强化学习笔记

强化学习要解决的是序贯决策问题，它不关心输入长什么样，只关心当前输入下应该采取什么动作才能实现最终的目标，再次强调当前采取什么动作与最终的目标有关。

强化学习和监督学习的共同点是两者都需要大量的数据进行训练，但是两者需要的数据类型不同。监督学习需要的是多样化的标签数据，强化学习需要的是带有回报的交互数据。监督学习的数据独立于算法本身，而强化学习不同，强化学习的数据是智能体与环境的交互数据，在交互中智能体逐渐的改善行为，产生更好的数据，从而学会技能。因此，相比较于监督学习只构建一个学习算法，强化学习还要构建一个用于与智能体进行交互的环境。

强化学习的特点：

* 没有监督数据，只有奖励信号
* 奖励信号不一定是实时的，而很有可能是延后的，有时甚至延后很多。
* 时间序列是一个重要因素
* 当前的行为影响到后续接收到的数据

强化学习中的一些概念：

* **奖励(reward)**：是信号的反馈，是一个标量，他反应个体在t时刻做的怎么样。个体的工作就是最大化累计奖励。**强化学习主要基于这样的“奖励假设”：所有问题解决的目标都可以被描述为最大化累积奖励。**
* **序列决策(Sequential Decision Making)：目标—选择一定的行为序列以最大化未来的总体奖励。**这些行为可能是一个长期的序列，奖励可能而且通常是延迟的，有时候宁愿牺牲即时（短期）的奖励以获取更多的长期奖励。
* **个体和环境（Agent & Environment）：可以从个体和环境两个方面来描述强化学习问题。在t时刻，**

**个体可以**

* + **1.有一个对于环境的观察评估**
  + **2.做出一个行为**
  + **3.从环境得到一个奖励信号**

**环境可以：**

* **1.接收个体的动作**
* **2.更新环境信息，同时使得个体可以得到下一个观测**
* **3.给个体一个奖励信号**
* **历史和状态(History & State)：**
  + **历史：是观测、行为、奖励的序列**
  + **状态：状态是所有决定将来的已有信息，是关于历史的一个函数：**
  + **环境状态：是环境的私有呈现，包括环境用来决定下一个观测/奖励的所有数据，通常对个体并不完全可见，也就是个体有时候并不知道环境状态的所有细节。即使有时候环境状态对个体可以是完全可见的，这些信息也可能包含着一些无关信息。**
  + **个体状态：是个体的内部呈现，包括个体可以使用的、决定未来动作的所有信息。个体状态是强化学习算法可以利用的信息，它可以是历史的一个函数：**
  + **信息状态：包括历史上所有有用的信息，又称Markov状态。**
* **马尔可夫属性(Markov Property)：一个状态是马尔可夫的，当且仅当：**

**也就是说，如果信息状态是克制的，那么所有的历史信息都可以丢掉，仅需要t时刻的信息状态就可以了。**

* **完全可观测的环境Fully Observable Environments**

**个体能够直接观测到环境状态，在这种条件下：**

**个体对环境的观测 = 个体状态 = 环境状态**

* 部分可观测的环境Partially Observable Environments

个体间接观测环境，比如一个扑克牌玩家只能看到自己的牌和其他已经出过的牌，而不知道整个环境（包括对手的牌）的状态。在这种情况下个体状态 环境状态

**强化学习个体的主要组成部分：**

**强化学习中的个体可以由以下三个组成部分中的一个或者多个组成：**

* **策略Policy：策略是决定个体行为的机制，是从状态到行为的一个映射，可以是确定性的，也可以是不确定性的。**
* **价值函数Value Function：是一个未来奖励的预测，用来评价当前状态的好坏程度。当面对两个不同状态时，个体可以用一个Value值来评估这两个状态可能获得的最终奖励区别，继而指导选择不同的行为，即制定不同的策略。同时，一个价值函数是基于某一个特定的策略的，不同的策略下同一状态的价值并不相同。某一策略下的价值函数用下式表示：**

**这里先不对其进行具体解释。**

* **模型Model：个体对环境的一个建模。它体现了个体是如何思考环境运行机制的，个体希望模型能模拟环境与个体的交互机制。**

**模型至少要解决两个问题：一个是状态转化概率。即预测下一个可能状态发生的概率：**

**另一个工作是预测可能获得的即时奖励：**

**模型并不是构建一个个体所必需的，很多强化学习算法中个体并不试图依赖/构建一个模型。**

**强化学习个体的分类**

**解决强化学习问题，个体可以有多种工具组合，比如通过建立对状态的价值的估计来解决问题，或者通过直接建立对策略的估计来解决问题，这些都是个体可以使用的工具箱里的工具。因此，根据个体内包含的“工具”进行分类，可以把个体分为三类：**

1. **仅基于价值函数的Valued Based：在这样的个体中，有对状态的价值估计函数，但是没有直接的策略函数，策略函数由价值函数间接得到。**
2. **仅直接基于策略的Policy Based：这样的个体中行为直接由策略函数产生，个体并不维护一个对各状态价值的估计函数。**
3. **演员-评判家形式Actor-Critic：个体既有价值函数，也有策略函数，两者相互结合解决问题。**

**此外，根据个体在解决强化学习问题时是否建立一个对环境的模型，可以将它分为两类：**

1. **不基于模型的个体：这类个体并不试图了解环境如何工作，而仅聚焦于价值或策略函数。**
2. **基于模型的个体：个体尝试建立一个描述环境运作过程的模型，以此来指导价值函数或策略函数的更新。**

**探索和利用Exploration & Exploitation**

**强化学习类似于一个试错的学习，个体需要从其与环境的交互中发现一个好的策略，同时又不至于在试错的过程中丢失太多的奖励。探索和利用是个体进行决策时需要平衡的两个方面。**

**一个形象的比方是，当你去一个餐馆吃饭，“探索”意味着你对尝试新餐厅感兴趣，很可能会去一家以前没有去过的新餐厅体验，“利用”则意味着你就在以往吃过的餐厅中挑一家比较喜欢的，而不去尝试以前没去过的餐厅。这两种做法通常是矛盾的，但对解决强化学习问题又都非常重要。**

**预测和控制Prediction & Control**

**在强化学习中，我们经常需要先解决关于预测(prediction)的问题，然后在此基础上解决关于控制(Control)的问题。**

* **预测：给定一个策略，评价未来，可以看成是求解在给定策略下的价值函数的过程。**
* **控制：找到一个好的策略来最大化未来的奖励。**