第 6 讲 熵率

马尔可夫链
熵率
例子：加权图上随机游动的熵率
马尔可夫链的函数
例题

随机过程 $\{X_i\}$ 是一个带下标（表示时间）的随机变量序列. 一般允许随机变量间具有任意的相关性. 为刻画一个过程我们需要知道所有有限的联合概率密度函数

$$P\{(X_1, X_2, \cdots, X_n) = (x_1, x_2, \cdots, x_n)\} = p(x_1, x_2, \cdots, x_n)$$

其中 $(x_1, x_2, \cdots, x_n) \in \mathcal{X}^n$， $n = 1, 2, \cdots$.

定义 1.1

如果随机变量序列的任意有限子集的联合分布关于时间下标的位移不变，即对于每个 n 和位移 l，以及任意的 $x_1, x_2, \cdots, x_n \in \mathcal{X}$，均满足

$$\begin{aligned} P\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\} \\ = P\{X_{l+1} = x_1, X_{l+2} = x_2, \cdots, X_{l+n} = x_n\}. \end{aligned}$$

则称该随机过程是平稳的.

定义 1.2: 如

对 $n=1,2,\cdots$，及所有的 $x_1,x_2,\cdots,x_n\in\mathcal{X}$，有

$$P\{X_{n+1}=x_{n+1}|X_n=x_n,X_{n-1}=x_{n-1},\cdots,X_1=x_1\}$$
$$=P\{X_{n+1}=x_{n+1}|X_n=x_n\}.$$

则称该离散随机过程为**马尔可夫链**或**马尔可夫过程**.

定义 1.3: 如

对 $n = 1, 2, \cdots$，及所有的 $x_1, x_2, \cdots, x_n \in \mathcal{X}$，有

$$P\{X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \cdots, X_1 = x_1\} = P\{X_{n+1} = x_{n+1} | X_n = x_n\}.$$

则称该离散随机过程为**马尔可夫链**或**马尔可夫过程**.

此时，随机变量的联合概率密度函数可以写为

$$p(x_1, x_2, \cdots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2)\cdots p(x_n|x_{n-1}).$$

定义 1.4

如果条件概率 $p(x_{n+1}|x_n)$ 不依赖于 n，即对 $n = 1, 2, \cdots$，有

$$P\{X_{n+1} = j|X_n = i\} = P\{X_2 = j|X_1 = i\} = P_{ij} \text{ 对任意} i, j \in \mathcal{X}.$$

则称该马尔可夫链是时齐的（或称是时间不变的）. 此时，称 P_{ij} 为从 i 到 j 的（一步）概率转移概率. 若将 P_{ij} 作为第 i 行第 j 列的元素，我们得到一个矩阵 $\mathbf{P} = (P_{ij})_{S\times S}$，称之为 $\{X_n\}$ 的（一步）概率转移矩阵. 以后，如果没有特别说明，我们提到的马尔可夫链都默认是时齐的.

若马尔可夫链可以从任意状态经过有限步转移到另一任意状态，且转移概率为正，则称该马尔可夫链是不可约的. 如果从一个状态转移到它自身的不同路径长度的最大公因子为 1，则称该马尔可夫链是非周期的. 用严格的数学语言来说，称 $\{n \geq 1 : p_{ii}^{(n)} > 0\}$ 的最大公约数为状态 i 的周期，若对于每个状态其周期均为 1，则称该马尔可夫链是非周期的.

若马尔可夫链可以从任意状态经过有限步转移到另一任意状态，且转移概率为正，则称该马尔可夫链是不可约的. 如果从一个状态转移到它自身的不同路径长度的最大公因子为 1，则称该马尔可夫链是非周期的. 用严格的数学语言来说，称 $\{n \geq 1 : p_{ii}^{(n)} > 0\}$ 的最大公约数为状态 i 的周期，若对于每个状态其周期均为 1，则称该马尔可夫链是非周期的.
对于一个马尔可夫链，如果在时间 n 随机变量的概率密度函数为 $p(x_n)$，那么在 $n+1$ 时刻，随机变量的概率密度函数为

$$p(x_{n+1}) = \sum_{x_n} p(x_n) P_{x_n x_{n+1}}.$$

若在 $n+1$ 时刻，状态空间上的分布与在 n 时刻的分布相同，则称此分布为平稳分布. 于是，平稳分布 μ 为下列方程组的解：

$$\mu_j = \sum_i \mu_i P_{ij} \text{ 对任意的 } j.$$

命题 1.5

如果马尔可夫链的初始状态服从平稳分布，那么该马尔可夫链为平稳过程.

