

马尔可夫链和熵率

- ① 马尔可夫链
- ② 熵率
- ③ 条件 Huffman 编码
- ④ 英文文本压缩

- 已讲解：独立同分布随机序列压缩
- 下面讲解：非独立同分布随机序列—马尔可夫链压缩

马尔可夫链和熵率

- ① 马尔可夫链
- ② 熵率
- ③ 条件 Huffman 编码
- ④ 英文文本压缩

马尔可夫链

考虑随机序列：

$$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}, X_{14}, X_{15} \dots$$

联合分布: $\Pr[X_1, X_2, \dots, X_n]$

马尔可夫链

考虑随机序列：

$$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}, X_{14}, X_{15} \dots$$

联合分布: $\Pr[X_1, X_2, \dots, X_n]$

令 $X_t \in \mathcal{X} = \{1, \dots, m\}$

- \mathcal{X} 为 状态空间
- $x \in \mathcal{X}$ 为 状态

马尔可夫链

考虑随机序列：

$$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}, X_{14}, X_{15} \dots$$

联合分布: $\Pr[X_1, X_2, \dots, X_n]$

令 $X_t \in \mathcal{X} = \{1, \dots, m\}$

- \mathcal{X} 为 状态空间
- $x \in \mathcal{X}$ 为 状态

● $s_t(x)$: t 时刻状态为 x 的 概率

● t 时刻的 状态空间向量为

$$\mathbf{s}_t = [s_t(1), s_t(2), \dots, s_t(m)]$$

平稳过程

定义（平稳随机过程）：对任意 n 和任意 k ，若联合概率分布满足以下平移不变性：

$$\Pr[X_1, \dots, X_n] = \Pr[X_{1+k}, \dots, X_{n+k}]$$

则称该随机过程是平稳的。

平稳过程

定义（平稳随机过程）：对任意 n 和任意 k ，若联合概率分布满足以下平移不变性：

$$\Pr[X_1, \dots, X_n] = \Pr[X_{1+k}, \dots, X_{n+k}]$$

则称该随机过程是平稳的。

平稳随机过程具有“平移不变”的概率分布：

$$\underbrace{X_1, X_2, X_3}_{\Pr[X_1, X_2, X_3]}, X_4, X_5, X_6, X_7, X_8, X_9, \underbrace{X_{10}, X_{11}, X_{12}}_{\Pr[X_8, X_9, X_{10}]}, X_{13}, X_{14}, X_{15} \dots$$

$$\Rightarrow \Pr[X_1, X_2, X_3] = \Pr[X_8, X_9, X_{10}]$$

马尔可夫链

定义 (一阶马尔可夫链) : 一个离散、平稳的随机过程 X_1, X_2, \dots 若满足:

$$\Pr[X_{t+1}|X_t, \dots, X_1] = \Pr[X_{t+1}|X_t],$$

则称为一阶马尔可夫链或马尔可夫过程。

马尔可夫链

定义 (一阶马尔可夫链)：一个离散、平稳的随机过程 X_1, X_2, \dots 若满足：

$$\Pr[X_{t+1}|X_t, \dots, X_1] = \Pr[X_{t+1}|X_t],$$

则称为一阶马尔可夫链或马尔可夫过程。

未来 取决于 现在 而非 过去
 X_{t+1} X_t X_{t-1}

状态转移矩阵

- 时不变马尔可夫链可由转移概率 $p_{i,j}$ 确定:

$$p_{j,i} = \Pr[X_{t+1} = j | X_t = i]$$

状态转移矩阵

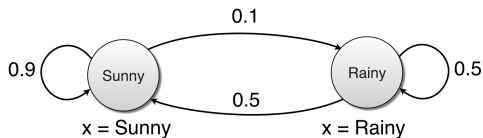
- 时不变马尔可夫链可由转移概率 $p_{i,j}$ 确定:

$$p_{j,i} = \Pr[X_{t+1} = j | X_t = i]$$

- 转移概率可以写成 $m \times m$ 的状态转移矩阵 P :

$$P = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,m} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m,1} & p_{m,2} & \cdots & p_{m,m} \end{bmatrix}$$

马尔可夫链举例

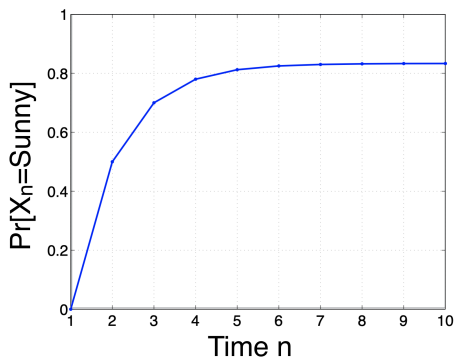


$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}$$

马尔可夫链举例

初始状态:

$$p_1 = \begin{bmatrix} \Pr[X_1 = \text{Sunny}] = 0 \\ \Pr[X_1 = \text{Rainy}] = 1 \end{bmatrix}$$

