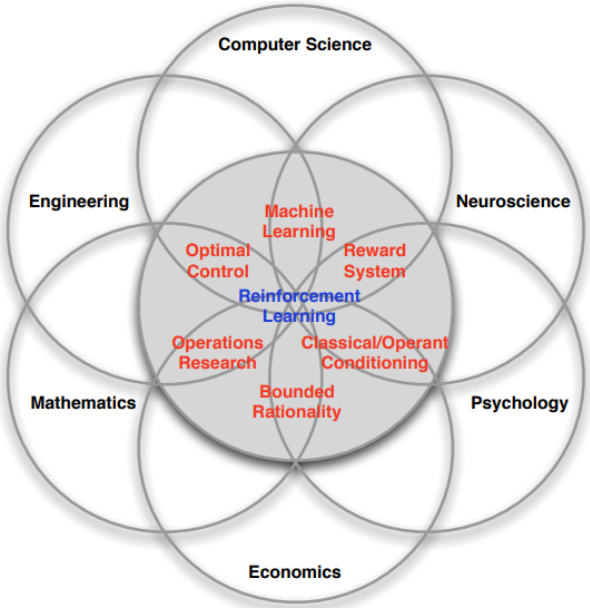
第一讲 强化学习介绍

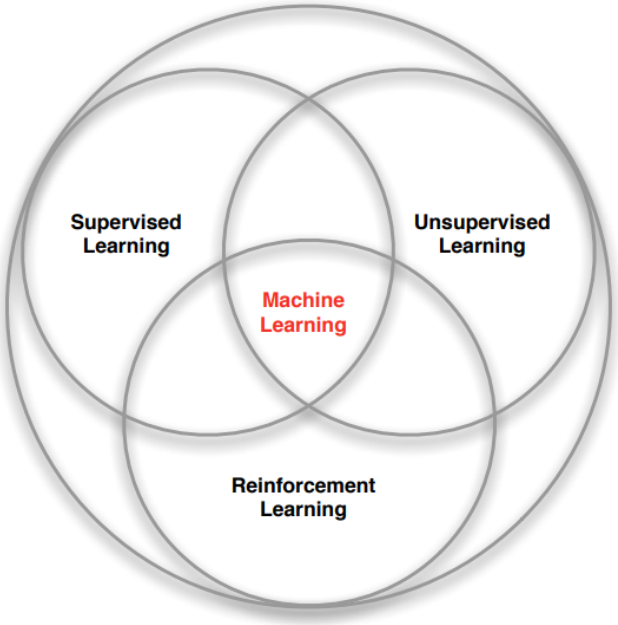
1.1 强化学习简介

学习的本质是什么？人类是通过和环境进行交互来学习的，如婴儿通过观察环境、与环境交互等方法进行学习。强化学习是一种最接近于人类的学习。强化学习的基本思想是通过试错（Trail-and-Error）来学习。试错学习分为两步：首先是智能体（agent）通过与特定的环境进行交互，观查结果，例如获得奖励或惩罚；然后，记住与特定环境交互的结果，得到自己的优化策略。优化策略的目标就是最大化奖励，即尽可能获得更多的奖励，获得更少的惩罚，这很像动物的趋利避害行为。所以，强化学习是一种以目标为导向，通过交互进行学习的学习方式。

强化学习是由神经科学、心里学、优化控制等学科发展而来的一门科学。下图描述了强化学习和各学科的关系。



强化学习是机器学习的一个分支，下图描述了强化学习和机器学习的关系。



强化学习的特点

与其他机器学习相比，强化学习有以下特点：

（1）强化学习没有监督数据，只有一个奖励信号。

（2）反馈是有延迟的，即延迟奖励，为了最大化奖励，可能牺牲当前奖励。

（3）时间序列是很重要的因素。

（4）当前的行为影响后续收到的数据。

1.2 强化学习的基本概念

1.2.1 智能体和环境

在强化学习中，学习者和决策者统称为智能体（agent)。除了智能体自身外，智能体打交道的任何东西都可以称为环境（environment）。例如在自动驾驶中，自动驾驶车辆称为智能体，其学习驾驶策略并执行学到的驾驶策略；除了自动驾驶车辆之外的其他东西称为环境。

1.2.2 奖励（Rewards)

