

目录

1. 概率回顾, Markov 链	5
1.1. 概率空间	5
1.2. 全概率公式	6
1.3. 独立性	7
1.4. Markov 链	8
1.5. 条件期望	10
1.5.1 关于随机事件的条件期望	10
1.5.2 关于随机变量的条件期望	11
1.5.3 条件期望的性质	11
2. 熵	13
2.1. 自信息 (Self-information)	13
2.2. 熵的定义	15
2.3. 联合熵与条件熵	16
2.4. 相对熵与互信息	18
2.5. 熵与互信息之间的关系	19
3. 熵的更多性质	20
3.1. 熵, 相对熵与互信息的链式法则	20
3.2. Jensen 不等式	21
3.3. Jensen 不等式的应用	22
3.4. 对数和不等式及其应用	25
4. 数据处理不等式与费诺不等式	28
4.1. 数据处理不等式	28
4.2. 充分统计量	29
4.3. 费诺不等式	30
5. 熵的公理化定义 *	34
6. 习题课 1	41
7. 渐近均分性	48
7.1. 大数定律	48
7.2. 渐近均分性定理	49
7.3. AEP 的推论: 数据压缩	50
7.4. 高概率集与典型集	51

8. 熵率	53
8.1. 马尔可夫链	53
8.2. 熵率	55
8.3. 例子：加权图上随机游动的熵率	57
8.4. 马尔可夫链的函数	58
9. 信源编码	61
9.1. 有关编码的几个例子	61
9.2. 唯一可译码	62
9.3. 即时码	64
9.4. Sardinas-Patterson 定理的证明 *	65
10. Kraft 不等式与最优码	68
10.1. 构造即时码	68
10.2. Kraft 不等式	69
10.3. 最优码	70
10.4. 最优码长的界	71
10.5. 惟一可译码的 Kraft 不等式	72
11. 赫夫曼码与最优码	74
11.1. r 元赫夫曼码	74
11.2. 赫夫曼码的相关讨论	74
11.3. 赫夫曼码的最优性	75
11.4. Shannon-Fano-Elias 编码	76
11.5. 香农码的竞争最优性	77
12. 习题课 2	79
13. 信道容量	80
13.1. 信道容量	80
13.2. 信道容量的几个例子	80
13.2.1 无噪声二元信道	80
13.2.2 二元对称信道	80
13.2.3 二元擦除信道	81
13.3. 对称信道	82
13.4. 并联信道	82
13.5. 信道容量的性质	83
13.6. 一些关键定义	83

14. 信道编码定理	86
14.1. 联合典型序列	86
14.2. 信道编码定理	88
15. 微分熵	91
15.1. 定义	91
15.2. 连续随机变量的 AEP	92
15.3. 微分熵与离散熵的关系	93
15.4. 联合微分熵与条件微分熵	94
15.5. 相对熵与互信息	95
15.6. 微分熵的性质, 相对熵和互信息的性质	96
16. 高斯信道	99
16.1. 高斯信道	99
16.1.1 定义	99
16.1.2 高斯信道编码定理	99
16.2. 几种常见的高斯信道	100
16.2.1 并联高斯信道	100
16.2.2 高斯彩色信道	101
17. 率失真理论	103
17.1. 率失真函数	103
17.2. 率失真函数的计算	104
17.2.1 二元信源	104
17.2.2 高斯信源	106
17.2.3 独立高斯随机变量的同步描述	107
18. 信息论与统计学习	109
18.1.	109
18.2. 监督学习	109
18.2.1 线性回归	111
18.2.2 逻辑斯蒂回归	112
18.3. 神经网络	113
18.3.1 单层网络	113
18.3.2 双层网络	114
18.3.3 深度网络	115
18.3.4 基于神经网络的回归与分类	115
18.3.5 神经网络训练	116

18.4. 自回归模型	116
18.4.1 训练自回归模型	116
18.4.2 关于自回归模型的注记	117
索引	118

第 1 讲 概率回顾, Markov 链

1.1 概率空间

定义 1.1

一个试验每个可能的结果称为 样本点, 所有样本点构成的集合 Ω 称为是这个试验的 **样本空间**. 样本空间 Ω 的子集 A 称为是一个 **事件**. 如果样本点 ω 包含在事件 A 中, 则记为 $\omega \in A$, 否则记为 $\omega \notin A$. 对于事件 A , 如果一次试验的结果 $\omega \in A$, 则称事件 A 发生, 否则称事件 A 不发生.

定义 1.2

Ω 的子集族 $\mathcal{F}$ 称为一个 σ -域, 如果其满足下列条件:

- (i) $\emptyset \in \mathcal{F}$;
- (ii) 如果 $A_1, A_2, \dots \in \mathcal{F}$ 则 $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$;
- (iii) 如果 $A \in \mathcal{F}$, 则 $A^c \in \mathcal{F}$.

定义 1.3

$(\Omega, \mathcal{F})$ 上的一个 **概率测度** P 是一个函数 $P: \mathcal{F} \rightarrow [0, 1]$, 满足:

- (i) $P(\Omega) = 1$;
- (ii) 如果 $A_1, A_2, \dots$ 是 $\mathcal{F}$ 中的一列不交的元素, 即对任意 $i \neq j$, $A_i \cap A_j = \emptyset$, 那么

$$P(\cup_{j=1}^{\infty} A_j) = \sum_{i=1}^{\infty} P(A_i).$$

由集合 Ω , Ω 的子集构成的 σ -域 $\mathcal{F}$, 以及 $(\Omega, \mathcal{F})$ 上的一个 概率测度 P 构成的一个三元组 $(\Omega, \mathcal{F}, P)$ 称为是一个 **概率空间**. σ -域 $\mathcal{F}$ 称为 **事件域**.

定义 1.4

可测空间 $(\Omega, \mathcal{F})$ 上的函数 $X: \Omega \rightarrow \mathbb{R}$ 称为是 (空间 $(\Omega, \mathcal{F})$ 上的) 一个 **随机变量**, 如果对任意 $x \in \mathbb{R}$, 有

$$\{\omega \in \Omega: X(\omega) \leq x\} \in \mathcal{F}.$$

这样的函数称为是 $\mathcal{F}$ -可测的.

- 设 X 是以 $\mathcal{X}$ 为字母表的一个随机变量.
- 我们用 $\{p_X(x), x \in \mathcal{X}\}$ 表示 X 的分布, 其中 $p_X(x) = \Pr\{X = x\}$.
- 当不会引起混淆时, 我们将 $p_X(x)$ 简记为 $p(x)$.

定义 1.5

设 X 是以有限或可数集 $\mathcal{X}$ 为字母表的一个随机变量, 分布为 $\{p(x), x \in \mathcal{X}\}$. X 的支集 (支撑集, 支撑) S_X 定义为 $\{x \in \mathcal{X} : p(x) > 0\}$.

如果 $S_X = \mathcal{X}$, 则称 X 是严格正的. 否则, 我们称 p 不是严格正的, 或者说 p 包含零概率质量. 上述所有符号都可以自然地推广到两个或多个随机变量的情况. 正如我们将会看到的, 包含零概率质量的概率分布非常微妙, 需要极其谨慎地处理.

1.2 全概率公式**定义 1.6**

设 A 和 B 是两个事件. 如果 $P(B) > 0$, 则在给定事件 B 发生的条件下事件 A 的条件概率为

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \quad (1.1)$$

我们称 $P(A)$ 为 A 的先验概率, 而 $P(A|B)$ 为 A 的后验概率.

样本空间 Ω 中的事件族 $A_1, A_2, \dots, A_n$ 称为是 Ω 的一个分割或分划, 如果

$$A_i \cap A_j = \emptyset \quad (i \neq j \text{ 时}), \quad \text{且} \quad \cup_{i=1}^n A_i = \Omega.$$

我们也称 $A_1, A_2, \dots, A_n$ 是一个完备事件组.

定理 1.7: 全概率公式

设 A 和 B 是样本空间 Ω 中的两个事件, 其中 $0 < P(A) < 1$, 则

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c).$$

一般地, 设事件 $A_1, A_2, \dots, A_n$ 为 Ω 的一个分划, 满足对任意 i 有 $P(A_i) > 0$. 则

$$P(B) = \sum_{i=1}^n P(B|A_i)P(A_i).$$

注记 1.8

实际上全概率公式的条件可以减弱为 $B \subset \cup_{i=1}^n A_i$. 实际上我们补充定义 $A_{n+1} = (\cup_{i=1}^n A_i)^c$, 注意到 $P(B|A_{n+1}) = 0$, 从而可得所需结论.

定理 1.9: Bayes 公式

设 A, B 是两个事件, 且 $P(A), P(B) > 0$, 则

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

利用 Bayes 公式和全概率公式, 我们有如下结论.

定理 1.10

设 $A_1, \dots, A_n$ 为样本空间 Ω 的一个分划, 满足对任意 i 均有 $\mathbb{P}(A_i) > 0$. 则对任意满足 $P(B) > 0$ 的事件 B , 有

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^n P(B|A_j)P(A_j)}.$$

1.3 独立性**定义 1.11**

我们称随机变量 X 和 Y 是独立的, 记做 $X \perp Y$, 如果对任意 $x \in \mathcal{X}, y \in \mathcal{Y}$, 有

$$p(x, y) = p(x)p(y)$$

定义 1.12: 相互独立

设 $n \geq 3$, 随机变量 $X_1, X_2, \dots, X_n$ 是 **相互独立的**, 如果对于任意 $x_1, x_2, \dots, x_n$, 有

$$p(x_1, x_2, \dots, x_n) = p(x_1)p(x_2) \cdots p(x_n).$$

定义 1.13: 两两独立

设 $n \geq 3$, 随机变量 $X_1, X_2, \dots, X_n$ 是 **两两独立的**, 如果对任意 $1 \leq i < j \leq n$, 有 X_i 和 X_j 相互独立.

相互独立可以推出两两独立, 但是反过来不成立.

定义 1.14: 条件独立

对于随机变量 X, Y 以及 Z , 称 X 和 Z 关于 Y **条件独立**, 记为 $X \perp Z | Y$, 如果对任意 x, y and z ,

$$p(x, y, z)p(y) = p(x, y)p(y, z).$$

或等价地,

$$p(x, y, z) = \begin{cases} \frac{p(x, y)p(y, z)}{p(y)} = p(x, y)p(z|y) & \text{如果 } p > 0, \\ 0 & \text{其他.} \end{cases} \quad (1.2)$$

命题 1.15

对于随机变量 X, Y , 和 Z , $X \perp Z | Y$ 当且仅当对任意 x, y 以及 z , 其中 $p(y) > 0$, 有

$$p(x, y, z) = a(x, y)b(y, z).$$

证明. “仅当”部分直接由 (1.2) 中关于条件独立的定义得出, 所以接下来我们证明“当”部分. 假设对于任意满足 $p(y) > 0$ 的 x, y, z 有

$$p(x, y, z) = a(x, y)b(y, z).$$

则对于上述 x, y, z , 我们有

$$p(x, y) = \sum_z p(x, y, z) = \sum_z a(x, y) b(y, z) = a(x, y) \sum_z b(y, z)$$

以及

$$p(y, z) = \sum_x p(x, y, z) = \sum_x a(x, y) b(y, z) = b(y, z) \sum_x a(x, y).$$

又

$$p(y) = \sum_z p(y, z) = \left(\sum_x a(x, y) \right) \left(\sum_z b(y, z) \right) > 0,$$

我们有

$$\frac{p(x, y)p(y, z)}{p(y)} = \frac{(a(x, y) \sum_z b(y, z)) (b(y, z) \sum_x a(x, y))}{\left(\sum_x a(x, y) \right) \left(\sum_z b(y, z) \right)} = a(x, y) b(y, z) = p(x, y, z).$$

对于满足 $p(y) = 0$ 的 x, y, z , 注意到

$$0 \leq p(x, y, z) \leq p(y) = 0,$$

我们可知

$$p(x, y, z) = 0.$$

综上由 (1.2) 我们有 $X \perp Z | Y$, 从而得证. □

1.4 Markov 链

定义 1.16: Markov 链

设 $X_1, X_2, \dots, X_n$ 是一列随机变量, 其中 $n \geq 3$, 称 $X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_n$ 构成一个马尔可夫链, 如果对任意 $x_1, x_2, \dots, x_n$, 有

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

或等价地,

$$p(x_1, x_2, \dots, x_n) = \begin{cases} p(x_1, x_2) p(x_3 | x_2) \cdots p(x_n | x_{n-1}) & \text{如果 } p(x_2), p(x_3), \dots, p(x_{n-1}) > 0 \\ 0 & \text{其他.} \end{cases}$$

注意到 $X \perp Z | Y$ 等价于我们有马尔可夫链 $X \rightarrow Y \rightarrow Z$.

命题 1.17

$X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_n$ 构成一个 Markov 链当且仅当 $X_n \rightarrow X_{n-1} \rightarrow \dots \rightarrow X_1$ 构成一个 Markov 链.

命题 1.18

$X_1 \rightarrow X_2 \rightarrow \cdots X_n$ 构成一个 Markov 链当且仅当

$$\begin{aligned} X_1 &\rightarrow X_2 \rightarrow X_3 \\ (X_1, X_2) &\rightarrow X_3 \rightarrow X_4 \\ &\vdots \\ (X_1, X_2, \cdots, X_{n-2}) &\rightarrow X_{n-1} \rightarrow X_n \end{aligned}$$

构成 Markov 链.

命题 1.19

$X_1 \rightarrow X_2 \rightarrow \cdots X_n$ 构成一个 Markov 链当且仅当任意满足 $p(x_2), p(x_3), \cdots, p(x_{n-1}) > 0$ 的 $x_1, x_2, \cdots, x_n$, 有

$$p(x_1, x_2, \cdots, x_n) = f_1(x_1, x_2)f_2(x_2, x_3) \cdots f_{n-1}(x_{n-1}, x_n).$$

命题 1.20: Markov 子链

设 $\mathcal{N}_n = \{1, 2, \cdots, n\}$, $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$ 是一个 Markov 链. 对任意 $\alpha \subseteq \mathcal{N}_n$, 我们将 $(X_i, i \in \alpha)$ 简记为 X_α . 对于任意两两不交的 $\alpha_1, \alpha_2, \cdots, \alpha_m \subseteq \mathcal{N}_n$, 使得对任意 $k_j \in \alpha_j$, $j = 1, \cdots, m$, 有

$$k_1 < k_2 < \cdots < k_m$$

成立. 那么我们有

$$X_{\alpha_1} \rightarrow X_{\alpha_2} \rightarrow \cdots \rightarrow X_{\alpha_m}$$

构成一个 Markov 链. 也就是说, $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$ 的一个子链也是一个 Markov 链.

命题 1.21

设 X_1, X_2, X_3 以及 X_4 是随机变量, 满足 $p(x_1, x_2, x_3, x_4)$ 是严格正的. 那么我们有

$$\left. \begin{aligned} X_1 \perp X_4 | (X_2, X_3) \\ X_1 \perp X_3 | (X_2, X_4) \end{aligned} \right\} \Rightarrow X_1 \perp (X_3, X_4) | X_2.$$

证明. 由 $X_1 \perp X_4 | (X_2, X_3)$, 我们有

$$p(x_1, x_2, x_3, x_4) = \frac{p(x_1, x_2, x_3)p(x_2, x_3, x_4)}{p(x_2, x_3)}. \quad (1.3)$$

另一方面, 由 $X_1 \perp X_3 | (X_2, X_4)$, 我们可以得到

$$p(x_1, x_2, x_3, x_4) = \frac{p(x_1, x_2, x_4)p(x_2, x_3, x_4)}{p(x_2, x_4)}. \quad (1.4)$$

对比上面两式, 我们有

$$p(x_1, x_2, x_3) = \frac{p(x_2, x_3)p(x_2, x_3, x_4)}{p(x_2, x_4)}.$$

上式两边对 x_3 求和, 我们有

$$p(x_1, x_2) = \frac{p(x_2)p(x_1, x_2, x_4)}{p(x_2, x_4)},$$

也就是说,

$$\frac{p(x_1, x_2, x_4)}{p(x_2, x_4)} = \frac{p(x_1, x_2)}{p(x_2)}.$$

于是对任意 x_1, x_2, x_3, x_4 , 有

$$\begin{aligned} p(x_1, x_2, x_3, x_4) &= \frac{p(x_1, x_2, x_4)p(x_2, x_3, x_4)}{p(x_2, x_4)} \\ &= \frac{p(x_1, x_2)p(x_2, x_3, x_4)}{p(x_2)}, \end{aligned}$$

也即 $X_1 \perp (X_3, X_4) | X_2$. □

思考

在上述定理中, 如果存在 x_1, x_2, x_3, x_4 使得 $p(x_1, x_2, x_3, x_4) = 0$, 结论是否依旧成立?

例 1.22

设 Y, Z 是两个独立的随机变量. 记 $X_1 = Y, X_2 = Z, X_3 = X_4 = (Y, Z)$. 则 $X_1 \perp X_4 | (X_2, X_3)$ 且 $X_1 \perp X_3 | (X_2, X_4)$, 但 $X_1 \not\perp (X_3, X_4) | X_2$.

1.5 条件期望

1.5.1 关于随机事件的条件期望

定义 1.23: 关于随机事件的条件期望

设 A 是一个概率为正的随机事件. 如果 Y 是一个离散型随机变量, 那么 Y 在给定事件 A 时的条件期望定义为

$$E[Y|A] = \sum yP(Y = y|A),$$

其中求和是在 Y 的支撑集上进行. 如果 Y 是一个连续型随机变量, 具有概率密度函数 f , 则

$$E[Y|A] = \int yf(y|A)dy,$$

其中条件概率密度函数 $f(y|A)$ 定义为条件分布函数 $F(y|A) = P(Y \leq y|A)$ 的导数, 由 Bayes 公式我们有

$$f(y|A) = \frac{P(A|Y = y)f(y)}{P(A)}.$$

定理 1.24

设 $A_1, \dots, A_n$ 是样本空间 Ω 的一个划分, 且对任意 i 有 $P(A_i) > 0$. 设 Y 是同一样本空间的一个随机变量. 则

$$E[Y] = \sum_{i=1}^n E[Y|A_i]P(A_i).$$

设 B 是样本空间 Ω 中的一个事件, 取 $Y = I_B$ 应用上面定理我们可以得到全概率公式.

1.5.2 关于随机变量的条件期望

设 X 是一个随机变量, Y 是一个离散型随机变量, 我们定义

$$E[Y|X = x] = \sum_y yP(Y = y|X = x).$$

类似地, 如果 Y 是连续的, 我们定义

$$E[Y|X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy.$$

定义 1.25

设 $g(x) = E[Y|X = x]$. 则给定 X 时 Y 的条件期望, 记为 $E[Y|X]$, 定义为随机变量 $g(X)$.

1.5.3 条件期望的性质

设 X 和 Y 是两个随机变量.

命题 1.26

如果 X 和 Y 是独立的, 那么

$$E[Y|X] = E[Y].$$

命题 1.27

设 h 是一个 Borel 可测函数, 那么

$$E[h(X)Y|X] = h(X)E[Y|X].$$

例 1.28

设 $Z \sim \mathcal{N}(0, 1)$, $Y = Z^2$. 求 $E[Y|Z]$.

解. 注意到由 $Y = Z^2$, 我们知

$$g(z) = E[Y|Z = z] = z^2.$$

于是 $E[Y|Z] = g(Z) = Z^2$. □

命题 1.29

设 c 是一个常数, 我们有

$$E[Y_1 + Y_2|X] = E[Y_1|X] + E[Y_2|X], \quad E[cY|X] = cE[Y|X].$$

例 1.30

设 $X_1, \dots, X_n$ 是独立同分布的, $S_n = X_1 + X_2 + \dots + X_n$. 求 $E[X_1|S_n]$.

解. 由对称性,

$$E[X_1|S_n] = E[X_2|S_n] = \dots = E[X_n|S_n],$$

由线性性, 我们有

$$E[X_1|S_n] + \cdots + E[X_n|S_n] = E[S_n|S_n] = S_n.$$

于是

$$E[X_1|S_n] = S_n/n = \bar{X}_n.$$

□

命题 1.31

$$E[E[Y|X]] = E[Y].$$

定理 1.32

设 X, Y 是同一样本空间上的两个随机变量, $E[Y^2] < \infty$. 设 h 是一个 Borel 可测函数, 满足 $E[h^2(X)] < \infty$. 那么随机变量 $Y - E[Y|X]$ 与 $h(X)$ 是不相关的. 也就是说,

$$E[(Y - E[Y|X])h(X)] = 0.$$

证明. 由 Cauchy-Schwarz 不等式,

$$E[Yh(X)] \leq \sqrt{E[Y^2]E[h^2(X)]} < \infty.$$

于是我们有

$$\begin{aligned} E[(Y - E[Y|X])h(X)] &= E[h(X)Y] - E[h(X)E[Y|X]] \\ &= E[h(X)Y] - E[E[h(X)Y|X]] = 0, \end{aligned}$$

从而得证.

□

第2讲 熵

2.1 自信息 (Self-information)

令 E 为给定事件空间中的一个事件, 其概率为 $p_E := \Pr(E)$, 其中 $0 \leq p_E \leq 1$. 令 $\mathcal{I}(E)$ ——称为事件 E 的自信息——表示当得知事件 E 发生时所获得的信息量 (或者等价地, 表示在得知 E 发生之前, 人们对 E 所抱有的不确定性量)。一个很自然会问的问题是: “ $\mathcal{I}(E)$ 应该具备哪些性质?” 尽管这个问题的答案可能因人而异, 但以下是人们普遍合理预期 $\mathcal{I}(E)$ 应当具备的几条性质.

1. $\mathcal{I}(E)$ 应当是 p_E 的递减函数.

换句话说, 这条性质表明 $\mathcal{I}(E) = I(p_E)$, 其中 $I(\cdot)$ 是定义在 $[0, 1]$ 上的实值函数. 此外, 人们自然会预期一个越不可能发生的事件 E 发生时, 获得的信息量就越多. 换句话说, $I(p_E)$ 是 p_E 的递减函数.

2. $I(p_E)$ 应当关于 p_E 连续.

直观上, 人们预期 p_E 的微小变化应该对应于事件 E 所携带信息量的微小变化.

3. 如果 E_1 和 E_2 是独立事件, 那么 $\mathcal{I}(E_1 \cap E_2) = \mathcal{I}(E_1) + \mathcal{I}(E_2)$, 或者等价地, $I(p_{E_1} \times p_{E_2}) = I(p_{E_1}) + I(p_{E_2})$.

这条性质表明, 当事件 E_1 和 E_2 相互独立时 (即它们在概率上互不影响), 这两个事件共同发生所获得的信息量, 应当等于每个单独事件发生所获得的信息量之和.

接下来, 我们将说明唯一满足上述性质 1-3 的函数是对数函数.

定理 2.1

存在唯一一个定义在 $[0, 1]$ 上的函数 $I(p)$, 满足以下条件:

1. $I(p)$ 关于 p 单调递减;
2. 在 $0 \leq p \leq 1$ 范围内, $I(p)$ 是关于 p 的连续函数;
3. $I(p_1 \times p_2) = I(p_1) + I(p_2)$. 该函数为 $I(p) = -c \cdot \log_b(p)$, 其中 c 是一个正常数, 对数的底数 b 是任何大于 1 的数.

证明. (1) 我们断言: 对于 $n = 1, 2, 3, \dots$,

$$I\left(\frac{1}{n}\right) = -c \cdot \log_b\left(\frac{1}{n}\right),$$

其中 $c > 0$ 是一个常数.

断言的证明: 首先注意到对于 $n = 1$, 由条件 3 可知该断言成立. 这是因为它可以推出 $I(1) = I(1) + I(1)$, 因此 $I(1) = 0 = -c \log_b(1)$.

现在令 n 为一个大于 1 的固定正整数. 条件 1 和条件 3 分别意味着

$$n < m \implies I\left(\frac{1}{n}\right) < I\left(\frac{1}{m}\right) \quad (2.1)$$

以及

$$I\left(\frac{1}{mn}\right) = I\left(\frac{1}{m}\right) + I\left(\frac{1}{n}\right) \quad (2.2)$$

其中 $n, m = 1, 2, 3, \dots$ 利用 (2.2), 我们可以通过 (对 k 的) 数学归纳法证明: 对所有非负整数 k ,

$$I\left(\frac{1}{n^k}\right) = k \cdot I\left(\frac{1}{n}\right) \quad (2.3)$$

现在, 对于任意正整数 r , 都存在一个非负整数 k 使得:

$$n^k \leq 2^r < n^{k+1}.$$

由 (2.1) 可得:

$$I\left(\frac{1}{n^k}\right) \leq I\left(\frac{1}{2^r}\right) < I\left(\frac{1}{n^{k+1}}\right),$$

将其与 (2.3) 结合, 可得:

$$k \cdot I\left(\frac{1}{n}\right) \leq r \cdot I\left(\frac{1}{2}\right) < (k+1) \cdot I\left(\frac{1}{n}\right).$$

因此, 由于 $I(1/n) > I(1) = 0$, 我们有:

$$\frac{k}{r} \leq \frac{I(1/2)}{I(1/n)} \leq \frac{k+1}{r}.$$

另一方面, 根据对数函数的单调性, 我们可得:

$$\log_b n^k \leq \log_b 2^r \leq \log_b n^{k+1} \iff \frac{k}{r} \leq \frac{\log_b(2)}{\log_b(n)} \leq \frac{k+1}{r}.$$

因此,

$$\left| \frac{\log_b(2)}{\log_b(n)} - \frac{I(1/2)}{I(1/n)} \right| < \frac{1}{r}.$$

由于 n 是固定的, 且 r 可以取任意大, 我们可以令 $r \rightarrow \infty$, 从而得到:

$$I\left(\frac{1}{n}\right) = c \cdot \log_b(n),$$

其中 $c = I(1/2)/\log_b(2) > 0$. 至此完成了该断言的证明.

(2) 我们断言: 对于正有理数 p , $I(p) = -c \cdot \log_b(p)$, 其中 $c > 0$ 是一个常数.

断言的证明: 一个正有理数 p 可以表示为两个整数的比值, 即 $p = r/s$, 其中 r 和 s 均为正整数. 那么由条件 3 可得:

$$I\left(\frac{1}{s}\right) = I\left(\frac{r}{s \cdot r}\right) = I\left(\frac{r}{s}\right) + I\left(\frac{1}{r}\right),$$

于是

$$I(p) = I\left(\frac{r}{s}\right) = I\left(\frac{1}{s}\right) - I\left(\frac{1}{r}\right) = c \cdot \log_b s - c \cdot \log_b r = -c \cdot \log_b p.$$

(3) 对于任意 $p \in [0, 1]$, 根据有理数在实数集中的连续性与稠密性, 可得:

$$I(p) = \lim_{a \uparrow p, a \text{ 为有理数}} I(a) = \lim_{b \downarrow p, b \text{ 为有理数}} I(b) = -c \cdot \log_b(p). \quad \square$$

上述常数 c 按照惯例通常归一化为 $c = 1$. 此外, 对数的底数 b 决定了衡量信息量时所使用的单位类型. 当底数 $b = 2$ 时, 信息量以比特 (bits, 即二进制位) 为单位表示. 当底数 $b = e$ 时——即使用自然对数 (ln)——信息量以纳特 (nats, 即自然单位或自然位) 为单位测量. 例如, 如果事件 E 代表抛掷一枚均匀硬币结果为“正面朝上”, 那么它的自信息量 (self-information) 为: $\mathcal{I}(E) = -\log_2(1/2) = 1 \text{ bit}$ 或者 $-\ln(1/2) = 0.693 \text{ nats}$.

2.2 熵的定义

熵是随机变量不确定度的度量. 设 X 是一个离散型的随机变量, 其字母表 (即概率论中的取值空间) 为 $\mathcal{X}$, 概率密度函数 $p(x) = \Pr(X = x)$, $x \in \mathcal{X}$. 为方便起见, 记概率密度函数为 $p(x)$ 以代替 $p_X(x)$. 由此, $p(x)$ 和 $p(y)$ 指两个不同的随机变量, 实际上分别表示两个不同的概率密度函数 $p_X(x)$ 和 $p_Y(y)$.

熵 (entropy) 一词源自于希腊语中的 entropia, 其意思为转折点, 变化. 熵的概念是由德国物理学家克劳修斯于 1865 年所提出. 物理学家玻尔兹曼于 1870 年将熵的概念引入统计力学. 1923 年, 德国科学家普朗克来中国讲学. 讲学时用到 “entropy” 一词, 胡刚复先生在翻译时灵机一动, 把 “商” 字加火旁来意译此词, 创造了 “熵” 字. 胡刚复先生说, 此名词的物理含意深奥须多字翻译, 又不合汉字规范, 于是他建议译成一个汉字 “熵”. 火代表与热学有关, 含义为温度 T 与热量 Q 之商, 熵读如商. 在 1948 年, 克劳德·艾尔伍德·香农将热力学的熵, 引入到信息论, 因此它又被称为香农熵 (Shannon entropy).

定义 2.2

一个离散型随机变量 X 的熵 $H(X)$ 定义为

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log p(x).$$

有时也将上面的量记为 $H(p)$. 其中对数的底是 2, 熵的单位用比特 (bit) 表示. 由于当 $x \rightarrow 0$ 时, $x \log x \rightarrow 0$, 今后我们约定 $0 \log 0 = 0$.

如果使用底为 b 的对数, 则相应的熵记为 $H_b(X)$. 当对数底为 e 时, 熵的单位用奈特 (nat) 表示. m 如无特别声明, 一般选取对数底为 2. 注意, 熵实际上是随机变量 X 的分布的泛函, 并不依赖于 X 的实际取值, 而仅依赖于其概率分布. 用 E 表示数学期望. 如果 $X \sim p(x)$, 则随机变量 $g(X)$ 的期望值可以表示为

$$\mathbb{E}_p g(X) = \sum_{x \in \mathcal{X}} g(x) p(x).$$

特别地, 当 $g(X) = \log \frac{1}{p(X)}$ 时, $\mathbb{E}_p g(X)$ 就是 X 的熵.

引理 2.3

$$H(X) \geq 0.$$

证明. 由 $0 \leq p(x) \leq 1$ 知 $\log \left(\frac{1}{p(x)} \right) \geq 0$. □

引理 2.4

$$H_b(X) = (\log_b a) H_a(X).$$

证明. 由 $\log_b p = (\log_b a) \log_a p$ 即可得到. □

例 2.5

设

$$X = \begin{cases} 1 & \text{概率为 } p \\ 0 & \text{概率为 } 1 - p. \end{cases}$$

于是

$$H(X) = -p \log p - (1 - p) \log(1 - p) =: H(p).$$

例 2.6

设

$$X = \begin{cases} a & \text{概率为 } \frac{1}{2} \\ b & \text{概率为 } \frac{1}{4} \\ c & \text{概率为 } \frac{1}{8} \\ d & \text{概率为 } \frac{1}{8} \end{cases}$$

则 X 的熵为

$$H(X) = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{4} \log \frac{1}{4} - \frac{1}{8} \log \frac{1}{8} - \frac{1}{8} \log \frac{1}{8} = \frac{7}{4} \text{ 比特.}$$

下面例子说明随机变量的熵可以是无穷的.

例 2.7

设 $A = \sum_{n=2}^{\infty} (n \log^2 n)^{-1}$. 定义随机变量 X 满足 $\Pr(X = n) = (An \log^2 n)^{-1}$, 其中 $n = 2, 3, \dots$, 则 $H(X) = +\infty$.

证明. 由定义, $p_n = \Pr(X = n) = 1/An \log^2 n$, $n \geq 2$. 于是

$$\begin{aligned} H(X) &= - \sum_{n=2}^{\infty} p(n) \log p(n) \\ &= - \sum_{n=2}^{\infty} (1/An \log^2 n) \log(1/An \log^2 n) \\ &= \sum_{n=2}^{\infty} \frac{\log(An \log^2 n)}{An \log^2 n} \\ &= \sum_{n=2}^{\infty} \frac{\log A + \log n + 2 \log \log n}{An \log^2 n} \\ &= \log A + \sum_{n=2}^{\infty} \frac{1}{An \log n} + \sum_{n=2}^{\infty} \frac{2 \log \log n}{An \log^2 n}. \end{aligned}$$

和式中第一项是有限的. 由于我们考虑对数以 2 为底, 最后一项中每一项均是非负的. 对于中间项, 我们有

$$\sum_{n=2}^{+\infty} \frac{1}{An \log n} > \int_2^{\infty} \frac{1}{Ax \log x} dx = K \ln \ln 2|_2^{\infty} = +\infty,$$

其中 $K = \frac{\ln 2}{A}$. 从而我们有 $H(X) = +\infty$. □

2.3 联合熵与条件熵**定义 2.8**

对于服从联合分布为 $p(x, y)$ 的一对离散随机变量 (X, Y) , 其**联合熵** $H(X, Y)$ (joint entropy) 定义为

$$H(X, Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y)$$

上式亦可表示为

$$H(X, Y) = -E \log p(X, Y).$$

一般地, 对于 n 个随机变量 $X_1, \dots, X_n$, 它们的联合熵定义为

$$H(X^n) = \sum_{x_1} \cdots \sum_{x_n} p(x_1, \dots, x_n) \log \frac{1}{p(x_1, \dots, x_n)}$$

定义 2.9

若 $(X, Y) \sim p(x, y)$, 条件熵 $H(Y|X)$ 定义为:

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \\ &= - \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \\ &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\ &= -E \log p(Y|X). \end{aligned}$$

定理 2.10: 链式法则

$$H(X, Y) = H(X) + H(Y|X).$$

证明. 我们有

$$\begin{aligned} H(X, Y) &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \\ &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log(p(x)p(y|x)) \\ &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\ &= - \sum_{x \in \mathcal{X}} p(x) \log p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\ &= H(X) + H(Y|X). \end{aligned}$$

从而得证. □

注记 2.11

注意 $H(Y|X) \neq H(X|Y)$, 但 $H(X) - H(X|Y) = H(Y) - H(Y|X)$, 稍后会用到这个性质.

例 2.12

设 (X, Y) 服从如下的联合分布:

Y \ X	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

X 的边缘分布为 $(1/2, 1/4, 1/8, 1/8)$, Y 的边缘分布为 $(1/4, 1/4, 1/4, 1/4)$, 因而 $H(X) = \frac{7}{4}$ 比特, $H(Y) = 2$ 比特. 同样地我们可以求得 $H(X|Y) = \sum_{i=1}^4 p(Y=i)H(X|Y=i) = \frac{11}{8}$ 比特, $H(Y|X) = \frac{13}{8}$ 比特, 以及 $H(X, Y) = \frac{27}{8}$ 比特.

2.4 相对熵与互信息

相对熵是两个随机分布之间距离的度量. 在统计学中, 它对应的是似然比的对数期望. 相对熵 $D(p||q)$ 度量当真实分布为 p 而假定分布为 q 时的无效性. 例如, 已知随机变量的真实分布为 p , 可以构造平均长度为 $H(p)$ 的码. 但是, 如果使用针对分布 q 的编码, 那么在平均意义上就需要 $H(p) + D(p||q)$ 比特来描述这个随机变量.

定义 2.13

两个概率质量函数 $p(x)$ 和 $q(x)$ 之间的**相对熵**或 **Kullback-Leibler 距离**定义为

$$D(p || q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(X)}{q(X)}. \quad (2.4)$$

在上述定义中, 我们采用约定 $0 \log \frac{0}{0} = 0$, 约定 $0 \log \frac{0}{q} = 0$, $p \log \frac{p}{0} = \infty$ (基于连续性). 因此, 若存在字符 $x \in \mathcal{X}$ 使得 $p(x) > 0$, $q(x) = 0$, 则有 $D(p||q) = \infty$.

稍后我们将证明相对熵总是非负的, 并且当且仅当 $p = q$ 时其为零. 但是, 由于相对熵并不对称, 也不满足三角不等式, 因此相对熵实际上并非两个分布之间的真正距离. 然而, 将相对熵视作分布之间的“距离”往往会很有用.

