《信息论基础》

信息安全专业
2015-2016学年第一学期

中国科学技术大学 刘斌
Email: flowice@ustc.edu.cn
办公室：科技实验楼西楼1711室

助教：刘志强
课程主页：
http://www.ubicom.ustc.edu.cn/teach/InfoTheory/index.html

课程安排

- 授课时间：40学时
  - 32学时上课，6学时复习和答疑，2学时考试
- 课后作业
  - 每周一交给助教，周四发回
- 评分标准
  - 期末考试：60分
  - 课后作业：32分，共8次作业，每次满分4分
    - 抄作业，该次作业按0分算
    - 迟交作业，该次作业满分按2分记
  - 平时表现：8分

课程简介

- 信息论是在对通信理论的研究中发展起来的
  1. 信源编码：临界数据压缩的值（熵 H）
  2. 信道编码：临界通信传输速率的值（信道容量 C）
- 信息论涉及许多学科：统计物理，计算机科学，概率与统计等
- 信息的理论安全性和实际安全性
- 本课程主要讲授香农（Shannon）信息论的基本理论

课程教材

- 教科书
  - 《信息论基础》（美）Thomas M. Cover，Joy A. Thomas 著，阮吉寿 张华 译，机械工业出版社

- 参考书
  - 《信息论与编码》，姜丹 著，中国科技大学出版社

课程内容

- 第2章：熵、相对熵和互信息
- 第3章：渐进均分性
- 第4章：随机过程的熵率
- 第5章：数据压缩
- 第7章：信道容量
- 第8章：微分熵
- 第9章：高斯信道
- 第10章：率失真理论
- 信息论和信息安全

第一章 绪论与概览

- 什么是信息？
  - 当今社会是信息社会
  - 信息的含义模糊和难于捉摸
- 如何准确的度量信息？
  - 一般来说，可以判断是否获得信息，但无法准确的度量信息
  - 应用数学工具，通过数学的运算来度量信息

Shannon信息论的三个基本论点

- 1948 Shannon 《通信的数学原理》"A Mathematical Theory of Communication"
- Shannon信息论的三个论点
  - ✓ **形式化假说：**通信的任务只是在接收端把发送端发出的消息从形式上复制出来，消息的语义、语用是接收者自己的事，与传送消息的通信系统无关。只保留了数学可描述的内容。
  - ✓ **非决定论：**一切有通信意义的消息的发生都是随机的，消息传递中遇到的噪声干扰也是随机的，通信系统的设计应采用概率论、随机过程、数理统计等数学工具。
  - ✓ **不确定性：**信息就是用来消除不确定性的东西

信息的度量

- 信息的度量和不确定性消除的程度有关
- 不确定性的程度与事件发生的概率有关
- 信息量与概率的关系
  - ✓ 信息量是概率的单调递减函数
  - ✓ 概率小，信息量大
  - ✓ 概率大，信息量小

第二章　熵、相对熵和互信息

- 离散随机变量：X
- 字母表（取值空间）：

$$\mathcal{X} = \{x_1, x_2, \cdots, x_i, \cdots, x_n\}$$

- 概率密度函数：$p_X(x) = p(x) = \Pr(X = x),\ x \in \mathcal{X}$

$$0 \le p(x_i) \le 1,\ \sum_{i=1}^{n} p(x_i) = 1$$

- **注意：** 大写字母X代表随机变量，小写字母x代表随机变量的一个取值（事件，消息，符号）。

自信息量的物理含义

- 自信息量表示事件发生后，事件给与观察者的信息量。
- 自信息量的大小取决于事件发生的概率。事件发生的可能性越大，它所包含的信息量就越小。反之，事件发生的概率越小，它能给与观察者的信息量就越大。