奖励R_t 是一个反馈信号，它是一个标量，衡量智能体在t 时刻表现的好坏程度。智能体的工作就是最大化累计奖励（cumulative reward）。强化学习有一个最基本的假设，即奖励假设。奖励假设是指所有的学习目标可以被描述为最大化期望累计奖励。奖励的例子：

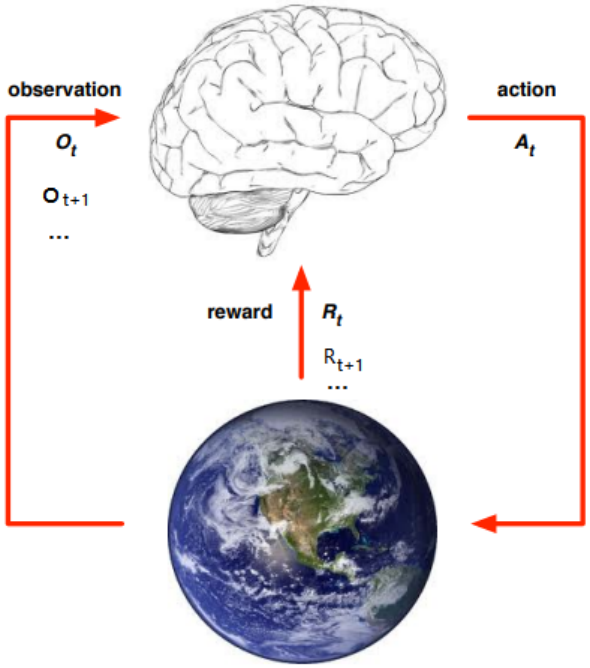
（1）炒股票：盈利为正奖励，亏损为负奖励

（2）玩游戏：赢了为正奖励，输了为负奖励

注：负奖励可以理解为惩罚，数学上通常描述为负数，所以称为负奖励。

1.2.3 智能体与环境的交互过程

强化学习是智能体与环境不断交互的过程。如下图所示，在t时刻，智能体从环境中得到观测值O_t和标量奖励R_t，执行动作A_t；环境接收到智能体的动作A_t后,更新信息，发出下一时刻的观测值O_{t+1}和下一个时刻的奖励R_{t+1}。在$t+1$时刻，智能体接收到O_{t+1}和R_{t+1}后，执行动作A_{t+1}。这种交互过程产生了一个时间序列：O_t， R_t， A_t， $O_{t+1}, R_{t+1}, A_{t+1} \ldots$。



> 注：
> 1.时间步长是在环境侧增加的
> 2.智能体只能通过动作对环境进行控制

1.2.4 序列决策（Sequential Decision Making）

智能体学习的目的是通过采取一系列的动作来最大化未来获得的总体奖励。所以，动作可能对结果有着长期的影响。奖励可能是有延迟的，有时候可能牺牲短期的奖励，获得长期的奖励。如，国内互联网的烧钱模式，前期大量的补贴客户，前期一直亏钱，当客户习惯于产品后，就能获得长期的收益。前期补贴客户的动作对后期客户习惯于使用产品产生了影响。

1.2.5 状态（state）

定义：在时刻t，历史是由观测、动作、奖励组成的一个时间系列，这个系列是从初始时刻到t时刻，即

$$H_t = O_1, R_1, A_1, \ldots, A_{t-1}, O_t, R_t$$

下一时刻会发生什么依赖于历史：

- 智能体当前采取什么动作依赖于它之前得到的历史

- 环境当前采取什么观测、奖励也依赖于历史

由于状态（state）是决定将来会发生什么的已有信息。因此，状态是一个关于历史的函数：

$$S_t = f(H_t)$$

环境状态S_t^e

- 环境状态是对环境的私有描述
- 环境状态是环境用来决定下一个观测、奖励的所有数据
- 对智能体来说，环境状态通常不是完全可见的
- 即使S_t^e对智能体完全可见，也可能包含一些无关信息

智能体状态S_t^a

- 智能体状态是智能体对自己内部的描述
- 智能体状态包含了智能体用来选取下一个动作的所有信息
- 智能体状态是强化学习算法可以利用的信息
- 它是关于历史的一个函数: $S_t^a = f(H_t)$

信息状态（Information State)