定义 2.14

考虑两个随机变量 X 和 Y , 它们的联合概率密度函数为 $p(x, y)$, 其边缘概率密度函数分别是 $p(x)$ 和 $q(x)$. 互信息 $I(X; Y)$ 为联合分布 $p(x, y)$ 和乘积分布 $p(x)p(y)$ 之间的相对熵, 即:

$$\begin{aligned} I(X; Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= D(p(x, y) || p(x)p(y)) \\ &= E_{p(x, y)} \log \frac{p(X, Y)}{p(X)p(Y)}. \end{aligned} \quad (2.5)$$

例 2.15

设 $\mathcal{X} = \{0, 1\}$, 考虑 $\mathcal{X}$ 上的两个分布 p 和 q . 设 $p(0) = 1 - r$, $p(1) = r$ 及 $q(0) = 1 - s$, $q(1) = s$, 则有

$$D(p\|q) = (1 - r) \log \frac{1 - r}{1 - s} + r \log \frac{r}{s}$$

以及

$$D(q\|p) = (1 - s) \log \frac{1 - s}{1 - r} + s \log \frac{s}{r}.$$

如果 $r = s$, 则 $D(p\|q) = D(q\|p) = 0$. 一般地, $D(p\|q) \neq D(q\|p)$. 例如, 若 $r = \frac{1}{2}$, $s = \frac{1}{4}$, 我们有

$$D(p\|q) = 0.2075 \text{ 比特}, D(q\|p) = 0.1887 \text{ 比特}.$$

2.5 熵与互信息之间的关系

可将互信息 $I(X; Y)$ 重新写为:

$$\begin{aligned} I(X; Y) &= \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x, y) \log \frac{p(x|y)}{p(x)} \\ &= - \sum_{x,y} p(x, y) \log p(x) + \sum_{x,y} p(x, y) \log p(x|y) \\ &= - \sum_x p(x) \log p(x) - \left(- \sum_{x,y} p(x, y) \log p(x|y) \right) \\ &= H(X) - H(X|Y) \end{aligned}$$

由此知互信息 $I(X; Y)$ 是在给定 Y 知识的条件下 X 的不确定度的缩减量.

对称地, 我们亦可以得到

$$I(X; Y) = H(Y) - H(Y|X),$$

因而, X 含有 Y 的信息量等同于 Y 含有 X 的信息量. 又由定理 3.1, 可得

$$I(X; Y) = H(X) + H(Y) - H(X, Y).$$

最后, 注意到

$$I(X; X) = H(X) - H(X|X) = H(X).$$

因此, 随机变量与自身的互信息为该随机变量的熵.

我们将前面关于熵和互信息的关系做一个总结如下.

定理 2.16

$$I(X; Y) = H(X) - H(X|Y) \quad (2.6)$$

$$I(X; Y) = H(Y) - H(Y|X) \quad (2.7)$$

$$I(X; Y) = H(X) + H(Y) - H(X, Y) \quad (2.8)$$

$$I(X; Y) = I(Y; X) \quad (2.9)$$

$$I(X; X) = H(X) \quad (2.10)$$

第3讲 熵的更多性质

3.1 熵, 相对熵与互信息的链式法则

定理 3.1: 链式法则

$$H(X, Y) = H(X) + H(Y|X).$$

定理 3.2: 熵的链式法则

设随机变量 $X_1, X_2, \dots, X_n$ 服从 $p(x_1, x_2, \dots, x_n)$, 则

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

证明. 反复应用定理3.1. □

定义 3.3

随机变量 X 和 Y 在给定随机变量 Z 时的条件互信息定义为

$$\begin{aligned} I(X; Y|Z) &= H(X|Z) - H(X|Y, Z) = E_{p(x,y,z)} \log \frac{p(X, Y|Z)}{p(X|Z)p(Y|Z)} \\ &= \sum_{x,y,z} p(x, y, z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)}. \end{aligned}$$

注意到 $I(X; Y|Z)$ 关于 X 和 Y 对称.

定理 3.4: 互信息的链式法则

$$I(X_1, X_2, \dots, X_n; Y) = \sum_{i=1}^n I(X_i; Y | X_{i-1}, X_{i-2}, \dots, X_1).$$

证明.

$$\begin{aligned} &I(X_1, X_2, \dots, X_n; Y) \\ &= H(X_1, X_2, \dots, X_n) - H(X_1, X_2, \dots, X_n | Y) \\ &= \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1) - \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1, Y) \\ &= \sum_{i=1}^n I(X_i; Y | X_{i-1}, X_{i-2}, \dots, X_1). \quad \square \end{aligned}$$

定义 3.5

对于联合概率密度函数 $p(x, y)$ 和 $q(x, y)$, 条件相对熵 $D(p(y|x) || q(y|x))$ 定义为条件概率密度函数

$p(y|x)$ 和 $q(y|x)$ 之间的平均相对熵, 其中取平均是关于概率密度函数 $p(x)$ 而言的. 更确切地,

$$\begin{aligned} D(p(y|x)||q(y|x)) &= \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} \\ &= E_{p(x,y)} \log \frac{p(Y|X)}{q(Y|X)}. \end{aligned}$$

定理 3.6: 相对熵的链式法则

$$D(p(x, y)||q(x, y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x)).$$

证明.

$$\begin{aligned} D(p(x, y)||q(x, y)) &= \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{q(x, y)} \\ &= \sum_x \sum_y p(x, y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} \\ &= \sum_x \sum_y p(x, y) \log \frac{p(x)}{q(x)} + \sum_x \sum_y p(x, y) \log \frac{p(y|x)}{q(y|x)} \\ &= D(p(x)||q(x)) + D(p(y|x)||q(y|x)). \quad \square \end{aligned}$$

3.2 Jensen 不等式

定义 3.7

若对于任意的 $x_1, x_2 \in (a, b)$ 及 $0 \leq \lambda \leq 1$, 满足

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2).$$

则称函数 $f(x)$ 在区间 (a, b) 上是凸的. 如果仅当 $\lambda = 0$ 或 $\lambda = 1$, 上式等号成立, 则称函数 f 是严格凸的.

定义 3.8

如果 $-f$ 为凸函数, 则称函数 f 是凹的.

定理 3.9

如果函数 f 在某个区间上存在非负 (正) 的二阶导数, 则 f 为该区间的凸函数 (严格凸函数).

定理 3.10

若给定凸函数 f 和一个随机变量 X , 则

$$Ef(X) \geq f(EX). \quad (3.1)$$

进一步, 若 f 是严格凸的, 那么式 (3.2) 中的等式蕴含 $X = EX$ 的概率为 1 (即 X 是一个常量).

3.3 Jensen 不等式的应用

定理 3.11

若给定凸函数 f 和一个随机变量 X , 则

$$Ef(X) \geq f(EX). \quad (3.2)$$

进一步, 若 f 是严格凸的, 那么式 (3.2) 中的等式蕴含 $X = EX$ 的概率为 1 (即 X 是一个常量).

定理 3.12: 信息不等式

设 $p(x), q(x) (x \in \mathcal{X})$ 为两个概率密度函数, 则

$$D(p||q) \geq 0.$$

当且仅当对任意的 x , $p(x) = q(x)$, 等号成立.

证明. 设 $A = \{x : p(x) > 0\}$ 为 $p(x)$ 的支撑集, 则

$$\begin{aligned} D(p||q) &= \sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \\ &= - \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \\ &\geq - \log \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \quad (\text{Jensen 不等式, 利用到 } f(t) = -\log t \text{ 是凸函数}) \\ &= - \log \sum_{x \in A} q(x) \\ &\geq - \log \sum_{x \in \mathcal{X}} q(x) \\ &= - \log 1 = 0. \end{aligned}$$

讨论上面式中等号成立的条件, 我们知等号成立当且仅当对任意的 x , 有 $p(x) = q(x)$, $D(p||q) = 0$. $\square$

注记 3.13

我们可以利用下面的不等式给出上面结果的另一个证明: 对任意 $a > 0$, $\ln a \geq 1 - \frac{1}{a}$ (这个不等式也被称为**基本不等式**), 等号成立当且仅当 $a = 1$. 这个不等式可以通过微积分的技巧证明. 我们记 B 为 $p(x)$ 的支撑集, 于是

$$\begin{aligned} D(p||q) &= (\log e) \sum_{x \in B} p(x) \ln \frac{p(x)}{q(x)} \\ &\geq (\log e) \sum_{x \in B} p(x) \left(1 - \frac{q(x)}{p(x)}\right) \\ &= (\log e) \left[\sum_{x \in B} p(x) - \sum_{x \in B} q(x) \right] \\ &\geq 0, \end{aligned}$$

讨论上面式中等号成立的条件, 我们知等号成立当且仅当对任意的 x , 有 $p(x) = q(x)$, $D(p||q) = 0$.

推论 3.14: 互信息的非负性

对任意两个随机变量 X 和 Y ,

$$I(X; Y) \geq 0.$$

当且仅当 X 与 Y 相互独立, 等号成立.

证明. $I(X; Y) = D(p(x, y) \| p(x)p(y)) \geq 0$, 当且仅当 $p(x, y) = p(x)p(y)$ (即 X 与 Y 为相互独立), 等号成立. $\square$

推论 3.15

$$D(p(y|x) \| q(y|x)) \geq 0.$$

当且仅当对任意的 y 以及满足 $p(x) > 0$ 的 x , 有 $p(y|x) = q(y|x)$, 等号成立.

推论 3.16

$$I(X; Y|Z) \geq 0.$$

当且仅当对给定随机变量 Z , X 和 Y 是条件独立的, 等号成立.

定理 3.17

$H(X) \leq \log |\mathcal{X}|$, 其中 $\mathcal{X}$ 表示 X 的字母表 $\mathcal{X}$ 中的元素个数, 当且仅当 X 服从 $\mathcal{X}$ 上均匀分布, 等号成立.

证明. 设 $u(x) = \frac{1}{|\mathcal{X}|}$ 为 $\mathcal{X}$ 上均匀分布的概率密度函数, $p(x)$ 是随机变量 X 的概率密度函数. 于是

$$D(p \| u) = \sum p(x) \log \frac{p(x)}{u(x)} = \log |\mathcal{X}| - H(X).$$

因而由相对熵的非负性,

$$0 \leq D(p \| u) = \log |\mathcal{X}| - H(X).$$

$\square$

推论 3.18

对于任意非负实数, 存在一个随机变量, 使得其熵恰好为该数.

证明. 考虑定义在有限字母表 $\mathcal{X}$ 上的随机变量 X . 由上面的定理我们知当 X 在 $\mathcal{X}$ 上一致分布的时候, $H(X) = \log |\mathcal{X}|$ 可以达到. 当 X 为一个确定值的时候 $H(X) = 0$ 可以达到. 对于 $0 \leq a \leq |\mathcal{X}|^{-1}$, 记

$$g(a) = H(\{1 - (|\mathcal{X}| - 1)a, a, \dots, a\}) = -l(1 - (|\mathcal{X}| - 1)a) - (|\mathcal{X}| - 1)l(a),$$

其中 $l: [0, \infty) \rightarrow \mathbb{R}$ 由下定义:

$$l(a) = \begin{cases} a \log a & \text{如果 } a > 0 \\ 0 & \text{如果 } a = 0. \end{cases}$$

注意到 $g(a)$ 是 a 的一个连续函数, 满足 $g(0) = 0$ 以及 $g(|\mathcal{X}|^{-1}) = \log |\mathcal{X}|$. 由介值定理, 我们知对任意 b , $0 < b < \log |\mathcal{X}|$, 我们可以找到一个 a , 使得得到的随机变量 X 满足 $H(X) = b$.

由上讨论我们知只要让 $|\mathcal{X}|$ 充分大, 我们可以使 $H(X)$ 达到任意的正值, 从而得证. $\square$

由上面的讨论我们知如果一个随机变量的字母表是有限的, 那么它的熵一定是有限的. 若字母表为无限的, 上一讲里我们已经看到了一个无穷大的例子. 这一讲, 我们通过下面两个新例子来说明字母表为无限的, 其熵既可能是有限的, 也可能是无限的..

例 3.19

设 X 为一个取值于自然数的随机变量, 且 $P\{X = i\} = 2^{-i}$. 则

$$H_2(X) = \sum_{i=1}^{\infty} i 2^{-i} = 2,$$

为一有限值.

例 3.20

设 Y 是取值于整数对的子集

$$\{(i, j) : 1 \leq i < \infty, 1 \leq j \leq \frac{2^{2^i}}{2^i}\}$$

满足对任意 i, j ,

$$P\{Y = (i, j)\} = 2^{-2^i}$$

的一个随机变量. 则

$$H_2(Y) = - \sum_{i=1}^{\infty} \sum_{j=1}^{2^{2^i}/2^i} 2^{-2^i} \log_2 2^{-2^i} = \sum_{i=1}^{\infty} 1 = \infty,$$

故其熵为无穷.

定理 3.21: 条件作用使熵减小

$$H(X|Y) \leq H(X).$$

当且仅当 X 与 Y 相互独立时, 等号成立.

证明. $0 \leq I(X; Y) = H(X) - H(X|Y)$. □

例 3.22

设随机变量 (X, Y) 有如下的联合分布:

Y \ X	1	2
1	0	3/4
2	1/8	1/8

于是 $H(X) = H(\frac{1}{8}, \frac{7}{8}) = 0.544$ 比特, $H(X|Y = 1) = 0$ 比特, $H(X|Y = 2) = 1$ 比特. 于是我们可以得到 $H(X|Y) = \frac{3}{4}H(X|Y = 1) + \frac{1}{4}H(X|Y = 2) = 0.25$ 比特.

定理 3.23: 熵的独立界

设 $X_1, X_2, \dots, X_n$ 服从 $p(x_1, x_2, \dots, x_n)$, 则

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

当且仅当 $X_1, X_2, \dots, X_n$ 相互独立, 等号成立.

证明. 由熵的链式法则, 我们有

$$\begin{aligned} H(X_1, X_2, \dots, X_n) &= \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1) \\ &\leq \sum_{i=1}^n H(X_i). \end{aligned}$$

等号成立当且仅当对任意 i 有 X_i 和 $X_{i-1}, \dots, X_1$ 独立. 于是对任意 $x_1, \dots, x_n$, 有

$$\begin{aligned} p(x_1, \dots, x_n) &= p(x_1, x_2, \dots, x_{n-1})p(x_n) \\ &= p(x_1, x_2, \dots, x_{n-2})p(x_{n-1})p(x_n) \\ &\vdots \\ &= p(x_1)p(x_2) \cdots p(x_n), \end{aligned}$$

也就是说, $X_1, X_2, \dots, X_n$ 相互独立. □

3.4 对数和不等式及其应用**定理 3.24: 对数和不等式**

对于非负数 $a_1, a_2, \dots, a_n$ 和 $b_1, b_2, \dots, b_n$,

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^n a_i \right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}.$$

当且仅当 $\frac{a_i}{b_i} = \text{常数}$, 等号成立.

这里我们约定 $0 \log 0 = 0$, $a \log \frac{a}{0} = \infty$ (当 $a > 0$), $0 \log \frac{0}{0} = 0$.

证明. 不失一般性, 我们可以假定 $a_i > 0$, $b_i > 0$. 由于对任意的正数 t 有 $f''(t) = \frac{1}{t} \log e > 0$, 可知函数 $f(t) = t \log t$ 严格凸. 从而由 Jensen 不等式, 我们有

$$\sum \alpha_i f(t_i) \geq f\left(\sum \alpha_i t_i\right).$$

其中 $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$. 令 $\alpha_i = \frac{b_i}{\sum_{j=1}^n b_j}$, $t_i = \frac{a_i}{b_i}$, 可得

$$\sum \frac{a_i}{\sum b_j} \log \frac{a_i}{b_i} \geq \sum \frac{a_i}{\sum b_j} \log \sum \frac{a_i}{\sum b_j}.$$

这就是对数和不等式. □

注记 3.25

利用基本不等式, 我们可以给出对数和不等式的另外一个证明如下:

记 $a = \sum a_i$, $b = \sum b_i$. 我们有

$$\begin{aligned}
 & \sum_{i=1}^n a_i \log \frac{a_i}{b_i} - \left(\sum_{i=1}^n a_i \right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \\
 &= \sum_{i=1}^n a_i \log \frac{a_i}{b_i} - \sum_{i=1}^n a_i \log \frac{a}{b} \\
 &= \sum_{i=1}^n a_i \log \frac{a_i b}{a b_i} \\
 &\geq \log e \sum_{i=1}^n a_i \left(1 - \frac{a b_i}{a_i b} \right) \\
 &= \log e \sum_{i=1}^n \left(a_i - a \frac{b_i}{b} \right) = 0.
 \end{aligned}$$

定理 3.26: 相对熵的凸性

$D(p||q)$ 关于对 (p, q) 是凸的, 即, 如果 (p_1, q_1) 和 (p_2, q_2) 为两对概率质量函数, 则对所有的 $0 \leq \lambda \leq 1$, 有

$$D(\lambda p_1 + (1 - \lambda)p_2 || \lambda q_1 + (1 - \lambda)q_2) \leq \lambda D(p_1 || q_1) + (1 - \lambda)D(p_2 || q_2).$$

证明. 由对数不等式我们有:

$$\begin{aligned}
 (\lambda p_1(x) + (1 - \lambda)p_2(x)) \log \frac{\lambda p_1(x) + (1 - \lambda)p_2(x)}{\lambda q_1(x) + (1 - \lambda)q_2(x)} \\
 \leq \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1 - \lambda)p_2(x) \log \frac{(1 - \lambda)p_2(x)}{(1 - \lambda)q_2(x)}
 \end{aligned}$$

对 x 求和我们可以得到所需要的结论. $\square$

定理 3.27: 熵的凹性

$H(p)$ 是关于 p 的凹函数.

证明. 我们有

$$H(p) = \log |\mathcal{X}| - D(p||u).$$

其中 u 为 $|\mathcal{X}|$ 个结果的均匀分布. 从而 H 的凹性可由 D 的凸性直接得到. $\square$

定理 3.28

设 $(X, Y) \sim p(x, y) = p(x)p(y|x)$. 如果固定 $p(y|x)$, 则互信息 $I(X; Y)$ 是关于 $p(x)$ 的凹函数; 而如果固定 $p(x)$, 则互信息 $I(X; Y)$ 是关于 $p(y|x)$ 的凸函数.

证明. 为了证明第一部分, 将互信息展开

$$I(X; Y) = H(Y) - H(Y|X) = H(Y) - \sum_x p(x)H(Y|X = x).$$

如果固定 $p(y|x)$, 则 $p(y)$ 时关于 $p(x)$ 的线性函数. 因而, 关于 $p(y)$ 的凹函数 $H(Y)$ 也是 $p(x)$ 的凹函数. 上式中的第二项是关于 $p(x)$ 的凹函数. 因此, 它们的差仍是关于 $p(x)$ 的凹函数.

为证明第二部分, 我们先固定 $p(x)$, 同时考虑两个不同的条件分布 $p_1(y|x)$ 和 $p_2(y|x)$. 对应的联合分布为 $p_1(x, y) = p(x)p_1(y|x)$ 和 $p_2(x, y) = p(x)p_2(y|x)$, 同时对应的边缘分布为 $p(x)$, $p_1(y)$ 以及 $p(x)$, $p_2(y)$. 我们考虑条件分布

$$p_\lambda(y|x) = \lambda p_1(y|x) + (1 - \lambda)p_2(y|x),$$

为 $p_1(y|x)$ 和 $p_2(y|x)$ 的一个凸组合, 其中 $0 \leq \lambda \leq 1$. 对应的联合分布为

$$p_\lambda(x, y) = \lambda p_1(x, y) + (1 - \lambda)p_2(x, y).$$

且 Y 的分布也是一个凸组合

$$p_\lambda(y) = \lambda p_1(y) + (1 - \lambda)p_2(y).$$

因此, 如果设 $q_\lambda(x, y) = p(x)p_\lambda(y)$ 为边际分布的乘积, 则有

$$q_\lambda(x, y) = \lambda q_1(x, y) + (1 - \lambda)q_2(x, y).$$

由于互信息是联合分布和边际分布乘积的相对熵, 有

$$I_\lambda(X; Y) = D(p_\lambda(x, y) \| q_\lambda(x, y)).$$

相对熵 $D(p \| q)$ 为关于二元对 (p, q) 的凸函数, 由此可知,

$$\begin{aligned} I(X; Y) &= D(p_\lambda(x, y) \| q_\lambda(x, y)) \\ &= D(\lambda p_1(x, y) + (1 - \lambda)p_2(x, y) \| \lambda q_1(x, y) + (1 - \lambda)q_2(x, y)) \\ &\leq \lambda D(p_1(x, y) \| q_1(x, y)) + (1 - \lambda)D(p_2(x, y) \| q_2(x, y)) \\ &= \lambda I_1(X; Y) + (1 - \lambda)I_2(X; Y). \end{aligned}$$

于是互信息是条件分布的凸函数. □

第 4 讲 数据处理不等式与费诺不等式

4.1 数据处理不等式

定理 4.1: 数据处理不等式

若 $X \rightarrow Y \rightarrow Z$, 则有 $I(X; Y) \geq I(X; Z)$.

上面结果说明, 数据处理过程中只会失掉一些信息, 绝不会创造出新的信息.

证明. 由链式法则, 将互信息以两种不同方式展开:

$$\begin{aligned} I(X; Y, Z) &= I(X; Z) + I(X; Y|Z) \\ &= I(X; Y) + I(X; Z|Y). \end{aligned}$$

由于在 Y 给定的条件下, X 和 Z 是条件独立的, 因此有 $I(X; Z|Y) = 0$. 又由于 $I(X; Y|Z) \geq 0$, 则有

$$I(X; Y) \geq I(X; Z),$$

当且仅当 $I(X; Y|Z) = 0$ (即 $X \rightarrow Z \rightarrow Y$ 构成马尔可夫链) 等号成立. 类似地可以证明 $I(Y; Z) \geq I(X; Z)$. $\square$

推论 4.2

如果 $Z = g(Y)$, 则 $I(X; Y) \geq I(X; g(Y))$.

推论 4.3

如果 $X \rightarrow Y \rightarrow Z$, 则 $I(X; Y|Z) \leq I(X; Y)$.

证明. 在数据处理不等式的证明中我们看到

$$I(X; Z) + I(X; Y|Z) = I(X; Y) + I(X; Z|Y).$$

由马尔可夫性 $I(X; Z|Y) = 0$, 又我们有 $I(X; Z) \geq 0$, 从而

$$I(X; Y|Z) \leq I(X; Y). \quad \square$$

注记 4.4

注意当 X, Y, Z 不构成马尔可夫链时, 有可能 $I(X; Y|Z) > I(X; Y)$.

设 X, Y 是相互独立的二元随机变量, $Z = X + Y$. 则 $I(X; Y) = 0$, 但 $I(X; Y|Z) = H(X|Z) - H(X|Y, Z) = H(X|Z) = P(Z = 1)H(X|Z = 1) = \frac{1}{2}$ 比特. (注意到当 $Z = 0$ 或 2 时, X 的值就确定下来了).

推论 4.5

如果 $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$ 构成一个马尔可夫链, 则对任意满足 $1 \leq i \leq j \leq k \leq n$ 的整数 i, j, k, l , 我们有

$$I(X_i; X_l) \leq I(X_j; X_k).$$

4.2 充分统计量

假定有一族以参数 θ 指示的概率密度函数 $\{f_\theta(x)\}$, 设 X 是从其中一个分布抽取的样本. 设 $T(X)$ 为任意一个统计量 (样本的函数), 如样本均值或样本方差, 那么 $\theta \rightarrow X \rightarrow T(X)$, 且由数据处理不等式, 对于 θ 的任何分布, 有

$$I(\theta; T(X)) \leq I(\theta; X)$$

然而, 若等号成立, 则表明无信息损失.

定义 4.6

如果对 θ 的任何分布, 在给定 $T(X)$ 的情况下, X 独立于 θ (即 $\theta \rightarrow T(x) \rightarrow X$ 构成马尔可夫链), 则称函数 $T(X)$ 关于分布族 $\{f_\theta(x)\}$ 的充分统计量.

这个定义等价于数据处理不等式中等号成立的条件, 即对 θ 的任意分布, 有

$$I(\theta; X) = I(\theta; T(X)).$$

因此充分统计量保持互信息不变, 反之亦然.

例 4.7

设 $X_1, X_2, \dots, X_n$ 是抛掷硬币过程所产生的独立同分布 (i.i.d.) 序列, 其中 $X_i \in \{0, 1\}$, 参数 $\theta = P\{X_i = 1\}$ 未知. 若给定 n , 则序列中出现 1 的个数, 即 $T(X_1, X_2, \dots, X_n) = \sum_{i=1}^n X_i$, 是关于 θ 的一个充分统计量. 事实上, 可以证明在给定 T 的情况下, 所有出现相同数目 1 的序列都是等可能的, 且独立于参数 θ . 具体讲

$$\begin{aligned} P\{(X_1, X_2, \dots, X_n) = (x_1, x_2, \dots, x_n) \mid \sum_{i=1}^n X_i = k\} \\ = \begin{cases} \frac{1}{\binom{n}{k}} & \text{如果 } \sum x_i = k \\ 0 & \text{其他} \end{cases} \end{aligned}$$

所以 $\theta \rightarrow \sum X_i \rightarrow (X_1, X_2, \dots, X_n)$ 构成马尔可夫链, T 是关于 θ 的充分统计量.

例 4.8

如果 X 服从均值为 θ , 方差为 1 的正态分布, 即

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2} = \mathcal{N}(\theta, 1)$$

且 $X_1, X_2, \dots, X_n$ 相互独立地服从该分布, 那么样本均值 $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ 为关于 θ 的充分统计量. 可以验证, 在给定 $\bar{X}_n$ 和 n 的条件下, $X_1, X_2, \dots, X_n$ 的条件分布不依赖于 θ .

例 4.9

如果 $f_\theta = \text{Uniform}(\theta, \theta + 1)$, 那么

$$T(X_1, X_2, \dots, X_n) = (\max\{X_1, X_2, \dots, X_n\}, \min\{X_1, X_2, \dots, X_n\}).$$

是关于 θ 的充分统计量.

定义 4.10

统计量 $T(X)$ 称为是一个关于 $\{f_\theta\}$ 极小充分统计量, 如果它可以表示为任意其它充分统计量 U 的函数. 由此我们可知

$$\theta \rightarrow T(X) \rightarrow U(X) \rightarrow X.$$

4.3 费诺不等式

我们观察与 X 相关的随机变量 Y , 相应的条件分布为 $p(y|x)$, 通过 Y 计算函数 $g(Y) = \hat{X}$, 其中 $\hat{X}$ 是对 X 的估计, 取值空间为 $\hat{\mathcal{X}}$. 我们并不要求 $\hat{X}$ 必须与 $\mathcal{X}$ 相同, 也允许函数 $g(Y)$ 是随机的. 我们希望得到事件 $\{\hat{X} \neq X\}$ 的概率的一个界. 我们记

$$P_e = \Pr\{\hat{X} \neq X\}.$$

注意到 $X \rightarrow Y \rightarrow \hat{X}$ 构成一个马尔可夫链.

定理 4.11: 费诺不等式

对任何满足 $X \rightarrow Y \rightarrow \hat{X}$ 的估计量 $\hat{X}$, 设 $P_e = P\{X \neq \hat{X}\}$, 有

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y)$$

或

$$P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}.$$

证明. 我们首先定义一个随机变量

$$E = \begin{cases} 1 & \text{如果 } \hat{X} \neq X \\ 0 & \text{如果 } \hat{X} = X \end{cases}$$

利用熵的链式法则将 $H(E, X|\hat{X})$ 以两种不同方式展开, 有

$$\begin{aligned} H(E, X|\hat{X}) &= H(X|\hat{X}) + H(E|X, \hat{X}) \\ &= H(E|\hat{X}) + H(X|E, \hat{X}) \end{aligned}$$

因为 E 是 X 和 $\hat{X}$ 的函数, 所以, 条件熵 $H(E|X, \hat{X})$ 等于 0. 由于条件作用使熵减少, 可知 $H(E|\hat{X}) \leq H(E) = H(P_e)$ (因为 E 是二值随机变量, 故 $H(E) = H(P_e)$). 对于剩余项 $H(X|E, \hat{X})$ 可以界定如下:

$$\begin{aligned} H(X|E, \hat{X}) &= P(E=0)H(X|\hat{X}, E=0) + P(E=1)H(X|\hat{X}, E=1) \\ &\leq (1 - P_e)0 + P_e \log |\mathcal{X}|. \end{aligned}$$

上述不等式成立是因为当 $E=0$ 时, $X = \hat{X}$; 当 $E=1$ 时, 条件熵的上界为 X 的可能取值数目的对数值. 综合这些结果, 可得

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X}).$$

因为 $X \rightarrow Y \rightarrow \hat{X}$ 构成马尔可夫链, 由数据处理不等式可知

$$I(X; \hat{X}) \leq I(X; Y),$$

从而 $H(X|\hat{X}) \geq H(X|Y)$. 于是, 有

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y).$$

□

推论 4.12

对任意两个随机变量 X 和 Y , 设 $p = P(X \neq Y)$,

$$H(p) + p \log |\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y).$$

推论 4.13

设 $P_e = \Pr(X \neq \hat{X})$, $\hat{X} : \mathcal{Y} \rightarrow \mathcal{X}$, 则

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y).$$

注记 4.14

假定没有任何关于 Y 的知识, 只能在毫无信息的情况下对 X 进行推测. 设 $X \in \{1, 2, \dots, m\}$ 且 $p_1 \geq p_2 \geq \dots \geq p_m$, 则对 X 的最佳估计是 $\hat{X} = 1$, 而此时误差概率 $P_e = 1 - p_1$. 费诺不等式变为

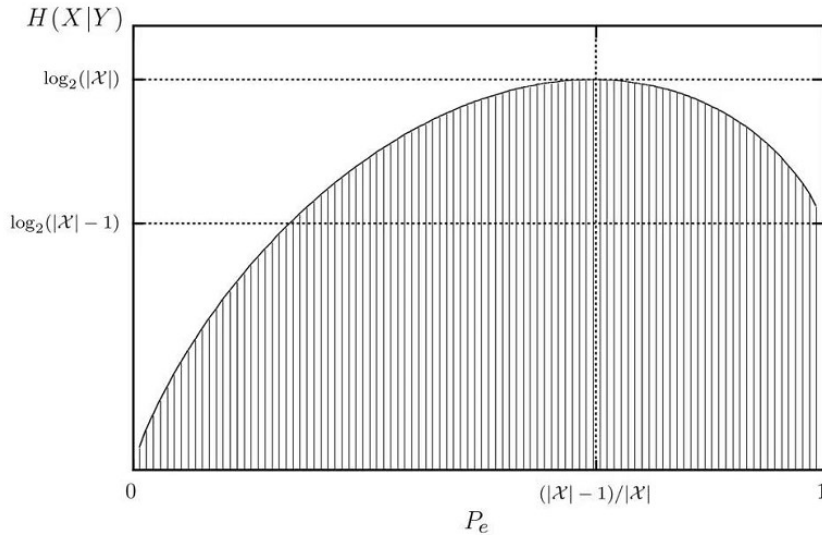
$$H(P_e) + P_e \log(m - 1) \geq H(X)$$

且概率密度函数

$$(p_1, p_2, \dots, p_m) = (1 - P_e, \frac{P_e}{m-1}, \dots, \frac{P_e}{m-1})$$

可以达到等号成立的界. 因此, 费诺不等式是精确的.

我们可以更仔细地研究一下费诺不等式中的 P_e 的上下界. 由费诺不等式我们有下图, 其中阴影部分表示可达区域.



由上图, 我们可知: 记 $f(x) := H(x) + x \cdot \log(|\mathcal{X}| - 1)$, 则当 $\log(|\mathcal{X}| - 1) < H(X|Y) \leq \log_2(|\mathcal{X}|)$ 时, 我们有 P_e 的上下界如下:

$$0 < \inf\{a : f(a) \geq H(X|Y)\} \leq P_e \leq \sup\{a : f(a) \geq H(X|Y)\} < 1,$$

而当 $0 < H(X|Y) \leq \log_2(|\mathcal{X}| - 1)$ 时, 我们有下界

$$P_e \geq \inf\{a : f(a) \geq H(X|Y)\} > 0.$$

我们可以利用基本不等式给出费诺不等式的另一个证明.

费诺不等式的另一个证明. 注意到 $X \rightarrow Y \rightarrow \hat{X}$ 构成了一个马尔可夫链, 由信息处理不等式我们有

$$I(X; Y) \geq I(X, \hat{X}),$$

从而 $H(X|Y) \leq H(X|\hat{X})$.

注意到 $P_e = \sum_{x \in \mathcal{X}} \sum_{\hat{x} \in \mathcal{X}, \hat{x} \neq x} P_{x, \hat{X}}(x, \hat{x})$ 以及 $1 - P_e = \sum_{x \in \mathcal{X}} \sum_{\hat{x} \in \mathcal{X}, \hat{x} = x} P_{x, \hat{X}}(x, \hat{x}) = \sum_{x \in \mathcal{X}} P_{X, \hat{X}}(x, x)$, 我们有:

$$\begin{aligned} & H(X|\hat{x}) - H(P_e) - P_e \log(|\mathcal{X}| - 1) \\ &= \sum_{x \in \mathcal{X}} \sum_{\hat{x} \in \mathcal{X}, \hat{x} \neq x} P_{x, \hat{X}}(x, \hat{x}) \log \frac{1}{P_{X|\hat{X}}(x|\hat{x})} + \sum_{x \in \mathcal{X}} P_{X, \hat{X}}(x, x) \log \frac{1}{P_{X|\hat{X}}(x|x)} \\ &\quad - \left[\sum_{x \in \mathcal{X}} \sum_{\hat{x} \in \mathcal{X}, \hat{x} \neq x} P_{x, \hat{X}}(x, \hat{x}) \right] \log \frac{|\mathcal{X}| - 1}{P_e} + \left[\sum_{x \in \mathcal{X}} P_{X, \hat{X}}(x, x) \right] \log(1 - P_e) \\ &= \sum_{x \in \mathcal{X}} \sum_{\hat{x} \in \mathcal{X}, \hat{x} \neq x} P_{x, \hat{X}}(x, \hat{x}) \log \frac{P_e}{P_{X|\hat{X}}(x|\hat{x})(|\mathcal{X}| - 1)} + \sum_{x \in \mathcal{X}} P_{X, \hat{X}}(x, x) \log \frac{1 - P_e}{P_{X|\hat{X}}(x|x)} \\ &\leq \log e \sum_{x \in \mathcal{X}} \sum_{\hat{x} \in \mathcal{X}, \hat{x} \neq x} P_{x, \hat{X}}(x, \hat{x}) \left[\frac{P_e}{P_{X|\hat{X}}(x|\hat{x})(|\mathcal{X}| - 1)} - 1 \right] \\ &\quad + \log e \sum_{x \in \mathcal{X}} P_{X, \hat{X}}(x, x) \left[\frac{1 - P_e}{P_{X|\hat{X}}(x|x) - 1} \right] \\ &= \log e \left[\frac{P_e}{(|\mathcal{X}| - 1)} \sum_{x \in \mathcal{X}} \sum_{\hat{x} \in \mathcal{X}: \hat{x} \neq x} P_{\hat{X}}(\hat{x}) - \sum_{x \in \mathcal{X}, x \neq \hat{x}} P_{X, \hat{X}}(x, \hat{x}) \right] \\ &\quad + \log e \left[(1 - P_e) \sum_{x \in \mathcal{X}} P_{\hat{X}}(x) - \sum_{x \in \mathcal{X}} P_{X, \hat{X}}(x, x) \right] \\ &= \log e \left[\frac{P_e}{|\mathcal{X}| - 1} (|\mathcal{X}| - 1) - P_e \right] + \log e \left[(1 - P_e) - (1 - P_e) \right] \\ &= 0. \end{aligned}$$

其中不等号是利用了基本不等式. □

如果字母表 $\mathcal{X}$ 元素个数有限, 那么由费诺不等式我们知道若 $P_e \rightarrow 0$, 则 $H(X|\hat{X}) \rightarrow 0$. 但如果 $\mathcal{X}$ 元素个数是可数多个, 结论就不一定成立了, 我们来看一个例子.

例 4.15

设 $\hat{X}$ 以概率 1 等于 0. 设 Z 是一个取值 $\{0, 1\}$ 的二元随机变量. 设 Y 是例 3.10 中定义的无穷熵的随机变量. 定义随机变量 X 满足

$$X = \begin{cases} 0 & \text{如果 } Z = 0, \\ Y & \text{如果 } Z = 1. \end{cases}$$

则我们有 $P_e = P(X \neq \hat{X}) = P(Z = 1)$. 于是对任意 $P_e > 0$ 都有

$$\begin{aligned} H(X|\hat{X}) &= H(X) \geq H(X|Z) \\ &= P(Z = 0)H(X|Z = 0) + P(Z = 1)H(X|Z = 1) \\ &= (1 - P_e) \cdot 0 + P_e \cdot H(Y) = \infty. \end{aligned}$$

于是当 $P_e \rightarrow 0$ 时, 我们不能得出 $H(X|\hat{X}) \rightarrow 0$.

设 X 和 X' 为两个独立同分布的随机变量, 则它们有相同的熵 $H(X)$, 那么 $X = X'$ 的概率为

$$P(X = X') = \sum_x p^2(x).$$

由此我们可以得到如下的不等式:

引理 4.16

如果 X 和 X' 为独立同分布, 则它们有相同的熵 $H(X)$, 我们有

$$P(X = X') \geq 2^{-H(X)}.$$

当且仅当 X 服从均匀分布, 等号成立.

