标题 **title**

作者 *author*

2023 年 7 月 31 日

前言

目录

第一部分

科学的逻辑

第二章　Markov 链与决策

§2.1　Markov 链

我们在第一次课说过，似然，或者说合情推理的合理程度，是一种概率的解释模型. 尽管概率论在数学上通常被形式化为 Kolmogorov 公理体系，但是公理体系并没有回答“概率”是什么. 概率的解释是一个哲学课题. 两个主要的例子：

- 频率解释：概率是无穷次独立重复试验的频率（大数定律）.
- 主观解释（*Bayes* 解释）：概率是对命题合理程度的信念（似然）.

主观解释对推理的假设是逻辑的、静态的，时间的概念并不出现在似然里面. 例如，考虑一个罐子，里面有除颜色之外不可区分的 N 个球，有 n 个白球，剩下的是黑球. 顺序从中拿出 N 个球，第 k 次拿出的球颜色是 W_k 或 B_k.

- $\Pr(W_iW_j) = \Pr(W_i|W_j)\Pr(W_j) = \Pr(W_j|W_i)\Pr(W_i)$. $(i<j)$
- $\Pr(W_i) = \Pr(W_j) = n/N \implies \Pr(W_i|W_j) = \Pr(W_j|W_i)$.

从似然的角度，$\Pr(W_i|W_j)$ 和 $\Pr(W_j|W_i)$ 不仅是可计算的，而且是相等的. 概率的计算告诉了我们，更早状态的信息依赖于未来的状态！逻辑上蕴含关系并不意味着实际上的因果关系，但是似然完全没有考虑这一点. 因此，我们需要引入一个带有时间的模型，这就是 Markov 链.

定义 2.1 (Markov 链) ***Markov*** **链**（马氏链）是一个随机变量序列 $\{X_t\}_{t=0}^{\infty}$. 包含如下概念：

- 状态空间 $\mathcal{S}$：X_t 所有可能值构成的集合，有限或者可数.
- 转移矩阵 $\mathcal{P}$：下一时刻系统状态之间转移的概率. $\mathcal{P} = (p_{ij})_{i,j\in\mathcal{S}}$，$p_{ij}$ 是从 i 状态转移到 j 状态的概率.

- *Markov* 性: 对任意时刻 $t = 1, \ldots, n$ 和任意状态 $j, k, j_0, \ldots, j_{t-1} \in \mathcal{S}$，如下等式成立

$$\begin{aligned} &\Pr(X_{t+1} = j | X_t = k, X_{t-1} = j_{t-1}, \ldots, X_0 = j_0) \\ =&\Pr(X_{t+1} = j | X_t = k) = p_{kj}. \end{aligned}$$

有时候也会考虑带初态的 *Markov* 链，此时 X_0 服从分布 $\lambda = (\lambda_s)_{s \in \mathcal{S}}$.

我们给出的定义是简化的 Markov 链，每个时刻之间的转移都是一样的转移矩阵，这样的 Markov 链被称为时齐的. 有时候也会考虑非时齐的 Markov 链，即每个时刻之间的转移矩阵不一样.

Markov 链是一种简化的带时间的概率模型，它最重要的性质是 Markov 性，即在固定现在的情况下，过去与未来相互独立. 这一性质的数学表述为：

命题 2.1 (Markov 性) 条件在 $X_n = i$ 下，$\{Y_m\}_{m=0}^{\infty} := \{X_{m+n}\}_{m=0}^{\infty}$ 是一个转移矩阵为 P 的 *Markov* 链，并且与 $(X_0, \ldots, X_{n-1})$ 相互独立.

证明留做习题。

我们考虑的 Markov 链还有时齐性，即状态的转移不依赖当前时间，只和当前的状态有关. 时齐性的数学表述为：

命题 2.2 设 $\{X_t\}_{t=0}^{\infty}$ 是一个 *Markov* 链，那么对任意的 $t, m, n \in \mathbb{N}$ 和 $i, j, k \in \mathcal{S}$，有 $\Pr(X_{m+n} = j | X_n = k) = \Pr(X_m = j | X_0 = k)$.

我们来看一个 Markov 链的例子。

例 2.1 (赌徒模型) 考虑公平对赌. 玩家 A 和 B 抛硬币来赌钱，A 赌正面，B 赌反面. 每一轮独立地抛硬币，正面朝上的概率和反面朝上的概率相等，都是 $1/2$. 赢的一方给输的一方一块钱. A 输 a 块钱破产，B 输 b 块钱破产，Z_i 是第 i 轮 A 的收入. $Z_0 = X_0 = 0$ 是 A 初始的收入. $X_n = Z_0 + \cdots + Z_n$ 是 A 的累计收入. 那么，$\{X_n\}_{n \geq 0}$ 是一个 *Markov* 链.

- 状态空间：$\mathcal{S} = \{-a, -a+1, \ldots, 0, 1, \ldots, b\}$.
- 转移概率：对 $-a < i < b-1$，$p_{i,i+1} = p_{i+1,i} = 1/2$；$p_{-a+1,-a} = p_{b-1,b} = 1/2$，$p_{-a,-a} = p_{b,b} = 1$；其他值为 0.

