人智原hw5

第2题 蒙特卡洛 (Monte Carlo)

(1)： 状态价值贝尔曼期望方程及求解

已知:

- 状态：A, B, Terminate
- 转移概率：$P_{AA}=1/4$, $P_{AB}=3/4$, $P_{BA}=1/2$, $P_{BT}=1/2$
- 状态奖励 (离开状态时获得)：$r_A=3$, $r_B=-3$
- 折扣因子：$\gamma=1$
- 终止状态价值：$V(\text{Terminate})=0$

贝尔曼期望方程为： $v(S)=r_S+\gamma\sum_{S'}P(S'|S)v(S')$

可得:

1. $v(A)=r_A+\gamma(P_{AA}v(A)+P_{AB}v(B))$ $v(A)=3+1\cdot(\frac{1}{4}v(A)+\frac{3}{4}v(B))$
2. $v(B)=r_B+\gamma(P_{BA}v(A)+P_{BT}v(\text{Terminate}))$ $v(B)=-3+1\cdot(\frac{1}{2}v(A)+\frac{1}{2}\cdot 0)$
   $v(B)=-3+\frac{1}{2}v(A)$

将 (2) 代入 (1)： $v(A)=3+\frac{1}{4}v(A)+\frac{3}{4}(-3+\frac{1}{2}v(A))$ $v(A)=3+\frac{1}{4}v(A)-\frac{9}{4}+\frac{3}{8}v(A)$
$v(A)-\frac{1}{4}v(A)-\frac{3}{8}v(A)=3-\frac{9}{4}$ $\frac{8-2-3}{8}v(A)=\frac{12-9}{4}$ $\frac{3}{8}v(A)=\frac{3}{4}$ $v(A)=\frac{3}{4}\cdot\frac{8}{3}=2$

将 $v(A)=2$ 代入 (2)： $v(B)=-3+\frac{1}{2}(2)=-3+1=-2$

因此，状态价值为 $v(A)=2$，$v(B)=-2$。

---

(2)： 蒙特卡洛预测

- 序列 1: A $\xrightarrow{+3}$ A $\xrightarrow{+2}$ B $\xrightarrow{-4}$ A $\xrightarrow{+4}$ B $\xrightarrow{-3}$ Terminate
  - 奖励: $R_1=3, R_2=2, R_3=-4, R_4=4, R_5=-3$
- 序列 2: B $\xrightarrow{-2}$ A $\xrightarrow{+3}$ B $\xrightarrow{-3}$ Terminate
  - 奖励: $R_1=-2, R_2=3, R_3=-3$
- 折扣因子 $\gamma=1$ (无折扣)

首次访问蒙特卡洛 (First-Visit Monte Carlo)

计算每个片段中，每个状态首次被访问后得到的累积回报 G_t。

- **片段 1:** $S_0(A)\xrightarrow{+3}S_1(A)\xrightarrow{+2}S_2(B)\xrightarrow{-4}S_3(A)\xrightarrow{+4}S_4(B)\xrightarrow{-3}S_5(\text{Terminate})$
  - 状态 A 首次出现在 $t=0$。回报 $G_0=3+2-4+4-3=2$ 。
  - 状态 B 首次出现在 $t=2$。回报 $G_2=-4+4-3=-3$ 。

- **片段 2:** $S_0(B) \xrightarrow{-2} S_1(A) \xrightarrow{+3} S_2(B) \xrightarrow{-3} S_3(\text{Terminate})$
  - 状态 B 首次出现在 $t=0$。回报 $G_0 = -2+3-3 = -2$ 。
  - 状态 A 首次出现在 $t=1$。回报 $G_1 = 3-3 = 0$ 。

汇总回报：

- 状态 A 的回报列表: $[2, 0]$ $v(A) = \frac{2+0}{2} = 1$
- 状态 B 的回报列表: $[-3, -2]$ $v(B) = \frac{-3+(-2)}{2} = \frac{-5}{2} = -2.5$

首次访问蒙特卡洛预测结果：$v(A) = 1$，$v(B) = -2.5$。

每次访问蒙特卡洛 (Every-Visit Monte Carlo)

