强化学习概述

一、概述

- 强化学习（reinforcement learning，RL）讨论的问题是智能体（agent）怎么在复杂、不确定的环 境（environment）里面去最大化它能获得的奖励
-

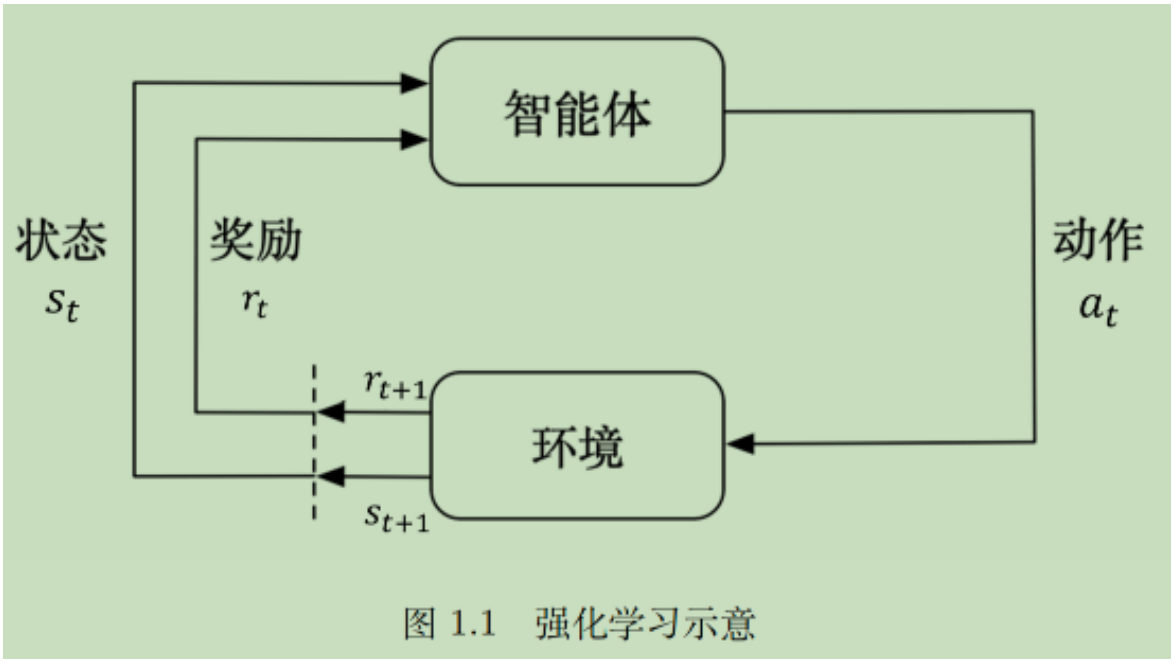


图 1.1　强化学习示意

- 强化学习中的智能体，并不会得到及时的反馈
- 强化学习和监督学习的区别如下
  - 强化学习输入的样本是序列数据
  - 学习器并没有告诉我们每一步正确的动作是什么，需要学习器自己发现哪些动作可以获得更多的奖励
  - 智能体获得能力的过程，是一个试错探索的过程
  - 强化学习的过程中，没有很强烈的监督信号，只有奖励信号，并且是延迟的
- 强化学习的特征
  - 试错探索，理解环境
  - 获得延迟奖励
  - 时间是一个非常重要的因素
  - 智能体的动作会影响它之后得到的数据。在训练智能体的过程中，很多时候我们也是通过正在学习的智能体与环境交互来得到数据的。所以如果在训练过程中，智能体不能保持稳定，就会使我们采集到的数据非常糟糕。我们通过数据来训练智能体，如果数据有问题，整个训练过程就会失败。所以在强化学习里面一个非常重要的问题就是，怎么让智能体的动作一直稳定地提升
  - 强化学习得到的模型可以有超人类的表现，而监督学习的上限是标注数据

