强化学习

定义：

强化学习定义为 鼓励学习 评价学习 或增强学习

是机器学习的重要组成部分，主要描述过程：

智能体Agent 与 环境Environment 交互过程种 学习 策略(Policy), 最大化总回报 (Rewards)

智能体的策略决定 其每次对于环境状态的动作

环境不仅返回回报 Rewards，同时返回给智能状态

1. 强化学习分为传统强化学习以及深度强化学习

区别为传统强化学习利用的算法例如

动态规划算法，蒙特卡洛算法，时序差分算法

深度强化学习，针对两种强化学习，价值学习value learning 以及 策略学习Policy learning

通过神经网络在某个模块的作用，简化，并强化Agent，策略动作采取

1. 了解其他机器学习算法

监督学习，非监督学习，以及半监督学习

最后一种较为主流的则是强化学习

理解强化学习与其他机器学习算法的区别

1. 强化学习，在与环境的交互过程中学习，数据是不断更新的，实时获取的

并且数据之间是有关联性，并非独立的

1. 然而监督学习以及半监督，无监督，其数据基本不会变化，且是已经固定采取好的

数据之间是独立的，单独存在的

总结两点在数据上：

强化学习：连续性(依赖性)，实时性 (更新)

机器学习：独立性，先验性(都是事先准备好的)

1. 强化学习问题建模

对于任何一个强化学习问题，都应该可以抽象为以下模型

Agent ->action-> Environment

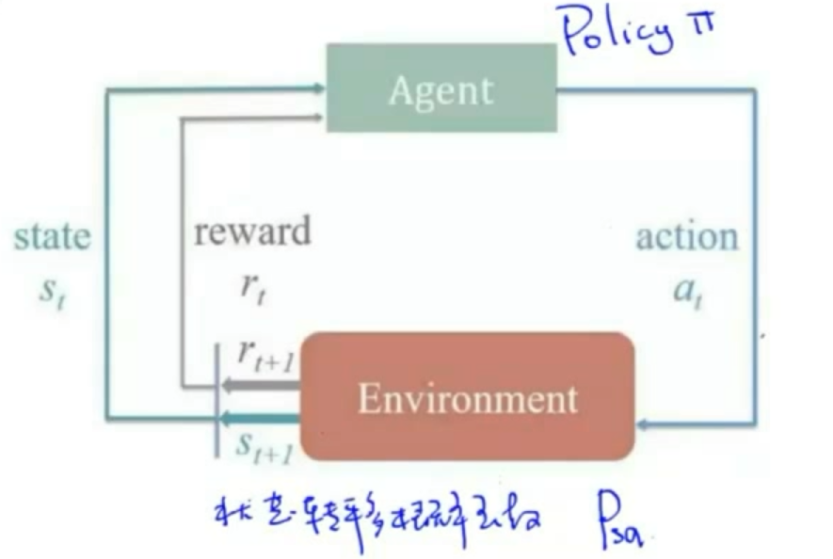
| |

|-----------< State Reward <-----------|

其中Agent 通过策略来选择动作

Environment 则是自身状态转移函数，来选择下一步状态，而Reward 通常是伴随着状态

也就是达到了什么样的状态，才给予什么样的奖励，下图个人理解应该Reward跟在state后面



1. 针对每个元素的具体介绍
2. Agent智能体, 表示强化学习的本体，就是我们需要探讨的个体

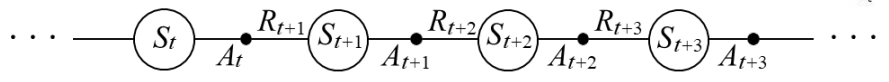
作为学习者与决策者

1. Environment除 智能体 以外的一切，主要指提供状态的环境，相关环境
2. State 状态，一个可以表示环境的数据，能被智能体感知到，并且解析
3. Action 动作 智能体会根据state 进行动作反馈，动作集合A，则表示智能体所有能够进行的动作
4. Reward，奖励，智能体在做出动作后，环境根据状态的改变反馈给Agent的奖励
5. Policy，状态到动作的函数映射，理解为智能体根据这个函数映射，得到状态后决策动作的过程
6. 强化学习交互过程

