

- 信息论的发展是以信息可以度量为基础的，度量信息的量称为信息量。
- 对于随机出现的事件，它的出现会给人们带来多大的信息量？
- 考虑到通信系统或很多实际的信息传输系统，对于所传输的消息如何用信息量的方法来描述？
- **第二章**信息的度量 将围绕这些问题展开讨论。

第二章 信息的度量

- 本章介绍信息论的一些基本概念，包括自信息量、平均自信息量、互信息量、平均互信息量、信息熵、熵的性质等，并解释了信息处理定理及其对信息处理的指导意义

本章问题的归纳:

- 什么叫不确定度?
- 什么叫自信息量?
- 什么叫平均不确定度?
- 什么叫信源熵?
- 什么叫平均自信息量?
- 什么叫条件熵?
- 什么叫联合熵?
- 联合熵、条件熵和熵的关系是什么?

- 什么叫后验概率？
- 什么叫互信息量？
- 什么叫平均互信息量？
- 什么叫疑义度？
- 什么叫噪声熵（或散布度）？
- 数据处理定理是如何描述的？
- 熵的性质有哪些？

第2章 信息的度量

1

自信息 平均自信息、信息熵

2

熵的性质 联合熵、条件熵

3

互信息 平均互信息

4

熵的关系(两个图)

5

数据处理定理

一、自信息 平均自信息 信息熵

- 1 信息度量的思路
- 2 自信息量的定义
- 3 信息量与不确定度
- 4 平均自信息量

1 信息度量的思路

信息的基本概念在于它的不确定性，任何已确定的事物都不含有信息。其**信息的直观认识**有：

- 第一个重要概念：信道上传送的是随机变量的值。
- 第二个重要概念：事件发生的概率越小，此事件含有的信息量就越大。
- 第三个重要概念：消息随机变量的随机性越大，此消息随机变量含有的信息量就越大。
- 第四个重要概念：两个消息随机变量的相互依赖性越大，它们的互信息量就越大。

由此可以合理地推算信源输出的信息量应该是输出事件的概率的减函数。

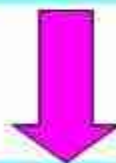
信息量的另一个直观属性是，某一输出事件的概率的微小变化不会很大地改变所传递的信息量，即信息量应该是信源输出事件概率的连续减函数。

- 若信源中事件 X_i 的出现所带来的信息量用 $I(x_i)$ 来表示并称之为事件 X_i 的自信息量,
- 则概率为 $p(x_i)$ 的信源输出 X_i 所包含的信息量 $I(x_i)$ 必须满足以下几个条件---信息度量的公理化条件

信息量的度量——公理化条件

- (1) $I(x_i)$ 是 $p(x_i)$ 的严格递减函数。当 $p(x_1) < p(x_2)$ 时, $I(x_1) > I(x_2)$, 概率越小, 事件发生的不确定性越大, 事件发生以后所包含的自信息量越大。
- (2) 极限情况下, 当 $p(x_i) = 0$ 时, $I(x_i) \rightarrow \infty$; 当 $p(x_i) = 1$ 时, $I(x_i) = 0$ 。
- (3) 从直观概念上讲, 由两个相对独立的不同的消息所提供的信息量, 应该等于它们分别提供的信息量之和, 即自信息量满足可加性。

问题: 什么函数能够同时满足以上条件呢?



对数函数

2 自信息量的定义

定义 2.1 随机事件的自信息量定义为该事件发生概率的对数的负值。
设事件 x_i 的概率为 $p(x_i)$ ，则它的自信息量定义为

$$I(x_i) \stackrel{def}{=} -\log p(x_i) = \log \frac{1}{p(x_i)}$$

自信息量的单位与所用对数的底有关

- (1) 通常取对数的底为2，信息量的单位为比特(bit, binary unit)
- (2) 若取自然对数（以e为底），信息量的单位为奈特(nat, natural unit)
- (3) 工程上用以10为底较方便。若以10为对数底，则自信息量的单位为哈特莱(Hartley)，也有的称为笛特(det)

这三个信息量单位之间的转换关系如下：

$$\begin{aligned} 1 \text{ nat} &= \log_2 e & 1.433 \text{ bit}, \\ 1 \text{ det} &= \log_2 10 & 3.322 \text{ bit} \end{aligned}$$

例子-1

$$\log 3 = 1.5850$$

$$\log 5 = 2.3219$$

$$\log 7 = 2.8074$$

$$\log 11 = 3.4594$$

- 一个**0, 1**等概的二进制随机序列，求任一码元的自信息量。
- 解：任一码元不是为**0**就是为**1**
- 因为 **$p(0) = p(1) = 1/2$**
- 所以 **$I(0) = I(1) = -\log(1/2) = 1(\text{bit})$**

例子-2

- 对于 2^n 进制的数字序列, 假设每一符号的出现完全随机且概率相等, 求任一符号的自信息量。
- 解: 设 2^n 进制数字序列任一码元 x_i 的出现概率为 $p(x_i)$, 根据题意:

$$p(x_i) = 1/2^n$$

$$I(x_i) = -\log(1/2^n) = n \text{ (bit)}$$

例子-3

- 英文字母中“e”的出现概率为0.105，“c”的出现概率为0.023，“o”的出现概率为0.001。分别计算它们的自信息量。

解：

“e”的自信息量 $I(e) = -\log_2 0.105 = 3.25 \text{ bit}$

“c”的自信息量 $I(c) = -\log_2 0.023 = 5.44 \text{ bit}$

“o”的自信息量 $I(o) = -\log_2 0.001 = 9.97 \text{ bit}$

自信息含义

$$I(x_i) = \log \frac{1}{p(x_i)}$$

- 当事件 x_i 发生以前：表示事件 x_i 发生的不确定性。
- 当事件 x_i 发生以后：表示事件 x_i 所含有（或所提供）的信息量。

信息量与不确定性的关系

- 信息量的直观定义：

收到某消息获得的信息量 = 不确定性减少的量

- 信源中某一消息发生的不确定性越大，一旦它发生，并为受信者收到后，消除的不确定性就越大，获得的信息也就越大。

4 平均自信息

概率空间

通常把一个随机变量的所有可能的取值和这些取值对应的概率称为它的概率空间。

$$\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ p(x_1) & p(x_2) & \cdots & p(x_n) \end{bmatrix}$$