转移矩阵可以画成图 2.1所示的形式.

在上面的赌徒模型中，A 的累计收入 $\{X_n\}_{n \geq 0}$ 形成了 Markov 链. 根据 Markov 性，未来双方的收入变化只取决于现在，而和过去运气无关. 与之相关的一个现象是赌徒谬

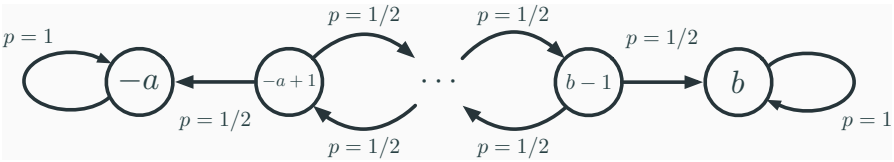

图 2.1: 赌徒模型的转移矩阵

误，即认为过去的运气会影响未来的运气. 例如，如果一个人连续输了很多次，那么他会认为自己未来运气会变好，赢的概率更大. 但是，根据 Markov 性，过去的运气不会影响未来的运气，因此这种想法是错误的. “风水轮流转” 在一场公平对赌中是不正确的认知. 那么，如何评估赌局的公平性?

如果对赌是公平的，那么我们应该认为两个人每一轮的累计收入分布都是一样的，即

$$\Pr(X_n = i | X_0 = 0) = \Pr(X_n = -i | X_0 = 0).$$

因此，我们需要能够计算多步转移的概率. 设 $p_{ij}^{(k)}$ 表示从状态 i 用 k 步转移到状态 j 的概率. k 步转移概率形成了一个矩阵 $\mathcal{P}^{(k)}$. 下面的定理给出了计算*多步转移概率*的方法.

定理 2.1 (Kolmogorov-Chapman 方程) $P^{(k+l)} = \mathcal{P}^{(k)}\mathcal{P}^{(l)}$.

证明 由 Markov 性、时齐性和全概率公式，$p_{ij}^{(k+l)} = \Pr(X_{k+l} = j | X_0 = i) = \sum_\alpha \Pr(X_{k+l} = j, X_k = \alpha | X_0 = i) = \sum_\alpha \Pr(X_k = \alpha | X_0 = i) \Pr(X_{k+l} = j | X_k = \alpha) = \sum_\alpha p_{i\alpha}^{(k)} p_{\alpha j}^{(l)}$. □

Kolmogorov-Chapman 方程有两个重要的特例，前向方程：$\mathcal{P}^{(k+1)} = \mathcal{P}^{(k)}\mathcal{P}$，以及后向方程：$\mathcal{P}^{(l+1)} = \mathcal{P}\mathcal{P}^{(l)}$. 见图 2.2 和图 2.3.

此外，利用归纳法，我们还有如下推论：

推论 2.1 $\mathcal{P}^{(k)} = \mathcal{P}^k$.

若已知初始分布向量为 λ，利用这一推论，我们可以计算它随时间的演化：

$$\lambda^\mathsf{T}, \lambda^\mathsf{T}\mathcal{P}, \ldots, \lambda^\mathsf{T}\mathcal{P}^n, \ldots$$

回到赌徒模型，如何计算公平对赌中 X_n 的概率分布？我们先来看一个简化的例子. 假设 $|p_{00} + p_{11} - 1| < 1$，考虑只有两个状态 0, 1，转移矩阵为

$$\mathcal{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}.$$

或者画成图 2.4 的形式.

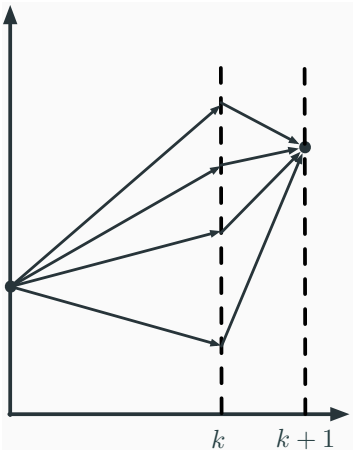

图 2.2: 前向方程（往前一步）

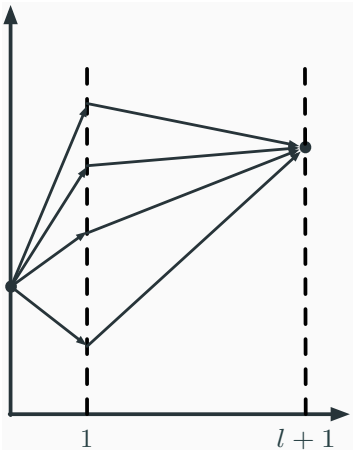

图 2.3: 后向方程（往回一步）

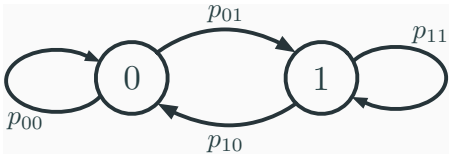

图 2.4: 只有两个状态的 Markov 链

可以归纳证明：

$$\begin{aligned}\mathcal{P}^n =& \frac{1}{2-p_{00}-p_{11}}\begin{pmatrix}1-p_{11} & 1-p_{00}\\ 1-p_{11} & 1-p_{00}\end{pmatrix}\\ &+\frac{(p_{00}+p_{11}-1)^n}{2-p_{00}-p_{11}}\begin{pmatrix}1-p_{00} & -(1-p_{00})\\ -(1-p_{11}) & 1-p_{11}\end{pmatrix}.\end{aligned}$$

