作业 3：动态规划

陈达贵　深蓝学院

2018-12-21

1　文字作业

1. (3 分) 在课堂上，我们针对 V 函数给出了策略评价、策略迭代和值迭代算法。现在要求

   - 给出 Q 函数的策略评价算法
   - 给出 Q 函数的策略迭代算法
   - 给出 Q 函数的值迭代算法

2. (1 分) 在策略提升阶段，课堂上我们只讲了贪婪的策略提升，除了贪婪的策略提升外，还有一种策略提升方法叫 ε-greedy，带 ε 的贪婪策略。具体指，我们在选择动作时，有 $1-\varepsilon$ 的概率选择贪婪的动作，有 ε 的概率随机选择动作。思考 ε 贪婪策略和贪婪策略有什么不同？各有什么优缺点?

2　编程作业

1. (6 分) 下图是一个格子迷宫，这个格子迷宫可以用一个简单的 MDP 来描述。其中我们可以用智能体的位置表示状态。在每个格子处，智能体可以做出四个动作: North, West, South, East。所有的动作都能

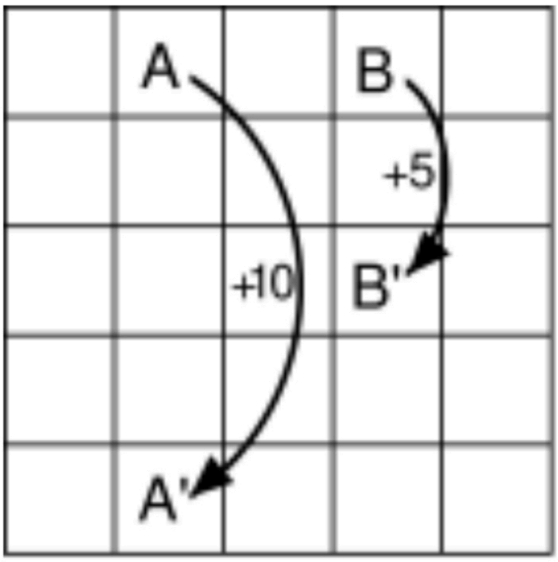
导致智能体以确定性的形式改变位置。其中 A 和 B 是两个特殊状态，奖励的定义如下, 其中 $\gamma = 0.9$:

- 除了 A、B 外，任何想要走出边界的动作都会撞墙，从而导致位置不变，获得-1 的奖励
- 除 A、B 外，任何非撞墙的动作都会获得 0 的奖励
- 在 A 处无论做什么动作，都会跳转到 A′，并获得 +10 的奖励
- 在 B 处无论做什么动作，都会跳转到 B′，并获得 +5 的奖励

我们需要实现以下功能, 假设值函数初始化为 0

- 实现环境逻辑
- 使用迭代式策略评价算法，计算随机策略的值函数。计算完后用表格的形式画出来
- 使用策略迭代算法，计算最优值函数。并画出最优值函数和最优策略，并画出收敛曲线
- 使用值迭代算法，计算最优值函数。并画出收敛曲线，和策略迭代对比
- 使用就地 (in-place) 的值迭代算法，计算最优值函数。并画出收敛曲线, 与前两者对比

(提示：环境部分代码已给；由于好策略的每一个状态的值函数都要大于坏的策略，所以收敛曲线的一种绘制方式画出平均值函数-迭代步数的曲线)


A
B
+5
+10
B'
A'