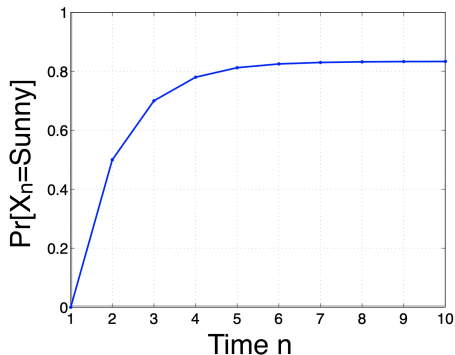


马尔可夫链举例

初始状态:

$$p_1 = \begin{bmatrix} \Pr[X_1 = \text{Sunny}] = 0 \\ \Pr[X_1 = \text{Rainy}] = 1 \end{bmatrix}$$

$$p_2 = Pp_1 = P \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$



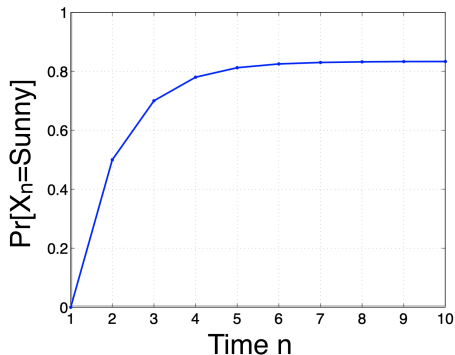
马尔可夫链举例

初始状态:

$$p_1 = \begin{bmatrix} \Pr[X_1 = \text{Sunny}] = 0 \\ \Pr[X_1 = \text{Rainy}] = 1 \end{bmatrix}$$

$$p_2 = Pp_1 = P \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$

$$p_3 = Pp_2 = P \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 0.7 \\ 0.3 \end{bmatrix}$$



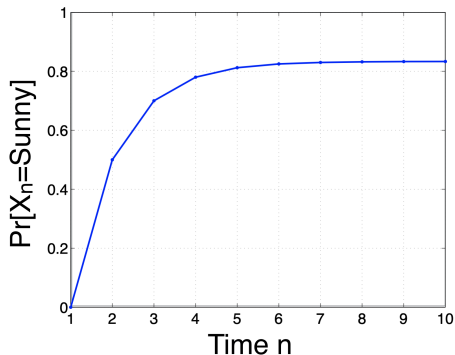
马尔可夫链举例

初始状态:

$$p_1 = \begin{bmatrix} \Pr[X_1 = \text{Sunny}] = 0 \\ \Pr[X_1 = \text{Rainy}] = 1 \end{bmatrix}$$

$$p_2 = Pp_1 = P \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$

$$p_3 = Pp_2 = P \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 0.7 \\ 0.3 \end{bmatrix}$$



马尔可夫链举例

初始状态:

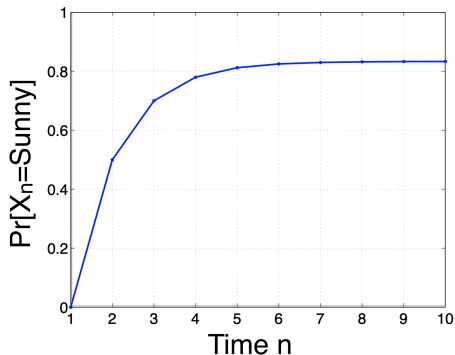
$$p_1 = \begin{bmatrix} \Pr[X_1 = \text{Sunny}] = 0 \\ \Pr[X_1 = \text{Rainy}] = 1 \end{bmatrix}$$

$$p_2 = Pp_1 = P \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$

$$p_3 = Pp_2 = P \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 0.7 \\ 0.3 \end{bmatrix}$$

$$p_4 = \begin{bmatrix} 0.78 \\ 0.22 \end{bmatrix}$$

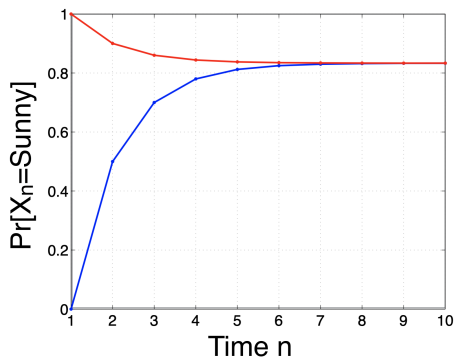
\vdots



马尔可夫链举例

初始状态:

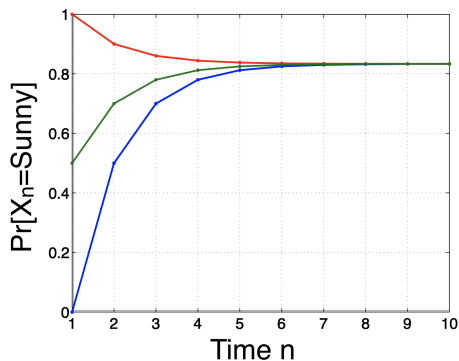
$$\mathbf{p}_1 = \begin{bmatrix} \Pr[X_1 = \text{Sunny}] = 1 \\ \Pr[X_1 = \text{Rainy}] = 0 \end{bmatrix}$$



马尔可夫链举例

初始状态:

$$\mathbf{p}_1 = \begin{bmatrix} \Pr[X_1 = \text{Sunny}] = 0.5 \\ \Pr[X_1 = \text{Rainy}] = 0.5 \end{bmatrix}$$



平稳分布

定义：满足以下条件的状态分布向量 z 称为平稳分布：

$$z = Pz \quad \Rightarrow \quad (P - I)z = 0 \quad (1)$$

平稳分布

定义：满足以下条件的状态分布向量 z 称为平稳分布：

$$z = Pz \Rightarrow (P - I)z = 0 \quad (1)$$

z 为概率分布向量，

$$[1 \ \cdots \ 1]z = 1. \quad (2)$$

平稳分布

定义：满足以下条件的状态分布向量 z 称为平稳分布：

$$z = Pz \Rightarrow (P - I)z = 0 \quad (1)$$

z 为概率分布向量，

$$[1 \ \cdots \ 1]z = 1. \quad (2)$$

求解：由 (1) 和 (2), (假设 Q 可逆)

$$\underbrace{\begin{bmatrix} 1 & \cdots & 1 \\ \{P - I\}_{\setminus r_1} \end{bmatrix}}_Q z = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \Rightarrow z = Q^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

其中 r_1 是 $P - I$ 的第一行 (任一行都行)。即，平稳分布 z 是 Q^{-1} 的第一列。

平稳分布的计算

计算下面这个马尔可夫链的平稳概率:

$$P = \begin{bmatrix} 0.8 & 0.5 & 0.7 \\ 0.1 & 0.5 & 0.2 \\ 0.1 & 0 & 0.1 \end{bmatrix}$$

马尔可夫链和熵率

- ① 马尔可夫链
- ② 熵率
- ③ 条件 Huffman 编码
- ④ 英文文本压缩

熵率

$$H(X_1) \leq \log |\mathcal{X}_1|$$

$$H(X_1, X_2) \leq \log |\mathcal{X}_1| + \log |\mathcal{X}_2|$$

$$\vdots$$

$$H(X_1, \dots, X_n) \leq \sum_{i=1}^n \log |\mathcal{X}_i|$$

平均比特数:

$$\frac{1}{n} H(X_1, \dots, X_n)$$

熵率的定义

定义: 随机过程 X_1, X_2, \dots 的熵率定义为 (假设极限存在):

$$H(\mathcal{X}) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, \dots, X_n)$$

熵率的定义

定义: 随机过程 X_1, X_2, \dots 的熵率定义为 (假设极限存在):

$$H(\mathcal{X}) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, \dots, X_n)$$

定义: 条件熵率定义为:

$$H'(\mathcal{X}) = \lim_{n \rightarrow \infty} H(X_n | X_{n-1}, \dots, X_1)$$

熵率的性质

性质: 对于平稳随机过程, $H(X_n|X_{n-1}, \dots, X_1)$ 关于 n 是非增的, 且有极限值 $H'(\mathcal{X})$.

熵率的性质

性质: 对于平稳随机过程, $H(X_n|X_{n-1}, \dots, X_1)$ 关于 n 是非增的, 且有极限值 $H'(\mathcal{X})$.

$$H(X_n|X_{n-1}, \dots, X_1) \leq H(X_n|X_{n-1}, \dots, X_2)$$

熵率的性质

性质: 对于平稳随机过程, $H(X_n|X_{n-1}, \dots, X_1)$ 关于 n 是非增的, 且有极限值 $H'(\mathcal{X})$.

$$\begin{aligned} H(X_n|X_{n-1}, \dots, X_1) &\leq H(X_n|X_{n-1}, \dots, X_2) \\ &= H(X_{n-1}|X_{n-2}, \dots, X_1) \quad \text{平稳特性} \end{aligned}$$

熵率的性质

性质: 对于平稳随机过程, $H(X_n|X_{n-1}, \dots, X_1)$ 关于 n 是非增的, 且有极限值 $H'(\mathcal{X})$.

