信息论与编码

马啸
maxiao@mail.sysu.edu.cn

计算机学院
中山大学

2021 年春季学期

1. 信源编码回顾
   - 离散无记忆信源
   - 熵
   - 信源编码定理
   - 编码类型
     - Shannon Code
     - Huffman Code: Optimal Prefix Code
     - Shannon-Fano-Elias, Arithmetic, LZ Codes

2. 信息不等式直观
   - 基础知识
   - 一些不等式的直观解释

信源编码回顾

离散无记忆信源

Definition 1

一个离散信源（source）可以用随机序列表示，即

$$\mathbf{X} = (X_1, X_2, \cdots, X_n, \cdots), \tag{1}$$

$X_t \in \mathcal{X}$。为方便起见，我们记$X^n \triangleq (X_1, X_2, \cdots, X_n)$。假定对于任意给定$n$，概率质量函数$P_{X^n}(x^n), x^n \in \mathcal{X}^n$ 是已知的。

信源编码的功能是用二进制序列表示信源产生的消息，目标是在允许的“错误”范围内，用尽可能少的二进制数位。

信源编码的压缩速率（平均每个信源符号所需要的二进制数位数）的极限完全由

$$\frac{1}{n} \log \frac{1}{P_{X^n}(X^n)}$$

的“谱线”（可以称为熵谱）的极限行为来决定。

离散无记忆信源

Definition 2 (离散无记忆信源)

设信源$\mathbf{X} = (X_1, X_2, ..., X_n, ...)$，满足：

1. **无记忆的：** $P_{X^n}(x^n) = \prod_{1 \le t \le n} P_{X_t}(x_t)$ 对于任意 $n > 1$
2. **平稳的：** $P_{X_t}(x) \equiv P_{X_1}(x) \triangleq P_X(x)$ 对于任意 $t > 1$

则称该信源为离散平稳无记忆信源，也称作独立同分布信源。

熵

Definition 3

离散随机变量X，概率质量函数为 $P_X(x)$， $x \in \mathcal{X}$，则的X 熵，记作 $H(X)$，

$$H(X) = \mathrm{E}(\log \frac{1}{P_X(X)}) = -\sum_{x \in \mathcal{X}} P_X(x) \log P_X(x). \tag{2}$$

熵

说明:

(1) 根据熵的定义，若 log 以2 为底，则熵的单位是“比特/符号”；若 log 以 e 为底，则熵的单位是“奈特/符号”。如果我们已知信源每秒发出的符号数，则熵的单位可以是“比特/秒”或“奈特/秒”。

(2) 我们约定 $0 \log 0 \triangleq 0$，这样约定是考虑到 $\lim_{x \to 0^+} x \log x = 0$ 。

(3) $I(x) \triangleq \log(1/P_X(x))$ 也被称为 x 的**自信息量**。所以，我们也可以说熵是自信息量的数学期望。概率小的样本点具有大的自信息量，但是在熵中权重较小；而概率大的样本点的自信息量小，但在熵中权重较大。

熵的性质

1. 对称性

概率矢量$p=(p_1,p_2,\cdots,p_n)$中，各分量的次序任意改变，熵不变。例如

$$H(p_1,p_2,\cdots,p_n)=H(p_2,p_1,\cdots,p_n)$$

说明熵仅与信源的总体概率特性有关，而与随机变量的取值无关，例如下列信源的熵都是相等的。

$$\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 \\ 1/3 & 1/2 & 1/6 \end{bmatrix}$$

$$\begin{bmatrix} Y \\ P \end{bmatrix} = \begin{bmatrix} y_1 & y_2 & y_3 \\ 1/3 & 1/6 & 1/2 \end{bmatrix}$$

$$\begin{bmatrix} Z \\ P \end{bmatrix} = \begin{bmatrix} z_1 & z_2 & z_3 \\ 1/2 & 1/3 & 1/6 \end{bmatrix}$$

熵的性质

2. 非负性

$$H(X) = H(p_1, p_2, \cdots, p_n) \geq 0$$

3. 确定性

$$H(0,1) = H(0,1,0,\cdots,0) = 0$$

4. 拓展性

$$\lim_{\varepsilon \to 0} H_{K+1}\left(P_1, P_2, \cdots, P_K - \varepsilon, \varepsilon\right) = H_K\left(P_1, P_2, \cdots, P_K\right)$$

