强化学习在自然语言处理下的应用篇

来自： AiGC面试宝典

宁静致远

2024年01月27日 20:47

一、强化学习基础面

1.1 介绍一下强化学习?

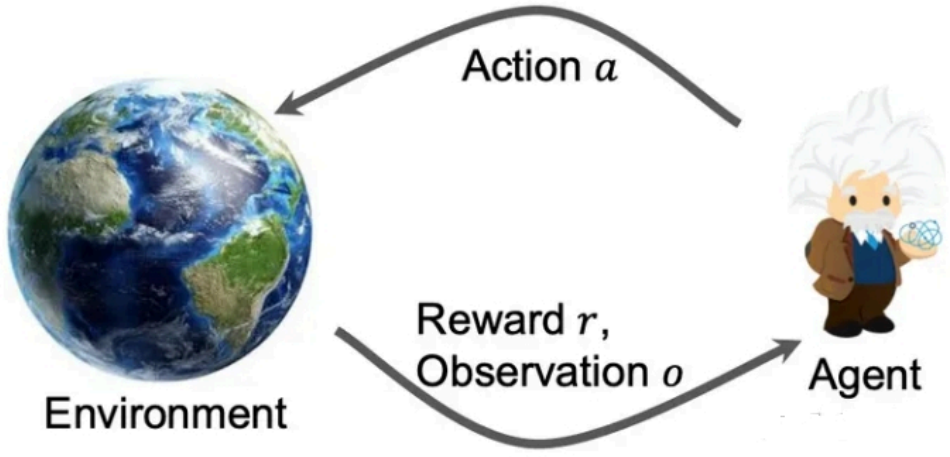
强化学习（Reinforcement Learning）是一种时序决策学习框架，通过智能体和环境交互

$$a_t = \pi(o_t)$$

得到的奖励

$$r_t = r(o_t, a_t)$$

从而来优化策略 π，使其能够在环境中自主学习。



1.2 介绍一下强化学习 的 状态（States） 和 观测（Observations）?

- 状态（States）：对于世界状态的完整描述
- 观测（Observations）：对于一个状态的部分描述，可能会缺失一些信息。当O=S时，称O为完美信息/fully observed；O<S时，称O为非完美信息/partially observed。

1.3 强化学习 有哪些 动作空间（Action Spaces），他们之间的区别是什么?

- 离散动作空间：当智能体只能采取有限的动作，如下棋/文本生成
- 连续动作空间：当智能体的动作是实数向量，如机械臂转动角度

其区别会影响policy网络的实现方式。

1.4 强化学习 有哪些 Policy策略?

- 确定性策略Deterministic Policy： at = u(st)，连续动作空间
- 随机性策略Stochastic Policy： at ~ π(·|st)，离散动作空间

1.5 介绍一下 强化学习 的 轨迹?

- 轨迹：指的是状态和行动的序列

$$\tau = (s_0, a_0, s_1, a_1, \dots)$$

1. 状态转换函数（transition function）：

$$s_{t+1} \sim P(\cdot|s_t, a_t)$$

1. 初始状态是从初始状态分布中采样的，一般表示为

$$s_0 \sim \rho(\cdot)$$

1.6 介绍一下 强化学习 的 奖赏函数?

$$r_t \sim R(s_t, a_t, s_{t+1}) \,/\, r_t \sim R(s_t, a_t)$$

智能体的目标是最大化**行动轨迹的累计奖励**：

$$R(\tau) = \sum_{t=0}^{\infty} \gamma^t r_t$$

1.7 介绍一下 强化学习问题?

- 核心问题：选择一种策略从而最大化**预期收益**

1. 假设环境转换和策略都是随机的，则T步行动轨迹概率：

$$P(\tau \mid \pi) = \rho_0 (s_0) \prod_{t=0}^{T-1} P (s_{t+1} \mid s_t, a_t) \pi (a_t \mid s_t)$$