$$\begin{aligned} H(X_n|X_{n-1}, \dots, X_1) &\leq H(X_n|X_{n-1}, \dots, X_2) \\ &= H(X_{n-1}|X_{n-2}, \dots, X_1) \quad \text{平稳特性} \end{aligned}$$

性质: 对于一个平稳随机过程:

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

熵率的性质

性质: 对于平稳随机过程, $H(X_n|X_{n-1}, \dots, X_1)$ 关于 n 是非增的, 且有极限值 $H'(\mathcal{X})$.

$$\begin{aligned} H(X_n|X_{n-1}, \dots, X_1) &\leq H(X_n|X_{n-1}, \dots, X_2) \\ &= H(X_{n-1}|X_{n-2}, \dots, X_1) \quad \text{平稳特性} \end{aligned}$$

性质: 对于一个平稳随机过程:

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

$$H(\mathcal{X}) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, \dots, X_n)$$

熵率的性质

性质: 对于平稳随机过程, $H(X_n|X_{n-1}, \dots, X_1)$ 关于 n 是非增的, 且有极限值 $H'(\mathcal{X})$.

$$\begin{aligned} H(X_n|X_{n-1}, \dots, X_1) &\leq H(X_n|X_{n-1}, \dots, X_2) \\ &= H(X_{n-1}|X_{n-2}, \dots, X_1) \quad \text{平稳特性} \end{aligned}$$

性质: 对于一个平稳随机过程:

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

$$\begin{aligned} H(\mathcal{X}) &= \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, \dots, X_n) \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n H(X_i|X_{i-1}, \dots, X_1) \quad \text{链式法则} \end{aligned}$$

熵率的性质

性质: 对于平稳随机过程, $H(X_n|X_{n-1}, \dots, X_1)$ 关于 n 是非增的, 且有极限值 $H'(\mathcal{X})$.

$$\begin{aligned} H(X_n|X_{n-1}, \dots, X_1) &\leq H(X_n|X_{n-1}, \dots, X_2) \\ &= H(X_{n-1}|X_{n-2}, \dots, X_1) \quad \text{平稳特性} \end{aligned}$$

性质: 对于一个平稳随机过程:

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

$$\begin{aligned} H(\mathcal{X}) &= \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, \dots, X_n) \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n H(X_i|X_{i-1}, \dots, X_1) \quad \text{链式法则} \\ &= \lim_{n \rightarrow \infty} H(X_n|X_{n-1}, \dots, X_1) \quad \text{平稳后条件熵率占主导} \end{aligned}$$

熵率的性质

性质: 对于平稳随机过程, $H(X_n|X_{n-1}, \dots, X_1)$ 关于 n 是非增的, 且有极限值 $H'(\mathcal{X})$.

$$\begin{aligned} H(X_n|X_{n-1}, \dots, X_1) &\leq H(X_n|X_{n-1}, \dots, X_2) \\ &= H(X_{n-1}|X_{n-2}, \dots, X_1) \quad \text{平稳特性} \end{aligned}$$

性质: 对于一个平稳随机过程:

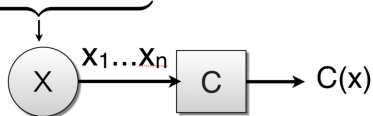
$$H(\mathcal{X}) = H'(\mathcal{X})$$

$$\begin{aligned} H(\mathcal{X}) &= \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, \dots, X_n) \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n H(X_i|X_{i-1}, \dots, X_1) \quad \text{链式法则} \\ &= \lim_{n \rightarrow \infty} H(X_n|X_{n-1}, \dots, X_1) \quad \text{平稳后条件熵率占主导} \\ &= H'(\mathcal{X}) \end{aligned}$$

平稳信源的压缩

平稳信源、马尔可夫信源等等

$X_1, X_2, X_3, \dots, X_n$



平稳信源的压缩

推论: 每符号最优码率 R^* 满足:

$$\frac{1}{n}H(X_1, \dots, X_n) \leq R^* \leq \frac{1}{n}H(X_1, \dots, X_n) + \frac{1}{n}$$

另外, 如果 X_1, \dots, X_n 是一个平稳随机过程,

$$\lim_{n \rightarrow \infty} R^* = H(\mathcal{X})$$

其中 $H(\mathcal{X})$ 是该随机过程的熵率。

马尔可夫信源的压缩

推论: 令 X_1, X_2, \dots 为平稳概率为 \mathbf{z} 、转移矩阵为 \mathbf{P} 的平稳马尔可夫链, 令 $X_1 \sim \mathbf{z}$, 那么熵率为:

$$H(\mathcal{X}) = H(X_n | X_{n-1}) = - \sum_{i=1}^m \sum_{j=1}^m z_i p_{j,i} \log p_{j,i}.$$

