第一章 信息论的基本概念

Basic Concepts of Information Theory

2007年9月19日
2007年9月26日
2007年10月10日

清华大学电子系—张林

概率空间概念回顾

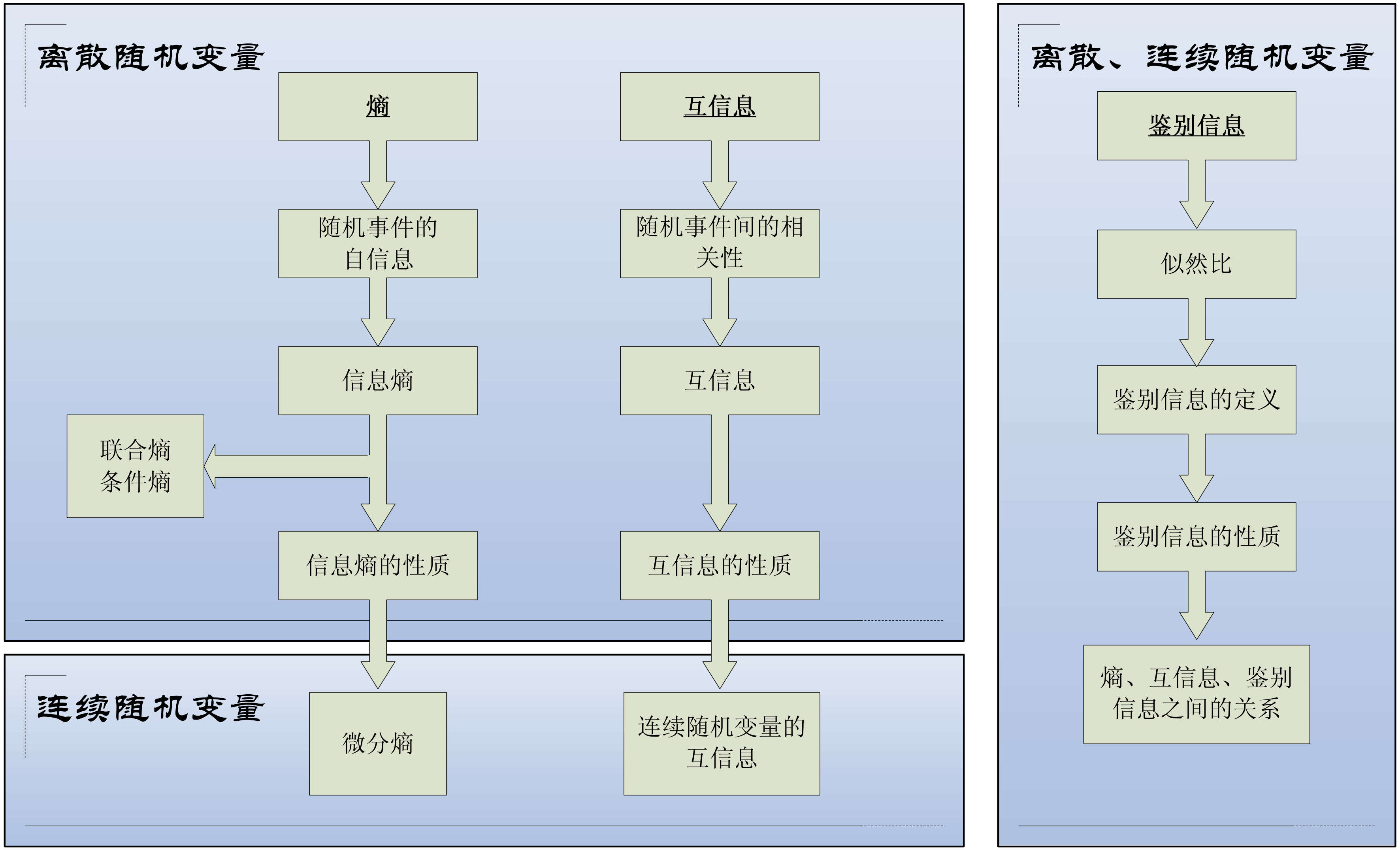
- 概率空间是一个三元组 $(\Omega, \mathcal{X}, P)$
  - Ω 为样本空间
  - $\mathcal{X}$为事件集，$E \in \mathcal{X}$，$E \subset \Omega$
  - P为概率度量，$P: \mathcal{X} \to [0,1]$

- 公理：

1. $\Phi, \Omega \in \mathcal{X}$
2. 若$E \in \mathcal{X}$，则$\mathrm{E}^{\mathrm{C}} \in \mathcal{X}$
3. 若$E_1, E_2, E_3 \ldots \in \mathcal{X}$，则$\bigcup_{i=1}^{\infty} E_i \in \mathcal{X}$

1. $P(\Omega) = 1$
2. $P(E^C) = 1 - P(E)$
3. 若$E_1, E_2, E_3 \ldots$彼此没有交集，则$\mathrm{P}\left(\bigcup_i E_i\right) = \sum_i P(E_i)$

本章知识脉络图


离散随机变量
熵
随机事件的自信息
信息熵
联合熵
条件熵
信息熵的性质
互信息
随机事件间的相关性
互信息
互信息的性质
连续随机变量
微分熵
连续随机变量的互信息
离散、连续随机变量
鉴别信息
似然比
鉴别信息的定义
鉴别信息的性质
熵、互信息、鉴别信息之间的关系

1.1 信息熵（Entropy）

1.1.1 随机事件的自信息
1.1.2 信息熵
1.1.3 信息熵的唯一性定理
1.1.4 联合熵与条件熵
1.1.5 信息熵的性质

1.1.1 随机事件的自信息

- 直觉的定义
  - 信息量等于传输该信息所用的代价
  - 两个相同的信源所产生的信息量两倍于单个信源的信息量
- 但是，直觉的定义立即会引起置疑：
  - 一卡车Beatles的单曲CD盘，承载的信息量很大吗？
  - “很高兴见到你”，“平安到达”，“生日快乐”，“妈妈，母亲节快乐！”等电文传达的信息与其长度等效吗？

信息是对不确定性的消除

- 天气预报消息量
  - 夏天预报下雪和冬天预报下雪，哪个消息含有更大信息量?
- 骗子股票分析员
- 特工00111如何为他提供的服务定价?
  - 用户来找00111是为了消除对某种不确定性
  - 所消除的不确定性越多，收费越高

随机事件的自信息

- 四个基本问题：
  - 随机性与概率的关系；
  - 概率为1的事件的信息量；
  - 概率为0的事件的信息量；
  - 两个独立事件的联合信息量。

↓

设a_1, a_2为两个随机事件，

（1）　若$P(a_1) > P(a_2)$，则 $f(a_1) < f(a_2)$

（2）　若$P(a_1) = 1$，则 $f(a_1) = 0$

（3）　若$P(a_1) = 0$，则 $f(a_1) = \infty$

（4）　若a_1, a_2为独立事件，则 $f(a_1, a_2) = f(a_1) + f(a_2)$

自信息

$$I(a_i) = \log \frac{1}{P(a_i)}$$

对数底与信息的单位

以2为底： bit (binary unit)

以e为底： nat (nature unit)