一个信息状态（又称Markov状态）包含了历史中的所有有用信息。
定义：
一个状态S_t具有马尔可夫性，当且仅当：

$$P[S_{t+1}|S_t] = P[S_{t+1}|S_1, \ldots, S_t]$$

从上述定义可知，马尔可夫状态具有以下性质：

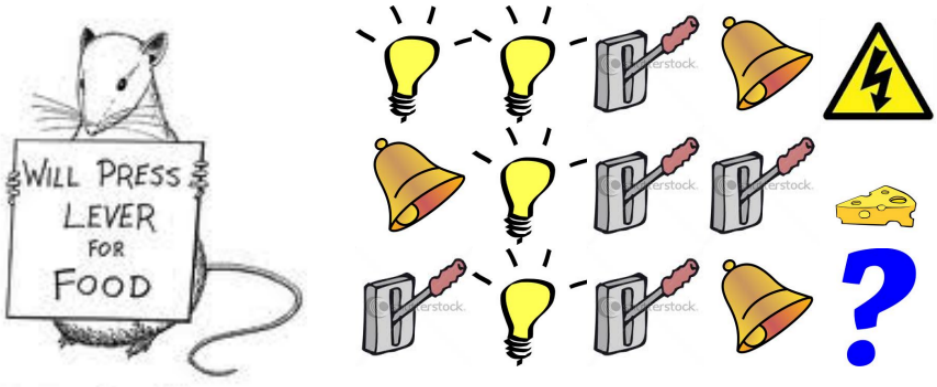
- 给定当前状态时，未来与过去无关，即知道了当前状态后，就不需要知道之前的状态。

$$H_{1:t} \to S_t \to H_{t:+\infty}$$

  一旦知道了状态，就不需要历史信息了。例如，环境状体S_t^e是马尔可夫的，因为，其包含了环境决定下一个观测/奖励的所有信息；同样，历史H_t也是马尔可夫的。

例1-1：
对小白鼠（智能体）做一序列4个动作的实验，分别有下面三个实验，其中前两个实验，最后的事件分

别是老鼠遭电击和获得一块奶酪，现在请分析比较这三个实验序列的特点，分析第三个实验序列中，老鼠是获得电击还是奶酪？



假如智能体状态 = 实验中的最后三个动作（不包括电击、获得奶酪，下同）。那么实验3的结果会是什么？（答案是：电击）

假如智能体状态 = 亮灯、响铃和拉电闸各自事件发生的次数。那么实验3的结果又是什么？（奶酪）

假如智能体状态 = 完整的实验序列。那结果又是什么？（未知）

环境是完全可观测的（fully observed）

状态是对环境的完整描述，包含了环境的所有信息。当智能体能够观测到环境的所有信息时，即智能体对环境的观测O_t 等于环境的状态S_t^e时，我们称这个环境是**完全可观测**。这种智能体对环境完全可观测的模型，数学上可以描述成**马尔可夫决策过程（Markov decision process，MDP）**。此时，智能体对环境的观测 = 智能体状态 = 环境状态，即：

$$O_t = S_t^a = S_t^e$$

环境是部分可观测（ Partially Observable Environments）

当智能体只能观测到环境的部分信息时，我们称环境是**部分可观测**。数学上可以描述成**部分可观测马尔可夫决策过程（partially observable Markov decision process, POMDP）**。此时，智能体的状态不等于环境状态，也不等于智能体的观测,即：

$$O_t \neq S_t^a \neq S_t^e$$

因此，智能体必须构建自己的状态。智能体可以用完整的历史来构建自己的状态，$S_t^a = H_t$，但这种方法比较原始。也可以用其他方法，如：

- Beliefs of environment state：虽然智能体不知道环境的状态，但智能体可以利用已有的经验数据，用智能体已知的各种环境状态的概率分布作为当前时刻的智能体状态的表示：

$$S_t^a = (P[S_t^e = s^1], \dots, P[S_t^e = s^n])$$

- Recurrent neural network（RNN）：不需要知道环境的概率，只需当前智能体状态以及当前时刻智能体的观测，输入循环神经网络(RNN)中，得到一个当前智能体状态的表示：

$$S_t^a = S_{t-1}^a W_s + O_t W_o$$

例1-2：
下图中，左边豹子在追羚羊，能观测到环境的全部信息，这时环境是完全可观测的，即$o_t = s_t$。 而右边一辆车挡住了豹子，只能看到了羚羊和车，只看到了环境的部分信息，这时环境是部分可观测的，智能体的观测不等于环境状态，即 $o_t \neq s_t$。