首先智能体感知环境 si

然后通过策略做出动作 ai

动作导致环境改变，此时返回给 ri+1 与 si+1



1. 强化学习(智能体)的目标

在与环境交互过程中，连续的时间序列中，寻找一种最优策略

该策略能够使得长期积累奖励最大化，总回报 G 最大化

1. 马尔可夫决策过程 (Markov Decision Process, MDP)
2. 马尔可夫性质，这里默认是一阶马尔可夫

在时间t+1步时，环境的反馈仅仅取决于上一个时间状态t的状态s与动作a

与时间t-1步骤以及前面的步骤均无关，没有关联性

这是一种强制 假设，对于许多现实生活问题不是这样的

这种强制性假设的目的是为了简化问题模型

否则所有状态都与 前几个状态相关，那么累加起来，问题变得十分复杂

事实上，这种简化问题的方式仍然能够解决问题，找到问题最优解的，因此没有影响

重点，**所有深度学习问题，都是建立在一阶马尔可夫性质上的**

1. 根据 MDP 以及 强化学习问题建模，进行符号集合总结
2. State 状态集合

S = {s1,s2......sm} 总共有m个不同的状态

但是有时候状态集合是连续的，也即是无穷的，集合表示方式可能是函数式，或更抽象

1. Action动作集合

A = {a1,a2.....an} 表示共有n个不同的动作

同样有时候动作集合是无穷的，表示为函数式，可能用神经网络标识，更抽象

1. 状态转移概率函数

P(s’|s,a) 表示在状态s下，执行动作a，转移到状态s’的概率

实际上很多情况的状态转移概率是固定的，环境给定的

也就是给的那个s与a，那么得到的状态只存在一个为1，当然游戏中的随机过程可能有随机性

比如给定机器人，给出命令向上走，但是由于履带问题，可能20%概率其他方向走

求和为1性质

1. 奖励函数

通常奖励是人们根据抵达的状态提前定义好的

但似乎有种情况是规定 R(s,a)由状态与动作共同决定

表明在状态s时，执行a后，将会给出奖励

比如在对方松懈时，攻击对方，会得到奖励，但我们并没有考虑是否有攻击到对方

但是如果按照更上层理解，你的攻击可能miss，事实上不应该得到奖励

但执行这个动作能够更好地给对方造成伤害，即使会miss

总之两种选择，具体理解

1. 策略函数 Policy

P(at|st)，一般用Π表示策略Π(at|st)

表示在时间t时，获得状态st，然后反馈动作at

具体如何得到at这个动作，就得看如何定义策略，就是如何定义映射函数

可以随机初始化获得策略，也可以使用随即策略

还可以通过神经网络来表示策略，总之找到合适地映射函数即可

目前深度强化学习选择通过价值来见解表示策略，并不是直接表示

1. 折扣因子 decent γ∈[0,1]

=0时，表示只关心后续一个奖励

=1时，表示关心后续所有状态地所有奖励，平等地看待后续所有奖励

介于两者，则是成比例关心，但仍然是越前面的奖励，占比越大

**总结因此，马尔可夫决策过程与强化学习过程结合，其决策过程表示为**

**MDP(S, A, Psa, Rsa,γ)**

1. MDP的流程

第一步，初始化 **(S, A, Psa, Rsa,γ)**

其中 S A Psa 均是给出问题时就已经确立的

然而Rsa，与γ则一般是需要认为定义

查看是否满足终止条件，满足终止条件则直接结束

比如设置时间长短限制，t>=T

或者游戏角色是否死亡 D=True

第二步，根据agent感知到的状态s，然后根据策略函数

at~Π(a|s), 一般情况选定最优可能的at

第三步，根据选定的动作，获得下一步的奖励rt+1与状态st+1

rt+1 = R(st,at) st+1 ~ P(st,at)

然后返回查看是否满足终止条件，如果不满足，则继续循环执行

个人对MDP Markov decision process过程的理解：

实际上是对强化学习执行流程的一个强化版本，结合了马尔可夫性质，并且给定了具体的数学推导符号

比如马尔可夫五元组 MDP(S,A,Rsa,Psa,γ)

给定了具体的状态集，动作集，奖励函数，策略函数，以及decent折扣因子γ

再次表现该过程产出的状态动作奖励序列

S0,a0,r1,s1,a1,r2,s2..................ST-1 aT-1 rT ST

Total Payoff

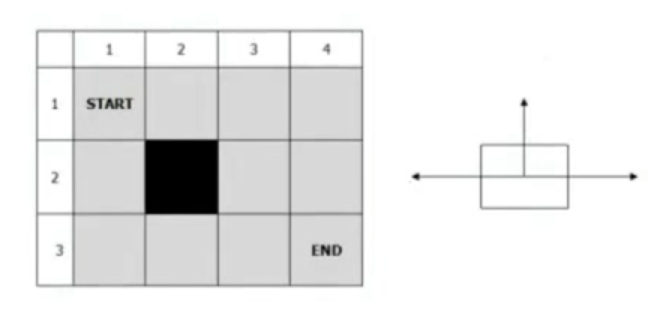
累计奖励，一般称之为回报，实际上是针对某个时刻t的累计奖励

例如，Gt = rt+1 + γrt+2 + γ^2 rt+3 + .... + γ^T-t-1 rT

注意累计奖励，回报是针对某一个时间点的，而并非整个序列，只是我们需要整个序列才能精确获取

机器人行走案例

其中(2,2)为障碍物，机器人碰到边缘不会移动，起始地点为(1,1),结束地点为(3,4)



