前言

强化学习，作为一门由计算机科学、神经科学、数学、心理学等多领域融合而成的学科，现在已经被DeepMind认为是实现AI(Artificial Intelligence)的关键技术。最为世人熟知的AlphaGo也是基于强化学习的方法，才实现了在Game of Go上对于人类的彻底超越。作为机器学习的一个特殊分支，熟悉机器学习的人应该很容易发现他和传统监督学习、非监督学习的区别，有学者甚至直接认为二者是强化学习的特殊情况。我的理解是刻意去找他们之间的关联并没有太多意义，而是需要理解每一种学习方式所适合的场景。对于强化学习，他解决的就是带有目标的决策问题，我们希望训练一个智能体来代替人在某些可以交互的环境下执行一系列决策，以获得最大回报。

现实中的一些例子

- 直升飞机特技动作:考虑智能体控制直升机控制器的多个按钮组合来完成一些特技动作
- 在Atari游戏中取得更高分：Atari提供了一系列非常适合训练强化学习算法的环境,DeepMind公司在几十个不同游戏上都训练出“超级玩家”
- 自动驾驶、机器人控制：在自然环境下，控制机械设备执行正确的操作，是AI的一个热门领域
- …

强化学习问题的基本概念

奖励（reward）

关于概念本身无需多言，比较有意思的是这三点：

- 奖励假说（Reward Hypothesis）：所有的目标都可以通过最大化期望聚集奖励值来描述
- 奖励的值：有的问题智能体在每一步时都会获得一个非零奖励，比如说自动驾驶；有的可能在任务结束时才可以给出非零的奖励，比如说围棋比赛。
- 无论如何，我们考虑的都不仅仅只是每一步的回报，而是最大化整个未来回报的期望。

环境和状态 （Environment and State)

环境是智能体所能感觉到的事物组成的一个整体，状态的概念感觉十分抽象，目前的理解是对于智能体决策有用的那一部分信息可以作为智能体所处环境的状态。

Environment State、Agent State、Observation
如何理解这三个概念的区别？如果把环境理解为客观物质世界（或者一些规则设定的场景），那么我理解的环境状态指的就是当前时刻客观世界中那些与任务相关的信息；Observation指的就是一个智能体在当前时刻观察环境能得到的信息，能得到通常意味着可能不能获得所有信息；智能体状态我目前的理解是智能体基于他已有的信息来产生自己对环境状态的理解，当然这种理解可能与真实情况并不完全一致。

Information State
信息状态的定义实际上就是具备马尔可夫性质的状态，这种性质也是RL的基础，我们希望RL问题中状态要满足这个性质。

Fully Observable Environments
完全可观测指的就是智能体可以观测到环境中与任务相关的所有信息，那么

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

我们学习的问题大部分背景都假设在这种性质的环境下

Partially Observable Environments
这种性质的环境我的理解就是S_t^e不能被智能体完全获得，智能体通过观察拿到的信息只是一部分信息。 形式化来说，这是一个POMDP问题。解决这种问题你就需要结合历史信息来定义智能体自己对于环境状态的理解，即S_t^a,然后训练的策略是从这种状态到动作的映射。当然你所定义的智能体状态，也是要满足Markov Property的。

An RL Agent

一个强化学习的智能体主要由以下三个部分组成：

- 策略（Policy）：策略是强化学习优化的对象，策略给出了决策的依据，策略分为确定性策略和随机策略
- 值函数（Value function）：预测了未来的总奖励，给出了状态和动作好坏衡量的定量值
- 模型（model）：一个模型预测环境的动态，即下一个状态和立即奖励值

智能体的分类
(1)

- Value Based：优化一个值函数，通过最优值函数可以推导出最优策略，比较经典的例子有Q-learning
- Policy Based：直接优化一个具体的策略，策略梯度算法已经有TRPO，PPO等一大批算法
- Actor Critic：核心仍然是直接优化一个具体的策略，但是会用值函数作为辅助优化的角色

(2)

- Model Free：通过和环境的交互产生的真实数据来学习最优策略
- Model Based：基于一个已有的环境模型或者是从真实数据中学习到的模型来进行规划

总结

目前的强化学习理论基础还不完善，一些应用也集中在虚拟的电子世界，理论的研究和实践仍然在不断进行中...