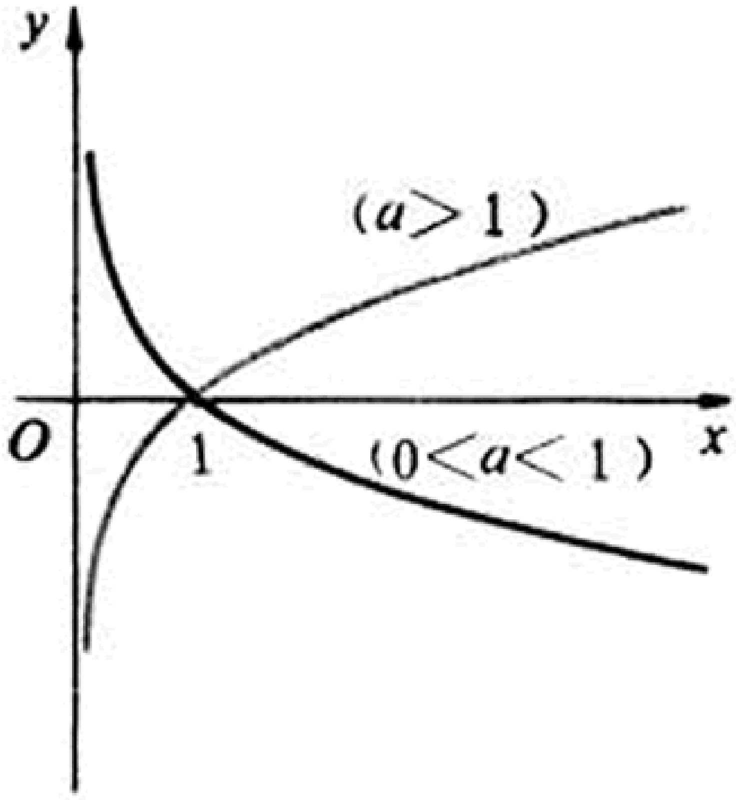
以10为底：Hart (Hartley)

换算关系：

1 nat=1.44 bit

1 Hart=3.32 bit

一般不加说明时，取以2为底。



$$y = \log_a x$$

关于自信息的评注

✓ 自信息大于等于零 $I(a_i)\geq 0$

$$\because 0\leq p(a_i)\leq 1,$$

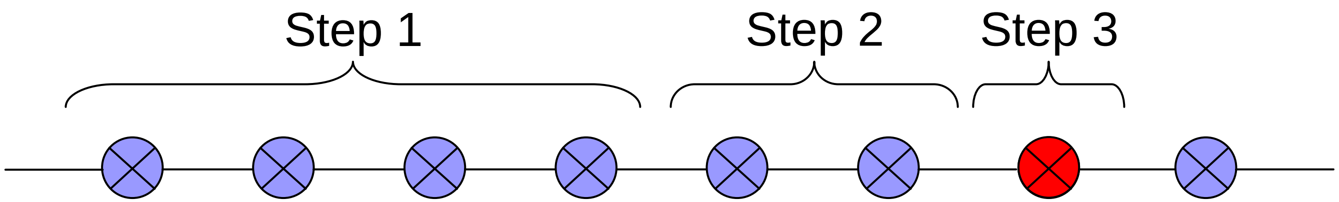
$$\therefore \log\left(1/p(a_i)\right)\geq 0,\ \text{证毕。}$$

✓不同底（单位）之间的自信息之间的换算关系

$$I_\alpha(a_i)=(\log_\alpha \beta)I_\beta(a_i)$$

证明： $\log_\beta p(a_i)=\log_\beta \alpha \log_\alpha p(a_i)$, 证毕。

例1.1 “比特”的意义



- 八个灯泡串联，其中一个灯丝断了。
- 如何用最少的步骤定位出哪一个坏了？
- 最少需要用三次二元判定来定位故障。因此，这个事件所含有的信息量是3比特。

例1.2 洗牌的信息

一副52张的扑克牌，现将其<u>充分洗牌</u>，试问：

（1）任意特定排列所给出的平均信息量是多少？

（2）若任意从这副牌中抽出13张，所示的点数都不同，应获得多少信息量？

解：

（1）获得某一个特定的排列的概率是多少？

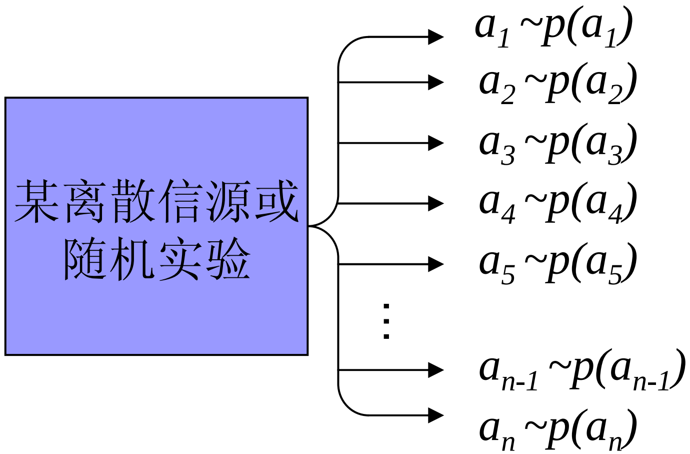
$$I(X)=\log\frac{1}{P\{\text{任意特定排列}\}}=\log\frac{1}{\frac{1}{52!}}=\log 52!=225.58\text{bit}$$

（2）获得“顺子”的概率是多少？

$$I(Y)=\log\frac{1}{P\{\text{得到一副“顺子”}\}}=\log\frac{1}{\frac{C_{52}^{1}C_{48}^{1}C_{44}^{1}\cdots C_{4}^{1}}{P_{52}^{13}}}=13.21\text{bit}$$

1.1.2 信息熵

- 上一节我们定义了对于随机事件的自信息
- 对于一个随机系统，我们如何定义信息的度量？



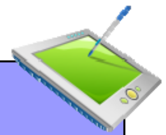
1. 每一个随机事件都有自信息 $I(a_i)$
2. 针对系统，取各随机事件自信息的统计平均：

$$E_p I(a_i) = \sum_i p(a_i) I(a_i)$$
$$= -\sum_i p(a_i) \log p(a_i)$$

定义1.1 离散随机变量的信息熵

离散随机变量 X 的信息熵 $H(X)$ 定义为：

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



- ✓ H(.)的综量是随机变量的分布，而非取值
- ✓ 0log0=0（x→0时，xlogx → 0），概率为0的事件不影响信息熵

例1.3

- 设随机变量X如下N元概率空间：

$$\begin{pmatrix} X \\ p(x) \end{pmatrix} = \begin{pmatrix} x_1, & x_2, \ldots, & x_N \\ p_1, & p_2, \ldots, & p_N \end{pmatrix}, \sum_{n=1}^{N} p_n = 1$$

