强化学习（一）

关键词：
#Evaluative_feedback
#Instructive_feedback
#greedy_selection
#softmax_selection
#n-armed_Bandit_Problem

强化学习最基本的建模

强化学习系统可以观察环境，并采取特定的行为，使得获得的奖励信号的**反馈**达到最大值。
换句话说，强化学习系统 ‘想要’ something，并且可以调整自己的行为策略，使得自己可以获得 ‘想要’ 的东西。

另一个特点是，系统通过不停地与环境交互，实现行为策略的调整。这种思想起源于对人类学习的思考，人类在婴幼儿时期，还无法理解语言的时候，正是通过不停与周围环境交互学习的。这是来源于**实践**的学习。而随着人类成长，建立起语言信号和万事万物的联系后（一个值得思考的问题是，人类如何建立语言信号与实际事物的联系），人类还可以通过语言学习（也就是从其他人的实践中学习）。

强化学习与监督学习的比较

监督学习会有明确的正确（correct）的行为（action）作为已知的信息。
即我们会有已知的 $(环境，正确的行为)$ 对（pair）。

我们希望构建一个映射 $f:环境 \rightarrow 行为$，将特定的环境映射到正确的行为上。

而强化学习中，我们并不会被告知 类似的 对（pair），我们的目的同样是构建一个映射 f，但是由于缺乏监督信息，我们构建映射的难度要大得多，因为我们需要在 行为空间（space of actions）中进行搜索，找到相对其他行为来说最好的行为。

N-Armed Bandit Problem

我们将先从最最简化的一种问题开始，逐步加深对强化学习的理解。

$我们有两个最基本的假设，一是行为空间中包含n个独立的、离散的行为，二是行为的好坏（value$

- 如果反馈的奖励信号是固定的
  那么，智能体（代理，$Agent$）只需要将 n 个行为都尝试一遍，找到最好的行为即可。
  （我们假设当智能体尝试一个行为后，可以获得反馈的奖励信号，奖励信号越高，意味着行为的价值也就越高）
- 如果反馈的奖励信号带有一定的随机性
  那么，上面的方法将不再适用，因为更高的奖励信号可能是由价值更低的行为带来的

（行为的价值是一种平均意义上的价值，即价值越高的行为带来的奖励信号的均值也就越高。），仅仅将 n 个行为都尝试一遍无法消除随机性。

一个朴素的思想是尝试很多次（每个行为不止尝试一次），并将某行为对应的所有的奖励信号取均值，作为该行为的价值。

$$Q_t(a) = \frac{r_1 + r_2 + r_3 + \ldots + r_{k_a}}{k_a}$$

上式中，a 即某个行为，$Q_t(a)$ 是 在经历 t 轮尝试后得到的 a 的价值，$r_1, r_2, \ldots, r_{k_a}$ 为这 t 轮尝试中，尝试行为 a 得到的奖励信号。

Evaluative feedback & Instructive feedback

评估性反馈（Evaluative feedback）常被用于强化学习中，反馈信号不能反映行为是否是最好（correct）的，只会评估行为的好坏。
指令式反馈（Instructive feedback）常被用于监督学习中，反馈信号会反映行为是否是最好（correct）的。

举个例子，当强化学习系统采取一个行为后，会从环境得到一个**标量**奖励信号，比如：3.7，0.1， 100等。系统没办法从标量的奖励信号中判断自己刚刚采取的行为是不是最好（correct）的，但是可以通过数值的大小感知这个行为的优劣。
而监督学习系统采取一个行为后，会得到一个二元奖励信号（如：0/1），其中一元（如 1）表示行为最好，另一元（如 0）表示行为并非最好。

Action Selection Policy

我们上面提到，如果反馈的奖励信号带有一定的随机性，那么我们需要尝试很多次才能很好地评估行为的价值。

一个自然的问题是，我们怎么决定 每次尝试 系统采取的行为。下面是一些常用的方法。

- Greedy Selection：
  每尝试一次行为后，更新行为的价值。在所有行为中，选择价值最大的进行下一次尝试。这种方法可能导致真正有价值的行为由于刚开始给予的反馈太小（随机性），一直没有再次尝试的机会。
- $\epsilon-$ Greedy Selection：
  每次尝试有 ϵ 的概率随机选择一个行为进行尝试。有 1 - ϵ 的概率选择当前最有价值的行为进行尝试。
- Softmax Selection：
  每次根据所有行为的价值，通过 softmax 操作求出概率，按概率分布采样。