强化学习基础：马尔科夫决策过程与贝尔曼方程

1. 强化学习概述

强化学习（Reinforcement Learning, RL）是机器学习的一个重要分支，其核心思想是智能体（Agent）通过与环境（Environment）的交互来学习最优行为策略，以最大化长期累积奖励。与监督学习和无监督学习不同，强化学习不依赖于标注数据或聚类结构，而是通过试错机制进行学习。
在强化学习中，智能体在每个时间步观察当前状态，采取某个动作，环境据此反馈一个奖励，并转移到下一个状态。目标是学习一个策略，使得从当前状态开始的未来奖励总和最大化。

2. 马尔科夫决策过程（Markov Decision Process, MDP）

为了形式化强化学习问题，通常使用马尔科夫决策过程（MDP）作为数学框架。MDP 提供了一种描述智能体与环境交互的结构化方式。

2.1 MDP 的五元组定义

一个 MDP 可以表示为五元组 (S, A, P, R, γ)，其中：

- S：状态集合（State Space），表示环境中所有可能的状态。
- A：动作集合（Action Space），表示智能体可以执行的所有动作。
- P：状态转移概率函数，定义为 $P(s' mid\ s, a)$，表示在状态 s 下执行动作 a 后转移到状态 s' 的概率。
- R：奖励函数，定义为 $R(s, a, s')$，表示从状态 s 执行动作 a 转移到 s' 时获得的即时奖励。
- γ：折扣因子，满足 $0 \le \gamma < 1$，用于衡量未来奖励的重要性。

2.2 马尔科夫性质

MDP 的关键假设是**马尔科夫性质**（Markov Property），即下一个状态的概率分布仅依赖于当前状态和动作，而与过去的历史无关。形式化表示为：

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

这一性质使得问题的建模和求解成为可能，因为它限制了状态转移的依赖范围。

3. 策略与价值函数

3.1 策略（Policy）

策略 $\pi(a\ mid\ s)$ 表示在状态 s 下选择动作 a 的概率分布。策略可以是确定性的（Deterministic），即每个状态对应唯一动作；也可以是随机性的（Stochastic）。
目标是找到一个最优策略 π^*，使得长期累积奖励最大。

3.2 价值函数（Value Function）

价值函数用于评估状态或状态-动作对的“好坏”。主要有两种价值函数：

3.2.1 状态价值函数 $V^{\pi}(s)$

表示从状态 s 开始，遵循策略 π 所能获得的期望累积折扣奖励：

$$V^{\pi}(s) = mathbb_{\pi}\left[\sum_{k=0}^{\infty}\gamma^{k} R_{t+k+1}, middle|, S_t = s\right]$$

其中，$mathbb_{\pi}$ 表示在策略 π 下的期望，R_{t+k+1} 是第 $t+k+1$ 步的即时奖励。

3.2.2 动作价值函数 $Q^{\pi}(s,a)$

表示在状态 s 下采取动作 a 后，遵循策略 π 所能获得的期望累积折扣奖励：

$$Q^{\pi}(s,a) = mathbb_{\pi}\left[\sum_{k=0}^{\infty}\gamma^{k} R_{t+k+1}, middle|, S_t = s, A_t = a\right]$$

4. 贝尔曼方程（Bellman Equation）

贝尔曼方程是强化学习中的核心数学工具，它将价值函数分解为当前奖励与后续状态价值的组合，体现了动态规划的思想。

4.1 贝尔曼期望方程

对于任意策略 π，其状态价值函数满足贝尔曼期望方程：

$$V^{\pi}(s) = \sum_{a \in A}\pi(a\ mid\ s)\sum_{s\prime \in S}P(s'\ mid\ s,a)\left[R(s,a,s') + \gamma V^{\pi}(s')\right]$$

该方程表明：当前状态的价值等于在该状态下采取所有可能动作的加权平均，每项包括即时奖励和下一状态的折扣后价值。

类似地，动作价值函数的贝尔曼期望方程为：

$$Q^{\pi}(s,a) = \sum_{s\prime \in S}P(s'\ mid\ s,a)\left[R(s,a,s') + \gamma\sum_{a\prime \in A}\pi(a'\ mid\ s')\,Q^{\pi}(s',a')\right]$$

4.2 贝尔曼最优方程

最优价值函数 $V^{*}(s)$ 和 $Q^{*}(s,a)$ 定义为所有策略中能达到的最大价值：

$$V^{*}(s) = max_{\pi}V^{\pi}(s)$$
$$Q^{*}(s,a) = max_{\pi}Q^{\pi}(s,a)$$

对应的贝尔曼最优方程如下：

$$V^{*}(s) = max_{a \in A}\sum_{s\prime \in S}P(s'\ mid\ s,a)\left[R(s,a,s') + \gamma V^{*}(s')\right]$$
$$Q^{*}(s,a) = \sum_{s\prime \in S}P(s'\ mid\ s,a)\left[R(s,a,s') + \gamma\, max_{a\prime \in A}Q^{*}(s',a')\right]$$

这些方程描述了最优价值函数的自洽性：最优价值等于当前最大可能收益加上未来最优价值的折扣和。

5. 最优策略与价值迭代

5.1 最优策略的存在性

在有限状态和动作空间的 MDP 中，至少存在一个确定性最优策略 π^{*}，使得：

$$\pi^{*}(a\ mid\ s) =$$

即在每个状态选择使 Q^{*} 最大的动作。

5.2 价值迭代算法

价值迭代是一种求解最优价值函数的动态规划方法。其更新规则基于贝尔曼最优方程：

$$V_{k+1}(s) \ = \ max_{a \in A} \sum_{s\prime \in S} P\,(s'\ mid\ s,a)\,[\,R(s,a,s') \ + \ \gamma\, V_k(s')\,]$$

从任意初始值函数 $V_0(s)$ 开始，反复应用上述更新，直到 V_k 收敛。收敛后的 V 即为最优状态价值函数 V^*。

6. 总结

强化学习通过马尔科夫决策过程建模序贯决策问题，利用贝尔曼方程将长期优化问题分解为递归结构。马尔科夫性质保证了状态转移的局部依赖性，而贝尔曼方程则提供了价值函数的递推关系，为策略评估和优化奠定了理论基础。通过求解贝尔曼最优方程，可以找到最优策略，实现智能体在复杂环境中的高效学习与决策。