

第 3 讲 熵的更多性质 (续)

对数和不等式及其应用

数据处理不等式

充分统计量

费诺不等式

对数和不等式及其
应用

数据处理不等式

充分统计量

费诺不等式

对数和不等式及其
应用

数据处理不等式

充分统计量

费诺不等式

- ▶ T. M. Cover and J. A. Thomas, Elements of Information Theory, 2.1-2.6.
- ▶ R. W. Yeung, Information and Network Coding, 2.2-2.7.
- ▶ F. Alajaji, P. Chen, An Introduction to Single-User Information Theory, 2.1-2.3.

对数和不等式及其应用

数据处理不等式

充分统计量

费诺不等式

对数和不等式及其
应用

数据处理不等式

充分统计量

费诺不等式

定理 1.1: 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

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

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 1.2

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

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

对数不等式及其应用

数据处理不等式

充分统计量

费诺不等式

定理 1.3: 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) 为两对概率密度函数, 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:

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

对 x 求和我们可以得到所需要的结论. □

定理 1.4: 熵的凹性

$H(p)$ 是关于 p 的凹函数.

定理 1.5: 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. □

定理 1.6

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.

证明.

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

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

证明.

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,

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

于是互信息是条件分布的凸函数.



对数和不等式及其应用

数据处理不等式

充分统计量

费诺不等式

对数和不等式及其
应用

数据处理不等式

充分统计量

费诺不等式

定理 2.1: 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 $I(Y; Z) \geq I(X; Z)$. □

推论 2.2

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

推论 2.3

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

对数和不等式及其
应用

数据处理不等式

充分统计量

费诺不等式

推论 2.4

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

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



注记 2.5

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

注记 2.6

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).

对数和不等式及其应用

数据处理不等式

充分统计量

费诺不等式

对数和不等式及其
应用

数据处理不等式

充分统计量

费诺不等式

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 3.1

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 3.2

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

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

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.

对数和不等式及其应用

数据处理不等式

充分统计量

费诺不等式

例 3.3

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 3.4

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 3.5

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

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

对数和不等式及其应用

数据处理不等式

充分统计量

费诺不等式

对数和不等式及其
应用

数据处理不等式

充分统计量

费诺不等式

罗伯特·马利欧·费诺 (Robert Mario Fano, 1917 年 11 月 11 日—2016 年 7 月 13 日) 是一名意裔美籍计算机科学家, 为麻省理工学院电机工程与计算机科学教授. 专长于信息论, 曾与克劳德·香农共同开发出香农-费诺编码, 并曾提出费诺不等式. 1976 年获得香农奖.



- ▶ 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 4.1: 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.

对数和不等式及其应用

数据处理不等式

充分统计量

费诺不等式

证明.

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 4.2

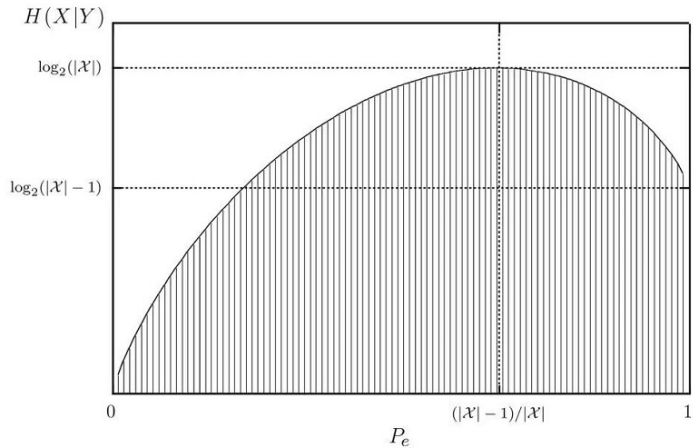
对任意两个随机变量 X 和 Y , 设 $p = P(X \neq Y)$,

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

推论 4.3

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

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



对数和不等式及其
应用

数据处理不等式

充分统计量

费诺不等式

注记 4.4

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.

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 4.5

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

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

□

推论 4.6

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

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

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

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. $\square$