インテリジェントシステム レポート課題 4

21T2166D 渡辺大樹

2024/07/08

1

(a) 状態価値関数 U(s) はここでは次の式で表すことができる。

$$U(s) = \max(R(s, stop), \sum_{s'} P(s'|s, spin)[R(s, spin, s') + \gamma U(s')])$$

状態 s=0 のときは

$$U(0) = \max(0, \sum_{s'} P(s'|0, spin)[R(0, spin, s') + \gamma U(s')])$$

となる。ここで

$$\sum_{s'} P(s'|0, spin)[R(0, spin, s') + \gamma U(s')]$$

を計算すると

$$\begin{aligned}\sum_{s'} P(s'|0, spin)[R(0, spin, s') + \gamma U(s')] &= \frac{1}{3}(U_0(2) + U_0(3) + U_0(4)) \\ &= \frac{1}{3}(0 + 0 + 0) \\ &= 0\end{aligned}$$

となる。よって

$$U(0) = \max(0, 0) = 0$$

となる。

状態 s=2 のときは

$$U(2) = \max(2, \sum_{s'} P(s'|2, spin)[R(2, spin, s') + \gamma U(s')])$$

となる。ここで

$$\sum_{s'} P(s'|2, spin)[R(2, spin, s') + \gamma U(s')]$$

を計算すると

$$\begin{aligned}\sum_{s'} P(s'|2, spin)[R(2, spin, s') + \gamma U(s')] &= \frac{1}{3}(U_0(4) + U_0(5) + 0) \\ &= \frac{1}{3}(0+0+0) \\ &= 0\end{aligned}$$

となる。よって

$$U(2) = \max(2, 0) = 2$$

状態 s=3 のときは

$$U(3) = \max(3, \sum_{s'} P(s'|3, spin)[R(3, spin, s') + \gamma U(s')])$$

となる。ここで

$$\sum_{s'} P(s'|3, spin)[R(3, spin, s') + \gamma U(s')]$$

を計算すると

$$\begin{aligned}\sum_{s'} P(s'|3, spin)[R(3, spin, s') + \gamma U(s')] &= \frac{1}{3}(U_0(5) + 0 + 0) \\ &= \frac{1}{3}(0+0+0) \\ &= 0\end{aligned}$$

となる。よって

$$U(3) = \max(3, 0) = 3$$

状態 s=4 のときは

$$U(4) = \max(4, \sum_{s'} P(s'|4, spin)[R(4, spin, s') + \gamma U(s')])$$

となる。ここで

$$\sum_{s'} P(s'|4, spin)[R(4, spin, s') + \gamma U(s')]$$

を計算すると

$$\begin{aligned}\sum_{s'} P(s'|4, spin)[R(4, spin, s') + \gamma U(s')] &= \frac{1}{3}(0 + 0 + 0) \\ &= \frac{1}{3}(0+0+0) \\ &= 0\end{aligned}$$

となる。よって

$$U(4) = \max(4, 0) = 4$$

状態 s=5 のときは

$$U(5) = \max(5, \sum_{s'} P(s'|5, spin)[R(5, spin, s') + \gamma U(s')])$$

となる。ここで

$$\sum_{s'} P(s'|5, spin)[R(5, spin, s') + \gamma U(s')]$$

を計算すると

$$\begin{aligned}\sum_{s'} P(s'|5, spin)[R(5, spin, s') + \gamma U(s')] &= \frac{1}{3}(0 + 0 + 0) \\ &= \frac{1}{3}(0 + 0 + 0) \\ &= 0\end{aligned}$$

となる。よって

$$U(5) = \max(5, 0) = 5$$

(b) $U_2(s)$ を考える。$U_2(s)$ は次のように表される。

$$U_2(s) = \max(R(s, stop), \sum_{s'} P(s'|s, spin)[R(s, spin, s') + \gamma U_1(s')])$$