注:

$$H(X_n | X_{n-1}) = \sum_{i=1}^m z_i H(X_n | X_{n-1} = s_i),$$

$$H(X_n | X_{n-1} = s_i) = - \sum_{j=1}^m p_{j,i} \log p_{j,i}.$$

马尔可夫链熵率的计算

计算以下马尔可夫链的熵率：

$$P = \begin{bmatrix} 0.9 & 0.2 \\ 0.1 & 0.8 \end{bmatrix}$$

马尔可夫链和熵率

- ① 马尔可夫链
- ② 熵率
- ③ 条件 Huffman 编码
- ④ 英文文本压缩

条件 Huffman 编码

考虑平稳随机过程 X_1, X_2, \dots : $\mathcal{X} = \{1, \dots, m\}$, 平稳概率为 z 、转移矩阵为 $P_{m \times m}$ 。

- 基于 P_{X_1} 编码 X_1 (Huffman)
- 基于 p_i 编码 X_n (给定条件 $x_{n-1} = i$) (Huffman)
(p_i : $P_{m \times m}$ 的第 i 列)

条件 Huffman 编码举例

令 $\mathcal{X} = \{1, 2, 3, 4\}$, $p_{X_1} = [1/2 \ 1/4 \ 1/8 \ 1/8]$, 转移概率矩阵为:

$$P = \begin{bmatrix} 1/2 & 1/4 & 1/8 & 1/8 \\ 1/4 & 1/8 & 1/8 & 1/2 \\ 1/8 & 1/8 & 1/2 & 1/4 \\ 1/8 & 1/2 & 1/4 & 1/8 \end{bmatrix}$$

条件 Huffman 编码举例

令 $\mathcal{X} = \{1, 2, 3, 4\}$, $p_{X_1} = [1/2 \ 1/4 \ 1/8 \ 1/8]$, 转移概率矩阵为:

$$P = \begin{bmatrix} 1/2 & 1/4 & 1/8 & 1/8 \\ 1/4 & 1/8 & 1/8 & 1/2 \\ 1/8 & 1/8 & 1/2 & 1/4 \\ 1/8 & 1/2 & 1/4 & 1/8 \end{bmatrix}$$

$$z = [1/4 \ 1/4 \ 1/4 \ 1/4],$$

条件 Huffman 编码举例

令 $\mathcal{X} = \{1, 2, 3, 4\}$, $p_{X_1} = [1/2 \ 1/4 \ 1/8 \ 1/8]$, 转移概率矩阵为:

$$P = \begin{bmatrix} 1/2 & 1/4 & 1/8 & 1/8 \\ 1/4 & 1/8 & 1/8 & 1/2 \\ 1/8 & 1/8 & 1/2 & 1/4 \\ 1/8 & 1/2 & 1/4 & 1/8 \end{bmatrix}$$

$$z = [1/4 \ 1/4 \ 1/4 \ 1/4], \quad H(z) = 2 \text{ bits}$$

条件 Huffman 编码举例

令 $\mathcal{X} = \{1, 2, 3, 4\}$, $p_{X_1} = [1/2 \ 1/4 \ 1/8 \ 1/8]$, 转移概率矩阵为:

$$P = \begin{bmatrix} 1/2 & 1/4 & 1/8 & 1/8 \\ 1/4 & 1/8 & 1/8 & 1/2 \\ 1/8 & 1/8 & 1/2 & 1/4 \\ 1/8 & 1/2 & 1/4 & 1/8 \end{bmatrix}$$

$$z = [1/4 \ 1/4 \ 1/4 \ 1/4], \quad H(z) = 2 \text{ bits}$$

$$R = 1 \cdot 1/2 + 2 \cdot 1/4 + 2(3 \cdot 1/8) = 1.75 \text{ bits.}$$

条件 Huffman 编码举例

令 $\mathcal{X} = \{1, 2, 3, 4\}$, $p_{X_1} = [1/2 \ 1/4 \ 1/8 \ 1/8]$, 转移概率矩阵为:

$$P = \begin{bmatrix} 1/2 & 1/4 & 1/8 & 1/8 \\ 1/4 & 1/8 & 1/8 & 1/2 \\ 1/8 & 1/8 & 1/2 & 1/4 \\ 1/8 & 1/2 & 1/4 & 1/8 \end{bmatrix}$$