平均自信息的定义

定义 2.3 随机变量 X 的每一个可能取值的自信息 $I(x_i)$ 的统计平均值定义为随机变量 X 的平均自信息量。又称为信息熵、信源熵，简称熵。

$$H(X) = E[I(x_i)] = -\sum_{i=1}^q p(x_i) \log p(x_i) \quad (2.3)$$

信源熵：表征信源的平均不确定度。

平均自信息：消除信源不确定度时所需要的信息的量度，即收到一个信源符号，全部解除了这个符号的不确定度。

信源熵的三种物理含义

- 信源熵是从平均意义上来表征信源的总体特性的一个量，因此信源熵有以下三种物理意义：

信源熵 $H(X)$ 是表示信源输出后每个消息/符号所提供的平均信息量；

信源熵 $H(X)$ 是表示信源输出前，信源的平均不确定性；

用信源熵 $H(X)$ 来表征变量 X 的随机性。

■ 例-4

- 有两个信源空间，其概率空间分别为：

$$\left[\begin{array}{c} X \\ P(X) \end{array} \right] \left\{ \begin{array}{c} x_1, x_2 \\ 0.99, 0.01 \end{array} \right\}$$

$$\left[\begin{array}{c} Y \\ P(Y) \end{array} \right] \left\{ \begin{array}{c} y_1, y_2 \\ 0.5, 0.5 \end{array} \right\}$$

- 信息熵分别为：
- $H(X) = -0.99\log 0.99 - 0.01\log 0.01 = 0.08$ 比特/符号
- $H(Y) = -0.5\log 0.5 - 0.5\log 0.5 = 1$ 比特/符号
- 可见 $H(Y) > H(X)$

例-5

一个布袋内放100个球，其中80个球是红色的，20个球是白色的，若随机摸取一个球，猜测其颜色，求平均摸取一次所能获得的自信息量。

这一随机事件的概率空间为：

$$\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \\ 0.8 & 0.2 \end{bmatrix}$$

为红球，信息量 $I(x_1) = -\log_2 p(x_1) = -\log_2 0.8 \text{ bit}$

为白球，信息量 $I(x_2) = -\log_2 p(x_2) = -\log_2 0.2 \text{ bit}$

每次摸出一个球后又放回袋中，再进行下一次摸取。随机摸取n次后总共所获得的信息量为

$$np(x_1)I(x_1) + np(x_2)I(x_2)$$

- 平均随机摸取一次所获得的信息量则为:

$$\begin{aligned} H(X) &= \frac{1}{n} [np(x_1)I(x_1) + np(x_2)I(x_2)] \\ &= -[p(x_1)\log_2 p(x_1) + p(x_2)\log_2 p(x_2)] \\ &= -\sum_{i=1}^2 p(x_i)\log_2 p(x_i) \\ &= 0.72 \text{ 比特/次} \end{aligned}$$

例 -6

- 每帧电视图像可以认为是由 500×600 个格点组成，所有像素均是独立变化，且每像素又取10个不同的灰度等级，并设亮度电平是等概率出现，问每帧图像含有多少信息量？若是一个广播员，在约10000个汉字中选1000个汉字来口述这电视图像，试问广播员描述此图像所广播的信息量是多少（假设汉字字汇是等概率分布，并彼此无依赖）？若要恰当地描述此图像，广播员在口述中至少需要多少篇这样的千字文？

- 电视屏上约有 $500 \times 600 = 3 \times 10^5$ 个格点，按每点有 10 个不同的灰度等级考虑，则共能组成 $n = 10^{3 \times 10^5}$ 个不同的画面。按等概率
- $1/10^{3 \times 10^5}$ 计算，平均每个画面可提供的信息量为

$$\begin{aligned} H(X) &= - \sum_{i=1}^n p(x_i) \log_2 p(x_i) = -\log_2 10^{-3 \times 10^5} \\ &= 3 \times 10^5 \times 3.32 \text{ 比特/画面} \end{aligned}$$

- 有一篇千字文章，假定每字可从万字表中任选，则共有不同的千字文

$$N=10000^{1000}=10^{4000} \text{ 篇}$$

仍按等概率 $1/10^{4000}$ 计算，平均每篇千字文可提供的信息量为

$$\begin{aligned} H(X) &= \log_2 N \\ &= 4 \times 10^3 \times 3.32 \\ &\approx 1.3 \times 10^4 \text{ 比特 / 千字文} \end{aligned}$$

比较：

- “一个电视画面”平均提供的信息量远远超过“一篇千字文”提供的信息量。

例-7

- 设信源符号集 $X=\{x_1, x_2, x_3\}$ ，每个符号发生的概率分别为 $P(x_1)=1/2$ ， $P(x_2)=1/4$ ， $P(x_3)=1/4$ 。求信源熵。

$$\begin{aligned} H(X) &= 1/2 \log_2 2 + 1/4 \log_2 4 + 1/4 \log_2 4 \\ &= 1.5 \text{ 比特/符号} \end{aligned}$$

[返回](#)

第2章 信息的度量

1

自信息 平均自信息、信息熵

2

熵的性质 联合熵、条件熵

3

互信息 平均互信息

4

熵的关系(两个图)

5

数据处理定理

二、熵的性质、联合熵、条件熵

- 1 熵函数
- 2 熵的性质
- 3 联合自信息量与条件自信息量
- 4 联合熵、条件熵
- 5 联合熵和条件熵的关系

1 熵函数

- 信息熵 $H(X)$ 是随机变量 X 的概率分布的函数，所以又称为熵函数。如果把概率分布 $p(x_i), i=1, 2, \dots, q$ 记为 p_1, p_2, \dots, p_q ，则熵函数又可以写成概率矢量 $\mathbf{p} = (p_1, p_2, \dots, p_q)$ 的函数的形式，记为 $H(\mathbf{p})$

- $$H(X) = -\sum_{i=1}^q p_i \log p_i = H(p_1, p_2, \dots, p_q) = H(\mathbf{p})$$

