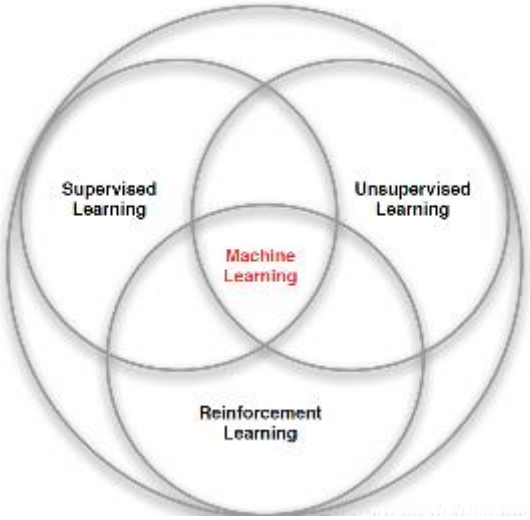
强化学习笔记——总述

自从 AlphaGO 战胜了世界围棋冠军李世石，人工智能迎来了有一个热潮，很多学者、学生趋之若鹜的学习研究人工智能。随之而来的很多专业性名词如机器学习、监督式学习、非监督式学习、深度学习、强化学习等。但是网上能找到的资料还是比较有限，故在此记录一下我在强化学习方面的学习过程，供大家参考。

强化学习基本概念

机器学习可以分为三大类：监督式学习(Supervised Learning)、非监督式学习(Unsupervised Learning)和强化学习(Reinforcement Learning)。三者关系如图所示：



监督学习（supervised learning）：通过已有的训练样本（即已知数据以及其对应的输出）来训练，从而得到一个最优模型，再利用这个模型将所有新的数据样本映射为相应的输出结果，对输出结果进行简单的判断从而实现分类的目的，那么这个最优模型也就具有了对未知数据进行分类的能力。监督学习中只要输入样本集,机器就可以从中推演出制定目标变量的可能结果.如协同过滤推荐算法,通过对训练集进行监督学习,并对测试集进行预测,从而达到预测的目的.

无监督学习（unsupervised learning）：我们事先没有任何训练数据样本，需要直接对数据进行建模。

强化学习（reinforcement learning）：强化学习就是学习采取怎样的动作去获取最大奖赏。学习者不会告诉说列奖赏。反复试验法和延迟奖赏是区别强化学习的最主要的两个特征。

强化学习区别于监督式学习，强化学习的 trial-error 模式要求 agent 去探索环境，然后对状态进行评估（evaluate），在每一个状态下 agent 可以选择多种 action，每次选择的依据可以是贪婪或者 softmax 等，但是得到的 reward 是无法表明当前的选择是正确的还是错误的，得到的只是一个 score，监督学习的 labels 可以给 agent 简洁明了的正确答案（correct or wrong），并且在 agent 在对环境充分的探索前即在每一种状态下选择的每个 action 的次数不够多时，无法充分求 expect，并且在 action 之间也无法进行对比择优。但是当监督学习的 label 信息有噪声干扰或者是利用一些 active learning 获得到的 labels 的时候，强化学习的 agent 与环境直接交互获取到的信息是更加可靠。

强化学习同样区别于非监督式学习，虽然强化学习和非监督式学习一样不依赖于正确的行为样例，但强化学习却是尝试最大化奖赏而不是像非监督式学习那样找出数据的隐藏结构特征。在强化学习中通过 Agent 的经验揭晓数据的结构特征对我们的探索是有利，但是它并不能解决强化学的奖赏最大化问题。因此，我们把强化学习并作三大机器学习之一。

强化学习的关键问题是探索（exploration）和利用（exploitation）的平衡问题。为了获取最大奖赏，agent 必须从已有的奖赏集中选择具有最大奖赏的动作，但为了发现具有最大奖赏值的动作，又必须尝试没有选择过的动作。因此，为了得到奖赏，Agent 必须利用已有的经验，同时又需要探索来发现更好的动作选择。令人难过的是，单单采用探索或者只有利用都会导致任务失败。Agent 必须尝试各种动作，并且渐渐趋近于那些表现好的动作。

强化学习的基本元素

在 agent 和 environment 下，强化学习的四个要素是：策略（policy）、奖赏（reward）、值函数（value function）和可选的环境模型（model）。