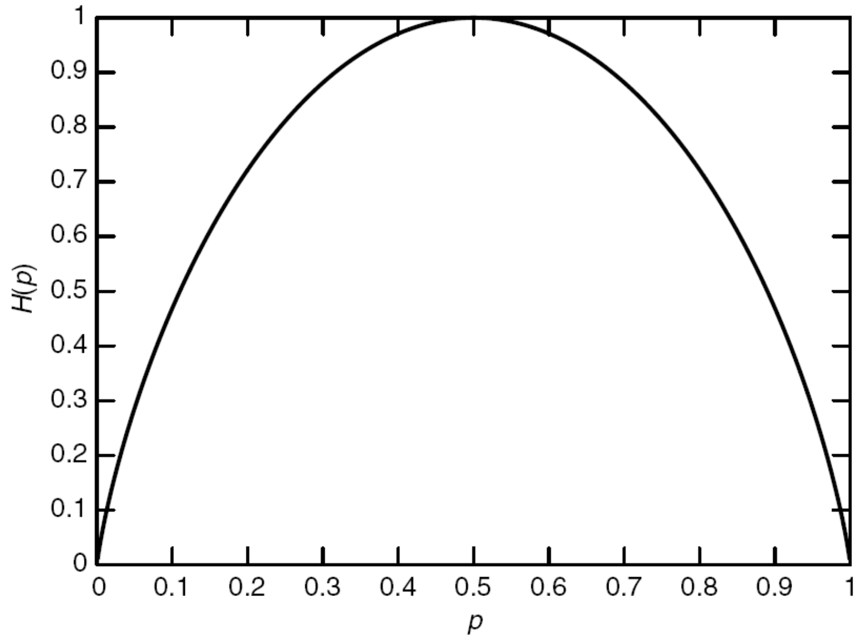
- 平均不确定性，即信息熵为：

$$H(X) = -\sum_{n=1}^{N} p_n \log p_n$$

例1.4

设

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



则有：

$$H(X) = -p\log p - (1-p)\log(1-p) \overset{\text{def}}{=\!=} H(p).$$

- ✓ H(.) 是随机变量的分布上凸（Concave）函数
- ✓ p=0→H(X)=0，确定性系统信息量为零
- ✓ p=0.5→H(X)最大熵

例1.5

设

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

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

- ✓ 平均意义下，确定X的取值所需要的问题数最少是1.75个
- ✓ 如果X是一个随机信源，表达之所需要的最小比特数存在于H(X)和H(X)+1之间
- ✓ 可见，熵与信息的有效表达密切相关

1.1.3 信息熵的唯一性定理

- 香农给出了信息熵函数满足的三个条件
  1. 连续性
  2. 等概时的单调增函数特性
  3. 可加性
- **定理1.1：** 满足上述三个条件的随机变量不确定性度量函数为：

$$f(p_1, p_2, \ldots, p_N) = -C\sum_{n=1}^{N} p_n \log p_n$$

参见板书证明

证明思路流程

1. 等概情况下熵函数的形式
2. 由等概走向有理数非等概的情况
3. 由有理数走向无理数

A. I. Khinchin给出的条件

1. 连续性
2. 可加性
3. 极值条件

$$\max f(p_1, p_2, ..., p_N) = f(\frac{1}{N}, \frac{1}{N}, \cdots, \frac{1}{N})$$

4. 零概率事件不影响不确定性

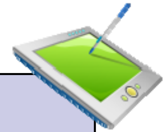
$$f(p_1, p_2, ..., p_N) = f(p_1, p_2, \cdots, p_N, 0)$$

Khinchin条件与香农条件等价

1.1.4 联合熵与条件熵

- **定义1.2**：一对离散随机变量(X,Y)的联合熵定义为：

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



✓ 多元扩展

$$H(X_1,X_2,\cdots X_n) = -\sum_{x\in\mathcal{X}_1}\sum_{x\in\mathcal{X}_2}\cdots\sum_{x\in\mathcal{X}_n} p(x_1,x_2,\cdots x_n)\log p(x_1,x_2,\cdots x_n)$$

✓ 问题：$H(X,Y)=H(X)+H(Y)$？

例1.6

袋子里装3个黑球，2个白球。进行两个随机试验 X 和 Y。

情况一：X — 从中随机取出一个球，看颜色，放回；

Y — 再从中随机取出一球，看颜色。

情况二：X—从中随机取出一个球，看颜色，不放回；

Y—再从中随机取出一球，看颜色。

研究联合试验（XY）的不确定性。

参见板书

- ✓ $H(X,Y) \leqslant H(X)+H(Y)$，等号在X、Y独立时成立。
- ✓ 联合信息小于等于独立观察信息量之和
- ✓ 缺少的信息哪里去了?

条件熵定义

- **定义1.3**：若$(X,Y)\sim p(x,y)$，则<u>条件熵</u>$H(Y|X)$ 定义为：

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

自己证明：

✓ 若 X,Y 统计独立，则 $H(X|Y)=H(X)$，$H(Y|X)=H(Y)$

定理 1.2： H(X,Y)=H(X)+H(Y|X)

证明参见板书

自己证明：

- ✓ $H(X,Y|Z)=H(X|Z)+H(Y|X,Z)$
- ✓ $H(X,Y)=H(Y)+H(X|Y)$
- ✓ 若X,Y统计独立，$H(X,Y)=H(X)+H(Y)$

推论1.3： 设随机变量$X_1,X_2,\ldots,X_n$满足分布$p(x_1,x_2,\ldots,x_n)$，则

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

证明参见板书

例1.7

设(X,Y)服从以下联合分布：

$Y \backslash X$	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 $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$
  $H(X)$=1.75比特
- Y的边缘分布为 $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$
  $H(Y)$=2比特

- $$\begin{aligned} H(X|Y) &= \sum_{i=1}^{4} p(Y=i) H(X|Y=i) \\ &= \frac{1}{4} H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right) \\ &\quad + \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) + \frac{1}{4} H(1, 0, 0, 0) \\ &= \frac{1}{4} \times \frac{7}{4} + \frac{1}{4} \times \frac{7}{4} + \frac{1}{4} \times 2 + \frac{1}{4} \times 0 \\ &= \frac{11}{8} \text{ bits.} \end{aligned}$$
- 类似地，$H(Y|X) = \frac{13}{8}$比特
- $H(X,Y) = \frac{27}{8}$ 比特

1.1.5 信息熵的性质

- **性质1.1 对称性** $H(p_1, p_2, \ldots, p_N) = H(p_{k(1)}, p_{k(2)}, \ldots, p_{k(N)})$
- **性质1.2 非负性** $H(\boldsymbol{p}) \geqslant 0$
- **性质1.3 可加性** $H(X,Y)=H(X)+H(Y|X)$
  - 定理1.2与推论1.3
- **性质1.4** 条件减少熵 $H(X|Y) \leqslant H(X)$
  - 知道了统计相关性的变量，则可以减少不确定性
  - 统计平均意义上条件减少不确定性，但是针对具体的Y取值则不一定

思考题： 证明 $H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$

最大离散熵定理

- **性质1.5** 离散随机变量X在等概率分布的时，熵取得最大值。

$$H(p_1, p_2, \ldots, p_N) \le H\left(\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N}\right) = \log N = \log\left|\mathcal{X}\right|$$