小结

本小节讲述了多种状态，其中的重点是理解信息状态（马尔可夫状态）的性质，即与历史无关，只与当前状态有关。另一个重点是，完全客观测时（MDP）：智能体对环境的观测 = 智能体状态 = 环境状态，即观测和状态相等；而部分可观测时（POMDP）：观测不等于状态。注意区分这两点。

1.3 智能体的组成部分

强化学习的智能体由以下三个组成部分中的一个或多个组成：

- **策略（Policy）**
  策略是从状态到动作的一个映射，智能体根据策略来选取动作。
- **价值函数 （Value Function）**
  用价值函数来评估当前状态的好坏程度。
- **模型（Model）**
  智能体对环境的建模，即对环境的动力学进行建模。

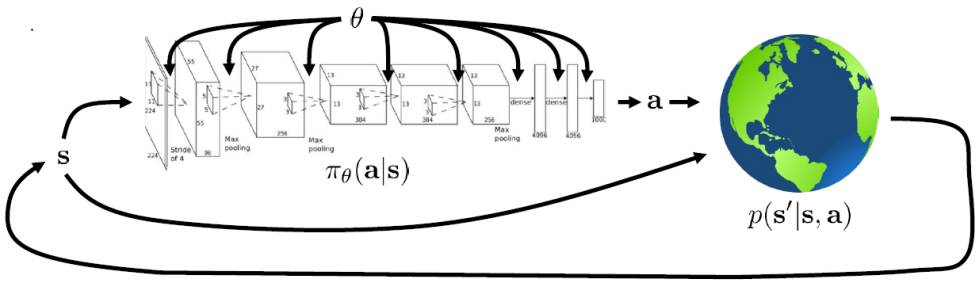
下面对这三个组成部分做详细介绍。

1.3.1 策略

策略就是告诉智能体在每个状态下该如何行动，即智能体在当前状态下根据策略来选取的动作，所以策略是从状态到动作的一个映射。强化学习的最终目标就是学到一个策略，告诉智能体该如何行动，在强化学习中策略一般用π来表示。策略分为：确定策略和随机策略。

- 确定策略（Deterministic policy）：$a = \pi(s)$， 处于状态s时，智能体选取的动作a是确定的。
- 随机策略（Stochastic policy）：$\pi\ (a|s) = P[A_t = a|S_t = s]$，处于状态s时，智能体以一定概率选取动作a。随机策略$\pi\ (a|s)$是智能体处于状态s时，关于选取动作a的概率分布。我们通常把随机策略表示成$\pi_\theta(a|s)$，其中θ是策略的参数。在深度强化学习中，θ就是神经网络的权值（weights）。

例1-3
假如用深度学习神经网络来表示策略$\pi_\theta(a|s)$，θ是神经网络的权值。如下图所示：



状态s输入到神经网络，神经网络输出动作a，然后s, a进入到环境中，环境根据动力学模型$P(s'|s,a)$输出下一个状态s'，虽然通常情况下，我们不知道$P(s'|s,a)$。

1.3.2 价值函数

价值函数$V(s)$定义为：在状态s时，对未来获得最终奖励的一个预测。我们用它来评估状态的好坏，当面对两个不同的状态时，智能体可以用价值函数来评估这两个状态可能获得的最终奖励区别，继而指导智能体选择不同的行为，即采取不同的策略。同时，一个价值函数是基于某一个特定策略的，同一状态不同策略下的价值也不一定相同。某一策略下的价值函数定义为：

$$V^\pi(s) = E_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots | S_t = s]$$

$V^\pi(s)$是从状态s开始，未来获得所有奖励的加权，加权系数γ将在下一讲讨论。

1.3.3 模型

智能体对环境的一个建模，即智能体对环境动力学进行建模。智能体希望这个模型能够完全模拟环境，即智能体执行动作a_t后，这个模型能够像环境一样给出，s_{t+1}和r_{t+1}。因此模型由两部组成：

