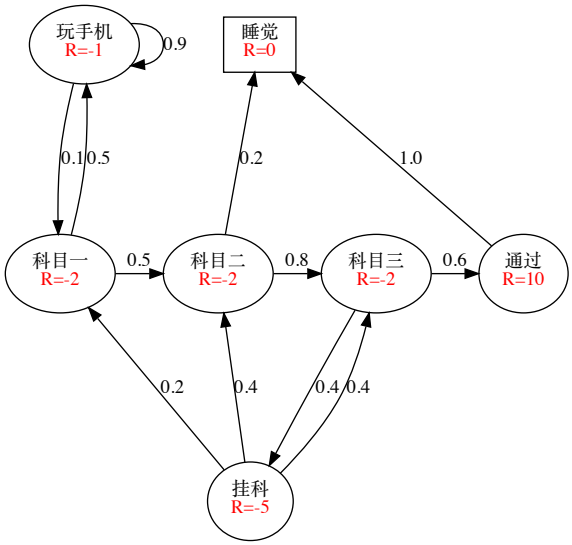
作业 2：马尔可夫决策过程

陈达贵　深蓝学院

2018-12-14

1　文字作业

1. (2 分) 计算 MRPs。在课堂上，我们说过小的 MRPs 是可以通过直接解贝尔曼方程解出来的，课程上也给出了几个例子。这里需要大家通过矩阵计算求出当 $\gamma = 0.5$ 时的各个状态的 V 函数值。



2. (2 分) 说明为什么下面 MDPs 的贝尔曼期望方程的两种形式等价，这两者在使用的时候有什么区别?

$$\begin{cases} v_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s)\left(\mathcal{R}(s,a) + \gamma \sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'} v_\pi(s')\right) \\ q_\pi(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'} \sum_{a'\in\mathcal{A}} \pi(a'|s') q_\pi(s',a') \end{cases}$$

$$\Leftrightarrow$$

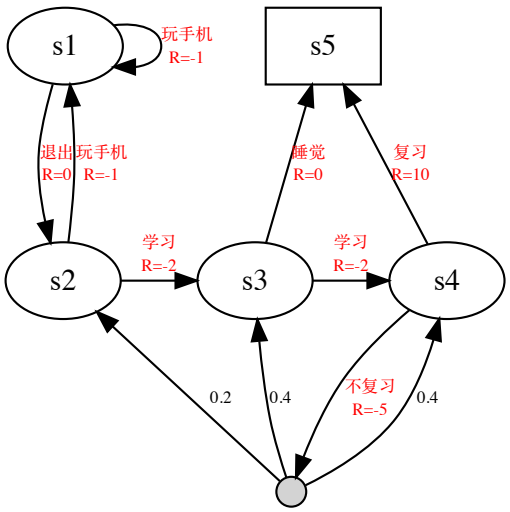
$$\begin{cases} v_\pi(s) = \mathbb{E}_\pi\left[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s\right] \\ q_\pi(s,a) = \mathbb{E}_\pi\left[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) | S_t = s, A_t = a\right] \end{cases}$$

3. (2 分) 课程中提到当得到最优 Q 函数之后，能够直接得到最优策略，那么如果已知最优 V 函数，能否得到最优策略呢? 如果能，写出两者之间的关系，如果不能说明为什么?

2 编程作业

1. (4 分)

   - 实现下图的环境，需要实现环境中的动态转移函数。
   - 实现一个 agent, 策略是随机的，通过仿真的方式，用回报值的经验平均去估计每个状态的值函数。验证仿真的结果和课件中计算的结果。（分别仿真 $\gamma = 0.5, 1$）
   - 强化学习中寻找最优策略的方法有很多种，其中全局遍历是朴素的解法，由于对于 MDPs 总存在最优的确定性策略，通过全局遍历所有确定性策略，并比较策略即可求出最优策略。在该 MDPs 中只有四个状态可以决策，每个状态只有两个可行的动作，所以总共有 2^4 个确定性策略，使用算法遍历所有策略，并输出最优策略。（分别考虑 $\gamma = 0.5, 1$）


s1
玩手机
R=-1
s5
退出
R=0
玩手机
R=-1
睡觉
R=0
复习
R=10
学习
R=-2
学习
R=-2
s2
s3
s4
0.2
0.4
不复习
R=-5
0.4