证明. 设 $X \sim p(x)$. 由 Jensen 不等式, 我们有

$$2^{E \log p(X)} \leq E 2^{\log p(X)},$$

从而有

$$2^{-H(X)} = 2^{\sum p(x) \log p(x)} \leq \sum p(x) 2^{\log p(x)} = \sum p^2(x). \quad \square$$

推论 4.17

设 X 和 Y 相互独立, 且 $X \sim p(x)$, $X' \sim r(x)$, $x, x' \in \mathcal{X}$, 那么

$$\begin{aligned} P(X = X') &\geq 2^{-H(p) - D(p \| r)} \\ P(X = X') &\geq 2^{-H(r) - D(r \| p)} \end{aligned}$$

证明. 我们有

$$\begin{aligned} 2^{-H(p) - D(p \| r)} &= 2^{\sum p(x) \log p(x) + \sum p(x) \log \frac{r(x)}{p(x)}} \\ &= 2^{\sum p(x) \log r(x)} \\ &\leq \sum p(x) 2^{\log r(x)} \\ &= \sum p(x) r(x) \\ &= P(X = X'), \end{aligned}$$

其中上面的不等号成立是由函数 $f(y) = 2^y$ 的凸性和 Jensen 不等式得到. $\square$

第5讲 熵的公理化定义 *

定理 5.1

设 $H(p_1, \dots, p_n)$ 为满足下列性质的函数:

- (1) H 在 p_i 上连续, 其中 $i = 1, \dots, n$.
- (2) 若对所有的 $i = 1, \dots, n$ 都有 $p_i = \frac{1}{n}$, 则 H 关于 n 单调递增.
- (3) (分组公理 / Grouping axiom) 对于 $P = (p_1, \dots, p_n)$, 令 $Q = (q_1, \dots, q_s)$ 定义为 $q_j \triangleq \sum_{i=n_{j-1}+1}^{n_j} p_i$ (其中 $j = 1, \dots, s$, 且 $0 = n_0 < n_1 < \dots < n_s = n$), 则:

$$H(P) = H(Q) + \sum_{j=1}^s q_j H\left(\frac{p_{n_{j-1}+1}}{q_j}, \dots, \frac{p_{n_j}}{q_j}\right).$$

那么 $H(p_1, \dots, p_n)$ 必然具有形如 $-K \cdot \sum_{i=1}^n p_i \cdot \log p_i$ 的形式, 其中 K 为常数因子.

在证明该定理之前, 我们简要讨论一下不确定性度量的这三个公理. 前两条性质看起来相当合理: 概率的微小变化应该对不确定性只有微小的影响; 而具有 n 个可能结果的实验的不确定性, 应该大于具有 $m < n$ 个可能结果的实验的不确定性. 在极极端情况 $n = 1$ 中 (例如实验结果已知), 此时不存在不确定性. 然而, 第三条公理并不那么显而易见. 我们可以用下面的例子来说明: 假设我们在酒店里寻找一个人. 设 $p_{n_{j-1}+i}$ (其中 $i \leq n_j - n_{j-1}$) 表示此人在 j 楼 i 号房间的概率, 那么 q_j 就是此人在 j 楼的概率. 当我们已经知道此人在 j 楼时, 寻找此人的剩余不确定性即为:

$$H\left(\frac{p_{n_{j-1}+1}}{q_j}, \dots, \frac{p_{n_j}}{q_j}\right).$$

证明. 定义 $A(n) \triangleq H\left(\frac{1}{n}, \dots, \frac{1}{n}\right)$. 设对某些 $s, m \in \mathbb{N}$, 有 $n = s^m$. 令 $Q = (\frac{1}{s}, \dots, \frac{1}{s})$, 其中 $\frac{1}{s} = q_j = \sum_{(j-1)s^{m-1}+1}^{js^{m-1}} \frac{1}{s^m}$ (对 $j = 1, \dots, s$). 应用分组公理可得:

$$A(s^m) = A(s) + s \cdot \frac{1}{s} \cdot A(s^{m-1}) = A(s) + A(s^{m-1})$$

迭代此过程可得:

$$A(s^m) = m \cdot A(s)$$

类似地, 对于某些 $t, \ell \in \mathbb{N}$, 我们可得:

$$A(t^\ell) = \ell \cdot A(t)$$

对于任意 $\ell \in \mathbb{N}$, 存在一个数 $m \in \mathbb{N}$, 使得 $s^m \leq t^\ell < s^{m+1}$. 由此我们可以推出:

$$\frac{m \cdot \log s}{\ell \cdot \log s} \leq \frac{\ell \cdot \log t}{\ell \cdot \log s} < \frac{(m+1) \cdot \log s}{\ell \cdot \log s}$$

因此:

$$\frac{m}{\ell} \leq \frac{\log t}{\log s} < \frac{m}{\ell} + \frac{1}{\ell}$$

由单调性 (公理 (2)) 可得 $A(s^m) \leq A(t^\ell) \leq A(s^{m+1})$, 进而有 $m \cdot A(s) \leq \ell A(t) \leq (m+1) \cdot A(s)$. 经过与上述相同的计算可得:

$$\frac{m}{\ell} \leq \frac{A(t)}{A(s)} \leq \frac{m}{\ell} + \frac{1}{\ell}$$

综合上述两个结果, 可得:

$$\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| \leq \frac{1}{\ell}$$

固定 s 并让 $\ell \rightarrow \infty$, 最终我们得到:

$$A(t) = \frac{A(s)}{\log s} \cdot \log t = K \cdot \log t$$

其中 $K \triangleq \frac{A(s)}{\log s}$ 为常数 (因为 s 是固定的)。至此, 定理对于均匀分布得到了证明。对于一般分布 $P = (p_1, \dots, p_n)$, 当 p_i 均为有理数时, 不失一般性可以假设所有的 p_i 具有相同的分母, 即 $p_i = \frac{n_i}{\sum_{i=1}^n n_i}$ (对 $i = 1, \dots, n$)。现将分组公理应用于 P 以及均匀分布 $(\frac{1}{\sum_{i=1}^n n_i}, \dots, \frac{1}{\sum_{i=1}^n n_i})$, 可得:

$$K \cdot \log \left(\sum_{i=1}^n n_i \right) = H(p_1, \dots, p_n) + \sum_{i=1}^n p_i \cdot K \log n_i$$

整理得:

$$\begin{aligned} H(p_1, \dots, p_n) &= K \left(\sum_{i=1}^n p_i \cdot \log \left(\sum_{j=1}^n n_j \right) - \sum_{i=1}^n p_i \cdot \log n_i \right) \\ &= K \left(- \sum_{i=1}^n p_i \cdot \log \frac{n_i}{\sum_{j=1}^n n_j} \right) \\ &= -K \sum_{i=1}^n p_i \cdot \log p_i \end{aligned}$$

最后, 由于要求 H 是连续的 (公理 (1)), 因此当 p_i 为实数时:

$$H(p_1, \dots, p_n) = -K \cdot \sum_{i=1}^n p_i \cdot \log p_i$$

同样成立。至此证明了: 满足定理三条公理的函数 $H(p_1, \dots, p_n)$ 必须具有 $H(p_1, \dots, p_n) = -K \sum_{i=1}^n p_i \cdot \log p_i$ 的形式。另一方面, 容易验证对于 $P = (\frac{1}{n}, \dots, \frac{1}{n})$, $-K \sum_{i=1}^n p_i \cdot \log p_i$ 关于 p_i ($i = 1, \dots, n$) 是连续的, 且关于 n 单调递增。同时分组公理也成立, 因为对于 $P = (p_1, \dots, p_n)$ 和 $Q = (q_1, \dots, q_s)$ (其中 $q_j = \sum_{i=n_{j-1}+1}^{n_j} p_i$), 有:

$$\begin{aligned} &H(Q) + \sum_{j=1}^s q_j H \left(\frac{p_{n_{j-1}+1}}{q_j}, \dots, \frac{p_{n_j}}{q_j} \right) \\ &= -K \cdot \left(\sum_{j=1}^s q_j \cdot \log q_j + \sum_{j=1}^s q_j \sum_{i=n_{j-1}+1}^{n_j} \frac{p_i}{q_j} \cdot \log \frac{p_i}{q_j} \right) \\ &= -K \cdot \left(\sum_{j=1}^s q_j \cdot \log q_j + \sum_{i=n_{j-1}+1}^{n_j} p_i \cdot \log p_i - \log q_j \cdot \sum_{i=n_{j-1}+1}^{n_j} p_i \right) \quad \square \\ &= -K \cdot \left(\sum_{i=1}^n p_i \cdot \log p_i \right) = H(P) \end{aligned}$$

辛钦 (Khinchin) 提出了以下一组公理, 它们同样可以在差一个常数因子的情况下确定熵:

- (1') $H(p, 1-p)$ 在 p 上连续.
- (2') $H(p_1, \dots, p_n)$ 是对称的 (对称性).
- (3') $H(p_1, \dots, p_n) \leq H(\frac{1}{n}, \dots, \frac{1}{n})$.

(4') (分组公理):

$$H(q_{11}, \dots, q_{1m_1}, q_{21}, \dots, q_{2m_2}, \dots, q_{n1}, \dots, q_{nm_n}) = H(p_1, \dots, p_n) + \sum_{i=1}^n p_i \cdot H\left(\frac{q_{i1}}{p_i}, \dots, \frac{q_{im_i}}{p_i}\right)$$

其中 $p_i = q_{i1} + \dots + q_{im_i}$.

这里的关键是用更弱的“对称性”性质替换了单调性. 辛钦的公理后来由法捷耶夫 (Faddeev) 进行了简化.

定理 5.2: (Faddeev 1956)

设 $H(p_1, \dots, p_n)$ 为满足下列性质的函数:

(1'') $H(p, 1-p)$ 对于 $0 \leq p \leq 1$ 是连续的.

(2'') $H(p_1, \dots, p_n)$ 对所有 $n \in \mathbb{N}$ 都是对称的.

(3'') 若 $p_n = q_1 + q_2 > 0$, 则:

$$H(p_1, \dots, p_{n-1}, q_1, q_2) = H(p_1, \dots, p_n) + p_n \cdot H\left(\frac{q_1}{p_n}, \frac{q_2}{p_n}\right).$$

则 $H(p_1, \dots, p_n) = -K \cdot \sum_{i=1}^n p_i \cdot \log p_i$, 其中常数因子 $K \in \mathbb{R}$.

为了证明该定理, 需要四个引理.

引理 5.3

$$H(1, 0) = 0.$$

证明. 由 (3'') 可得 $H(\frac{1}{2}, \frac{1}{2}, 0) = H(\frac{1}{2}, \frac{1}{2}) + \frac{1}{2}H(1, 0)$. 进一步由 (2'') 和 (3'') 可得 $H(\frac{1}{2}, \frac{1}{2}, 0) = H(0, \frac{1}{2}, \frac{1}{2}) = H(0, 1) + H(\frac{1}{2}, \frac{1}{2})$. 因此 $\frac{1}{2}H(1, 0) = H(1, 0)$, 当然也就有 $H(1, 0) = 0$. $\square$

引理 5.4

$$H(p_1, \dots, p_n, 0) = H(p_1, \dots, p_n)$$

证明. 出于对称性, 我们可以不失一般性假设 $p_n > 0$. 根据性质 (3'') 以及引理 5.3, 可得:

$$H(p_1, \dots, p_n, 0) = H(p_1, \dots, p_n) + p_n \cdot H(1, 0) = H(p_1, \dots, p_n).$$

$\square$

引理 5.5

$$H(p_1, \dots, p_{n-1}, q_1, \dots, q_m) = H(p_1, \dots, p_n) + p_n \cdot H\left(\frac{q_1}{p_n}, \dots, \frac{q_m}{p_n}\right),$$

其中 $p_n = q_1 + \dots + q_m > 0$.

证明. 对 m 使用数学归纳法: 当 $m = 2$ 时, 引理的表述即为性质 (3''). 由引理 5.4 可知, 我们只需考虑对于所有 $i = 1, \dots, m$ 均有 $q_i > 0$ 的情况. 假设存在某个 m 使得对于所有的 n 该断言都成立. 那么利用条件 (3), 我们有:

$$H(p_1, \dots, p_{n-1}, q_1, \dots, q_{m+1})$$

$$\begin{aligned}
&= H(p_1, \dots, p_{n-1}, q_1, p') + p' \cdot H\left(\frac{q_2}{p'}, \dots, \frac{q_{m+1}}{p'}\right) \\
&= H(p_1, \dots, p_n) + p_n \cdot H\left(\frac{q_1}{p_n}, \frac{p'}{p_n}\right) + p' \cdot H\left(\frac{q_2}{p'}, \dots, \frac{q_{m+1}}{p'}\right)
\end{aligned}$$

其中 $p' = q_2 + \dots + q_{m+1}$ 。此外,

$$H\left(\frac{q_1}{p_n}, \dots, \frac{q_{m+1}}{p_n}\right) = H\left(\frac{q_1}{p_n}, \frac{p'}{p_n}\right) + \frac{p'}{p_n} \cdot H\left(\frac{q_2}{p'}, \dots, \frac{q_{m+1}}{p'}\right)$$

将此式代入前一个等式中, 我们便得到了 $m+1$ 情况下的引理断言。 $\square$

引理 5.6

我们有

$$H(q_{11}, \dots, q_{1m_1}, q_{21}, \dots, q_{2m_2}, \dots, q_{n1}, \dots, q_{nm_n}) = H(p_1, \dots, p_n) + \sum_{i=1}^n p_i \cdot H\left(\frac{q_{i1}}{p_i}, \dots, \frac{q_{im_i}}{p_i}\right)$$

其中对每一个 $i = 1, \dots, n$, 都有 $p_i = q_{i1} + \dots + q_{im_i}$ 。

证明. 根据引理 5.5,

$$\begin{aligned}
&H(q_{11}, \dots, q_{1m_1}, q_{21}, \dots, q_{2m_2}, \dots, q_{n1}, \dots, q_{nm_n}) \\
&= p_n \cdot H\left(\frac{q_{n1}}{p_n}, \dots, \frac{q_{nm_n}}{p_n}\right) + H(q_{11}, \dots, q_{1m_1}, \dots, q_{n-1,1}, \dots, q_{n-1,m_{n-1}}, p_n)
\end{aligned}$$

重复迭代此过程即可得到所需的结果。 $\square$

定理 5.2 的证明. 对 $n \geq 2$, 令

$$F(n) = H\left(\frac{1}{n}, \dots, \frac{1}{n}\right),$$

并令

$$F(1) = 0.$$

(因为 $F(1) = H(1) = H(1, 0) = 0$.)

由引理 5.6, 取

$$m_1 = \dots = m_n = m,$$

可以得到

$$F(mn) = F(m) + F(n).$$

进一步应用引理 5.5, 可以得到相应的关系.

现在, 由于

$$\lim_{n \rightarrow \infty} H\left(\frac{1}{n}, \frac{n-1}{n}\right) = H(0, 1) = 0,$$

我们有

$$\lim_{n \rightarrow \infty} \left(F(n) - \frac{n-1}{n} F(n-1)\right) = 0.$$

于是

$$\lim_{n \rightarrow \infty} (F(n) - F(n-1)) = 0.$$

综上,

$$F: \mathbb{N} \rightarrow \mathbb{R}$$

是一个满足

$$F(nm) = F(n) + F(m)$$

以及

$$\lim_{n \rightarrow \infty} (F(n) - F(n-1)) = 0$$

的乘性函数.

下面的引理中我们将证明, $F(n)$ 必然具有

$$c \log n$$

的形式, 其中 c 是一个常数. 有了这一点以后, 我们就可以沿着 Shannon 定理证明中的思路继续, 从而得到所需的结果.

引理 5.7

设

$$f : \mathbb{N} \rightarrow \mathbb{R}$$

是一个满足以下条件的乘法函数:

1.

$$f(nm) = f(n) + f(m);$$

2.

$$\lim_{n \rightarrow \infty} (f(n+1) - f(n)) = 0.$$

那么存在某个常数 $c \in \mathbb{R}$, 使得

$$f(n) = c \cdot \log n.$$

这个引理最初由 Erdős 提出, 后来由 Faddeev 重新发现. 下面的证明归功于 Rényi. 该证明使用了另外一个引理, 其中加入了关于该函数的一个附加条件.

引理 5.8

设

$$g : \mathbb{N} \rightarrow \mathbb{R}$$

是一个满足以下条件的乘法函数:

1.

$$g(nm) = g(n) + g(m);$$

2.

$$\lim_{n \rightarrow \infty} (g(n+1) - g(n)) = 0;$$

3. 存在某个 $N > 1$, 使得

$$g(N) = 0.$$

那么

$$\lim_{n \rightarrow \infty} \frac{g(n)}{\log n} = 0.$$

在证明引理 5.8 之前, 我们先证明引理 5.7 可以很容易地由引理 5.8 得到. 为此, 对于固定的 $N > 1$, 我们定义

$$g(n) \triangleq f(n) - \frac{f(N) \cdot \log n}{\log N},$$

其中 f 满足引理 5.7 所要求的性质。

显然, g 满足引理 5.8 中的性质 (ii) 和 (iii), 并且性质 (i) 也成立, 这是因为

$$\begin{aligned} g(nm) &= f(nm) - \frac{f(N) \cdot \log(nm)}{\log N} \\ &= f(n) - \frac{f(N) \cdot \log n}{\log N} + f(m) - \frac{f(N) \cdot \log m}{\log N} \\ &= g(n) + g(m). \end{aligned}$$

由引理 5.8,

$$0 = \lim_{n \rightarrow \infty} \frac{g(n)}{\log n} = \lim_{n \rightarrow \infty} \left(\frac{f(n)}{\log n} - \frac{f(N)}{\log N} \right),$$

因此

$$\lim_{n \rightarrow \infty} \frac{f(n)}{\log n} = \frac{f(N)}{\log N} \triangleq c,$$

其中 $c \in \mathbb{R}$ 为某个常数。

由于 N 可以任意选取, 我们有

$$\frac{f(N')}{\log N'} = c$$

对所有 $N' > 1$ 都成立, 因此引理 5.7 得证。

引理 5.8 的证明. 首先定义函数

$$G: \mathbb{N} \cup \{-1, 0\} \rightarrow \mathbb{R}$$

如下:

$$G(-1) \triangleq 0,$$

并且对于 $k = 0, 1, 2, \dots$, 定义

$$G(k) \triangleq \max_{N^k \leq n < N^{k+1}} |g(n)|.$$

此外, 对于 $k = 0, 1, 2, \dots$, 令

$$\delta_k \triangleq \max_{N^k \leq n < N^{k+1}} |g(n+1) - g(n)|.$$

由于

$$\lim_{n \rightarrow \infty} (g(n+1) - g(n)) = 0,$$

对于任意 $\varepsilon > 0$, 我们可以找到一个 $n(\varepsilon)$, 使得当 $N^k > n \geq n(\varepsilon)$ 时,

$$|g(n+1) - g(n)| \leq \varepsilon,$$

从而有 $\lim_{k \rightarrow \infty} \delta_k = 0$. 进一步注意到, 当

$$N^k \leq n < N^{k+1}$$

时,

$$\frac{g(n)}{\log n} \leq \frac{G(k)}{k \log N}.$$

因此, 只需证明 $\lim_{k \rightarrow \infty} \frac{G(k)}{k \log N} = 0$, 或者更好地, 由于 N 是固定的, 只需证明

$$\lim_{k \rightarrow \infty} \frac{G(k)}{k} = 0.$$

于是取 n 和 k , 满足 $N^k \leq n < N^{k+1}$, 并令

$$n' \triangleq N \left\lfloor \frac{n}{N} \right\rfloor.$$

我们有

$$g(n) = g(n') + \sum_{\ell=n'}^{n-1} (g(\ell+1) - g(\ell)).$$

因此

$$\begin{aligned} |g(n)| &\leq |g(n')| + \sum_{\ell=n'}^{n-1} |g(\ell+1) - g(\ell)| \\ &\leq |g(n')| + N\delta_k. \end{aligned}$$

根据引理 14 中的性质 (i) 和 (iii), 我们有

$$g(n') = g\left(N \cdot \left\lfloor \frac{n}{N} \right\rfloor\right) = g(N) + g\left(\left\lfloor \frac{n}{N} \right\rfloor\right) = g\left(\left\lfloor \frac{n}{N} \right\rfloor\right),$$

因此

$$|g(n)| \leq \left|g\left(\left\lfloor \frac{n}{N} \right\rfloor\right)\right| + N\delta_k.$$

该不等式特别适用于满足 $|g(n)| = G(k)$ 的 n , 并且由于 $N^{k-1} \leq \left\lfloor \frac{n}{N} \right\rfloor < N^k$, 我们得到

$$G(k) \leq G(k-1) + N\delta_k.$$

进一步地, 对于任意 $m \in \mathbb{N}$

$$\sum_{k=0}^m G(k) \leq \sum_{k=0}^m G(k-1) + N \cdot \sum_{k=0}^m \delta_k.$$

我们可以得到

$$G(m) = \sum_{k=0}^m (G(k) - G(k-1)) \leq N \sum_{k=0}^m \delta_k,$$

从而

$$\frac{G(m)}{m} \leq N \frac{\sum_{k=0}^m \delta_k}{m}.$$

由于 $\lim_{k \rightarrow \infty} \delta_k = 0$, 同样有 $\lim_{m \rightarrow \infty} \frac{1}{m} \sum_{k=0}^m \delta_k = 0$. 因此 $\lim_{m \rightarrow \infty} \frac{G(m)}{m} = 0$, 证明完毕。 $\square$

第6讲 习题课 1

例 6.1

不断抛掷一枚均匀硬币，直到第一次出现正面为止。令 X 表示所需的抛掷次数。

(a) 求熵 $H(X)$ (以比特为单位)。以下表达式可能会有用：

$$\sum_{n=0}^{\infty} r^n = \frac{1}{1-r}, \quad \sum_{n=0}^{\infty} nr^n = \frac{r}{(1-r)^2}.$$

(b) 根据该分布抽取了一个随机变量 X 。构造一个形式为“ X 是否包含在集合 S 中？”的“有效”的“是-否”问题序列，以确定 X 的值。比较 $H(X)$ 与确定 X 所需的期望问题数。

证明. (a) 抛掷到首次出现正面时的次数 X 服从参数 $p = 1/2$ 的几何分布，其中 $P(X = n) = pq^{n-1}$, $n \in \{1, 2, \dots\}$. 因此 X 的熵为：

$$\begin{aligned} H(X) &= - \sum_{n=1}^{\infty} pq^{n-1} \log(pq^{n-1}) \\ &= - \left[\sum_{n=0}^{\infty} pq^n \log p + \sum_{n=0}^{\infty} npq^n \log q \right] \\ &= \frac{-p \log p}{1-q} - \frac{pq \log q}{p^2} \\ &= \frac{-p \log p - q \log q}{p} \\ &= H(p)/p \text{ 比特 (bits).} \end{aligned}$$

如果 $p = 1/2$, 则 $H(X) = 2$ 比特。

(b) 直观来看，最好的问题显然是那些获得“是”或“否”的几率各占一半的问题。因此，我们可以猜测，最“有效”的一系列问题为： $X = 1$ 吗？如果不是， $X = 2$ 吗？如果不是， $X = 3$ 吗？……得到的期望问题数等于 $\sum_{n=1}^{\infty} n(1/2^n) = 2$ 。这应该增强了这样一种直觉： $H(X)$ 是衡量 X 的不确定性的一种度量。事实上，在此情况下，熵恰好等于确定 X 所需的平均问题数，且在一般情况下 $E(\text{问题数}) \geq H(X)$ 。这个问题可以理解为信源编码问题。令 0 = 否，1 = 是， X = 信源，而 Y = 编码后的信源。那么上述过程中的问题集可以写成 (X, Y) 对的集合： $(1,1)$, $(2,01)$, $(3,001)$ 等。事实上，这个通过直觉得出的代码正是最小化预期问题数的最佳（哈夫曼 Huffman）码。□

例 6.2

设离散随机变量 X_1 和 X_2 的概率质量函数分别为 $p_1(\cdot)$ 和 $p_2(\cdot)$, 字母表分别为 $\mathcal{X}_1 = \{1, 2, \dots, m\}$, $\mathcal{X}_2 = \{m+1, \dots, n\}$. 设

$$X = \begin{cases} X_1 & \text{概率为 } \alpha \\ X_2 & \text{概率为 } 1 - \alpha. \end{cases}$$

(1) 试求 $H(X)$ 关于 $H(X_1)$ 、 $H(X_2)$ 和 α 的表达式。

(2) 试关于 α 最大化 $H(X)$, 证明 $2^{H(X)} \leq 2^{H_1(X)} + 2^{H_2(X)}$.

证明. 注意到

$$X = \begin{cases} X_1 & \text{概率为 } \alpha \\ X_2 & \text{概率为 } 1 - \alpha, \end{cases}$$

我们定义

$$\theta = f(X) = \begin{cases} 1 & \text{若 } X = X_1, \\ 2 & \text{若 } X = X_2, \end{cases}$$

那么我们有

$$\begin{aligned} H(X) &= H(X, f(X)) = H(\theta) + H(X|\theta) \\ &= H(\theta) + p(\theta = 1)H(X|\theta = 1) + p(\theta = 2)H(X|\theta = 2) \\ &= H(\alpha) + \alpha H(X_1) + (1 - \alpha)H(X_2) \end{aligned}$$

其中 $H(\alpha) = -\alpha \log \alpha = (1 - \alpha) \log(1 - \alpha)$. 记 $f(\alpha) = H(\alpha) + \alpha H(X_1) + (1 - \alpha)H(X_2)$, 则

$$f'(\alpha) = \log\left(\frac{1 - \alpha}{\alpha}\right) + H(X_1) - H(X_2).$$

解方程 $f'(\alpha_0) = 0$ 可以得 $\alpha_0 = \frac{2^{H(X_1)}}{2^{H(X_1)} + 2^{H(X_2)}}$. 不难看出 f 在 α_0 处取到最大值, 于是

$$\begin{aligned} H(X) &\leq -\frac{2^{H(X_1)}}{2^{H(X_1)} + 2^{H(X_2)}} \log \frac{2^{H(X_2)}}{2^{H(X_1)} + 2^{H(X_2)}} - \frac{2^{H(X_2)}}{2^{H(X_1)} + 2^{H(X_2)}} \log \frac{2^{H(X_1)}}{2^{H(X_1)} + 2^{H(X_2)}} \\ &\quad + \frac{2^{H(X_2)}}{2^{H(X_1)} + 2^{H(X_2)}} H(X_1) + \frac{2^{H(X_1)}}{2^{H(X_1)} + 2^{H(X_2)}} H(X_2) \\ &= \log(2^{H(X_1)} + 2^{H(X_2)}). \end{aligned}$$

从而 (2) 得证. □

例 6.3

设 $p_i > 0$, $i = 1, 2, \dots, m$. 证明服从分布 $(p_1, \dots, p_i, \dots, p_j, \dots, p_m)$ 的随机变量的熵, 小于或等于服从分布 $(p_1, \dots, \frac{p_i + p_j}{2}, \dots, \frac{p_i + p_j}{2}, \dots, p_m)$ 的随机变量的熵。

解. 这个问题取决于对数函数的凸性。设

$$\begin{aligned} P_1 &= (p_1, \dots, p_i, \dots, p_j, \dots, p_m) \\ P_2 &= (p_1, \dots, \frac{p_i + p_j}{2}, \dots, \frac{p_i + p_j}{2}, \dots, p_m) \end{aligned}$$

那么,

$$\begin{aligned} H(P_2) - H(P_1) &= -2 \left(\frac{p_i + p_j}{2} \right) \log \left(\frac{p_i + p_j}{2} \right) + p_i \log p_i + p_j \log p_j \\ &= -(p_i + p_j) \log \left(\frac{p_i + p_j}{2} \right) + p_i \log p_i + p_j \log p_j. \end{aligned}$$

由于 $f(x) = x \log x$ 在 $x > 0$ 时是严格凸函数, 以下不等式成立:

$$\frac{f(p_i) + f(p_j)}{2} = \frac{1}{2}(p_i \log p_i + p_j \log p_j) \geq f\left(\frac{p_i + p_j}{2}\right) = \frac{p_i + p_j}{2} \log \frac{p_i + p_j}{2}.$$

因此, $H(P_2) \geq H(P_1)$. □

例 6.4

设 X, Y 和 Z 为三个服从 $\text{Bernoulli}(\frac{1}{2})$ 的二元随机变量, 且两两独立. 在上述约束条件下, $H(X, Y, Z)$ 的最小值是多少?

证明. 我们有

$$\begin{aligned} H(X, Y, Z) &= H(X, Y) + H(Z|X, Y) \\ &\geq H(X, Y) \\ &= 2 \text{ 比特}. \end{aligned}$$

下面我们验证这个下界可以达到, 从而 $H(X, Y, Z)$ 的最小值为 2 比特.

设 $X, Y \sim \text{Bernoulli}(1/2)$, $Z = X \oplus Y$, 其中 $\oplus$ 表示模 2 加法. □

例 6.5

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

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

求 $I(X_1; X_2, X_3, \dots, X_n)$.

证明. 首先注意到 $X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_n$ 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

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

例 6.6

设 $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ 构成一个马尔可夫链. 证明:

$$I(X_1; X_3) + I(X_2; X_4) \leq I(X_1; X_4) + I(X_2; X_3).$$

证明. 注意到

$$\begin{aligned} I(X_1; X_3, X_4) &= I(X_1; X_3) + I(X_1; X_4 | X_3) \\ &= I(X_1; X_4) + I(X_1; X_3 | X_4) \end{aligned}$$

于是

$$I(X_1; X_4) - I(X_1; X_3) = I(X_1; X_4 | X_3) - I(X_1; X_3 | X_4).$$

类似地,

$$I(X_2; X_3) - I(X_2; X_4) = I(X_2; X_3 | X_4) - I(X_2; X_4 | X_3).$$

从而我们有 (由于 $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ 构成一个马尔可夫链, $I(X_1; X_4|X_3) = 0$, $I(X_2; X_4|X_3) = 0$.)

$$\begin{aligned} & I(X_1; X_4) + I(X_2; X_3) - I(X_1; X_3) + I(X_2; X_4) \\ &= I(X_1; X_4|X_3) - I(X_1; X_3|X_4) + I(X_2; X_3|X_4) - I(X_2; X_4|X_3) \\ &= I(X_2; X_3|X_4) - I(X_1; X_3|X_4) \end{aligned}$$

又由

$$\begin{aligned} I(X_1, X_2; X_3|X_4) &= I(X_1; X_3|X_4) + I(X_2; X_3|X_1, X_4) \\ &= I(X_2; X_3|X_4) + I(X_1; X_3|X_2, X_4), \end{aligned}$$

我们知

$$\begin{aligned} & I(X_1; X_4) + I(X_2; X_3) - I(X_1; X_3) + I(X_2; X_4) \\ &= I(X_2; X_3|X_1, X_4) - I(X_1; X_3|X_2, X_4). \end{aligned}$$

注意到 (这里我们再次用到 $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ 构成一个马尔可夫链这一事实) $I(X_1; X_3|X_2, X_4) = I(X_1; X_3, X_4|X_2) - I(X_1; X_4|X_2) = 0$, 于是

$$I(X_1; X_4) + I(X_2; X_3) - I(X_1; X_3) + I(X_2; X_4) = I(X_2; X_3|X_1, X_4) \geq 0.$$

从而得证. □

例 6.7

记 $\pi(n)$ 为不超过 n 的素数个数. 注意到对于任一正整数 n , 其有唯一的形如下的素因子分解

$$n = \prod_{i=1}^{\pi(n)} p_i^{X_i}.$$

其中 $p_1, p_2, p_3, \dots$ 为素数. 即 $p_1 = 2, p_2 = 3, p_3 = 5, \dots$. $X_i = X_i(n)$ 为一个非负整数, 其表示 n 的素因子分解中 p_i 的重数. 设 N 为在 $\{1, 2, \dots, n\}$ 上均匀分布的一个随机变量.

(1) 证明 $X_i(N)$ 为一个整数值随机变量, 满足

$$0 \leq X_i(N) \leq \log n.$$

(2) 证明 $\log n = H(N) \leq \pi(n) \log(\log n + 1)$. 从而 $\pi(n) \geq \frac{\log n}{\log(\log n + 1)}$. 特别地当 $n \rightarrow \infty$ 时有 $\pi(n) \rightarrow \infty$.

证明. (1) 由于 $p_i^{X_i} | N$, 所以 $2^{X_i(N)} \leq p_i^{X_i(N)} \leq N \leq n$. 从而 $X_i(N) \leq \log n$.

(2) 注意到对于每个 N , 其对应的 $(X_1(N), \dots, X_{\pi(n)}(N))$ 是唯一的, 所以

$$\begin{aligned} \log n &= H(N) \\ &= H(X_1(N), X_2(N), \dots, X_{\pi(n)}(N)) \\ &\leq H(X_1(N)) + \dots + H(X_{\pi(n)}(N)) \\ &\leq \pi(n) \log(\log(n + 1)), \end{aligned}$$

从而 $\pi(n) \geq \frac{\log n}{\log(\log n + 1)}$. 特别地当 $n \rightarrow \infty$ 时有 $\pi(n) \rightarrow \infty$. □

注记 6.8

上例的结论可以进一步加强. 设 $N \sim \text{Uniform}\{1, 2, \dots, n\}$, 其可以唯一表示为

$$N = M^2 \cdot p_1^{Y_1} \cdot p_2^{Y_2} \cdot \dots \cdot p_{\pi(n)}^{Y_{\pi(n)}},$$

其中 M 是使得 $M^2|N$ 的最大整数. 这样随机变量 Y_i 取值范围是 $\{0, 1\}$. 由于 $M^2|N$, $1 \leq M \leq \sqrt{n}$. 从而我们有

$$\begin{aligned} \log n &= H(N) \\ &= H(M, Y_1, \dots, Y_{\pi(n)}) \\ &\leq H(M) + H(Y_1) + H(Y_2) + \dots + H(Y_{\pi(n)}) \\ &\leq \frac{1}{2} \log n + \pi(n). \end{aligned}$$

从而 $\pi(n) \geq \frac{1}{2} \log n$.

例 6.9

假设 X 是一个取值于自然数集 $\mathbb{N} = \{1, 2, \dots\}$ 上的随机变量, 其概率质量函数 (PMF) 为 P . 我们将尝试寻找使该随机变量的熵有限的条件.

- (1) 假设 $\mathbb{E}[X] < \infty$. 设 Q 为成功概率为 r 的几何随机变量的 PMF, 即对于所有 $x \in \mathbb{N}$, 有 $Q(x) = (1-r)^{x-1}r$, 其中 $r \in (0, 1)$. 请用 $H(X)$ 、 $\mathbb{E}[X]$ 和 r 来表示相对熵 $D(P||Q)$.
- (2) 利用上一问的结论证明:

$$H(X) \leq \mathbb{E}[X] \cdot h_2\left(\frac{1}{\mathbb{E}[X]}\right).$$

当 X 服从几何分布时等号成立. 由此我们知如果 $\mathbb{E}[X] < \infty$ 则 X 的熵是有限的. 其中, $h_2(\cdot)$ 表示二元熵函数, 定义为 $h_2(x) = -x \log_2 x - (1-x) \log_2 (1-x)$.

- (3) 现在假设我们不知道是否有 $\mathbb{E}[X] < \infty$ 但是我们知道 $\mathbb{E}[\log X] < \infty$. 证明这也能推出 $H(X) < \infty$.

