初探强化学习

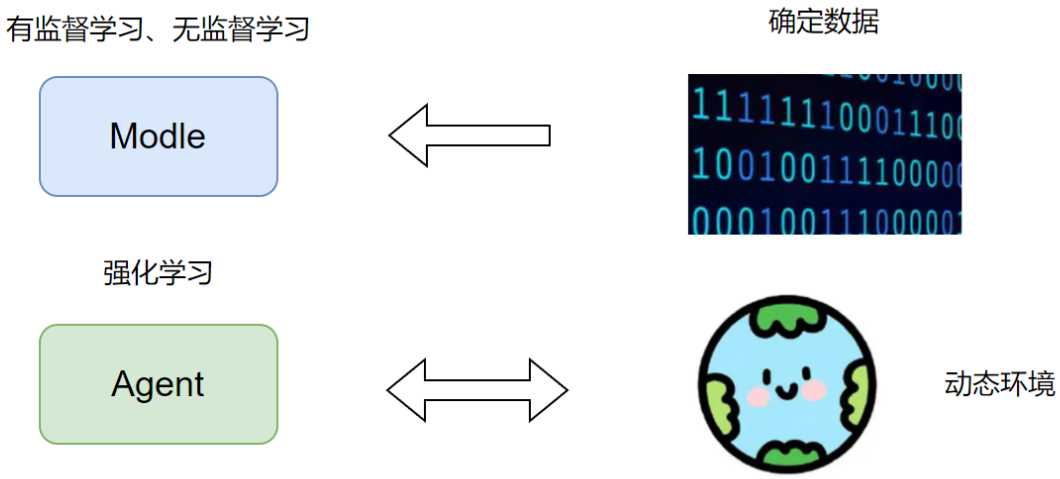
强化学习简介

机器学习可以分为两种类型，预测型和决策型

预测型任务简单的来说可以根据训练样本是否有label，可以分为**有监督学习**和**无监督学习**，这是最简单直接的区别。概括的说有监督学习通过训练样本来预测值，而无监督学习常用来进行分类，它在学习时并不知道其分类结果是否正确（因为没有告诉它什么是正确的）....扯远了，今天的主角是强化学习

决策型任务在动态环境中采取行动，在不断变化的状态中进行感知和决策获得即时奖励，随着时间的推移最大化累计奖励从而确定最优的决策，这就是**强化学习**

预测型任务和决策型任务的区别在于对环境是否存在交互：



决策下达到环境中，就会直接改变环境，所以未来发展也随之改变，做了决策就要担责任

而预测根据数据仅产生预测信号、不考虑环境的改变，预测信号如何使用是不考虑的，它不会反作用回数据

总之，决策往往会带来“后果”，而预测仅仅产生一个针对输入数据的信号，并期望它和未来可观测到的信号一致，这不会使未来情况发生任何改变。

在机器学习领域，有一类重要的任务和人生选择很相似，**序贯决策**任务，序贯决策是指决策者序贯地做出一个个决策，并接续看到新的观测，直到最终任务结束。实现序贯决策的机器学习方法就是**强化学习**

强化学习定义：

概括的说，强化学习是机器通过与环境交互来实现目标的一种计算方法

交互过程：

机器和环境的一轮交互是指，①机器在环境的一个状态下做一个动作决策，②把这个动作作用到环境当中，③这个环境发生相应的改变并且将相应的奖励反馈和下一轮状态传回机器。

这种交互是迭代进行的，机器的目标是最大化在多轮交互过程中获得的累积奖励的期望。

智能体：

强化学习用**智能体**（agent）这个概念来表示做决策的机器。相比于有监督学习中的“模型”，强化学习中的“智能体”强调机器不但可以感知周围的环境信息，还可以通过做决策来直接改变这个环境，而不只是给出一些预测信号。所以一个智能体要具有三个关键的要素：

- 感知：智能体能够在某种程度上感知环境的状态，从而知道自己所处的现状
- 决策：根据所处的状态智能体决定去实现什么样的**行动**然后在作用于环境
- 奖励：环境根据智能体采取的动作而改变的状态，产生一个标量信号作为奖励反馈

⚠️奖励往往是一个标量，因为多维度的向量无法简单直观的判断出行为的好坏

强化学习系统要素：

- 历史：历史是已经发生过的感知观察、决策行动、奖励的序列

$$H_t = O_1, R_1, A_1, O_2, R_2, A_2, \ldots, O_{t-1}, R_{t-1}, A_{t-1}, O_t, R_t$$

即一直到时间 t 为止的所有可观测变量

所以可以根据历史可以预测接下来发生什么（观史可以知兴替😼），智能体来决定行动

- 状态：状态是一种很关键的要素，它是一种用于确定接下来发生的事情的信息

状态是关于历史的函数

$$S_t = f(H_t)$$

- 策略：策略是智能体在特定时间的行为方式，即根据环境所处状态而决策采取的行动——策略

策略是从状态到行动的映射，分为确定性策略和随机性策略

确定性策略就是一个函数，它根据状态然后来决定策略

$$a = \pi(s)$$

这个函数通常用π()表示，而随机策略从状态s到行为a其实是条件概率分布，环境的下一刻状态的概率分布将由当前状态和智能体的动作来共同决定

$$\pi(a|s) = P(A_t = a \mid S_t = s)$$

- 奖励：奖励是定义强化学习的目标标量，选取标量的原因是让我们智能体能立即感知到什么是“好”的

$$R(s, a)$$

它由状态和行为共同决定，指决策行动后对环境产生的后果是不是“好”的，好的就可以得到奖励

- 价值函数：价值函数是对于未来累积奖励的预测，用于评估在给定的策略下的**状态**好坏

$$Q_\pi(s, a)$$

关键字：**未来**、**累积**、**奖励**

就以下棋为例，在进行下一个决策的时候我们可能需要牺牲一个棋子♟️，很明显这个行为的奖励是糟糕的，但是牺牲掉这个棋子后未来的一步两步或者三步，就能取得胜利，那么想较于当前状态糟糕的奖励就显得微不足道了。所以价值函数就是用来预测长期累积的奖励的

- 环境模型：用于模拟智能体所处环境的模型，常用来根据智能体对环境采取的行动而返回状态或者给予奖励

基于模型的强化学习和模型无关的强化学习的根本区别在于学习过程中有没有环境模型

强化学习的方法分类：

- 基于价值：知道什么是好的什么是坏的
    - 没有策略
    - 价值函数
- 基于策略：知道怎么行动
    - 策略
    - 没有价值函数
- Actor-Critic：学生听老师的
    - 策略
    - 价值函数