计算每个片段中，每次访问某状态后得到的累积回报 G_t。

- **片段 1:** $S_0(A) \xrightarrow{+3} S_1(A) \xrightarrow{+2} S_2(B) \xrightarrow{-4} S_3(A) \xrightarrow{+4} S_4(B) \xrightarrow{-3} S_5(\text{Terminate})$
  - 访问 A ($t=0$): $G_0 = 3+2-4+4-3 = 2$.
  - 访问 A ($t=1$): $G_1 = 2-4+4-3 = -1$.
  - 访问 B ($t=2$): $G_2 = -4+4-3 = -3$.
  - 访问 A ($t=3$): $G_3 = 4-3 = 1$.
  - 访问 B ($t=4$): $G_4 = -3$.
- **片段 2:** $S_0(B) \xrightarrow{-2} S_1(A) \xrightarrow{+3} S_2(B) \xrightarrow{-3} S_3(\text{Terminate})$
  - 访问 B ($t=0$): $G_0 = -2+3-3 = -2$.
  - 访问 A ($t=1$): $G_1 = 3-3 = 0$.
  - 访问 B ($t=2$): $G_2 = -3$.

汇总回报：

- 状态 A 的回报列表: $[2, -1, 1, 0]$ $v(A) = \frac{2-1+1+0}{4} = \frac{2}{4} = 0.5$
- 状态 B 的回报列表: $[-3, -3, -2, -3]$ $v(B) = \frac{-3-3-2-3}{4} = \frac{-11}{4} = -2.75$

每次访问蒙特卡洛预测结果：$v(A) = 0.5$，$v(B) = -2.75$。

第3题 时序差分 (Temporal Difference)

(1)： TD(0) V 值更新

- 初始 V 值：所有非终止状态 $V(s) = 0$。
- 片段：$4 \to 1 \to 4 \to 7 \to$ 终止
- 参数：$\alpha = 0.5, \gamma = 1$
- TD(0) 更新规则：$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$

逐步更新： 初始 $V(1) = \ldots = V(7) = 0$ 。

1. **转移** $4 \to 1$: ($S_t = 4, S_{t+1} = 1, R_{t+1} = -1$) $V(4) \leftarrow V(4) + 0.5[-1 + 1 \cdot V(1) - V(4)]$
   $V(4) \leftarrow 0 + 0.5[-1 + 1 \cdot 0 - 0] = -0.5$. 当前 $V(4) = -0.5$.
2. **转移** $1 \to 4$: ($S_t = 1, S_{t+1} = 4, R_{t+1} = -1$) $V(1) \leftarrow V(1) + 0.5[-1 + 1 \cdot V(4) - V(1)]$
   $V(1) \leftarrow 0 + 0.5[-1 + 1 \cdot (-0.5) - 0] = 0.5 \cdot (-1.5) = -0.75$. 当前

$V(1) = -0.75, V(4) = -0.5$.

3. **转移** $4 \to 7$: ($S_t = 4, S_{t+1} = 7, R_{t+1} = -1$) $V(4) \leftarrow V(4) + 0.5[-1 + 1 \cdot V(7) - V(4)]$
   $V(4) \leftarrow -0.5 + 0.5[-1 + 1 \cdot 0 - (-0.5)] = -0.5 + 0.5[-0.5] = -0.5 - 0.25 = -0.75$. 当前 $V(1) = -0.75, V(4) = -0.75$.

4. **转移** $7 \to$ 终止: ($S_t = 7, S_{t+1} =$ 终止, $R_{t+1} = -1$) $V(7) \leftarrow V(7) + 0.5[-1 + 1 \cdot V(\text{终止}) - V(7)]$
   $V(7) \leftarrow 0 + 0.5[-1 + 1 \cdot 0 - 0] = -0.5$. 当前 $V(1) = -0.75, V(4) = -0.75, V(7) = -0.5$.

该片段结束后的 V 值:

- $V(1) = -0.75$
- $V(2) = 0$
- $V(3) = 0$
- $V(4) = -0.75$
- $V(5) = 0$
- $V(6) = 0$
- $V(7) = -0.5$

---

(2)： SARSA Q 值更新

- 初始状态：4
- 参数：$\alpha = 1, \gamma = 1$
- SARSA 更新规则：$Q(S, A) \leftarrow Q(S, A) + \alpha[R + \gamma Q(S', A') - Q(S, A)]$ 因为 $\alpha = 1, \gamma = 1$ ，简化为 $Q(S, A) \leftarrow R + Q(S', A')$
- 策略：确定性贪心策略
- 初始 Q 表 (行为：上U, 右R, 下D, 左L；状态：1-7)：