证明. (1) 展开相对熵, 我们有:

$$\begin{aligned} D(P||Q) &= \sum_{x \in \mathbb{N}} P(x) \log \frac{P(x)}{Q(x)} \\ &= \sum_{x \in \mathbb{N}} P(x) \log P(x) + \sum_{x \in \mathbb{N}} P(x) \log \frac{1}{Q(x)} \\ &= -H(X) + \sum_{x \in \mathbb{N}} P(x) \log \frac{1}{(1-r)^{x-1}r} \\ &= -H(X) + (\mathbb{E}[X] - 1) \log \frac{1}{1-r} + \log \frac{1}{r}. \end{aligned}$$

- (2) 设 Q 与上一问的分布相同. 利用 KL 散度的非负性 $D(P||Q) \geq 0$, 我们有:

$$H(X) \leq (\mathbb{E}[X] - 1) \log \left(\frac{1}{1-r} \right) + \log \left(\frac{1}{r} \right).$$

该上界对任意 $r \in (0, 1)$ 均成立。因此，我们选择使右式极小化的 r 值。对 r 求导并令导数为 0：

$$\frac{\mathbb{E}[X] - 1}{1 - r} - \frac{1}{r} = 0 \implies r = \frac{1}{\mathbb{E}[X]}.$$

将此 r 代入上界表达式中，可得：

$$\begin{aligned} H(X) &\leq (\mathbb{E}[X] - 1) \log \left(\frac{1}{1 - \frac{1}{\mathbb{E}[X]}} \right) + \log \left(\frac{1}{\frac{1}{\mathbb{E}[X]}} \right) \\ &= \mathbb{E}[X] \cdot \left[- \left(1 - \frac{1}{\mathbb{E}[X]} \right) \log \left(1 - \frac{1}{\mathbb{E}[X]} \right) - \frac{1}{\mathbb{E}[X]} \log \left(\frac{1}{\mathbb{E}[X]} \right) \right] \\ &= \mathbb{E}[X] \cdot h_2 \left(\frac{1}{\mathbb{E}[X]} \right). \end{aligned}$$

由于期望 $\mathbb{E}[X]$ 有限，且二元熵函数受限于 1，因此熵被一个有限值所界定，故而是有限的。当 $D(P||Q) = 0$ 时取得上界，这意味着 $P = Q$ ，即 X 服从参数为 $r = \frac{1}{\mathbb{E}[X]}$ 的几何分布。

(3) 对于任意 $\alpha > 1$ ，令：

$$\sum_{x=1}^{\infty} \frac{1}{x^\alpha} = c_\alpha$$

由于当 $\alpha > 1$ 时级数收敛，我们知道存在这样一个有限的 c_α 。对于 $x \in \mathbb{N}$ ，选取分布 Q 为 $Q(x) = \frac{1}{c_\alpha x^\alpha}$ 。我们可以将 P 与 Q 之间的 KL 散度展开为：

$$D(P||Q) = -H(X) + \mathbb{E} \left[\log \frac{1}{Q(X)} \right]$$

利用 KL 散度的非负性，可得：

$$H(X) \leq \mathbb{E} [\log(c_\alpha X^\alpha)] = \log c_\alpha + \alpha \mathbb{E} [\log X]$$

由于 c_α 、 α 和 $\mathbb{E}[\log X]$ 均为有限值，因此熵被一个有限值所界定，故也是有限的。 $\square$

例 6.10

我们将证明：如果 (Y, Y') 是独立同分布的离散随机变量，则：

$$H(Y - Y') - H(Y) \leq 2(H(Y' + Y) - H(Y))$$

我们将按照以下步骤证明该不等式。

(a) **互信息的预处理不等式 (Data Processing Inequality for Mutual Information)** 设 X_1, X_2 为离散随机变量。对于某些函数 $F(\cdot), G(\cdot)$ ，令 $Y_1 = F(X_1)$ 且 $Y_2 = G(X_2)$ 。证明：

$$I(X_1; X_2) \geq I(Y_1; Y_2)$$

(b) **次模性 (Submodularity)** 假设存在函数 F, G, R 使得 $X_0 = F(X_1) = G(X_2)$ 且 $X_{12} = R(X_1, X_2)$ ，其中 X_1, X_2 为离散随机变量。利用前一部分证明：

$$H(X_{12}) + H(X_0) \leq H(X_1) + H(X_2)$$

(c) **鲁扎三角形不等式 (Ruzsa Triangle Inequality)** 设 X, Y, Z 为独立的离散随机变量。利用 (b) 部分证明：

$$1. H(X - Z) \leq H(X - Y) + H(Y - Z) - H(Y)$$

$$2. H(X - Z) \leq H(X + Y) + H(Y + Z) - H(Y)$$

[提示: 在 (b) 部分中使用 $X_1 = (X - Y, Y - Z)$, $X_2 = (X, Z)$, $X_{12} = (X, Y, Z)$ 且 $X_0 = X - Z$ 。]

(d) 熵的和与差 (Sum and Difference of Entropy) 利用前一部分得出结论: 对于独立同分布的 (Y, Y') 随机变量, 有:

$$H(Y - Y') - H(Y) \leq 2(H(Y' + Y) - H(Y))$$

证明. (a)

$$\begin{aligned} I(X_1; X_2) &= H(X_1) - H(X_1 | X_2) \\ &\stackrel{(i)}{=} H(X_1) - H(X_1 | X_2, Y_2) \\ &\stackrel{(ii)}{\geq} H(X_1) - H(X_1 | Y_2) \\ &= I(Y_2; X_1) \\ &= H(Y_2) - H(Y_2 | X_1) \\ &\stackrel{(iii)}{=} H(Y_2) - H(Y_2 | X_1, Y_1) \\ &\stackrel{(iv)}{\geq} H(Y_2) - H(Y_2 | Y_1) \\ &= I(Y_1; Y_2) \end{aligned}$$

其中:

(i) 来自 $Y_2 = G(X_2)$ 的事实。

(ii) 来自“条件化降低熵”。

(iii) 来自 $Y_1 = F(X_1)$ 的事实。

(iv) 来自“条件化降低熵”。

(b) 显而易见 $H(X_{12}) \leq H(X_1, X_2)$, 因此 $H(X_1) + H(X_2) - H(X_{12}) \geq H(X_1) + H(X_2) - H(X_1, X_2) = I(X_1; X_2)$ 。结合上述 (a) 部分 $I(X_1; X_2) \geq I(Y_1; Y_2) = I(X_0; X_0) = H(X_0)$, 即完成了证明。

(c) 1. 令 $X_1 = (X - Y, Y - Z)$, $X_2 = (X, Z)$, $X_0 = X - Z$ 且 $X_{12} = (X, Y, Z)$ 。因此对于函数 F, G, R , 我们有 $X_0 = F(X_1) = G(X_2)$ 且 $X_{12} = R(X_1, X_2)$ 。利用上述 (b) 部分:

$$H(X, Y, Z) + H(X - Z) \leq H(X - Y, Y - Z) + H(X, Z)$$

根据独立性展开, 并利用条件化降低熵整理可得:

$$H(X - Z) \leq H(X - Y) + H(Y - Z) - H(Y)$$

2. 将 Y 替换为 $-Y$, 并注意到 $H(Y) = H(-Y)$, 即可得到该结果。

(d) 将 (c)-2 用于独立同分布的 X, Y, Z , 并利用 $H(Y) = H(X)$ 以及 $H(X + Y) = H(Y + Z)$, 得到:

$$H(X - Z) + H(X) \leq 2H(X + Y)$$

现在替换 $X = Y'$ 且 $Z = Y''$, 即可得到:

$$H(Y' - Y'') - H(Y) \leq 2(H(Y' + Y) - H(Y)). \quad \square$$

第7讲 渐近均分性

在信息论中,与大数定律类似的是渐近均分性 (AEP),它是弱大数定律的直接结果。大数定律针对独立同分布 (i.i.d.) 随机变量,当 n 很大时, $\frac{1}{n} \sum_{i=1}^n X_i$ 近似于期望值 EX 。渐近均分性表明 $\frac{1}{n} \log \frac{1}{p(X_1, X_2, \dots, X_n)}$ 近似于熵 H , 其中 $X_1, X_2, \dots, X_n$ 为 i.i.d. 随机变量, $p(X_1, X_2, \dots, X_n)$ 是观察序列 $X_1, X_2, \dots, X_n$ 出现的概率。因而, 当 n 很大时, 一个观察序列出现的概率 $p(X_1, X_2, \dots, X_n)$ 近似等于 2^{-nH} 。

这促使我们将全体序列组成的集合划分成两个子集, 其一是典型集, 其中样本熵近似于真实熵; 其二是非典型集, 包含其余的序列。我们将主要关注典型集, 这是因为任何基于典型序列的性质都是以高概率成立的, 并且决定着大样本的平均行为。

首先举个例子。设随机变量 $X \in \{0, 1\}$ 的概率密度函数为 $p(1) = p$ 和 $p(0) = q$ 。若 $X_1, X_2, \dots, X_n$ 为 i.i.d., 且服从 $p(x)$, 则序列 $x_1, x_2, \dots, x_n$ 出现的概率为 $\prod_{i=1}^n p(x_i)$ 。比如, 序列 $(1, 0, 1, 1, 0, 1)$ 出现的概率是 $p^{\sum x_i} q^{n - \sum x_i} = p^4 q^2$ 。很显然, 并非所有长度为 n 的 2^n 个序列都具有相同的概率。

然而, 我们能够预测出实际观测到的序列的概率, 即可以求出观测结果 $X_1, X_2, \dots, X_n$ 的概率

$$p(X_1, X_2, \dots, X_n),$$

其中 $X_1, X_2, \dots, X_n$ 为 i.i.d. $\sim p(x)$ 。这是一个自引用的问题, 但仍然是可以明确定义的。显然, 我们是在寻求服从同一概率分布的事件的概率, 而结论是 $p(X_1, X_2, \dots, X_n)$ 将以高的概率接近于 2^{-nH} 。

对此, 概括为“几乎一切事件都令人同等的意外。”换言之, 当 $X_1, X_2, \dots, X_n$ 为 i.i.d. $\sim p(x)$ 则

$$Pr\{(X_1, X_2, \dots, X_n) : p(X_1, X_2, \dots, X_n) = 2^{-n(H \pm \epsilon)}\} \approx 1$$

在这个例子中, $p(X_1, X_2, \dots, X_n) = p^{\sum X_i} q^{n - \sum X_i}$ 可以简单地说序列中 1 出现的个数近似等于 np (以很高的概率), 且所有这样的序列 (粗略地) 有相同的概率 $2^{-nH(p)}$ 。

7.1 大数定律

定义 7.1: 随机变量的收敛

设 $X_1, X_2, \dots$ 为一随机变量序列。下面我们考虑序列 $\{X_n\}$ 收敛到 X 的三种不同方式:

1. 如果对任意的 $\epsilon > 0$, $P\{|X_n - X| > \epsilon\} \rightarrow 0$, 则称为依概率收敛。
2. 如果 $E(X_n - X)^2 \rightarrow 0$, 则称为均方收敛。
3. 如果 $P\{\lim_{n \rightarrow \infty} X_n = X\} = 1$, 则称为概率 1 (或称几乎处处) 收敛。

引理 7.2: 马尔可夫不等式

对任意非负随机变量 X 以及任意的 $t > 0$, 我们有

$$P\{X \geq t\} \leq \frac{EX}{t}. \quad (7.1)$$

证明. 如果 X 有分布 $F(x)$,

$$EX = \int_0^\infty x dF = \int_0^t x dF + \int_t^\infty x dF \geq \int_t^\infty x dF \geq \int_t^\infty t dF = tP\{X \geq t\}.$$

从而我们有 (7.1). □

引理 7.3: 切比雪夫不等式

设随机变量 Y 的均值和方差分别为 μ 和 σ^2 . 则有对任意 $\epsilon > 0$,

$$P\{|Y - \mu| > \epsilon\} \leq \frac{\sigma^2}{\epsilon^2}.$$

证明. 在马尔可夫不等式中我们取 $X = (Y - \mu)^2$, 可得

$$\begin{aligned} P\{(Y - \mu)^2 > \epsilon^2\} &\leq \frac{P\{(Y - \mu)^2 \geq \epsilon^2\}}{E(Y - \mu)^2} \\ &= \frac{E(Y - \mu)^2}{\epsilon^2} \\ &= \frac{\sigma^2}{\epsilon^2}, \end{aligned}$$

又注意到 $P\{(Y - \mu)^2 > \epsilon^2\} = P\{|Y - \mu| > \epsilon\}$, 我们有

$$P\{|Y - \mu| > \epsilon\} \leq \frac{\sigma^2}{\epsilon^2}. \quad \square$$

定理 7.4: 弱大数定律

设 $Z_1, Z_2, \dots, Z_n$ 为 i.i.d. 随机变量序列, 其均值和方差分别为 μ 和 σ^2 , 令 $\bar{Z}_n = \frac{1}{n} \sum_{i=1}^n Z_i$ 为样本均值. 证明:

$$P\{|\bar{Z}_n - \mu| > \epsilon\} \leq \frac{\sigma^2}{n\epsilon^2}.$$

因此, 当 $n \rightarrow \infty$ 时, $P\{|\bar{Z}_n - \mu| > \epsilon\} \rightarrow 0$. 这就是著名的弱大数定律.

证明. 在切比雪夫不等式中取 $Y = \bar{Z}_n$, 且注意到 $E\bar{Z}_n = \mu$ 以及 $\text{Var}(\bar{Z}_n) = \frac{\sigma^2}{n}$ ($\bar{Z}_n$ 是 n 个独立同分布的随机变量 $\frac{Z_i}{n}$, 每个的方差为 $\frac{\sigma^2}{n^2}$), 我们有

$$P\{|\bar{Z}_n - \mu| > \epsilon\} \leq \frac{\sigma^2}{n\epsilon^2}.$$

从而得证. □

7.2 渐近均分性定理**定理 7.5: AEP**

若 $X_1, X_2, \dots, X_n$ 为 i.i.d. $\sim p(x)$, 则

$$-\frac{1}{n} \log p(X_1, X_2, \dots, X_n) \rightarrow H(X) \text{ 依概率.}$$

证明. 独立随机变量的函数依然是独立随机变量. 因此, 由于 X_i 是 i.i.d., 从而 $\log p(X_i)$ 也是 i.i.d. 的. 从而由大数定律,

$$\begin{aligned} -\frac{1}{n} \log p(X_1, X_2, \dots, X_n) &= -\frac{1}{n} \sum_i \log p(X_i) \\ &\rightarrow -E \log p(X) \text{ 依概率} \\ &= H(X). \end{aligned}$$

这就证明了该定理. □

定义 7.6

关于 $p(x)$ 的典型集 $A_\epsilon^{(n)}$ 是序列 $(x_1, x_2, \dots, x_n) \in \mathcal{X}^n$ 的集合, 且满足性质:

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \dots, x_n) \leq 2^{-n(H(X)-\epsilon)}.$$

作为渐近均分性的一个推论, 可以证明典型集 $A_\epsilon^{(n)}$ 有如下性质:

定理 7.7

1. 如果 $(x_1, x_2, \dots, x_n) \in A_\epsilon^{(n)}$, 则 $H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \dots) \leq H(X) + \epsilon$.
2. 当 n 充分大时, $P\{A_\epsilon^{(n)}\} > 1 - \epsilon$.
3. $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$, 其中 $|A|$ 表示集合 A 中的元素个数.
4. 当 n 充分大时, $|A_\epsilon^{(n)}| \geq (1 - \epsilon)2^{n(H(X)-\epsilon)}$.

证明. 性质 (1) 的证明可直接由 $A_\epsilon^{(n)}$ 的定义, 利用对数函数的单调性得到. 第二个性质由定理 7.5 可以得到. 为证明性质 (3), 我们有

$$\begin{aligned} 1 &= \sum_{x \in \mathcal{X}^n} p(x) \\ &\geq \sum_{x \in A_\epsilon^{(n)}} p(x) \\ &\geq \sum_{x \in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} \\ &= 2^{-n(H(X)+\epsilon)} |A_\epsilon^{(n)}| \end{aligned}$$

因此 $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$. □

证明. 为证明 (4), 我们注意到对于充分大的 n , $P\{A_\epsilon^{(n)}\} > 1 - \epsilon$, 从而

$$\begin{aligned} 1 - \epsilon &< P\{A_\epsilon^{(n)}\} \\ &\leq \sum_{x \in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} \\ &= 2^{-n(H(X)-\epsilon)} |A_\epsilon^{(n)}|, \end{aligned}$$

其中第二个不等式由典型集的定义可得. 从而, $|A_\epsilon^{(n)}| \geq (1 - \epsilon)2^{n(H(X)-\epsilon)}$, 从而定理得证. □

7.3 AEP 的推论: 数据压缩

设 $X_1, X_2, \dots, X_n$ 为 i.i.d $\sim p(x)$ 的随机变量. 我们将 $\mathcal{X}^n$ 中的所有序列划分成两个集合: 典型集 $A_\epsilon^{(n)}$ 及其补集. 我们现在要对 $\mathcal{X}^n$ 中的元素进行 01 编码, 我们希望能够有较短的平均码字长度.

将每个集合中的所有元素按某种顺序 (比如字典序) 排列. 然后给集合中的序列指定下标可以表示 $A_\epsilon^{(n)}$ 中的每个序列. 由于 $A_\epsilon^{(n)}$ 中序列个数 $\leq 2^{n(H+\epsilon)}$, 则表示它们需要的比特数不超过 $n(H+\epsilon) + 1$ (加 1 是因为 $n(H+\epsilon)$ 可能是非整数). 在所有这些序列前加上 0, 于是可知表示 $A_\epsilon^{(n)}$ 需要的比特数不超过 $n(H+\epsilon) + 2$. 类似地, 对不属于 $A_\epsilon^{(n)}$ 中的序列, 我们表示它们需要的比特数不超过 $n \log |\mathcal{X}| + 1$, 再在这些序列前加 1. 这样我们就获得了关于所有序列的一个编码方案.

注意, 上述编码方案有如下特征:

- 编码是 1-1 的, 且可以很方便地译码. 起始位作为标识位, 标明紧随码字的长度.
- 没有考虑非典型集的元素实际上少于 $\mathcal{X}^n$ 中元素的个数. 但这足以产生一个有效的描述. (后面我们会看到, 这是由于非典型集占的比例很小)
- 典型序列具有较短的描述长度.

下面用记号 x^n 表示序列 $x_1, x_2, \dots, x_n$. 设 $l(x^n)$ 表示相应于 x^n 的码字长度. 若 n 充分大, 使得 $P\{A_\epsilon^{(n)}\} \geq 1 - \epsilon$, 于是, 码字长度的数学期望为

$$\begin{aligned}
 E(l(X^n)) &= \sum_{x^n} p(x^n) l(x^n) \\
 &= \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) l(x^n) + \sum_{x^n \in (A_\epsilon^{(n)})^c} p(x^n) l(x^n) \\
 &\leq \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) (n(H + \epsilon) + 2) + \sum_{x^n \in (A_\epsilon^{(n)})^c} p(x^n) (n \log |\mathcal{X}| + 2) \\
 &= P\{A_\epsilon^{(n)}\} (n(H + \epsilon) + 2) + P\{(A_\epsilon^{(n)})^c\} (n \log |\mathcal{X}| + 2) \\
 &\leq n(H + \epsilon) + \epsilon n (\log |\mathcal{X}|) + 2\epsilon + 2 \\
 &= n(H + \epsilon').
 \end{aligned}$$

其中 $\epsilon' = \epsilon + \epsilon \log |\mathcal{X}| + \frac{2\epsilon+2}{n}$, 适当选取 ϵ 和 n 时, ϵ' 可以任意小. 至此, 我们证明了如下的定理.

定理 7.8

设 X^n 为服从 $p(x)$ 的 i.i.d. 序列, $\epsilon > 0$, 则存在一个编码将长度为 n 的序列 x^n 映射为比特串, 使得映射是 1-1 的 (因而可逆), 且对于充分大的 n , 有

$$E\left[\frac{1}{n} l(X^n)\right] \leq H(X) + \epsilon.$$

上述定理说明, 在平均意义下, 我们可以用大约 $nH(X)$ 比特来表示序列 X^n .

7.4 高概率集与典型集

定义 7.9

对每个 $n = 1, 2, \dots$, 设 $B_\delta^{(n)} \subset \mathcal{X}^n$ 为满足如下条件的元素个数最小的集合:

$$P\{B_\delta^{(n)}\} \geq 1 - \delta.$$

定理 7.10

设 $X_1, X_2, \dots, X_n$ 为服从 $p(x)$ 的 i.i.d. 序列. 对于 $\delta < \frac{1}{2}$ 以及任意的 $\delta' > 0$, 如果 $P\{B_\delta^{(n)}\} > 1 - \delta$, 则对于充分大的 n , 有

$$\frac{1}{n} \log |B_\delta^{(n)}| > H - \delta'.$$

证明. 设 $A_\varepsilon^{(n)}$ 为一个典型集, 则 $P(A_\varepsilon^{(n)}) > 1 - \varepsilon$, 从而 $P(A_\varepsilon^{(n)} \cap B_\delta^{(n)}) \geq 1 - \varepsilon - \delta$. 我们有

$$\begin{aligned}
 1 - \varepsilon - \delta &\leq P(A_\varepsilon^{(n)} \cap B_\delta^{(n)}) \\
 &= \sum_{A_\varepsilon^{(n)} \cap B_\delta^{(n)}} p(x^n) \\
 &\leq \sum_{A_\varepsilon^{(n)} \cap B_\delta^{(n)}} 2^{-n(H-\varepsilon)} \\
 &= |A_\varepsilon^{(n)} \cap B_\delta^{(n)}| 2^{-n(H-\varepsilon)} \\
 &\leq |B_\delta^{(n)}| 2^{-n(H-\varepsilon)}.
 \end{aligned}$$

前式两边取对数, 我们有

$$\frac{1}{n} \log |B_\delta^{(n)}| \geq (H - \varepsilon) + \frac{1}{n} \log(1 - \varepsilon - \delta).$$

取 ε 充分小, 以及 n 充分大, 使得 $\varepsilon - \frac{1}{n} \log(1 - \varepsilon - \delta) < \delta'$, 从而得证. $\square$

下面引进一个新记号以表示在一阶指数意义下的相等概念.

定义 7.11: 记

$a_n \doteq b_n$ 表示

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0.$$

于是, 我们有: 如果 $\delta_n \rightarrow 0$ 且 $\varepsilon_n \rightarrow 0$, 则

$$|B_{\delta_n}^{(n)}| \doteq |A_{\varepsilon_n}^{(n)}| \doteq 2^{nH}.$$

例 7.12

为说明 $A_\varepsilon^{(n)}$ 与 $B_\delta^{(n)}$ 的区别, 考虑一个 Bernoulli 序列 $X_1, X_2, \dots, X_n$, 其参数 $p = 0.9$. 此时, 典型序列元素中 1 所占的比例近似等于 0.9. 然而, 这并不包括全部都是 1 的序列, 虽然其出现的概率最大. 集合 $B_\delta^{(n)}$ 包含所有很可能出现的序列, 因而包括全部为 1 的序列.

第 8 讲 熵率

8.1 马尔可夫链

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

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

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

定义 8.1

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 $x_1, x_2, \dots, x_n \in \mathcal{X}$, 均满足

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

则称该随机过程是平稳的.

定义 8.2: 如

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

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

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

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

定义 8.3

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

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 8.4

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.

下面我们用归纳法对 m 归纳证明

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

假设上式对 $m-1$ 成立, 下面我们来考虑 m 的情形.

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

从而我们知该马尔可夫链是平稳的. □

若有限状态马尔可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 $n \rightarrow \infty$ 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 8.5

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

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

8.2 熵率

定义 8.6

当如下极限存在时, 随机过程 $\{X_i\}$ 的熵率定义为

$$H(X) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, X_2, \dots, X_n). \quad (8.1)$$

例 8.7: 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^n 个序列, 并且都等可能出现. 因此, $H(X_1, X_2, \dots, X_n) = \log m^n$, 从而熵率为 $H(\mathcal{X}) = \log m$ 比特/字符.

例 8.8

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

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

例 8.9

若 $X_1, X_2, \dots, X_n$ 独立但不同分布, 则

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

我们可以选择 $X_1, X_2, \dots$ 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:

$$H'(\mathcal{X}) = \lim_{n \rightarrow \infty} H(X_n | X_{n-1}, X_{n-2}, \dots, X_1) \quad (8.2)$$

定理 8.10

对于平稳随机过程, 式 (8.1) 和 (8.2) 中的极限均存在且相等:

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

定理 8.11

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

证明.

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

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 $H(X_n|X_{n-1}, \dots, X_1)$ 是非负且递减的数列, 故知极限 $H'(\mathcal{X})$ 存在. $\square$

引理 8.12

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

定理 8.10 的证明. 由链式法则, 我们有

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

于是由引理 8.12, 我们知

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

研究随机过程熵率的重要意义体现在平稳遍历过程的 AEP. 我们可以证明更一般的 AEP, 即对任意的遍历过程,

$$-\frac{1}{n} \log p(X_1, X_2, \dots, X_n) \rightarrow H(\mathcal{X}) \quad (4-24)$$

以概率 1 收敛. 由此, 上一讲中的所有定理可容易地推广到一般的平稳遍历过程. 上一讲中的 i. i. d. 情形类似, 我们可定义典型集, 并采用同样的讨论方法, 可以证明典型集的概率近似为 1, 且大约有 $2^{nH(\mathcal{X})}$ 个长度为 n 的典型序列, 其每个序列出现的概率大约为 $2^{-nH(\mathcal{X})}$. 所以, 大约使用 $nH(\mathcal{X})$ 比特可表示长度为 n 的典型序列. 这体现出熵率可以表征平稳遍历过程的平均描述长度的重要意义.

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

$$H(\mathcal{X}) = H'(\mathcal{X}) = \lim_{n \rightarrow \infty} H(X_n|X_{n-1}, \dots, X_1) = \lim_{n \rightarrow \infty} H(X_n|X_{n-1}) = H(X_2|X_1).$$

其中第三个等式是因为

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

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 8.13

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.

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

□

例 8.14

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 8.15

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 $n \rightarrow \infty$ 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 (8.1) 和 (8.2) 中给出的 $H(\mathcal{X})$ 。

8.3 例子：加权图上随机游动的熵率

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 $\{1, 2, \dots, m\}$ ，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 $\{X_n\}$ ， $X_n \in \{1, 2, \dots, m\}$ 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

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

例 8.16

假定一个“王”在 8×8 的 (国际象棋) 棋盘上做随机游动. “王”这个棋子在棋盘内部时可有 8 个移位, 在边缘时有 5 个移位, 在角落时有 3 个移位. 据此及前面所述结论可知, 平稳概率分别是 $\frac{8}{420}$, $\frac{5}{420}$ 和 $\frac{3}{420}$. 从而, 熵率为 $0.92 \log 8$. 因子 0.92 是由于边缘效应产生的; 如果这个棋子在无限的棋盘上游动, 则可得其熵率为 $\log 8$.

类似地, 可以求得“车” (它的熵率为 $\log 14$ 比特, 因为“车”总是有 14 个可能的移位)、“相”、“王后”的熵率. “王后”的可能移位恰是“车”和“相”的可能移位的合成, 那么“王后”比起“车”和“相”, 拥有更多还是更少的自由度呢?

注记 8.17

易知图上的平稳随机游动是时间可逆的, 即是说任何状态序列向前和向后的概率是相等的:

$$\begin{aligned}
 &\Pr(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) \\
 &= \Pr(X_n = x_1, X_{n-1} = x_2, \cdots, X_1 = x_n)
 \end{aligned}$$

出乎意料的是, 反命题亦成立, 即任何时间可逆的马尔可夫链均可以表示为某个无向加权图上的随机游动.

8.4 马尔可夫链的函数

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. 一个自然的问题是: 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. 由于 $H(Y_n|Y_{n-1}, \cdots, Y_1)$ 收敛速度可能很慢, 很难知道是否已经接近极限.

引理 8.18

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

证明. 对 $k = 1, 2, \dots$, 有

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

两边取极限, 我们知

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

从而引理得证. □

引理 8.19

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

证明. 首先注意到

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

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

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

注意到上面的正项级数收敛, 所以其通项趋于 0, 于是

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

引理得证. □

定理 8.20

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

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

且

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

一般地,我们也可以考虑 X_i 的随机函数 Y_i (即非确定性的函数). 给定马尔可夫过程 $X_1, X_2, \dots, X_n$, 由此定义新过程 $Y_1, Y_2, \dots, Y_n$, 其中每个 Y_i 服从 $p(y_i|x_i)$, 且条件独立于其他所有的 $X_j, j \neq i$, 即

$$p(x^n, y^n) = p(x_1) \prod_{i=1}^{n-1} p(x_{i+1}|x_i) \prod_{i=1}^n p(y_i|x_i)$$

这样的过程称为隐马尔可夫模型 (HMM), 它已广泛应用于语音识别、手写体识别等等。以上对马尔可夫链的函数的讨论同样适用于隐马尔可夫模型。通过对隐含的马尔可夫状态加入条件, 我们可以估计出隐马尔可夫模型熵率的下界.

第9讲 信源编码

9.1 有关编码的几个例子

定义 9.1

关于随机变量 X 的信源编码 C 是从 X 的取值空间 $\mathcal{X}$ 到 $\mathcal{D}^*$ 的一个映射, 其中 $\mathcal{D}^*$ 表示 D 元字母表 $\mathcal{D}$ 上有限长度的字符串所构成的集合. 用 $C(x)$ 表示 x 的码字并用 $l(x)$ 表示 $C(x)$ 的长度.

定义 9.2

设随机变量 X 的概率密度函数为 $p(x)$, 定义信源编码 $C(x)$ 的期望长度 $L(C)$ 为

$$L(C) = \sum_{x \in \mathcal{X}} p(x) l(x).$$

其中 $l(x)$ 表示对应于 x 的码字长度.

例 9.3

设随机变量 X 的分布及其码字分配如下:

$$P\{X = 1\} = \frac{1}{2} \text{ 码字 } C(1) = 0$$

$$P\{X = 2\} = \frac{1}{4} \text{ 码字 } C(2) = 10$$

$$P\{X = 3\} = \frac{1}{8} \text{ 码字 } C(3) = 110$$

$$P\{X = 4\} = \frac{1}{8} \text{ 码字 } C(4) = 111$$

易知 X 的熵 $H(X)$ 为 1.75 比特, 而期望长度 $L(C) = El(X)$ 也是 1.75 比特.

例 9.4

设随机变量 X 的分布及其码字分配如下:

$$P\{X = 1\} = \frac{1}{3} \text{ 码字 } C(1) = 0$$

$$P\{X = 2\} = \frac{1}{3} \text{ 码字 } C(2) = 10$$

$$P\{X = 3\} = \frac{1}{3} \text{ 码字 } C(3) = 11$$

易知 X 的熵 $H(X)$ 为 $\log 3 = 1.58$ 比特, 而期望长度 $L(C) = El(X)$ 是 1.67 比特, 此时 $El(X) > H(X)$.

定义 9.5

如果编码将 X 的取值空间中的每个元素映射成 $\mathcal{D}^*$ 中不同的字符串, 即

$$x \neq x' \Rightarrow C(x) \neq C(x')$$

则称这个编码是非奇异的.

定义 9.6

编码 C 的扩展 C^* 是从 $\mathcal{X}$ 上的有限长字符串到 $\mathcal{D}$ 上的有限长字符串的映射, 定义为

$$C(x_1x_2\cdots x_n) = C(x_1)C(x_2)\cdots C(x_n).$$

例 9.7

如果 $C(x_1) = 00$ 且 $C(x_2) = 11$, 则 $C(x_1x_2) = 0011$.

9.2 唯一可译码**定义 9.8**

如果一个编码的扩展编码是非奇异的, 则称该编码是唯一可译的.

定理 9.9

若非奇异码 C 中的所有码字 w_i 都具有相同的长度, 则 C 是唯一可译码。

证明. 设 l 为公共码字长度. 若某个序列 $\mathbf{t} \in C^*$ 可以分解为两种码字拼接形式:

$$u_1u_2\cdots u_m = v_1v_2\cdots v_n$$

其中每个 $u_i, v_j \in C$, 则序列 $\mathbf{t}$ 的总长度满足:

$$l \cdot m = |\mathbf{t}| = l \cdot n$$

由此可得 $m = n$ 。

由于 u_1 和 v_1 均由序列 $\mathbf{t}$ 的前 l 个符号 (字符) 组成, 因此必定有 $u_1 = v_1$ 。同理, 对所有的 $i = 1, 2, \dots, m$, 均有 $u_i = v_i$ 。□

若码 C 中的所有码字都具有相同的长度, 我们称其为**区块码**。

例 9.10

由下式给出的二进制编码 C :

$$s_1 \mapsto w_1 = 0, \quad s_2 \mapsto w_2 = 01, \quad s_3 \mapsto w_3 = 011$$

具有变长特征, 但仍然是唯一可译码。在序列 $\mathbf{t}$ 中, 每个符号 0 都标志着一个码字 w_i 的开始, 而下标 i 等于 1 加上其后跟随的 1 的个数。例如,

$$t = 001011010011 = 0.01.011.01.0.011 \Rightarrow \mathbf{s} = s_1s_2s_3s_2s_1s_3.$$

实际上, 这里我们把符号 $0 \in T$ 当作标点符号来使用。

下面我们介绍由 Sardinas 和 Patterson 于 1957 年设计出的一种判断唯一可译码的测试方法。

我们设编码的字母表为 T 。我们用 T^+ 表示所有非空字符串构成的集合:

$$T^+ = \bigcup_{n>0} T^n.$$

我们利用归纳法定义一个由非空单词集合 $C_0, C_1, \dots$ 构成的序列, 使得对所有 n , 都有 $C_n \subseteq T^+$ 。具体地, 我们定义 $C_0 = C$, 并且对于每个 $n \geq 1$, 定义

$$C_n = \{w \in T^+ \mid uw = v, \text{ 其中 } u \in C, v \in C_{n-1}, \text{ 或者 } u \in C_{n-1}, v \in C\}. \quad (9.1)$$

然后定义

$$C_\infty = \bigcup_{n=1}^{\infty} C_n.$$

这个定义乍看起来可能有些令人望而生畏, 但如果我们一步一步来看, 就会变得更加清楚: 我们从 $C_0 = C$ 开始, 然后利用前一项 C_{n-1} 构造每一个 C_n ($n \geq 1$), 最后取

$$C_\infty = C_1 \cup C_2 \cup \dots.$$

注意, 当 $n = 1$ 时, C_n 的定义可以进一步简化: 由于 $C_{n-1} = C_0 = C$, 所以在 (9.1) 中的两个条件实际上是相同的, 即

$$C_1 = \{w \in T^+ \mid uw = v, \text{ 其中 } u, v \in C\}.$$

我们还注意到, 如果 $C_{n-1} = \emptyset$, 那么 $C_n = \emptyset$ 。因此不断迭代可以得到

$$C_{n+1} = C_{n+2} = \dots = \emptyset.$$

例 9.11

设 $C = \{0, 01, 011\}$ 。则根据 (9.1) 可得 $C_1 = \{1, 11\}$: 我们有 $1 \in C_1$, 因为 $01.1 = 011$ (其中 $01, 011 \in C = C_0$); 且 $11 \in C_1$, 因为 $0.11 = 011$ (其中 $0, 011 \in C = C_0$)。在下一步中, 当 $n = 2$ 时, 经检验可知不存在 $w \in T^+$ 满足 $uw = v$, 其中 $u \in C, v \in C_1$ (或反过来)。因此 $C_2 = \emptyset$, 从而对所有 $n \geq 2$ 都有 $C_n = \emptyset$, 于是 $C_\infty = C_1 = \{1, 11\}$ 。

命题 9.12

如果 C 的码字长度为 $l_1, \dots, l_q$, 且对于某个 n , 有 $w \in C_n$, 那么 $|w| \leq l = \max(l_1, \dots, l_q)$ 。从而每个 C_n 都是有限集, 且集合序列 $C_0, C_1, \dots$ 是最终周期的。

证明. 对 n 使用数学归纳法. 若 $n = 0$, 则 $C_n = C$ 。因此 $|w| \leq l$ 。若 $n > 0$, 则 $uw = v$ (其中 $v \in C_{n-1}$ 或 C) , 因此根据归纳法或 l 的定义可知 $|w| \leq |v| \leq l$ 。长度满足 $|w| \leq l$ 的非空 r 进制词只有 $N = r + r^2 + \dots + r^l = r(r^l - 1)/(r - 1)$ 个, 所以对每个 n , 都有 $|C_n| \leq N$ 。此类字 w 的子集最多只有 2^N 个不同的非空集合, 因此在集合序列 $C_0, \dots, C_{2^N}$ 中必然存在重复, 即存在 $i < j \leq 2^N$ 使得 $C_j = C_i$ 。根据等式 (1.3), 每一个 C_n 仅取决于 C 和 C_{n-1} 。因此对所有 $k \geq 0$, 都有 $C_{j+k} = C_{i+k}$; 从而每个 C_n 都等于 C_0 或 C_1 或 ... 或 C_{j-1} 。于是 $C_\infty = C_0 \cup C_1 \cup \dots \cup C_{j-1}$ 。因此, 一旦我们在连续集合序列 $C_0, C_1, \dots$ 中找到重复项, 就构造出了 C_∞ 的所有元素。□

定理 9.13: Sardinas-Patterson 定理

编码 C 是唯一可译码, 当且仅当集合 C 和 C_∞ 的交为空集。

例 9.14

如果 $C = \{0, 01, 011\}$ (如例 1.8 和例 1.9 所示), 那么 $C_\infty = \{1, 11\}$, 它与 C 互不相交。正如我们之前已经看到的, 由定理 9.13 可知 C 是唯一可译码。

例 9.15

设 C 为三进制编码 $\{01, 1, 2, 210\}$. 我们求得 $C_1 = \{10\}$, $C_2 = \{0\}$, 以及 $C_3 = \{1\}$, 因此 $1 \in C \cap C_\infty$, 从而 C 不是唯一可译码 (没有必要再计算 $n > 3$ 时的 C_n). 为了找到一个非唯一译码的例子, 注意到 $10 \in C_1$ 是因为 $2 \in C$ 且 $2.10 = 210 \in C$; 接着 $0 \in C_2$ 是因为 $1 \in C$ 且 $1.0 = 10 \in C_1$; 然后 $1 \in C_3$ 是因为 $0 \in C_2$ 且 $0.1 = 01 \in C$. 将这些等式结合起来, 我们得到:

$$210.1 = 2.10.1 = 2.1.0.1 = 2.1.01$$

因此, 码序列 $\mathbf{t} = 2101$ 既可以被解码为 210.1 , 也可以被解码为 $2.1.01$.