熵的性质

5. 可加性

$$
\begin{aligned}
&H_M\left(P_1 Q_{11}, P_1 Q_{21}, \cdots, P_1 Q_{m_1 1}, P_2 Q_{12}, P_2 Q_{22}, \cdots,\right.\\
&\left.P_2 Q_{m_2 2}, \cdots, P_K Q_{1K}, P_K Q_{2K}, \cdots, P_K Q_{m_k K}\right)\\
&= H_K\left(P_1, P_2, \cdots, P_K\right) + \sum_{k=1}^{K} P_k H_{m_k}\left(Q_{1k}, Q_{2k}, \cdots, Q_{m_k k}\right)
\end{aligned}
$$

其中

$$
\begin{aligned}
&\sum_{k=1}^{K} P_k = 1, P_k \geqslant 0\\
&\sum_{j=1}^{m_k} Q_{jk} = 1, Q_{jk} \geqslant 0\\
&M = \sum_{k=1}^{K} m_k
\end{aligned}
$$

可加性可以从概率树的角度描述。

链式法则： 设N 维随机变量集$(X_1 X_2 \cdots X_n)$，则有

$$
H\left(X_1 X_2 \cdots X_n\right) = H\left(X_1\right) + H\left(X_2 \mid X_1\right) + \cdots + H\left(X_n \mid X_1 \cdots X_{n-1}\right)
$$

熵的性质

6. 极值性

   当随机变量X 的各个取值概率相等时，熵最大。因为出现任何取值的可能性相等，不确定性最大，即

$$H(X) \le H(\frac{1}{M}, \frac{1}{M}, \cdots, \frac{1}{M}) = \log M$$

信源编码基本框架

一般地，一个时间离散的信源可以表示为一个随机变量序列：$X_1, X_2, \ldots, X_n, \ldots$，其中$X_t$ 取值在$\mathcal{X}$ 上，其统计规律可以用一族联合分布律$\{P_{\mathbf{X}}(\mathbf{x})\}, n = 1, 2, \ldots$ 来表征。设$\mathcal{D} = \{0, 1, \ldots, D-1\}$ 是字符集，我们用$\mathcal{D}^*$ 表示由$\mathcal{D}$ 构成的字符串的全体，包括空字符串，即$\mathcal{D}^* = \bigcup_{\ell \geq 0} \mathcal{D}^{\ell}$。信源编码的一般框架可以描述为：

$$\text{编码} \quad \phi_n : \mathcal{X}^n \longmapsto \mathcal{D}^*$$

$$\text{译码} \quad \psi_n : \mathcal{D}^* \longmapsto \mathcal{X}^n$$

$$\text{码率} \quad R_n = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}^n} P_{\mathbf{X}}(\mathbf{x}) \ell\left(\phi_n(\mathbf{x})\right)$$

$$\text{译码错误} \quad \epsilon_n = \Pr\left\{\psi_n\left(\phi_n(\mathbf{X})\right) \neq \mathbf{X}\right\}$$

信源编码基本框架

R_n 中的$\ell(\phi_n(\mathbf{x}))$ 表示码字$\phi_n(\mathbf{x})$ 的长度。由此，我们知道码率表示在统计意义下每个信源符号所用的码字的平均长度。离散信源编码的问题就是通过证明ϕ_n 与ψ_n 的存在性，寻找满足$\lim_{n\to\infty} \epsilon_n = 0$ 的码率R_n 的下极限。

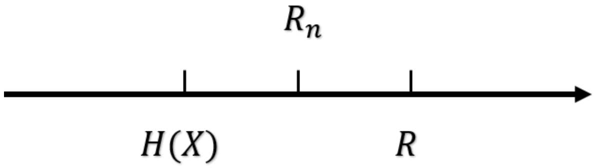
此外，根据序列或码字的长度是否固定，编译码大致可以分为四种类型：

1. 定长 $\longmapsto$ 定长
2. 定长 $\longmapsto$ 变长
3. 变长 $\longmapsto$ 定长
4. 变长 $\longmapsto$ 变长

信源编码定理

Theorem 4 (离散信源无失真编码定理)

给定一个离散无记忆信源，即一个独立同分布（*IID*）的随机变量序列$X_1, X_2, \cdots$，其熵为 $H(X)$。设码率$R > H(X)$，则存在固定码长编码(ϕ_n, ψ_n)，使得 R_n 满足$R_n \leq R$，并且$\lim_{n\to\infty} \epsilon_n = 0$。若允许变码长编码，则可以使得 $\epsilon_n = 0$。



