人工智能基础第八次作业

3.

(1) 首次访问的蒙特卡洛预测:

$V(A)=\frac{1}{2}\times[(+3+2-4+4-3)+(+3-3)]=1$

$V(B)=\frac{1}{2}\times[(-4+4-3)+(-2+3-3)]=-2.5$

每次访问的蒙特卡洛预测:

$V(A)=\frac{1}{4}\times[(+3+2-4+4-3)+(+2-4+4-3)+(+4-3)+(+3-3)]=0.5$

$V(B)=\frac{1}{4}\times[(-4+4-3)+(-3)+(-2+3-3)+(-3)]=-2.75$

(2) 共8次状态转移过程

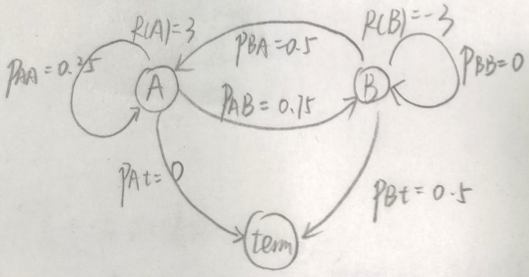
其中 A→A 1次 A→B 3次 B→A 2次 B→terminate 2次

故 $P_{AA}=\frac{1}{1+3}=0.25$, $P_{AB}=\frac{3}{1+3}=0.75$ $P_{At}=0$

$P_{BA}=\frac{2}{2+2}=0.5$, $P_{Bt}=\frac{2}{2+2}=0.5$ $P_{BB}=0$

$R(A)=\frac{1}{4}\times(3+2+4+3)=3$

$R(B)=\frac{1}{4}\times(-4-3-2-3)=-3$



(3) 状态价值的贝尔曼期望方程 $V_\pi(S)=\sum_{a\in A}\pi(a|s)\left(r_s^a+\gamma\sum_{s'\in S}P_{ss'}^a V_\pi(s')\right)$

其中 S=A, B, term

$$\begin{bmatrix}V(A)\\V(B)\\V(t)\end{bmatrix}=\begin{bmatrix}R(A)\\R(B)\\R(t)\end{bmatrix}+1\times\begin{bmatrix}P_{AA}&P_{AB}&P_{At}\\P_{BA}&P_{BB}&P_{Bt}\\P_{tA}&P_{tB}&P_{tt}\end{bmatrix}\begin{bmatrix}V(A)\\V(B)\\V(t)\end{bmatrix}$$

$$\begin{bmatrix} V(A) \\ V(B) \\ V(T) \end{bmatrix} = \begin{bmatrix} 3 \\ -3 \\ 0 \end{bmatrix} + 1 \times \begin{bmatrix} 0.25 & 0.75 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} V(A) \\ V(B) \\ V(T) \end{bmatrix}$$

得 $\begin{cases} V(A) = 2 \\ V(B) = -2 \end{cases}$

4.

1) 4→5:

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

$= 0 + 0.5 \times (-1 + 0 - 0) = -0.5$

5→4: $V(5) = V(5) + 0.5 \times (-1 + 1 \times V(4) - V(5)) = 0 + 0.5 \times (-1 - 0.5 - 0) = -0.75$

4→3: $V(4) = V(4) + 0.5 \times (-1 + 1 \times V(3) - V(4)) = -0.5 + 0.5 \times (-1 + 0 + 0.5) = -0.75$

3→T: $V(3) = V(3) + 0.5 \times (-1 + 1 \times V(T) - V(3)) = 0 + 0.5 \times (-1 + 0 - 0) = -0.5$

更新后V值为

0	0	0
-0.5	-0.75	-0.25
0	0	0

2)

① SARSA算法：

状态4：向下运动 $Q(4, \downarrow) = Q(4, \downarrow) + 1 \times (-1 + 1 \times Q(7, \leftarrow) - Q(4, \downarrow))$

（下一状态：7，向左） $= -2 + 1 \times (-1 + 1 \times (-2) - (-2)) = -3$

状态7：向左运动 $Q(7, \leftarrow) = Q(7, \leftarrow) + 1 \times (-1 + 1 \times Q(6, \uparrow) - Q(7, \leftarrow))$

（下一状态：6，向上） $= -2 + 1 \times (-1 + 1 \times (-2) - (-2)) = -3$

状态6：向上运动 $Q(6, \uparrow) = Q(6, \uparrow) + 1 \times (-1 + 1 \times Q(3, \uparrow) - Q(6, \uparrow))$

（下一状态：3，向上） $= -2 + 1 \times (-1 + 1 \times (-1) - (-2)) = -2$

状态3：向上运动 $Q(3,\uparrow) = Q(3,\uparrow) + 1 \times (-1 + 1 \times Q(1,a) + Q(3,\uparrow))$

（下状态估计）

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

$= -1$

更新后的Q表

-4	-3	-1	-3	-4	-2	-4
-3	-3	-2	-4	-2	-3	-3
-4	-3	-4	-2	-2	-3	-4
-3	-2	-3	-3	-4	-3	-3

② Q-learning算法：

因为也采用贪心方式选择动作，故更新过程和更新后的Q表与SARSA算法相同