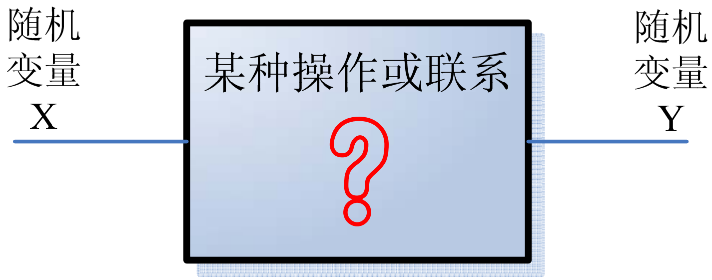
证明参见板书

1.2 互信息与鉴别信息

- 1.2.1 互信息
- 1.2.2 互信息的基本性质
- 1.2.3 似然比与鉴别信息
- 1.2.4 熵、互信息与鉴别信息的关系

1.2.1 互信息

- 事物是普遍联系的，随机变量之间也存在相关关系



- 如何从信息的角度刻画 X与Y之间的相关程度？
  - 单独观察X，得到的信息量是$H(X)$
  - 已知Y之后，X 的信息量变为$H(X|Y)$
  - 了解了Y之后，X 的信息量减少了$H(X)-H(X|Y)$
  - 这个减少量是<u>得知</u>Y 取值之后<u>提供</u>的关于X 的信息

离散互信息的定义

- **定义1.4:** 定义离散随机变量X与Y之间的互信息$I(X;Y)$为

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

✓ 也可以直接定义互信息为

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

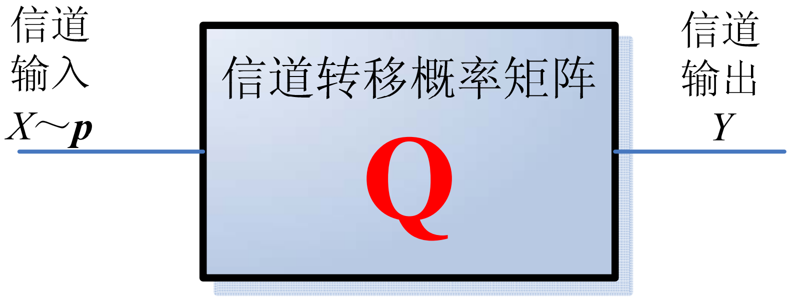
✓ $I(X;Y) = H(X) + H(Y) - H(X,Y)$

✓ 若X, Y独立，则$I(X;Y)=0$

✓ 若X, Y一一映射，则$I(X;Y) = H(X)$

从信道的角度看互信息定义

- 假定X是信道的输入，Y是信道的输出



- $I(X;Y)$表示了一个信道输入与输出之间的依存关系

<u>信道的传输能力</u>

- 用$\boldsymbol{p}$和$\mathbf{Q}=[q(y_j|x_k)]_{J\times K}$表示的互信息

$$I(X;Y)=I(\mathbf{p};\mathbf{Q})=\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}p(x)q(y|x)\log\frac{q(y|x)}{\sum_{x\in\mathcal{X}}p(x)q(y|x)}$$

多变量的互信息

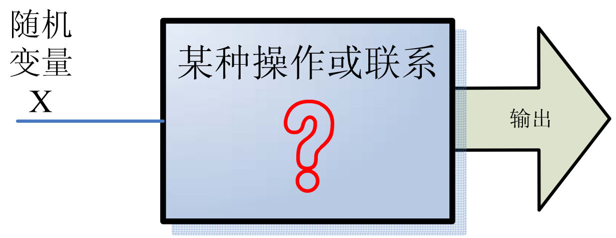
- **定义1.5**：设有随机变量X,Y,Z，则定义

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

✓ 也可以直接定义互信息为

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

$$=\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}}p(x,y,z)\log\frac{p(x,y,z)}{p(x)p(y,z)}$$



多变量的条件互信息

- **定义1.6:** 设有随机变量X,Y,Z，则定义

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

✓ 也可以直接定义条件互信息为

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

✓ 条件互信息非负。

✓ $I(X;Y|Z)=H(X|Z)-H(X,Y|Z)+H(Y|Z)$

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

$$=H(X,Z)+H(Y,Z)-H(X,Y,Z)-H(Z)$$

$I(X;Y;Z)$

- **定义1.7**：三个随机变量互相之间的互信息定义为

$$I(X;Y;Z)=\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}}p(x,y,z)\log\frac{p(x,y)p(y,z)p(x,z)}{p(x)p(y)p(z)p(x,y,z)}$$

- ✓ $I(X;Y;Z)$没有明确的物理意义，是为了数学上的对称性而定义的一个中间量，可正可负。
- ✓ 对于一些推导非常有用。
- ✓ $I(X;Y;Z)=I(X;Y)-I(X;Y|Z)$ （参见“朱书”.pp.34）
- ✓ $I(X;Y;Z)=I(Y;Z)-I(Y;Z|X)$
- ✓ $I(X;Y;Z)=I(X;Z)-I(X;Z|Y)$

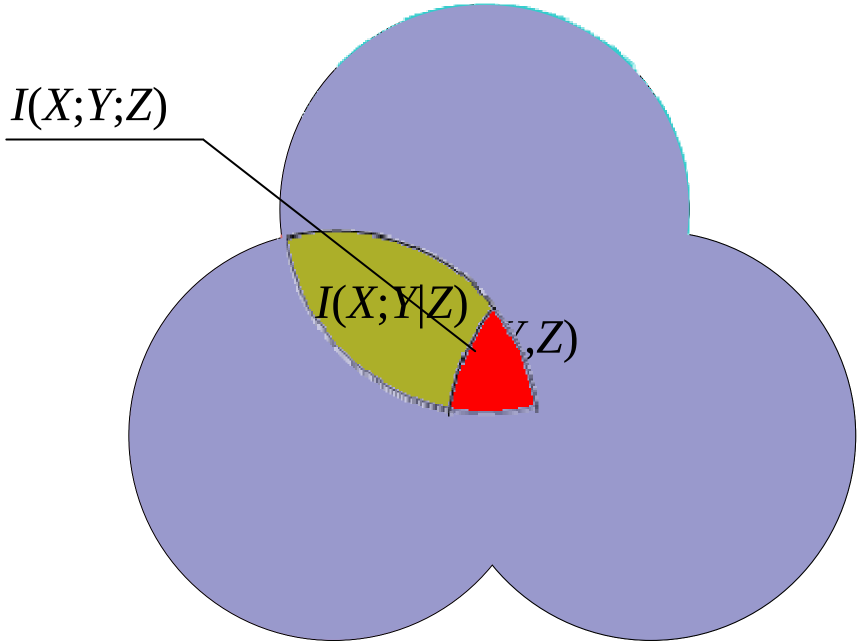
例 1.8

求证当随机变量X与Z统计独立时，有$I(X;Y) \leq I(X;Y|Z)$。

证明：

$$I(X;Y;Z)=I(X;Y)-I(X;Y|Z)=I(X;Z)-I(X;Z|Y)=-I(X;Z|Y)$$

由于$I(X;Z|Y) \geq 0$，所以$I(X;Y)-I(X;Y|Z) \leq 0$，**证毕**。



1.2.2 互信息的基本性质

证明参见板书

- **性质1.6** 对称性 $I(X;Y)=I(Y;X)$
  - “互信息”中的“互（Mutual）”字蕴涵着对称性
- **性质1.7** 非负性 $I(X;Y)\geqslant 0$
  - 了解一个随机变量对于了解另外一个随机变量总有一些帮助
- **性质1.8** 极值性 $I(X;Y)\leqslant \mathbf{min}(H(X),H(Y))$
  - 两个随机变量的互信息不可能比自身还大
- **性质1.9** 可加性 $I(X_1,X_2,\cdots,X_n;Y)=\sum_{i=1}^{n} I(X_i;Y\mid X_1,\cdots,X_{i-1})$
  - 互信息可以分步获得