信源编码定理的逆定理

Theorem 5 (离散信源无失真编码逆定理)

设 $R < H(X)$。则对于任何定长编码，若其码率 $R_n \le R < H(X)$，则必有 $\lim_{n\to\infty} \epsilon_n = 1$。

Consider a discrete memoryless source X with distribution $P_X(x), x \in \mathcal{X}$. A code of source $\mathcal{X}$ over the alphabet $\mathcal{D}$ consists of the following essentials.

$$\begin{aligned} \text{Encoding} \quad \phi: \quad & \mathcal{X} \to \mathcal{D}^* \\ & x \mapsto c(x) \\ \text{Average length} \quad & L = \textstyle\sum_{x \in \mathcal{X}} P_X(x)\ell(c(x)) \end{aligned}$$

Definition 6

1. A source code is called non-singular if $c(x) \neq c(x')$ for $x \neq x'$.
2. A source code is called uniquely decodable if $c(\mathbf{x}) \neq c(\mathbf{x}')$ for $\mathbf{x} \neq \mathbf{x}'$, where $c(\mathbf{x}) = c(x_1)c(x_2)\ldots c(x_n)$.
3. A source code is called to be a prefix code or instantaneous code if no codeword is a prefix of any other codeword.

Table: source codes

X	Singular	non-singular	Uniquely decodable	Instantaneous
1	0	0	10	0
2	0	010	00	10
3	0	01	11	110
4	0	10	110	111

Table: source codes

X	Singular	non-singular	Uniquely decodable	Instantaneous
1	0	0	10	0
2	0	010	00	10
3	0	01	11	110
4	0	10	110	111

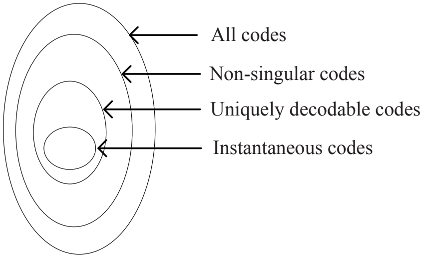


Figure: Classes of codes.

Shannon Code

Consider the following method for generating a code for a random variable X that takes on M values $\{1, 2, ..., M\}$ with probabilities $p_1, p_2, \cdots, p_M$. Assume that the probabilities are ordered so that $p_1 \geq p_2 \geq \cdots \geq p_M$. Define

$$F_i = \sum_{k=1}^{i-1} p_k$$

the sum of the probabilities of all symbols less than i. Then the codeword for i is the number $F_i \in [0, 1]$ rounded off to ℓ_i bits, where $\ell_i = \lceil \log \frac{1}{p_i} \rceil$.

Shannon Code

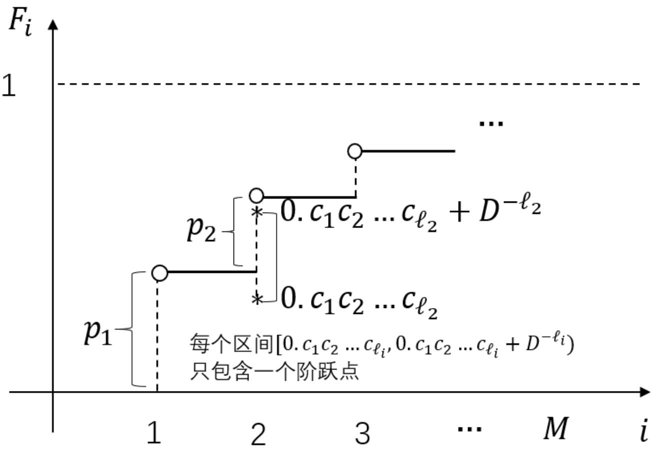


Figure: Shannon Code 图解.

Suppose that a probability mass function $\mathbf{q}$ is used in practice instead of the true probability mass function $\mathbf{p}$, then the Shannon code is of lengths $\ell_i = \lceil \log \frac{1}{q_i} \rceil$.

Theorem 7

The average length under $\mathbf{p}$ *of the Shannon code assignment* $\ell_i = \lceil \log \frac{1}{q_i} \rceil$ *satisfies*

$$H(\mathbf{p}) + D(\mathbf{p}\|\mathbf{q}) \le \mathbf{E}_{\mathbf{p}}[\ell(X)] < H(\mathbf{p}) + D(\mathbf{p}\|\mathbf{q}) + 1.$$

Remark: Using the wrong distribution incurs a penalty of $D(\mathbf{p}\|\mathbf{q})$ in the average description length.

