清华大学实验报告

系别 [redacted] 班号 [redacted] 姓名 [redacted] (同组姓名:_____)

作实验日期 2021年 12 月 18 日 [redacted] 教师评定:

2021.12.18

第8次作业.

2. (1) 首先计算更新后的状态价值函数 $V_2(s)$.

$V_2(A) = r_s^a + \gamma \sum_{s' \in S} p_{ss'}^a V_1(s') = -8 + 0.5 \times 1.0 \times V_1(B) = -8 + 0.5 \times 1.0 \times 2 = -7$

$V_2(B) = \max\{2 + 0.5 \times 1.0 \times V_1(A), -2 + 0.5 \times 1.0 \times V_1(C)\} = \max\{3, -1\} = 3$

$V_2(C) = \max\{8 + 0.5 \times 1.0 \times V_1(B), 0 \times 0.75 + 4 \times 0.25 + 0.5 \times 0.25 \times V_1(A) + 0.5 \times 0.75 \times V_1(C)\} = \max\{9, 2\} = 9.$

则 $V_2(S) = (-7, 3, 9)^T$

由确定性贪心策略: $\pi_2(a = ab \mid S = A) = 1$,

$\pi_2(a = ba \mid S = B) = 1$, $\pi_2(a = bc \mid S = B) = 0$

$\pi_2(a = cb \mid S = C) = 1$, $\pi_2(a = ca \mid S = C) = 0$.

(2) 同样地, 计算 $V_2'(s)$

$V_2'(A) = V_2(A) = -7$

$V_2'(B) = \max\{2 + 0.5 \times 1.0 \times V_2'(A), -2 + 0.5 \times 1.0 \times V_1(C)\} = \max\{-1.5, -1\} = -1$

$V_2'(C) = \max\{8 + 0.5 \times 1.0 \times V_2'(B), 0 \times 0.75 + 4 \times 0.25 + 0.5 \times 0.25 \times V_2'(A) + 0.5 \times 0.75 \times V_1(C)\} = \max\{7.5, 0.875\} = 7.5$

则 $V_2(S) = (-7, -1, 7.5)^T$

由确定性贪心策略: $\pi_2'(a = ab \mid S = A) = 1$.

$\pi_2'(a = ba \mid S = B) = 0$, $\pi_2(a = bc \mid S = B) = 1$

$\pi_2'(a = cb \mid S = C) = 1$. $\pi_2(a = ca \mid S = C) = 0$.

4. (1) $V(S_t) \leftarrow V(S_t) + \alpha(R_{t+1} + \gamma V(S_{t+1}) - V(S_t))$

$4 \to 5$ $V'(4) = V(4) + 0.5(-1 + 1 \times V(5) - V(4)) = 0 + 0.5 \times (-1) = -0.5$

$5 \to 4$ $V'(5) = V(5) + 0.5(-1 + 1 \times V'(4) - V(5)) = 0 + 0.5 \times (-1 + 1 \times -0.5) = -0.75$

$4 \to 3$ $V''(4) = V'(4) + 0.5(-1 + 1 \times V(3) - V'(4)) = -0.5 + 0.5(-1 + 0.5) = -0.75$

$3 \to$ terminate $V'(3) = V(3) + 0.5(-1 + 1 \times 0 - V(3)) = -0.5$

更新后的 V 值.

0	0	0
-0.5	-0.75	-0.75
0	0	0

(2)

	1	2	3	4	5	6	7
上	-4	-3	-1	-3	-4	-2	-4
右	-3	-3	-2	-4	-2	-3	-3
下	-4	-3	-4	-2	-2	-3	-4
左	-3	-2	-3	-3	-4	-3	-2

SARSA算法：

① 4 向下→ 7 向左→

$Q(4,下) = Q(4,下) + \alpha[-1 + 1 \times Q(7,左) - Q(4,下)]$

$= -2 + 1 \times [-1 + 1 \times (-2) - (-2)] = -3$

② 7 向左→ 6 向上→

$Q(7,左) = Q(7,左) + 1 \times [-1 + 1 \times Q(6,上) - Q(7,左)]$

$= -2 + 1 \times [-1 + 1 \times (-2) - (-2)] = -3$

③ 6 向上→ 3 向上→

$Q(6,上) = Q(6,上) + 1 \times [-1 + 1 \times Q(3,上) - Q(6,上)]$

$= -2 + 1 \times [-1 + 1 \times (-1) - (-2)] = -2$

④ 3 向上→ 终止

$Q(3,上) = Q(3,上) + 1 \times [-1 + 1 \times Q(terminate, a) - Q(3,上)]$

$= -1 + 1 \times [-1 + 1 \times 0 - (-1)] = -1$

故更新后的Q表为

	1	2	3	4	5	6	7
上	-4	-3	-1	-3	-4	-2	-4
右	-3	-3	-2	-4	-2	-3	-3
下	-4	-3	-4	-3	-2	-3	-4
左	-3	-2	-3	-3	-4	-3	-3

Q-learning算法：

由于两种算法均使用贪心方式选择动作，故得到的Q表应完全相同，更新后的Q表如上表所示。