证明.
设初始分布为 μ. 若初始状态服从平稳分布，$P(X_0 = i) = \mu_i$. 于是 $P(X_1 = j) = \sum_i P(X_0 = i)P_{ij} = \sum_i \mu_i P_{ij} = \mu_j$. 下面我们来归纳证明 $P(X_n = j) = \mu_j$. 如果我们有 $P(X_k = j) = \mu_j$，下面我们来证明 $P(X_{k+1} = j) = \mu_j$. 我们有

$$P(X_{k+1} = j) = \sum_i P(X_k = i)P(X_{k+1} = j | X_k = i) = \sum_i \mu_i P_{ij} = \mu_j.$$

从而得证.

证明.
下面我们用归纳法对 m 归纳证明

$$P(X_{k+1}=j_1,\cdots,X_{k+m}=j_m)=P(X_1=j_1,\cdots,X_{k+m}=j_m).$$

假设上式对 $m-1$ 成立，下面我们来考虑 m 的情形.

$$\begin{aligned}
& P(X_{k+m}=j_m,\cdots,X_{k+1}=j_1) \\
= \; & P(X_{k+m}=j_m|X_{k+m-1}=j_{m-1},\cdots,X_{k+1}=j_1)P(X_{k+m-1}=j_{m-1},\cdots,X_{k+1}=j_1) \\
= \; & P(X_{k+m}=j_m|X_{k+m-1}=j_{m-1})P(X_{k+m-1}=j_{m-1},\cdots,X_{k+1}=j_1) \\
= \; & P(X_m=j_m|X_{m-1}=j_{m-1})P(X_{m-1}=j_{m-1},\cdots,X_1=j_1) \\
= \; & P(X_m=j_m|X_{m-1}=j_{m-1},\cdots,X_1=j_1)P(X_{m-1}=j_{m-1},\cdots,X_1=j_1) \\
= \; & P(X_m=j_m,X_{m-1}=j_{m-1},\cdots,X_1=j_1).
\end{aligned}$$

从而我们知该马尔可夫链是平稳的. $\square$

若有限状态马尔科夫链是不可约的和非周期的，则它的平稳分布惟一，从任意的初始分布出发，当 $n \to \infty$ 时，X_n 的分布必趋向于此平稳分布. 有兴趣的读者可参见随机过程相关书籍.

例 1.6

考虑一个两状态马尔可夫链，它的概率转移矩阵为

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}.$$

设向量 μ 表示平稳分布，则通过解方程 $\mu P = \mu$ 可求得 μ. 在时刻 n 时状态 X_n 的熵为

$$H(X_n) = H(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}).$$

马尔可夫链

熵率

例子：加权图上随机游动的熵率

马尔可夫链的函数

例题

定义 2.1

当如下极限存在时，随机过程 $\{X_i\}$ 的熵率定义为

$$H(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \cdots, X_n). \tag{2.1}$$

例 2.2: 打字机

假定一台打字机可以输出 m 个等可能的字母. 由此打字机可以产生长度为 n 的 m 个序列，并且都等可能出现. 因此，$H(X_1, X_2, \cdots, X_n) = \log m^n$，从而熵率为 $H(\mathcal{X}) = \log m$ 比特/字符.

例 2.3

设 $X_1, X_2, \cdots, X_n$ 为一列独立同分布的随机变量序列. 此时，有

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{H(X_1, X_2, \cdots, X_n)}{n} = \lim_{n\to\infty} \frac{nH(X_1)}{n} = H(X_1).$$

例 2.4

若 $X_1, X_2, \cdots, X_n$ 独立但不同分布，则

$$H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i).$$

我们可以选择 $X_1, X_2, \cdots$ 的一个分布序列，使 $\frac{1}{n}\sum H(X_i)$ 的极限不存在. 例如我们考虑二值随机分布序列，其中 $p_i = P(X_i = 1)$ 不是常数，而为 i 得函数. 如果我们定义

$$p_i = \begin{cases} 0.5 & 2k < \log\log i \leq 2k+1 \\ 0 & 2k+1 < \log\log i \leq 2k+2 \end{cases}$$

此时，该序列的情况是，满足 $H(X_i) = 1$ 的随机变量序列（可以任意长）之后，紧接着是更长以指数变化的序列满足 $H(X_i) = 0$. 所以，$H(X_i)$ 的累计平均值将在 0 与 1 之间震荡，从而不存在极限. 因此，该过程的 $H(\mathcal{X})$ 无定义.