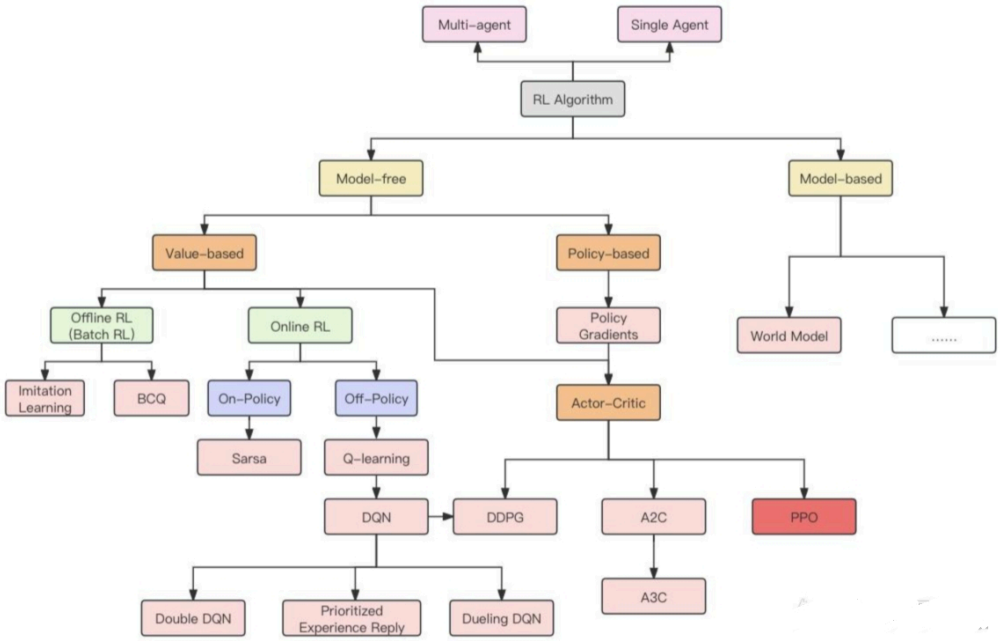
1. 预期收益：

$$J(\pi) = \int_{\tau} P(\tau \mid \pi) R(\tau) = \underset{\tau \sim \pi}{\mathrm{E}} [R(\tau)]$$

1. 核心优化问题：找到最优策略

$$\pi^* = \arg\max_{\pi} J(\pi)$$

二、RL发展路径（至PPO）



2.1 介绍一下 强化学习 中 优化方法 Value-based?

- value-based：状态的值 V(s) 或者 状态行动对(state-action pair) 的值Q(s,a)，作为一种累积奖赏的估计，可以通过最大化值函数来优化得到最优策略

1. 最优值函数（Optimal Value Function）：

$$V^*(s) = \max_\pi \mathop{\mathrm{E}}_{\tau\sim\pi}[R(\tau) \mid s_0 = s]$$

1. 最优动作-值函数（Optimal Action-Value Function）：

$$Q^*(s,a) = \max_\pi \mathop{\mathrm{E}}_{\tau\sim\pi}[R(\tau) \mid s_0 = s, a_0 = a]$$

最优动作：

$$a^*(s) = \arg\max_a Q^*(s,a)$$

1. 两者的关系：

$$V^\pi(s) = \mathop{\mathrm{E}}_{a\sim\pi}[Q^\pi(s,a)]\ ;\ \ V^*(s) = \max_a Q^*(s,a)$$

2.2 介绍一下 强化学习 中 贝尔曼方程?

- 中心思想：当前值估计=当前奖赏+未来值估计

$$V^\pi(s) = \mathop{\mathrm{E}}_{\substack{a\sim\pi \\ s'\sim P}}[r(s,a) + \gamma V^\pi(s')]$$

$$Q^\pi(s,a) = \mathop{\mathrm{E}}_{s'\sim P}\left[r(s,a) + \gamma \mathop{\mathrm{E}}_{a'\sim\pi}[Q^\pi(s',a')]\right]$$

所以，最优值函数的贝尔曼公式为：

$$V^*(s) = \max_a \mathop{\mathrm{E}}_{s' \sim P} \left[r(s,a) + \gamma V^*\left(s'\right) \right]$$

$$Q^*(s,a) = \mathop{\mathrm{E}}_{s' \sim P} \left[r(s,a) + \gamma \max_{a'} Q^*\left(s', a'\right) \right]$$

2.3 介绍一下 强化学习 中 优势函数Advantage Functions?

强化学习中，有时不需要知道一个行动的绝对好坏，而只需要知道它相对于其他action的相对优势。即

$$A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$$