例1.9

设有同一规格的硬币25枚，其中24枚是标准的，重量相同；而另一枚是假的，重量较标准轻，但其外观上与标准的一样，难于分辨真伪。试求在不用砝码的天平上至少称多少次，才能发现其中的假硬币。

解： 在实验前，由于25枚硬币中每一枚都可能以等概率为假币，因此，确定检测出假币的整个实验X_0的信息量为

$$H(X_0)=\log 25$$

设用天平称两枚硬币的试验X：每一次试验有三种可能的结果：

$$\begin{pmatrix} X \\ p(x) \end{pmatrix} = \begin{pmatrix} 左偏 & 右偏 & 平衡 \\ p(左偏) & p(右偏) & p(平衡) \end{pmatrix}$$

$$H(X) \le \log 3$$

例1.9 （续）

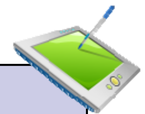
设进行了k次称量，联合实验$X_1,X_2,\ldots,X_n$给出的识别假币的信息量为

$$I(X_1\cdots X_k;X_0)\le H(X_1\cdots X_k)\le\sum_{i=1}^{k}H(X_i)=kH(X)\le k\log 3$$

设k次称量确定了X_0，即 $I(X_1,X_2,\ldots,X_k;X_0)=H(X_0)$

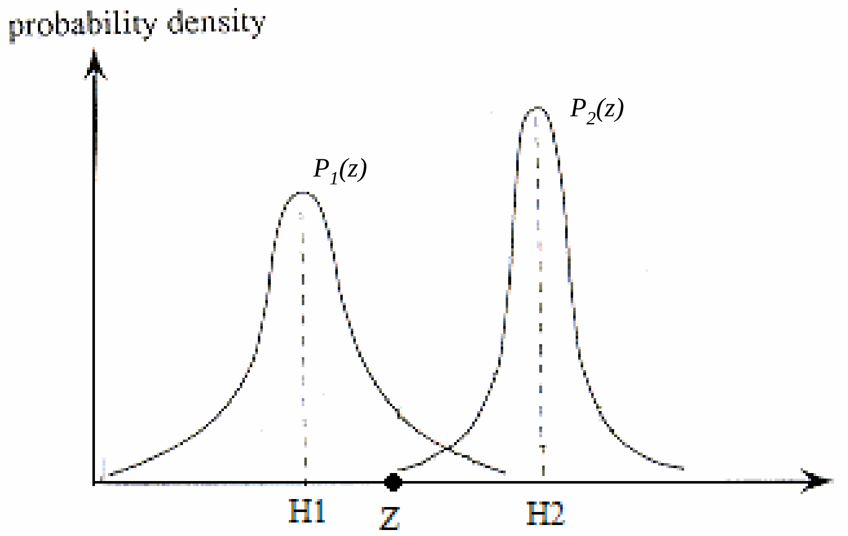
$$\log 25\le k\log 3$$

$$k\ge\frac{\log 25}{\log 3}=\log_3 25,\qquad 3^k\ge 25, k\ge 3$$



- ✓ 通过基于互信息与熵的推导，我们得到了最优方法所需要次数的极限
- ✓ 如何达到这个极限？信息论没有给出答案。
- ✓ 这是信息论解决问题的典型范式。

1.2.3 似然比与鉴别信息



- 经典的二元检测问题回顾
- 最大后验检测（MAP）

$$\frac{P(H_2|z)}{P(H_1|z)} < 1, H_1\text{为真。}$$

$$\frac{P(H_2|z)}{P(H_1|z)} > 1, H_2\text{为真。}$$

- 但是，后验概率往往很难得到，因此利用贝叶斯公式

$$P(H_i|z) = \frac{P(z|H_i)P(H_i)}{P(z)}$$

可以将MAP变成似然比检测（LRT）

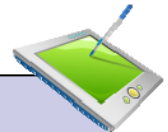
$$\frac{p(z|H_2)}{p(z|H_1)} \underset{H_2}{\overset{H_1}{\lessgtr}} \frac{p(H_1)}{p(H_2)}$$

称 $\Lambda(z) = \dfrac{P(z|H_2)}{P(z|H_1)} = \dfrac{P_2(z)}{P_1(z)}$ 为似然比

对数似然比与鉴别信息

- 取似然比的对数，称为对数似然比：$\log \Lambda(z) = \log \dfrac{P_2(z)}{P_1(z)}$
- **定义1.8**：两个随机分布$p(x)$和$q(x)$之间的鉴别信息定义为

$$D(\mathbf{p} \| \mathbf{q}) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$



✓ 物理意义：
  - 观察者对随机变量X的了解由分布$q(x)$变为$p(x)$时获得的信息量
  - 当实际分布为$p(x)$而估计为$q(x)$时，$D(p\|q)$衡量了这种估计的偏差程度

✓ 也称为Kullback-Leibler距离、交叉熵（Cross-Entropy）

✓ 不满足对称性和三角不等式，因此不是严格意义上的“距离”

✓ **p=q**，鉴别信息为零；鉴别信息的非负性（朱书.pp.51）。

✓ 思考：鉴别信息具有分步可加性吗？

例 1.10：鉴别信息不对称

设$\mathcal{X}$＝{0,1}，考查两个分布**p**和**q**。设 $p(0)=1-r$，$p(1)=r$，而 $q(0)=1-s$，$q(1)=s$。

求鉴别信息：

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

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

若$r=s$,$D(p||q) = D(q||p) = 0$ ；若$r=1/2$，$s=1/4$，

$$D(p||q) = \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{3}{4}} + \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{4}} = 1 - \frac{1}{2}\log 3 = 0.2075 \text{ bit}$$

$$D(q||p) = \frac{3}{4}\log\frac{\frac{3}{4}}{\frac{1}{2}} + \frac{1}{4}\log\frac{\frac{1}{4}}{\frac{1}{2}} = \frac{3}{4}\log 3 - 1 = 0.1887 \text{ bit}$$

