作业 3 参考答案

人工智能导论课（2023 春季学期）

编辑于 2023/5/23

1. 网格游戏

使用动态规划的方法求解最优解 $V_t(s)$, 表示状态 s 的最大值当还有 t 步游戏结束以后。$V^*(Y)=25$，$V^*(X)=\frac{25}{8}$。在 6 次后 (或第 7 次迭代时) 所有状态 V 值收敛。

2. 骰子游戏

1)

状态 3 至 6 的行动都是停止，则 $V(s_i)=i$，对于所有 $i\in\{3,4,5,6\}$。因为游戏停止，玩家收取相应投数量的金钱。对于 $V(s_1),V(s_2)$ 可以利用 Bellman 等式求解，即：

$$
\begin{aligned}
V(s_1) &= -1+\frac{1}{6}[V(s_1)+V(s_2)+3+4+5+6] \\
V(s_2) &= -1+\frac{1}{6}[V(s_1)+V(s_2)+3+4+5+6]
\end{aligned}
$$

解得 $V(s_1)=V(s_2)=3$。

2)

根据 Bellman 等式，可以推出：

$$
\begin{aligned}
\pi(s) &= \arg\max_a Q(s,a) \\
Q(s,a) &= \sum T(s,a,s')[R(s,a,s')+\gamma V(s')]
\end{aligned}
$$

给定上一步算出的状态值，对于每个状态 S_i，采取继续投掷的回报值是 $-1+\frac{1}{6}(3+3+3+4+5+6)=3$，而采取停止的回报值是 i。因为每个状态选择的是最大回报值的行动，所以，更新后的策略是：状态 s_1,s_2 采取继续的行动，s_4,s_5,s_6 都采取停止，在状态 s_3 采取继续和停止行动的回报值是相等的，所以两个行动采取哪一个都可以。

3)

从 2) 中可以发现 $\pi(s)$ 和 $\pi'(s)$ 这两个策略等效，所以策略迭代算法已经收敛，$\pi(s)$ 是最优策略。

3. 强化学习算法

1)

Q 状态值原本应该是状态值的加权平均，但是我们不知道权值，只有样本，所以用样本值的移动平均法（时间差分法，或叫指数平均法）来近似。利用样本更新 Q 状态值的公式如下所示：

$$\begin{aligned} Q(s,a) \leftarrow & (1-\alpha)Q(s,a) + \alpha\text{SampleVal} \\ \text{SampleVal} = & r(s,a,s') + \gamma \max_{a'} Q(s',a') \end{aligned}$$

计算本题中的两个 Q 状态值，实际只需要相关的三个样本进行以下的计算：

$$\begin{aligned} Q(A,\text{前进}) \leftarrow & (1-\alpha)Q(A,\text{前进}) + \alpha(r + \gamma \max_a Q(B,a)) = 0.5(0) + 0.5(2+0) = 1 \\ Q(C,\text{停止}) \leftarrow & (1-\alpha)Q(C,\text{停止}) + \alpha(r + \gamma \max_a Q(A,a)) = 0.5(0) + 0.5(0+1) = 0.5 \\ Q(C,\text{前进}) \leftarrow & (1-\alpha)Q(C,\text{前进}) + \alpha(r + \gamma \max_a Q(A,a)) = 0.5(0) + 0.5(2+1) = 1.5 \end{aligned}$$

a) 0.5

b) 1.5

2)

为了提高表达的泛化性，Q 状态值用特征函数的线性组合来表示，即，$Q(s,a) = \sum_i w_i f_i(s,a)$，给定样本，根据梯度下降法更新权值 w_i：

$$\begin{aligned} w_i \leftarrow & w_i + \alpha(\text{diff}) f_i(\cdot) \\ \text{diff} = & \text{sampleVal} - Q(s,a) \\ \text{sampleVal} = & [r + \gamma \max_{a'} Q(s',a')] \end{aligned}$$

相关的特征函数值可计算：$f_1(A,\text{前进}) = 1, f_2(A,\text{前进}) = 1$。

a)

初始时 $w_i = 0$，计算原始的 $Q(A,\text{前进}) = w_1 f_1(A,\text{前进}) = 1 + w_2 f_2(A,\text{前进}) = 0$，同样地，$Q(B,\text{前进}) = Q(B,\text{停止}) = 0$。获得第一个样本，$\text{diff} = [r + \gamma \max_a Q(B,a)] - Q(A,\text{前进}) = 4$，继而得到：

$$
\begin{aligned}
w_1 =& w_1 + \alpha(\text{diff}) f_1(A, \text{前进}) = 2 \\
w_2 =& w_2 + \alpha(\text{diff}) f_2(A, \text{前进}) = 2
\end{aligned}
$$

- $w_1 = 2$
- $w_2 = 2$

b)

获得第二个样本，此时，$w_1 = w_2 = 2$，相应的特征函数值，$f_1(B, \text{停止}) = 1$, $f_2(B, \text{停止}) = -1$；$Q(B, \text{停止}) = 2 \cdot 1 + 2(-1) = 0$, $Q(A, \text{前进}) = 2 \cdot 1 + 2 \cdot 1 = 4$, $Q(A, \text{停止}) = 2 \cdot 1 + 2 \cdot (-1) = 0$。

而且，$\text{diff} = [r + \gamma \max_a Q(A, a)] - Q(B, \text{停止}) = [0 + 1 \cdot 4] - 0 = 4$，继而得到：

$$
\begin{aligned}
w_1 =& w_1 + \alpha(\text{diff}) f_1(B, \text{停止}) = 2 + 0.5 \cdot 4 \cdot 1 = 4 \\
w_2 =& w_2 + \alpha(\text{diff}) f_2(B, \text{停止}) = 2 + 0.5 \cdot 4 \cdot (-1) = 0
\end{aligned}
$$

- $w_1 = 4$
- $w_2 = 0$