我们也可以定义熵率的一个相关的量:

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_{n-1}, X_{n-2}, \cdots, X_1) \tag{2.2}$$

定理 2.5

对于平稳随机过程，式 (2.1) 和 (2.2) 中的极限均存在且相等：

$$H(\mathcal{X}) = H'(\mathcal{X}).$$

定理 2.6

对于平稳随机过程，$H(X_n|X_{n-1},\cdots,X_1)$ 随 n 递减且存在极限 $H'(\mathcal{X})$.

定理 2.7

对于平稳随机过程，$H(X_n|X_{n-1},\cdots,X_1)$ 随 n 递减且存在极限 $H'(\mathcal{X})$.

证明.

$$\begin{aligned} H(X_{n+1}|X_1,X_2,\cdots,X_n) &\leq H(X_{n+1}|X_n,\cdots,X_2) \\ &= H(X_n|X_{n-1},\cdots,X_1) \end{aligned}$$

其中的不等式由条件作用使熵减小这个性质得到，而等式由该过程的平稳性得到. 由于 $H(X_n|X_{n-1},\cdots,X_1)$ 是非负且递减的数列，故知极限 $H'(\mathcal{X})$ 存在. □

引理 2.8

若 $a_n \to a$，且 $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i$，则 $b_n \to a$.

定理2.5的证明.
由链式法则，我们有

$$\frac{H(X_1, X_2, \cdots, X_n)}{n} = \frac{1}{n}\sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1).$$

于是由引理2.8，我们知

$$H(\mathcal{X}) = \lim \frac{H(X_1, X_2, \cdots, X_n)}{n} = \lim H(X_n|X_{n-1}, \cdots, X_1) = H'(\mathcal{X}). \quad \square$$

下面我们来计算平稳马尔可夫链这一特殊的随机过程的熵率. 我们有

$$H(\mathcal{X}) = H'(\mathcal{X}) = \lim H(X_n|X_{n-1},\cdots,X_1) = \lim H(X_n|X_{n-1}) = H(X_2|X_1).$$

其中第三个等式是因为

$$\begin{aligned}
& H(X_n|X_{n-1},\cdots,X_1) \\
= & -\sum_{x_1,\cdots,x_n} p(x_1,x_2,\cdots,x_{n-1})p(x_n|x_{n-1},x_{n-2},\cdots,x_1)\log p(x_n|x_{n-1},x_{n-2},\cdots,x_1) \\
= & -\sum_{x_1,\cdots,x_n} p(x_1,x_2,\cdots,x_{n-1})p(x_n|x_{n-1})\log p(x_n|x_{n-1}) \\
= & -\sum_{x_{n-1},x_n} p(x_n|x_{n-1})\log p(x_n|x_{n-1}) \sum_{x_1,\cdots,x_{n-2}} p(x_1,x_2,x_{n-1}) \\
= & -\sum_{x_{n-1},x_n} p(x_{n-1})p(x_n|x_{n-1})\log p(x_n|x_{n-1}) \\
= & -\sum_{x_{n-1},x_n} p(x_{n-1},x_n)\log p(x_n|x_{n-1}) \\
= & \ H(X_n|X_{n-1}).
\end{aligned}$$

条件熵可以根据给出的平稳分布计算得到. 注意到，平稳分布 μ 为下列方程组的解：

$$\mu_j = \sum_i \mu_i P_{ij} \text{ 对任意的} j.$$

定理 2.9

设 $\{X_i\}$ 为平稳马尔可夫链，其平稳分布为 μ，转移矩阵为 P. 则熵率为

$$H(\mathcal{X}) = -\sum_{ij} \mu_i P_{ij} \log P_{ij}.$$

证明.

$$H(\mathcal{X}) = H(X_2|X_1) = \sum_i \mu_i \Big(\sum_j -P_{ij} \log P_{ij}\Big). \qquad \square$$

例 2.10

我们可以计算前面提到的两状态马尔可夫链的熵率为

$$H(\mathcal{X}) = H(X_2|X_1) = \frac{\beta}{\alpha+\beta}H(\alpha) + \frac{\alpha}{\alpha+\beta}H(\beta).$$

注记 2.11

若马尔可夫链是不可约的，那么该马尔可夫链存在状态空间上的惟一平稳分布，并且给定任意的初始分布，当 $n \to \infty$ 时，分布必趋向于此平稳分布. 此时，即使初始分布不是平稳分布，熵率也如式 (2.1) 和 (2.2) 中给出的 $H(\mathcal{X})$.

