作业 1：强化学习概述

陈达贵　深蓝学院

2018-12-7

1　文字作业

1. (3 分) 在使用强化学习方法解决问题的时候，最关键的一点就是能够定义强化学习问题。良好的强化学习定义是解决问题的先决条件。选择你认为可以建模成强化学习的**两个**场景，并回答下面的问题

   (a) 为什么这个场景需要用强化学习来建模？用监督学习，赌博机问题建模可不可以？

   (b) 在这个强化学习问题中，环境和智能体分别指什么？

   (c) 如何定义状态、动作和奖励？思考定义的合理性

   (d) 是全观测的还是部分观测的？为什么？

   (e) 环境模型已知还是未知？

   (f) 你所认为的最佳策略应该是确定性的还是随机性的？

2. (1 分) 自我对弈在训练各种对抗性的强化学习问题时（比如棋类）游戏，除了跟随机的对手对弈外，还可以通过自己跟自己对弈来训练。自我对弈（self-play）可能会导致什么样的结果？会学到和之前不同的策略吗？

3. (2 分) 贪婪策略贪婪策略 (greedy policy) 指每一步动作都按照当前最大的值函数去执行。一个贪婪策略相比非贪婪的策略有什么优劣? 可能会导致什么问题?（hint：从学习过程和学习完成之后两个方面来考虑）

2 编程作业

1. 熟悉 python 基本用法，numpy 基本用法

2. (4 分) *TicTacToe* 是一个简单的对抗游戏，棋盘大小为 3×3，谁先将棋子连成线（横、竖、斜），谁就获得胜利。(× 先手) 这里要求大家实现以下功能：

   (a) 用数值的方式表示状态、动作、奖励（+1/0/-1 区分胜/平/负）

   (b) 环境类，环境能够根据智能体的动作给出反馈。即实现成员函数 $\text{step}(a) \rightarrow s, r$

   (c) 智能体类，并包含一个随机策略，即从剩下的空位中随机采样一个位置下。函数形式 $\text{policy}(s) \rightarrow a$

   (d) 通过仿真的方式，大量对弈，统计在都执行随机策略的情况下，先手和后手胜利的概率。

   注：框架代码已经给出，见 code1.py

   注：本作业主要希望大家从代码层面体会环境和智能体之间的关系，用编程函数的形式去实现数学函数 $\mathcal{R}, \mathcal{P}, \pi$