1 马尔科夫决策过程

对于强化学习问题, 常使用马尔科夫决策过程的数学框架来进行描述. 我们在这一章里对这一框架进行简单介绍, 并引出核心的贝尔曼最优方程.

1.1 有限马尔科夫决策过程

定义 1.1 随机事件的马尔科夫性质 给定一个随机过程的当前状态 s_t 和历史的所有状态 $s_1, \cdots, s_{t-1}$, 如果未来时刻的状态仅依赖于当前状态 s_t 而与历史状态无关, 即

$$p(s_{t+1}|s_1, \cdots, s_t) = p(s_{t+1}|s_t)$$

则称该随机过程具有**马尔科夫性质**

定义 1.2 马尔科夫过程 具有马尔科夫性质的随机过程称作**马尔科夫过程**.

马尔科夫过程具有广泛的应用. 在物理和化学中, 马尔科夫过程可用于对动力学过程进行建模. 例如, 化学反应就可以看成马尔科夫过程. 氧分子和氢分子能否碰撞生成水分子, 具有一定的随机性与概率, 但与氧分子和氢分子的历史位置和来源无关.

定义 1.3 有限马尔科夫过程 状态的可能取值的数目有限的马尔科夫过程称作**有限马尔科夫过程**. 此时的条件概率可以用状态转移矩阵描述:

$$p(s_{t+1}|s_t) = p(s'|s) = \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nn} \end{bmatrix}$$

其中矩阵的行号代表转移前的状态, 列号代表转移后的状态.

定义 1.4 马尔可夫决策过程 在马尔科夫过程的基础上对每个状态增加可自由控制的行动 a 和回报 r 就是**马尔科夫决策过程 (Markov Decision Process, MDP)**. 此时, 随机过程的行为由概率 $p(s', r|s, a)$ 刻画, 即体系处于状态 s 并采取行动 a 后状态变成 s' 并获得回报 r 的概率.

结合上述概念就是有限马尔科夫决策过程. 基于 $p(s', r|s, a)$ 可以计算一些其它的重要的概率, 例如:

1. 状态转移概率

$$p(s'|s, a) = \sum_{r \in \mathcal{R}} p(s', r|s, a)$$

2. 状态-行动的回报期望

$$r(s, a) = \sum_{r \in \mathcal{R}, s' \in \mathcal{S}} r p(s', r|s, a)$$

3. 状态-行动-后继状态的回报期望

$$r(s,a,s') = \sum_{r\in\mathcal{R}} rp(s',r|s,a) = \sum_{r\in\mathcal{R}} r\frac{p(s',r|s,a)}{p(s'|s,a)}$$

MDP 将马尔可夫性质应用于时序决策建模. 对于强化学习问题, 如果智能体的状态转移和回报符合马尔可夫性, 并进一步假设智能体的策略也具有马尔可夫性, 则强化学习问题就可以描述成马尔可夫决策过程, 进而可以利用动态规划, 随机采样等方法来求解使回报最大化的智能体策略. 具体地, MDP 框架把目标导向行为的强化学习概括成智能体与环境之间来回传递的三种信号, 即行动 a, 状态 s 和回报 r.

1.2 贝尔曼方程

1.2.1 目标函数

在马尔科夫决策过程的普适框架下, 强化学习在 t 时刻的决策目标, 是使之后接收的累积回报最大化. 定义目标函数

$$G_t = R_{t+1} + \cdots + R_t = \sum_{k=t+1}^{T} R_k$$

其中 T 是终止时刻. 这种定义适合**分幕制 (回合制)** 任务, 即具有明确开始和结束状态的任务, 例如棋类游戏等.

强化学习有时也涉及持续性任务, 时间无穷长, 没有明确终止状态, 例如恒温器调节温度等任务. 此时, 为防止目标函数发散, 一般引入折扣项 γ, 即

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

折扣项的引入能让算法在短期收益和长期收益达成平衡. 事实上, 上面两种目标函数可以统一地写成

$$G_t = \sum_{k=0}^{T} \gamma^k R_{t+k+1}$$

$\gamma = 1$ 和 $T \to \infty$ 分别代表了两种情形. 从上述式子中我们可以得到目标函数 G_t 的一个重要性质:

$$G_t = R_{t+1} + \gamma G_{t+1}$$

1.2.2 策略与价值函数

策略是指智能体怎样根据状态采取行动的规则.

定义 1.5 策略函数 在马尔科夫决策过程的框架下, **策略函数** π 给出了当状态为 s 时采取行动 a 的概率为 $\pi(a|s)$.

$$\pi(a|s)$$

这指导算法以概率分布的形式来选择行动, 因此是随机性策略.