那么强化学习MDP五元组进行确认

1. 状态集合

总共有11个状态，每一个位置都代表一个状态

1. 动作集合

总共有三个可执行动作，前，左，右

1. 状态转移函数

给定向前的动作，但是只要0.8的概率正确执行

向前的动作有0.8的概率，向左与向右有0.1的概率

因此其状态转移函数，针对给定的动作与状态，得到的状态是不确定的

1. 奖励函数

规定只有到达最后一个状态才能获得奖励

因此 R(s,a)，其下一个状态为终止状态，这样有两个正向的R(s,a)，均处于终止地点的相接状态

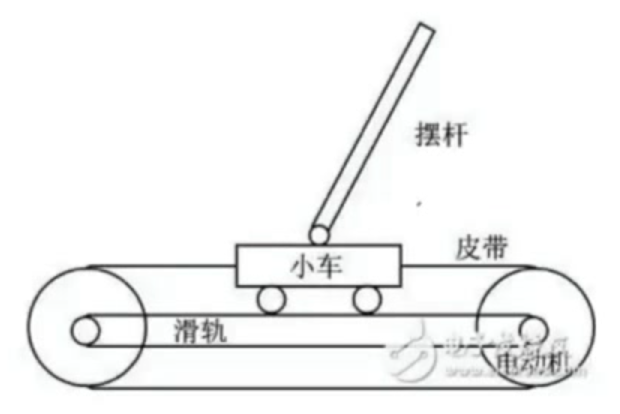
个人理解：这里可能不需要考虑下一个状态随机问题

1. 折扣因子：γ

MDP序列： S0,a0,r1,s1,a1,r2,s2..................ST-1 aT-1 rT ST

状态s0的累计奖励：G0 = r1 + γr2 + γ^2 r3 + .... + γ^T-1 rT

第二个例子的MDP过程，倒立摆



倒立摆控制系统，看似十分简单，实际在火箭与导弹飞行过程中发挥着重大的作用

该系统的完善性能够平衡火箭飞行方向，不偏轨

更难得则是一根杆子上有两个节点，尝试让两者保持平衡

另外机器人行走，保持身体平衡也能利用该系统，说明该问题的基础性与重要性

1. 状态集合

摆杆相对垂直位置的角度， θ

摆杆的角速度 w

小车在滑轨的位置 x

小车的速度 v

1. 动作集合

A = {L，R}，小车往左移动或者往右移动

1. 状态转移概率

该问题状态转移概率，需要根据该问题进行物理建模才能写出来，不必深了解

P(θ’,w’,x’,v’|θ,w,x,v)

1. 奖励函数

通常奖励函数的定义是能够影响问题的最优策略的

比如越靠近中间位置，偏角越垂直，此时奖励越大

但是具体定义可能比较困难，暂时不了解

1. 折扣因子

MDP序列：

S0,a0,r1,s1,a1,r2,s2..................ST-1 aT-1 rT ST

如果没有时间限制，能量限制，那么理论上可能一直运动下去，没有结束状态

累计奖励 Total Payoff

基于模型与免模型的深度强化学习

1. 模型

指的是是否对环境建模，具体指的是状态转移概率，以及奖励函数

1. 基于模型的强化学习 (Model-based)

智能体知道在任何状态下的任何执行动作所获得的回报，即R(s,a)已知

且在下一个状态的概率也是已知的，Psa也是也是已知的，那么针对这种模型环境已知的问题

可以直接使用动态规划算法来求解最优策略，对环境进行建模的强化学习算法

然而实际情况下是很多环境是无法具体建模的，我们就是很难知道P与R

对于每一个状态以及每一个动作，只是知道某个具体点的奖励Reward

1. 面模型的强化学习 (Model-free)

实际上不对环境建模也可以找到最优策略，这种情况是比较符合现实生活问题的

因为大部分时候我们是无法对环境建模的，能做的只有观察环境

在这种情况下，一般是通过求解最大累计奖励 来求解 最优策略