注意到，$\lim_{n\to\infty} p_{i0}^{(n)} = (1-p_{11})/(2-p_{00}-p_{11})$，$\lim_{n\to\infty} p_{i1}^{(n)} = (1-p_{00})/(2-p_{00}-p_{11})$. 随着时间的推移，Markov 链初始状态对概率分布的影响逐渐消失. 这个规律具有普遍性，这就是遍历定理.

定理 2.2 (遍历定理) 设 *Markov* 链的状态空间为 $\mathcal{S}=\{1,\dots,N\}$，转移矩阵为 $\mathcal{P}=(p_{ij})$. 如果对于某一个 n_0 有

$$\min_{ij} p_{ij}^{(n_0)} > 0, \tag{2.1}$$

那么存在分布 $\lambda=(\lambda_1,\dots,\lambda_N)$ 使得

$$\lambda_i > 0, \quad \sum_i \lambda_i = 1, \tag{2.2}$$

并且对于每一个 $j \in \mathcal{S}$ 和任意 $i \in \mathcal{S}$ 都有

$$p_{ij}^{(n)} \to \lambda_j, n \to \infty. \tag{2.3}$$

反之，如果存在满足 (2.2) 和 (2.3) 的 λ，则存在满足 (2.1) 的 n_0. 式 (2.2) 的 λ 满足

$$\lambda^{\mathsf{T}} = \lambda^{\mathsf{T}} \mathcal{P}. \tag{2.4}$$

条件 (2.1) 表明超过某个步数 n_0 之后，从 i 出发到达 j 的概率总是正的，这个条件被称为*遍历*. 条件 (2.2) 表明每一个状态被访问到的概率都是正的，没有 "死状态". 遍历定理表明遍历的 Markov 链从任何状态出发都是不可逆的，最终会把每个状态都走过一遍（遍历），变成一个混合均匀的状态. 这可以用来解释物理学中的扩散现象.

满足条件 (2.4) 的分布被称为*平稳分布*. 平稳分布为初始状态时，Markov 链的演化与时间无关：

命题 2.3 *设 $\{X_n\}$ 是 Markov 链，如果 X_0 是平稳分布，那么 $(X_k, \ldots, X_{k+l})$ 的联合分布不依赖于 k.*

如果 Markov 链是遍历的，那么平稳分布是唯一的：

命题 2.4 *设 $\{X_n\}$ 是遍历的 Markov 链，那么它有唯一平稳分布 μ。*

证明 假设 μ 是另外一个平稳分布，那么 $\mu_j = \sum_\alpha \mu_\alpha p_{\alpha j} = \cdots = \sum_\alpha \mu_\alpha p_{\alpha j}^{(n)}$. 因为 $p_{\alpha j}^{(n)} \to \lambda_j$，所以 $\mu_j = \sum_\alpha (\mu_\alpha \lambda_j) = \lambda_j$. □

非遍历 Markov 链也可能存在（唯一）平稳分布，考虑如下转移矩阵：

$$\mathcal{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

它有唯一平稳分布 $\lambda = (1/2, 1/2)^{\mathsf{T}}$.

§2.2 Markov 奖励过程（MRP）

我们接下来的目标就是在 Markov 链上建立决策理论. 每一阶段我们可以选择某个行动，这个行动在 Markov 链会产生一些奖励. 我们的目标是选择恰当的行动方式是的我们的总奖励最大. 首先我们定义奖励的过程.

定义 2.2 *一个 **Markov** 奖励过程（MRP）是四元组 $\langle \mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma \rangle$：*

- $\mathcal{S}$ 是一个有穷的状态集合.
- $\mathcal{P}$ 是一个状态转移矩阵，从 i 转移到 j 的概率记为 $\mathcal{P}_{i,j}$.
- $\mathcal{R}$ 是一个奖励函数，$\mathcal{R}_s = \mathbb{E}[R_{t+1}|S_t = s]$：当 t 时刻位于状态 s 时下一时刻（离开）获得的奖励的期望，R_{t+1} 是下一阶段所处状态的奖励.
- γ 是一个折扣系数，$\gamma \in [0,1]$.