二、序列决策

- 奖励是由环境给的一种标量的反馈信号（scalar feedback signal），这种信号可显示智能体在某一步采取某个策略的表现如何。强化学习的目的就是最大化智能体可以获得的奖励，智能体在环境里面存在的目的就是最大化它的期望的累积奖励
- 在一个强化学习环境里面，智能体的目的就是选取一系列的动作来最大化奖励，所以这些选取的动作必须有长期的影响
- 状态和观测有什么关系?
    - **状态**是对世界的完整描述，不会隐藏世界的信息。**观测**是对状态的部分描述，可能会遗漏一些信息
    - 环境有自己的函数 *s**et = f e (H**t)* 来更新状态，在智能体的内部也有一个函数 *s**at = f a (H**t)* 来更新状态。当智能体的状态与环境的状态等价的时候，即当智能体能够观察到环境的所有状态时，我们称这个环境是**完全可观测的（fully observed）**在这种情况下面，强化学习通常被建模成一个**马尔可夫决策过程（Markov decision process，MDP）**的问题
    - 在马尔可夫决策过程中，*o**t = s**et = s**at* 。但是有一种情况是智能体得到的观测并不能包含环境运作的所有状态，因为在强化学习的设定里面，环境的状态才是真正的所有状态。比如智能体在玩 *black jack* 游戏，它能看到的其实是牌面上的牌。或者在玩雅达利游戏的时候，观测到的只是当前电视上面这一帧的信息，我们并没有得到游戏内部里面所有的运作状态。也就是当智能体只能看到部分的观测，我们就称这个环境是**部分可观测的（partially observed）**。在这种情况下，强化学习通常被建模成**部分可观测马尔可夫决策过程（partially observable Markov decision process, POMDP）**的问题
    - 部分可观测马尔可夫决策过程是马尔可夫决策过程的一种泛化。部分可观测马尔可夫决策过程依然具有马尔可夫性质，但是假设智能体无法感知环境的状态，只能知道部分观测值。比如在自动驾驶中，智能体只能感知传感器采集的有限的环境信息
    - 部分可观测马尔可夫决策过程可以用一个七元组描述：(*S, A, T, R, Ω, O, γ*)。其中 *S* 表示状态空间，为隐变量，*A* 为动作空间 T(s'|s, a) 为状态转移概率，*R* 为奖励函数，Ω(o|s, a) 为观测概率，*O* 为观测空间，*γ* 为折扣因子

三、动作空间

- 离散动作空间
- 连续动作空间

四、强化学习智能体的组成成分和类型

组成成分

- 策略（Policy）
    - 智能体的动作模型，将输入状态转变为动作
    - 分类
        - 随机性策略（更优）
            - 输出所有动作的概率
        - 确定性策略
            - 输出一个明确的动作
- 价值函数（Value Function）

    - 用于评估状态的好坏，其中包括一个折扣因子（尽可能短的时间，尽可能多的奖励）
    - 类型
        - $V_\pi(s) \doteq \mathbb{E}_\pi\left[G_t \mid s_t = s\right] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s\right]$，对于所有的$s \in S$
        - $Q_\pi(s,a) \doteq \mathbb{E}_\pi\left[G_t \mid s_t = s, a_t = a\right] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s, a_t = a\right]$
- 模型（Model）
    - 模型决定了下一步的状态，使用当前的状态和当前的动作，对下一步状态的预估，由状态转移矩阵和奖励函数两部分组成
    - 状态转移矩阵
        - $p_{ss'}^a = p\left(s_{t+1} = s' \mid s_t = s, a_t = a\right)$
        - $R(s,a) = \mathbb{E}\left[r_{t+1} \mid s_t = s, a_t = a\right]$

智能体类型

- 学习目标分类法
    - 基于价值的智能体（Value Based Agent）
        - 显式地学习价值函数，隐式学习策略
        - 应用在离散环境下（围棋或某些游戏领域），对于动作集合庞大，如机器人操作，效果不好
        - Q Learning / Sarsa 算法
    - 基于策略的智能体（Policy Based Agent）
        - 直接学习策略
        - Policy Gradient 算法
    - 结合上面两个，就有了「演员-评论员智能体」（Actor Critic Agent）
- 模型分类法
    - 有模型强化学习智能体（Model Based）
        - 学习状态的转移（创建一个虚拟世界来理解真实世界）
        - 这种方法好，但却非常难