我们考虑一个连通图上的随机游动，假定该图有 m 个顶点，标记为 $\{1,2,\cdots,m\}$，其中连接节点 i 和 j 的边权重为 W_{ij}. 假定此图是无向的，即 $W_{ij}=W_{ji}$. 若节点 i 和 j 没有连接边，则设 $W_{ij}=0$.

有一个粒子在图中由一个节点到另一个节点作随机游动. 设随机游动 $\{X_n\}$，$X_n\in\{1,2,\cdots,m\}$ 为图的一个顶点序列. 若 $X_n=i$，那么下一个顶点 j 只可能是与节点 i 相连的所有节点中的一个，且转移概率为连接 i 和 j 的边权重所占所有与 i 相连的边的权重之和的比例. 因此，$P_{ij}=W_{ij}/\sum_k W_ik$.

对此情形，平稳分布有一个非常简单的形式. 设

$$W_i = \sum_j W_{ij}$$

为连接节点 i 的所有的边的权重总和，再设

$$W = \sum_{i,j:j\geq i} W_{ij}$$

为图中所有的边权重总和，则 $\sum_i W_i = 2W$.

现在我们猜测平稳分布为

$$\mu_i = \frac{W_i}{2W}.$$

不难验证上述分布的确为平稳分布. 详细地说，我们有

$$\sum_i \mu_i P_{ij} = \sum_i \frac{W_i}{2W}\frac{W_{ij}}{W_i} = \sum_i \frac{1}{2W} W_{ij} = \frac{W_j}{2W} = \mu_j.$$

因此，状态 i 的平稳概率为连接节点 i 的各边权重总和占所有的边权重总和的比例. 此平稳分布仅依赖于总权重以及与该节点相连的所有的边权重之和.

通过计算，熵率为

$$
\begin{aligned}
H(\mathcal{X}) &= H(X_2|X_1) \\
&= -\sum_i \mu_i \sum_j P_{ij} \log P_{ij} \\
&= -\sum_i \frac{W_i}{2W} \sum_j \frac{W_{ij}}{W_i} \log \frac{W_{ij}}{W_i} \\
&= -\sum_i \sum_j \frac{W_{ij}}{2W} \log \frac{W_{ij}}{W_i} \\
&= -\sum_i \sum_j \frac{W_{ij}}{2W} \log \frac{W_{ij}}{2W} + \sum_i \sum_j \frac{W_{ij}}{2W} \log \frac{W_i}{2W} \\
&= H(\cdots, \frac{W_{ij}}{2W}, \cdots) - H(\cdots, \frac{W_i}{2W}, \cdots).
\end{aligned}
$$

如果所有的边有相同的权重，则平稳分布可设置成在节点 i 上为 $E_i/2E$，其中 E_i 表示连接节点 i 的边数，E 表示该图的边的总数. 此时，随机游动的熵率为

$$H(\mathcal{X}) = \log(2E) - H(\frac{E_1}{2E}, \frac{E_2}{2E}, \cdots, \frac{E_m}{2E}).$$

- 设 $X_1, X_2, \cdots, X_n, \cdots$ 为平稳马尔可夫链.
- 再设 $Y_i = \phi(X_i)$ 是一个随机过程, 其中 ϕ 是一个 (可测) 函数.
- 此时熵率 $H(\mathcal{Y})$ 为多少?
- 若 $Y_1, Y_2, \cdots, Y_n$ 也构成一个马尔可夫链, 问题就会简单许多.
- 但实际情况往往并非如此.
- 由于原马尔可夫链是平稳的, 则 $Y_1, Y_2, \cdots, Y_n$ 也是平稳的, 从而可以明确定义熵率.
- 若要计算 $H(\mathcal{Y})$, 我们可能会先对每个 n 计算出 $H(Y_n|Y_{n-1}, \cdots, Y_1)$ 的值, 然后求其极限.
- 在实际应用中, 我们希望了解 $H(Y_n|Y_{n-1}, \cdots, Y_1)$ 与 $H(\mathcal{Y})$ 的接近程度.