2 熵函数的性质

熵函数的性质—重要

(1).对称性:

$$H(p_1, p_2, \dots, p_q) = H(p_2, p_1, \dots, p_q) = \dots = H(p_q, p_1, \dots, p_{q-1})$$

说明熵函数仅与信源的总体统计特性有关。

(2).确定性:

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

在概率矢量中，只要有一个分量为1，其它分量必为0，它们对熵的贡献均为0，因此熵等于0。也就是说确定信源的不确定度为0。

(3).非负性:

$$H(\mathbf{p}) = H(p_1, p_2, \dots, p_q) \geq 0$$

对确定信源，等号成立。信源熵是自信息的数学期望，自信息是非负值，所以信源熵必定是非负的。

(4).扩展性:

$$\lim_{\varepsilon \rightarrow 0} H_{q+1}(p_1, p_2, \dots, p_q - \varepsilon, \varepsilon) = H_q(p_1, p_2, \dots, p_q)$$

这个性质的含义是增加一个基本不会出现的小概率事件，信源的熵保持不变。

(5).连续性:

$$\lim_{\varepsilon \rightarrow 0} H(p_1, p_2, \dots, p_{q-1} - \varepsilon, p_q + \varepsilon) = H(p_1, p_2, \dots, p_q)$$

即信源概率空间中概率分量的微小波动, 不会引起熵的变化。

(6).递推性:

$$H(p_1, p_2, \dots, p_{n-1}, q_1, q_2, \dots, q_m) = H(p_1, p_2, \dots, p_n)$$

$$H(p_1, p_2, \dots, p_{n-1}, q_1, q_2, \dots, q_m) = H(p_1, p_2, \dots, p_n) + p_n H\left(\frac{q_1}{p_n}, \frac{q_2}{p_n}, \dots, \frac{q_m}{p_n}\right)$$

这性质表明, 假如有一信源的 n 个元素的概率分布已知, 其中某个元素又被划分成 m 个元素, 这 m 个元素的概率之和等于该元素的概率, 这样得到的新信源的熵增加, 熵增加了一项是由于划分产生的不确定性。

(7).极值性 (最大熵定理)

$$H(p_1, p_2, \dots, p_n) \leq H\left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right) = \log n$$

式中 n 是随机变量 X 的可能取值的个数。

极值性表明离散信源中各消息等概率出现时熵最大，这就是最大离散熵定理。连续信源的最大熵则与约束条件有关。

(8).上凸性:

$H(\mathbf{p})$ 是严格的上凸函数, 设

$$\mathbf{p} = (p_1, p_2, \dots, p_q), \mathbf{p}' = (p_1', p_2', \dots, p_q'), \sum_{i=1}^q p_i = 1, \sum_{i=1}^q p_i' = 1$$

则对于任意小于1的正数 α , ($0 < \alpha < 1$) 有以下不等式成立:

$$H[\alpha \mathbf{p} + (1-\alpha)\mathbf{p}'] > \alpha H(\mathbf{p}) + (1-\alpha)H(\mathbf{p}')$$

凸函数在定义域内的极值必为极大值, 可以利用熵函数的这个性质可以证明熵函数的极值性。

例-8

设信源 $\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{Bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ 0.2 & 0.19 & 0.18 & 0.17 & 0.16 & 0.17 \end{Bmatrix}$, 求这
信源的熵

3 联合自信息量与 条件自信息量

联合自信息量

■ 概率空间 $\left[\begin{matrix} XY \\ P(XY) \end{matrix} \right] = \left\{ \begin{matrix} x_1y_1, \dots, x_1y_m, & x_2y_1, \dots, & x_2y_m, \dots, & x_ny_1, \dots, & x_ny_m \\ p(x_1y_1), \dots, p(x_1y_m), & p(x_2y_1), \dots, & p(x_2y_m), \dots, & p(x_ny_1), \dots, & p(x_ny_m) \end{matrix} \right\}$

■ 其中 $0 \leq p(x_iy_j) \leq 1 \ (i=1,2,\dots,n; j=1,2,\dots,m)$ $\sum_{i=1}^n \sum_{j=1}^m p(x_iy_j) = 1$

■ 则联合自信息量为 $I(x_iy_j) = \log_2 \frac{1}{p(x_iy_j)}$

■ 当 **X** 和 **Y** 相互独立时, $p(x_iy_j) = p(x_i)p(y_j)$

$$I(x_iy_j) = \log_2 \frac{1}{p(x_i)p(y_j)} = \log_2 \frac{1}{p(x_i)} + \log_2 \frac{1}{p(y_j)} = I(x_i) + I(y_j)$$

■ 两个随机事件相互独立时, 同时发生得到的信息量, 等于各自自信息量之和。

条件自信息量

设 y_j 条件下, 发生 x_i 的条件概率 $p(x_i / y_j)$
那么它的条件自信息量 $I(x_i / y_j)$ 定义为:

$$I(x_i / y_j) = -\log p(x_i / y_j)$$

表示在特定条件下 (y_j 已定) 随机事件 x_i 所带来的信息量

自信息量、条件自信息量和联合自信息量之间的关系:

$$\begin{aligned} I(x_i y_j) &= \log_2 \frac{1}{p(x_i) p(y_j / x_i)} = I(x_i) + I(y_j / x_i) \\ &= \log_2 \frac{1}{p(y_j) p(x_i / y_j)} = I(y_j) + I(x_i / y_j) \end{aligned}$$

4 联合熵与条件熵

联合熵

一个随机变量的不确定性可以用熵来表示，这一概念可以方便地推广到多个随机变量。

定义2.4 二维随机变量 XY 的概率空间表示为

$$\begin{bmatrix} XY \\ P(XY) \end{bmatrix} = \begin{bmatrix} x_1 y_1 & \cdots & x_i y_j & \cdots & x_n y_m \\ p(x_1 y_1) \cdots p(x_i y_j) \cdots p(x_n y_m) \end{bmatrix}$$

其中 $p(x_i y_j)$ 满足概率空间的非负性和完备性：

$$0 \leq p(x_i y_j) \leq 1, \sum_{i=1}^n \sum_{j=1}^m p(x_i y_j) = 1$$

二维随机变量 XY 的**联合熵**定义为联合自信息的数学期望，它是二维随机变量 XY 的不确定性的度量。

$$H(XY) \stackrel{\text{def}}{=} \sum_{i=1}^n \sum_{j=1}^m p(x_i y_j) I(x_i y_j) = - \sum_{i=1}^n \sum_{j=1}^m p(x_i y_j) \log p(x_i y_j)$$