状態 s=0 のときは

$$U_2(0) = \max(0, \sum_{s'} P(s'|0, spin)[R(0, spin, s') + \gamma U_1(s')])$$

となる。ここで

$$\sum_{s'} P(s'|0, spin)[R(0, spin, s') + \gamma U_1(s')]$$

を計算すると

$$\begin{aligned}\sum_{s'} P(s'|0, spin)[R(0, spin, s') + \gamma U_1(s')] &= \frac{1}{3}(U_1(2) + U_1(3) + U_1(4)) \\ &= \frac{1}{3}(2 + 3 + 4) \\ &= 3\end{aligned}$$

となる。よって

$$U_2(0) = \max(0, 3) = 3$$

状態 s=2 のときは

$$U_2(2) = \max(2, \sum_{s'} P(s'|2, spin)[R(2, spin, s') + \gamma U_1(s')])$$

となる。ここで

$$\sum_{s'} P(s'|2, spin)[R(2, spin, s') + \gamma U_1(s')]$$

を計算すると

$$\begin{aligned}\sum_{s'} P(s'|2, spin)[R(2, spin, s') + \gamma U_1(s')] &= \frac{1}{3}(U_1(4) + U_1(5) + 0) \\ &= \frac{1}{3}(4 + 5 + 0) \\ &= 3\end{aligned}$$

となる。よって

$$U_2(2) = \max(2, 3) = 3$$

状態 s=3 のときは

$$U_2(3) = \max(3, \sum_{s'} P(s'|3, spin)[R(3, spin, s') + \gamma U_1(s')])$$

となる。ここで

$$\sum_{s'} P(s'|3, spin)[R(3, spin, s') + \gamma U_1(s')]$$

を計算すると

$$\begin{aligned}\sum_{s'} P(s'|3, spin)[R(3, spin, s') + \gamma U_1(s')] &= \frac{1}{3}(U_1(5) + 0 + 0) \\ &= \frac{1}{3}(5 + 0 + 0) \\ &= \frac{5}{3}\end{aligned}$$

となる。よって

$$U_2(3) = \max(3, \frac{5}{3}) = 3$$

状態 s=4 のときは

$$U_2(4) = \max(4, \sum_{s'} P(s'|4, spin)[R(4, spin, s') + \gamma U_1(s')])$$

となる。ここで

$$\sum_{s'} P(s'|4, spin)[R(4, spin, s') + \gamma U_1(s')]$$

を計算すると

$$\begin{aligned}\sum_{s'} P(s'|4, spin)[R(4, spin, s') + \gamma U_1(s')] &= \frac{1}{3}(0 + 0 + 0) \\ &= \frac{1}{3}(0 + 0 + 0) \\ &= 0\end{aligned}$$

となる。よって

$$U_2(4) = \max(4, 0) = 4$$

状態 s=5 のときは

$$U_2(5) = \max(5, \sum_{s'} P(s'|5, spin)[R(5, spin, s') + \gamma U_1(s')])$$

となる。ここで

$$\sum_{s'} P(s'|5, spin)[R(5, spin, s') + \gamma U_1(s')]$$

を計算すると

$$\begin{aligned}
\sum_{s'} P(s'|5, spin)[R(5, spin, s') + \gamma U_1(s')] &= \frac{1}{3}(0+0+0) \\
&= \frac{1}{3}(0+0+0) \\
&= 0
\end{aligned}$$

となる。よって

$$U_2(5) = \max(5, 0) = 5$$

Bellman update を繰り返し、同様に $U_3(s), U_4(s)$ を求めると、

$$\begin{aligned}
U_3(0) &= 3 \\
U_3(2) &= 3 \\
U_3(3) &= 3 \\
U_3(4) &= 4 \\
U_3(5) &= 5 \\
U_4(0) &= 3 \\
U_4(2) &= 3 \\
U_4(3) &= 3 \\
U_4(4) &= 4 \\
U_4(5) &= 5
\end{aligned}$$

となり、$U(s)$ は

$$\begin{aligned}
U(0) &= 3 \\
U(2) &= 3 \\
U(3) &= 3 \\
U(4) &= 4 \\
U(5) &= 5
\end{aligned}$$

