Introduction

在强化学习中我们考虑的是在不需要有明确的指导下如何通过历史数据(经验)去学习如何行动。一个强化学习的 agent 必须同它的世界交互，并且学习如何最大化长期的累积奖励，其实也是为了学习好的决策序列。

近年来强化学习变得越来越流行，可能是因为该学科取得了很大的进度，例如 Deep-Q-Network。更多的，在人工智能的其他领域看到了大量借鉴和利用强化学习概念而取得成功的案例。例如，在围棋领域 alphaGO 使用强化学习方法达到了超越人类的境界，并且强化学习的概念也经常被借鉴在训练生成对抗网络中(Generative Adversarial Networks)。

在人工智能和机器学习领域的基本挑战是在不确定的情况下去学习如何做出好的决策。

很多人经常想要知道强化学习和其他类型的学习有什么不同。在监督学习中我们会有一份数据集，数据集由样本集合和标签集合组成。在监督学习的设定中，我们会有一份训练集合，对于每一个样本我们都会给出正确的标签(分类问题)或者正确的输出值(回归问题)。与监督学习不同，无监督学习的数据集没有提供标签，无监督学习提供的方法是要找在某些数据之中潜在的结构。比起做出预测，强化学习需要即时的处理决策问题同时比较并能够被采用的行动。一个强化学习 agent 能够和世界进行交互，并且每次都应该即时收到部分的反馈信号，一般将其称为 Reward(奖励)。假设 agent 所执行的行动实际上是"best"的，而 agent 是不能给出任何最好的迹象(条件，信号)，并且 agent 必须以某种方式学会去选择该最优动作，其最终目的是选择最优动作以使得长期累积的奖励值最大。因此，由于通过 reward 信号的反馈是弱/不完整(weak/incomplete)，我们可以考虑强化学习是位于监督学习(带有标签数据的强反馈)和无监督学习(无标签的 0 反馈)之间的。

我们需要克服强化学习设置的一系列的挑战并可能需要在之间做出权衡。agent 必须能够去选择最优的行动以最大化它所接收的奖励信号。然而随着 agent 需要同环境交互学习，探索未知的环境又是必须的，很自然的必须要在探索和利用之间做出平衡，agent 需要在发现新的更好的策略以及获得较低奖励的风险之间做出决定，或者就是利用目前已知的最好的策略。另一个我们需要面对的问题是 agent 是否能泛化它的的经验?也就是说它是否能在未知状态学会一些动作的好坏?(That is,can it learn whether some actions are good/bad in previously unseen states?)最后，我们需要去考虑 agent 动作的延迟序列,它是指,如果 agent 收到一个高额的奖励,那么造成这个奖励的原因是它当下所执行的行动还是由于它更早之

前所执行的动作?

2 Overview of reinforcement learning

2.1 Sequential Decision Making

通常我们考虑 RL 问题是做出一系列好的决策问题。将其形式化：动作的集合:$actions\{a_t\}$,观察序列:$observations\{o_t\}$,奖励序列:$rewards\ \{r_t\}$,定义在 t 时刻的历史序列:$h_t=(a_1,o_1,r_1,...,a_t,o_t,r_t)$,agent 选择下一次的行动可以被看为是历史的函数：$a_{t+1}=f(h_t)$。序列决策问题可以被认为通过定义和近似计算函数 f 来解决。

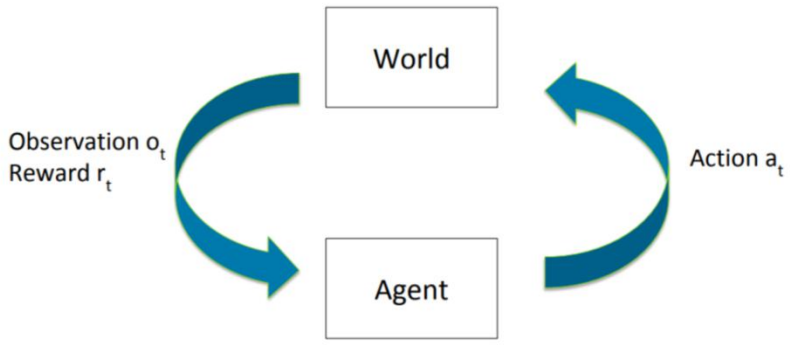


Figure 1: Overview of how an agent interacts with its world.