条件熵

定义2.5 给定 X 时, Y 的条件熵:

$$\begin{aligned} H(Y|X) &= \sum_i p(x_i) H(Y|x_i) = - \sum_i \sum_j p(x_i) p(y_j|x_i) \log p(y_j|x_i) \\ &= - \sum_i \sum_j p(x_i y_j) \log p(y_j|x_i) \end{aligned}$$

其中, $H(Y|X)$ 表示已知 X 时, Y 的平均不确定性。

5 联合熵和条件熵的关系

(1) 联合熵与信息熵、条件熵的关系:

$$H(YX) = H(X) + H(Y|X)$$

这个关系可以方便地推广到 N 个随机变量的情况:

$$H(X_1X_2\cdots X_N) = H(X_1) + H(X_2|X_1) + \cdots + H(X_N|X_1X_2\cdots X_{N-1})$$

称为**熵函数的链规则**。

推论: 当二维随机变量 X, Y 相互独立时, 联合熵等于 X 和 Y 各自熵之和: $H(XY) = H(X) + H(Y)$

(2). 条件熵与信息熵的关系:

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

(3). 联合熵和信息熵的关系:

$$H(XY) \leq H(X) + H(Y) \quad \text{当} X, Y \text{相互独立时等号成立。}$$

第2章 信息的度量

1

自信息 平均自信息、信息熵

2

熵的性质 联合熵、条件熵

3

互信息 平均互信息

4

熵的关系(两个图)

5

数据处理定理

三、互信息、平均互信息

- 1 互信息的定义
- 2 平均互信息的定义
- 3 平均互信息性质
- 4 通信系统中的互信息

1 互信息

定义 2.2 一个事件 y_j 所给出关于另一个事件 x_i 的信息定义为互信息，用 $I(x_i; y_j)$ 表示。

$$I(x_i; y_j) = I(x_i) - I(x_i | y_j) = \log \frac{p(x_i | y_j)}{p(x_i)} \quad (2.2)$$

定义：后验概率与先验概率比值的对数。

$$I(x_i; y_j) = \log \frac{p(x_i | y_j)}{p(x_i)}$$

① 信源X的先验概率 $p(x_i)$

由于信宿事先不知道信源在某一时刻发出的是哪一个符号,所以每个符号消息是一个随机事件。信源发出符号通过有干扰的信道传递给信宿。通常信宿可以预先知道信息X发出的各个符号消息的集合,以及它们的概率分布,即预知信源X的先验概率 $p(x_i)$ 。

② 后验概率

当信宿收到一个符号消息 y_j 后,信宿可以计算信源各消息的条件概率 $p(x_i/y_j)$, $i=1,2,\dots,N$,这种条件概率称为后验概率。

某地二月份天气出现的概率分别为 晴 $1/2$,
阴 $1/4$, 雨 $1/8$, 雪 $1/8$. 某一天有人告诉你:
“今天不是晴天”, 把这句话作为收到的消息 y_1 ,
求收到 y_1 后, y_1 与各种天气的互信息量.

例-11

某地二月份天气构成的信源为

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} x_1(\text{晴}), & x_2(\text{阴}), & x_3(\text{雨}), & x_4(\text{雪}) \\ \frac{1}{2}, & \frac{1}{4}, & \frac{1}{8}, & \frac{1}{8} \end{bmatrix}$$

收到消息 y_1 : “今天不是晴天”

收到 y_1 后: $p(x_1/y_1)=0$, $p(x_2/y_1)=1/2$,
 $p(x_3/y_1)=1/4$, $p(x_4/y_1)=1/4$

- 计算 y_1 与各种天气之间的互信息量
- 对天气 x_1 , 不必再考虑
- 对天气 x_2 ,
$$I(x_2; y_1) = \log_2 \frac{p(x_2/y_1)}{p(x_2)} = \log_2 \frac{1/2}{1/4} = 1(\text{比特})$$
- 对天气 x_3 ,
$$I(x_3; y_1) = \log_2 \frac{p(x_3/y_1)}{p(x_3)} = \log_2 \frac{1/4}{1/8} = 1(\text{比特})$$
- 对天气 x_4 ,
$$I(x_4; y_1) = \log_2 \frac{p(x_4/y_1)}{p(x_4)} = \log_2 \frac{1/4}{1/8} = 1(\text{比特})$$
- 结果表明从 y_1 分别得到了各1比特的信息量;
- 或者说 y_1 使 x_2 , x_3 , x_4 的不确定度各减少量1比特。

2 平均互信息

为了从整体上表示从一个随机变量 Y 所给出关于另一个随机变量 X 的信息量，我们定义互信息 $I(x_i; y_j)$ 在的 XY 联合概率空间中的统计平均值为随机变量 X 和 Y 间的**平均互信息**：

定义2.6

$$\begin{aligned} I(X; Y) &= \sum_{i=1}^n \sum_{j=1}^m p(x_i y_j) I(x_i; y_j) = \sum_{i=1}^n \sum_{j=1}^m p(x_i y_j) \log \frac{p(x_i | y_j)}{p(x_i)} \\ &= \sum_{i=1}^n \sum_{j=1}^m p(x_i y_j) \log \frac{1}{p(x_i)} - \sum_{i=1}^n \sum_{j=1}^m p(x_i y_j) \log \frac{1}{p(x_i | y_j)} \\ &= H(X) - H(X | Y) \end{aligned}$$

说明:

在通信系统中，若发端的符号是 X ，而收端的符号是 Y ， $I(X; Y)$ 就是在接收端收到 Y 后所能获得的关于 X 的信息。

性质:

$$1) \quad I(X; Y) = H(X) - H(X/Y)$$

$$2) \quad I(Y; X) = H(Y) - H(Y/X)$$

$$I(Y; X) = I(X; Y)$$

$$I(X; Y) = H(X) - H(X/Y)$$

证明:

$$\begin{aligned} I(X; Y) &= \sum_{i,j} p(y_j) p(x_i / y_j) \log \frac{p(x_i / y_j)}{p(x_i)} \\ &= \sum_{i,j} p(y_j) p(x_i / y_j) \log P(x_i / y_j) \\ &\quad - \sum_{i,j} p(y_j) p(x_i / y_j) \log P(x_i) \\ &= \sum_{i,j} p(x_i y_j) \log P(x_i / y_j) \\ &\quad - \sum_i p(x_i) \log P(x_i) \\ &= H(X) - H(X/Y) \end{aligned}$$

$$I(Y; X) = H(Y) - H(Y/X)$$

证明-自行练习

3 平均互信息性质

(1) 非负性: $I(X;Y) \geq 0$

平均互信息是非负的, 说明给定随机变量 Y 后, 一般来说总能消除一部分关于 X 的不确定性。

(2) 互易性 (对称性): $I(X;Y) = I(Y;X)$

对称性表示 Y 从 X 中获得关于的信息量等于 X 从 Y 中获得关于的信息量。

(3) 平均互信息和各类熵的关系:

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

当 X, Y 统计独立时, $I(X;Y) = 0$

(4) 极值性: $I(X;Y) \leq H(X), I(X;Y) \leq H(Y)$

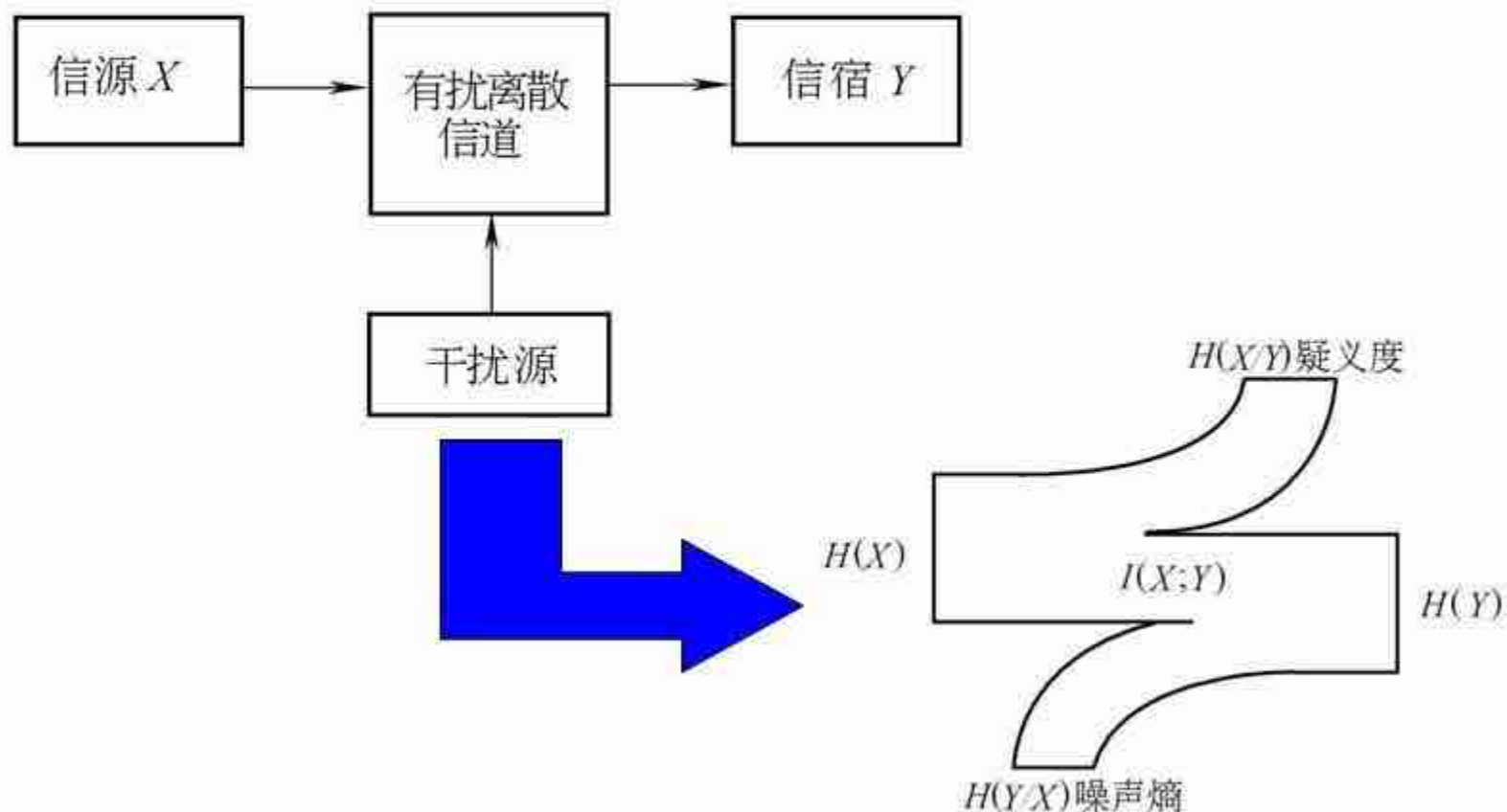
极值性说明从一个事件提取关于另一个事件的信息量，至多只能是另一个事件的平均自信息量那么多，不会超过另一事件本身所含的信息量。

(5) 凸函数性:

定理2.1 当条件概率分布 $\{p(y_j | x_i)\}$ 给定时，平均互信息 $I(X;Y)$ 是输入分布 $\{p(x_i)\}$ 的上凸函数。

定理2.2 对于固定的输入分布 $\{p(x_i)\}$ ，平均互信息量 $I(X;Y)$ 是条件概率分布 $\{p(y_j | x_i)\}$ 的下凸函数。

4 通信系统中的平均互信息



什么叫疑义度/损失熵?