Theorem 8 (Optimal properties)

For any distribution, there exists a binary optimal prefix code $\mathcal{C}$ such that

1. *If $p_j > p_k$, then $\ell_j \leq \ell_k$.*
2. *The two longest codewords have the same length.*
3. *The two longest codewords differ only in the last bit and corresponds to the two least likely symbols.*

Outline of proof:

1. If $p_j > p_k$, we swap their codewords to construct a new code $\mathcal{C}'$. Then
$$L' - L = \sum p_i \ell_i' - \sum p_i \ell_i = (p_j \ell_k + p_k \ell_j) - (p_j \ell_j + p_k \ell_k) = (p_j - p_k)(\ell_k - \ell_j)$$
Since $\mathcal{C}$ is optimal, then $L' - L \geq 0$. Hence $\ell_j \leq \ell_k$ as $p_j > p_k$.
2. If the two longest codewords are not of the same length, then we can delete the last bit of the longer one preserving the prefix condition and achieving lower average length.
3. If there is a maximal length codeword without a sibling, then we can delete the last bit preserving the prefix condition and achieving lower average length.

Theorem 9

If a binary code $\mathcal{C}^*$ is constructed by Huffman coding, then it is a binary optimal code.

Outline of proof: this theorem can be proved by induction.

Let $m = |\mathcal{X}|$ and the code for the source $\mathcal{X}$ is denoted by $\mathcal{C}_m$. Without loss of generality, we assume that $p_1 \geq p_2 \geq \ldots \geq p_m$.

(1) Let $\mathcal{C}_m$ be a code satisfying the optimal properties. Based on $\mathcal{C}_m$, a code $\mathcal{C}_{m-1}$ for $m-1$ symbols is construct as follows.

$$
\begin{array}{ccccccc}
 & \mathcal{C}_{m-1} & & & & \mathcal{C}_m & \\
p_1 & c'_1 & \ell'_1 & & p_1 & c_1 = c'_1 & \ell_1 = \ell'_1 \\
p_2 & c'_2 & \ell'_2 & & p_2 & c_2 = c'_2 & \ell_2 = \ell'_2 \\
\vdots & \vdots & \vdots & \Leftarrow & \vdots & \vdots & \vdots \\
p_{m-2} & c'_{m-2} & \ell'_{m-2} & & p_{m-2} & c_{m-2} = c'_{m-2} & \ell_{m-2} = \ell'_{m-2} \\
p_{m-1} + p_m & c'_{m-1} & \ell'_{m-1} & & p_{m-1} & c_{m-1} = c'_{m-1}0 & \ell_{m-1} = \ell'_{m-1} + 1 \\
 & & & & p_m & c_m = c'_{m-1}1 & \ell_m = \ell'_{m-1} + 1
\end{array}
$$

Shannon-Fano-Elias Coding

Assume that for a source $\mathcal{X} = \{1, 2, \ldots, m\}$, $p(x) > 0$ for all x. The cumulative distribution function is defined as

$$F(x) = \sum_{a \leq x} p(a). \tag{3}$$

Let $\bar{F}(x)$ be a modified cumulative distribution function as

$$\bar{F}(x) = \sum_{a < x} p(a) + \frac{1}{2} p(x). \tag{4}$$

The value of $\bar{F}(x)$ can be used as codeword for x. Since $\bar{F}(x)$ is a real number expressible only for an infinite number of bits, we round off $\bar{F}(x)$ to $\ell(x)$ bits and denote it by $\lfloor \bar{F}(x) \rfloor_{\ell(x)}$.

What should the length $\ell(x)$ be?

