インテリジェントシステム レポート課題 4

21T2166D 渡辺大樹

2023/07/21

1

(a)

1 回 Bellman update を行った状態価値関数 $U_1(s)$ の一般式は

$$U_1(s) = \max_{a \in \{a_1, a_2\}} \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma U_0(s')]$$

となる。

今回初期値として与えられる U_0 はすべて 0 なので

$$U_1(s) = \max_{a \in \{a_1, a_2\}} \sum_{s'} P(s'|s,a) R(s,a,s')$$

と書いてしまって計算する。またこの環境での s の取りうる値も s_1, s_2 のどちらかであるため $U_1(s_0)$ は 0 である。

まず $U_1(s_1)$ を計算する。$U_1(s_1)$ は状態行動価値関数 $Q_1(s_1, a) = \sum_{s'} P(s'|s,a)R(s,a,s')$ を用いて

$$U_1(s_1) = \max_{a \in \{a_1, a_2\}} Q_1(s_1, a)$$

と表せる。$Q(s,a)$ は課題資料中の表から計算することで

$$Q_1(s_1, a_1) = 1, Q_1(s_1, a_2) = 2$$

となるため、二つから最大値を取って

$$U_1(s_1) = 2$$

と計算できる。

s_2 でも同様な計算を行うと

$$Q_1(s_2, a_1) = 2, Q_1(s_2, a_2) = -10$$

となり、$U_1(s_2)$ は

$$U_1(s_2) = 2$$

となる。

表で示すと以下のようになる。

	s_1	s_2	s_0
U_1	2	2	0

(b)

(a) から Bellman update をもう一度行うと状態価値関数 $U_2(s)$ の一般式は

$$U_2(s) = \max_{a \in \{a_1, a_2\}} \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma U_1(s')]$$

となる。また (a) と同様 $U_2(s_0) = 0$ である。

まず $U_2(s_1)$ を計算する。状態行動価値関数 $Q_2(s,a)$ は

$$Q_2(s,a) = \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma U_1(s')]$$

となる。課題資料中の表、また (a) の回答より $Q_2(s_1,a)$ を a についてそれぞれ計算することで

$$Q_2(s_1,a_1) = 2, Q_2(s_1,a_2) = 3$$

が得られる。これの最大値を取るので

$$U_2(s_1) = 3$$

となる。

同様にして $U_2(s_2)$ も計算していくと

$$Q_2(s_2,a_1) = 2, Q_2(s_2,a_2) = -9$$

となる。したがって最大値を取ることで

$$U_2(s_2) = 2$$

を得る。

表で表すと以下のようになる。

	s_1	s_2	s_0
U_2	3	2	0

(c)

方策評価によって得られる価値関数 $U^{\pi_i} = U_i(s)$ は

$$U_i(s) = \sum_{s'} P(s'|s,\pi_i(s))[R(s,\pi_i(s),s') + \gamma U_i(s')]$$

となる。今回求めるのは $U^{\pi_i} = U_0(s)$ であるので、式は

$$U_0(s) = \sum_{s'} P(s'|s,\pi_0(s))R(s,\pi_0(s),s')$$

としてしまう。課題資料にある π_0 を用いて s_1 から計算していく。

$U_0(s_1)$ は

$$\begin{aligned} U_0(s_1) &= P(s_1|s_1,a_1)R(s_1,a_1,s_1) \\ &= 1 \end{aligned}$$

となる。同様に $U_0(s_2) = -10, U_0(s_0) = 0$ と計算できる。

表で表すと以下のようになる。

	s_1	s_2	s_0
U_0	1	-10	0

(d)

方策 $\pi_1(s)$ は

$$\pi_1(s) = \arg\max_a \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma U_0(s')]$$

となる。$\gamma = \frac{1}{2}$ で、$U_0(s)$ は前問の値を使う。

状態行動価値関数 $Q(s,a)$ を $s = s_1, s_2$ について a_1, a_2 それぞれで計算すると

$$\begin{aligned} Q(s_1,a_1) &= P(s_1|s_1,a_1)[R(s_1,a_1,s_1) + \frac{1}{2}U_0(s_1)] \\ &= 1.5 \\ Q(s_1,a_2) &= P(s_1|s_1,a_2)[R(s_1,a_2,s_1) + \frac{1}{2}U_0(s_1)] \\ &\quad + P(s_2|s_1,a_2)[R(s_1,a_2,s_2) + \frac{1}{2}U_0(s_2)] \\ &= -0.25 \\ Q(s_2,a_1) &= P(s_1|s_2,a_1)[R(s_2,a_1,s_1) + \frac{1}{2}U_0(s_1)] \\ &\quad + P(s_2|s_2,a_1)[R(s_2,a_1,s_2) + \frac{1}{2}U_0(s_2)] \\ &= -1.25 \\ Q(s_2,a_2) &= P(s_0|s_2,a_2)[R(s_2,a_2,s_0) + \frac{1}{2}U_0(s_0)] \\ &= -10 \end{aligned}$$

よって以上の結果から

$$\pi_1(s_1) = a_1, \pi_1(s_2) = a_1$$

となる。

表で表すと以下のようになる。

	s_1	s_2
π_1	a_1	a_1

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

$$U(s_2) = -10, U(s_4) = -10, U(s_5) = 100$$

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、

$$U(s_3) = -20p + 96q$$

$p + q = 1$ より、p または q で整理すると

$$\begin{aligned} &= 96 - 116p \\ &= 116q - 20 \end{aligned}$$

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

$$U(s_1) = \max(86 - 116p, -5)$$

となる。この関数の解は

$$U(s_1) = \begin{cases} 86 - 116p & (p < \frac{91}{116}) \\ -5 & \text{otherwise} \end{cases}$$

となる。