Y关于X的后验不确定度。表示收到变量Y后，对随机变量X仍然存在的不确定性，代表了信道中损失的信息，故称为疑义度。

$$I(X; Y) = H(X) - H(X/Y)$$

这里的平均互信息量表征了对接收的每一个符号的正确性所产生怀疑的程度，所以条件熵 $H(X/Y)$ 又称之为疑义度

什么叫噪声熵或散布度？

条件熵 $H(Y/X)$ 唯一地确定信道噪声所需要的平均信息量，故又称**噪声熵**或**散布度**。是由信道噪声造成的

$$I(Y; X) = H(Y) - H(Y/X)$$

说明：

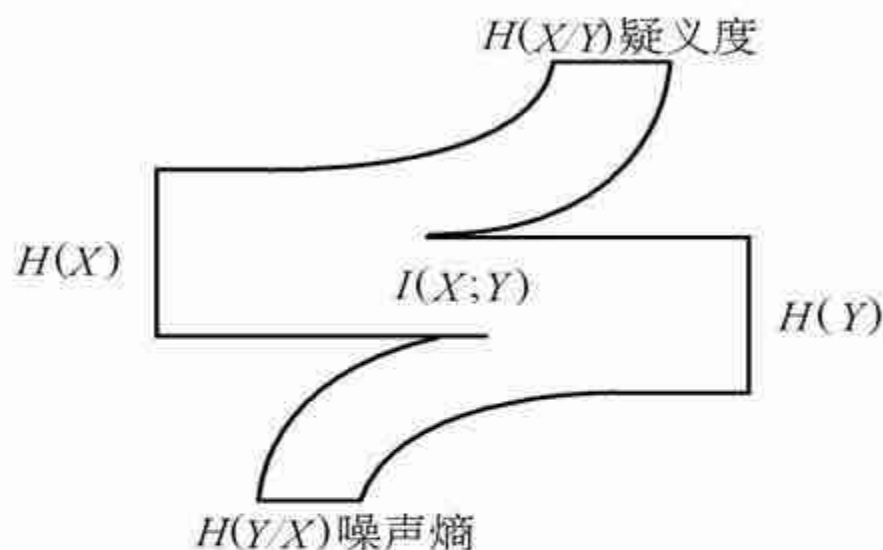
平均互信息量可看作在有扰离散信道上传递消息时，唯一地确定接收符号 y 所需要的平均信息量 $H(Y)$ ，减去当信源发出符号为已知时需要确定接收符号 y 所需要的平均信息量 $H(Y/X)$ 。

特例

- ① 如果 X 与 Y 是相互独立的，无法从 Y 中去提取关于 X 的信息，即 $H(X/Y)=H(X)$ ，故称为全损离散信道。
- ② 如果 Y 是 X 的确定的一一对应函数， $I(X;Y)=H(X)$ ，已知 Y 就完全解除了关于 X 的不确定度，所获得的信息就是 X 的不确定度或熵。这可看成无扰离散信道，疑义度 $H(X/Y)$ 为零，噪声熵也为零。

在一般情况下， X 和 Y 既非相互独立，也不是一一对应，那么从 Y 获得 X 的信息必在零与 $H(X)$ 之间，即常小于 X 的熵。

从通信过程看互信息的性质



$$I(X; Y) = H(X) - H(X/Y)$$

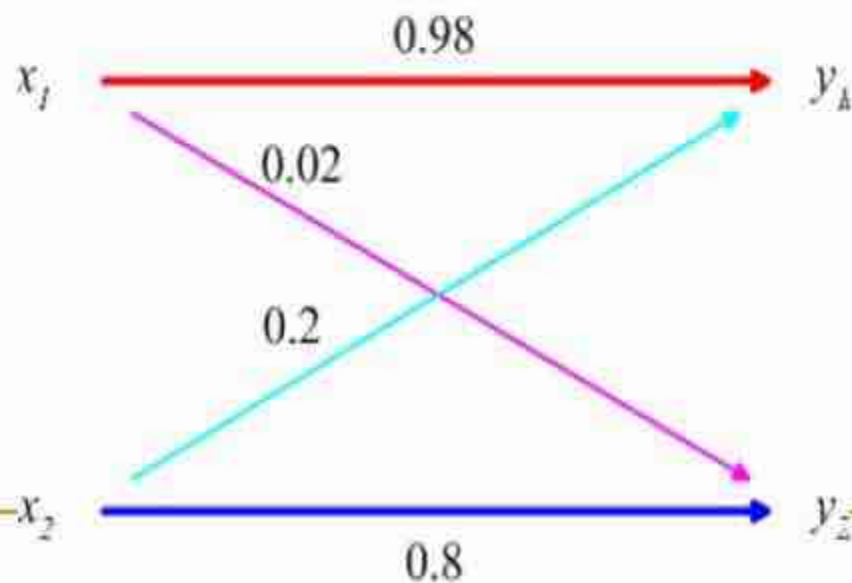
$$I(Y; X) = H(Y) - H(Y/X)$$

$$I(X; Y) = H(X) + H(Y) - H(XY)$$

例-12

- 已知信源概率空间为 $\left[\begin{matrix} X \\ P(X) \end{matrix} \right] \left\{ \begin{matrix} x_1, x_2 \\ 0.5, 0.5 \end{matrix} \right\}$ ，其在如下

的信道上传输，求在该信道上传输的平均互信息量 $I(X; Y)$ ，疑义度 $H(X|Y)$ ，噪声熵 $H(Y|X)$ ，联合熵 $H(XY)$ 。



返回

求解步骤:

- 求联合概率
- 求Y的各消息概率
- 求X的后验概率
- 求信息熵和联合熵
- 求平均互信息
- 求疑义度
- 求噪声熵

第2章 信息的度量

1

自信息 平均自信息、信息熵

2

熵的性质 联合熵、条件熵

3

互信息 平均互信息

4

熵的关系(两个图)