当序列决策过程的模型是已知的情况下，一些确定的和有限设定的 AI 技术例如 A*搜索和 minimax 算法能够被用于寻找最优的动作序列。

然而，当我们有一个很大状态空间(可能是无穷的)或者将随机性引入我们的模型环境中，就算是暴力搜索也变得不可能了。在这样的设定中，有必要结合泛化来使得任务可行，一个有效的例子是 Atrai,在后续中会介绍。撇开暴力求解不谈，agent 必须在短期和长期的奖励中做出战略性决策并选取适当的动作。

2.2 Modeling the world

设 S 为我们真实世界可能存在的状态集合，设$\{s_t\}$为观察到的状态序列,以时间 t 为其索引，设 A 为可能的动作集合。通常我们想要了解的世界动态转移函数$P(s_{t+1}\mid s_t,a_t,...,s_1,a_1)$，其实是在 S 之上的概率分布，它是一个描述之前状态和动作的函数。在强化学习中，我们通常假设其具有马尔可夫属性：

$$P(s_{t+1}\mid s_t,a_t,...,s_1,a_1)=P(s_{t+1}\mid s_t,a_t)$$

在实际中对我们来说它是非常灵活的。使得马尔可夫属性成立的一个有用的技巧是使用历史h_t作为我们的状态。

通常我们认为奖励r_t是在状态$s_t \xrightarrow{a_t} s_{t+1}$时收到的。一个奖励函数通常被用于预测奖励：$R(s,a,s') = E[r_t \mid s_t = s, a_t = a, s_{t+1} = s']$，我们通常认为奖励函数由$R(s) = E[r_t \mid s_t = s]$ or $R(s,a) = E[r_t \mid s_t = s, a_t = a]$组成。也有退化为在给定$s_t = s$，$r_t$是一个固定值：$r_t \mid s_t = s$

一个模型通常由上述的动态转移函数和奖励函数组成。

2.3 Components of an reinforcement learning agent

首先，假设 agent 状态为一个历史的函数：$s_t^a = g(h_t)$。一个 RL agent 一般有明确的下述一个或多个定义：**policy,value function,optionally a model**。

policy：π是状态到动作的映射，也就是输出的是 action,所以$\pi(s_t^a) \in A$,有时也被描述为在动作空间之上的随机分布：$\pi(a_t \mid s_t^a)$。当 agent 想要获得一个行动同时π是随机的，它将会根据概率$P(a_t = a) = \pi(a \mid s_t^a)$去选取一个动作$a \in A$。

确定的 policy:$\pi(s) = a$

随机的 policy:$\pi(a \mid s) = \Pr(a_t = a \mid s_t = s)$

value function： Feture rewards from being in a state and / or action when following a particular policy,值函数是一个 state 的奖励值(这个奖励是包含未来奖励的，future reward)，或者是一个跟随固定策略的动作的奖励值(同样包含未来奖励)，给定一个 policy π，折扣因子$\gamma \in [0,1]$,那么值函数V^π可以被描述为总折扣奖励的期望值：

$$V^\pi(s) = E_\pi[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + ... \mid s_t = s]$$

