浅谈强化学习

xcmu

Overview

- 强化学习简介
- 强化学习建模
- 马尔科夫决策过程
- 强化学习模型案例
- 强化学习的分类
- 一个强化学习的应用

强化学习简介
- 定义

强化学习(Reinforcement Learning, RL)又称为鼓励学习、评价学习或增强学习，是机器学习的重要组成部分之一，用于描述和解决智能体(Agent)在与环境(Environment)交互过程中通过学习策略(Policy)以达成回报(Rewards)最大化或实现特定目标的问题。

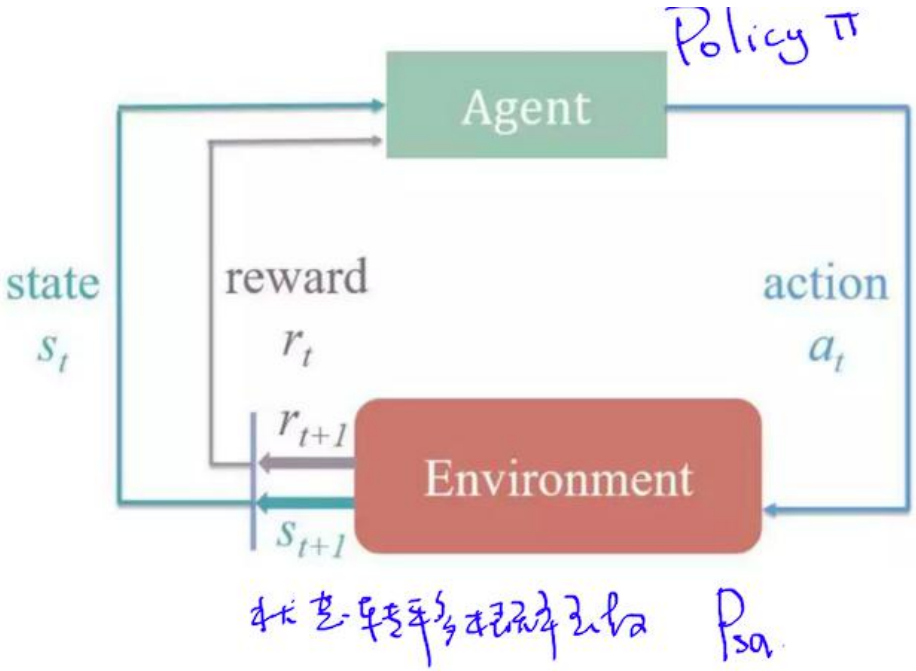
强化学习简介

- 强化学习与有监督无监督的区别与联系

强化学习和监督学习的区别：强化学习不需要事先准备好训练数据，更没有输出作为监督来指导学习过程。强化学习只有奖励值，但这个奖励值和监督学习的输出不一样，它并不是事先给出的，而是延后给出的。同时，强化学习的每一步与时间顺序前后关系密切，而监督学习的训练数据一般是相互独立的，相互之间没有依赖关系。

强化学习与非监督学习的区别：非监督学习只有输入数据，没有输出值也没有奖励，同时非监督学习的数据之间也是相互独立的，相互之间没有依赖关系。

强化学习建模
– 基本模型


Policy π
Agent
state
s_t
reward
r_t
action
a_t
r_{t+1}
Environment
s_{t+1}
P_{sa}

强化学习建模
- 基本组成元素

- 智能体(Agent)：强化学习的本体，作为学习者或决策者存在；
- 环境(Environment)：智能体以外的一切，主要指状态；
- 状态(States)：表示环境的数据，状态集是环境中所有可能的状态；
- 动作(Actions)：智能体可以作出的动作，动作集是智能体可以作出的所有动作；
- 奖励(Rewards)：智能体在执行一个动作后，获得的正负奖励信号；
- 策略(Policy)：从状态到动作的映射，智能体基于某种状态选择某种动作的过程。