`Agent`:表示一个具备行为能力的物体，比如机器人，无人车，人等等

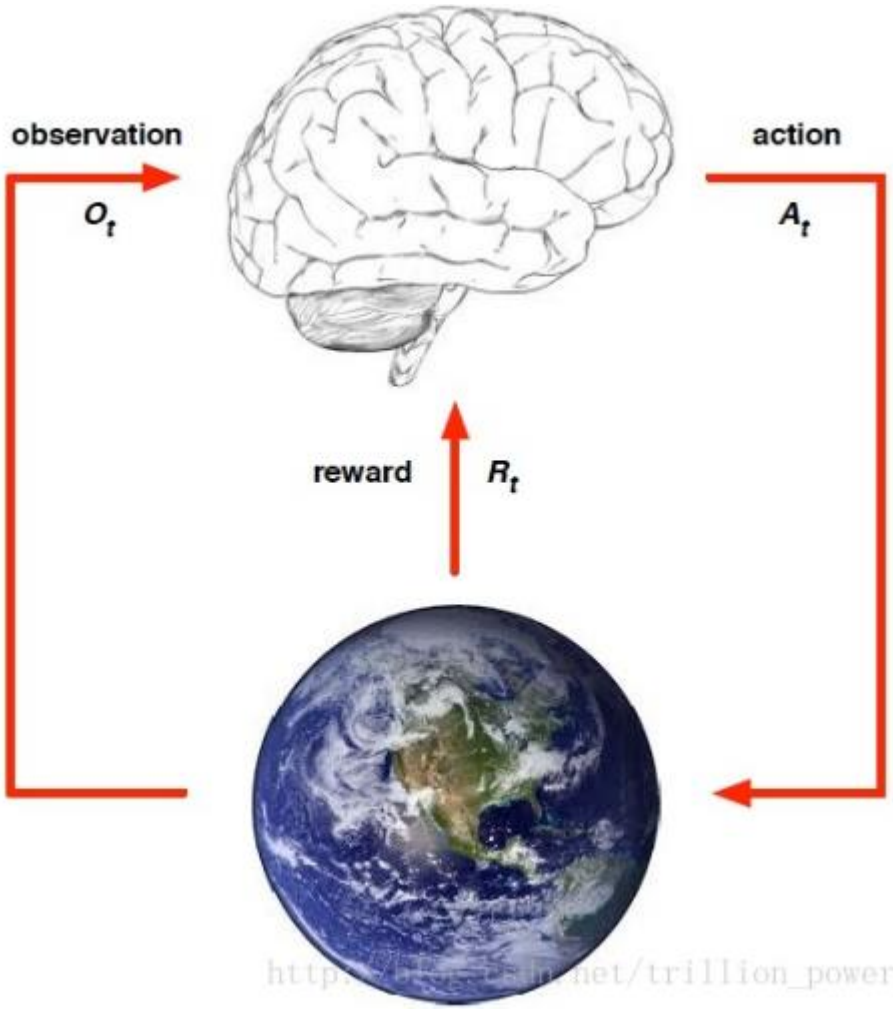
`Environment`：表示除 agent 之外的环境，不依赖于 agent 的一切物体。

`策略（policy）`：定义了学习 agent 在给定时间内的行为方式，简单的说，一个策略就是从环境感知的状态到这些状态中可采用动作的一个映射。对应心理学中被称为刺激-反应的规则或联系的一个集合。

`奖赏（reward）`：定义了强化学习的目标，简单的说，它把环境中感知到的状态（或状态-动作对）映射为单独的一个数值，即为奖赏，表示该状态的可取程度。强化学习 agent 的唯一目标就是最大化在长期运作过程中收获的总奖赏。奖赏函数定义了对 agent 来说什么事好和坏的动作。

`值函数（value function）`：因为强化学习基本上可以总结为通过最大化 reward 来得到一个最优策略。但是如果只是瞬时 reward 最大会导致每次都只会从动作空间选择 reward 最大的那个动作，这样就变成了最简单的贪心策略(Greedy policy)，所以为了很好地刻画

是包括未来的当前 reward 值最大（即使从当前时刻开始一直到状态达到目标的总 reward 最大)。因此就使用了值函数（value function）来描述这一变量。



上面这张图（来自 David Silver）可以很清楚的看到整个交互过程。事实上，这就是人与环境交互的一种模型化表示。在每个时间点 time-step Agent 都会从可以选择的动作集合 A 中选择一个 a_t 执行.这个动作集合可以是连续的比如机器人的控制也可以是离散的比如游戏中的几个按键。动作集合的数量将直接影响整个任务的求解难度，因此 DeepMind 才从玩最简单的游戏做起。

总的来说，强化学习解决过程就是针对一个具体问题得到一个最优的策略（policy)，使得在该策略下获得的总奖赏（total reward）最大。所谓的 policy 其实就是一系列 action。

下一次我们将通过 K 臂赌博机问题来系统的了解强化学习。