$$z = [1/4 \ 1/4 \ 1/4 \ 1/4], \quad H(z) = 2 \text{ bits}$$

$$R = 1 \cdot 1/2 + 2 \cdot 1/4 + 2(3 \cdot 1/8) = 1.75 \text{ bits.}$$

Huffman 编码: $X_1 \rightarrow \{0 \ 10 \ 110 \ 111\}$ 。

条件 Huffman 编码举例

令 $\mathcal{X} = \{1, 2, 3, 4\}$, $p_{X_1} = [1/2 \ 1/4 \ 1/8 \ 1/8]$, 转移概率矩阵为:

$$P = \begin{bmatrix} 1/2 & 1/4 & 1/8 & 1/8 \\ 1/4 & 1/8 & 1/8 & 1/2 \\ 1/8 & 1/8 & 1/2 & 1/4 \\ 1/8 & 1/2 & 1/4 & 1/8 \end{bmatrix}$$

$$z = [1/4 \ 1/4 \ 1/4 \ 1/4], \quad H(z) = 2 \text{ bits}$$

$$R = 1 \cdot 1/2 + 2 \cdot 1/4 + 2(3 \cdot 1/8) = 1.75 \text{ bits.}$$

Huffman 编码: $X_1 \rightarrow \{0 \ 10 \ 110 \ 111\}$. 给定条件 x_{n-1} Huffman 编码 x_n :

$$x_{n-1} = 1 \Rightarrow x_n \in \{0 \ 10 \ 110 \ 111\}$$

条件 Huffman 编码举例

令 $\mathcal{X} = \{1, 2, 3, 4\}$, $p_{X_1} = [1/2 \ 1/4 \ 1/8 \ 1/8]$, 转移概率矩阵为:

$$P = \begin{bmatrix} 1/2 & 1/4 & 1/8 & 1/8 \\ 1/4 & 1/8 & 1/8 & 1/2 \\ 1/8 & 1/8 & 1/2 & 1/4 \\ 1/8 & 1/2 & 1/4 & 1/8 \end{bmatrix}$$

$$z = [1/4 \ 1/4 \ 1/4 \ 1/4], \quad H(z) = 2 \text{ bits}$$

$$R = 1 \cdot 1/2 + 2 \cdot 1/4 + 2(3 \cdot 1/8) = 1.75 \text{ bits.}$$

Huffman 编码: $X_1 \rightarrow \{0 \ 10 \ 110 \ 111\}$. 给定条件 x_{n-1} Huffman 编码 x_n :

$$x_{n-1} = 1 \Rightarrow x_n \in \{0 \ 10 \ 110 \ 111\}$$

$$x_{n-1} = 2 \Rightarrow x_n \in \{10 \ 110 \ 111 \ 0\}$$

条件 Huffman 编码举例

令 $\mathcal{X} = \{1, 2, 3, 4\}$, $p_{X_1} = [1/2 \ 1/4 \ 1/8 \ 1/8]$, 转移概率矩阵为:

$$P = \begin{bmatrix} 1/2 & 1/4 & 1/8 & 1/8 \\ 1/4 & 1/8 & 1/8 & 1/2 \\ 1/8 & 1/8 & 1/2 & 1/4 \\ 1/8 & 1/2 & 1/4 & 1/8 \end{bmatrix}$$

$$z = [1/4 \ 1/4 \ 1/4 \ 1/4], \quad H(z) = 2 \text{ bits}$$

$$R = 1 \cdot 1/2 + 2 \cdot 1/4 + 2(3 \cdot 1/8) = 1.75 \text{ bits.}$$

Huffman 编码: $X_1 \rightarrow \{0 \ 10 \ 110 \ 111\}$ 。 给定条件 x_{n-1} Huffman 编码 x_n :

$$x_{n-1} = 1 \Rightarrow x_n \in \{0 \ 10 \ 110 \ 111\}$$

$$x_{n-1} = 2 \Rightarrow x_n \in \{10 \ 110 \ 111 \ 0\}$$

$$x_{n-1} = 3 \Rightarrow x_n \in \{110 \ 111 \ 0 \ 10\}$$

$$x_{n-1} = 4 \Rightarrow x_n \in \{111 \ 0 \ 10 \ 110\}$$

条件 Huffman 编码举例

令 $\mathcal{X} = \{1, 2, 3, 4\}$, $p_{X_1} = [1/2 \ 1/4 \ 1/8 \ 1/8]$, 转移概率矩阵为:

$$P = \begin{bmatrix} 1/2 & 1/4 & 1/8 & 1/8 \\ 1/4 & 1/8 & 1/8 & 1/2 \\ 1/8 & 1/8 & 1/2 & 1/4 \\ 1/8 & 1/2 & 1/4 & 1/8 \end{bmatrix}$$