强化学习建模
- 学习过程与目标

Step 1：智能体感知环境状态；

Step 2：智能体根据某种策略做出动作；

Step 3：动作作用于环境导致环境状态改变；

Step 4：同时，环境向智能体发出一个反馈信号。

智能体寻找在连续时间序列里的最优策略。最优策略是指使得长期累积奖励最大化的策略。

马尔科夫决策过程
– 马尔科夫性质

在时间步t+1时，环境的反馈仅取决于上一时间步t的状态s和动作a，与时间步t-1以及t-1步之前的时间步都没有关联性

$$p(s_{t+1} \mid a_t s_t a_{t-1} s_{t-1} \cdots a_0 s_0) = p(s_{t+1} \mid a_t, s_t)$$

马尔科夫决策过程
- MDP的基本组成部分

- 状态集合：$S = \{s_1, s_2, \cdots, s_m\}$

  区别：s_i $i = 1, 2, \cdots, m$ 环境有限的状态。

- 动作集合：s_t $t = 1, 2, \cdots$ 时刻的状态 $s_t \in S$

  $A = \{a_1, a_2, \cdots, a_n\}$ 区别：a_i s_i $(A|s_i)$：状态 s_i 下的所有合法的动作

- 状态转移概率函数：

  $P_{sa}: S \times A \times S \to [0,1]$ $P_{sa}(s')$：在状态 s 时执行动作 a 转移到状态 s' 的概率

  $\sum_{s' \in S} P_{sa}(s') = 1$ 且 $P_{sa}(s') \in [0,1]$，$\forall s' \in S$

- 奖励函数：

  $R_{sa}: S \times A \to R$（实数） $R_{sa}(s,a)$：在状态 s 时执行动作 a 所得奖励

  $R_{sas}: S \times A \times S \to R$（实数） $R_{sa}(s,a,s')$ 在状态 s 时执行动作 a 转移到状态 s' 时得到的奖励

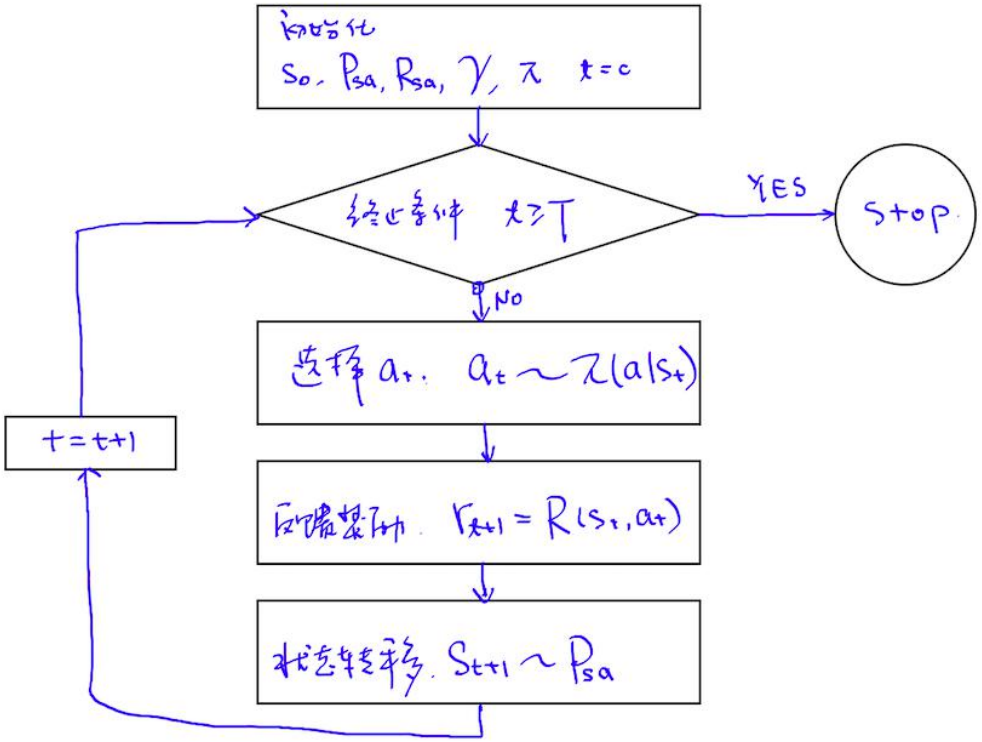
- 策略函数：

  $\pi: S \times A \to [0,1]$：$\pi(a|s)$：在状态 s 的前提，执行动作 a 的概率

  $$\pi(a|s) = \begin{cases} 1 \text{ 或 } 0 & (\pi(s) = a) \text{ 确定性策略} \\ [0,1] & \text{随机性策略} \end{cases}$$