1.2.3 熵、互信息、鉴别信息的关系

- **定理1.4** 离散随机变量的熵与鉴别信息的关系满足

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

其中，**u**为均匀分布。

✓ $\log|\mathcal{X}|$是最大熵（性质1.5），在分布为u的时候取得

✓ $D(\mathbf{p}||\mathbf{u})$是均匀分布与实际分布之间的差异的度量

- **定理1.5** 离散随机变量的互信息与鉴别信息的关系满足

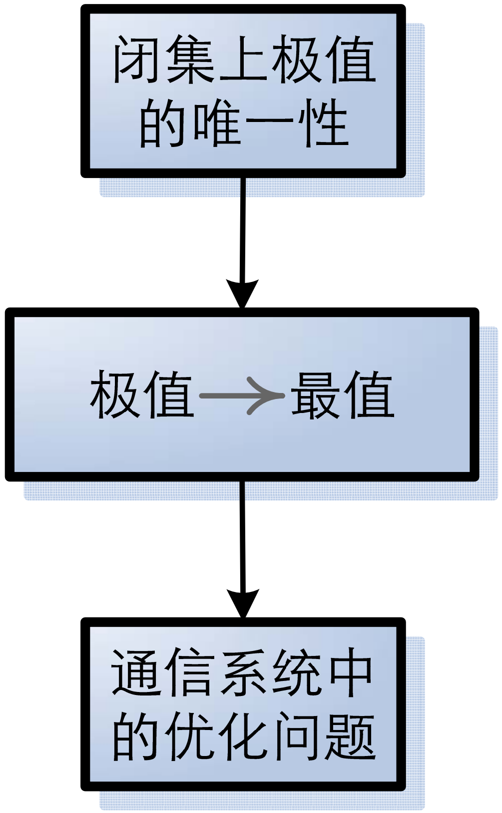
$$I(X;Y)=D(p(x,y)\|p(x)p(y))$$

✓当随机变量X,Y分布$p(x,y)$时，我们假定二者独立，这种假定距离实际情况差异有多大？由上述定义给出。

证明见板书

1.3 熵、互信息、鉴别信息的凸性

- 为什么要研究凸性？

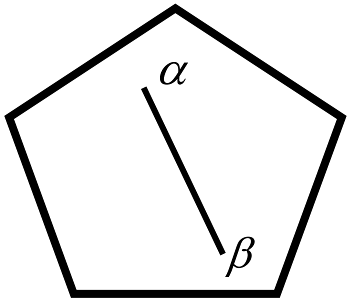


凸集概念回顾

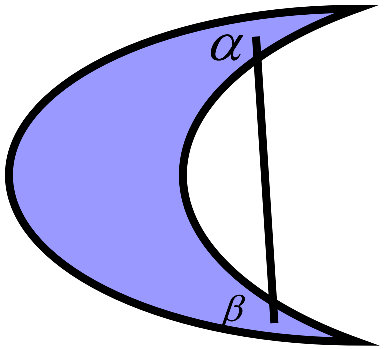
- 若对集合D中任意两点$\alpha \in D$ 和 $\beta \in D$ ，均有：

$$\lambda \alpha + (1-\lambda)\beta \in D, \forall 0 \le \lambda \le 1, \lambda \in \mathbf{R}$$

则称集合 D 是凸集。



凸集



非凸集

- ✓ 实数域是凸集
- ✓ 整数域非凸集
- ✓ 有理数域非凸集
- ✓ <u>概率矢量集合是凸集</u>

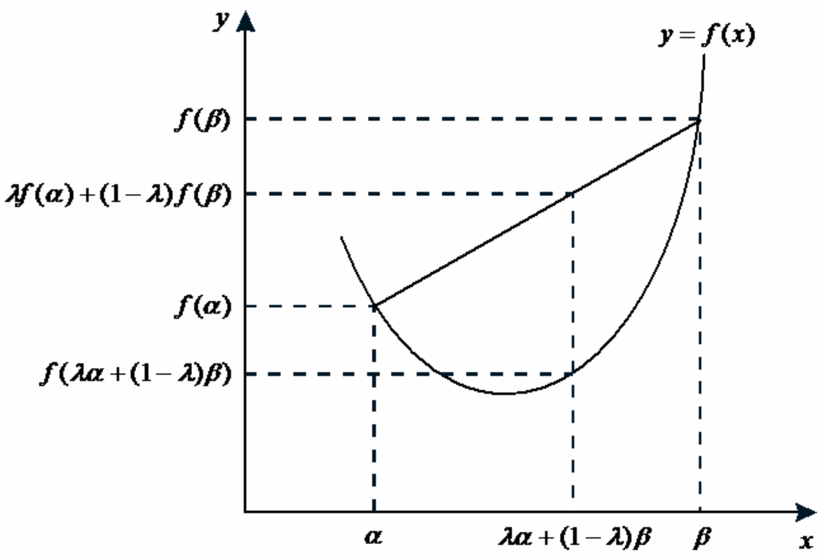
凸函数概念回顾

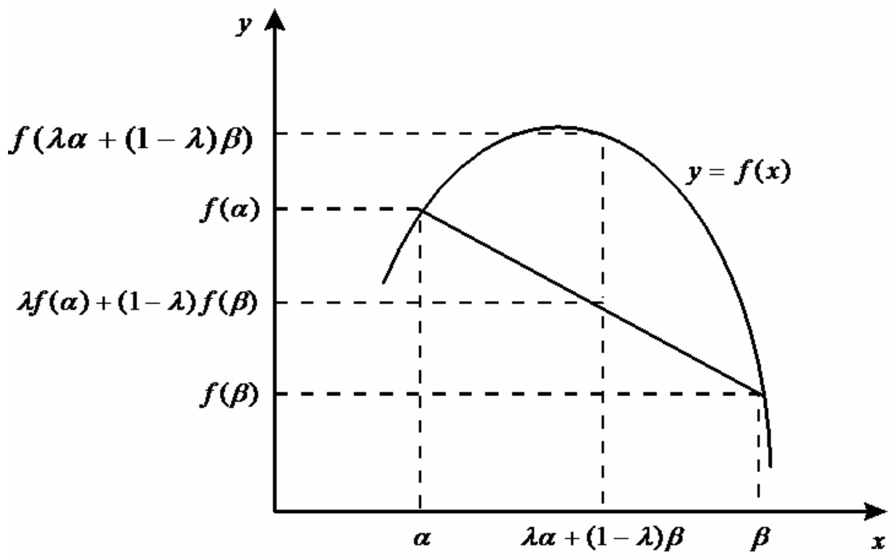
- 定义在凸集D上的函数 $f(x)$ 称为下凸函数(Convex)，如果

$$f(\lambda\alpha+(1-\lambda)\beta)\le\lambda f(\alpha)+(1-\lambda)f(\beta)$$

- 定义在凸集 D 上的函数 $f(x)$ 称为上凸函数(Concave)，如果

$$f(\lambda\alpha+(1-\lambda)\beta)\ge\lambda f(\alpha)+(1-\lambda)f(\beta)$$





凸函数的性质

- 若f(x)下凸，那么-f(x)上凸，C-f(x)上凸
- 上凸＋上凸＝上凸
- 下凸＋下凸＝下凸
- f(x)与f(ax)凸性一致
- f(**x**)与f(A**x**)凸性一致
- 线性函数既上凸也下凸