接下来我们看一个例子：学生 MRP。

例 2.2 (学生 MRP) 见图 2.5.*[lhy: 补全]*

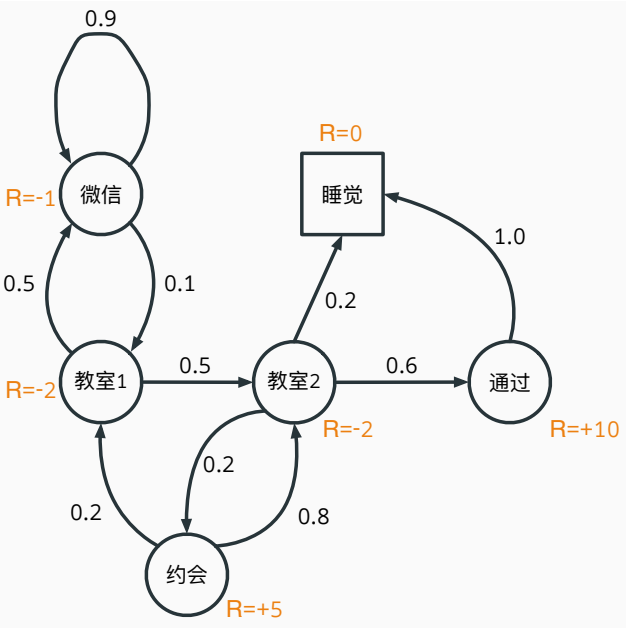

图 2.5: 学生 MRP

MRP 中，t 时刻以后的总回报G_t 定义为

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

$\gamma \in [0,1]$ 衡量了未来下一时段 1 的奖励在当前时刻的价值. 未来 $k+1$ 时刻的奖励对当前时刻 t 的作用是 $\gamma^k R_{t+k+1}$. 若 $\gamma \to 0$，表示对奖励进行“短视”的评估；反之更“远见”.

许多 MRP 和后面学习的 MDP 都有与时间无关的折扣系数 $\gamma < 1$，原因例如，对未来不确定性对冲，这直接对英语直接对应于利润率. 另外，动物和人类对即时回报具有偏

好. 有时也使用非折扣化的 MRP（即 $\gamma = 1$），例如当所有的转移序列都会有固定的终止时间.

在 MRP 中，状态*价值函数* $v(s)$ 表示从状态 s 出发的期望回报

$$v(s) = \mathbb{E}(G_t | S_t = s).$$

价值函数 $v(s)$ 衡量了状态 s 的长期效益. 这一定义蕴含了 Markov 性：只从当前起考虑未来收益，不考虑历史收益（沉没成本）的影响；也蕴含了时齐性：价值函数的定义不依赖于时刻 t（无穷阶段情形）.

价值函数可以被分解为两部分：即时回报 R_{t+1} 以及下一个状态开始的折扣价值 $\gamma v(S_{t+1})$，具体来说，我们有

$$\begin{aligned} v(s) &= \mathbb{E}(G_t | S_t = s) \\ &= \mathbb{E}(R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots | S_t = s) \\ &= \mathbb{E}(R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \dots) | S_t = s) \\ &= \mathbb{E}(R_{t+1} + \gamma G_{t+1} | S_t = s) \\ &= \mathbb{E}(R_{t+1} + \gamma v(S_{t+1}) | S_t = s) \\ &= \mathcal{R}_s + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{s,s'} v(s'). \end{aligned}$$

我们因此得到了 **Bellman 方程**：

定理 2.3 (Bellman 方程) $v(s) = \mathcal{R}_s + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{s,s'} v(s')$.

Bellman 方程可以用矩阵形式表达：

$$v = \mathcal{R} + \gamma \mathcal{P} v.$$

这里 v 是列向量 $v = (v(s))_{s \in \mathcal{S}}$.

Bellman 方程是一个线性方程，可以被直接解:

$$v = \mathcal{R} + \gamma \mathcal{P} v \implies (I - \gamma \mathcal{P}) v = \mathcal{R} \implies v = (I - \gamma \mathcal{P})^{-1} \mathcal{R}.$$

对于 n 个状态的 Markov 链，计算复杂度为 $\mathcal{O}(n^3)$. 对于较小的 MRP 可以直接解，太大的 MRP 开销太大. 对于大型 MRP，可以采用迭代算法，例如：

- 动态规划
- Monte-Carlo 评估
- 时序差分学习

§2.3 Markov 决策过程（MDP）

接下来我们定义 Markov 决策过程. MDP 是 MRP 的扩展，它增加了行动的概念，是一个定义了决策的 MRP. 它可以看做一个任意状态都具有 Markov 性的环境.

定义 2.3 ***Markov*** **决策过程**（*MDP*）是一个 *MDP* 是五元组 $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$，其中

- $\mathcal{S}$ 是一个有限的状态集合.
- $\mathcal{A}$ 是一个有限的行动（*action*）集合.
- $\mathcal{P}$ 是状态转移概率矩阵，

$$\mathcal{P}_{ss'}^{a} = \Pr(S_{t+1} = s' | S_t = s, A_t = a).$$

- $\mathcal{R}$ 是一个奖励函数，$\mathcal{R}_s^a = \mathbb{E}(R_{t+1} | S_t = s, A_t = a)$，$R_{t+1}$ 是进行某一行动到达某一状态后的奖励.
- γ 是一个折扣系数 $\gamma \in [0, 1]$.