- 折扣因子：$\gamma \in [0,1]$
  - $\gamma = 0$：贪婪法，价值只由当前延时奖励决定；
  - $\gamma = 1$：所有后续状态奖励和当前状态奖励同等重要；
  - $\gamma \in (0,1)$：当前延时奖励的权重比后续奖励的权重大。

马尔科夫决策过程
- MDP的基本流程



产生一个状态-动作-奖励序列：

$$s_0\ a_0\ r_1, s_1, a_1\ r_2, s_2\ a_2\ r_3, \cdots, s_t\ a_t\ r_{t+1}, \cdots s_{T-1}\ a_{T-1}\ r_T\ s_T$$

累积奖励(Total Payoff):

$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots + \gamma^{T-t-1} r_T$$

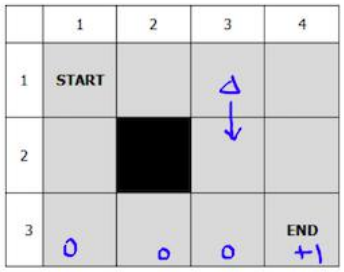
这里

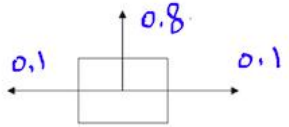
$$\gamma \in [0, 1]$$

强化学习模型案例
– 离散空间

例 1 机器人行走

想象一个机器人在图中所示的网格中行走，其中格子(2,2)为障碍物。机器人碰到墙（边缘）或障碍物会保持不动。机器人初始状态为格子(1,3)，若机器人移动到格子(3,4)，则过程结束。





- 状态集合：由 11 个状态构成，分别为除障碍物以外的每一个格子

$S = \{(1,1), (1,2), \cdots (3,4)\}$ $m=12$

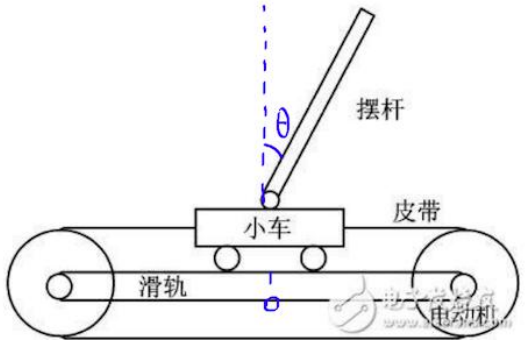
- 动作集合：由 4 个动作构成，每一个移动方向为一个动作

$A = \{N, S, E, W\}$ $A = \{up, down, left, right\}$

强化学习模型案例
– 连续空间

例 2 倒立摆

倒立摆控制系统是一个复杂的、不稳定的、非线性系统，是进行控制理论教学及开展各种控制实验的理想实验平台。对倒立摆系统的研究能有效的反映控制中的许多典型问题：如非线性问题、鲁棒性问题、镇定问题、随动问题以及跟踪问题等。通过对倒立摆的控制，用来检验新的控制方法是否有较强的处理非线性和不稳定性问题的能力。同时，其控制方法在军工、航天、机器人和一般工业过程领域中都有着广泛的用途，如机器人行走过程中的平衡控制、火箭发射中的垂直度控制和卫星飞行中的姿态控制等。