If $\ell(x)=\lceil\log\frac{1}{p(x)}\rceil+1$, then we have

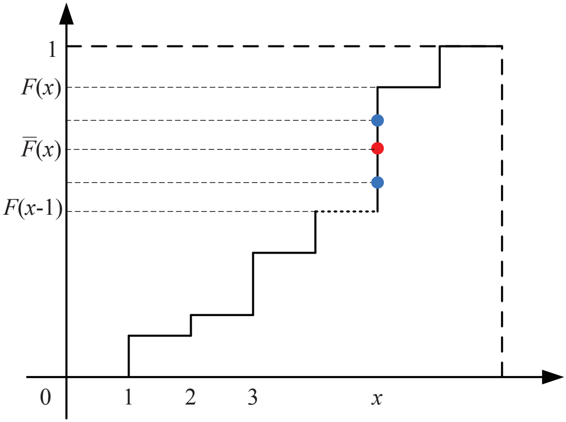
$$\bar{F}(x)-\lfloor\bar{F}(x)\rfloor_{\ell(x)}<2^{-\ell(x)}<\frac{p(x)}{2}=\bar{F}(x)-F(x-1) \tag{5}$$

Then $\lfloor\bar{F}(x)\rfloor_{\ell(x)}$ lies within the lower-half step corresponding to x. Thus $\ell(x)$ bits suffices to describe x.
On the other hand,

$$\lfloor\bar{F}(x)\rfloor_{\ell(x)}+2^{-\ell(x)}<\lfloor\bar{F}(x)\rfloor_{\ell(x)}+\frac{p(x)}{2}=\lfloor\bar{F}(x)\rfloor_{\ell(x)}+F(x)-\bar{F}(x)<F(x). \tag{6}$$

Let the bits corresponding to $\lfloor\bar{F}(x)\rfloor_{\ell(x)}$ be $z_1z_2\ldots z_\ell$. Then the interval corresponding to the codeword $z_1z_2\ldots z_\ell$ is $[0.z_1z_2\ldots z_\ell, 0.z_1z_2\ldots z_\ell+\frac{1}{2^\ell}]$. Such intervals are disjoint from the above two inequalities. Hence the code is prefix-free. And the average length is

$$L=\sum p(x)\left(\lceil\log\frac{1}{p(x)}\rceil+1\right)<\sum p(x)\left(\log\frac{1}{p(x)}+2\right)=H(X)+2. \tag{7}$$


1
F(x)
F̄(x)
F(x-1)
0
1
2
3
x

算术码编码

算术码编码的主要思想如下：设信源符号集包含 N 个符号，对每个符号从 $1 \sim N$ 进行编号。设每个符号出现的概率为 p_i，此处 $1 \leq i \leq N$。在初始区间给每个符号分配一个初始子区间，其长度等于对应符号的概率。每个序列的首个信源符号概率确定本序列编码的初始区间，后续信源符号的编码过程是对选定区间进行再分割的过程。

LZ78 算法：编码

设信源符号集 $\mathbf{A} = \{a_1, a_2, \cdots, a_k\}$ 共 K 个符号，设输入信源符号序列为 $\mathbf{u} = (u_1, u_2, \cdots, u_L)$ 。编码时将此序列分成不同的段。分段的规则为：尽可能取最少个相连的信源符号，并保证各段都不相同。

开始时，先取一个符号作为第一段，然后继续分段。若出现与前面相同的符号时，就再取紧跟后面的一个符号一起组成一个段，使之与前面的段不同。这些分段构成字典。当字典达到一定大小后，再分段时就应查看有否与字典中的短语相同，若有重复就添加符号，以便与字典中短语不同，直至信源符号序列结束。这样，不同的段内的信源符号可看成一短语，可得不同段所对应的短语字典表。

码字构成：前面字段所在的段号+末尾的一个符号对应的号。设 $\mathbf{u}$ 构成的字典中的短语共有 $M(\mathbf{u})$ 个。若编为二元码，段号所需码长 $n = \lceil \log M(\mathbf{u}) \rceil$，每个符号需要的码长为 $\lceil \log K \rceil$。单符号的码字段号为0。

LZ78 算法：译码

LZ78 编码的编码方法很便捷，译码也很简单，可以一边译码一边建立字典，只需要传输字典的大小，无需传输字典本身。当编码的信源序列较短时，LZ 算法性能似乎会变坏,但是当序列增长时，编码效率会提高，平均码长会逼近信源熵。

LZ78 算法

将有 K 个符号，长为 L 的信源序列 $\mathbf{u}$ 分为 $M(\mathbf{u})$ 个码段后，设最长的段的长度为 ℓ_{max}，可以证明，每个源符号的平均码长有

$$H(U)+\frac{\log K}{\ell_{max}}<\bar{n}<H(U)+\frac{\log K+2}{\ell_{max}}$$

将编码的信源序列趋于无穷时，ℓ_{max} 也趋于无穷，平均码长趋近于信源熵。

信息不等式直观

熵

$$H(X) = \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)}$$

熵是描述随机变量所需比特数的下确界，是熵密度的数学期望。我们可以认为 $H(X)$ 是随机变量的不确定性度量，是模糊度，是揭示 X 所需要的信息量，是得到 X 后所得到的信息量。