由于 Sardinas-Patterson 定理的证明相当长. 在此处, 我们将给出两个典型的推导过程来说明其中涉及的思想.

($\Rightarrow$) 假设 $C \cap C_\infty \neq \emptyset$, 不妨设 $w \in C \cap C_2$; 因此 $uw = v$, 其中 $u \in C$ 且 $v \in C_1$ (反之亦然, 为了简单起见, 我们假设第一种情况成立, 第二种情况类似). 那么 $u'v = v'$, 其中 $u', v' \in C$, 因此序列 $\mathbf{t} = u'uw \in T^*$ 可以表示三个信源符号构成的序列 (因为 $u', u, w \in C$), 或者表示一个信源符号 (因为 $u'uw = u'v = v' \in C$). 因此解码不是唯一的.

($\Leftarrow$) 假设我们存在形式为 $\mathbf{t} = u_1u_2 = v_1v_2$ 的非唯一解码实例, 其中 $u_1, u_2, v_1, v_2 \in C$. 我们不可能有 $|u_1| = |v_1|$, 因为这会导致 $u_1 = v_1$, 进而得到 $u_2 = v_2$. 因此, 如有必要可以通过重新编号, 假设 $|u_1| > |v_1|$, 即 $u_1 = v_1w$, 其中 $|w| > 0$. 于是 $w \in C_1$, 又因为 $wu_2 = v_2$, 所以 $u_2 \in C_2$. 因此 $u_2 \in C \cap C_\infty$, 这就说明 C 与 C_∞ 不相交.

我们已经将唯一可译性定义为所有有限码序列 $\mathbf{t}$ 都可以被唯一解码, 但我们也可以考虑更强的要求, 即对于所有码序列 (无论有限还是无限) 这一结论都成立. Even, Levenshtein 和 Riley 证明的一个定理表明, 当且仅当 $C \cap C_\infty = \emptyset$ 同时对于某个 $n \geq 1$ 有 $C_n = \emptyset$ 时, 该情况才会发生. 这些也是 C 成为有界延迟唯一可译码 (uniquely decodable with bounded delay) 的充要条件, 这意味着存在一个常数 d , 使得如果两个码序列的前 d 个符号相同, 则它们具有相同的第一个码字; 因此解码可以在延迟最多 d 个符号后开始.

9.3 即时码

在给出即时码的定义之前, 我们先来看几个例子.

例 9.16

考虑由下式给出的二进制编码 C :

$$s_1 \mapsto 0, \quad s_2 \mapsto 01, \quad s_3 \mapsto 11.$$

我们有 $C_1 = C_2 = \dots = \{1\}$, 因此 $C_\infty = \{1\}$; 由此可得 $C \cap C_\infty = \emptyset$, 根据定理编码 C 是唯一可译码. 现在假设我们接收到一个以 $t = 0111\dots$ 开头的有限消息. 尽管我们知道该消息可以被唯一译码, 但在读到这串连续 1 的末尾之前, 我们无法开始译码: 如果这块连续 1 的个数为偶数, 则 t 的分解必然为 $0.11.11.11\dots$, 译码后的消息必然以 $\mathbf{s} = s_1s_3s_3s_3\dots$ 开头; 然而, 如果 1 的个数为奇数, 则分解必然为 $01.11.11.11\dots$, 因此 $\mathbf{s} = s_2s_3s_3s_3\dots$. 在实际应用中, 这种译码延迟可能会带来困难. 我们称编码 C 是非即时的.

现在给出即时码的正式定义.

定义 9.17

如果对于码字序列 $w_{i_1}, w_{i_2}, \dots, w_{i_n}$, 无论 $\mathbf{t}$ 后续的符号是什么, 任何以该序列开头的码序列 $\mathbf{t} = w_{i_1} w_{i_2} \dots w_{i_n} \dots$ 都能被唯一解码为 $\mathbf{s} = s_{i_1} s_{i_2} \dots s_{i_n} \dots$, 则称码 C 是**即时码**.

例 9.18

考虑由以下映射给出的二进制码 $\mathcal{D}$:

$$s_1 \mapsto 0, s_2 \mapsto 10, s_3 \mapsto 11,$$

它是例 9.16 中码 C 的反向码 (reverse code). 我们可以看出它是唯一可译的, 这既可以通过 Sardinas-Patterson 定理得出, 也可以因为 C 是唯一可译的而得出. 它同时也是即时码, 这意味着我们可以一边接收消息 $\mathbf{t}$ 一边进行解码: 数字 0 表示码字 w_1 (我们将其解码为 s_1); 数字 1 则表示 $w_2 = 10$ 或 $w_3 = 11$ 的开头, 一旦知道下一个符号, 就可以立即解码为 s_2 或 s_3 . 因此, 消息 $\mathbf{t}$ 中的任何码字一旦到达就可以被立即解码, 没有任何延迟.

例 9.19

考虑二元码 $C = \{0, 01, 011, 111\}$. 我们有 $C_1 = \{1, 11\}$ 且 $C_2 = \{1, 11\}$, 因此对于所有 $n \geq 1$ 有 $C_n = \{1, 11\}$; 从而 $C_\infty = \{1, 11\}$, 它与 C 不相交, 故根据 Sardinas-Patterson 定理, C 是唯一可译的. 等待直到数字 1 的序列结束; 如果共有 k 个 1, 且满足 $k \equiv 0, 1$ 或 $2 \pmod{3}$, 则对应地 (唯一地) 解码为 $0.(111)^{k/3}$ 、 $01.(111)^{(k-1)/3}$ 或 $011.(111)^{(k-2)/3}$.

定义 9.20

若码中无任何码字是其它码字的前缀, 则称该编码为**前缀码**.

定理 9.21

编码 C 是即时码, 当且仅当它是前缀码。

证明. ($\Rightarrow$) 如果 C 不是前缀码, 设 w_i 是 w_j 的前缀, 那么以 $\mathbf{t} = w_i \dots$ 开头的代码序列在译码时既可以译为 $\mathbf{s} = s_i \dots$, 也可以译为 $\mathbf{s} = s_j \dots$, 因此 C 不是即时码。

($\Leftarrow$) 如果 C 是前缀码, 且 $\mathbf{t}$ 以 $w_i \dots$ 开头, 那么 $\mathbf{s}$ 必须以 s_i 开头, 因为没有其他码字 w_j ($j \neq i$) 是 w_i 的前缀或以 w_i 为前缀. 我们可以像这样继续下去, 在接收到 $\mathbf{t}$ 中的后续码字时依次进行译码, 因此 C 是即时码. $\square$

9.4 Sardinas-Patterson 定理的证明 *

回想一下, 如果只要对于所有的 $u_i, v_j \in C$, 当 $u_1 \dots u_l = v_1 \dots v_m$ 时, 都有 $l = m$ 且对每个 i 都有 $u_i = v_i$, 则称码 $C \subseteq T^*$ 是唯一可译的. 给定一个码 C , 对每个 $n \geq 1$, 我们定义 $C_0 = C$,

$$C_n = \{w \in T^+ \mid uw = v \text{ 其中 } u \in C, v \in C_{n-1} \text{ 或 } u \in C_{n-1}, v \in C\}$$

且 $C_\infty = \bigcup_{n=1}^{\infty} C_n$. 定理表明, 当且仅当 $C \cap C_\infty = \emptyset$ 时, C 是唯一可译的.

首先我们需要一些记号. 如果 $u, v \in T^*$ 且 u 是 v 的前缀, 即对于某个 $w \in T^*$ 有 $v = uw$, 我们记作 $u \leq v$; 我们也可以记作 $w = u^{-1}v$, 意味着 w 是通过从 v 中删除前缀 u 得到的. (注意单独的 u^{-1} 是没有定义的.) 此外, 如果 $u \neq v$, 我们就记作 $u < v$. 现在我们可以开始证明了。

首先, (T_{n-1}) 是成立的, 因为由 (S_n) 可得:

$$v_{n-1}y_{n-1} = z_{n-1}$$

其中要么 $y_{n-1} = \varepsilon$ 且 $z_{n-1} = u_{n-1}v_n$, 要么 $y_{n-1} = v_n$ 且 $z_{n-1} = u_{n-1}$; 在任何一种情况下, 因为 $u_{n-1}, v_n \in C$, 所以都有 $y_{n-1}, z_{n-1} \in C^*$ 。现在我们将证明对于 $2 \leq k \leq n-1$, (T_k) 蕴含 (T_{k-1}) 。假设 (T_k) 成立。命题 (S_k) 给出:

$$u_{k-1}v_k = v_{k-1} \quad \text{或} \quad v_{k-1}v_k = u_{k-1},$$

因此, 或者

$$v_{k-1}y_k = u_{k-1}v_ky_k = u_{k-1}z_k \quad \text{或} \quad v_{k-1}z_k = u_{k-1}v_ky_k = u_{k-1}y_k.$$

这两个断言的形式均为

$$v_{k-1}y_{k-1} = z_{k-1},$$

其中分别为

$$y_{k-1} = y_k, \quad z_{k-1} = u_{k-1}z_k \quad \text{或} \quad y_{k-1} = z_k, \quad z_{k-1} = u_{k-1}y_k$$

在这两种情况下, y_{k-1} 和 z_{k-1} 都是 C^* 中的元素 (因为 y_k 和 z_k 是, 且 $u_{k-1} \in C$), 因此 (T_{k-1}) 得证。由此可知, 每一个 (T_k) 都是成立的, 因此取 $k=1$ 得到

$$v_1y_1 = z_1 \quad \text{对于某些} \quad y_1, z_1 \in C^*.$$

然后 (S_1) 蕴含

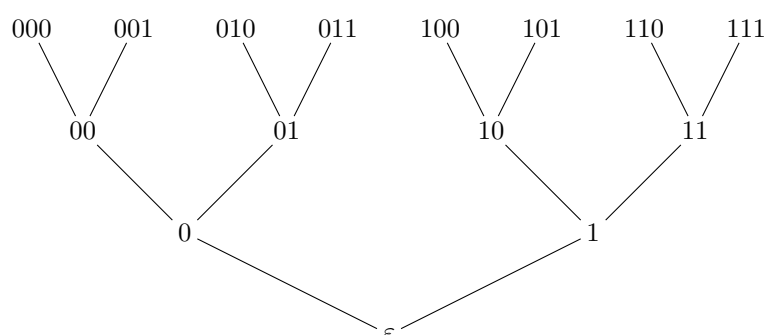
$$u'y_1 = uv_1y_1 = uz_1,$$

其中 $y_1, z_1 \in C^*$ 且 u, u' 是不同的码字。因此, 方程 $u'y_1 = uz_1$ 给出了将同一码序列分解为码字的两种不同方式, 从而 C 不是唯一可译的。

第 10 讲 Kraft 不等式与最优码

10.1 构造即时码

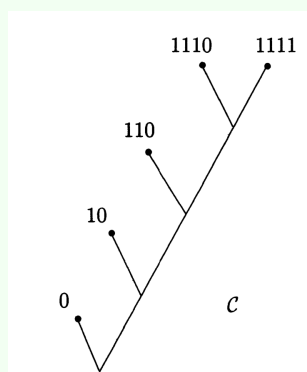
为了理解即时码的构造, 将 T 中的字构成的集合 T^* 视为一个图是很有帮助的. 图是一个由点 (称为顶点) 构成的集合, 其中某些顶点对由边相连. 在这种情况下, 顶点就是字 $w \in T^*$, 并且每个字 w 都与由在 w 末尾添加单个符号 $t_i \in T$ 所形成的 r 个字 $wt_1, \dots, wt_r$ 通过边相连. 我们可以将这个图可视化为向上生长的树, 其中空字 ε 在底部, 长度为 l 的字位于 ε 上方第 l 层; 在图论中, 这样的图被称为 r 元有根树. (树是没有圈的连这里根是 ε .) 下图展示了当 $T = \mathbb{Z}_2$ 时, 直到第 $l = 3$ 层的二叉树 T^* ,



码 C 可以被看作是树 T^* 的顶点的有限集合. 当且仅当顶点 w_i 被顶点 w_j 支配 (也就是说, 在 T^* 中存在一条从 w_i 到 w_j 的向上路径) 时, 字 w_i 才是 w_j 的前缀. 我们知 C 是即时码当且仅当 C 中没有任何顶点 w_i 被 C 中的另一个顶点 w_j ($i \neq j$) 所支配. 我们可以利用这个准则来构造即时码, 即在 T^* 中逐个选择顶点, 使得没有任何顶点支配 (或被支配于) 任何先前选择的顶点.

例 10.1

让我们为含有五个符号 $s_1, \dots, s_5$ 的信源 $\mathcal{S}$ 寻找一个二元即时码 C . 首先, 让我们尝试 $s_1 \mapsto w_1 = 0$, 因此 0 是 C 中的一个顶点. 如果 C 要成为一个前缀码, 那么 C 中的其他任何顶点都不能支配 0, 因此它们必须全都支配 1 (即其他码字必须以 1 开头). 如果我们尝试 $s_2 \mapsto w_2 = 1$, 那么无法添加更多的码字, 因为它们会支配 w_1 或 w_2 . 相反, 让我们尝试 $s_2 \mapsto w_2 = 10$. 接着 $s_3 \mapsto w_3 = 11$ 是不可行的, 因为它不允许有进一步的码字, 所以让我们尝试 $s_3 \mapsto w_3 = 110$. 继续下去, 我们发现有以下可能性: $s_4 \mapsto w_4 = 1110, s_5 \mapsto w_5 = 1111$. 这就给出了一个二元即时码 $C = \{0, 10, 110, 1110, 1111\}$, 其字长分别为 $l_i = 1, 2, 3, 4, 4$, 如下图所示. (这并不是唯一的可能方案. 例如, 二进制码 $\{00, 01, 10, 110, 111\}$ 同样是即时的.)



10.2 Kraft 不等式

定理 10.2: Kraft 不等式

对于 D 元字母表上的即时码（前缀码），码字长度 $l_1, l_2, \dots, l_m$ 必定满足不等式

$$\sum_i D^{-l_i} \leq 1$$

反之，若给定满足以上不等式的一组码字长度，则存在一个相应的即时码，其码字长度就是给定的长度。

证明. 考虑每一节点均含 D 个子节点的 D 叉树. 假定树枝代表码字的字符. 例如，源于根节点的 D 条树枝代表着码字第一个字符的 D 个可能值. 另外，每个码字均由树的一片叶子表示. 因此，始于根节点的路径可描绘出码字中的所有字符. 码字的前缀条件表明树中无一码字是其他任一码字的祖先. 因而，在这样的编码树中，每一码字都去除了其可能成为码字的所有后代.

令 l_{max} 为码字集中最长码字长度. 考虑在树中 l_{max} 层的所有节点，可知其中有些是码字，有些是码字的后代，而另外的节点不是码字，也不是码字的后代. 在树中 l_i 层的码字拥有 l_{max} 层中的 $D^{l_{max}-l_i}$ 个后代. 所有这些后代互不相交. 而且，这些集合中的总结点数必定不超过 $D^{l_{max}}$. 因此，对所有码字求和，可得

$$\sum D^{l_{max}-l_i} \leq D^{l_{max}}.$$

于是

$$\sum D^{-l_i} \leq 1,$$

这就是 Kraft 不等式.

反之，若给定任意一组满足 Kraft 不等式的码字长度 $l_1, l_2, \dots, l_m$ ，总可以构造出如下的一个编码树：将第一个深度为 l_1 的节点（依字典序）标为码字 1，同时去除其所有后代. 然后在剩余的节点中找出第一个深度为 l_2 的节点，将其标为码字 2，同时去除树中所有属于它的所有后代，等等. 按此方法继续下去，则可构造出一个码字长度为 $l_1, l_2, \dots, l_m$ 的前缀码. $\square$

定理 10.3: 推广的 Kraft 不等式

对任意构成前缀码的可数无限码字集，码字长度也满足推广的 Kraft 不等式：

$$\sum_{i=1}^{\infty} D^{-l_i} \leq 1.$$

反之，若给定任意满足推广的 Kraft 不等式的 $l_1, l_2, \dots$ ，则可构造出具有相应码长的前缀码。

证明. 我们不妨设 D 元字母表为 $\{0, 1, \dots, D-1\}$ ，第 i 个码字为 $y_1 y_2 \dots y_{l_i}$. 记 $0.y_1 y_2 \dots y_{l_i}$ 是以 D 进制表示的实值小数，即

$$0.y_1 y_2 \dots y_{l_i} = \sum_{j=1}^{l_i} y_j D^{-j}.$$

由此，这个码字对应一个区间

$$[0.y_1 y_2 \dots y_{l_i}, 0.y_1 y_2 \dots y_{l_i} + \frac{1}{D^{l_i}}),$$

这是一个实数集合, 集合中的所有数的 D 进制展开都以 $0.y_1y_2\cdots y_{l_i}$ 开始. 由前缀条件我们知所有这些区间互不相交. 因此, 它们的区间长度总和小于或等于 1. 因此我们证明了

$$\sum_{i=1}^{\infty} D^{-l_i} \leq 1.$$

反之, 和有限情形类似. 首先将长度下表重新排列, 使得 $l_1 \leq l_2 \leq \cdots$. 然后从单位区间的左端开始, 依次进行分配, 即可获得满足条件的码字集. 例如, 如果想构造一个二元编码使得 $l_i = i$, $i = 1, 2, \cdots$, 那么将区间 $[0, \frac{1}{2})$, $[\frac{1}{2}, \frac{3}{4})$, $\cdots$ 分配给字符, 使其对应码字为 $0, 10, \cdots$. $\square$

10.3 最优码

前面我们已经证明了满足前缀条件的任何一个码字集合满足 Kraft 不等式, 并且当一组码字长度集合满足 Kraft 不等式时, 存在这样的码字集, 它们的长度集合正好就是给定的长度集合. 下面我们来考虑求解前缀码的最小期望长度问题.

由前面的讨论我们知该问题等价于寻找一个前缀码, 它的码字长度集合满足 Kraft 不等式, 其码字期望长度 $L = \sum p_i l_i$ 达到最小. 这就化为了一个最优化的问题: 在所有整数 $l_1, l_2, \cdots, l_m$ 上, 最小化

$$L = \sum p_i l_i$$

其约束条件为

$$\sum D^{-l_i} \leq 1.$$

先利用微积分的知识做一个简单的分析, 以此说明达到最小值时 l^* 应具有的形式. 取消 l_i 必须是整数的限制, 并假定约束条件中的等号成立. 于是, 利用拉格朗日乘子法, 将带约束的最小化问题转化为求

$$J = \sum p_i l_i + \lambda (\sum D^{-l_i} - 1)$$

的最小化问题. 关于 l_i 求微分, 可得

$$\frac{\partial J}{\partial l_i} = p_i - \lambda D^{-l_i} \ln D.$$

令偏导数为 0, 得

$$D^{-l_i} = \frac{p_i}{\lambda \ln D}.$$

将此代入约束条件中以求得合适的 λ , 可得 $\lambda = 1/\ln D$, 因而 $p_i = D^{-l_i}$. 即最优码长为

$$l_i^* = -\log_D p_i.$$

若可以取码字长度为非整数, 则此时的期望码字长度为

$$L^* = \sum p_i l_i^* = -\sum p_i \log_D p_i = H_D(X).$$

定理 10.4

随机变量 X 的任一 D 元即时码的期望长度必定大于等于熵 $H_D(X)$, 即

$$L \geq H_D(X)$$

当且仅当 $D^{-l_i} = p_i$, 等号成立.

证明. 我们将期望长度与熵的差写成如下形式

$$\begin{aligned} L - H_D(X) &= \sum p_i l_i - \sum p_i \log_D \frac{1}{p_i} \\ &= -\sum p_i \log_D D^{-l_i} + \sum p_i \log_D p_i \end{aligned}$$

设 $r_i = D^{-l_i} / \sum_j D^{-l_j}$, $c = \sum D^{-l_i}$, 由相对熵的非负性以及 $c \leq 1$ (利用 Kraft 不等式), 可得

$$\begin{aligned} L - H_D(X) &= \sum p_i \log_D \frac{p_i}{r_i} - \log_D c \\ &= D(\mathbf{p} \parallel \mathbf{r}) + \log_D \frac{1}{c} \\ &\geq 0. \end{aligned}$$

因此, $L \geq H$, 当且仅当 $p_i = D^{-l_i}$ (即对所有的 i , $-\log_D p_i$ 为整数), 等号成立. $\square$

定义 10.5

对于某个 n , 如果概率分布的每一个概率值均等于 D^{-n} . 则称这个概率分布是 D 进制的.

因此, 当且仅当 X 的分布是 D 进制的, 上述定理等号成立.

10.4 最优码长的界

定理 10.6

设 $l_1^*, l_2^*, \dots, l_m^*$ 是关于信源分布 $\mathbf{p}$ 和一个 D 元字母表的一组最优码长, L^* 为最优码的相应期望长度 ($L^* = \sum p_i l_i^*$), 则

$$H_D(X) \leq L^* < H_D(X) + 1.$$

证明. 设 $l_i = \lceil \log_D \frac{1}{p_i} \rceil$, 则 l_i 满足 Kraft 不等式且

$$H_D(X) \leq L = \sum p_i l_i < H_D(X) + 1.$$

但由于 L^* 是最优码的期望长度, 它不大于 $L = \sum p_i l_i$. 再由定理 10.4 可知 $L^* \geq H_D$, 从而得证. $\square$

定义 L_n 为每输入字符期望码字长度, 也就是说, 如果设 $l(x_1, x_2, \dots, x_n)$ 是与 $(x_1, x_2, \dots, x_n)$ 相应的二进制码字长度 (为简便起见, 在本小节余下的部分, 假定 $D = 2$), 则

$$L_n = \frac{1}{n} p(x_1, x_2, \dots, x_n) l(x_1, x_2, \dots, x_n) = \frac{1}{n} El(X_1, X_2, \dots, X_n).$$

将前面推导的界应用于此时的编码, 有

$$H(X_1, X_2, \dots, X_n) \leq El(X_1, X_2, \dots, X_n) < H(X_1, X_2, \dots, X_n) + 1.$$

若 $X_1, X_2, \dots, X_n$ 是独立同分布的, 因此 $H(X_1, X_2, \dots, X_n) = \sum H(X_i) = nH(X)$. 上式两边同除以 n , 得

$$H(X) \leq L_n < H(X) + \frac{1}{n}.$$

因此, 通过足够大的分组长度, 可以获得一个编码, 可以使其每字符期望码长任意地接近熵.

若此随机过程不是独立同分布的, 我们亦然可以类似讨论, 我们有估计

$$H(X_1, X_2, \dots, X_n) \leq El(X_1, X_2, \dots, X_n) < H(X_1, X_2, \dots, X_n) + 1.$$

上式两边同时除以 n ，且定义 L_n 为每字符期望描述长度，可得

$$\frac{H(X_1, X_2, \dots, X_n)}{n} \leq L_n < \frac{H(X_1, X_2, \dots, X_n)}{n} + \frac{1}{n}.$$

如果随机过程是平稳的，则 $H(X_1, X_2, \dots, X_n)/n \rightarrow H(\mathcal{X})$. 于是我们有如下定理.

定理 10.7

每字符最小期望码字长满足

$$\frac{H(X_1, X_2, \dots, X_n)}{n} \leq L_n^* < \frac{H(X_1, X_2, \dots, X_n)}{n} + \frac{1}{n}.$$

进一步，若 $X_1, X_2, \dots, X_n$ 是平稳随机过程，则

$$L_n^* \rightarrow H(\mathcal{X}).$$

其中 $H(\mathcal{X})$ 为随机过程的熵率.

最后我们讨论当面对的对象非真实分布时，期望描述长度会变得怎么样？下面考虑概率密度函数 $q(x)$ 的香农编码，相应的码长为 $l(x) = \lceil \log \frac{1}{q(x)} \rceil$. 假定真实分布的概率密度函数是 $p(x)$.

定理 10.8: 偏码

码字长度分配 $l(x) = \lceil \log \frac{1}{q(x)} \rceil$ 关于 $p(x)$ 的期望码长满足

$$H(p) + D(p||q) \leq E_p l(X) < H(p) + D(p||q) + 1.$$

证明. 期望码长为

$$\begin{aligned} E l(x) &= \sum_x p(x) \lceil \log \frac{1}{q(x)} \rceil \\ &< \sum_x p(x) (\log \frac{1}{q(x)} + 1) \\ &= \sum_x p(x) \log \frac{p(x)}{q(x)} \frac{1}{p(x)} + 1 \\ &= \sum_x p(x) \log \frac{p(x)}{q(x)} + \sum_x p(x) \log \frac{1}{p(x)} + 1 \\ &= D(p||q) + H(p) + 1. \end{aligned}$$

类似地，可以得到期望码长的下界. □

10.5 惟一可译码的 Kraft 不等式

定理 10.9: McMillan

任意惟一可译的 D 元码的码字长度必然满足 Kraft 不等式

$$\sum D^{-l_i} \leq 1.$$

反之，若给定满足上述不等式的一组码字长度，则可以构造出具有同样码字长度的惟一可译码.

证明. 考虑编码 C 的 k 次扩展 C^k (即原先惟一可译码 C 的 k 次串联所形成的码). 由惟一可译的定义, 该码的 k 次扩展是非奇异的. 由于所有长度为 n 的不同 D 元串的数目仅为 D^n , 故由惟一可译性可知, 在码的 k 次扩展中, 长度为 n 的码序列数目必定不超过 D^n . 由此讨论我们来证明 Kraft 不等式.

设字符 $x \in \mathcal{X}$ 所对应的码字长度记为 $l(x)$. 对于扩展码, 码序列的长度为

$$l(x_1, x_2, \dots, x_k) = \sum_{i=1}^k l(x_i).$$

我们的目标是要证明

$$\sum_{x \in \mathcal{X}} D^{-l(x)} \leq 1.$$

我们考虑上式左边的 k 次方, 从而有

$$\begin{aligned} \left(\sum_{x \in \mathcal{X}} D^{-l(x)} \right)^k &= \sum_{x_1 \in \mathcal{X}} \sum_{x_2 \in \mathcal{X}} \dots \sum_{x_k \in \mathcal{X}} D^{-l(x_1)} D^{-l(x_2)} \dots D^{-l(x_k)} \\ &= \sum_{x_1, x_2, \dots, x_k \in \mathcal{X}^k} D^{-l(x_1)} D^{-l(x_2)} \dots D^{-l(x_k)} \\ &= \sum_{x^k \in \mathcal{X}^k} D^{-(x^k)}. \end{aligned}$$

我们将上式右边按字符长度重新求和, 有

$$\sum_{x^k \in \mathcal{X}^k} D^{-(x^k)} = \sum_{m=1}^{kl_{max}} a(m) D^{-m},$$

其中 l_{max} 是最大词长, $a(m)$ 为编码后码长为 m 的 x^k 的个数.

由码的惟一可译性, 我们有 $a(m) \leq D^m$, 从而我们有

$$\begin{aligned} \left(\sum_{x \in \mathcal{X}} D^{-l(x)} \right)^k &= \sum_{m=1}^{kl_{max}} a(m) D^{-m} \\ &\leq \sum_{m=1}^{kl_{max}} D^m D^{-m} \\ &= kl_{max}, \end{aligned}$$

从而

$$\sum_j D^{-l_j} \leq (kl_{max})^{1/k}.$$

由于上式对任意 k 成立, 当 $k \rightarrow \infty$ 时上式也成立. 由于 $(kl_{max})^{1/k} \rightarrow 1$, 我们有

$$\sum_{x \in \mathcal{X}} D^{-l(x)} \leq 1.$$

从而我们有 Kraft 不等式. □

推论 10.10

可数无限字母表 $\mathcal{X}$ 上的一个惟一可译码也满足 Kraft 不等式.

证明. 由于惟一可译码的任一子集也是惟一可译码, 因此无限码字集的有限子集亦满足 Kraft 不等式, 故

$$\sum_{i=1}^{\infty} D^{-l_i} = \lim_{N \rightarrow \infty} \sum_{i=1}^N D^{-l_i} \leq 1.$$

反过来, 给定满足 Kraft 的一组码字长度我们可以构造出相应的即时码, 由于即时码是惟一可译的, 所以我们就得到了所需要的惟一可译码. 因而, McMillan 定理对于无限字母表情形同样成立. □

第 11 讲 赫夫曼码与最优码

1952 年赫夫曼提出一种构造最佳码的方法称为赫夫曼码。

下面首先给出二元赫夫曼码的编码方法，它的编码步骤如下：

1. 将 q 个信源符号按概率分布 $P(s_i)$ 的大小，以递减次序排列起来，设

$$p_1 \geq p_2 \geq p_3 \geq \cdots \geq p_q.$$

2. 将 0 和 1 分别分配给概率最小的两个信源符号，并将这两个概率最小的信源符号合并成一个新符号，并用这两个最小概率之和作为新符号的概率，从而得到只包含 $q-1$ 个符号的新信源，称为信源 S 的缩减信源 S_1 。
3. 把缩减信源 S_1 的符号仍按概率大小以递减依次序排列，再将其最后两个概率最小的符号合并成一个新符号，并分别用 0 和 1 码符号表示，这样得到了 $q-2$ 个符号的缩减信源 S_2 。
4. 依次继续下去，直至缩减信源最后只剩两个符号为止。将这最后两个符号分别用 0 和 1 码符号表示。最后这两个符号的概率之和必为 1。然后从最后一级缩减信源开始，依编码路由后向前返回，就得到各信源符号所对应的码符号序列，即得对应的码字。

11.1 r 元赫夫曼码

上面我们讨论了二元赫夫曼码，它的编码方式同样可以推广到 r 元编码中来。不同的只是把 r 个符号（概率最小的）合并成一个新信源信号，并分别用 $0, 1, \cdots, (r-1)$ 等码元表示。为了使短码得到充分利用，使平均码长最短，必须使最后一步的缩减信源有 r 个信源符号。因此对于 r 元编码，信源 S 的符号个数 q 必须满足

$$q = (r-1)\theta + r.$$

式中， θ 表示缩减的次数， $(r-1)$ 为每次缩减所减少的信源符号个数。对于二元气码，信源符号个数 q 必须满足

$$q = \theta + 2.$$

可见对于二元气码， q 等于任意正整数时，总能找到一个 θ 满足上式。而对于 r 元码，则不一定。此时我们不妨虚设 t 个信源符号 $s_{q+1}, s_{q+2}, \cdots, s_{q+t}$ 并使它们对应的概率为零，即 $p_{q+1} = p_{q+2} = \cdots = p_{q+t} = 0$ 。此时， $q+t = (r-1)\theta + r$ 。

11.2 赫夫曼码的相关讨论

赫夫曼码的优点有：

- 赫夫曼码的编码方法保证了概率大的符号对应于短码，概率小的符号对应于长码，充分利用了短码；
- 缩减信源的最后两个码字总是最后一位不同，从而保证了赫夫曼码是即时码。

赫夫曼码的缺点有：

- 每个信源符号所对应的码长不同。一般情况下，信源符号以恒定速度输出，信道也是恒速传输的。通过编码后，会造成编码输出每秒的比特数不是常量，因而不能直接由信道来传送。
- 信源符号与码字之间不能用某种有规律的数学方法对应起来。

从赫夫曼码的编码过程可以看出, 赫夫曼码的编码方式得到的码一定是即时码, 因为这种编码方法不会使任一码字的前缀为码字.

注意到赫夫曼码编码方法得到的码并非使唯一的. 首先因为, 每次对缩减信源最后两个概率最小的符号, 用 0 和 1 码是任意的, 所以可以得到不同的码. 但它们只是码字的具体形式不同, 而其码长 l_i 不变, 平均码长 $\bar{L}$ 也不变, 所以没有本质差别. 其次, 若当缩减信源中缩减合并的符号的概率与其他信源符号概率相同时, 从编码方法上来说, 它们概率次序的排列哪个在前哪个在后是没有区别的, 但得到的码使不同的赫夫曼码. 对于这两种码, 他们的码长 l_i 各不同, 然而平均码长 $\bar{L}$ 是相同的.

那么, 在这些不同的码中, 选哪一个码好呢? 我们引进码字长度 l_i 偏离平均长度 $\bar{L}$ 的方差 σ^2 , 即

$$\sigma^2 = E[(l_i - \bar{L})^2] = \sum_{i=1}^q P_i(l_i - \bar{L})^2.$$

当方差 σ^2 比较小时, 码序列长度变化比较小, 相对来说效果会更好.

11.3 赫夫曼码的最优性

引理 11.1

对于任意一个分布, 必然存在满足如下性质的一个即时最优码 (即有最小期望长度):

1. 其长度序列与按概率分布列排列的次序相反, 即, 若 $p_j > p_k$, 则 $l_j \leq l_k$.
2. 最长的两个码字具有相同的长度.
3. 最长的两个码字在最后一位上有所差别, 且对应于两个最小可能发生的字符.

证明. 考虑一个最优码 C_m :

- 若 $p_j > p_k$, 则 $l_j \leq l_k$. 此时通过交换码字即可得此结论.

设 C'_m 为将 C_m 中的码字 j 和 k 交换所得到的编码, 则

$$\begin{aligned} L(C'_m) - L(C_m) &= \sum p_i l'_i - \sum p_i l_i \\ &= p_j l_k + p_k l_j - p_j l_j - p_k l_k \\ &= (p_j - p_k)(l_k - l_j). \end{aligned}$$

但 $p_j - p_k > 0$, 由于 C_m 是最优的, 可得 $L(C'_m) - L(C_m) \geq 0$, 故必有 $l_k \geq l_j$. 从而最优码本身 C_m 必定满足性质 1.

- 最长的两个码字具有相同的长度.

通过修建码字获得结论. 如果两个码字最长码长度不相同, 那么将较长的码字最后一位删除, 它仍可保持前缀性质, 但此时具有更短的期望码字长. 因此, 最长的两个码字长度必定相等. 由性质 1 我们知, 最长的码字对应于那些最小可能发生的符号信源字符.

- 两个最长码仅在最后一位有所差别, 并且分别对应两个最小可能发生的信源字符.

并非所有的最优码都满足这个性质, 但通过重排可以获得满足该性质的最优码. 如果存在长度最长的码字, 则删除码字的最后一位, 所得的码字仍然满足前缀性质. 从而期望的码字长度有所减小, 这与编码的最优性矛盾. 因此, 在任何一个最优编码中, 最大长度码字有兄弟.

总之, 我们已证明: 若 $p_1 \geq p_2 \geq \cdots \geq p_m$, 则存在长度列为 $l_1 \leq l_2 \leq \cdots \leq l_{m-1} = l_m$ 的一个最优码, 且码字 $C(x_{m-1})$ 和 $C(x_m)$ 仅最后一位有所区别. $\square$

满足上述引理的最优码我们称之为典则码. 对于 m 个字符上的概率密度函数 $\mathbf{p} = (p_1, p_2, \dots, p_m)$, $p_1 \geq p_2 \geq \dots \geq p_m$, 我们定义其 $m-1$ 个字符上的赫夫曼合并为 $\mathbf{p}' = (p_1, p_2, \dots, p_{m-1} + p_m)$. 用 $C_{m-1}^*(\mathbf{p}')$ 表示 $\mathbf{p}'$ 的最优码, 而用 $C_m^*(\mathbf{p})$ 表示 $\mathbf{p}$ 的典则最优码.

定理 11.2

赫夫曼码是最优的, 即如果 C^* 为赫夫曼码而 C' 为其他码, 则有 $L(C^*) \leq L(C')$.

证明. 我们利用归纳法, 对信源的字符个数做归纳. 当信源有一个和两个字符时, 显然赫夫曼码是最优的. 下面我们假设对于 $m-1$ 个字符的信源, 赫夫曼码是最优的. 对于 m 个字符上的概率密度函数 $\mathbf{p} = (p_1, p_2, \dots, p_m)$, $p_1 \geq p_2 \geq \dots \geq p_m$, 我们定义其 $m-1$ 个字符上的赫夫曼合并为 $\mathbf{p}' = (p_1, p_2, \dots, p_{m-1} + p_m)$. 用 $C_{m-1}^*(\mathbf{p}')$ 表示 $\mathbf{p}'$ 的最优码, 而用 $C_m^*(\mathbf{p})$ 表示 $\mathbf{p}$ 的典则最优码. 记 $L^*(\mathbf{p})$ 和 $L^*(\mathbf{p}')$ 为对应的概率密度函数所对应的最优码平均码长.

