上海科技大学

2024-2025 强化学习应用实践

Project1 Part-A

姓名: 周守琛
学号: 2021533042

2025 年 3 月 17 日

- **状态集合** S: $S = \{s_1, s_2, s_3\}$
- **动作集合** A: $A = \{a_1, a_2\}$
- **终止状态**: s_3 为终止状态，不再执行动作。
- **转移概率** $P(s'|s,a)$:
  - 从 s_1:
  $$P(s_2|s_1,a_1) = 0.5, \quad P(s_3|s_1,a_1) = 0.5$$
  $$P(s_2|s_1,a_2) = 0.7, \quad P(s_3|s_1,a_2) = 0.3$$
  - 从 s_2:
  $$P(s_1|s_2,a_1) = 0.6, \quad P(s_3|s_2,a_1) = 0.4$$
  $$P(s_1|s_2,a_2) = 0.8, \quad P(s_3|s_2,a_2) = 0.2$$
  - 从 s_3: 无后续转移（所有 $P(s'|s_3,a) = 0$）。
- **奖励函数** $R(s,a)$:
$$R(s_1,a_1) = 1, \quad R(s_1,a_2) = 2$$
$$R(s_2,a_1) = 3, \quad R(s_2,a_2) = 0$$
$$R(s_3,a) = 0 \quad \forall a$$
- **折扣因子** γ: 0.9
- **初始策略** $\pi(s,a) = \frac{1}{|A|}$: 所有非终止状态的动作选择均匀随机：
$$\pi(s_1,a_1) = \pi(s_1,a_2) = 0.5, \quad \text{其余同理。}$$
- **初始值函数/状态-动作值函数**:
$$V(s) = 0 \quad \forall s, \quad Q(s,a) = 0 \quad \forall s,a$$
- **学习率** $\alpha = 0.1$。

1. 策略迭代（20 分） 根据初始策略 $\pi(s,a)=\dfrac{1}{2}$，手动推导策略迭代第一轮的迭代步骤，包括策略评估和策略改进，写出：

1. 第一轮策略评估的状态值函数 $V(s)$。(10 分)
2. 改进后的新策略 $\pi'(s)$。(10 分)

Solution

We can get the value function's update rule:

$$V(s)=\sum_a \pi(s,a)\left[R(s,a)+\gamma\sum_{s'}P(s'|s,a)V(s')\right]$$

1. To get the value function $V(s)$ after the first policy evaluation, apply the formula to the initial policy, we have

$$\begin{aligned}
V(s_1) &\leftarrow \underbrace{\frac{1}{2}\left[1+0.9\times(0.5\times 0+0.5\times 0)\right]}_{a=a_1}+\underbrace{\frac{1}{2}\left[2+0.9\times(0.7\times 0+0.3\times 0)\right]}_{a=a_2}=1.5\\
V(s_2) &\leftarrow \underbrace{\frac{1}{2}\left[3+0.9\times(0.6\times 0+0.4\times 0)\right]}_{a=a_1}+\underbrace{\frac{1}{2}\left[0+0.9\times(0.8\times 0+0.2\times 0)\right]}_{a=a_2}=1.5\\
V(s_3) &\leftarrow 0
\end{aligned}$$

2. To update the policy, we have the update rule

$$\pi(s)=\arg\max_a Q(s,a)=\arg\max_a\sum_{s'}P(s'|s,a)\left[R(s,a)+\gamma V(s')\right]=\arg\max_a R(s,a)+\gamma\sum_{s'}P(s'|s,a)V(s')$$

- $s=s_1,a=a_1$: $\sum_{s'}P(s'|s_1,a_1)V(s')=1+0.9\times\left[\underbrace{0.5\times 1.5}_{s'=s_2}+\underbrace{0.5\times 0}_{s'=s_3}\right]=1.675$
- $s=s_1,a=a_2$: $\sum_{s'}P(s'|s_1,a_2)V(s')=2+0.9\times\left[\underbrace{0.7\times 1.5}_{s'=s_1}+\underbrace{0.3\times 0}_{s'=s_3}\right]=2.945$
- $s=s_2,a=a_1$: $\sum_{s'}P(s'|s_2,a_1)V(s')=3+0.9\times\left[\underbrace{0.6\times 1.5}_{s'=s_1}+\underbrace{0.4\times 0}_{s'=s_3}\right]=3.81$
- $s=s_2,a=a_2$: $\sum_{s'}P(s'|s_2,a_2)V(s')=0+0.9\times\left[\underbrace{0.8\times 1.5}_{s'=s_1}+\underbrace{0.2\times 0}_{s'=s_3}\right]=1.08$

So the updated policy after the first policy improvement is

$$\pi'(s_1)=a_2,\quad \pi'(s_2)=a_1$$

Since s_3 is the terminal state, so it has no policy.

2. 价值迭代 (20 分) 根据给定的环境和初始值函数 $V(s)=0\forall s$，推导价值迭代第一轮的迭代步骤，写出：

1. 每个状态的值函数 $V(s)$。(10 分)
2. 相应的策略 $\pi(s)$。(10 分)

Solution

We can get the value function's update rule:

$$V(s)=\max_a\left(R(s,a)+\gamma\sum_{s'}P(s'|s,a)V(s')\right)$$

1. To get the value function $V(s)$ after the first policy evaluation, apply the formula to the initial policy, we have

$$V(s_1)\leftarrow\max\left(\underbrace{1+0.9\times(0.5\times 0+0.5\times 0)}_{a=a_1},\underbrace{2+0.9\times(0.7\times 0+0.3\times 0)}_{a=a_2}\right)=\max(1,2)=2$$

$$V(s_2)\leftarrow\max\left(\underbrace{3+0.9\times(0.6\times 0+0.4\times 0)}_{a=a_1},\underbrace{0+0.9\times(0.8\times 0+0.2\times 0)}_{a=a_2}\right)=\max(3,0)=3$$

$$V(s_3)\leftarrow 0$$

2. The update policy is

$$\pi(s)=\arg\max_a V(s)$$

So the updated policy after the first policy improvement is

$$\pi(s_1)=a_2,\quad \pi(s_2)=a_1$$