状态集合：

$$S=\{(s,\dot{s},\theta,\dot{\theta})\}$$

动作集合：

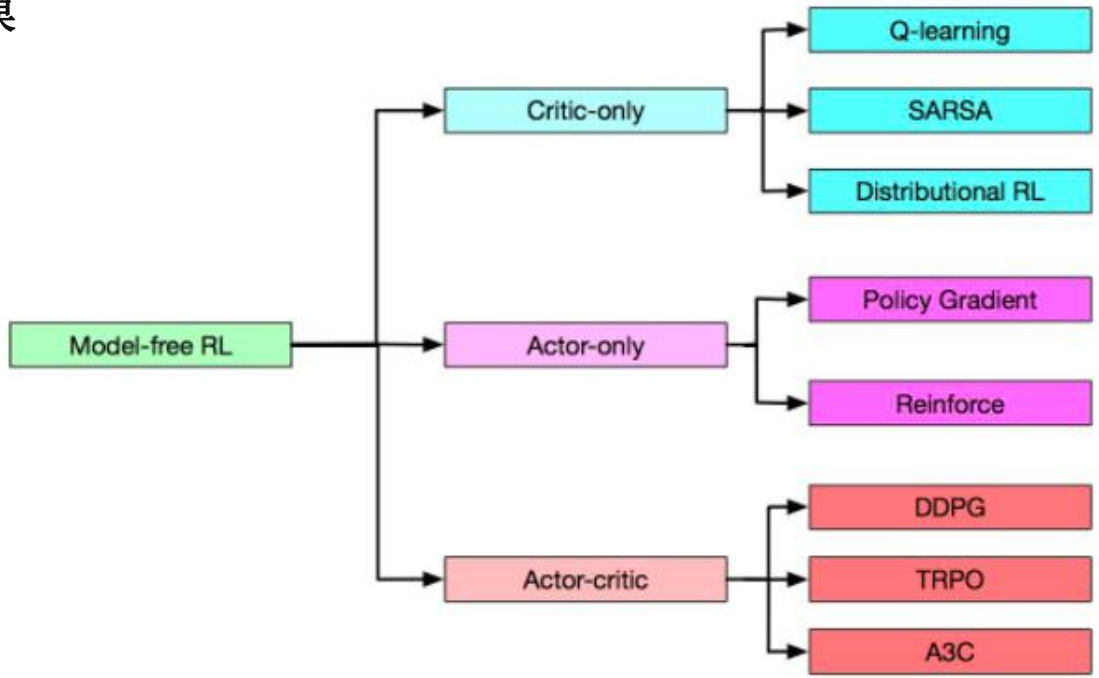
$$A=\{L,R\}$$

强化学习的分类

强化学习的算法目前一般分为两类：

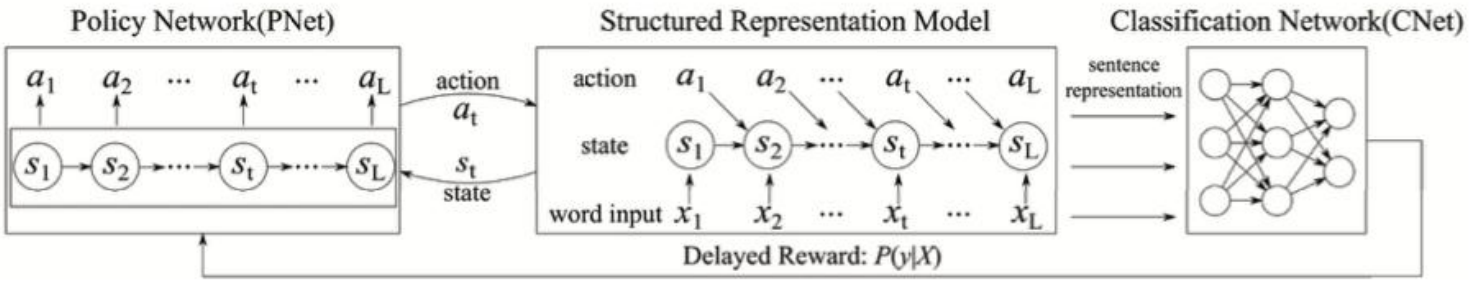
- Model-free RL
- Model-based RL

Model-free 以及Model-based 的最大区别是：**是否有对环境建模**



一个强化学习的应用

- 特征选择



Learning Structured Representation for Text Classificationxvia Reinforcement Learning
（AAAI 2018）