练后得分逐渐稳定在一百分以上