基于 $\mathbf{p}'$ 的赫夫曼码我们可以构造一个对应 m 字符信源的编码: 取 C_{m-1}^* 中的权重为 $p_{m-1} + p_m$ 的码字, 对其进行扩展, 在尾部加 0 形成字符 $m-1$ 相应码字, 加 1 形成字符 m 的相应码字. 由此得到的新码是对应的赫夫曼码, 码字平均长度为

$$L(\mathbf{p}) = L^*(\mathbf{p}') + p_{m-1} + p_m.$$

另外, 我们从 $\mathbf{p}$ 的典则码出发, 将两个最小概率 p_{m-1} 和 p_m 对应的字符 $m-1$ 和 m 合并, 可以构造关于分布 $\mathbf{p}'$ 的新码, 其平均长度为

$$L(\mathbf{p}') = L^*(\mathbf{p}) - p_{m-1} - p_m.$$

于是

$$(L(\mathbf{p}') - L^*(\mathbf{p}')) + (L(\mathbf{p}) - L^*(\mathbf{p})) = 0$$

而由最优码有最短的平均码字, 故我们有 $L(\mathbf{p}) = L^*(\mathbf{p})$. 从而我们知上面得到的 m 元字符信源的赫夫曼码是最优码. $\square$

11.4 Shannon-Fano-Elias 编码

不失一般性, 假定取 $\mathcal{X} = \{1, 2, \dots, m\}$. 假设对所有的 x , 有 $p(x) > 0$. 定义累积分布函数 $F(x)$ 为

$$F(x) = \sum_{a \leq x} p(a).$$

考虑修正的累积分布函数

$$\bar{F}(x) = \sum_{a < x} p(a) + \frac{1}{2}p(x).$$

由于所有的概率值是正的, 若 $a \neq b$, 则 $\bar{F}(a) \neq \bar{F}(b)$. 若已知 $\bar{F}(x)$, 则可以确定 x . 故 $\bar{F}(x)$ 可以作为 x 的编码. 但是在一般情况下, $\bar{F}(x)$ 是一个无理数, 其需要无限多比特才能表示. 所以使用 $\bar{F}(x)$ 的精确值作为对 x 的编码并非切实可行. 那么如果我们要使用近似值, 那么需要精确到什么程度呢?

假定将 $\bar{F}(x)$ 舍入取 $l(x)$ 位 (记为 $\lfloor \bar{F}(x) \rfloor_{l(x)}$). 于是, 取 $\bar{F}(x)$ 的前 $l(x)$ 位作为 x 的码. 由舍入定义, 可得

$$\bar{F}(x) - \lfloor \bar{F}(x) \rfloor_{l(x)} < \frac{1}{2^{l(x)}}.$$

若 $l(x) = \lceil \log \frac{1}{p(x)} \rceil + 1$, 则

$$\frac{1}{2^{l(x)}} < \frac{p(x)}{2} = \bar{F}(x) - F(x-1).$$

于是 $\lfloor \bar{F}(x) \rfloor_{l(x)}$ 和 $\lfloor \bar{F}(x-1) \rfloor_{l(x)}$ 不同. 从而, 使用 $l(x)$ 比特足以表示 x .

这里除了要求码字和字符一一对应之外, 还需要码字集是无前缀的. 为验证该编码是否为前缀码是否为前缀码, 考虑每个码字 $z_1 z_2 \cdots z_l$, 注意到它实际上代表的不是一个点, 而是一个区间 $[0.z_1 z_2 \cdots z_l, 0.z_1 z_2 \cdots z_l + \frac{1}{2^l}]$. 码是无前缀的当且仅当码字对应的区间互不相交.

下面我们来证明上述码字集无前缀. 对应任意码字的区间长度为 $2^{-l(x)}$, 由 (11.4) 可知所有区间长度均小于 x 对应的阶梯高度的 $1/2$. 区间的下端位于对应阶梯的下一半中, 于是区间的上端位于对应阶梯的顶部之下, 故而在累积分布函数之中, 任一码字对应的区间都真包含于相应字符所对应的阶梯中. 所以不同码字对应的区间不相交, 于是我们知此码是无前缀的. 注意, 该程序没有要求字符按其概率大小顺序排列.

我们可以计算编码的期望长度:

$$L = \sum_x p(x) l(x) = \sum_x p(x) (\lceil \log \frac{1}{p(x)} \rceil + 1) < H(X) + 2.$$

因此, 该编码方案的期望码长不会超过熵值 2 比特.

11.5 香农码的竞争最优性

定理 11.3

设 $l(x)$ 为香农码的相应码字长度, 而 $l'(x)$ 表示其他唯一可译码的相应码字长度. 则

$$P(l(X) \geq l'(X) + c) \leq \frac{1}{2^{c-1}}.$$

证明.

$$\begin{aligned} P(l(X) \geq l'(X) + c) &= P(\lceil \log \frac{1}{p(X)} \rceil \geq l'(X) + c) \\ &\leq P(\log \frac{1}{p(X)} \geq l'(X) + c - 1) \\ &= P(p(X) \leq 2^{-l'(X) - c + 1}) \\ &= \sum_{x: p(x) \leq 2^{-l'(x) - c + 1}} p(x) \\ &\leq \sum_{x: p(x) \leq 2^{-l'(x) - c + 1}} 2^{-l'(x) - (c-1)} \\ &\leq \sum_x 2^{-l'(x)} 2^{-(c-1)} \\ &\leq 2^{-(c-1)}. \end{aligned}$$

最后一步我们用到了 Kraft 不等式. □

定理 11.4

对二进制概率密度函数 $p(x)$, 设 $l(x) = \log \frac{1}{p(x)}$ 为信源的二维香农码的码字长度, $l'(x)$ 为信源任何其他唯一可译二元码的码字长度. 则

$$P(l(X) < l'(X)) \geq P(l(X) \geq l'(X)).$$

当且仅当对所有的 x , 有 $l'(x) = l(x)$ 等号成立. 于是码长分配 $l(x) = \log \frac{1}{p(x)}$ 是唯一竞争最优的.

证明. 定义函数 $\text{sgn}(t)$ 如下:

$$\text{sgn}(t) = \begin{cases} 1 & \text{当 } t > 0 \\ 0 & \text{当 } t = 0 \\ -1 & \text{当 } t < 0 \end{cases}$$

我们不难看出

$$\text{sgn}(t) \leq 2^t - 1, \quad t = 0, \pm 1, \pm 2, \dots$$

于是我们可以得到

$$\begin{aligned} P\{l'(X) < l(X)\} - P\{l'(X) > l(X)\} &= \sum_{x:l'(x) < l(x)} p(x) - \sum_{x:l'(x) > l(x)} p(x) \\ &= \sum_x p(x) \text{sgn}(l(x) - l'(x)) \\ &= E \text{sgn}(l(X) - l'(X)) \\ &\leq \sum_x p(x) (2^{l(x) - l'(x)} - 1) \\ &= \sum_x 2^{-l(x)} (2^{l(x) - l'(x)} - 1) \\ &= \sum_x 2^{-l'(x)} - \sum_x 2^{-l(x)} \\ &= \sum_x 2^{-l'(x)} - 1 \leq 1 - 1 = 0. \end{aligned}$$

讨论上面不等式等号成立的条件我们知当且仅当对所有的 x , 有 $l'(x) = l(x)$ 等号成立. □

推论 11.5

对于非二进的概率密度函数,

$$E \text{sgn}(l(x) - l'(x) - 1) \leq 0.$$

其中 $l(x) = \lceil \log \frac{1}{p(x)} \rceil$, $l'(x)$ 为信源其他任何一个编码.

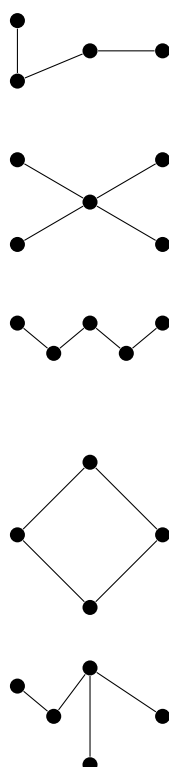
第 12 讲 习题课 2

例 12.1

考虑在一个有 4 条边的连通图上的随机游走。

- (a) 哪个图的熵率 (entropy rate) 最高?
- (b) 哪个图最低?

解. 共有五个有 4 条边的图.



其中熵率分别为 $1/2 + 3/8 \log(3) \approx 1.094$ 、1、0.75、1 以及 $1/4 + 3/8 \log(3) \approx 0.844$.

(a) 由上可知, 第一个图实现了最大熵率, 其熵率为 1.094.

(b) 由上可知, 第三个图实现了最小熵率, 其熵率为 0.75.

□

例 12.2

假设消息的概率分布以递减的顺序给出: $p_1 \geq p_2 \geq \dots \geq p_m$. 证明: 对任意的二元赫夫曼码, 如果最可能出现的消息字符的概率 $p_1 > 2/5$, 则该字符分配的码字长度必为 1.

第13讲 信道容量

13.1 信道容量

定义 13.1

离散信道是由输入字母表 $\mathcal{X}$ ，输出字母表 $\mathcal{Y}$ 和概率转移矩阵 $p(y|x)$ 构成的系统，其中 $p(y|x)$ 表示发送字符 x 的条件下收到输出字符 y 的概率。如果输出的概率分布仅依赖于它的所对应的输入，而与先前信道的输入或者输出条件独立，就称这个信道是无记忆的。

定义 13.2

离散无记忆信道的“信息”信道容量定义为

$$C = \max_{p(x)} I(X; Y).$$

这里的最大值取自所有可能的输入分布 $p(x)$ 。

13.2 信道容量的几个例子

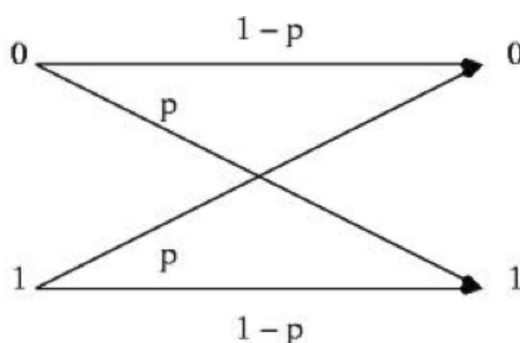
13.2.1 无噪声二元信道

对于无噪声二元信道，它的二元输入在输出端能精确地重现。

在这种情况下，任何一个传输的比特都能被无误差地接收到。因此，每次使用该信道，都可以毫无误差地传输一个比特，信道容量就是一个比特。当然，也可以计算得到信道容量 $C = \max I(X; Y) = 1$ 比特，且在 $p(x) = (\frac{1}{2}, \frac{1}{2})$ 时达到。

13.2.2 二元对称信道

考虑下图所示的二元对称信道。这个二元信道的输入字符以概率 p 互补。这是一个有误差信道的最简单模型。



在出现错误时，发出的信息 0 接受到的结果是 1，或者正好相反。从接收到的比特我们并不能看出哪里发生了错误。从某种意义上说，所有接收到的比特都不可靠。我们稍后将证明，我们仍然可以使用这样的通信信道以非 0 的传输码率发送信息，并且误差概率可以任意小。

我们可以给出互信息的一个界

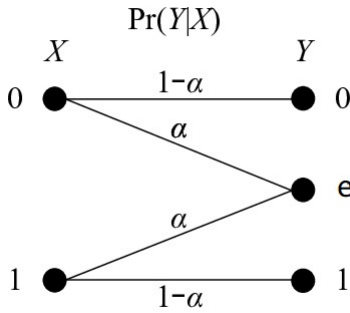
$$\begin{aligned}
 I(X;Y) &= H(Y) - H(Y|X) \\
 &= H(Y) - \sum p(x)H(Y|X=x) \\
 &= H(Y) - \sum p(x)H(p) \\
 &= H(Y) - H(p) \\
 &\leq 1 - H(p).
 \end{aligned}$$

其中最后一个不等式成立是因为 Y 是一个二元随机变量. 当输入分布是均匀分布时等号成立. 因此, 参数为 p 的二元对称信道的信道容量为

$$C = 1 - H(p) \text{ 比特.}$$

13.2.3 二元擦除信道

有一种信道类似于二元对称信道, 会损失一些比特 (不是被损坏), 这种信道称为二元擦除信道. 在二元擦除信道中, 比例为 α 的比特被擦除掉, 并且接收者知道是哪些比特已经被擦除掉了. 如下图所示, 二元擦除信道有两个输入和三个输出.



我们尝试利用和二元对称信道类似的方法, 计算二元擦除信道的容量如下:

$$\begin{aligned}
 C &= \max_{p(x)} I(X;Y) \\
 &= \max_{p(x)} (H(Y) - H(Y|X)) \\
 &= \max_{p(x)} H(Y) - H(\alpha).
 \end{aligned}$$

初看, 似乎 $H(Y)$ 的最大值是 $\log 3$. 但是实际上在绝大多数情况下 (除了 $\alpha = 1/3$ 时), 无论如何选取初始分布 $p(x)$, 我们都无法让 $H(Y)$ 达到这个值. 所以我们需要另辟蹊径.

设 E 代表事件 $\{Y = e\}$, 设 $P(X = 1) = \pi$, 由表达式

$$H(Y) = H(Y, E) = H(E) + H(Y|E)$$

我们有

$$H(Y) = H((1 - \pi)(1 - \alpha), \alpha, \pi(1 - \alpha)) = H(\alpha) + (1 - \alpha)H(\pi).$$

因此

$$\begin{aligned}
 C &= \max_{p(x)} H(Y) - H(\alpha) \\
 &= \max_{\pi} (1 - \alpha)H(\pi) + H(\alpha) - H(\alpha) \\
 &= \max_{\pi} (1 - \alpha)H(\pi) \\
 &= 1 - \alpha.
 \end{aligned}$$

其中, 当 $\pi = 1/2$ 时, 达到该信道容量.

13.3 对称信道

二元对称信道是 $C = 1 - H(p)$ 比特/传输, 二元擦除信道的容量是 $C = 1 - \alpha$ 比特/传输. 下面考虑具有如下转移矩阵的信道:

$$p(y|x) = \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.5 & 0.3 & 0.2 \\ 0.2 & 0.5 & 0.3 \end{bmatrix}$$

上述矩阵中的第 x 行第 y 列的元素表示条件概率 $p(y|x)$, 即传输 x 收到 y 的概率. 在该信道中, 概率转移矩阵中所有的行都可以通过其他行置换得到, 每一列也如此. 这样的信道称为对称的.

例 13.3

另一个对称信道的例子是

$$Y = X + Z \pmod{c}$$

其中 Z 服从整数集 $\{0, 1, 2, \dots, c-1\}$ 上的某个分布, X 与 Z 拥有相同的字母表, 并且 Z 独立于 X .

设 $\mathbf{r}$ 表示转移矩阵中的一行, 则有

$$I(X; Y) = H(Y) - H(Y|X) = H(Y) - H(\mathbf{r}) \leq \log |\mathcal{Y}| - H(\mathbf{r}),$$

当输出是均匀分布时等号成立. 而且, $p(x) = 1/|\mathcal{X}|$ 可以使 Y 达到均匀分布, 这可以由如下式子看出

$$p(y) = \sum_{x \in \mathcal{X}} p(y|x)p(x) = \frac{1}{|\mathcal{X}|} \sum_x p(y|x) = c \frac{1}{|\mathcal{X}|} = \frac{1}{|\mathcal{Y}|}$$

其中 c 是概率转移矩阵的一列中所有元素之和, 不难证明 $c = \frac{|\mathcal{X}|}{|\mathcal{Y}|}$.

定义 13.4

如果信道转移矩阵 $p(y|x)$ 的任何两行互相置换; 任何两列也互相置换, 那么称该信道是对称的. 如果转移矩阵的每一行 $p(\cdot|x)$ 都是其它每行的置换, 而所有列的元素和 $\sum_x p(y|x)$ 相等, 则称这个信道是弱对称的.

定理 13.5

对于弱对称信道,

$$C = \log |\mathcal{Y}| - H(\text{转移矩阵的行}),$$

当输入字母表上的分布为均匀时达到该容量.

13.4 并联信道

我们考虑两个信道 $\{\mathcal{X}_1, p_1(y_1|x_1), \mathcal{Y}_1\}$ 和 $\{\mathcal{X}_2, p_2(y_2|x_2), \mathcal{Y}_2\}$ 并联后的信道容量 C , 其中要求每次发送字符时, 要么是通过信道 1, 要么是通过信道 2 发送, 而不能同时发送, 假定两者的输出字母表不相交.

命题 13.6

$$2^C = 2^{C_1} + 2^{C_2}.$$

证明. 我们考虑下面的信道:

$$X = \begin{cases} X_1 & \text{概率 } \alpha \\ X_2 & \text{概率 } (1 - \alpha). \end{cases}$$

设

$$\theta(X) = \begin{cases} 1 & X = X_1 \\ 2 & X = X_2 \end{cases}$$

由于 $\mathcal{Y}_1 \cap \mathcal{Y}_2 = \emptyset$, 我们有 $X \rightarrow Y \rightarrow \theta$, 从而

$$\begin{aligned} I(X; Y, \theta) &= I(X; \theta) + I(X; Y | \theta) \\ &= I(X; Y) + I(X; \theta | Y). \end{aligned}$$

由 $\mathcal{Y}_1 \cap \mathcal{Y}_2 = \emptyset$ 我们有 $I(X; \theta | Y) = 0$, 从而

$$\begin{aligned} I(X; Y) &= I(X; \theta) + I(X; Y | \theta) \\ &= H(\theta) - H(\theta | X) + \alpha I(X_1; Y_1) + (1 - \alpha) I(X_2; Y_2) \\ &= H(\alpha) + \alpha I(X_1; Y_1) + (1 - \alpha) I(X_2; Y_2). \end{aligned}$$

于是我们不难看出

$$C = \sup_{\alpha} \{H(\alpha) + \alpha C_1 + (1 - \alpha) C_2\}.$$

于是我们把问题化归为求函数最值问题. 不难验证, 当 $H'(\alpha) + C_1 - C_2 = 0$ 时, 即 $\alpha = 2^{C_1} / (2^{C_1} + 2^{C_2})$ 时达到最大值, 此时

$$2^C = 2^{C_1} + 2^{C_2}. \quad \square$$

13.5 信道容量的性质

- 由 $I(X; Y) \geq 0$ 我们知 $C \geq 0$.
- 由于 $C = \max I(X; Y) \leq \max H(X) \leq \log |\mathcal{X}|$, 所以 $C \leq \log |\mathcal{X}|$.
- 同理知 $C \leq \log |\mathcal{Y}|$.
- $I(X; Y)$ 是关于 $p(x)$ 的一个连续函数.
- $I(X; Y)$ 是关于 $p(x)$ 的凹函数. 由于 $I(X; Y)$ 是闭凸集上的凹函数, 因此局部最大值也是全局最大值. 最大值可以利用标准的非线性最优化技术求解.

13.6 一些关键定义

取自下标集 $\{1, 2, \dots, M\}$ 的消息 W , 产生信号 $X^n(W)$, 这个信号以随机序列 $Y^n \sim p(y^n | x^n)$ 的方式被接收者收到. 然后, 接收者使用适当的译码规则 $\hat{W} = g(Y^n)$ 猜测消息 W . 如果 $\hat{W}$ 与所传输的消息 W 不同, 则表明接受者出错. 下面我们严格定义这些思路.

定义 13.7

用 $(\mathcal{X}, p(y|x), \mathcal{Y})$ 表示的离散信道由两个有限集 $\mathcal{X}$ 和 $\mathcal{Y}$ 以及一簇概率密度函数 $p(y|x) (x \in \mathcal{X})$ 构成, 其中对任意 x 与 y , 有 $p(y|x) \geq 0$, 以及对任意的 x , 有 $\sum_y p(y|x) = 1$, 而 X 和 Y 分别看作信道的输入和输出.

定义 13.8

离散无记忆信道 (DMC) 的 n 次扩展是指信道 $(\mathcal{X}^n, p(y^n|x^n), \mathcal{Y}^n)$, 其中

$$p(y_k|x^k, y^{k-1}) = p(y_k|x_k), \quad k = 1, 2, \dots, n.$$

注记 13.9

如果信道不带反馈, 也就是说, 如果输入字符不依赖于过去的输出字符, 即 $p(x_k|x^{k-1}, y^{k-1}) = p(x_k|x^{k-1})$, 那么离散无记忆信道的 n 次扩展的信道转移函数就可以转化为

$$p(y^n|x^n) = \prod_{i=1}^n p(y_i|x_i).$$

为证明上式, 我们首先注意到

$$p(y^n|x^n) = p(y_n|y^{n-1}, x^n)p(y^{n-1}|x^n) = p(y_n|x_n)p(y^{n-1}|x^n).$$

为计算 $p(y^{n-1}|x^n)$, 我们利用下式

$$p(y^{n-1}|x^n)p(x_n|x^{n-1}) = p(y^{n-1}, x_n|x^{n-1}) = p(x_n|y^{n-1}, x^{n-1})p(y^{n-1}|x^{n-1}).$$

注意到信道不带反馈, $p(x_n|x^{n-1}) = p(x_n|y^{n-1}, x^{n-1})$. 从而

$$p(y^{n-1}|x^n) = p(y^{n-1}|x^{n-1}).$$

再利用归纳法我们可得所需结论.

定义 13.10

信道 $(\mathcal{X}, p(y|x), \mathcal{Y})$ 的 (M, n) 码由以下部分构成:

1. 下标集 $\{1, 2, \dots, M\}$.
2. 编码函数 $X^n: \{1, 2, \dots, M\} \rightarrow \mathcal{X}^n$, 生成码字 $x^n(1), x^n(2), \dots, x^n(M)$. 所有码字的集合称作码簿.
3. 译码函数

$$g: \mathcal{Y}^n \rightarrow \{1, 2, \dots, M\}$$

其为一个确定性规则, 为每个收到的字符向量指定一个猜测.

定义 13.11: 条件概率误差

设

$$\lambda_i = P\{g(Y^n) \neq i | X^n = x^n(i)\} = \sum_{y^n} p(y^n|x^n(i))I(g(y^n) \neq i)$$

为已知下标 i 被发送的条件下的条件误差概率, 其中 $I(\cdot)$ 为示性函数.

定义 13.12

(M, n) 码的最大误差概率 $\lambda^{(n)}$ 定义为

$$\lambda^{(n)} = \max_{i \in (1, 2, \dots, M)} \lambda_i.$$

定义 13.13

(M, n) 码的平均误差概率 $P_e^{(n)}$ 定义为

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^M \lambda_i.$$

定义 13.14

(M, n) 码的码率定义为

$$R = \frac{\log M}{n} \text{ 比特/传输.}$$

定义 13.15

如果存在一个 $(\lceil 2^{nR} \rceil, n)$ 码序列, 满足当 $n \rightarrow \infty$ 时, 最大误差概率 $\lambda^{(n)} \rightarrow 0$, 则称码率 R 是可达的.

定义 13.16

信道的容量定义为所有可达码率的上确界.

第 14 讲 信道编码定理

14.1 联合典型序列

定义 14.1

服从分布 $p(x, y)$ 的联合典型序列 $\{(x^n, y^n)\}$ 所构成的集合 $A_\epsilon^{(n)}$ 是指其经验熵与真实熵 ϵ 接近的 n 长序列构成的集合, 即:

$$A_\epsilon^{(n)} = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \\ |-\frac{1}{n} \log p(x^n) - H(X)| < \epsilon, \\ |-\frac{1}{n} \log p(y^n) - H(Y)| < \epsilon, \\ |-\frac{1}{n} \log p(x^n, y^n) - H(X, Y)| < \epsilon\}$$

其中

$$p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i).$$

定理 14.2: 联合 AEP

设 (X^n, Y^n) 为服从 $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$ 的 i.i.d. 的 n 长序列, 那么:

- 对任意 $\epsilon > 0$, 当 n 充分大时, 有 $P((X^n, Y^n) \in A_\epsilon^{(n)}) > 1 - \epsilon$.
- $|A_\epsilon^{(n)}| \leq 2^{n(H(X, Y) + \epsilon)}$.
- 如果 $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$, 即 $\tilde{X}^n$ 与 $\tilde{Y}^n$ 是独立的且与 $p(x^n, y^n)$ 有相同的边际分布, 那么

$$P\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\} \leq 2^{-n(I(X; Y) - 3\epsilon)}.$$

而且, 对于充分大的 n ,

$$P\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\} \geq (1 - \epsilon)2^{-n(I(X; Y) + 3\epsilon)}.$$

证明. 1. 首先我们来证明, 包含在典型集中的序列具有很高的概率. 由弱大数定律,

$$-\frac{1}{n} \log P(X^n) \rightarrow -E[\log p(X)] = H(X) \text{ 依概率.}$$

因此, 给定 $\epsilon > 0$, 存在 n_1 , 使得对于任意 $n > n_1$, 使得对于任意 $n > n_1$,

$$P(|-\frac{1}{n} \log p(X^n) - H(X)| \geq \epsilon) < \frac{\epsilon}{3}.$$

类似地, 由弱大数定律, 我们有

$$-\frac{1}{n} \log p(Y^n) \rightarrow -E[\log p(Y)] = H(Y) \text{ 依概率,}$$

以及

$$-\frac{1}{n} \log p(X^n, Y^n) \rightarrow -E[\log p(X, Y)] = H(X, Y) \text{ 依概率.}$$

从而, 存在 n_2 和 n_3 , 使得对任意 $n \geq n_2$,

$$P(|-\frac{1}{n} \log p(Y^n) - H(Y)| \geq \epsilon) < \frac{\epsilon}{3}.$$

以及对任意 $n \geq n_3$,

$$P(|-\frac{1}{n} \log p(X^n, Y^n) - H(X, Y)| \geq \epsilon) < \frac{\epsilon}{3}.$$

选取 $n > \max(n_1, n_2, n_3)$, 则我们知

$$\begin{aligned} P((X^n, Y^n) \in A_\epsilon^{(n)}) &\geq 1 - P(|-\frac{1}{n} \log p(X^n) - H(X)| \geq \epsilon) \geq \epsilon \\ &\quad - P(|-\frac{1}{n} \log p(Y^n) - H(Y)| \geq \epsilon) \geq \epsilon \\ &\quad - P(|-\frac{1}{n} \log p(X^n, Y^n) - H(X, Y)| \geq \epsilon) \geq \epsilon \\ &> 1 - \epsilon \end{aligned}$$

从而定理的第一部分得证. 2. 为证明定理的第二部分, 我们注意到

$$\begin{aligned} 1 &= \sum p(x^n, y^n) \\ &\geq \sum_{A_\epsilon^{(n)}} p(x^n, y^n) \\ &\geq |A_\epsilon^{(n)}| 2^{-n(H(X, Y) + \epsilon)}. \end{aligned}$$

因此, $|A_\epsilon^{(n)}| \leq 2^{n(H(X, Y) + \epsilon)}$. 3. 现在, 如果 $\tilde{X}^n$ 和 $\tilde{Y}^n$ 相互独立, 但是与 X^n 和 Y^n 分别具有相同的边际分布, 那么

$$\begin{aligned} P\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\} &= \sum_{(x^n, y^n) \in A_\epsilon^{(n)}} p(x^n) p(y^n) \\ &\leq 2^{n(H(X, Y) + \epsilon)} 2^{-n(H(X) - \epsilon)} 2^{-n(H(Y) - \epsilon)} \\ &= 2^{-n(I(X; Y) - 3\epsilon)}. \end{aligned}$$

对充分大的 n , $P(A_\epsilon^{(n)}) \geq 1 - \epsilon$, 因此

$$\begin{aligned} 1 - \epsilon &\leq \sum_{(x^n, y^n) \in A_\epsilon^{(n)}} p(x^n, y^n) \\ &\leq |A_\epsilon^{(n)}| 2^{-n(H(X, Y) - \epsilon)} \end{aligned}$$

以及

$$|A_\epsilon^{(n)}| \geq (1 - \epsilon) 2^{n(H(X, Y) - \epsilon)}.$$

和上界讨论的估计类似, 我们可以证明, 对于充分大的 n , 有

$$\begin{aligned} P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}) &= \sum_{(x^n, y^n) \in A_\epsilon^{(n)}} p(x^n) p(y^n) \\ &\geq (1 - \epsilon) 2^{n(H(X, Y) - \epsilon)} 2^{-n(H(X) + \epsilon)} 2^{-n(H(Y) + \epsilon)} \\ &= (1 - \epsilon) 2^{-n(I(X; Y) + 3\epsilon)}. \quad \square \end{aligned}$$

14.2 信道编码定理

定理 14.3: 信道编码定理

对于离散无记忆信道, 小于信道容量 C 的所有码率都是可达的. 具体来说, 对任意码率 $R < C$, 存在一个 $(2^{nR}, n)$ 码序列, 它的最大误差概率为 $\lambda^{(n)} \rightarrow 0$.
反之, 任何满足 $\lambda^{(n)} \rightarrow 0$ 的 $(2^{nR}, n)$ 码序列必定有 $R \leq C$.

证明. 我们在这一节证明可达性, 逆定理的证明放在后面.

固定 $p(x)$, 根据分布 $p(x)$ 随机生成一个 $(2^{nR}, n)$ 码. 具体来说, 根据分布

$$p(x^n) = \prod_{i=1}^n p(x_i)$$

独立生成 2^{nR} 个码字, 将它们作为信息 $\{1, \dots, M\}$ 的编码. 将 2^{nR} 个码字展开为矩阵的行:

$$\mathcal{C} = \begin{pmatrix} x_1(1) & x_2(1) & \cdots & x_n(1) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \cdots & x_n(2^{nR}) \end{pmatrix}$$

该矩阵的每一项都是依据 i.i.d. 服从 $p(x)$ 生成的, 即 $P(X^n(w) = x^n) = \prod_{i=1}^n p(x_i)$ 且 $X^n(w)$ 是独立的, 其中 $w \in \{1, \dots, M\}$. 依次我们生成一个特定码 $\mathcal{C}$ 的概率就是

$$P(\mathcal{C}) = P(X^n(w) = x_1(w)x_2(w) \cdots x_n(w), w \in \{1, 2, \dots, M\}) \quad (14.1)$$

$$= \prod_{w=1}^{2^{nR}} \prod_{i=1}^n p(x_i(w)). \quad (14.2)$$

我们考虑下面的系列事件:

1. 如14.1所示, 按分布 $p(x)$ 生成一个码簿 $\mathcal{C}$.
2. 我们将码簿 $\mathcal{C}$ 告诉发送者和接收者, 并且假定两者都知道该信道的信道转移矩阵 $p(y|x)$.
3. 依如下的均匀分布选取一条消息 W

$$P(W = w) = 2^{-nR}, w = 1, 2, \dots, 2^{nR}.$$

4. 第 w 个码字 $X^n(w)$ 是 $\mathcal{C}$ 的第 w 行, 通过该信道被发送.
5. 接收者收到的序列 Y^n 服从分布

$$p(y^n|x^n(w)) = \prod_{i=1}^n p(y_i|x_i(w)).$$

6. 接收者猜测所发送的消息是什么. 如果满足下面两个条件, 则接收者认为 $\hat{W}$ 就是所发送的下标.

- $(X(\hat{W}), Y^n)$ 是联合典型的.
- 不存在其他的下标 $W' \neq \hat{W}$ 满足 $(X^n(W'), Y^n) \in A_\epsilon^{(n)}$.

如果这样的 $\hat{W}$ 不存在, 或者有超过一个这样的 $\hat{W}$, 则断言发生了错误 (在这种情况下, 假定接收者给出一个哑下标, 例如 0).

7. 如果 $\hat{W} \neq W$, 则说明译码错误, 我们记事件 $\{\hat{W}(Y^n) \neq W\}$ 为 $\mathcal{E}$.

下面我们来计算误差概率. 设 W 服从 $\{1, 2, \dots, 2^{nR}\}$ 上的均匀分布, 并用上面的译码方法得到 $\hat{W}(y^n)$. 设 $\mathcal{E} = \{\hat{W}(Y^n) \neq W\}$ 表示误差事件. 现在计算平均误差概率, 也就是

$$\begin{aligned} P(\mathcal{E}) &= \sum_{\mathcal{C}} P(\mathcal{C}) P_e^{(n)}(\mathcal{C}) \\ &= \sum_{\mathcal{C}} P(\mathcal{C}) \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(\mathcal{C}) \\ &= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathcal{C}} P(\mathcal{C}) \lambda_w(\mathcal{C}). \end{aligned}$$

注意到对每个编码 $\mathcal{C}$, 我们可以将 $\mathcal{C}$ 的第 1 行和第 i 行对换, 得到一个新码字 $\mathcal{C}'$. 这时 $P(\mathcal{C}) = P(\mathcal{C}')$, 而 $\lambda_1(\mathcal{C}) = \lambda_w(\mathcal{C}')$, 对 $\mathcal{C}$ 求和, 则 $\mathcal{C}'$ 也遍历所有码簿, 从而

$$\sum_{\mathcal{C}} P(\mathcal{C}) \lambda_1(\mathcal{C}) = \sum_{\mathcal{C}'} P(\mathcal{C}') \lambda_w(\mathcal{C}').$$

从而我们有

$$P(\mathcal{E}) = \sum_{\mathcal{C}} P(\mathcal{C}) \lambda_1(\mathcal{C}) = P(\mathcal{E} | W = 1).$$

定义下列事件:

$$E_i = \{(X^n(i), Y^n) \text{ 在 } A_\epsilon^{(n)} \text{ 中}\}, i \in \{1, 2, \dots, 2^{nR}\}.$$

E_i 表示第 i 个码字与 Y^n 为联合典型的这一事件.

如果 E_1^C 发生, 或者 $E_2 \cup E_3 \cup \dots \cup E_{2^{nR}}$ 发生, 则译码会发生错误. 反过来, 若译码错误, 则 $E_1^C, E_2, E_3, \dots, E_{2^{nR}}$ 中必有一个发生. 于是

$$\begin{aligned} P(\mathcal{E}) &= P(\mathcal{E} | W = 1) \\ &= P(E_1^C \cup E_2 \cup E_3 \cup \dots \cup E_{2^{nR}} | W = 1) \\ &= P(E_1^C | W = 1) + \sum_{i=2}^{2^{nR}} P(E_i | W = 1). \end{aligned}$$

由联合 AEP 的性质, 我们知当 n 充分大时,

$$P(E_1^C | W = 1) \leq \epsilon.$$

从编码的过程我们知 $X^n(1)$ 和 $X^n(i) (i \neq 1)$ 是独立的, 所以 Y^n 与 $X^n(i)$ 也是独立的. 因此, 根据联合 AEP 的性质, $X^n(i)$ 与 Y^n 是联合典型的概率 $\leq 2^{-n(I(X;Y)-3\epsilon)}$. 从而, 如果 n 充分大且 $R < I(X;Y) - 3\epsilon$ 时,

$$\begin{aligned} P(\mathcal{E}) &= P(\mathcal{E} | W = 1) \leq P(E_1^C | W = 1) + \sum_{i=2}^{2^{nR}} P(E_i | W = 1) \\ &\leq \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \\ &= \epsilon + (2^{nR} - 1) 2^{-n(I(X;Y)-3\epsilon)} \\ &\leq \epsilon + 2^{3n\epsilon} 2^{-n(I(X;Y)-R)} \\ &\leq 2\epsilon. \end{aligned}$$

因此, 如果 $R < I(X;Y)$, 对于任意 $\epsilon > 0$, 可以适当选取 n , 使得平均误差概率不超过 2ϵ .

为完成证明, 我们需要进行下面一系列操作:

- 我们将证明中的 $p(x)$ 换成 $p^*(x)$, 即达到信道容量时关于 X 的分布. 此时, 条件 $R < I(X; Y)$ 可以由可达性条件 $R < C$ 所替代.
- 去除码簿上的平均. 由于所有的码簿上的平均误差概率比较小, 所以存在一个码簿 $\mathcal{C}^*$ 具有较小的平均误差概率. 于是 $P_e^{(n)}(\mathcal{E}|\mathcal{C}^*) \leq 2\epsilon$. 同时我们注意到

$$P(\mathcal{E}|\mathcal{C}^*) = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i(\mathcal{C}^*).$$

- 抛弃最佳码簿 $\mathcal{C}^*$ 中最差的一半码字. 由于这个码字的算术平均误差概率 $P_e^n(\mathcal{C}^*)$ 小于 2ϵ , 我们有

$$P(\mathcal{E}|\mathcal{C}^*) \leq \frac{1}{2^{nR}} \sum \lambda_i(\mathcal{C}^*) \leq 2\epsilon.$$

这说明至少有一半的下标 i 及其对应的码字 $X^n(i)$ 的条件误差概率 λ_i 小于 4ϵ (否则, 这些码字本身的和就将大于 2ϵ). 因此, 所有码字中最佳的一半的最大误差概率必定小于 4ϵ . 码字总数有 2^{nR-1} 个. 抛弃一半码字使得码率由 R 变为 $R - \frac{1}{n}$, 当 n 充分大时, 这是可以忽略的.