5

数据处理定理

四、熵的关系

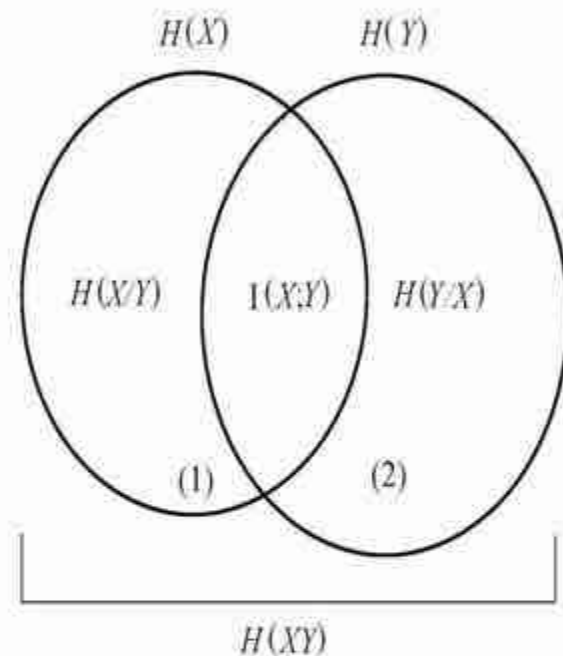
图中两圆外轮廓表示联合熵 $H(XY)$ ，圆(1)表示 $H(X)$ ，圆(2)表示 $H(Y)$ ，则
 $H(XY)=H(X)+H(Y/X)=H(Y)+H(X/Y)$
 $H(X)\geq H(X/Y)$ ， $H(Y)\geq H(Y/X)$
 $I(X;Y)=H(X)-H(X/Y)=H(Y)-H(Y/X)$
 $=H(X)+H(Y)-H(XY)$
 $H(XY)\leq H(X)+H(Y)$

如果 X 与 Y 互相独立，则

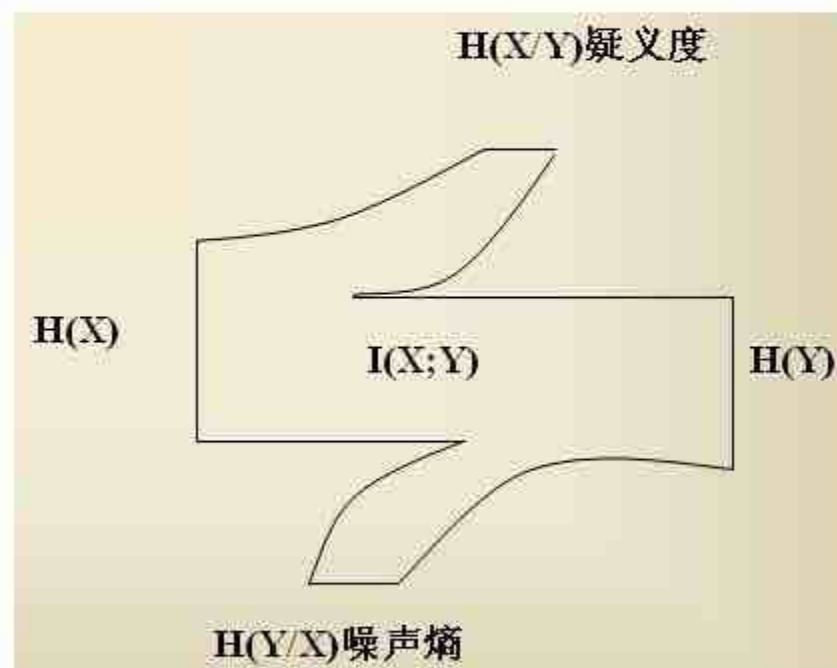
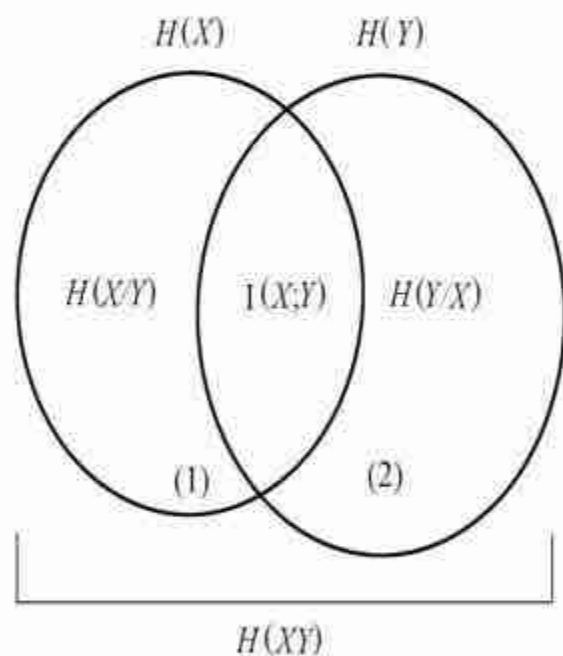
$$I(X;Y)=0$$

$$H(XY)=H(X)+H(Y)$$

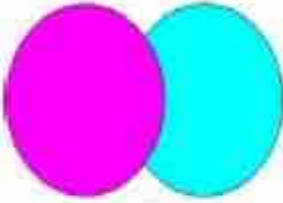
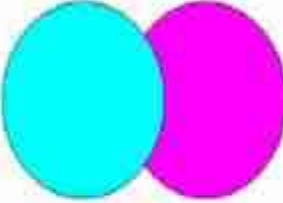
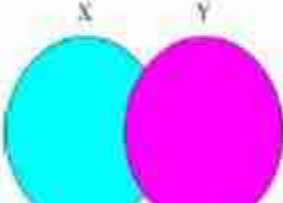

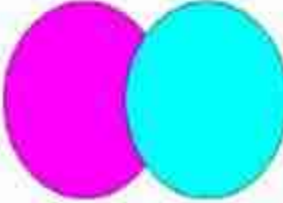
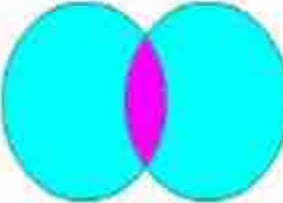
$$H(X)=H(X/Y), H(Y)=H(Y/X)$$



