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

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

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

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

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 $Q_1(s_1, a) = \sum_{s'} P(s'|s,a)R(s,a,a')$ 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

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

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

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

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_1	3	2	0