第 7 次作业

2022 年 12 月 10 日

题 2

甲乙两人进行比赛, 设每局比赛甲胜的概率是 p, 乙胜的概率是 q, 平局的概率是 r, 有 $p+q+r=1$, 设每局比赛后, 胜者记 +1 , 负者记 -1, 平局不记分。当两人中一人积到 2 分时, 比赛结束。用 x_n 表示比至第 n 局结束, 甲获得的分数, 则序列 $\{X_1, X_2, \cdots\}$ 为一个马尔可夫过程。

(1) 请给出状态空间和状态转移矩阵。

状态空间 $S=\{s_0, s_1, s_2, s_3, s_4\}$，其中 s_0 代表第 n 局结束，甲获得 2 分；s_1 代表第 n 局结束，甲获得 1 分；s_2 代表第 n 局结束，甲获得 0 分；s_3 代表第 n 局结束，甲获得-1 分；s_4 代表第 n 局结束，甲获得-2 分。

其中 2 分与-2 分为终止状态分别代表甲赢与乙赢。

状态转移矩阵为

$$P=\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ p & r & q & 0 & 0 \\ 0 & p & r & q & 0 \\ 0 & 0 & p & r & q \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

(2) 问在甲积 1 分的情况下, 再赛两局可以结束比赛的概率是多少?

$$S_{t+1}=P^T S_t=\begin{bmatrix} 1 & p & 0 & 0 & 0 \\ 0 & r & p & 0 & 0 \\ 0 & q & r & p & 0 \\ 0 & 0 & q & r & 0 \\ 0 & 0 & 0 & q & 1 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}=\begin{bmatrix} p \\ r \\ q \\ 0 \\ 0 \end{bmatrix}$$

$$S_{t+2}=\left(P^T\right)^2 S_t=\begin{bmatrix} 1 & p & 0 & 0 & 0 \\ 0 & r & p & 0 & 0 \\ 0 & q & r & p & 0 \\ 0 & 0 & q & r & 0 \\ 0 & 0 & 0 & q & 1 \end{bmatrix}\begin{bmatrix} 1 & p & 0 & 0 & 0 \\ 0 & r & p & 0 & 0 \\ 0 & q & r & p & 0 \\ 0 & 0 & q & r & 0 \\ 0 & 0 & 0 & q & 1 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}=\begin{bmatrix} p+pr \\ r^2+pq \\ 2qr \\ q^2 \\ 0 \end{bmatrix}$$

故而，在**两局内**结束的概率为 p+pr，**刚好**再赛两局结束的概率为 pr。

题 3

假设 $\gamma=0.9$, 收益序列是首项 $R_1=2$ 的无限 6 循环序列（即 $R_2=R_3=\cdots=6$ ）。那么 G_1 和 G_0 分别是多少?

$G_t=R_{t+1}+\gamma R_{t+2}+\gamma^2 R_{t+3}+\cdots=\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}$
$G_0=R_1+\gamma R_2+\gamma^2 R_3+\cdots=2+6\sum_{k=1}^{\infty}0.9^k=lim_{n\rightarrow\infty}2+6\times 9\times(1-0.9^n)=56$
$G_1=R_2+\gamma R_3+\gamma^2 R_4+\cdots=6\sum_{k=0}^{\infty}0.9^k=lim_{n\rightarrow\infty}60\times(1-0.9^n)=60$

题 4

对下面网格图，非终止状态集合 $S=\{1,2,\ldots,7\}$ 。每个状态有四种可能的动作 { up、down、left、right }, 对于每次转移 $\mathrm{R_t}=-1$, 每个动作会导致状态转移, 但当动作会导致智能体移出网格时, 状态保持不变。$\gamma=1$, 若 π 是等概率随机策略，那么行动价值 $\mathrm{q}_\pi(4,$ left) $\mathrm{q}_\pi(7,$ right) 是多少?

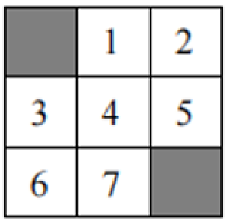

图 1

先写出状态转移矩阵，其中左上角终止状态记为 0，右下角终止状态记为 8。

$$
\begin{array}{cc}
 & \begin{array}{ccccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\
P= & \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.25 & 0.25 & 0.25 & 0 & 0.25 & 0 & 0 & 0 & 0 \\
0 & 0.25 & 0.5 & 0 & 0 & 0.25 & 0 & 0 & 0 \\
0.25 & 0 & 0 & 0.25 & 0.25 & 0 & 0.25 & 0 & 0 \\
0 & 0.25 & 0 & 0.25 & 0 & 0.25 & 0 & 0.25 & 0 \\
0 & 0 & 0.25 & 0 & 0.25 & 0.25 & 0 & 0 & 0.25 \\
0 & 0 & 0 & 0.25 & 0 & 0 & 0.5 & 0.25 & 0 \\
0 & 0 & 0 & 0 & 0.25 & 0 & 0.25 & 0.25 & 0.25 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\end{array}
$$

若理解为，智能体每个动作都会带来 $\mathrm{R_t}=-1$，且终止状态价值为 0，由

$$v(s)=r_s+\gamma\sum_{s'\in\mathbf{S}}p_{ss'}v\left(s'\right)$$

可得出 1,3,5,7 价值相同，记为 a，2，6 价值相同，记为 b。4 价值记为 c。

$$
\begin{aligned}
V(1) &= -1 + 0.25V(2) + 0.25V(4) + 0.25V(1) \\
V(2) &= -1 + 0.25V(1) + 0.5V(2) + 0.25V(5) \\
V(4) &= -1 + 0.5(V(1) + V(3) + V(5) + V(7))
\end{aligned}
$$

即为

$$
\begin{aligned}
2a + 5 &= b \\
b &= a - 2 \\
c &= a - 1
\end{aligned}
$$

故而:

$$
V = \begin{bmatrix} 0 & -7 & -9 \\ -7 & -8 & -7 \\ -9 & -7 & 0 \end{bmatrix}
$$

故而

$$
\begin{aligned}
&q_\pi(4, left) = r_4^{left} + \gamma v(3) = -1 + v(3) = -8 \\
&q_\pi(7, right) = r_7^{right} + \gamma v(\text{终点}) = -1
\end{aligned}
$$