定义 1.6 状态价值函数 状态 s 在策略 π 下的状态价值函数则是从 s 出发根据策略采取行动所能取得回报的期望:

$$v_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s]$$

价值函数衡量了状态 s 下策略 π 的优劣程度.

定义 1.7 行动价值函数 状态 s 在策略 π 下采取行动 a 的价值函数则是从 s 出发采取行动 a 所能取得回报的期望:

$$q_\pi(s,a) = \mathbb{E}_\pi[G_t|S_t = s, A_t = a]$$

行动价值函数衡量了在状态 s 下采取行动 a 的优劣程度.

状态价值函数和行动价值函数满足如下关系:

$$v_\pi(s) = \sum_{a\in\mathcal{A}} q_\pi(s,a)\pi(a|s)$$

1.2.3 贝尔曼方程

强化学习的贝尔曼方程是描述价值函数迭代关系的核心公式. 我们结合前面两小节的公式可得

$$\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi[G_t|S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] \\
&= \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}, r\in\mathcal{R}} p(s', r|s, a)\left(r + \gamma\mathbb{E}_\pi[G_{t+1}|S_{t+1} = s']\right) \\
&= \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}, r\in\mathcal{R}} p(s', r|s, a)\left(r + \gamma v_\pi(s')\right)
\end{aligned}$$

类似地可得

$$q_\pi(s,a) = \sum_{s'\in\mathcal{S}, r\in\mathcal{R}} p(s', r|s, a)\left[r + \gamma \sum_{a'\in\mathcal{A}} q_\pi(s', a')\pi(a'|s')\right]$$

定义 1.8 贝尔曼方程 上述关于状态价值函数和行动价值函数的迭代方程称作**贝尔曼方程**.

贝尔曼方程将复杂的长期决策问题分解为*当前行动与未来状态*的迭代关系, 从而将问题转化为类似递归的形式.

1.2.4 最优策略与贝尔曼最优方程

在所有策略中, 总存在一个 (或一组) 最优策略 π^* 使得智能体在与环境交互时获得的回报的期望 $\mathbb{E}_\pi[G_t]$. 记最优策略下的状态价值函数为 $v_*(s)$, 行动价值函数为 $q_*(s,a)$, 则有

$$v_*(s) = \max_\pi v_\pi(s), \quad q_{*(s,a)=\max_\pi q_p i(s,a)}$$

$v_*(s)$ 就是从状态 s 出发能得到的最大的平均累计回报. 如果我们先采取行动 a, 然后遵循策略 π, 那么得到的累计回报事实上就是行动价值函数 $q_\pi(s,a)$. 为了使这一值最大, 我们可以改变 a 和 π, 以得到最大的取值作为 $v_*(s)$ 的结果, 即

$$v_*(s) = \max_{a,\pi} q_\pi(s,a) = \max_a q_*(s,a)$$

即, 最优策略下的某一状态 s 的价值等于该状态下最优行动的价值. 利用马尔可夫决策过程的特性将上式展开可得

$$\begin{aligned} v_*(s) &= \max_a q_*(s,a) \\ &= \max \mathbb{E}_{\pi^*}[G_t | S_t = s, A_t = a] \\ &= \max \mathbb{E}_{\pi^*}[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] \\ &= \max \mathbb{E}_{\pi^*}[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a] \\ &= \max_a \sum_{s' \in \mathcal{S}, r \in \mathcal{R}} p(s', r | s, a) \left(r + \gamma v_*(s')\right) \end{aligned}$$

类似地可得

$$q_*(s,a) = \sum_{s' \in \mathcal{S}, r \in \mathcal{R}} p(s', r | s, a) \left(r + \gamma \max_{a' \in \mathcal{A}} q_*(s', a')\right)$$

定义 1.9 贝尔曼最优方程 上述关于最优策略下状态价值函数和行动价值函数的迭代方程称作**贝尔曼最优方程**.

原则上可以基于贝尔曼最优方程求解 $v_*(s)$ 与 $q_*(s,a)$. 不过由于 max 的存在, 方程组是非线性的, 实际求解比较困难.

如果 $v_*(s)$ 与 $q_*(s,a)$ 已经求得, 那么最优策略就变得非常容易, 只需利用 $v_*(s)$ 向前搜索一步

$$a^* = \arg\min_a \mathbb{E}_{\pi^*}[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a] = \arg\min_a q_*(s,a)$$

即可得出最优行动 a^*. 我们将在下一节介绍求解这两个函数的具体办法.