2023-12-05及之前

多臂老虎机问题

1.一个理解上不自信的地方：强化学习任务的优化目标是 “最大化智能体策略在与动态环境交互的过程中的**价值**”

那么最优策略的数学表达应该是：

最优策略 $= arg\ min\ E_{(状态，动作)\sim系统度量}[奖励函数(状态,动作)]$

这里是照抄书上的公式，但是我觉得定义中“最大化价值”和数学表达中的$arg\ min$似乎由冲突?

2.实现了$\epsilon - Greedy$算法，UCB算法和$ThompsonSampling$算法。说实话代码是直接抄书的，因为感觉自己用面向对象编程实现人工智能算法的思维还不熟练，还是抄一遍更有效率。当然没有不带脑子的抄代码，每一行代码都带着思考的看过了，感觉在实现这些算法的思路和方法上有了更新的理解，上面说到的三种算法的思想内核也掌握了。在实现$算法的时候基本上就可以做到：看一眼书找找感觉，然后就可以自己写出来了。总的来说还是比较有
收获的。

马尔可夫决策过程

1.马尔可夫过程，随机过程选讲里面讲了，居然还记得。

2.马尔可夫奖励过程中的回报函数：站在**当前**时间点看**接下来**一系列动作（到终止之前的所有动作，那要是动作不终止怎么办）的回报，距离当前时间点越远，回报的加权越小。

2023-12-06

3.一个理解上不自信的地方

我对回报函数用途的理解如下：回报函数表示了从当前时间步开始，假设状态接下来根据一个特定的规律变化，那么这个规律的价值可以表示为回报G

似乎想明白了/*不理解的是，回报函数 $G_t = R_t + \gamma R_{t+1} + ...$ 中为什么要有当前时间步的回报，这个时间步下的这个状态和上述规律无关。换句话说：我啥都没干为啥就有回报了，就因为我碰巧运气好在这个状态下吗?

或者说这个函数并不是表达上述规律的价值，而是表示当前状态的价值，比如说对于在海面上随风漂泊的小船，越靠近岸边的状态价值越高*/

4.价值函数

价值：回报的期望

我对**价值**的理解：因为同一个状态有不同的发展可能，所以会有不同的概率取到不同的回报，那么为了更好的评价这个状态的吸引人程度，需要计算从这个状态出发获得回报的期望。

寻找最优状态路径的问题就变成了寻找下一个价值最高的状态的问题（有种动态规划问题的味道，将一个大的复杂问题切分成一个个最小可解决的问题）

价值函数的自变量就是不同的状态，因变量就是对应的期望

价值最大化对应的就是全局激励的期望最大化

5.贝尔曼方程

$$当前价值 = 当前回报 + 折扣 * 转移概率 * 转移后价值$$

6.马尔可夫奖励过程：可以算出每个状态的价值（算出价值函数）

7.超爱书上p24里举的例子，马尔可夫奖励过程是随波逐流的小船，飘到岸边有大的奖励，马尔可夫决策过程是有水手的小船，飘到岸边有较大的奖励

8.状态价值函数：在当前状态下根据策略π执行到结束获得的激励的加权和。

9.动作价值函数：和状态价值函数相比增加了一个条件，这个条件就是在状态条件下**执行特定动作**，也就是说在这个状态下执行这个动作的价值，个人认为其意义是十分简单明了的。

10.MDP变换为MRP的方法：对MDP的奖励根据策略中动作的概率进行加权，得到的奖励就是对应MRP的奖励。

蒙特卡罗方法

有一说一这玩意我大一下上python课的时候就接触过，大致思想就是“随机尝试，然后根据尝试的成功失败次数来估计一个值”，希望真的有这么简单

1.啊真就那么简单，使用策略π采样个几条序列，然后记录状态出现次数和状态对应回报，具体的计算方法就是：次数就计数就行，回报就记录每次出现后积累的奖励就行，奖励就和MDP以及MRP中出现的一样算

2023-12-7

2.好吧在代码实现蒙特卡洛方法的时候出了一点问题，思想应该没有问题，当时代码实现的过程中有点小问题，暂且不管

3.发现问题了，P和R长得太像以至于弄混了，哈哈，他妈的，一个上午就这么没了

占用度量

1.状态访问分布：该策略下该状态在所有时间步被访问到的概率的和，需要乘上$1-\gamma$来保证概率和为1

2.占用度量：表示**状态动作对**被访问到的概率

3.状态访问分布和占用度量之间的关系：后者是前者乘该策略下该状态条件下出现该动作的概率

4.rho是啥