我们继续前面学生的例子，此时变成学生 MDP：

例 2.3 (学生 MDP) 见图 2.6.

一个策略 π 是给定状态下行动的分布，

$$\pi(a|s) = \Pr(A_t = a | S_t = s).$$

一个策略完全决定了一个智能体在 MDP 环境中的行为. 它蕴含着 Markov 性：MDP 的策略取决于当前状态，而非历史状态；也蕴含着时齐性：MDP 的策略不依赖于时刻 t.

MDP 与 Markov 链、MDP 的关系由策略给出。

命题 2.5 给定一个 *MDP* $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ 和一个策略 π，$\langle \mathcal{S}, \mathcal{P}^{\pi} \rangle$ 是一个 *Markov* 链，$\langle \mathcal{S}, \mathcal{P}^{\pi}, \mathcal{R}^{\pi}, \gamma \rangle$ 是一个 *MRP*，其中

$$\mathcal{P}_{s,s'}^{\pi} = \mathbb{E}_{a \sim \pi(\cdot|s)}(\mathcal{P}_{s,s'}^{a}) = \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{P}_{s,s'}^{a},$$

$$\mathcal{R}_s^{\pi} = \mathbb{E}_{a \sim \pi(\cdot|s)}(\mathcal{R}_s^a) = \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{R}_s^a.$$

证明 利用全概率公式。 □

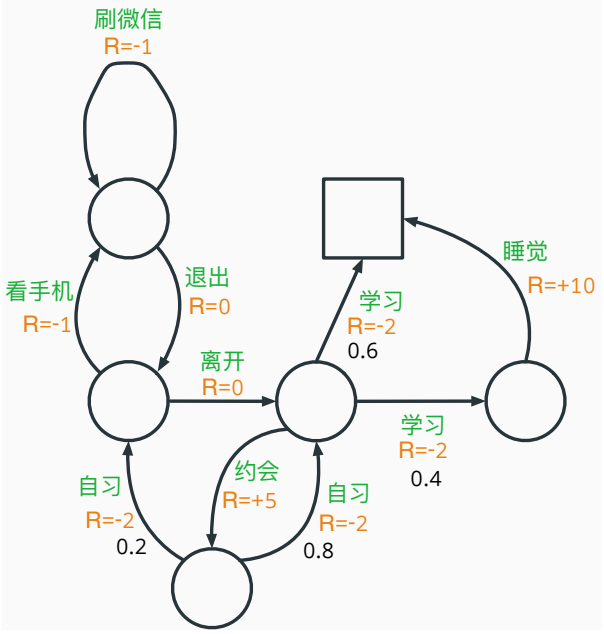

图 2.6: 学生 MDP

在 MDP 中，状态-价值函数和行动-价值函数是两个重要的价值函数，它们分别描述了从某一状态出发，遵从某一策略的期望回报。状态-价值函数 $v_\pi(s)$ 是从状态 s 出发，遵从策略 π 的期望回报

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

行动-价值函数 $q_\pi(s,a)$ 是从状态 s 出发，采取行动 a，遵从策略 π 的期望回报

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

注意，以上定义都具有 Markov 性和时齐性.

下面我们给出状态-价值函数和行动-价值函数的 Bellman 方程。状态-价值函数可以被分解为：即时回报加后续状态的折扣价值，

$$v_\pi(s) = \mathbb{E}_\pi(R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s).$$

行动-价值函数可以被类似地分解，

$$q_\pi(s,a) = \mathbb{E}_\pi(R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1})|S_t = s, A_t = a).$$

二者之间的关系（全概率公式、一步转移概率）：

$$q_\pi(s,a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{s,s'}^a v_\pi(s').$$

$$v_\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}(q_\pi(s,a)) = \sum_{a\in\mathcal{A}} \pi(a|s) q_\pi(s,a),$$

因此，我们得到 MDP 的 Bellman 期望方程：

命题 2.6 (状态-价值函数和行动-价值函数的 Bellman 方程)

$$v_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s) \left(\mathcal{R}_s^a + \gamma \sum_{s'\in\mathcal{S}} \mathcal{P}_{s,s'}^a v_\pi(s') \right),$$

$$q_\pi(s,a) = \mathcal{R}_s^a + \gamma \sum_{s'\in\mathcal{S}} \mathcal{P}_{s,s'}^a \sum_{a'\in\mathcal{A}} \pi(a'|s') q_\pi(s',a').$$

状态-价值函数的 Bellman 方程可以被写成矩阵形式：

$$v_\pi = \mathcal{R}^\pi + \gamma \mathcal{P}^\pi v_\pi = (I - \gamma \mathcal{P}^\pi)^{-1} \mathcal{R}^\pi.$$

接下来我们讨论最优策略和最优价值函数。

定义 2.4 (最优状态-价值函数和最优行动-价值函数) **最优状态-价值函数** $v_\star(s)$ 是所有决策中最大的状态-价值函数

$$v_\star(s) = \max_\pi v_\pi(s).$$

最优行动-价值函数 $q_\star(s,a)$ 是所有决策中最大的行动-价值函数

$$q_\star(s,a) = \max_\pi q_\pi(s,a).$$

最优价值函数确定了 MDP 中的最佳收益，解 MDP 即确定达到最优价值函数的策略.