- 状态转移概率，用来预测下一个状态

$$P^a_{ss'} = P(S_{t+1} = s'|S_t = s, A_t = a)$$

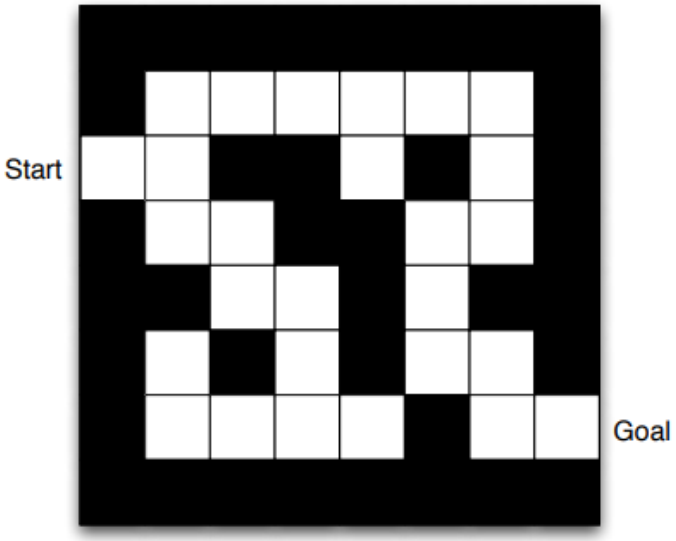
- 奖励函数，用来预测下一个（即时）奖励

$$R^a_s = E[R_{t+1}|S_t = s, A_t = a]$$

> 注：
>
> - 只有在基于模型（Model Based）的强化学习算法中,才需要构建模型。很多强化学习算法不需要模型。
> - 环境实际运行机制称为环境动力学，模型是智能体对环境动力学的建模。

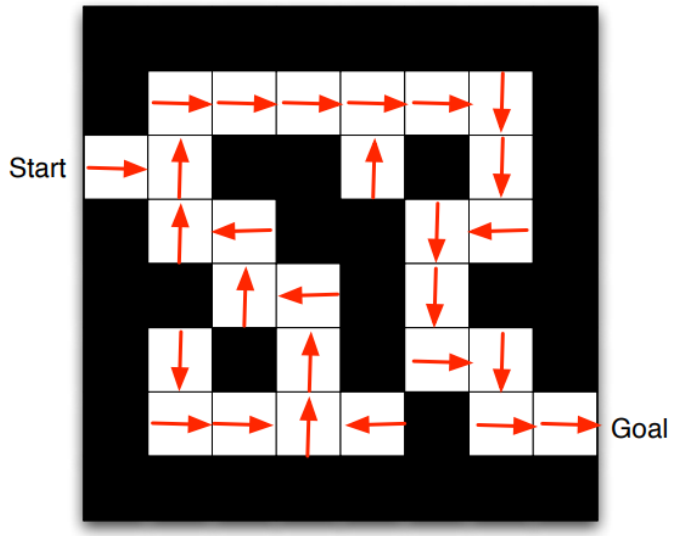
例1-4
如下图所示，智能体走迷宫，智能体从起点（start）开始，到达目标点（Goal)。

- 动作: 上，下，左，右
- 奖励: 每走一步为-1
- 状态: 智能体当前位置

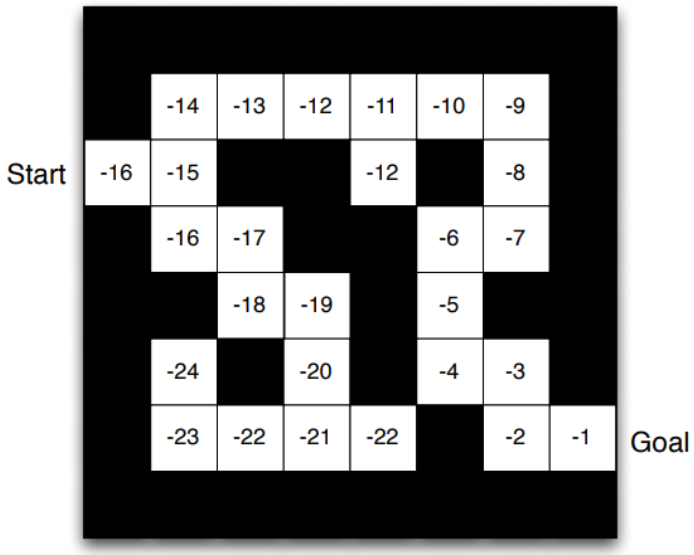


我们可以从两种强化学习的方法来看智能体怎么选取动作，即基于策略（policy based）的方法和基于价值的方法（value based）。

- **基于策略（policy based）的方法:**
  下图中的箭头表示每个状态 s的策略 $\pi(s)$。所以，每一个状态都有一个最优策略。比如在开始位置时的策略是向右走，第二个格子位置是向上走。按着箭头的方向就能得到从起点到终点的最优策略。



