实验2：V值预测问题与最优策略

1. 实验目的
2. 熟悉并掌握强化学习的内容；
3. 深入理解V值预测问题与最优策略，掌握基本原理；
4. 掌握强化学习设计与编程实现的基本技能
5. 实验环境
6. 接入Internet 的实验主机；
7. Windows 7/8/10操作系统，IOS, linux操作系统等
8. 实验内容

值预测问题给定以下⽹格环境(可⾃⾏定义该 MDP 问题(状态空间，状态转移函数，收益函数等)或图中加⼀些障碍物等)，给定任意随机性策略，在对环境的MDP 模型未知的情况下，分别使⽤ MC 和TD(0)学习算法计算该策略下每个状态的V 值。



在下⾯环境 MDP 定义基础上，在当前环境的 MDP 模型未知的情况下，分别使⽤ MC 和 TD(0)学习算法计算出各状态下的最优 V 值以及最优策略。

附加题：

在下⾯环境 MDP 定义基础上，在当前环境的 MDP 模型未知的情况下，使⽤ n-step 学习和 TD(lamda)算法计算出各状态下的最优 V 值以及最优策略。

1. 实验步骤

【具体方法参考课程使用书籍】

对问题的分析，解决方案的各个环节的功能要求，描述解决方案时，应当注意的条件及限制规则。

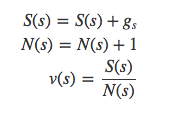
问题分析：该问题有一个reward为10的终结态和reward为0的终结态。各点走向应该尽量向+10的终结态靠近。但是因为还有一个0的终结态，考虑到路上行走所花的代价，应该尽量权衡两者利弊。

蒙特卡罗算法：

基于蒙特卡罗算法，需要一系列随机样本。这里我选择的随机方式是随机选择一个状态，让它向任意方向移动，直到移动到终结态，记录下它每次移动的动作，到达的状态以及获得的reward，随机样本数目为100。然后基于所有获得的一系列样本进行计算其衰减奖励：

../../../../../../Desktop/屏幕快照%202017-11-07%20下午11.3

然后更新状态价值：



时差学习算法：

利用马尔可夫性质，只利用下一步信息，每探索一步都进行状态价值的更新：

../../../../../../Desktop/屏幕快照%202017-11-07%20下午11.4

这里为当前状态，s′为下一步状态。r是系统获得的奖励，α是学习率，γ是衰减因子

1. 实验方式

每位同学独立上机编程实验，实验指导教师现场指导。

1. 参考内容

强化学习

参考书籍：《Reinforcement Learning:An Introduction》

作者：Richard S. Sutton and Andrew G. Barto

1. 相关软件下载

Code::Blocks <http://www.codeblocks.org/> (C++编辑器)

PyCharm <http://www.jetbrains.com/pycharm/> （Python编辑器）

1. 实验报告要求
2. 原题描述
3. 问题的分析与公式原理
4. 算法实现：代码以及必要的注释
5. 实验结果以及结果分析
6. 实验心得（未来的展望与思考）