然而，每个状态取到最大价值的策略 π 可能并不是同一个. 幸运的是，确实存在一个这样的最优策略. 定义一个策略的偏序：

$$\pi \geq \pi' \iff \forall s \in \mathcal{S}\ v_\pi(s) \geq v_{\pi'}(s).$$

我们有如下定理：

定理 2.4 (MDP 解的存在性) 对任意 *MDP*,

- 存在一个最优策略 $\pi_\star$ 使得 $\forall \pi\ \pi_\star \geq \pi$.
- 最优策略取得最优状态-价值函数：$v_{\pi_\star}(s) = v_\star(s)$.
- 最优策略取得最优行动-价值函数：$q_{\pi_\star}(s,a) = q_\star(s,a)$.

证明 我们给出一个构造性证明，即找出最优决策。可以通过最大化 $q_\star(s,a)$ 来寻找：

- 固定 s.
- 找到一个 $a_\star$ 使得 $q_\star(s,a_\star)=\max_a q_\star(s,a)$，令 $\pi_\star(a_\star|s)=1$.
- 对 $\forall a\neq a_\star$，$\pi_\star(a|s)=0$.

首先，根据选法，$\pi_\star$ 取得最优行动-价值函数. 由 $v_\pi(s)=\mathbb{E}_{a\sim\pi(\cdot|s)}(q_\pi(s,a))\leq\mathbb{E}_{a\sim\pi(\cdot|s)}(q_\star(s,a))\leq q_\star(s,a_\star)=v_{\pi_\star}(s)$ 知 $\pi_\star$ 取得最优状态-价值函数. □

这个证明还有一个推论：

推论 2.2 *对任意 MDP，总存在一个非随机的最优决策.*

如果我们知道 $q_\star(s,a)$，我们就能获得最优决策. 最优价值函数由 Bellman 最优性方程联系：

$$
\begin{aligned}
v_\star(s)&=\max_a q_\star(s,a),\\
q_\star(s,a)&=\mathcal{R}_s^a+\gamma\sum_{s'\in\mathcal{S}}\mathcal{P}_{s,s'}^a v_\star(s'),\\
v_\star(s)&=\max_a\left\{\mathcal{R}_s^a+\gamma\sum_{s'\in\mathcal{S}}\mathcal{P}_{s,s'}^a v_\star(s')\right\},\\
q_\star(s,a)&=\mathcal{R}_s^a+\gamma\sum_{s'\in\mathcal{S}}\mathcal{P}_{s,s'}^a\max_{a'}q_\star(s',a').
\end{aligned}
$$

Bellman 最优性方程不是线性的. 因此没有解析形式的解. 但是 MDP 的数值解是可以多项式时间求出来的. 我们一般采用迭代算法求解：

- 价值迭代
- 策略迭代
- Q-learning
- Sarsa

Bellman 方程是强化学习、经济学动态优化的核心. Bellman 方程的推导是 Markov 链中最为常用的技巧：考虑从当前状态转移到下一状态，利用全概率公式，一步转移会将两个状态之间的概率（期望）用递推公式联系起来. 在随机过程中，有大量这样的例子：前向方程、*Wald* 等式、调和函数. 后面的*HMM* 也是类似的例子.

§2.4 隐 Markov 模型（HMM）

我们考虑 Markov 链上的另一种应用. 在统计学和机器学习中，我们有时候要处理一类含时间的数据. 最简单的情况是回归，即数据完全由所处时刻决定. 但是通常，现在的数据依赖于过去的数据. 因此，一种最简单的考虑就是数据依赖于 Markov 链，这就是隐 *Markov* 模型.

定义 2.5 (隐 Markov 模型) 一个隐 ***Markov*** 模型（*HMM*）是一列随机变量 $X_1, X_2, \ldots, X_t$，满足：

- X_t 的分布仅依赖于隐状态 Z_t，即 $\Pr(X_1, \ldots, X_t | Z_1, Z_2, \ldots, Z_t) = \prod_i \Pr(X_i | Z_i)$.
- $\{Z_t\}$ 构成一条 *Markov* 链.

示意见图 2.7.