- **基于价值（value based）的方法:**
- 下图中的数字表示每个状态 s的价值$V_\pi(s)$。所以，每个状态都有一个价值，我们的目标就是朝着价值越来越大的方向走。比如开始是-16，向右走一步为-15，第二个格子可以向上和向下走，价值分别是-14和-16，因此选择向上走。朝着价值增加的方向就能最快走到终点。



由上面例子可知，在智能体学习的过程中,智能体通常会由策略、价值函数、模型这三个部分中一个或多个部分组成。智能体通过与环境不断地交互的过程中，得到经验数据，从中学习到策略、模型或价值；然后利用学到的策略继续和环境交互，产生新的数据，进一步优化策略、模型或价值。智能体不断地学习，直到逼近最优解，如上例中的箭头和数字，就是最优策略和最优价值。

1.3.4 智能体分类

在上例中，智能体可以学习策略或者价值来求解强化学习的问题。因此，根据智能体的学习内容，我们可以把智能体分为如下三类：

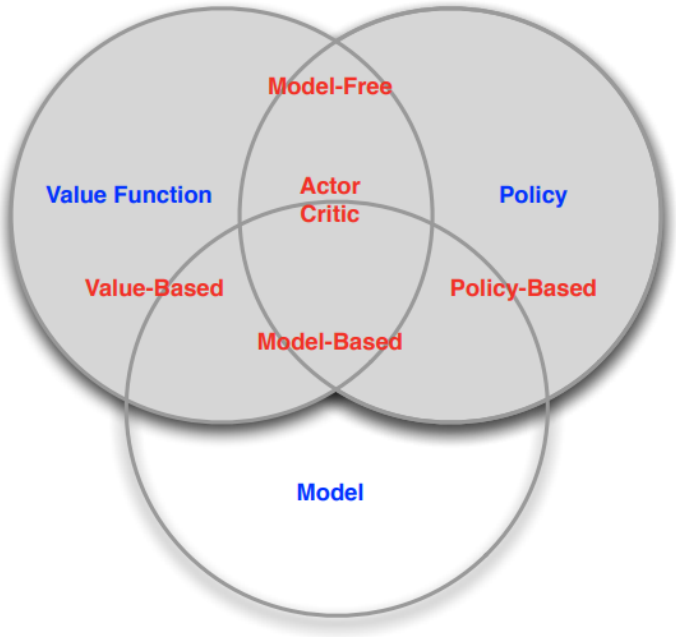
- **基于策略（policy based）的智能体** 直接学习策略$\pi(a|s)$，不需要学习价值函数。
- **基于价值（value based）的智能体** 学习价值函数，通过价值函数隐式地得到策略。

- **演员-评论家（Actor Critic）的智能体** 是基于策略和基于价值的结合，既学习策略，也学习价值函数。

根据智能体是否需要对环境动力学进行建模，可以把智能体分为如下两类：

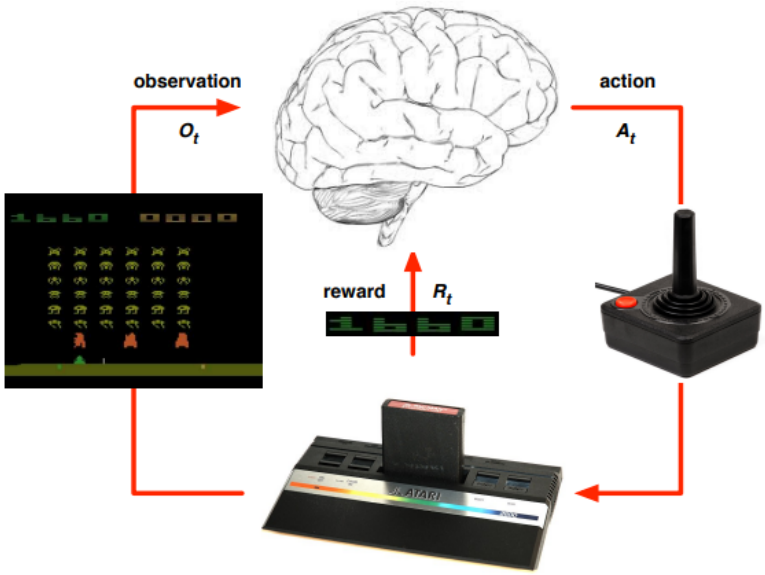
- **基于模型（model based）的智能体** 通过对环境进行建模，以此来学习策略或价值函数。
- **不于模型（model free）的智能体** 不需要对环境建模，通过学习价值函数和策略函数进行决策。