$$z = [1/4 \ 1/4 \ 1/4 \ 1/4], \quad H(z) = 2 \text{ bits}$$

$$R = 1 \cdot 1/2 + 2 \cdot 1/4 + 2(3 \cdot 1/8) = 1.75 \text{ bits.}$$

Huffman 编码: $X_1 \rightarrow \{0 \ 10 \ 110 \ 111\}$ 。 给定条件 x_{n-1} Huffman 编码 x_n :

$$x_{n-1} = 1 \Rightarrow x_n \in \{0 \ 10 \ 110 \ 111\}$$

$$x_{n-1} = 2 \Rightarrow x_n \in \{10 \ 110 \ 111 \ 0\}$$

$$x_{n-1} = 3 \Rightarrow x_n \in \{110 \ 111 \ 0 \ 10\}$$

$$x_{n-1} = 4 \Rightarrow x_n \in \{111 \ 0 \ 10 \ 110\}$$

马尔可夫序列: 1 2 4 1 1 3 4 2...

条件 Huffman 编码举例

令 $\mathcal{X} = \{1, 2, 3, 4\}$, $p_{X_1} = [1/2 \ 1/4 \ 1/8 \ 1/8]$, 转移概率矩阵为:

$$P = \begin{bmatrix} 1/2 & 1/4 & 1/8 & 1/8 \\ 1/4 & 1/8 & 1/8 & 1/2 \\ 1/8 & 1/8 & 1/2 & 1/4 \\ 1/8 & 1/2 & 1/4 & 1/8 \end{bmatrix}$$

$$z = [1/4 \ 1/4 \ 1/4 \ 1/4], \quad H(z) = 2 \text{ bits}$$

$$R = 1 \cdot 1/2 + 2 \cdot 1/4 + 2(3 \cdot 1/8) = 1.75 \text{ bits.}$$

Huffman 编码: $X_1 \rightarrow \{0 \ 10 \ 110 \ 111\}$ 。 给定条件 x_{n-1} Huffman 编码 x_n :

$$x_{n-1} = 1 \Rightarrow x_n \in \{0 \ 10 \ 110 \ 111\}$$

$$x_{n-1} = 2 \Rightarrow x_n \in \{10 \ 110 \ 111 \ 0\}$$

$$x_{n-1} = 3 \Rightarrow x_n \in \{110 \ 111 \ 0 \ 10\}$$

$$x_{n-1} = 4 \Rightarrow x_n \in \{111 \ 0 \ 10 \ 110\}$$

马尔可夫序列: 1 2 4 1 1 3 4 2...

\Rightarrow Huffman: 00 01 11 00 00 10 11 01... L=16/8=2 bits

条件 Huffman 编码举例

令 $\mathcal{X} = \{1, 2, 3, 4\}$, $p_{X_1} = [1/2 \ 1/4 \ 1/8 \ 1/8]$, 转移概率矩阵为:

$$P = \begin{bmatrix} 1/2 & 1/4 & 1/8 & 1/8 \\ 1/4 & 1/8 & 1/8 & 1/2 \\ 1/8 & 1/8 & 1/2 & 1/4 \\ 1/8 & 1/2 & 1/4 & 1/8 \end{bmatrix}$$

$$z = [1/4 \ 1/4 \ 1/4 \ 1/4], \quad H(z) = 2 \text{ bits}$$

$$R = 1 \cdot 1/2 + 2 \cdot 1/4 + 2(3 \cdot 1/8) = 1.75 \text{ bits.}$$

Huffman 编码: $X_1 \rightarrow \{0 \ 10 \ 110 \ 111\}$ 。 给定条件 x_{n-1} Huffman 编码 x_n :

$$x_{n-1} = 1 \Rightarrow x_n \in \{0 \ 10 \ 110 \ 111\}$$

$$x_{n-1} = 2 \Rightarrow x_n \in \{10 \ 110 \ 111 \ 0\}$$

$$x_{n-1} = 3 \Rightarrow x_n \in \{110 \ 111 \ 0 \ 10\}$$

$$x_{n-1} = 4 \Rightarrow x_n \in \{111 \ 0 \ 10 \ 110\}$$

马尔可夫序列: 1 2 4 1 1 3 4 2...