に収束する。

(c) $U_4(s)$ を用いた最適な行動 $\pi(s) \in \{spin, stop\}$ を求める。

状態 s=0 のときは

$$\pi(0) = \arg\max_{\mathbf{a}\in\{spin,stop\}} \sum_{s'} P(s'|0,\mathbf{a})[R(0,\mathbf{a},s') + \gamma U_4(s')]$$

となる。ここで

$$\sum_{s'} P(s'|0,spin)[R(0,spin,s') + \gamma U_4(s')]$$

を計算すると

$$\begin{aligned}\sum_{s'} P(s'|0,spin)[R(0,spin,s') + \gamma U_4(s')] &= \frac{1}{3}(U_4(2) + U_4(3) + U_4(4)) \\ &= \frac{1}{3}(3+3+4) \\ &= \frac{10}{3}\end{aligned}$$

となる。stop を選択したときは報酬が 0 なので

$$R(0,stop) = 0$$

となる。よって

$$\pi(0) = \arg\max_{\mathbf{a}\in\{spin,stop\}} (\frac{10}{3}, 0)$$

となり、$\pi(0) = spin$ となる。

状態 s=2 のときは

$$\pi(2) = \arg\max_{\mathbf{a}\in\{spin,stop\}} \sum_{s'} P(s'|2,\mathbf{a})[R(2,\mathbf{a},s') + \gamma U_4(s')]$$

となる。ここで

$$\sum_{s'} P(s'|2,spin)[R(2,spin,s') + \gamma U_4(s')]$$

を計算すると

$$\begin{aligned}\sum_{s'} P(s'|2,spin)[R(2,spin,s') + \gamma U_4(s')] &= \frac{1}{3}(U_4(4) + U_4(5) + 0) \\ &= \frac{1}{3}(4+5+0) \\ &= 3\end{aligned}$$

となる。stop を選択したときは

$$R(2,stop) = 2$$

となる。よって

$$\pi(2) = \arg\max_{\mathbf{a}\in\{spin,stop\}} (3, 2)$$

となり、$\pi(2) = spin$ となる。

状態 s=3 のときは

$$\pi(3) = \arg\max_{\mathbf{a}\in\{spin, stop\}} \sum_{s'} P(s'|3, \mathbf{a})[R(3, \mathbf{a}, s') + \gamma U_4(s')]$$

となる。ここで

$$\sum_{s'} P(s'|3, spin)[R(3, spin, s') + \gamma U_4(s')]$$

を計算すると

$$\begin{aligned}\sum_{s'} P(s'|3, spin)[R(3, spin, s') + \gamma U_4(s')] &= \frac{1}{3}(U_4(5) + 0 + 0)\\ &= \frac{1}{3}(5 + 0 + 0)\\ &= \frac{5}{3}\end{aligned}$$

となる。stop を選択したときは

$$R(3, stop) = 3$$

となる。よって

$$\pi(3) = \arg\max_{\mathbf{a}\in\{spin, stop\}} (\frac{5}{3}, 3)$$

となり、$\pi(3) = stop$ となる。

状態 s=4 のときは

$$\pi(4) = \arg\max_{\mathbf{a}\in\{spin, stop\}} \sum_{s'} P(s'|4, \mathbf{a})[R(4, \mathbf{a}, s') + \gamma U_4(s')]$$

となる。ここで

$$\sum_{s'} P(s'|4, spin)[R(4, spin, s') + \gamma U_4(s')]$$

を計算すると

$$\begin{aligned}\sum_{s'} P(s'|4, spin)[R(4, spin, s') + \gamma U_4(s')] &= \frac{1}{3}(0 + 0 + 0)\\ &= \frac{1}{3}(0 + 0 + 0)\\ &= 0\end{aligned}$$

となる。stop を選択したときは

$$R(4, stop) = 4$$

となる。よって

$$\pi(4) = \arg\max_{\mathbf{a}\in\{spin, stop\}} (0, 4)$$

となり、$\pi(4) = stop$ となる。

状態 s=5 のときは

$$\pi(5) = \arg\max_{\mathbf{a}\in\{spin,stop\}} \sum_{s'} P(s'|5,\mathbf{a})[R(5,\mathbf{a},s') + \gamma U_4(s')]$$