相对熵

设 $Q_X(x)$，$x \in \mathcal{X}$ 也是概率向量，即 $Q_X(x) \geq 0$，$\sum_{x\in\mathcal{X}} Q_X(x) = 1$。我们定义

$$D(P||Q) = \sum_{x\in\mathcal{X}} P_X(x) \log \frac{P_X(x)}{Q_X(x)}$$

称之为**相对熵**。通常情况下，$D(P||Q) \neq D(Q||P)$。

相对熵的含义可以粗略解释为，当 Shannon 编码器使用 Q 而不是 P 确定码长时所带来的码长代价。

条件熵

考虑随机向量 $(X,Y)\sim P_{X,Y}(x,y)$，其中$(x,y)\in\mathcal{X}\times\mathcal{Y}$，我们可以定义**条件熵**

$$H(X|Y)=\sum_{y\in\mathcal{Y}}P_Y(y)\sum_{x\in\mathcal{X}}P_{X|Y}(x|y)\log\frac{1}{P_{X|Y}(x|y)}$$

条件熵的含义可以解释为，若已知 Y 的条件下，描述 X 需要的比特数。

信息密度

给定 $P_{X,Y}(x,y)$，我们可以计算 $P_X(x)$，$P_Y(y)$，$P_{Y|X}(y|x)$，$P_{X|Y}(x|y)$。若不知道 X，则描述 Y 需 $H(Y)$ 比特。若已知 X，则描述 Y 需要 $H(Y|X)$ 比特。

给定 $X=x$，我们有 Y 的条件分布律 $P_{Y|x}(y|x)$，$y \in \mathcal{Y}$。我们定义

$$i(x;y) = \log \frac{P(y|x)}{P(y)}$$

为 (x,y) 点对应的**信息密度**。其可以理解为描述 y 的比特数在已知 $Y=y$ 前后的差： $\log \frac{1}{P(y)} - \log \frac{1}{P(y|x)}$，也可以看作自信息量的“减少”或模糊度的“减少”。

平均互信息

定义 $i(X;Y)$ 的均值为**互信息**，即

$$I(X;Y)=\sum_{x,y}p(x,y)\log\frac{p(y|x)}{p(y)}$$

可以证明

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

平均互信息

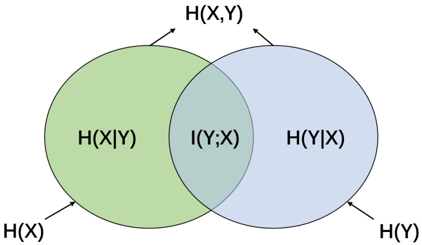


Figure: 互信息

从信源编码的角度，$I(X;Y)$ 可以看作是已知 X 条件下，描述 Y 的平均比特数减少，也可以认为是 X 中包含 Y 的信息量，还可以看作是已知 X的条件下，Y 的模糊度的减少量。

不等式的直观解释

1. $H(X) \geq 0$

   直观解释：一个信源序列的最小压缩速率是不可能小于0 的。

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

   直观解释：根据信道输出至少可以区分1 个信道输入序列，即，$2^{nI(X;Y)} \geq 1$，故而有上式成立。

3. $H(X) \leq \log|\mathcal{X}|$

   直观解释：假定信源X 的符号集是$\mathcal{X}$，对于一个n 长的信源序列，我们可以把n 个信源符号看作一个整体进行编码，即，用二进制展开表示$\mathcal{X}^n$ 中所有的序列。在这种方案下，每个n 长序列对应$\lceil\log(|\mathcal{X}|^n)\rceil$ 比特。当n 趋于无穷时，此方案的压缩速率恒为$\log|\mathcal{X}|$，而$H(X)$ 是压缩速率的下限，故而有上式成立。当X 服从均匀分布时，等号成立。

不等式的直观解释

4. $H(X,Y) \le H(X) + H(Y)$

   直观解释：假定有两个信源X 和Y, 不考虑关联性，分别对X 和Y 进行最优压缩，此方案的压缩速率为$H(X)+H(Y)$, 而$H(X,Y)$ 是这两个信源的压缩速率的下限，故而有上式成立。