\Rightarrow Huffman: 00 01 11 00 00 10 11 01... $L=16/8=2$ bits

\Rightarrow 条件 Huffman: 0 10 0 111 0 110 10 0... $L=14/8=1.75$ bits

条件 Huffman 编码举例

令 $\mathcal{X} = \{1, 2, 3, 4\}$, $p_{X_1} = [1/2 \ 1/4 \ 1/8 \ 1/8]$, 转移概率矩阵为:

$$P = \begin{bmatrix} 1/2 & 1/4 & 1/8 & 1/8 \\ 1/4 & 1/8 & 1/8 & 1/2 \\ 1/8 & 1/8 & 1/2 & 1/4 \\ 1/8 & 1/2 & 1/4 & 1/8 \end{bmatrix}$$

$$z = [1/4 \ 1/4 \ 1/4 \ 1/4], \quad H(z) = 2 \text{ bits}$$

$$R = 1 \cdot 1/2 + 2 \cdot 1/4 + 2(3 \cdot 1/8) = 1.75 \text{ bits.}$$

Huffman 编码: $X_1 \rightarrow \{0 \ 10 \ 110 \ 111\}$. 给定条件 x_{n-1} Huffman 编码 x_n :

$$x_{n-1} = 1 \Rightarrow x_n \in \{0 \ 10 \ 110 \ 111\}$$

$$x_{n-1} = 2 \Rightarrow x_n \in \{10 \ 110 \ 111 \ 0\}$$

$$x_{n-1} = 3 \Rightarrow x_n \in \{110 \ 111 \ 0 \ 10\}$$

$$x_{n-1} = 4 \Rightarrow x_n \in \{111 \ 0 \ 10 \ 110\}$$

马尔可夫序列: 1 2 4 1 1 3 4 2...

\Rightarrow Huffman: 00 01 11 00 00 10 11 01... L=16/8=2 bits

\Rightarrow 条件 Huffman: 0 10 0 111 0 110 10 0... L=14/8=1.75 bits

\Rightarrow 相似地, 给定条件 x_{n-1} 解码 x_n

马尔可夫链和熵率


- ① 马尔可夫链
- ② 熵率
- ③ 条件 Huffman 编码
- ④ 英文文本压缩

英文文本压缩

考虑英文文本压缩问题，简单起见，考虑以下包含 27 个符号的集合：

A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z, -

space

A diagram showing a box containing the 27 symbols: A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z, -. An arrow points from the word 'space' above to the hyphen '-' symbol inside the box.

忽略大小写 ($a \rightarrow A$) 和特殊符号 ($. , ? !$ 等等)

由于 $|\mathcal{X}| = 27$ ，因此可以压缩到码率 $R = \log 27 \approx 4.75$.

问题：对于英文文本，可以无损压缩的最低的码率是多少？

首先考虑：如果一段文本每个符号独立同分布会怎样？英文中最常见的字母是什么？最不常见的字母是什么？

假设独立同分布

最常见的字母是 A 和 E，最不常见的字母是 Q 和 Z。

Letter	Probability	Letter	Probability	Letter	Probability
A	8.29%	J	0.21%	S	6.33%
B	1.43	K	0.48	T	9.27
C	3.68	L	3.68	U	2.53
D	4.29	M	3.23	V	1.03
E	12.8	N	7.16	W	1.62
F	2.20	O	7.28	X	0.20
G	1.71	P	2.93	Y	1.57
H	4.54	Q	0.11	Z	0.09
I	7.16	R	6.90		

$$H(X) = - \sum_{i=1}^{27} p_i \log p_i \approx 4.17 \text{ bits} < 4.75 \text{ bits}$$

i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	-	0.1928



假设马尔可夫性

条件可以减小信息熵:

$$H(X_{10}) > H(X_{10}|X_9) > H(X_{10}|X_9X_8) > H(X_{10}|X_9X_8X_7) > \dots$$

我们知道，英语中常常出现固定的两个字母搭配:

- THIS THAT THE THEN THOUGH
- LOOK GOOD BOOK

英文中 Q 之后通常是什么字母?

- May I ask a **q**uestion?
- This is the library, please be **q**uiet?

一阶马尔可夫

考虑一段英文文本序列：

$X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9 \dots$

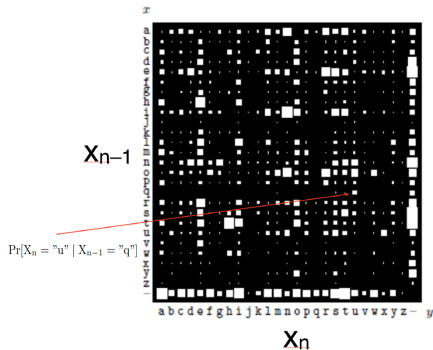
我们可以衡量条件概率

$$\Pr[X_n = a \mid X_{n-1} = b]$$

例如，字母 Q 之后通常是字母 U，

$$\Pr[X_n = \text{"u"} \mid X_{n-1} = \text{"q"}]$$

图中方块大小表示概率大小



学术词频查询——Google N-Gram Viewer



可以查找更高阶的短语吗？

对于单词而言，一般最高统计三阶

总结

- 马尔可夫链是一个非 IID 随机序列
- 马尔可夫信源可以被进一步压缩（码率更低）

作业

- 复习授课内容
- 预习信道容量
- 独立完成习题
 - 3.16