Jeson不等式与对数求和不等式

证明参见板书

- **引理1.6** 如果f是下凸函数，且X是离散随机变量，则

$$\mathbf{E}f(X) \geqslant f(\mathbf{E}X)$$

并且，若f是严格下凸函数，上式中等号说明X为常数，即X与$\mathbf{E}X$以概率1相等。

- **引理1.7** 对于非负实数$a_1,a_2,\ldots,a_n$和$b_1,b_2,\ldots,b_n$有

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

等号当且仅当a_i/b_i为常数的时候成立。

鉴别信息的凸性

- **定理1.8** $D(\mathbf{p}\|\mathbf{q})$是$(\mathbf{p},\mathbf{q})$的下凸函数，即若存在分布对$(\mathbf{p_1},\mathbf{q_1})$和$(\mathbf{p_2},\mathbf{q_2})$，则

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

对于所有 $0\le\lambda\le1$ 成立

证明参见板书

熵函数的上凸性质

- **定理1.9** 熵函数是随机分布 p 的上凸函数。

证明见板书

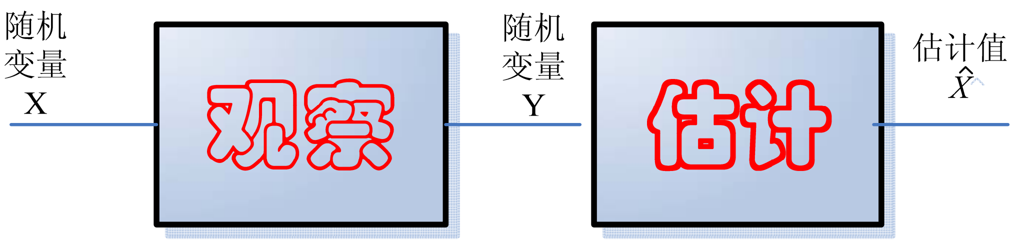
互信息函数的凸性

- **定理1.10** 互信息函数是随机分布 $\boldsymbol{p}$ 的上凸函数，是信道转移概率矩阵**Q**的下凸函数。

证明见板书

1.4 Fano不等式与数据估计

- 在通信与信号处理中，“估计”是我们经常遇到的一类重要问题
- 假定我们知道了随机变量Y，希望藉此推断一个相关的随机变量 X，判断的准确程度如何？
- 直觉上，估计的准确程度与条件熵 $H(X|Y)$ 有关。



定义错误概率：

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

Fano不等式

证明见板书

- **定理1.11** 对于任意估计 $\hat{X}$，满足 $X \to Y \to \hat{X}$ 且 $P_e = \Pr(X \neq \hat{X})$

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

✓ **物理意义**

  ✓ $H(X|Y)$：用 Y 估计 X 产生的信息损失

  ✓ $H(P_e)$：误差的不确定性

  ✓ $\log|\mathcal{X}|$：估计错误时系统剩余的不确定性

✓ $P_e=0$ 标志着 $H(X|Y)=0$，与直观感觉一致。

✓ Fano不等式可以弱化为

$$1 + P_e \log|\mathcal{X}| \geq H(X|Y) \text{或} \quad P_e \geq \frac{H(X|Y)-1}{\log|\mathcal{X}|}$$

Fano不等式的简单数学应用

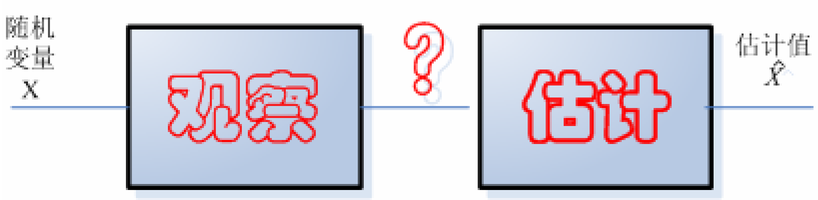
- **推论1.12** 对于任意两个随机变量X和Y，设p＝Pr($X \neq Y$)

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

- **推论1.13** 若估计$\hat{X} = Y$，则Fano不等式强化为

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

Fano不等式中的等号



- 假定没有观察到Y
- 完全凭借X的先验分布进行估计
- 不妨设$X\in\{1,2,\ldots,m\}$，且$p_1 \geqslant p_2 \geqslant \ldots \geqslant p_m$
- 显然，此时的最优估计是 $\hat{X}=1$，而$P_e=1-p_1$
- 可以证明，此时Fano不等式取得等号的条件是

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

$$\because H(X|Y) = H(X) = H(p_1, p_2, \cdots, p_m)$$

$$= H(1-P_e, \frac{P_e}{m-1}, \frac{P_e}{m-1}, \cdots, \frac{P_e}{m-1})$$

$$= H(P_e) + P_e \log(m-1)$$

一般意义下的估计

- **定义1.9:** 称映射$\boldsymbol{T}$: $\mathcal{X}^n \rightarrow \Theta$为基于样本$X^n$对$\theta$的估计，也记做：$T(X_1, X_2, \ldots, X_n)$

- **定义1.10:** 称估计是无偏的，若$E_\theta T(x_1, x_2, \ldots, x_n) - \theta = 0$

- **定义1.11:** 称估计是一致的，若$n \rightarrow \infty$时，有

$$T(X_1, X_2, \ldots, X_n) \rightarrow \theta$$

- **定义1.12:** 称估计$\boldsymbol{T_1}$优于估计$\boldsymbol{T_2}$，若对于所有的θ，有

$$E\left(T_1(X_1, X_2, \ldots, X_n) - \theta\right)^2 \leq E\left(T_2(X_1, X_2, \ldots, X_n) - \theta\right)^2$$

Fisher信息和Cramer-Rao界

- **定义1.13:** 定义Fisher信息为

$$J(\theta) = E_\theta\left[\frac{\partial}{\partial\theta}\ln f(X;\theta)\right]^2$$

- **定理1.14:** (Cramer-Rao不等式) 对于任何无偏估计***T***(X)，其均方误差满足

$$\mathrm{var}(T) \geq \frac{1}{J(\theta)}$$

1.5 连续随机变量的熵、互信息

1.5.1 连续随机变量的微分熵

1.5.2 微分熵的变换特性

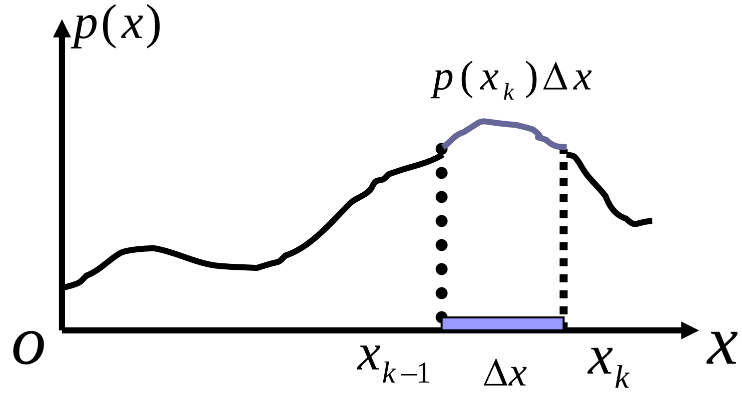
1.5.3 连续随机变量的互信息

1.5.1 连续随机变量的微分熵

- 数学概念中的抽象：
  - 可数→不可数
  - 整数→有理数→无理数
- 物理世界中，连续是更加普遍的概念
  - 电压、电流的测量
  - 语音信号
  - 图像信号
- 问题：如何描述连续随机变量的不确定性?

