

## Chapter 3

# Finite Markov Decision Processes

### 3.8 GridWorld

#### 3.8.1 题目说明

本题的大意是存在一个  $5 \times 5$  的矩阵，对于每个元素 (cell)，可以有”东南西北”四个方向的前进方向，奖励函数满足如下性质：

1. 存在两个特殊的状态  $A(0, 1), B(0, 3)$ ，这两个点无论往哪个方向的 reward 均为 +10，且将会移动到  $A'(4, 1), B'(2, 3)$ 。
2. 如果在边界的状态，若尝试把它向边界外面移动，则会弹回原位，且 reward=-1。
3. 折扣系数  $\gamma = 0.9$ ，此题采用随机策略。

#### 3.8.2 解题思路

题目提示采用  $\gamma$  折扣的积累奖赏的策略评估算法，根据公式 (3.14)：

$$v_{\pi}(s) = \sum_{a \in A} \pi(a|s) \sum_{s', r} P(s', r|s, a) [r + \gamma v_{\pi}(s')] \quad (3.1)$$

由于在本题中，一个状态执行一个动作到达另一个状态是确定的，故  $\sum_{s',r} P(s',r|s,a) = 1$  故原式可简化成：

$$V_{\pi}(s) = \sum_{a \in A} \pi(a|s) [R_{s \rightarrow s'}^a + \gamma V_{\pi}(s')] \quad (3.2)$$

其中  $V_{\pi}(s)$  是一个  $5 \times 5$  的价值函数矩阵， $V_{\pi}(s)_{ij}$  代表第  $i$  行第  $j$  列 cell 的价值函数，同理  $R$  也是相应的 reward 矩阵， $s'$  代表变化后的状态。

在上面的前提下我们可以得出每一项的具体数值：

- $\pi(a|s)$  代表在状态  $s$  下选额动作  $a$  的概率，由于随机选择”东南西北”四个方向，故：

$$\pi('north'|s) = \pi('south'|s) = \pi('west'|s) = \pi('east'|s) = 0.25 \quad (3.3)$$

- $R$  是一个  $5 \times 5 \times 4$  的矩阵，每一项对应每个 cell 选择每个动作的 reward。
- 每一遍新的迭代每个 cell 的初始值为 0，其值的计算公式为：

$$v_{\pi}(s) = \sum \pi(a|s)(r + \gamma v_{\pi}(s')) \quad (3.4)$$

即该 cell 东南西北四个相邻 cell 的值与相应的奖励之和，再与选择该动作概率的乘积之和。

综上所述我们通过遍历  $V_{\pi}(s)$  来更新每一个  $v_{\pi}(s)$ ，设置一个阈值  $\theta$ ，若  $\text{sum}(|V_{\pi}(s) - V_{\pi}(s')|) < \theta$ ，可停止算法。

### 3.8.3 实验结果

设  $\theta = 10^{-6}$  经过 121 次迭代后， $V_{\pi}(s)$  收敛得到右边的矩阵，如下截图：

```
after running: 121
[[ 3.3  8.8  4.4  5.3  1.5]
 [ 1.5  3.   2.3  1.9  0.5]
 [ 0.1  0.7  0.7  0.4 -0.4]
 [-1.   -0.4 -0.4 -0.6 -1.2]
 [-1.9 -1.3 -1.2 -1.4 -2. ]]
```

图 3.1: example 3.8 实验结果

上面的矩阵说明，经过随机策略后，若我们想达到最大化期望收益，要选择东南西北四个方向中最大的一个值作为移动方向。通过观察，特殊点  $A, B$  具有最大价值函数，与我们主观认知是相同的。说明机器已经通过随机策略初步掌握了 GridWorld 的玩法。

代码已经上传到[github](#).