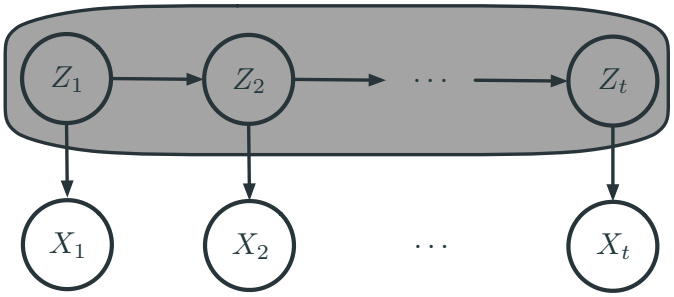

图 2.7: 隐 Markov 模型

我们可以更具体地说写出来 HMM 的结构。一个 HMM 包含：

- $\mathcal{Z}$: 有限的状态集合.
- $\mathcal{X}$: 有限的观测集合.
- $T : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$，$\mathcal{Z}$ 的转移概率.
- $M : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$，给定状态时的观测概率（条件概率）.
- $\lambda : \mathcal{Z} \to \mathbb{R}_{\geq 0}$，初始状态的先验概率分布列.

如果随机过程 $\{X_t\}$ 的值域是有限集，我们则可以用矩阵表达 HMM：

- T 是 $\{Z_t\}$ 的转移矩阵.

- M 是观测矩阵：$M_{i,k} = \Pr(X_t = k | Z_t = i)$.
- λ 是一个概率向量.

2.4.1 评估问题

给定一个特定的 HMM，它对实际观测序列的拟合程度有多好？为了讨论，我们引入记号随机向量 $X = (X_1, \ldots, X_t)$，$Z = (Z_1, \ldots, Z_t)$. 我们考虑 HMM 的评估问题：给定一个 HMM $\mathcal{M}$，以及它的观测历史 $x = (x_1, x_2, \ldots, x_t)$，计算 $\Pr(X = x|\mathcal{M})$. 关键困难是我们不知道状态历史 $Z = (z_1, z_2, \ldots, z_t)$.

直接使用条件概率进行推导，我们可以得到如下算法：

$$\Pr(X = x|\mathcal{M}) = \sum_{Z=(z_1,\ldots,z_t)\in\mathcal{Z}} \Pr(X = x|Z = z, \mathcal{M}) \Pr(Z = z|\mathcal{M}),$$

$$\Pr(X = x|Z = z, \mathcal{M}) = \prod_{i=1}^{t} \Pr(X_i = x_i|Z_i = z_i) = M_{z_1,x_1} \cdot M_{z_2,x_2} \ldots M_{z_t,x_t},$$

$$\begin{aligned}\Pr(Z = z|\mathcal{M}) &= \Pr(Z_1 = z_1) \prod_{i=2}^{t} \Pr(Z_i = z_i|Z_{i-1} = z_{i-1}) \\ &= \lambda_{z_1} \cdot T_{z_1,z_2} \cdot T_{z_2,z_3} \ldots T_{z_{t-1},z_t}.\end{aligned}$$

这一方法的时间复杂度是 $\mathcal{O}(t|\mathcal{Z}|^t)$. 然而，这一算法中，$t$ 在指数上，因此是不可接受的. 我们需要更好的算法.

接下来，我们采用前向方程的思路，从前 k 步的结果推出前 $k+1$ 步的结果. 因此可以列出递推方程. 为了方便，引入记号：$X_{i:j} = (X_i, \ldots, X_j)$. 然后，定义 $\alpha_k(z) := \Pr(X_{1:k} = x_{1:k}, Z_k = z|\mathcal{M})$，我们有递推：

- $\alpha_1(z) = \lambda(z) M_{z,x_1}$.
- $\alpha_{k+1}(z) = \sum_{z'\in\mathcal{Z}} \alpha_k(z') T_{z',z} M_{z,x_{k+1}}$.

于是，$\Pr(X = x|\mathcal{M}) = \sum_{z\in\mathcal{Z}} \alpha_t(z)$.

这一方法的时间复杂度是 $\mathcal{O}(t|\mathcal{Z}|^2)$.

镜像地，我们可以使用后向方程的思路，从前 $k+1$ 步的结果推出前 k 步的结果. 同样可以列出递推方程. 定义 $\beta_k(z) := \Pr(X_{k+1:t} = x_{k+1:t}|Z_k = z, \mathcal{M})$，我们有递推：

- 当 $k = t$，$\beta_k(z) = 1$.

- 当 $1 \leq k < t$，$\beta_k(z) = \sum_{z' \in \mathcal{Z}} T_{z,z'} M_{z',x_{k+1}} \beta_{k+1}(z')$.

于是，$\Pr(X = x|\mathcal{M}) = \sum_{z \in \mathcal{Z}} \lambda(z) M_{z,x_1} \beta_1(z)$. 这一方法的时间复杂度是时间复杂度 $O(t|\mathcal{Z}|^2)$.

2.4.2 解释问题

接下来我们讨论 HMM 的*解释问题*. 给定一个 HMM $\mathcal{M} = (\mathcal{Z}, \mathcal{X}, T, M, \lambda)$, 一列观测历史 $x = (x_1, x_2, \ldots, x_t)$, 解释问题旨在寻找一个状态序列，能最好地解释这些历史观察. 具体来说我们考虑如下四个问题：

1. 过滤：计算 $\Pr(Z_k = s|X_{1:k} = x_{1:k}, \mathcal{M})$.
2. 平滑：计算 $\Pr(Z_k = s|X = x, \mathcal{M})$，$k < t$.
3. 预测：计算 $\Pr(Z_k = s|X = x, \mathcal{M})$，$k > t$.
4. 解码：找到最有可能的状态序列 $z = (z_1, z_2, \ldots, z_t)$.

