强化学习笔记[1]简介

这是Albert在学习强化学习时的笔记，他学习时使用的文献包括斯坦福CS229，Rich Sutton和Andrew Barto合著的Reinforcement Learning一书，以及一系列近期的论文。相关的文档和完整代码可以在Albert的Github repo上找到，希望这份笔记可以在大家学习强化学习时起到帮助。

所以，什么是强化学习？

Albert和很多人一样，很喜欢玩游戏，他是一个星际选手。每次玩星际争霸的时候，他需要观察屏幕上的游戏情况，使用鼠标和键盘，控制最多上百个单位，进行移动，建造，攻击等操作。虽然对人类来说很自然，结合天赋和刻苦的训练，一个人在天梯最终总可以成为韩宗，但是打一个像星际争霸这样的游戏对于电脑来说可不是简单的事儿。在打游戏时，除了简单的操作和攻击，还要思考一些长期的目标，比如：对方狗我了吗？在扩展经济和进行骚扰之间如何抉择？以现在观察到的建筑，对方会空投我吗？如果空投了我要怎么防守？

星际游戏可能非常复杂，尤其对于只能看到一个个像素的电脑来说

传统的教电脑打游戏的策略是基于规则的，比如让星际专家来hardcode电脑的策略：如果看到运输机，那就要防守空投。这种方法的问题是，在极其复杂的情况下，很多条件是相互

关联的，并且这种方法太过依赖专家的知识，而很难发展出AI自己的智能。

而强化学习的思路是：让AI自己与环境进行trial-and-error的尝试，发展出自己的策略，逐渐达成接近以致超越人类的智能。一个强化学习问题只包含三个要素：机器感知，动作，和目标。

强化学习基本要素

除了最基础的环境和玩家之外，一个强化学习问题有四个基本要素：

1. 策略 (policy)
2. 奖励函数 (reward function)
3. 价值函数 (value function)
4. (可选择的)环境模型 (model)

这些基本要素将在我们介绍马尔可夫过程时详细介绍。简单来说，策略是在获得当前的状态时选择做出的动作; 奖励函数是在当前状态下做出动作时，系统环境所给予玩家的奖励; 而价值函数，则是从现在直到游戏结束时，所有奖励函数的加权累积。如果要做一个类比的话，奖励函数就像“愉悦”或是“不爽”的心情，而价值函数则像是长期获得的满足感或是挫败感。如果我现在开始每天喝酒打牌，虽然短期的奖励函数值会很高，但长远来看却可能因为考试挂科而非常伤心，价值函数的值很低。

现代的强化学习认为，对于价值函数的估计是一个极其重要的内容，我们将在后来的笔记中看到这一点。

下一期：马尔可夫过程和一些最简单的强化学习算法

私货时间：[Please follow and star my github repo!]