令Δx→0的简单推广

- 将连续随机变量的取值切分为 Δx 的小区间
- 采用黎曼积分（Riemann Integral）的方法求解



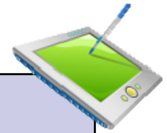
- $H(X) = -\sum_k p(x_k)\Delta x \log p(x_k)\Delta x = -\sum_k [p(x_k)\log p(x_k)]\Delta x - \sum_k [p(x_k)\log \Delta x]\Delta x$
- $\lim_{\substack{k\to\infty \\ \Delta x\to 0}} H(X) = -\int p(x)\log p(x)dx - \int p(x)\lim_{\Delta x\to 0}\log \Delta x dx = h(X) - \infty$

$$\boxed{h(X) = -\int_{-\infty}^{+\infty} p(x)\log p(x)dx}$$

微分熵的定义与评述

- **定义1.14** 定义连续随机变量的微分熵为

$$h(X)=-\int_{-\infty}^{+\infty}p(x)\log p(x)dx$$



✓ $\Delta x\to 0$ 时，$H(X)\to\infty$表明：<u>连续熵不存在</u>。连续随机变量所包含的信息量为无限大，我们不可能全部获取，我们关心的只是其中足以满足我们所需要的一部分。

✓从物理上层面上看，以 $-\log\ \Delta x$ 作为参考点，$h(X)$是相对值。实际通信中关心的是熵差，所以重点研究它也符合信息理论研究的实际需求。

微分熵定义的多变量扩展

- 联合熵：$h(X,Y) = -\iint p(x,y)\log p(x,y)dxdy$
- 条件熵：$h(X\,|\,Y) = -\iint p(x,y)\log p(x\,|\,y)dxdy = -\int p(y)\int p(x\,|\,y)\log p(x\,|\,y)dxdy$
- 不等式关系

$$h(X,Y) = h(X) + h(Y\,|\,X) = h(Y) + h(X\,|\,Y)$$

$$h(X\,|\,Y) \le h(X),\quad h(Y\,|\,X) \le h(Y)$$

$$h(X,Y) \le h(X) + h(Y)$$

- 但是注意：$h(X)$不一定为正

微分熵为负值的例子

- **例1.11**设随机变量

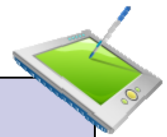
$$X: p(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b, \quad b-a<1 \\ 0 & x>b, \qquad x<a \end{cases}$$

则$$h(X) = -\int_a^b \frac{1}{b-a} \log \frac{1}{b-a} dx = \log(b-a) < 0$$

1.5.2 微分熵的性质

- **定理1. 15（微分熵的变换性质）**

$$h(aX)=h(X)+\log|a|$$



思考：

✓对于高维随机矢量，定理1.15变为什么形式？

✓在什么变换下，微分熵才能获得不变性？

例1.12 高斯分布的微分熵

- 高斯分布：$X: p(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp[-\frac{(x-m)^2}{2\sigma^2}]$
- 高斯分布微分熵的计算

$$h(X)=-\int_{-\infty}^{+\infty}p(x)\log p(x)dx$$

$$=-\int_{-\infty}^{+\infty}p(x)\log\{\frac{1}{\sqrt{2\pi\sigma^2}}\exp[-\frac{(x-m)^2}{2\sigma^2}]\}dx$$

$$=-\int_{-\infty}^{+\infty}p(x)\log\frac{1}{\sqrt{2\pi\sigma^2}}dx+\log e\int_{-\infty}^{+\infty}p(x)\frac{(x-m)^2}{2\sigma^2}dx$$

$$=\log\sqrt{2\pi\sigma^2}+\frac{\log e}{2\sigma^2}\int_{-\infty}^{+\infty}p(x)(x-m)^2dx$$

$$=\log\sqrt{2\pi\sigma^2}+\frac{1}{2}\log e=\frac{1}{2}\log 2\pi e\sigma^2$$

$$\boxed{h(X)=\frac{1}{2}\log 2\pi e\sigma^2}$$

定理1.16

若给定连续随机变量X的均值m与方差σ²，则当其服从高斯分布时，微分熵最大。

证明：

设$p(x)$为满足均值m，方差σ²的高斯分布PDF，q(x)为任意满足均值m，方差σ²的PDF。有：

$$-\int_{-\infty}^{+\infty} q(x)\log p(x)dx = -\int_{-\infty}^{+\infty} q(x)\log\{\frac{1}{\sqrt{2\pi\sigma^2}}\exp[-\frac{(x-m)^2}{2\sigma^2}]\}dx$$

$$= -\int_{-\infty}^{+\infty} q(x)\log\frac{1}{\sqrt{2\pi\sigma^2}}dx + \log e\int_{-\infty}^{+\infty} q(x)\frac{(x-m)^2}{2\sigma^2}dx = \log\sqrt{2\pi\sigma^2} + \frac{1}{2}\log e = \frac{1}{2}\log 2\pi e\sigma^2$$

$$= -\int_{-\infty}^{+\infty} p(x)\log p(x)dx$$

于是：$h_{q(x)}(X) - h_{p(x)}(X) = -\int_{-\infty}^{+\infty} q(x)\log q(x)dx + \int_{-\infty}^{+\infty} p(x)\log p(x)dx$

$$= -\int_{-\infty}^{+\infty} q(x)\log q(x)dx + \int_{-\infty}^{+\infty} q(x)\log p(x)dx = \int_{-\infty}^{+\infty} q(x)\log\frac{p(x)}{q(x)}dx$$

$$\le \int_{-\infty}^{+\infty} q(x)(\frac{p(x)}{q(x)}-1)dx = \int_{-\infty}^{+\infty} p(x)dx - \int_{-\infty}^{+\infty} q(x)dx = 0$$

1.5.3 连续随机变量的互信息

- 采用类似于微分熵的推广方法求连续随机变量的互信息

$$I(X;Y)=\lim_{\substack{\Delta x\to 0\\ \Delta y\to 0}}\sum_{i=-\infty}^{+\infty}\sum_{j=-\infty}^{+\infty}p(x_i,y_j)\Delta x_i\Delta y_j\log\frac{p(x_i,y_j)\Delta x_i\Delta y_j}{p(x_i')\Delta x_i\, p(y_j')\Delta y_j}$$

$$=\iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy$$

- **定义1.15** 定义连续随机变量X，Y之间的互信息为

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

对于微分熵与连续变量互信息的历史说明

- 最早的微分熵由Shannon在1948年的论文中给出
- 严格的对于微分熵和连续变量互信息的定义由Kolmogorov和Pinsker给出

•A. N. Kolmogorov. On the Shannon theory of information transmission in the case of continuous signals. IRE Trans. Inf. Theory, IT-2:102–108, Sept. 1956.

•M. S. Pinsker. Information and Stability of Random Variables and Processes. Izd. Akad. Nauk, 1960. Translated by A. Feinstein, 1964.

第一章知识要点

- 熵、互信息、鉴别信息的定义与性质
- 三者之间的关系
- Jensen不等式与对数和不等式
- 熵、互信息、鉴别信息的凸性
- Fano不等式的证明与意义
- 熵与互信息在连续随机变量条件下的推广及其与离散条件下性质的区别