强化学习智能体的分类如下图所示：



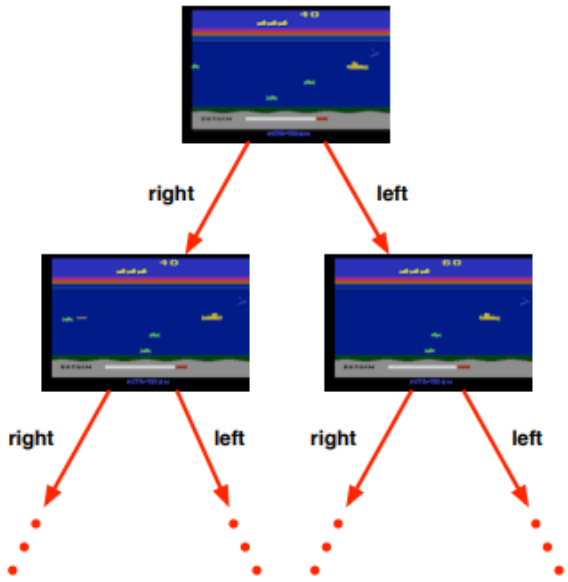
1.4 学习和规划（learning and planning）

学习和规划是序列决策的两个基本问题。
在强化学习中， 初始时环境是未知的，智能体不知道环境的动力学模型，智能体通过与环境不断地进行交互，逐渐改善其行为策略。如下图智能体玩Atari例子中，智能体不知道游戏规则，智能体直接通过玩游戏来交互来学习，选择游戏杆执行的动作，观察屏幕上的像素值和分数，不断优化策略。



在规划中， 环境如何工作（环境动力学模型）对于智能体是已知或近似已知的，智能体并不与环境发生实际的交互，而是利用其构建的模型进行计算，从而优化其策略。如下图的例子中，智能体知道游戏规则，即智能体大脑中有完美的环境模型，它不需要与环境交互，可以在智能体大脑内模拟整个决策过程，如在状态s时，如果执行动作$a = left$，下一个状态会是什么？得分是多少？智能体都能计算出来。

在强化学习中，一个解决问题的常见思路是，智能体先学习环境如何工作，也就是了解环境的动力学模型，即学习得到一个模型，然后利用这个模型进行规划。



1.5 探索和利用（Exploration and Exploitation）

强化学习的智能体的学习过程中，探索和利用是需要考虑的一对矛盾。

- **探索**是指智能体去探索未知环境，通过尝试未知的动作来得到更优的策略，但也可能会得到一个更差的策略。
- **利用**是指智能体不去尝试新的动作，而是利用已知的能带来最大奖励的动作，但也可能错失能带来更大奖励的新动作。

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。这两种做法通常是一对矛盾，但对解决强化学习问题又都非常重要。

在智能体刚开始学习的时候，智能体并没有策略，所以它只能通过随机地探索来学习，当探索的未知的动作越来越多时，智能体的策略不断地改善，智能体随之要增加利用，减少探索。 继续上面吃饭的例子，一条街上你吃过10%餐厅时，应该继续去尝试没吃过的餐厅，你能吃到更好吃的餐厅的概率很大；而当街上90%餐厅你吃过时，去吃你已知最好吃的餐厅，很大概率是这条街最好吃的餐厅。

强化学习类似于一个试错的学习，智能体需要从其与环境的交互中学习到一个好的策略，同时又不至于在试错的过程中丢失太多的奖励。探索和利用是智能体在学习过程中需要平衡的一对矛盾。

参考：

1.David Silver第一课
2.叶强《David Silver强化学习公开课中文讲解及实践》
3.CS285 第四课