となる。ここで

$$\sum_{s'} P(s'|5,spin)[R(5,spin,s') + \gamma U_4(s')]$$

を計算すると

$$\begin{aligned}\sum_{s'} P(s'|5,spin)[R(5,spin,s') + \gamma U_4(s')] &= \frac{1}{3}(0+0+0)\\ &= \frac{1}{3}(0+0+0)\\ &= 0\end{aligned}$$

となる。stop を選択したときは

$$R(5,stop) = 5$$

となる。よって

$$\pi(5) = \arg\max_{\mathbf{a}\in\{spin,stop\}} (0,5)$$

となり、$\pi(5) = stop$ となる。

以上より、最適な行動は

$$\begin{aligned}\pi(0) &= spin\\ \pi(2) &= spin\\ \pi(3) &= stop\\ \pi(4) &= stop\\ \pi(5) &= stop\end{aligned}$$

となる。

(d) 初期方策を以下の図で定めたときの価値関数 $U^{\pi_0}(s)$ を考える。

	s = 0	s = 2	s = 3	s = 4	s = 5
π_0	spin	stop	spin	stop	spin

この方策のもとで価値関数 $U^{\pi_0}(s)$ は次の線形方程式を解くことで求めることができる。

$$
\begin{aligned}
U^{\pi_0}(0) &= \sum_{s'} P(s'|0, spin)[R(0, spin, s') + \gamma U^{\pi_0}(s')] \\
&= \frac{1}{3}(U^{\pi_0}(2) + U^{\pi_0}(3) + U^{\pi_0}(4)) \\
U^{\pi_0}(2) &= 2 \\
U^{\pi_0}(3) &= \sum_{s'} P(s'|3, spin)[R(3, spin, s') + \gamma U^{\pi_0}(s')] \\
&= \frac{1}{3}(U^{\pi_0}(5) + 0 + 0) \\
U^{\pi_0}(4) &= 4 \\
U^{\pi_0}(5) &= \sum_{s'} P(s'|5, spin)[R(5, spin, s') + \gamma U^{\pi_0}(s')] \\
&= \frac{1}{3}(0 + 0 + 0) \\
&= 0
\end{aligned}
$$

となる。これを解くと

$$
\begin{aligned}
U^{\pi_0}(0) &= \frac{1}{3}(2 + 0 + 4) = 2 \\
U^{\pi_0}(2) &= 2 \\
U^{\pi_0}(3) &= \frac{1}{3}(0 + 0 + 0) = 0 \\
U^{\pi_0}(4) &= 4 \\
U^{\pi_0}(5) &= 0
\end{aligned}
$$

となる。

(e) 続いて、$U^{\pi_0}(s)$ を用いて方策改善を行う。

$$
\pi_1(s) = \arg \max_{\mathbf{a} \in \{spin, stop\}} \sum_{s'} P(s'|s, \mathbf{a})[R(s, \mathbf{a}, s') + \gamma U^{\pi_0}(s')]
$$

となる。これを計算すると

$$
\begin{aligned}
\pi_1(0) &= \arg \max_{\mathbf{a} \in \{spin, stop\}} \sum_{s'} P(s'|0, \mathbf{a})[R(0, \mathbf{a}, s') + \gamma U^{\pi_0}(s')] \\
&= \arg \max_{\mathbf{a} \in \{spin, stop\}} (\frac{1}{3}(U^{\pi_0}(2) + U^{\pi_0}(3) + U^{\pi_0}(4)), 0) \\
&= \arg \max_{\mathbf{a} \in \{spin, stop\}} (\frac{1}{3}(2 + 0 + 4), 0) \\
&= \arg \max_{\mathbf{a} \in \{spin, stop\}} (\frac{6}{3}, 0) \\
&= \arg \max_{\mathbf{a} \in \{spin, stop\}} (2, 0) \\
&= spin
\end{aligned}
$$

