强化学习 7 日打卡总结

经过学习，从对强化学习有了进一步的了解。

强化学习是机器学习中的一个领域，强调如何基于环境而行动，以取得最大化的预期利益。核心思想智能体 agent 在环境 environment 中学习，根据环境的状态 state（或观测到的 observation），执行动作 action，并根据环境的反馈 reward（奖励）来指导更好的动作。

主要学习了基于表格型方法求解 RL，基于神经网络方法求解 RL， 基于策略梯度求解 RL，连续动作空间上求解 RL。受益匪浅。

强化学习　监督学习　非监督学习

强化学习来和监督学习最大的区别是它是没有监督学习已经准备好的训练数据输出值的。强化学习只有奖励值，但是这个奖励值和监督学习的输出值不一样，它不是事先给出的，而是延后给出的，比如上面的例子里走路摔倒了才得到大脑的奖励值。同时，强化学习的每一步与时间顺序前后关系紧密。而监督学习的训练数据之间一般都是独立的，没有这种前后的依赖关系。

强化学习和非监督学习的区别。也还是在奖励值这个地方。非监督学习是没有输出值也没有奖励值的，它只有数据特征。同时和监督学习一样，数据之间也都是独立的，没有强化学习这样的前后依赖关系。

强化学习其特点总结为：

1. 没有监督标签。只会对当前状态进行奖惩和打分，其本身并不知道什么样的动作才是最好的。
2. 评价有延迟。往往需要过一段时间，已经走了很多步后才知道当时选择是好是坏。有时候需要牺牲一部分当前利益以最优化未来奖励。
3. 时间顺序性。每次行为都不是独立的数据，每一步都会影响下一步。目标也是如何优化一系列的动作序列以得到更好的结果。即应用场景往往是连续决策问题。
4. 与在线学习相比，强化学习方法可以是在线学习思想的一种实现，但是在线学习的数据流一定是增加的，而强化学习的数据可以做减少（先收集，更新时按丢掉差数据的方向）。而且在线学习对于获得的数据是用完就丢，强化学习是存起来一起作为既往的经验

基本元素元素分别是：

states ，就是节点 {A, B, C, D, E, F}

action ，就是从一点走到下一点 {A -> B, C -> D, etc}

reward function ，就是边上的 cost

policy，就是完成任务的整条路径 {A -> C -> F}

有一种走法是这样的，在 A 时，可以选的 (B, C, D, E)，发现 D 最优，就走到 D，此时，可以选的 (B, C, F)，发现 F 最优，就走到 F，此时完成任务。

这个算法就是强化学习的一种，叫做 epsilon greedy，是一种 Policy based 的方法，当然了这个路径并不是最优的走法。