E_π表示在当前已遭遇的状态 s 之上而采用 policy π而产生的长期累积奖励的期望值，而折扣因子γ被用于描述即时奖励和延迟奖励的权重。

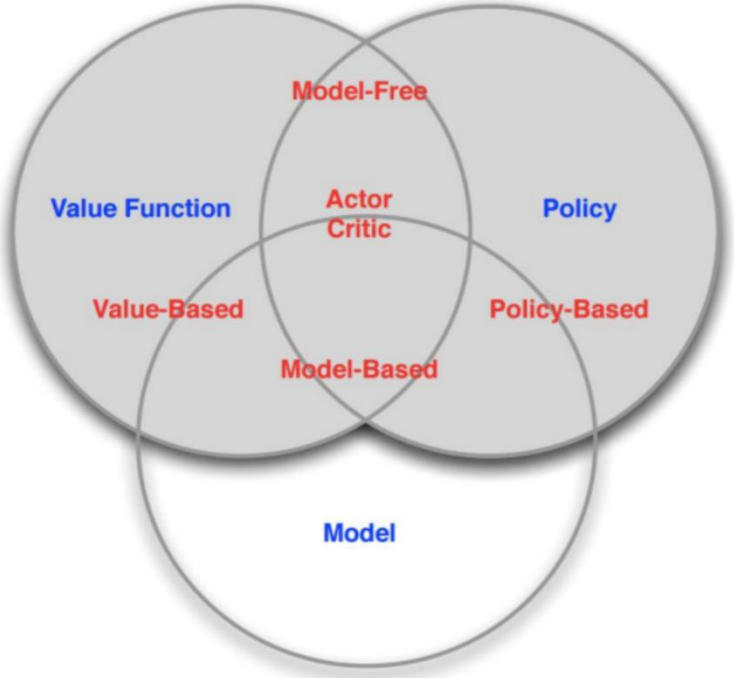
最后 RL agent 可能会有一个在 2.2 小节中描述的模型。如果 agent 有一个模型，我们将其称为 model-based agent，如果 agent 没有包含模型，我们称其为 model-free agent。

目前为止，我们已经讨论了 RL 基本设定的定义，但我们没有做出任何关于o_t和s_t的关系的假设。当$o_t \neq s_t$时，我们称其为部分可观察，对于 RL 算法而言在共同的可观察的部分去保持一个基于真实世界的概率分布，定义为s_t^a,它被认为是可信任的状态。

然而对于这门课程的主要任务，我们将会假设全可观察情况，$o_t = s_t$，我们将会假设$s_t^a = s_t$。

2.4 Taxonomy of reinforcement learning agents

Agent type	Policy	Value Function	Model
Value Based	Implicit	✓	?
Policy Based	✓	X	?
Actor Critic	✓	✓	?
Model Based	?	?	✓
Model Free	?	?	X



2.5 continue domains

简单而言,我们仅关注在离散的时间步骤之下的离散状态和行动空间。然而还是有很多应用，尤其是 Robotics and Control，它是最适合在基于连续状态和行动空间的基础上建模的，具有连续的时间。上述的讨论可以被泛化到连续的设定之中。

2.6 Markov Assumption

状态 s_t 具有 markov 当且仅当： $p(s_{t+1} \mid s_t, a_t) = p(s_{t+1} \mid h_t, a_t)$，意味着将来和过

去是相互独立的。但实际中没法用 Markov 来建模，因为往往无法抛弃历史来分析。那么为什么 Markov Assumption 受到欢迎?

因为它总是可以做出假设：$s_t = h_t$，并且在实践中也经常假设观察到的数据是有效的历史统计，即：$s_t = o_t$

2.7 Reference CH1

强化学习的两个主要特点：

①试错搜索，强化学习不同于监督学习那样通过打上正确或者错误的标签来告知采用怎样的行动，它是通过在“尝试-错误”过程中学习正确的行为，从而发现最优的策略。

②延迟奖励，在大部分有挑战，有意义的环境中，某一步动作可能不仅仅是影响当前的局面，它也有可能对未来的局面造成深远意义的影响,以至于对全局的奖励均造成影响,所以决策要充分的考虑长期的奖励和眼前的利益。

③信用分配,要确定某次奖励(高奖励,低奖励?)与之前那些动作有较为直接的关系，是一个比较难的问题。

强化学习也和无监督学习不同,无监督学习是想要去发现无标签数据集的隐藏结构，强化学习并不依赖训练样本的正确性，它不是去发现隐藏的结构，而是想要去获取最大化的奖励。因此，强化学习是除了监督学习和无监督学习外的第三类机器学习。

强化学习的基本挑战：

1.RL 系统中状态的表示。

2.通过泛化可以训练状态的转移。

3.利用时间信用分配以确定哪些行动对结果很重要。

4.探索不是最优的状态和动作。