各种熵之间的关系图示



熵的关系-表

名称	符号	关系	图示	名称	符号	关系	图示
无条件熵	$H(X)$	$H(X) = H(X,Y) - I(X;Y)$ $\geq H(X Y)$ $H(X) = H(X,Y) - H(Y X)$	 $H(X)$	条件熵	$H(Y X)$	$H(Y X) = H(X,Y) - H(X)$ $= H(Y) - I(X;Y)$	 $H(Y X)$
	$H(Y)$	$H(Y) = H(X,Y) - I(X;Y)$ $\geq H(Y X)$ $H(Y) = H(X,Y) - H(X Y)$	 $H(Y)$	联合熵	$H(X,Y) = H(Y,X)$	$H(X,Y) = H(X) + H(Y X)$ $= H(Y) + H(X Y)$ $= H(X) + H(Y) - I(X;Y)$ $= H(X Y) + H(Y X) + I(X;Y)$	 $H(X,Y)$
条件熵	$H(X Y)$	$H(X Y) = H(X,Y) - H(Y)$ $= H(X) - I(X;Y)$	 $H(X Y)$	交互熵	$I(X;Y) = I(Y;X)$	$I(X;Y) = H(X) - H(X Y)$ $= H(Y) - H(Y X)$ $= H(X,Y) - H(Y X) - H(X Y)$ $= H(X) + H(Y) - H(X,Y)$	 $I(X,Y)$

第2章 信息的度量

1

自信息 平均自信息、信息熵

2

熵的性质 联合熵、条件熵

3

互信息 平均互信息

4

熵的关系(两个图)

5

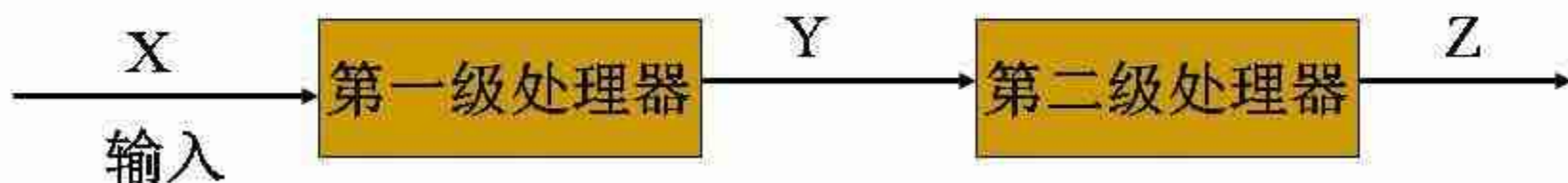
数据处理定理

五、数据处理定理

- 1 数据处理中信息的变化
- 2 平均条件自信息和平均联合自信息
- 3 数据处理定理

1 数据处理中信息的变化

当消息通过多级处理器时，随着处理器数目的增多，输入消息与输出消息之间的平均互信息量趋于变小。



级联处理器

结论：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，所谓**信息不增性**。

- 在一些实际通信系统中，常常出现串联信道。
例如：
 - 微波中继接力通信就是一种串联信道。
 - 信宿收到数据后再进行数据处理，数据处理系统可看成一种信道，它与前面传输数据的信道构成串联信道。

2 平均条件互信息和平均联合自信息

为了证明数据处理定理，我们需要引入三元随机变量 X, Y, Z 的平均条件互信息和平均联合互信息的概念。

定义2.7 平均条件互信息

$$I(X; Y | Z) = E \{ I(xy | z) \} = \sum_x \sum_y \sum_z p(xyz) \log \frac{p(x | yz)}{p(x | z)}$$

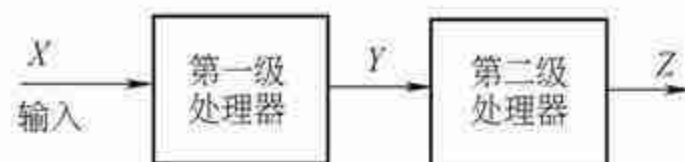
它表示随机变量 Z 给定后，从随机变量 Y 所得到得关于随机变量 X 的信息量。

定义2.8 平均联合互信息

$$I(X; YZ) = E \{ I(x; yz) \} = \sum_x \sum_y \sum_z p(xyz) \log \frac{p(x | yz)}{p(x)}$$

它表示从二维随机变量 YZ 所得到得关于随机变量 X 的信息量。

3 数据处理定理



定理2.3 (数据处理定理)

如果随机变量 X, Y, Z 构成一个马尔可夫链，则有以下关系成立：

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

等号成立的条件是对于任意的 X, Y, Z ，有

$$p(x|yz) = p(x|z) \quad p(z|xy) = p(z|x)$$

数据处理定理再一次说明，在任何信息传输系统中，最后获得的信息至多是信源所提供的信息，如果一旦在某一过程中丢失一些信息，以后的系统不管如何处理，如不触及丢失信息的输入端，就不能再恢复已丢失的信息，这就是信息不增性原理，它与热熵不减原理正好对应，反映了信息的物理意义。

结论

- 两级串联信道输入与输出消息之间的平均互信息量既不会超过第 I 级信道输入与输出消息之间的平均互信息量，也不会超过第 II 级信道输入与输出消息之间的平均互信息量。
- 当对信号/数据/消息进行多级处理时，每处理一次，就有可能损失一部分信息，也就是说数据处理会把信号/数据/消息变成更有用的形式，但是绝不会创造出新的信息。这就是所谓的信息不增原理。

练习-1

设离散无记忆信源，其 $\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} x_1=0 & x_2=1 & x_3=2 & x_4=3 \\ 3/8 & 1/4 & 1/4 & 1/8 \end{bmatrix}$
发出的消息为
**(202120130213001203210110321010021032011223
210)**

- 1.此消息的自信息量是多少？
- 2.在此消息中平均每个符号携带的信息量是多少？

练习-2

- 例 二进制通信系统用符号“0”和“1”，由于存在失真，传输时会产生误码，用符号表示下列事件： u_0 ：一个“0”发出； u_1 ：一个“1”发出； v_0 ：一个“0”收到； v_1 ：一个“1”收到。

给定下列概率： $p(u_0)=1/2$ ， $p(v_0/u_0)=3/4$ ， $p(v_0/u_1)=1/2$ ，求：

- (1) 已知发出一个“0”，收到符号后得到的信息量；
- (2) 已知发出的符号，收到符号后得到的信息量；
- (3) 知道发出的和收到的符号能得到的信息量；
- (4) 已知收到的符号，被告知发出的符号得到的信息量。

思考题

- 1、设有12枚同值硬币，其中有一枚假币，且只知道假币的重量与真币的重量不同，但不知究竟是轻还是重。现采用天平比较左右两边轻重的方法来测量（不提供砝码），为了在天平上称出哪一枚是假币，试问至少必须称多少次？请运用信息论的知识说明你的理由