结合所有这些改进, 我们构造了一个码率为 $R' = R - \frac{1}{n}$ 的码, 它的最大误差概率 $\lambda^{(n)} \leq 4\epsilon$. 这就证明了任何小于信道容量的码率是可达的. $\square$

我们首先来证明 $P_e^{(n)} = 0$ 蕴含结论 $R \leq C$. 假定有一个零误差概率的 $(2^{nR}, n)$ 码, 也就是说, 译码器输出的 $g(Y^n)$ 依概率 1 等于输入的下标 W . 那么, 输入下标 W 完全由输出序列决定 (即 $H(W|Y^n) = 0$). 为了获得更强的界, 随意假定 W 服从 $\{1, 2, \dots, 2^{nR}\}$ 上的均匀分布, 于是, $H(W) = nR$. 从而我们有如下的一串不等式:

$$\begin{aligned} nR = H(W) &= H(W|Y^n) + I(W; Y^n) \\ &= I(W; Y^n) \\ &\leq I(X^n, Y^n) \\ &\leq \sum_{i=1}^n I(X_i; Y_i) \\ &\leq nC. \end{aligned}$$

其中第一个不等式是由数据处理不等式可得, 第二个不等式和第三个不等式我们将在下一讲中给出证明. 因此, 对于任何零误差的 $(2^{nR}, n)$ 码以及所有的 n ,

$$R \leq C.$$

第 15 讲 微分熵

15.1 定义

定义 15.1

设 X 是一个随机变量，其累积分布函数为 $F(x) = P(X \leq x)$. 如果 $F(x)$ 是连续的，则称该随机变量是连续的. 当 $F(x)$ 的导数存在时，令 $f(x) = F'(x)$. 若 $\int_{-\infty}^{\infty} f(x)dx = 1$ ，则称 $f(x)$ 是 X 的概率密度函数. 另外，使 $f(x) > 0$ 的所有 x 构成的集合称为 X 的支撑集.

定义 15.2

一个以 $f(x)$ 为密度函数的连续型随机变量 X 的微分熵 $h(X)$ 定义为

$$h(X) = - \int_S f(x) \log f(x) dx.$$

其中 S 是这个随机变量的支撑集.

例 15.3: 均匀分布

考虑一个服从 $[0, a]$ 上的均匀分布的随机变量，它的密度函数在 $[0, a]$ 熵为 $1/a$ ，而在其他地方为 0. 此时，该随机变量的微分熵为

$$h(X) = - \int_0^a \frac{1}{a} \log \frac{1}{a} dx = \log a.$$

例 15.4: 正态分布

设 $X \sim \phi(x) = (1/\sqrt{2\pi\sigma^2})e^{-x^2/2\sigma^2}$. 如果我们以奈特为单位计算微分熵，我们有

$$\begin{aligned} h(\phi) &= - \int \phi \ln \phi \\ &= - \int \phi(x) \left[-\frac{x^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2} \right] \\ &= \frac{EX^2}{2\sigma^2} + \frac{1}{2} \ln 2\pi\sigma^2 \\ &= \frac{1}{2} + \frac{1}{2} \ln 2\pi\sigma^2 \\ &= \frac{1}{2} \ln 2\pi e\sigma^2 \text{ 奈特.} \end{aligned}$$

改变对数的底，我们可得

$$h(\phi) = \frac{1}{2} \log 2\pi e\sigma^2 \text{ 比特.}$$

15.2 连续随机变量的 AEP

定理 15.5

设 $X_1, X_2, \dots, X_n$ 使一个服从于密度函数 $f(x)$ 的独立同分布的随机变量序列. 那么下面的极限依概率收敛.

$$-\frac{1}{n} \log f(X_1, X_2, \dots, X_n) \rightarrow E[-\log f(X)] = h(X) \text{ 依概率.}$$

定义 15.6

对 $\epsilon > 0$ 及任意的 n , 定义 $f(x)$ 的典型集 $A_\epsilon^{(n)}$ 如下:

$$A_\epsilon^{(n)} = \{(x_1, x_2, \dots, x_n) \in S^n : |-\frac{1}{n} \log f(X_1, x_2, \dots, x_n) - h(X)| \leq \epsilon\}.$$

其中 $f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i)$.

定义 15.7

集合 $A \subset \mathbb{R}^n$ 的体积 $\text{Vol}(A)$ 定义为

$$\text{Vol}(A) = \int_A dx_1 dx_2 \cdots dx_n.$$

定理 15.8

典型集 $A_\epsilon^{(n)}$ 有如下的性质:

1. 对于充分大的 n , $P(A_\epsilon^{(n)}) > 1 - \epsilon$.
2. 对于所有的 n , $\text{Vol}(A_\epsilon^{(n)}) \leq 2^{n(h(X)+\epsilon)}$.
3. 对于充分大的 n , $\text{Vol}(A_\epsilon^{(n)}) \geq (1 - \epsilon)2^{n(h(X)-\epsilon)}$.

证明. 根据 AEP, 我们有依概率有 $-\frac{1}{n} \log f(X^n) = -\frac{1}{n} \sum \log f(x_i) \rightarrow h(X)$, 故性质 1 得证. 另外,

$$\begin{aligned} 1 &= \int_{S^n} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_n \\ &\geq \int_{A_\epsilon^{(n)}} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_n \\ &\geq \int_{A_\epsilon^{(n)}} 2^{-n(h(X)+\epsilon)} dx_1 dx_2 \cdots dx_n \\ &= 2^{-n(h(X)+\epsilon)} \int_{A_\epsilon^{(n)}} dx_1 dx_2 \cdots dx_n \\ &= 2^{-n(h(X)+\epsilon)} \text{Vol}(A_\epsilon^{(n)}). \end{aligned}$$

因此, 性质 2 获证.

我们进一步论证该典型集的体积至少有这么大. 如果 n 足够大使得性质 1 成立, 那么

$$\begin{aligned}
 1 - \epsilon &\leq \int_{A_\epsilon^{(n)}} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_n \\
 &\leq \int_{A_\epsilon^{(n)}} 2^{-n(h(X)-\epsilon)} dx_1 dx_2 \cdots dx_n \\
 &= 2^{-n(h(X)-\epsilon)} \int_{A_\epsilon^{(n)}} dx_1 dx_2 \cdots dx_n \\
 &= 2^{-n(h(X)-\epsilon)} \text{Vol}(A_\epsilon^{(n)}).
 \end{aligned}$$

从而性质 3 成立. 从而当 n 充分大时, 我们有

$$(1 - \epsilon) 2^{n(h(X)-\epsilon)} \leq \text{Vol}(A_\epsilon^{(n)}) \leq 2^{n(h(X)+\epsilon)}.$$

□

定理 15.9

在一阶指数意义下, 在所有概率 $\geq 1 - \epsilon$ 的集合中, $A_\epsilon^{(n)}$ 为使体积最小者.

15.3 微分熵与离散熵的关系

考虑一个密度函数是 $f(x)$ 的随机变量 X . 假定将 X 的定义域等长度地分割成长为 Δ 的若干小区间, 并且假定密度函数在这些小区间内是连续的. 由中值定理可知, 在每个小区间内存在一个值 x_i , 满足

$$f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)dx.$$

考虑量化后的随机变量 X^Δ , 其定义是

$$X^\Delta = x_i \text{ 当 } i\Delta \leq X < (i+1)\Delta.$$

则 $X^\Delta = x_i$ 的概率为

$$p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = f(x_i)\Delta.$$

由于 $\sum f(x_i)\Delta = \int f(x) = 1$, 所以量化后的随机变量 X^Δ 的熵为

$$\begin{aligned}
 H(x^\Delta) &= - \sum_{i=-\infty}^{\infty} p_i \log p_i \\
 &= - \sum_{i=-\infty}^{\infty} f(x_i)\Delta \log(f(x_i)\Delta) \\
 &= - \sum \Delta f(x_i) \log f(x_i) - \sum f(x_i)\Delta \log \Delta \\
 &= - \sum \Delta f(x_i) \log f(x_i) - \log \Delta
 \end{aligned}$$

如果 $f(x) \log x$ 是黎曼可积的, 则我们可以得到下面的定理.

定理 15.10

如果随机变量 X 的密度函数 $f(x)$ 是黎曼可积的, 那么

$$H(X^\Delta) + \log \Delta \rightarrow h(f) = h(X), \text{ 当 } \Delta \rightarrow 0.$$

于是, 连续随机变量 X 经过 n 比特量化处理 (此时分割的小区间长度为 $\frac{1}{2^n}$) 后的熵大约为 $h(X) + n$.

例 15.11

如果 X 服从 $[0, \frac{1}{8}]$ 上的均匀分布, 那么在二进制表示中, X 取值的小数点右边取值必定为 0. 因而, 在精确到 n 位的意义下, 描述 X 仅需 $n - 3$ 比特, 这与 $h(X) = -3$ 相一致.

一般来讲, 在精确到 n 位的意义下, $h(X) + n$ 是为了描述 X 所需要的平均比特数.

15.4 联合微分熵与条件微分熵**定义 15.12**

联合密度函数为 $f(x_1, x_2, \dots, x_n)$ 的一组随机变量 $X_1, X_2, \dots, X_n$ 的联合微分熵定义为

$$h(X_1, X_2, \dots, X_n) = - \int f(x^n) \log f(x^n) dx^n.$$

定义 15.13

如果 X, Y 的联合密度函数为 $f(x, y)$, 定义条件微分熵 $h(X|Y)$ 为

$$h(X|Y) = - \int f(x, y) \log f(x|y) dx dy.$$

由于一般地 $f(x|y) = f(x, y)/f(y)$, 我们可以写为

$$h(X|Y) = h(X, Y) - h(Y).$$

我们要注意微分熵可能为无穷的情况, 需要分开讨论.

定理 15.14: 多元正态分布的熵

设 $X_1, X_2, \dots, X_n$ 满足多元正态分布, 平均值为 μ , 协方差矩阵为 K . 则

$$h(X_1, X_2, \dots, X_n) = h(\mathcal{N}_n(\mu, K)) = \frac{1}{2} \log(2\pi e)^n |K| \text{ 比特},$$

其中 $|K|$ 记为 K 的行列式.

证明. $X_1, X_2, \dots, X_n$ 的概率密度函数为 $f(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^n |K|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu)}$. 则

$$\begin{aligned}
 h(f) &= - \int f(\mathbf{x}) \left[-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu) - \ln(\sqrt{2\pi})^n |K|^{\frac{1}{2}} \right] d\mathbf{x}. \\
 &= \frac{1}{2} E \left[\sum_{i,j} (X_i - \mu_i)(K^{-1})_{ij}(X_j - \mu_j) \right] + \frac{1}{2} \ln(2\pi)^n |K| \\
 &= \frac{1}{2} E \left[\sum_{i,j} (X_i - \mu_i)(X_j - \mu_j)(K^{-1})_{ij} \right] + \frac{1}{2} \ln(2\pi)^n |K| \\
 &= \frac{1}{2} \sum_{i,j} E[(X_i - \mu_i)(X_j - \mu_j)](K^{-1})_{ij} + \frac{1}{2} \ln(2\pi)^n |K| \\
 &= \frac{1}{2} \sum_{i,j} K_{ji}(K^{-1})_{ij} + \frac{1}{2} \ln(2\pi)^n |K| \\
 &= \frac{1}{2} \sum_j I_{jj} + \frac{1}{2} \ln(2\pi)^n |K| \\
 &= \frac{n}{2} + \frac{1}{2} \ln(2\pi)^n |K| \\
 &= \frac{1}{2} \ln(2\pi e)^n |K| \text{ 奈特} \\
 &= \frac{1}{2} \log(2\pi e)^n |K| \text{ 比特}. \quad \square
 \end{aligned}$$

15.5 相对熵与互信息

定义 15.15

两个概率密度函数 f 和 g 的相对熵 (或 Kullback-Leibler 距离) $D(f||g)$ 定义为

$$D(f||g) = \int f \log \frac{f}{g}.$$

注意到 $D(f||g)$ 有限当且仅当 f 的支集包含在 g 的支集中. (和以前一样我们默认 $0 \log \frac{0}{0} = 0$.)

定义 15.16

两个联合分布为 $f(x, y)$ 的随机变量 X 和 Y 的互信息定义为

$$I(X; Y) = \int f(x, y) \log \frac{f(x, y)}{f(x)f(y)} dx dy.$$

由定义我们知和离散的情形类似,

$$I(X; Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) = h(X) + h(Y) - h(X, Y)$$

以及

$$I(X; Y) = D(f(x, y) || f(x)f(y)).$$

设 $\mathcal{X}$ 为随机变量 X 的值域, $\mathcal{P}$ 为 $\mathcal{X}$ 的一个分割是指存在有限个不相交的集合 P_i 使得 $\cup_i P_i = \mathcal{X}$. X 关于 $\mathcal{P}$ 的量化 (记为 $[X]_{\mathcal{P}}$) 是定义如下的离散随机变量:

$$P([X]_{\mathcal{P}} = i) = P(X \in P_i) = \int_{P_i} dF(x).$$

对于任意两个分割分别为 $\mathcal{P}$ 和 $\mathcal{Q}$ 的随机变量 X 和 Y , 可以利用 (2.5) 来计算它们对应的量化随机变量的互信息. 于是, 对任意成对的随机变量, 我们可以如下定义其互信息:

定义 15.17

任何随机变量 X 与 Y 间的互信息如下:

$$I(X; Y) = \sup_{\mathcal{P}, \mathcal{Q}} I([X]_{\mathcal{P}}; [Y]_{\mathcal{Q}}),$$

其中上确界遍历所有可能的有限剖分 $\mathcal{P}$ 和 $\mathcal{Q}$.

15.6 微分熵的性质, 相对熵和互信息的性质

定理 15.18

我们有

$$D(f||g) \geq 0.$$

等号成立当且仅当 $f = g$, a.e..

证明. 设 f 的支撑集为 S . 则

$$\begin{aligned} -D(f||g) &= \int_S f \log \frac{g}{f} \\ &\leq \log \int_S f \frac{g}{f} \\ &= \log \int_S g \\ &\leq \log 1 = 0 \end{aligned}$$

当且仅当 Jensen 不等式中的等号成立, 即当且仅当 $f = g$ a.e. 等号成立. □

推论 15.19

$I(X; Y) \geq 0$, 等号成立当且仅当 X 和 Y 独立.

定理 15.20: 微分熵的链式法则

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

推论 15.21

$$h(X_1, X_2, \dots, X_n) \leq \sum h(X_i),$$

等号成立当且仅当 $X_1, X_2, \dots, X_n$ 为独立的.

推论 15.22: 阿达玛不等式

设 $\mathbf{x} \sim \mathcal{N}(0, K)$ 是一个多元正态分布, 那么将熵的定义公式代入上面的不等式中, 我们可以得到

$$|K| \leq \prod_{i=1}^n K_{ii}.$$

定理 15.23

$$h(X + c) = h(X).$$

定理 15.24

$$h(aX) = h(X) + \log |a|.$$

证明. 令 $Y = aX$. 则 $f_Y(y) = \frac{1}{|a|} f_X(\frac{y}{a})$, 且经过积分变量替换, 有

$$\begin{aligned} h(aX) &= - \int f_Y(y) \log f_Y(y) dy \\ &= - \int \frac{1}{|a|} f_X\left(\frac{y}{a}\right) \log\left(\frac{1}{|a|} f_X\left(\frac{y}{a}\right)\right) dy \\ &= - \int f_X(x) \log f_X(x) dx + \log |a| \\ &= h(X) + \log |a| \end{aligned}$$

从而得证. □

类似地, 对于取值为向量的随机变量, 可以证明下面的推论.

推论 15.25

$$h(A\mathbf{X}) = h(\mathbf{X}) + \log |\det(A)|.$$

定理 15.26

设随机向量 $\mathbf{X} \in \mathbb{R}^n$ 的均值为零, 协方差矩阵为 $K = E\mathbf{X}\mathbf{X}^t$, 则 $h(\mathbf{X}) \leq \frac{1}{2} \log(2\pi e)^n |K|$, 当且仅当 $\mathbf{X} \sim \mathcal{N}(0, K)$ 等号成立.

证明. 设 $g(\mathbf{x})$ 是对任意的 i 和 j 均满足 $\int g(\mathbf{x}) x_i x_j d\mathbf{x} = K_{ij}$ 的密度函数. 令 ϕ_K 是服从高斯分布 $\mathcal{N}(0, K)$ 随机向量的密度函数. 注意到 $\log \phi_K(\mathbf{x})$ 是一次二次型, 并且 $\int x_i x_j \phi_K(\mathbf{x}) d\mathbf{x} = K_{ij}$, 则

$$\begin{aligned} 0 &\leq D(g \parallel \phi_K) \\ &= \int g \log(g/\phi_K) \\ &= -h(g) - \int g \log \phi_K \\ &= -h(g) - \int \phi_K \log \phi_K \\ &= -h(g) + h(\phi_K) \end{aligned}$$

其中所作的替换 $\int g \log \phi_K = \int \phi_K \log \phi_K$ 是由于二次型 $\log \phi_K(\mathbf{x})$ 关于 g 和 ϕ_K 具有相同的矩. □

定理 15.27: 估计误差与微分熵

对任意随机变量 X 及其估计 $\hat{X}$,

$$E(X - \hat{X})^2 \geq \frac{1}{2\pi e} e^{2h(X)}.$$

其中等号成立的充分必要条件是 X 为高斯分布而 $\hat{X}$ 为其均值. 这里 $h(X)$ 的量纲为比特.

证明. 令 $\hat{X}$ 为 X 的一个估计, 此时

$$\begin{aligned} E(X - \hat{X})^2 &\geq \min_{\hat{X}} E(X - \hat{X})^2 \\ &= E(X - E(X))^2 \\ &= \text{var}(X) \\ &\geq \frac{1}{2\pi e} e^{2h(X)} \end{aligned}$$

不等式成立当且仅当 X 为高斯分布而 $\hat{X}$ 为其均值. □

推论 15.28

当边信息 Y 以及估计 $\hat{X}(Y)$ 已知时, 可以推出

$$E(X - \hat{X}(Y))^2 \geq \frac{1}{2\pi e} e^{2h(X|Y)}.$$

第 16 讲 高斯信道

16.1 高斯信道

16.1.1 定义

高斯信道是一个时间离散信道，在时刻 i ，输出信号 Y_i 是输入信号 X_i 与噪声 Z_i 之和，其中 Z_i 为独立同分布序列，且均服从方差为 N 的高斯分布。于是，

$$Y_i = X_i + Z_i, \quad Z_i \sim \mathcal{N}(0, N)$$

假设噪声 Z_i 与信号 X_i 独立。

对于输入最通常的限制是在能量或者功率方面的约束。我们一般考虑对平均功率的约束，即对于在信道上传输的任意码字 $(x_1, x_2, \dots, x_n)$ ，我们要求 $\frac{1}{n} \sum_{i=1}^n x_i^2 \leq P$ 。

定义 16.1

功率限制为 P 的高斯信道的信息容量为

$$C = \max_{EX^2 \leq P} I(X; Y).$$

我们来看看如何计算信道容量。将 $I(X; Y)$ 展开，由于 Z 与 X 相互独立，由微分熵的平移不变性，我们可得

$$\begin{aligned} I(X; Y) &= h(Y) - h(Y|X) \\ &= h(Y) - h(X + Z|X) \\ &= h(Y) - h(Z|X) \\ &= h(Y) - h(Z) \end{aligned}$$

此时， $h(Z) = \frac{1}{2} \log 2\pi e N$ 。又由于 X 与 Z 独立以及 $EZ = 0$ ，所以

$$EY^2 = E(X + Z)^2 = EX^2 + 2EXEZ + EZ^2 = P + N$$

假设给定 $EY^2 = P + N$ ，则我们知， Y 的熵的上界为 $\frac{1}{2} \log 2\pi e(P + N)$ 。于是

$$\begin{aligned} I(X; Y) &= h(Y) - h(Z) \\ &\leq \frac{1}{2} \log 2\pi e(P + N) - \frac{1}{2} \log 2\pi e N \\ &= \frac{1}{2} \log \left(1 + \frac{P}{N}\right). \end{aligned}$$

因此，高斯信道的信道容量为

$$C = \max_{EX^2 \leq P} I(X; Y) = \frac{1}{2} \log \left(1 + \frac{P}{N}\right),$$

并且最大值在 $X \sim \mathcal{N}(0, P)$ 时达到。

16.1.2 高斯信道编码定理

一个功率限制为 P 的高斯信道所对应的 (M, n) 码由以下几个要素构成：

1. 下标集 $\{1, 2, 3, \dots, M\}$.

2. 编码函数 $x: \{1, 2, \dots, M\} \rightarrow \mathcal{X}^n$, 其相应的码字为 $x^n(1), x^n(2), \dots, x^n(M)$, 且满足功率限制 P , 即

$$\sum_{i=1}^n x_i^2(\omega) \leq nP, \omega = 1, 2, \dots, M.$$

3. 译码函数

$$g: \mathcal{Y}^n \rightarrow \{1, 2, \dots, M\}$$

该编码的码率和误差概率与前面我们离散情形时的定义相同.

定义 16.2

对于一个功率限制为 P 的高斯信道, 如果存在一个 $(2^{nR}, n)$ 码, 满足功率限制条件, 使得最大误差概率 $\lambda^n \rightarrow 0$, 则称码率 R 关于该功率限制为 P 的高斯信道是可达的.

定理 16.3

一个功率限制为 P 且噪声方差为 N 的高斯信道的容量为

$$C = \frac{1}{2} \log(1 + \frac{P}{N}) \text{ 比特/传输}$$

它是所有可达码率的上界.

16.2 几种常见的高斯信道

16.2.1 并联高斯信道

假设有一组并联高斯信道, 每个信道的输出是输入与高斯噪声之和. 对于信道 j ,

$$Y_j = X_j + Z_j, j = 1, 2, \dots, k,$$

其中

$$Z_j \sim \mathcal{N}(0, N_j).$$

并且我们假设噪声在信道与信道之间是相互独立的. 假定在所使用的总功率方面存在一个公共的功率限制, 即,

$$E \sum_{j=1}^k X_j^2 \leq P.$$

我们希望将功率分配于各信道之中以使总容量达到最大.

信道的信息容量 C 为

$$C = \max_{f(x_1, x_2, \dots, x_k): \sum E X_i^2 \leq P} I(X_1, X_2, \dots, X_k; Y_1, Y_2, \dots, Y_k)$$

注记 16.4

信息容量是所有可达码率的上确界, 这一事实的证明与单个高斯信道的容量定理的证明方法相同, 故我们略去.

由于 $Z_1, Z_2, \dots, Z_k$ 是相互独立的,

$$\begin{aligned}
 & I(X_1, X_2, \dots, X_k; Y_1, Y_2, \dots, Y_k) \\
 &= h(Y_1, Y_2, \dots, Y_k) - h(Y_1, Y_2, \dots, Y_k | X_1, X_2, \dots, X_k) \\
 &= h(Y_1, Y_2, \dots, Y_k) - h(Z_1, Z_2, \dots, Z_k | X_1, X_2, \dots, X_k) \\
 &= h(Y_1, Y_2, \dots, Y_k) - h(Z_1, Z_2, \dots, Z_k) \\
 &= h(Y_1, Y_2, \dots, Y_k) - \sum_i h(Z_i) \\
 &\leq \sum_i (h(Y_i) - h(Z_i)) \\
 &\leq \sum_i \frac{1}{2} \log(1 + \frac{P_i}{N_i}).
 \end{aligned}$$

其中 $P_i = EX_i^2$, $\sum P_i = P$. 等号在如下条件达到时成立

$$(X_1, X_2, \dots, X_k) \sim \mathcal{N}(0, P), \quad P = \begin{pmatrix} P_1 & 0 & \dots & 0 \\ 0 & P_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & P_k \end{pmatrix}.$$

于是, 我们问题就简化为在满足约束条件 $\sum P_i = P$ 下, 寻找一个功率分配方法使得容量达到最大, 这是一个标准的优化问题, 可以利用拉格朗日乘子法得到解决. 相应的函数为

$$J(P_1, \dots, P_k) = \sum \frac{1}{2} \log(1 + \frac{P_i}{N_i}) + \lambda(\sum P_i - P).$$

对 P_i 求导, 我们有

$$\frac{1}{2} \frac{1}{P_i + N_i} + \lambda = 0,$$

于是

$$P_i = v - N_i.$$

然而, 由于 P_i 必须非负, 所以, 并不能总找到一个如此形式的解. 这样, 可利用库恩-塔克条件来验证如下解

$$P_i = (v - N_i)^+$$

是使得容量达到最大的分配方法, 其中 v 的选取满足

$$\sum (v - N_i)^+ = P.$$

这里 $(x)^+$ 表示对 x 取正的部分:

$$(x)^+ = \begin{cases} x & \text{若 } x \geq 0, \\ 0 & \text{若 } x < 0. \end{cases}$$

16.2.2 高斯彩色信道

设 K_Z 为噪声的协方差阵, K_X 为输入信号的协方差阵. 那么, 对于输入信号的功率限制可以写为

$$\frac{1}{n} \sum_i EX_i^2 \leq P,$$

或等价地

$$\frac{1}{n} \text{tr}(K_X) \leq P.$$

这里的功率限制依赖于 n , 因此, 我们不得不对每个 n 单独计算容量.

与独立信道情形相同, 我们有

$$I(X_1, X_2, \dots, X_n; Y_1, Y_2, \dots, Y_n) = h(Y_1, Y_2, \dots, Y_n) - h(Z_1, Z_2, \dots, Z_n)$$

这里 $h(Z_1, \dots, Z_n)$ 由噪声分布唯一确定, 而不依赖于输入信号分布的选择. 所以, 计算信道容量等价于将 $h(Y_1, \dots, Y_n)$ 最大化. 由于输入信号和噪声是相互独立的, 所以, 输出 Y 的协方差阵为 $K_Y = K_X + K_Z$, 这时由我们前面的结果知当 Y 服从正态分布时输出信号的熵达到最大, 这时输入分布也是一个正态分布. 我们可以计算得

$$h(Y_1, \dots, Y_n) = \frac{1}{2} \log((2\pi e)^n |K_X + K_Z|).$$

于是, 问题就化为在 K_X 的迹约束条件下, 选取 K_X 使得 $|K_X + K_Z|$ 达到最大. 为达此目的, 将 K_Z 分解成对角型,

$$K_Z = Q\Lambda Q^t, \text{ 其中 } QQ^t = I.$$

那么,

$$\begin{aligned} |K_X + K_Z| &= |K_X + Q\Lambda Q^t| \\ &= |Q||Q^t K_X Q + \Lambda||Q^t| \\ &= |Q^t K_X Q + \Lambda| \\ &= |A + \Lambda|, \end{aligned}$$

其中 $A = Q^t K_X Q$. 于是我们有,

$$\begin{aligned} \text{tr}(A) &= \text{tr}(Q^t K_X Q) \\ &= \text{tr}(Q^t Q K_X) \\ &= \text{tr}(K_X). \end{aligned}$$

于是问题就简化为在条件 $\text{tr}(A) \leq nP$ 下, 求 $|A + \Lambda|$ 的最大值.

我们利用前面讲到的阿达马不等式, 此不等式说明任意正定矩阵 K 得行列式一定小于它的对角元素得乘积, 即

$$|K| \leq \prod_i K_{ii},$$

等号成立当且仅当 K 为对角阵. 于是,

$$|A + \Lambda| \leq \prod_i (A_{ii} + \lambda_i),$$

当且仅当 A 为对角阵时等号成立.

由于 A 受到迹的约束,

$$\frac{1}{n} \sum_i A_{ii} \leq P,$$

且 $A_{ii} \geq 0$, 所以 $\prod_i (A_{ii} + \lambda_i)$ 的最大值在

$$A_{ii} + \lambda_i = v$$

时达到. 然而, 考虑到约束条件, 不可能总是存在正的 A_{ii} 满足上述方程. 和上一节的讨论类似, 在不满足的情况下根据库恩-塔克条件可以证明最优解对应于取

$$A_{ii} = (v - \lambda_i)^+$$

时的解. 其中 v 满足 $\sum A_{ii} = nP$. 此时 A 的值使 Y 的熵达到最大, 因此, 互信息达到最大.

第 17 讲 率失真理论

17.1 率失真函数

假设某信源产生序列 $X_1, X_2, \dots, X_n$ 是 i.i.d. $\sim p(x)$, $x \in \mathcal{X}$. 在本章的证明中, 假设字母表是有限的, 但大多数离散情形下的证明都可以推广到连续的随机变量. 信源序列 X^n 的编码用下标 $f_n(X^n) \in \{1, 2, \dots, 2^{nR}\}$ 表示, X^n 的译码用估计形式 $\hat{X}^n \in \hat{\mathcal{X}}$ 表示.

定义 17.1

失真函数或者失真度量是指从信源字母表与再生字母表的乘积空间到非负实数集上的映射

$$d: \mathcal{X} \times \hat{\mathcal{X}} \rightarrow \mathcal{R}^+.$$

失真 $d(x, \hat{x})$ 是用来刻画使用 $\hat{x}$ 表示 x 时的代价度量.

定义 17.2

称失真度量是有界的, 如果失真的最大值有限:

$$d_{\max} := \max_{x \in \mathcal{X}, \hat{x} \in \hat{\mathcal{X}}} d(x, \hat{x}) < \infty.$$

定义 17.3

x^n 与 $\hat{x}^n$ 序列间的失真定义为

$$d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^n d(x_i, \hat{x}_i).$$

定义 17.4

一个 $(2^{nR}, n)$ 率失真码包括一个编码函数

$$f_n: \mathcal{X}^n \rightarrow \{1, 2, \dots, 2^{nR}\}$$

和一个译码 (再生) 函数

$$g_n: \{1, 2, \dots, 2^{nR}\} \rightarrow \hat{\mathcal{X}}^n.$$

关于这个 $(2^{nR}, n)$ 码的失真定义为

$$D = E[d(X^n, g_n(f_n(X^n)))] = \sum_{x^n} p(x^n) d(x^n, g(f_n(x^n))).$$

将 n 元组 $g_n(1), g_n(2), \dots, g_n(2^{nR})$ 记为 $\hat{X}^n(1), \hat{X}^n(2), \dots, \hat{X}^n(2^{nR})$, 它构成一个码簿, 且 $f_n^{-1}(1), f_n^{-1}(2), \dots, f_n^{-1}(2^{nR})$ 为相应的分配区域.

定义 17.5

称率失真对 (R, D) 是可达的, 若存在一个 $(2^{nR}, n)$ 率失真码序列 (f_n, g_n) , 满足

$$\lim_{n \rightarrow \infty} E[d(X^n, g_n(f_n(X^n)))] \leq D.$$

定义 17.6

全体可达率失真对 (R, D) 所组成的集合的闭包称为信源的率失真区域.

定义 17.7

对于给定的失真 D , 满足 (R, D) 包含于信源的率失真区域中的所有码率 R 的下确界称为是率失真函数 $R(D)$.

定义 17.8

对于给定的码率 R , 满足 (R, D) 包含于信源的率失真区域中的所有失真 D 的下确界称为是失真率函数 $D(R)$.

定义 17.9

设信源 X 的失真度量为 $d(x, \hat{x})$, 定义其信息率失真函数 $R^{(I)}(D)$ 为

$$R^{(I)}(D) = \min_{p(\hat{x}|x): \sum_{(x, \hat{x})} p(x)p(\hat{x}|x)d(x, \hat{x}) \leq D} I(X, \hat{X}).$$

其中的最小值取自使联合分布 $p(x, \hat{x}) = p(x)p(\hat{x}|x)$ 满足期望失真限制的所有条件分布 $p(\hat{x}|x)$.

定理 17.10

对于独立同分布的信源 X , 若公共分布为 $p(x)$ 且失真函数 $d(x, \hat{x})$ 有界, 那么其率失真函数与对应的信息率失真函数相等. 于是,

$$R(D) = R^{(I)}(D) = \min_{p(\hat{x}|x): \sum_{(x, \hat{x})} p(x)p(\hat{x}|x)d(x, \hat{x}) \leq D} I(X, \hat{X})$$

为在失真 D 下的最小可达码率.

17.2 率失真函数的计算**17.2.1 二元信源****定理 17.11**

Bernoulli(p) 信源在汉明失真度量下的率失真函数为

$$R(D) = \begin{cases} H(p) - H(D), & 0 \leq D \leq \{p, 1-p\} \\ 0, & D > \min\{p, 1-p\}. \end{cases}$$

证明. 考虑在汉明失真度量下的二元信源 $X \sim \text{Bernoulli}(p)$. 不失一般性, 假设 $p < \frac{1}{2}$. 我们来计算量率失真函数

$$R(D) = \min_{p(\hat{x}|x): \sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X, \hat{X}).$$

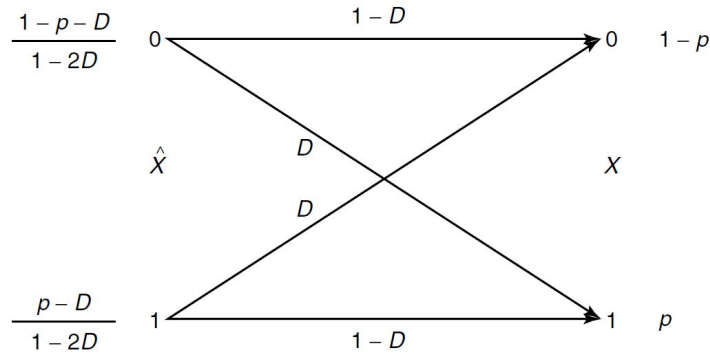
用 $\oplus_2$ 表示模 2 加法运算, 则 $X \oplus_2 \hat{X} = 1$ 等价于 $X \neq \hat{X}$. 我们无法直接最小化 $I(X; \hat{X})$, 而且先获得率失真函数的一个下界, 然后证明这个下界是可达的. 对于任何一个满足失真限制的联合分布, 我们有

$$\begin{aligned} I(X; \hat{X}) &= H(X) - H(X|\hat{X}) \\ &= H(p) - H(X \oplus \hat{X}|\hat{X}) \\ &\geq H(p) - H(X \oplus \hat{X}) \\ &\geq H(p) - H(D). \end{aligned}$$

由于 $P(X \neq \hat{X}) \leq D$ 且 $H(D)$ 在 $D \leq 1/2$ 时单增的. 于是,

$$R(D) \geq H(p) - H(D).$$

我们下面说明, 若能找到一个满足失真限制且有 $I(X; \hat{X}) = R(D)$ 的联合分布, 则这个下界实际上是率失真函数. 由于 $0 \leq D \leq p$, 选取 $(X, \hat{X})$ 使其联合分布满足如下图所示的二元对称信道, 则可以达到上式中的率失真函数.



我们来解释如何得到上面所述的信道输入处 $\hat{X}$ 的分布, 使输出分布 X 服从上图中指定的分布. 令 $r = P(\hat{X} = 1)$, 并且对 r 的选取满足

$$r(1-D) + (1-r)D = p$$

或

$$r = \frac{p-D}{1-2D}.$$

若 $D \leq p \leq 1/2$, 则 $p(\hat{X} = 1) \geq 0$, 且 $P(\hat{X} = 0) \geq 0$. 于是我们有

$$I(X; \hat{X}) = H(X) - H(X|\hat{X}) = H(p) - H(D)$$

且期望失真为 $P(X \neq \hat{X}) = D$.