不等式的直观解释

5. $H(X,Y)=H(X\mid Y)+H(Y)$

   直观解释：若想要对信源(X,Y)进行压缩，可以先对Y进行压缩，再对X进行压缩。H $(X\mid Y)$表示已知Y的条件下，对X进行压缩的最低压缩速率。我们通过随机装箱的办法对左式进行具体说明：先随机地将信源产生的X^n扔进一些箱子中，再将箱号作为压缩结果。在箱子的数目大约为$2^{nH(X|Y)}$的情况下，由于信源产生的典型的X^n大约有$2^{nH(X)}$个，因此每个箱子中会有$2^{nH(X)}/2^{nH(X|Y)}=2^{nI(X;Y)}$个。而根据对$Y^n$的观察，可以区分$2^{nI(X;Y)}$个$X^n$，因此要知道$X^n$只需要知道$X^n$在哪一个箱子即可。对箱子编号需要$nH(X\mid Y)$比特，压缩速率为$H(X\mid Y)$。采用这种分步的方案，和直接对$(X,Y)$进行压缩的最低压缩速率是相等的。

不等式的直观解释

6. $D(P\|Q) \geq 0$

   $D(P\|Q) = H(P,Q) - H(X), X$ 为服从概率分布P 的信源。根据定义，

$$D(P\|Q) = \mathbf{E}\left(\log \frac{P_X(X)}{Q_X(X)}\right) = \sum_{x\in\mathcal{X}} P_X(x)\log \frac{P_X(x)}{Q_X(x)}$$
$$H(P,Q) = \mathbf{E}\left(\log \frac{1}{Q_X(X)}\right) = \sum_{x\in\mathcal{X}} P_X(x)\log \frac{1}{Q_X(x)}$$

   相对熵就是交叉熵与熵之间的差。交叉熵表示在假定信源概率分布为Q 的情况下，对X 采用香农编码方案进行压缩的速率。而熵是信源的最低压缩速率。前者不会小于后者，因此相对熵一定大于等于零。

不等式的直观解释

7. $H(X \mid Y) \leq H(X)$

   直观解释：假定有一个信源同时产生关于(X, Y) 的n 长序列对，但是只需要对X 进行压缩。左式表示考虑了对Y 的观察，对X 进行压缩可以达到的最低压缩速率，而右式则是忽略了对Y 的观察，可达的最低压缩速率，即X 的熵。

不等式的直观解释

8. (Fano 不等式) 设X 是一个系统的输入，Y 是一个系统的输出，$\widehat{X}$ 是根据Y 对X 进行估计得到的结果，则$X \to Y \to \widehat{X}$ 构成一个马尔科夫链。记$P_e = \Pr\{\widehat{X} \neq X\}$，有

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

直观解释：不等式的左边$H(X \mid Y)$ 表示观测到$Y_1, Y_2, \cdots, Y_n$ 的条件下描述X 所需的最小比特数，而不等式的右边对应一种描述X 的方法所需的比特数。分三步：
1) 根据$Y_1, Y_2, \cdots, Y_n$ 可以估计得到$\widehat{X}_1, \widehat{X}_2, \cdots, \widehat{X}_n$;
2) 确定哪些位置的$\widehat{X}$ 是错的，平均需要$H(P_e)$ 比特;
3) 在错的位置描述X ，平均需要$\log(|\mathcal{X}| - 1)$ 比特。

不等式的直观解释

9. 设$X \to Y \to Z$ 构成一个马尔科夫链，有

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

直观解释：互信息越大，根据系统输出可以区分的系统输入序列越多，系统的可区分度越高。显然，根据Y 可区分的信道输入序列不会少于根据Z 可区分的信道输入序列。

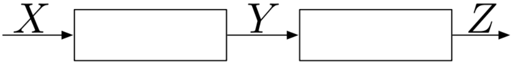


Figure: 互信息不增性

例子

Coin weighing. Suppose one has n coins, among which there may or may not be one counterfeit coin. If there is a counterfeit coin, it may be either heavier or lighter than the other coins. The coins are to be weighed by a balance.

(a) Find an upper bound on the number of coins n so that k weighings will find the counterfeit coin (if any) and correctly declare it to be heavier or lighter.

(b) *(Difficult)* What is the coin weighing strategy for $k = 3$ weighings and 12 coins?

谢谢！