首先考虑过滤：$\Pr(Z_k = s|X_{1:k} = x_{1:k}, \mathcal{M})$。回顾记号 $\alpha_k(s) = \Pr(X_{1:k} = x_{1:k}, Z_k = s|\mathcal{M})$，我们有

$$\begin{aligned}\Pr(Z_k = s|X_{1:k} = x_{1:k}, \mathcal{M}) &= \frac{\Pr(X_{1:k} = x_{1:k}, Z_k = s|\mathcal{M})}{\Pr(X_{1:k} = x_{1:k}|\mathcal{M})} \\ &= \frac{\alpha_k(s)}{\sum_{z \in \mathcal{Z}} \alpha_k(z)}.\end{aligned}$$

这一推导给出了一个计算过滤的算法.

然后是平滑：$\Pr(Z_k = s|X = x, \mathcal{M})$，$k < t$。回顾记号 $\alpha_k(s) = \Pr(X_{1:k} = x_{1:k}, Z_k = s|\mathcal{M})$，以及 $\beta_k(s) = \Pr(X_{k+1:t} = x_{k+1:t}|Z_k = s, \mathcal{M})$. 可以证明：

$$\Pr(z_k = s|X = x, \mathcal{M}) = \frac{\beta_k(s)\alpha_k(s)}{\sum_{z \in \mathcal{Z}} \alpha_t(z)}.$$

这一推导给出了一个计算平滑的算法.

之后是预测：$\Pr(Z_k = s|X = x, \mathcal{M})$，$k > t$。首先用过滤计算 $\lambda = \Pr(Z_t = s|X = x, \mathcal{M})$. 然后用 λ 作为 Markov 的初始状态，利用 Kolmogorov-Chapman 方程向前计算 $k - t$ 步.

最后是解码。定义

$$\delta_k(s) = \max_{Z_{1:k-1}} \Pr(Z_{1:k} = (z_{1:k-1}, s), X_{1:k} = x_{1:k}|\mathcal{M}).$$

根据一步转移，我们有

$$\delta_{k+1}(s) = \max_{q \in \mathcal{Z}} \{\delta_k(q) T_{q,s}\} M_{s,x_{k+1}}.$$

问题转化为: 记录最高概率的路径，这是一个动态规划问题.

我们有 *Viterbi* 算法，如下:

- 初始化:
  - $\delta_1(s) = \lambda(s) M_{s,z_1}$.
  - $\mathrm{Pre}_1(s) = \varnothing$.
- 对 $k = 1, 2, \ldots, t-1$，$s \in \mathcal{Z}$:
  - $\delta_{k+1}(s) = \max_{q \in \mathcal{Z}} \{\delta_k(q) T_{q,s}\} M_{s,x_{k+1}}$.
  - $\mathrm{Pre}_{k+1}(s) = \mathrm{argmax}_{q \in \mathcal{Z}} \{\delta_k(q) T_{q,s}\}$.
- $z_t = \mathrm{argmax}_{s \in \mathcal{Z}} \delta_t(s)$.
- 对 $1 \le k < t$，$z_k = \mathrm{Pre}_{k+1}(z_{k+1})$.
- 时间复杂度: $\mathcal{O}(t|\mathcal{Z}|^2)$.

第二部分

信息与数据

第三部分

决策与优化

第四部分

逻辑与博弈

第五部分

认知逻辑

参考文献

[Bre57] Leo Breiman. The Individual Ergodic Theorem of Information Theory. *The Annals of Mathematical Statistics*, 28(3):809–811, 1957.

[CT12] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[Huf52] David A. Huffman. A Method for the Construction of Minimum-Redundancy Codes. *Proceedings of the IRE*, 40(9):1098–1101, September 1952.

[Inf] Information | Etymology, origin and meaning of information by etymonline. https://www.etymonline.com/word/information.

[Jay02] Edwin T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2002.

[KL51] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[LLG+19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, October 2019.

[McM53] Brockway McMillan. The Basic Theorems of Information Theory. *The Annals of Mathematical Statistics*, 24(2):196–219, June 1953.

[RHW86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, pages 318–362. MIT Press, Cambridge, MA, USA, January 1986.

[Rob49] Robert M. Fano. *The Transmission of Information*. March 1949.

[Sha48] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948.

[Shi96] A. N. Shiryaev. *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer, New York, NY, 1996.

[Tin62] Hu Kuo Ting. On the Amount of Information. *Theory of Probability & Its Applications*, 7(4):439–447, January 1962.

[Uff22] Jos Uffink. Boltzmann's Work in Statistical Physics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2022 edition, 2022.

[李 10] 李贤平. 概率论基础. 高等教育出版社, 2010.

索引