となる。同様に計算すると

$$\begin{aligned}\pi_1(2) &= \arg\max_{\mathbf{a}\in\{spin,stop\}}(\frac{1}{3}(U^{\pi_0}(4)+U^{\pi_0}(5)+0),2)\\&= \arg\max_{\mathbf{a}\in\{spin,stop\}}(\frac{1}{3}(4+0+0),2)\\&= \arg\max_{\mathbf{a}\in\{spin,stop\}}(\frac{4}{3},2)\\&= \arg\max_{\mathbf{a}\in\{spin,stop\}}(1.\dot{3},2)\\&= stop\end{aligned}$$

$$\begin{aligned}\pi_1(3) &= \arg\max_{\mathbf{a}\in\{spin,stop\}}(\frac{1}{3}(U^{\pi_0}(5)+0+0),3)\\&= \arg\max_{\mathbf{a}\in\{spin,stop\}}(\frac{1}{3}(0+0+0),3)\\&= \arg\max_{\mathbf{a}\in\{spin,stop\}}(0,3)\\&= stop\end{aligned}$$

$$\begin{aligned}\pi_1(4) &= \arg\max_{\mathbf{a}\in\{spin,stop\}}(\frac{1}{3}(0+0+0),4)\\&= \arg\max_{\mathbf{a}\in\{spin,stop\}}(0,4)\\&= stop\end{aligned}$$

$$\begin{aligned}\pi_1(5) &= \arg\max_{\mathbf{a}\in\{spin,stop\}}(\frac{1}{3}(0+0+0),5)\\&= \arg\max_{\mathbf{a}\in\{spin,stop\}}(0,5)\\&= stop\end{aligned}$$

となる。

以上より、方策改善によって得られた方策は

	s = 0	s = 2	s = 3	s = 4	s = 5
π_1	spin	stop	stop	stop	stop

となる。

2

(a)

状態 s_2, s_4, s_5 はそれぞれ s_4, s_5, s_0 にしか遷移しないため価値関数 U はそれぞれ

$$U(s_2) = 80, U(s_4) = 90, U(s_5) = 100$$

となる。

(b)

状態 s_3 での状態価値関数 $U(s_3)$ は

$$
\begin{aligned}
U(s_3) &= \max_{\mathbf{a}\in\{a,b\}} \sum_{s'} P(s'|s_3,\mathbf{a})[R(s_3,\mathbf{a},s') + \gamma U(s')] \\
&= \sum_{s'} P(s'|s_3,b)[R(s_3,b,s') + \gamma U(s')] \\
&= P(s_4|s_3,b)[R(s_3,b,s_4) + \gamma U(s_4)] \\
&\quad + P(s_5|s_3,b)[R(s_3,b,s_5) + \gamma U(s_5)]
\end{aligned}
$$

となる。$P(s_4|s_3,b) = p, P(s_5|s_3,b) = q$ を代入し、ほかの値も課題資料中の表の値を用いると、

$$U(s_3) = 80p + 96q$$

$p + q = 1$ より、p または q で整理すると

$$
\begin{aligned}
&= 96 - 16p \\
&= 16q + 80
\end{aligned}
$$

となる。

(c)

状態 s_1 での状態価値関数 $U(s_1)$ は

$$
\begin{aligned}
U(s_1) &= \max_{\mathbf{a}\in\{a,b\}} \sum_{s'} P(s'|s_1,\mathbf{a})[R(s_1,\mathbf{a},s') + \gamma U(s')] \\
&= \max(P(s_3|s_1,a)[R(s_1,a,s_3) + \gamma U(s_3)], P(s_2|s_1,b)[R(s_1,b,s_2) + \gamma U(s_2)])
\end{aligned}
$$

である。ここに資料中の値と前問で出した答えを代入すると

$$U(s_1) = \max(86 - 16p, 85)$$

となる。この関数の解は

$$U(s_1) = \begin{cases} 86 - 16p & (p <= \frac{1}{16}) \\ 85 & \text{otherwise} \end{cases}$$

となる。