GridWorld

AAIS in PKU　陈伟杰 1901111420

October 18, 2019

1 Problem Setting

设置一个 5×5 的自由行走的网格空间 $\mathbb{S}$, 行动空间 $\mathbb{A} = \{\uparrow, \downarrow, \leftarrow, \rightarrow\}$，定义行走规则和边界条件如下：

$$\begin{aligned} (s,a) &\to (s', r=0) \quad \text{if } s' \in \mathbb{S} \\ (s,a) &\to (s, r=-1) \quad \text{if } s' \notin \mathbb{S} \end{aligned} \tag{1}$$

其中设置两个传送点 $A = (0,1)$ 和 $B = (0,3)$，及其对应的到达点 $A' = (4,1)$ 和 $B' = (2,3)$，定义如下：

$$\begin{aligned} (s = A, \forall a) &\to (s' = A', r = 10) \\ (s = B, \forall a) &\to (s' = B', r = 5) \end{aligned} \tag{2}$$

2 Experiment Result

在实验中采用随机行走作为策略，即 $\pi(a|s) = \frac{1}{4}$。且对于 GridWorld 问题，对于给定状态和行动，可以精确得到下一状态和对应的回报，即 $p(s', r|s, a) = 1$。因此，$q_\pi(s,a)$ 和 $v_\pi(s)$ 的 Bellman 方程可以写成：

$$\begin{aligned} q_\pi(s,a) &= r_a + \frac{1}{4}\gamma \sum_{a' \in \mathbb{A}} q_\pi(s', a') \\ v_\pi(s) &= \frac{1}{4} \sum_{a \in \mathbb{A}} q_\pi(s,a) \end{aligned} \tag{3}$$

实验中设置 $\gamma = 0.9$，$\epsilon = 1e-4$，当 $|q_\pi^{k+1}(s,a) - q_\pi^k(s,a)| < \epsilon$ 时，迭代停止。最终得到结果：

$$v_\pi(\mathbb{S}) = \begin{array}{|c|c|c|c|c|} \hline 3.31 & 8.79 & 4.43 & 5.32 & 1.49 \\ \hline 1.52 & 2.99 & 2.25 & 1.91 & 0.55 \\ \hline 0.05 & 0.74 & 0.67 & 0.36 & -0.40 \\ \hline -0.97 & -0.44 & -0.35 & -0.59 & -1.18 \\ \hline -1.85 & -1.35 & -1.23 & -1.42 & -1.98 \\ \hline \end{array} \tag{4}$$

对于最终得到的 $q_\pi(s,a)$ 矩阵，假如在每个 s 选择 greedy 方法，可得到最优策略 $\pi_*(a|s) = argmax_a q_\pi(s,a)$:

$$\pi_*(a|\mathbb{S}) = \begin{array}{|c|c|c|c|c|} \hline \rightarrow & \times & \leftarrow & \times & \leftarrow \\ \hline \uparrow & \uparrow & \uparrow & \uparrow & \leftarrow \\ \hline \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\ \hline \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\ \hline \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\ \hline \end{array} \tag{5}$$

这与求解最优值 Bellman 方程得到的最优策略相比，主要区别在于 (1,3) 点的行动显得更短视，容易陷入局部最优解，即 $B \to B' \to B$ 的循环。

3 README

代码包括 GridWorld.py 和 test.py 两个程序，需要 python3 环境。直接运行 test.py 可以得到上述结果。