2025-2026-1 学期强化学习课程 - 第一次作业

1120231863　左逸龙

October, 15 2025

1 马尔可夫决策过程

该 MDP 由五元组 $\mathcal{M}=(\mathcal{S},\mathcal{A},\mathcal{P},\mathcal{R},\gamma)$ 组成，其中:

- 状态集 $\mathcal{S}$: 本题的核心状态是**当前时刻所拥有的钱的数量**，因此状态集 $\mathcal{S}=\{0,10,20,30,40\}$。
- 动作集 $\mathcal{A}$: 本题当中，在游戏未结束前，我们有两种动作，即**玩老虎机 A** 和**玩老虎机 B**。不过需要注意的是，在不同状态之下，可以采取的动作有所不同，具体而言动作集如下:

$$\mathcal{A}(s)=\begin{cases}\emptyset \text{ if } s=0\vee s=40\\ \{A\} \text{ if } s=10\\ \{A,B\} \text{ if } s=20\vee s=30\end{cases}$$

- 状态转移概率 $\mathcal{P}$: 可以由一 $5\times5\times2$ 的矩阵表示，其中第一维表示起始状态s_t，第二维表示转移状态s_{t+1}，第三维表示动作a_t。

  依据题意，玩老虎机 A，即采取动作 A，有 0.05 的概率净赚 10 元，有 0.95 的概率输掉 10 元，因此:

$$P(s_t,s_{t+1},a_t=A)=\begin{cases}0.05 \text{ if } s_t\in\{10,20,30\}\wedge s_{t+1}=s_t+10\\ 0.95 \text{ if } s_t\in\{10,20,30\}\wedge s_{t+1}=s_t-10\\ 0 \text{ otherwise}\end{cases}$$

  同理，玩老虎机 B，即采取动作 B，有:

$$P(s_t,s_{t+1},a_t=B)=\begin{cases}0.01 \text{ if } s_t\in\{20,30\}\wedge s_{t+1}=s_t+10\\ 0.99 \text{ if } s_t\in\{20,30\}\wedge s_{t+1}=s_t-20\\ 0 \text{ otherwise}\end{cases}$$

- 奖励函数 $\mathcal{R}$: 可以由一个 5×2 的矩阵表示，其中第一维表示起始状态s_t，第二维表示动作a_t。根据期望公式可得:

$$R(s_t,a_t)=\begin{cases}-10+20\times0.05=-9 \text{ if } a_t=A\\ -20+30\times0.01=-19.7 \text{ if } a_t=B\end{cases}$$

- 折扣因子 γ: 依据题意，本题折扣因子 $\gamma=1$。

2 Gridworld 小游戏

(a) 最短路径策略

$r_s = -1$，以下阐述原因：

首先，累计回报$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$，其中$\gamma$为折扣因子，本题中$\gamma = 1$，$R_{t+k+1}$为状态$s_{t+k+1}$的奖励，有：

$$R(s) = \begin{cases} r_g \text{ if } s = 3 \\ r_r \text{ if } s = 14 \\ r_s \text{ otherwise} \end{cases}$$

假设智能体 Agent 在时刻 T 到达终点 3，则有：

$$G_t = r_g + r_s \times (T - t)$$

Agent 的优化目标是最大化累计回报 G_t，当 $r_s = -1$ 时，Agent 必须最小化 $T - t$，即最小化路径长度，使得最优策略可以返回到格子 3 的最短路径；当 $r_s = 0$ 时，Agent 可以不考虑路径长度，因此最优策略可以为任意路径；当 $r_s = 1$ 时，Agent 将尽可能最大化路径长度，与最短路径背道而驰。综上，$r_s = -1$。

在此条件下，每个格子的最优价值如下表所示：

1	**2**	**3**	**−4**
0	**1**	**2**	**−3**
−1	**0**	**1**	**−2**
−2	**−3**	**0**	**−1**

该表可通过求解贝尔曼方程得到，即：

$$V(s) = \max_{a \in \mathcal{A}(s)} R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a)V(s')$$

(b) 奖励变化的影响

在新网格世界当中新的价值函数如下表所示:

8	**7**	**6**	**13**
9	**8**	**7**	**12**
10	**9**	**8**	**11**
11	**−6**	**9**	**10**

(c) 奖励变化的一般表达式

1. $V^{\pi}_{\text{new}} = c \times V^{\pi}_{\text{old}}$。这是因为在无限长的马尔可夫链中，$V^{\pi} = (I - \gamma P^{\pi})^{-1} R^{\pi}$，当奖励变为 $R'^{\pi} = cR^{\pi}$ 时，价值函数也等比例变化。

2. 当 $c < 0$ 时，会使得最优策略发生变化，理由如下:

   首先让我们考虑原始的 MDP，令其最优策略为 π_*，由此引发的最优价值函数为 $V^{\pi_*}_{\text{old}}$。最优策略 π_* 优于任意其他策略 π，因此满足如下关系:

   $$V^{\pi_*}_{\text{old}}(s) \geq V^{\pi}_{\text{old}}(s),\ \forall s \in \mathcal{S}$$

   现在让我们考虑新的 MDP，由于 $V^{\pi}_{\text{new}} = c \times V^{\pi}_{\text{old}}$，因此:

   - 当 $c \geq 0$ 时，$c \times V^{\pi_*}_{\text{old}}(s) \geq c \times V^{\pi}_{\text{old}}(s) \Longrightarrow V^{\pi_*}_{\text{new}}(s) \geq V^{\pi}_{\text{new}}(s)$，因此 $\pi_* \geq \pi$，即原策略仍然是新的 MDP 的最优策略。

   - 而当 $c < 0$ 时，$c \times V^{\pi_*}_{\text{old}}(s) \leq c \times V^{\pi}_{\text{old}}(s) \Longrightarrow V^{\pi_*}_{\text{new}}(s) \leq V^{\pi}_{\text{new}}(s) \Longrightarrow V^{\pi_*}_{\text{new}}(s) \geq V^{\pi}_{\text{new}}(s)$不恒成立，因此 $\pi_* \geq \pi$不恒成立，即原策略不一定是新的 MDP 的最优策略，此时最优策略大概率会发生变化，除非所有策略的价值都相同，而这是不太可能的。

   综上，当 $c < 0$ 时，最优策略会发生变化。

(d) 正奖励的影响

类似于 (a) 小节当中针对 $r_s = 1$ 的分析，此时的最优策略是尽可能不踏入终点，即绿色的 3 与红色的 14，从而最大化路径长度。考虑到格子图当中从非阴影格子出发时，总能够找到一条路径返回起点，因此 Agent 通过进入这一循环路径，可以使得格子的价值无限次累加正奖励 $r_s = 2$，进而把所有非阴影格子的价值变为正无穷。