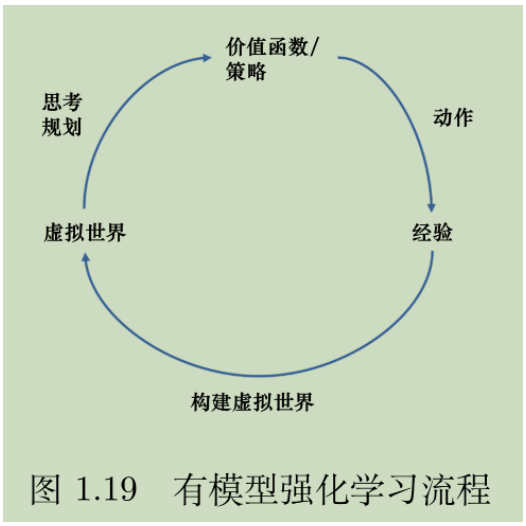
    - 

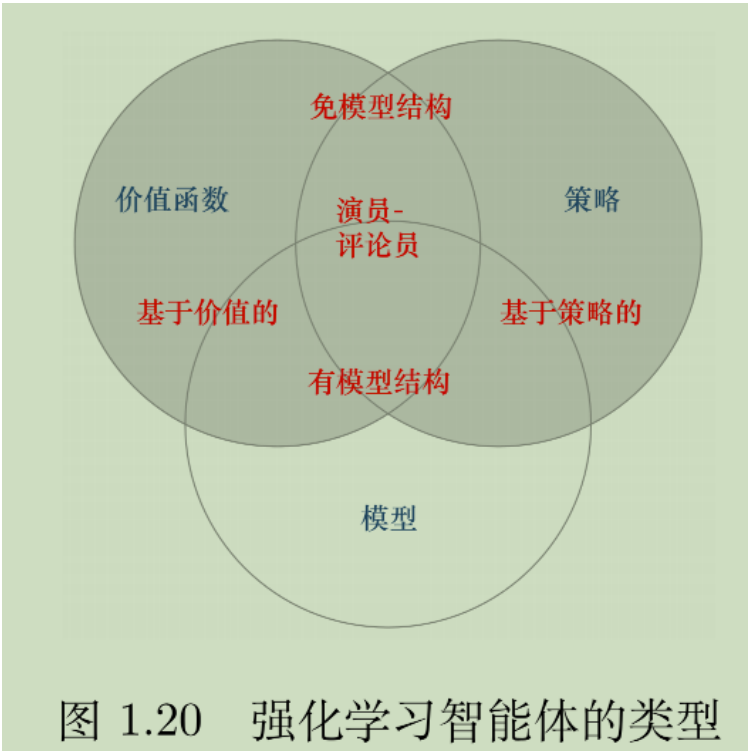
图 1.19　有模型强化学习流程
  - 无模型强化学习智能体（Model Free）
    - 没有学习状态转移，不需要对真实环境建模
    - 属于数据驱动型方法，需要大量采样进行估计状态、动作和奖励函数，从而优化动作策略
    - 这种方法泛化能力优于上面的方法，但上面的方法可以缓解此方法的数据匮乏问题
    - 这种方法是主流的方法，更简单、资料丰富，AlphaGo 就是这种方法
- 综合
  - 

图 1.20　强化学习智能体的类型

五、学习和规划

- 学习和规划是序列决策的两个基本问题

六、探索和利用

- 强化学习里，探索和利用是两个很核心的问题

- 探索
  - 优化策略
- 利用
  - 尝试新的动作

七、强化学习轮子

- Gym
  - OpenAI 开源的一个环境仿真库
    - 离散控制环境
    - 连续控制环境

八、关键定义

- 强化学习（reinforcement learning，RL）
  - 智能体可以在与复杂且不确定的环境进行交互时，尝试使所获得的奖励最大化的算法
- 动作（action）
  - 环境接收到的智能体基于当前状态的输出
- 状态（state）
  - 智能体从环境中获取的状态
- 奖励（reward）
  - 智能体从环境中获取的反馈信号，这个信号指定了智能体在某一步采取了某个策略以后是否得到奖励，以及奖励的大小
- 探索（exploration）
  - 在当前的情况下，继续尝试新的动作。其有可能得到更高的奖励，也有可能一无所有
- 开发（exploitation）
  - 在当前的情况下，继续尝试已知的可以获得最大奖励的过程，即选择重复执行当前动作
- 深度强化学习（deep reinforcement learning）
  - 不需要手动设计特征，仅需要输入状态就可以让系统直接输出动作的一个端到端（end-to-end）的强化学习方法。通常使用神经网络来拟合价值函数（valuefunction）或者策略网络（policy network）
- 全部可观测（full observability）、完全可观测（fully observed）和部分可观测（partially observed）
  - 当智能体的状态与环境的状态等价时，我们就称这个环境是全部可观测的
  - 当智能体能够观察到环境的所有状态时，我们称这个环境是完全可观测的
  - 一般智能体不能观察到环境的所有状态时，我们称这个环境是部分可观测的
- 部分可观测马尔可夫决策过程（partially observable Markov decision process，POMDP）
  - 即马尔可夫决策过程的泛化。部分可观测马尔可夫决策过程依然具有马尔可夫性质，但是其假设智能体无法感知环境的状态，能知道部分观测值
- 动作空间（action space）、离散动作空间（discrete action space）和连续动作空间（continuous action space）
  - 在给定的环境中，有效动作的集合被称为动作空间，智能体的动作数量有限的动作空间称为离散动作空

间，反之，则被称为连续动作空间

- 基于策略的（policy-based）
  - 智能体会制定一套动作策略，即确定在给定状态下需要采取何种动作，并根据这个策略进行操作。强化学习算法直接对策略进行优化，使制定的策略能够获得最大的奖励
- 基于价值的（valued-based）
  - 智能体不需要制定显式的策略，它维护一个价值表格或者价值函数，并通过这个价值表格或价值函数来执行使得价值最大化的动作
- 有模型（model-based）结构
  - 智能体通过学习状态的转移来进行决策
- 免模型（model-free）结构
  - 智能体没有直接估计状态的转移，也没有得到环境的具体转移变量，它通过学习价值函数或者策略网络进行决策