若 $D \geq p$, 则可通过令 $\hat{X} = 0$ 的概率为 1 而达到码率 $R(D) = 0$. 此时, $I(X; \hat{X}) = 0$, 且期望失真为 $D = p$. 同样地, 若 $D \geq 1-p$, 则可通过令 $\hat{X} = 1$ 的概率为 1 而达到码率 $R(D) = 0$. 因此, 二元信源的率失真函数为

$$R(D) = \begin{cases} H(p) - H(D), & 0 \leq D \leq \min\{p, 1-p\} \\ 0, & D > \min\{p, 1-p\}. \end{cases}$$

从而得证. □

17.2.2 高斯信源

定理 17.12

一个 $\mathcal{N}(0, \sigma^2)$ 信源在平方误差失真度量下的率失真函数为

$$R(D) = \begin{cases} \frac{1}{2} \log \frac{\sigma^2}{D}, & 0 \leq D \leq \sigma^2, \\ 0, & D > \sigma^2. \end{cases}$$

证明. 设 $X \sim \mathcal{N}(0, \sigma^2)$, 由推广到连续型字母表情形的率失真定理, 我们有

$$R(D) = \min_{f(\hat{x}|x): E(\hat{X} - X)^2 \leq D} I(X; \hat{X})$$

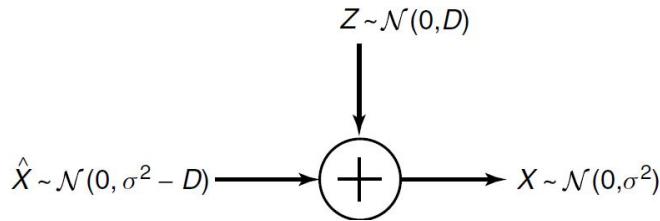
与前面的例子类似, 首先获得率失真函数的一个下界, 然后证明这个下界是可达的. 由于 $E(X - \hat{X})^2 \leq D$, 我们有

$$\begin{aligned} I(X; \hat{X}) &= h(X) - h(X|\hat{X}) \\ &= \frac{1}{2} \log(2\pi e)\sigma^2 - h(X - \hat{X}|\hat{X}) \\ &\geq \frac{1}{2} \log(2\pi e)\sigma^2 - h(X - \hat{X}) \\ &\geq \frac{1}{2} \log(2\pi e)\sigma^2 - h(\mathcal{N}(0, E(X - \hat{X})^2)) \\ &= \frac{1}{2} \log(2\pi e)\sigma^2 - \frac{1}{2} \log(2\pi e)E(X - \hat{X})^2 \\ &\geq \frac{1}{2} \log(2\pi e)\sigma^2 - \frac{1}{2} \log(2\pi e)D = \frac{1}{2} \log \frac{\sigma^2}{D}. \end{aligned}$$

因此,

$$R(D) \geq \frac{1}{2} \log \frac{\sigma^2}{D}. \quad (17.1)$$

为了求得达到这个下界时的条件密度 $f(\hat{x}|x)$, 通常更为简便的办法是着眼考虑条件密度函数 $f(x|\hat{x})$, 对此, 有时称作测试信道. 和二元信源情形一样, 我们希望去寻找使等号成立的 $f(x|\hat{x})$. 选取如下图所示的联合分布.



如果 $D \leq \sigma^2$, 取

$$X = \hat{X} + Z, \quad \hat{X} \sim \mathcal{N}(0, \sigma^2 - D), \quad Z \sim \mathcal{N}(0, D).$$

其中 $\hat{X}$ 与 Z 独立. 对于该联合分布, 计算可得

$$I(X; \hat{X}) = \frac{1}{2} \log \frac{\sigma^2}{D}$$

以及 $E(X - \hat{X})^2 = D$, 于是这个联合分布可以达到 (17.1) 中的下界. 如果 $D > \sigma^2$, 以概率 1 选取 $\hat{X} = 0$, 则由此可得 $R(D) = 0$. 因此, 高斯信源在平方误差失真下的率失真函数为

$$R(D) = \begin{cases} \frac{1}{2} \log \frac{\sigma^2}{D}, & 0 \leq D \leq \sigma^2, \\ 0, & D > \sigma^2. \end{cases}$$

□

17.2.3 独立高斯随机变量的同步描述

- 本小节考虑 m 个独立 (但服从不同的分布) 的正态随机信源 $X_1, \dots, X_m$ 的表示问题, 其中 $X \sim \mathcal{N}(0, \sigma_i^2)$, 失真为平方误差失真.
- 假设用 R 比特来表示这个随机向量.
- 自然有这样一个问题: 如何分配这些比特到各成员, 才能使总失真最小?
- 将信息率失真函数的定义推广到向量情形, 我们有

$$R(D) = \min_{f(\hat{x}^m|x^m): E d(X^m, \hat{X}^m) \leq D} I(X^m; \hat{X}^m),$$

其中 $d(X^m, \hat{X}^m) = \sum_{i=1}^m (x_i - \hat{x}_i)^2$.

由前面的例子的讨论, 我们有

$$\begin{aligned} I(X^m; \hat{X}^m) &= h(X^m) - h(X^m | \hat{X}^m) \\ &= \sum_{i=1}^m h(X_i) - \sum_{i=1}^m h(X_i | X^{i-1}, \hat{X}^m) \\ &\geq \sum_{i=1}^m h(X_i) - \sum_{i=1}^m h(X_i | \hat{X}_i) \\ &= \sum_{i=1}^m I(X_i; \hat{X}_i) \\ &\geq \sum_{i=1}^m R(D_i) \\ &= \sum_{i=1}^m \left(\frac{1}{2} \log \frac{\sigma_i^2}{D_i} \right)^+, \end{aligned}$$

其中 $D_i = E(X_i - \hat{X}_i)^2$. 等号可由前面例子类似地选取 $f(x^m|\hat{x}^m) = \prod_{i=1}^m f(x_i|\hat{x}_i)$ 以及分别选取分布 $\hat{X}_i \sim \mathcal{N}(0, \sigma_i^2 - D_i)$ 得到.

因此, 求解率失真问题可简化如下的最优化问题 (为了简便起见, 使用奈特为单位):

$$R(D) = \min_{\sum D_i = D} \sum_{i=1}^m \max\left\{ \frac{1}{2} \log \frac{\sigma_i^2}{D_i}, 0 \right\}.$$

用拉格朗日乘法, 我们建立函数

$$J(D) = \sum_{i=1}^m \frac{1}{2} \log \frac{\sigma_i^2}{D_i} + \lambda \sum_{i=1}^m D_i.$$

同时关于 D_i 求偏导数, 并令其等于 0, 我们有

$$\frac{\partial J}{\partial D_i} = -\frac{1}{2} \frac{1}{D_i} + \lambda = 0$$

或

$$D_i = \lambda'.$$

此时, 由库恩-塔克条件可导出

$$\frac{\partial J}{\partial D_i} = \frac{1}{2} \frac{1}{D_i} + \lambda,$$

其中 λ 的选取满足

$$\frac{\partial J}{\partial D_i} \begin{cases} = 0 & \text{如果 } D_i < \sigma_i^2, \\ \leq 0 & \text{如果 } D_i \geq \sigma_i^2. \end{cases}$$

定理 17.13: 并联高斯信源的率失真

设 $X_i \sim \mathcal{N}(0, \sigma_i^2)$ ($i = 1, 2, \dots, m$) 为独立的高斯随机变量, 假定失真度量为 $d(x^m, \hat{x}^m) = \sum_{i=1}^m (x_i - \hat{x}_i)^2$, 则率失真函数为

$$R(D) = \sum_{i=1}^m \frac{1}{2} \log \frac{\sigma_i^2}{D_i},$$

其中

$$D_i = \begin{cases} \lambda & \text{如果 } \lambda < \sigma_i^2, \\ \sigma_i^2 & \text{如果 } \lambda \geq \sigma_i^2. \end{cases}$$

其中 λ 的选取满足 $\sum_{i=1}^m D_i = D$.

第 18 讲 信息论与统计学习

18.1

统计学习——通常被视为机器学习的一个子领域或同义词——致力于利用数据训练概率模型。这些模型随后被用于预测、样本生成以及发现数据紧凑且有意义的表示。

统计学习的核心问题表述非常简单：给定从未知分布 P_{data} 中采样得到的数据集，以及一个概率分布族（模型） $\mathcal{M} = \{p_\theta; \theta \in \Theta\}$ ，识别出 $\mathcal{M}$ 中与 P_{data} 最接近的分布 p_{θ^*} 。

因此，构建学习问题的第一步是指定一个概率模型，并配以衡量概率分布之间差异的适当散度量。在这些度量中，相对熵（或交叉熵）使用最为广泛，这主要是因为最小化相对熵等价于最大化观察数据的对数似然。然而，在某些设定下，直接最大化对数似然在计算上是不可行的，此时人们会转而优化一个合适的下界。在其他场景中，则会采用备选的信息论度量。

我们首先介绍监督学习，并通过经典的线性回归和逻辑回归预测模型阐明其应用。这一讨论作为神经网络的铺垫，而神经网络是统计学习中最强大且应用最广的模型之一。

随后，我们转向生成模型，其主要目标是生成与训练数据集类似的样本。我们探讨的模型包括：自回归模型（如大语言模型 LLM）；通过优化对数似然的证据下界（ELBO）进行训练的变分自编码器（VAE）和扩散模型；使用各种 f -散度（第 1.7.1 节）作为训练目标的生成对抗网络（GAN）；以及使用 Fisher 散度训练的基于分数的模型。

统计学习在概念上与参数估计问题密切相关。在参数估计中，我们假设存在真实参数 $\theta^* \in \Theta$ 满足 $p_{\theta^*} = P_{\text{data}}$ ，目标是从数据中估计 θ^* 。统计学习问题在关键方面脱离了这一传统范式：我们不假设 $P_{\text{data}} \in \mathcal{M}$ 。事实上，真实的数据生成分布可能位于 $\mathcal{M}$ 之外。相反，模型的选择反映了关于数据生成分布的先验知识、建模假设或归纳偏置。此外，参数空间（也称为模型类）可能比传统参数估计中复杂得多，甚至可能由高度过参数化的深度神经网络族构成。

由于学习的目标是在未来数据上取得良好表现，而不仅仅是像传统统计学那样推断数据生成机制，因此产生了一种潜在的张力：学习模型必须在拟合训练数据与泛化到新数据的能力之间取得平衡。过拟合的模型可能在训练数据集上表现优异，却无法泛化到未见过的样本。这种权衡通常通过正则化来解决，正则化限制了学习模型的有效复杂度，或惩罚过复杂的解，以提高泛化能力。最终得到的泛化能力通常通过将观察到的数据划分为训练集和测试集，并根据模型在测试集上的表现进行评估来衡量。我们将在下一章讨论这一学习核心方面的根本极限。

本讲中，我们主要考虑数据集结果，例如样本集 $\{x_i, i = 1, \dots, n\}$ 或输入-输出对 $\{(x_i, y_i), i = 1, \dots, n\}$ ，除非另有说明，输入-输出对 (x_i, y_i) 的取值范围为 $\mathbb{R}^K \times \mathbb{R}^M$ 。我们偶尔会使用下标 NN 来表示由神经网络建模的函数或概率分布。

18.2 监督学习

为具体起见，下文中我们假设变量为离散随机变量。相同的讨论适用于连续随机变量，只需将 p 替换为 f ，将熵替换为微分熵。

在监督学习中，我们给定一个条件概率模型 $\mathcal{M} = \{p_\theta(y|x), \theta \in \Theta\}$ ，它描述了在给定输入（或特征） X 的情况下输出（或标签） Y 如何随机生成；以及一个从未知联合 pmf $P_{\text{data}}(x, y)$ 中独立同分布（i.i.d.）采样的输入-输出对数据集 $\{(x_i, y_i), i = 1, 2, \dots, n\}$ 。目标是在相对熵意义下，找到与真实条件 pmf $P_{\text{data}}(y|x)$ 最接近的条件 pmf $p_{\theta^*}(y|x) \in \mathcal{M}$ 。

我们首先将这一目标与交叉熵联系起来。

引理 18.1

最小化条件相对熵 $D(P_{\text{data}}(y|x) || p_{\theta}(y|x) | P_{\text{data}}(x))$ 这一目标等价于最小化条件交叉熵 $H(P_{\text{data}}(y|x), p_{\theta}(y|x) | P_{\text{data}}(x))$ 。

证明. 我们展开条件相对熵:

$$D(P_{\text{data}}(y|x) || p_{\theta}(y|x) | P_{\text{data}}(x)) = \mathbb{E}_{P_{\text{data}}(x,y)}[\log P_{\text{data}}(Y|X)] - \mathbb{E}_{P_{\text{data}}(x,y)}[\log p_{\theta}(Y|X)].$$

第一项是数据生成分布下的负条件熵 $-H(Y|X)$, 它不是 θ 的函数。因此, 最小化条件相对熵等价于最小化

$$-\mathbb{E}_{P_{\text{data}}(x,y)}[\log p_{\theta}(Y|X)], \quad (18.1)$$

这正是条件交叉熵 $H(P_{\text{data}}(y|x), p_{\theta}(y|x) | P_{\text{data}}(x))$ 。□

在实际应用中, 数据的真实概率质量函数是未知的, 因此我们改为最小化条件交叉熵的经验估计, 或者等价地, 最大化数据的条件对数似然。

定义 18.2: 条件对数似然

给定数据集 $\{(x_i, y_i), i = 1, \dots, n\}$, 我们将数据的条件对数似然定义为:

$$l_n(\theta) = \sum_{i=1}^n \log p_{\theta}(y_i | x_i). \quad (18.2)$$

由于对 n 的依赖性在上下文中很明确, 在本讲的剩余部分我们将省略下标 n 。

定理 18.3

设 $(X_1, Y_1), \dots, (X_n, Y_n)$ 为从 $P_{\text{data}}(x, y)$ 中采样的独立同分布的样本。若条件交叉熵 $H(P_{\text{data}}(y|x), p_{\theta}(y|x) | P_{\text{data}}(x)) < \infty$, 则我们有

$$\frac{1}{n} l_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log p_{\theta}(Y_i | X_i) \xrightarrow{P} -H(P_{\text{data}}(y|x), p_{\theta}(y|x) | P_{\text{data}}(x)) \quad \text{依概率收敛}. \quad (18.3)$$

证明. 该定理是弱大数定律的直接推论。□

该结果为使用数据的条件对数似然来估计最优参数 θ^* 提供了合理性依据。这也支持了在监督学习中, 使用条件相对熵作为距离度量, 因为最大似然估计具备许多良好的统计性质。

注记 18.4

监督学习只是几种学习范式之一。本章我们还将遇到无监督学习: 在无监督学习中, 我们给定一个没有关联标签的数据集 $\{x_i, i = 1, 2, \dots, n\}$, 旨在从给定的模型中选择一个概率分布 $p_{\theta}(x)$, 使其与未知的真实数据生成分布 P_{data} 最接近。与监督学习类似, 相对熵是常用的差异度量, 但正如我们将看到的, 也会使用其他目标函数。

接下来, 我们通过经典的线性回归和逻辑回归模型展示监督学习。

18.2.1 线性回归

在回归分析中, 我们指定一个条件 pdf $f(y|x_1, x_2, \dots, x_K)$, 旨在根据输入向量 $(x_1, x_2, \dots, x_K)$ 预测输出 (响应) 变量 Y 。预测器 $r(x_1, x_2, \dots, x_K) = \mathbb{E}(Y | X_1 = x_1, X_2 = x_2, \dots, X_K = x_K)$ 被称为回归函数。例如, 输入向量可以代表历史股票价格以及宏观经济指标, 输出是对未来股票收益的预测; 或者它可以包括历史用电量以及日期、月份和时段信息, 输出是对电力需求量的预测。

回归函数是从输入-输出对数据集 $\{(x_{1i}, x_{2i}, \dots, x_{Ki}), y_i, i = 1, \dots, n\}$ 中估计得到的。估计该函数可能具有挑战性, 特别是在高维设定下, 通常需要对给定输入的 Y 的条件概率密度函数施加结构性假设。一种特别易于处理的方法是假设概率密度函数为高斯分布。在此假设下, 条件期望在输入变量上是线性的, 自然引出了经典的线性回归模型。

定义 18.5

线性回归概率模型由所有形如以下形式的条件概率密度函数组成:

$$f_\theta(y|x_1, x_2, \dots, x_K) = \mathcal{N}\left(y; w_0 + \sum_{k=1}^K w_k x_k, \sigma^2\right), \quad (18.4)$$

其中 $w_0, w_1, \dots, w_K \in \mathbb{R}$ 被称为回归系数 (或权重), σ^2 是参数完全已知时预测误差的方差, 参数向量为 $\theta = (w_0, w_1, \dots, w_K, \sigma^2)$ 。

在该模型下, 回归函数 $r(x) = \mathbb{E}(Y|X = x) = w_0 + \sum_{k=1}^K w_k x_k$ 在 $(x_1, x_2, \dots, x_K)$ 上是线性的。定义 $w^T = [w_0, w_1, \dots, w_K]$ 和 $x^T = [1, x_1, \dots, x_K]$, 我们可以将参数更紧凑地重写为 $\theta = (w, \sigma^2)$, 回归函数重写为 $r(x) = w^T x$ 。

为了估计 (或学习) 模型参数, 我们使用监督学习, 即寻找使数据的条件对数似然最大化的参数 $(\hat{w}, \hat{\sigma}^2)$:

$$l(w, \sigma^2) = \sum_{i=1}^n \log \mathcal{N}(y_i; w^T x_i, \sigma^2) \quad (18.5)$$

$$= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{\log e}{2\sigma^2} \sum_{i=1}^n (y_i - w^T x_i)^2. \quad (18.6)$$

该优化问题具有闭式解 (解析解)。

定理 18.6

使式 (1.5) 最大化的参数值由著名的最小二乘公式给出:

$$\hat{w} = (X^T X)^{-1} X^T y, \quad (18.7)$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{w}^T x_i)^2, \quad (18.8)$$

其中

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{K1} \\ 1 & x_{12} & \cdots & x_{K2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{Kn} \end{pmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}. \quad (18.9)$$

量 $n\hat{\sigma}^2$ 被称为残差平方和 (RSS)。

参数学习完成后, 估算的回归函数 $\hat{r}(x)$ 可用于根据未见的输入 x' 预测输出 $\hat{Y}'$, 即 $\hat{Y}' = \hat{w}^T x'$ 。

18.2.2 逻辑斯蒂回归

在分类问题中,我们指定一个条件概率质量函数 $p(y|x_1, x_2, \dots, x_K)$, 其中 $(x_1, x_2, \dots, x_K)$ 是输入向量, 类标签 y 取值于大小为 M 的有限集 $\mathcal{M}$ 。目标是在给定输入向量的情况下预测类别 Y 。例如, 输入向量可以表示一封电子邮件或从中提取的特征, 类标签 $y \in \{\text{垃圾邮件}, \text{非垃圾邮件}\}$; 或者表示手写字符的特征, 类标签 $y \in \{A, B, \dots, Z\}$ 。

为了执行分类, 我们从输入-输出数据集 $\{(x_{1i}, x_{2i}, \dots, x_{Ki}), y_i, i = 1, \dots, n\}$ 中估计条件概率质量函数 $p(y|x_1, x_2, \dots, x_K)$ 的模型。我们从逻辑斯蒂回归 (Logistic Regression) 开始, 这是二分类问题 ($\mathcal{M} = \{0, 1\}$) 中经典且广泛使用的模型。

定义 18.7

逻辑斯蒂回归概率模型由所有形如以下形式的条件概率质量函数组成:

$$p(y|x) = \text{Bern} \left(y; \sigma \left(w_0 + \sum_{j=1}^K w_j x_j \right) \right), \quad (18.10)$$

其中

$$\sigma(a) = \frac{1}{1 + e^{-a}}, \quad a \in \mathbb{R} \quad (18.11)$$

被称为逻辑斯蒂函数 (或 Sigmoid 函数)。我们定义 $w^T = [w_0, w_1, \dots, w_K]$, $x^T = [1, x_1, \dots, x_K]$, 以及 $p_i = \sigma(w^T x_i), i = 1, \dots, n$ 。

给定标注数据集中 $\{(x_i, y_i), i = 1, \dots, n\}$, 其中 $x_i \in \mathbb{R}^K$, $y_i \in \{0, 1\}$, 我们希望学习参数 w 。这是监督学习的另一个实例, 每个输入 x_i 都配有一个已知的标签 y_i 。在逻辑回归中, 参数 w 通过最大化数据的条件对数似然来确定:

$$l(w) = \sum_{i=1}^n \log \text{Bern}(y_i; p_i) \quad (18.12)$$

$$= \sum_{i=1}^n (y_i \log p_i + (1 - y_i) \log(1 - p_i)). \quad (18.13)$$

因此, 数据的负对数似然等于 n 倍的平均交叉熵 (或平均对数损失)。

与线性回归不同, $\hat{w}$ 没有闭式解。然而, 对数似然 $l(w)$ 关于 w 是凹函数, 可以通过数值方法最大化, 例如使用梯度上升——一种形如以下形式的迭代优化算法:

$$w^{(t+1)} = w^{(t)} + \gamma \nabla l(w^{(t)}), \quad t = 1, 2, \dots \quad (18.14)$$

其中 $\gamma > 0$ 被称为学习率。利用导数关系:

$$\frac{d\sigma}{da} = \frac{e^{-a}}{(1 + e^{-a})^2} = \sigma(a)(1 - \sigma(a)), \quad (18.15)$$

我们可以证明梯度具有如下简单形式:

$$\nabla l(w) = \sum_{i=1}^n (y_i - p_i) x_i. \quad (18.16)$$

逻辑回归可以扩展到两个以上的类别。设 $a \in \mathbb{R}^M$, 定义 Softmax 函数为:

$$\text{softmax}(a) = \left(\frac{e^{a_1}}{\sum_{m=1}^M e^{a_m}}, \dots, \frac{e^{a_M}}{\sum_{m=1}^M e^{a_m}} \right). \quad (18.17)$$

注意到这自然地将逻辑函数扩展到了 $M > 2$ 的情形。

我们用 $\text{Cat}(y; p_1, \dots, p_M)$ 表示分类概率质量函数，其中：

$$p_m = \text{softmax}_m(a) = \frac{e^{a_m}}{\sum_{m'=1}^M e^{a_{m'}}}, \quad m = 1, \dots, M. \quad (18.18)$$

定义 18.8

多分类逻辑回归概率模型由形如以下形式的条件 pmf 组成：

$$p(y|x) = \text{Cat}(y; \text{softmax}(w_1^T x, \dots, w_M^T x)), \quad (18.19)$$

其中 $x^T = [1, x_1, \dots, x_K]$ 是输入向量， $w_m^T = [w_{m0}, w_{m1}, \dots, w_{mK}]$ 是权重向量， $y \in \{1, 2, \dots, M\}$ 是输出（类别）。

与前面一样，给定标注数据集 $\{(x_i, y_i), i = 1, \dots, n\}$ ，我们希望学习使条件对数似然最大化的参数 $\hat{w}_1, \dots, \hat{w}_M$ ：

$$l(w_1, \dots, w_M) = \sum_{i=1}^n \sum_{m=1}^M y_{im} \log p_{im}, \quad (18.20)$$

其中对于 $m = 1, \dots, M, i = 1, \dots, n$ ：

$$p_{im} = \text{softmax}_m(w_1^T x_i, \dots, w_M^T x_i) \quad (18.21)$$

且

$$y_{im} = \begin{cases} 1 & \text{若 } y_i = m, \\ 0 & \text{否则.} \end{cases} \quad (18.22)$$

该优化问题同样是凸优化问题，可以通过梯度上升法求解。

18.3 神经网络

首先，我们介绍单层神经网络，作为线性回归和逻辑回归模型的自然扩展。

18.3.1 单层网络

单层神经网络由一组输入 x 和一组输出 y 组成，表示为：

$$y_m = \phi_m \left(w_{m0} + \sum_{k=1}^K w_{mk} x_k \right), \quad m = 1, \dots, M, \quad (18.23)$$

其中 $w_{m0}, w_{m1}, \dots, w_{mK}$ 为权重， ϕ_m 是与输出 m 关联的激活函数。

单层神经网络的示例包括：

- 线性回归：此处 ϕ_m 为恒等函数。
- 逻辑回归：此处 $\phi_1(a) = \sigma(a)$ 。
- 感知机（Perceptron）：单层神经网络最早的例子，激活函数为赫维赛德阶跃函数（Heaviside step function）：

$$\phi_m(a) = \begin{cases} 1 & \text{若 } a \geq 0, \\ 0 & \text{若 } a < 0. \end{cases} \quad (18.24)$$

- 多分类逻辑回归：此处 $\phi_1, \dots, \phi_M$ 被替换为向量值函数 $\phi(a) = \text{softmax}(a)$ 。

18.3.2 双层网络

单层神经网络的输出是输入加权和的函数，这一事实限制了其拟合函数的能力。增加表达能力的一种方法是添加一层固定的非线性基函数，使得：

$$y_m = \phi_m \left(w_{m0} + \sum_{k=1}^K w_{mk} \psi_k(x) \right), \quad m = 1, \dots, M. \quad (18.25)$$

例如，假设单一输入 x ，基函数可以是多项式 x^k 、“高斯”函数或 Sigmoid 函数 $\psi_k(x) = \exp\left(-\frac{(x-\mu_k)^2}{2s^2}\right)$ 、 $\psi_k(x) = \sigma\left(\frac{x-\mu_k}{s}\right)$ 等。

通过增加基函数的数量，此类模型可以任意好地逼近一大类函数。此外，由于模型对于权重仍然是线性的，1.1.1 和 1.1.2 节中线性回归和逻辑回归的推导可以通过将每个输入变量 x_k 替换为 $\psi_k(x)$ 自然延续。然而，在实际应用中，这种方法的有效性受到限制，因为基函数是固定的，无法适应训练数据。通过从数据中学习基函数而不是先验地固定它们，可以克服这一局限。由此产生的架构由一个隐藏层和一个输出层组成。按照标准机器学习术语，输入层不计为网络层；因此，该架构被称为双层神经网络。

在第一层中，我们将输入进行线性组合得到：

$$a_m^{(1)} = w_{m0}^{(1)} + \sum_{k=1}^K w_{mk}^{(1)} x_k, \quad m = 1, \dots, M_1. \quad (18.26)$$

然后，我们使用非线性激活函数转换每个 $a_m^{(1)}$ ，得到第一层的输出：

$$z_m^{(1)} = \phi^{(1)}(a_m^{(1)}), \quad m = 1, \dots, M_1. \quad (18.27)$$

在第二层中，我们将 $z_k^{(1)}$ 值进行线性组合得到：

$$a_m^{(2)} = w_{m0}^{(2)} + \sum_{k=1}^{M_1} w_{mk}^{(2)} z_k^{(1)}, \quad m = 1, \dots, M. \quad (18.28)$$

最后，我们使用另一个激活函数转换每个 $a_m^{(2)}$ 以获得网络输出：

$$y_m = \phi_m^{(2)}(a_m^{(2)}), \quad m = 1, \dots, M. \quad (18.29)$$

隐藏层中使用的激活函数的例子包括逻辑 Sigmoid 函数 $\sigma(a)$ 、赫维赛德阶跃函数，以及修正线性单元 (ReLU) $\phi(a) = \max(0, a)$ 。输出层中可以使用相同的激活函数，但根据应用的不同通常会做出不同的选择。例如，在回归中通常使用恒等函数，而在多分类中 Softmax 函数很常见。

上述双层神经网络已被证明是通用函数逼近器 (Universal Function Approximator)。例如，考虑一个具有恒等输出激活函数的单输出双层神经网络，表达为：

$$y = \sum_{m=1}^M w_m^{(2)} \phi \left(w_{m0}^{(1)} + \sum_{k=1}^K w_{mk}^{(1)} x_k \right), \quad (18.30)$$

其中 ϕ 是激活函数（如逻辑 Sigmoid 函数 σ ）。关于该网络逼近能力的以下结果在 (Funahashi 1989) 中得到证明。

定理 18.9

设 $\phi(a)$ 为有界单调递增函数。设 $K \subset \mathbb{R}^K$ 为紧集， $\hat{g}(x)$ 为 K 上的实值连续函数。则对于任意

$\epsilon > 0$, 存在整数 M 和权重 $\{w_{m0}^{(1)}, w_{mk}^{(1)}, w_m^{(2)}, m = 1, \dots, M, k = 1, \dots, K\}$, 使得:

$$\hat{g}(x) = \sum_{m=1}^M w_m^{(2)} \phi \left(w_{m0}^{(1)} + \sum_{k=1}^K w_{mk}^{(1)} x_k \right) \quad (18.31)$$

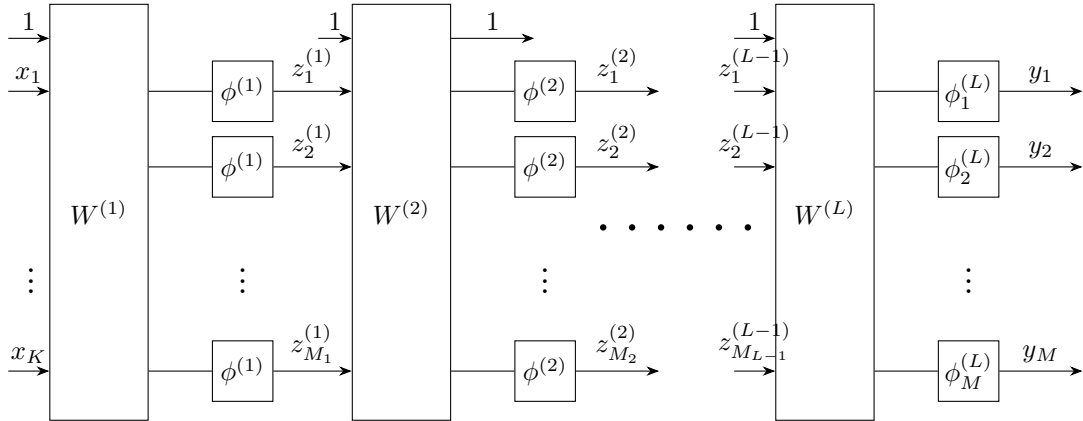
满足 $\max_{x \in K} |g(x) - \hat{g}(x)| < \epsilon$ 。

尽管上面结果确立了双层神经网络的表达能力, 但它们仅保证了存在能够以任意精度逼近任何给定函数的神经网络。特别是, 它们没有给出达到所需逼近精度所需的网络规模, 也没有说明如何在实践中有效地训练此类网络。

通过引入额外的隐藏层可以缓解其中的一些局限性, 这通常会减少表示复杂函数所需的参数数量。尽管此类深层网络曾经在计算上不可行, 但硬件和优化方法的进步, 结合大规模数据集的可用性, 使得训练具有数十甚至数百层以及数十亿参数的网络成为可能。

18.3.3 深度网络

我们将双层网络的描述扩展到 $L > 2$ 层, 如下所示。设输入向量为 $z^{(0)} = x$, 其中 $x = [1, x_1, \dots, x_K]^T$ 。对于每个层 $l = 1, 2, \dots, L-1$, 设输出向量 $z^{(l)}$ 的长度为 M_l ; 对于层 L , 设输出向量 y 的长度为 M 。



定义权重矩阵, 则层 $l = 1, \dots, L-1$ 的输出可以表示为:

$$z^{(l)} = \begin{bmatrix} 1 \\ \phi^{(l)}(W^{(l)} z^{(l-1)}) \end{bmatrix}, \quad (18.32)$$

其中 $\phi^{(l)}(a) = [\phi^{(l)}(a_1), \dots, \phi^{(l)}(a_{M_l})]^T$ 。网络的最终输出由下式给出:

$$y = [\phi_1^{(L)}(a_1^{(L)}), \dots, \phi_M^{(L)}(a_M^{(L)})]^T, \quad (18.33)$$

其中 $a^{(L)} = W^{(L)} z^{(L-1)}$ 。

18.3.4 基于神经网络的回归与分类

神经网络 (NN) 可用于回归和分类, 形式如下:

- 回归: 具有神经回归函数的概率模型形式为:

$$p(y|x) = \mathcal{N}(y; r_{\text{NN}}(x), \sigma_{\text{NN}}^2), \quad (18.34)$$

其中 $r_{\text{NN}}(x)$ 由神经网络建模, 其输出层具有单个恒等激活函数单元。

- 二分类：概率模型形式为：

$$p(y|x) = \text{Bern}(y; p_{\text{NN}}(x)), \quad (18.35)$$

其中神经网络的输出具有单个逻辑 Sigmoid 激活单元。

- 多分类：概率模型形式为：

$$p(y|x) = \text{Cat}(y; p_{\text{NN}}(x)), \quad (18.36)$$

其中输出层使用 Softmax 激活函数。

18.3.5 神经网络训练

与线性回归和逻辑回归类似，训练神经网络基于数据集，旨在寻找一组能最大化对数似然（或更一般地，优化所选目标函数）的权重。与这些简单模型不同，由此产生的优化问题通常是高度非凸的，因此无法保证找到全局最优解。在实践中，神经网络采用可微激活函数，并使用基于梯度的优化方法来寻找局部最优解或驻点。通过重复应用微分的链式法则（称为反向传播），梯度的计算变得高效；同时，通过随机梯度下降（SGD）方法进一步加速训练，其中梯度在每次迭代中在随机选择的数据小批量（mini-batch）上求值。

18.4 自回归模型

我们现在转向生成模型，其目标是训练一个概率模型，然后使用它生成类似于训练数据集的新样本。例如，在自然语言处理中，目标是生成反映语言结构和统计特性的文本；在音乐创作中，生成与给定流派一致的乐曲；在计算机视觉中，合成具有期望内容的图像或视频。

我们考虑的第一类生成模型是**自回归模型**，它特别适合由顺序信息源产生的数据，如文本、语音和音乐。

定义 18.10

自回归概率模型 $\mathcal{P}$ 由所有通过概率链式法则表达的概率质量函数（或概率密度函数）组成：

$$p_{\theta}(\mathbf{x}) = p_{\theta}(x_1)p_{\theta}(x_2|x_1)\dots p_{\theta}(x_K|x^{K-1}), \quad (18.37)$$

其中 $\mathbf{x} \in \mathcal{X}^K$ （对于某个有限集 $\mathcal{X}$ ），且 θ 在参数集 Θ 中取值。

变量 x_k 的顺序对应于数据生成的顺序，例如文本块中单词或 Token 的顺序。

18.4.1 训练自回归模型

给定一组自回归概率分布 $\mathcal{P}$ 和从未知 pmf P_{data} 中独立同分布采样的数据集 $\{x_i, i = 1, \dots, n\}$ ，我们希望选择最接近 P_{data} 的 pmf $p_{\theta} \in \mathcal{M}$ 。

首先假设 $x \in \mathcal{X}^K$ 。我们可以简单地通过频数表估计每个条件概率 $p_{\theta}(x_k|x^{k-1})$ 来“训练”模型。然而，随着 k 的增加，存储这些估计所需的表大小呈指数级增长，这使得该方法不可行。为了解决这个问题，我们可以用有限记忆的马尔可夫过程来逼近自回归模型。Shannon (1948) 曾使用这种方法来估计英文文本的熵。然而，它产生的采样文本与自然语言几乎没有相似之处。这表明生成可以通过英语测试的句子需要具有长记忆的模型。

另一种训练自回归模型的方法是对每个条件分布进行参数化。例如，假设二值数据 $x \in \{0, 1\}^K$ ，我们可以为每个步骤 k 使用单独的逻辑回归模型：

$$P_{\theta_k}(x_k|x^{k-1}) = \text{Bern}(x_k; \sigma(w_{0k} + \sum_{j=1}^{k-1} w_{jk}x_j)). \quad (18.38)$$

并使用监督学习训练每个模型。

这种方法需要训练 K 个不同的模型，效率低下且扩展性差。对每个条件使用更强大的模型（如专用神经网络）只会恶化这种扩展问题，使其在计算上难以承受。更重要的是，这种方法未能捕获一种直觉，即序列的底层规则（如文本中的语法）在不同位置上应该是一致的。

一种更有效且具备可扩展性的解决方案是使用具有共享参数集 θ 的单个神经网络来对所有条件分布建模。然后，通过在单一模型 $p_\theta(x)$ 下最大化数据集的对数似然来联合训练整个模型。这就是 Transformer 等现代架构背后的原理。

对于文本等离散数据，网络通常输出词表 $\mathcal{X}$ 上 Softmax 分布的 Logits。对于连续数据，网络可以配置为输出更复杂分布的参数，例如高斯混合模型。

18.4.2 关于自回归模型的记号

1. **困惑度 (Perplexity)** 在自然语言处理中——特别是大语言模型 (LLM) ——性能通常在测试数据集上使用经验困惑度进行评估。困惑度是每个 Token 负对数似然的指数。较低的困惑度表明模型更好，因为它对应于数据上更高的平均单 Token 概率。
2. **预测** 训练好的自回归模型 $p_\theta(x)$ 可用于预测。给定观察到的序列前缀，模型可以通过评估分布预测下一个元素。
3. **采样** 模型的预测分布为序列生成提供了基础。一种常用的方法是祖先采样 (Ancestral Sampling)：首先从 $p_{\hat{\theta}}(x_1)$ 中抽取 a_1 ；然后从 $p_{\hat{\theta}}(x_2|a_1)$ 中抽取 a_2 ；继续顺序采样，直到获得完整序列 a^K 。
4. **推理** 在人工智能中，根据模型进行预测和样本生成通常被称为“推理” (Inference)。注意不要将此用法与统计学中的“统计推断”概念混淆。
5. **顺序概率估计与通用压缩** 训练好的自回归模型充当强大的顺序概率估计器。它可以与算术编码相结合以进行通用数据压缩。

索引

马尔可夫链, 8

事件, 5

数据处理不等式, 28

样本空间, 5