动作	S=1	S=2	S=3	S=4	S=5	S=6	S=7
上 (U)	-4	-3	-1	-3	-4	-2	-4
右 (R)	-3	-3	-2	-4	-2	-3	-3
下 (D)	-4	-3	-4	-2	-2	-3	-4
左 (L)	-3	-2	-3	-3	-4	-3	-2

逐步更新 (所有转移奖励 $R = -1$)：

1. **当前** $S = 4$**。**
   - 根据 Q 表为 $S = 4$ 选择贪心动作 A： $Q(4, U) = -3$, $Q(4, R) = -4$, $Q(4, D) = -2$, $Q(4, L) = -3$。最大值为 $Q(4, D) = -2$。所以 $A =$ 下 (D)。
   - 执行动作 $A =$ 下。$S = 4 \to S' = 7$。奖励 $R = -1$。
   - 根据 Q 表为 $S' = 7$ 选择贪心动作 A'： $Q(7, U) = -4$, $Q(7, R) = -3$ (撞墙), $Q(7, D) = -4$ (至终止), $Q(7, L) = -2$。最大值为 $Q(7, L) = -2$。所以 $A' =$ 左 (L)。
   - 更新 $Q(S, A) = Q(4,$ 下$)$:
     $Q(4,$ 下$) \leftarrow R + Q(S', A') = -1 + Q(7,$ 左$) = -1 + (-2) = -3$ 。 ($Q(4, D)$ 从 -2 变为 -3)

2. **当前** $S = 7$**,** $A =$ 左 (L) **(上一步选择的** A'**)。**

- 执行动作 $A = \text{左}$。$S = 7 \to S' = 6$。奖励 $R = -1$。
- 根据 Q 表为 $S' = 6$ 选择贪心动作 A'： $Q(6, U) = -2$, $Q(6, R) = -3$, $Q(6, D) = -3$ (撞墙), $Q(6, L) = -4$ (撞墙)。最大值为 $Q(6, U) = -2$。所以 $A' = \text{上}\ (\mathrm{U})$。
- 更新 $Q(S, A) = Q(7, \text{左})$：
  $Q(7, \text{左}) \leftarrow R + Q(S', A') = -1 + Q(6, \text{上}) = -1 + (-2) = -3$ 。 ($Q(7, L)$ 从 -2 变为 -3)

3. **当前** $S = 6$, $A = \text{上}\ (\mathrm{U})$。

- 执行动作 $A = \text{上}$。$S = 6 \to S' = 3$。奖励 $R = -1$。
- 根据 Q 表为 $S' = 3$ 选择贪心动作 A'： $Q(3, U) = -1$ (至终止), $Q(3, R) = -2$, $Q(3, D) = -4$, $Q(3, L) = -3$ (撞墙)。最大值为 $Q(3, U) = -1$。所以 $A' = \text{上}\ (\mathrm{U})$。
- 更新 $Q(S, A) = Q(6, \text{上})$：
  $Q(6, \text{上}) \leftarrow R + Q(S', A') = -1 + Q(3, \text{上}) = -1 + (-1) = -2$ 。 ($Q(6, U)$ 保持 -2, 未变)

4. **当前** $S = 3$, $A = \text{上}\ (\mathrm{U})$。

- 执行动作 $A = \text{上}$。$S = 3 \to S' = \text{终止}$。奖励 $R = -1$。
- S' 是终止状态，$Q(S', A') = Q(\text{终止}, \text{任意动作}) = 0$。
- 更新 $Q(S, A) = Q(3, \text{上})$： $Q(3, \text{上}) \leftarrow R + Q(S', A') = -1 + 0 = -1$ 。 ($Q(3, U)$ 保持 -1, 未变)

片段结束。

一个 episode 结束后更新的 Q 表： $Q(4, \text{下})$ 从 -2 更新为 -3。 $Q(7, \text{左})$ 从 -2 更新为 -3。 其他值不变。

动作	S=1	S=2	S=3	S=4	S=5	S=6	S=7
上 (U)	-4	-3	-1	-3	-4	-2	-4
右 (R)	-3	-3	-2	-4	-2	-3	-3
下 (D)	-4	-3	-4	**-3**	-2	-3	-4
左 (L)	-3	-2	-3	-3	-4	-3	**-3**