引理 4.1

$$H(Y_n|Y_{n-1},\cdots,Y_2,X_1) \leq H(\mathcal{Y}).$$

证明.

对 $k=1,2,\cdots,$ 有

$$\begin{aligned} H(Y_n|Y_{n-1},\cdots,Y_2,X_1) &= H(Y_n|Y_{n-1},\cdots,Y_2,Y_1,X_1) \\ &= H(Y_n|Y_{n-1},\cdots,Y_1,X_1,X_0,X_{-1},\cdots,X_{-k}) \\ &= H(Y_n|Y_{n-1},\cdots,Y_1,X_1,X_0,X_{-1},\cdots,X_{-k},Y_0,\cdots,Y_{-k}) \\ &\le H(Y_n|Y_{n-1},\cdots,Y_1,Y_0,\cdots,Y_{-k}) \\ &= H(Y_{n+k+1}|Y_{n+k},\cdots,Y_1) \end{aligned}$$

两边取极限，我们知

$$\begin{aligned} H(Y_n|Y_{n-1},\cdots,Y_1,X_1) &\le \lim_k H(Y_{n+k+1}|Y_{n+k},\cdots,Y_1) \\ &= H(\mathcal{Y}). \end{aligned}$$

从而引理得证. $\square$

引理 4.2

$$H(Y_n|Y_{n-1},\cdots,Y_1) - H(Y_n|Y_{n-1},\cdots,Y_1,X_1) \to 0.$$

证明.

首先注意到

$$
\begin{aligned}
&H(Y_n|Y_{n-1},\cdots,Y_1)-H(Y_n|Y_{n-1},\cdots,Y_1,X_1)\\
=&I(X_1;Y_n|Y_{n-1},\cdots,Y_1)
\end{aligned}
$$

由互信息的性质，可得 $I(X_1;Y_1,Y_2,\cdots,Y_n)\le H(X_1)$ 且 $I(X_1;Y_1,Y_2,\cdots,Y_n)$ 随 n 递增. 因此 $\lim I(X_1;Y_1,Y_2,\cdots,Y_n)$ 存在且满足 $\lim_{n\to\infty} I(X_1;Y_1,Y_2,\cdots,Y_n)\le H(X_1)$. 由链式法则，

$$
\begin{aligned}
H(X_1)&\ge \lim_{n\to\infty} I(X_1;Y_1,Y_2,\cdots,Y_n)\\
&=\lim_{n\to\infty}\sum_{i=1}^{n} I(X_i;Y_i|Y_{i-1},\cdots,Y_1)\\
&=\sum_{i=1}^{\infty} I(X_1;Y_i|Y_{i-1},\cdots,Y_1).
\end{aligned}
$$

注意到上面的正项级数收敛，所以其通项趋于 0，于是

$$
\lim_{n\to\infty} I(X_1;Y_n|Y_{n-1},\cdots,Y_1)=0,
$$

引理得证.

定理 4.3

设 $X_1, X_2, \cdots, X_n, \cdots$ 为平稳马尔可夫链, 且 $Y_i = \phi(X_i)$, 那么

$$H(Y_n|Y_{n-1}, \cdots, Y_1, X_1) \le H(\mathcal{Y}) \le H(Y_n|Y_{n-1}, \cdots, Y_1)$$

且

$$\lim H(Y_n|Y_{n-1}, \cdots, Y_1, X_1) = H(\mathcal{Y}) = \lim H(Y_n|Y_{n-1}, \cdots, Y_1).$$

例 5.1

设 $Z_1, Z_2, \cdots$ 为独立同分布的随机变量，服从 $\{0, 1\}$ 上的均匀分布. 对于 $1 \le i \le n$，设

$$X_i = \sum_{j=1}^{i} Z_j.$$

求 $I(X_1; X_2, X_3, \cdots, X_n)$.

证明.

首先注意到 $X_1 \to X_2 \to \cdots \to X_n$ 构成一个马尔可夫链. (思考：为什么？)

由互信息的链式法则, 我们有

$$\begin{aligned}
I(X_1; X_2, X_3, \cdots, X_n) &= \sum_{i=2}^{n} I(X_1; X_i | X_2, \cdots, X_{i-1}) \\
&= I(X_1; X_2). \\
&= I(Z_1; Z_1 + Z_2) \\
&= H(Z_1 + Z_2) - H(Z_1 + Z_2 | Z_1) \\
&= \frac{3}{2} - 1 = 1/2\text{比特}.
\end{aligned}$$

□