例 布袋中装有手感觉完全一样的球，但颜色和数量不同，问下面三种情况下随意拿出一个球的不确定程度的大小。

（1）99个红球和1个白球（2）50个红球和50个白球

（3）红球、白球、黑球、黄球各25个

自信息量需满足的条件

- 自信息量是事件发生概率的函数

$$I(x) = I(p(x))$$

- 自信息量函数必须满足以下条件：
  - ✓ 若 $p(x_i) > p(x_j)$ ，则 $I(p(x_i)) < I(p(x_j))$
  - ✓ 若 $p(x) = 0$ ，则 $I(p(x)) \to \infty$
  - ✓ 若 $p(x) = 1$ ，则 $I(p(x)) = 0$
  - ✓ 对于两个统计独立事件，

$$I(x_i, x_j) = I(x_i) + I(x_j)$$

自信息量的数学表达式

- 定义 事件x的自信息量为

$$I(x) = -\log p(x)$$

- I(x)实质上是无量纲的
- 为研究问题方便， I(x)的量纲根据对数的底来定义
  - ✓ 对数取2为底，自信息量的单位是比特(bit)；
  - ✓ 取e为底（自然对数），单位为奈特(nat)；
  - ✓ 取10为底（常用对数），单位为哈特(hart)

自信息量单位的转换

- 对数的换底公式

$$\log_a x = \frac{\log_b x}{\log_b a}$$

$$1\text{奈特} = \log_2 e = 1.443\text{比特}，\quad 1\text{哈特} = \log_2 10 = 3.322\text{比特}$$

$$1\text{比特} = 0.693\text{奈特}，\quad 1\text{比特} = 0.301\text{哈特}$$

- 一般情况下，我们在课程中使用2为底的对数，信息量的单位是比特。

自信息量的例子

$$I(x) = -\log p(x)$$

例 英文字母中“e”出现的概率为0.105，“d”出现的概率为0.035，“y”出现的概率为0.012。分别计算它们的自信息量。

解：“e”的自信息量 $I(e) = -\log 0.105 = 3.25$比特

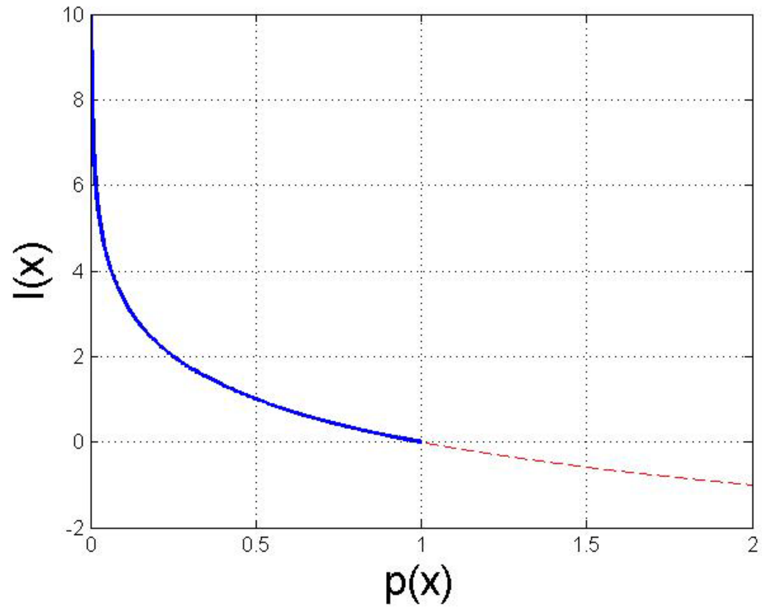
“d”的自信息量 $I(d) = -\log 0.035 = 4.84$比特

“y”的自信息量 $I(y) = -\log 0.012 = 6.38$比特

自信息量的性质

$$I(x) = -\log p(x)$$

- 自信息量是非负的
- 确定事件的信息量为零
- 自信息量是概率的单调递减函数
- I(x)基于随机变量X的特定取值x，不能作为整个随机变量X的信息测度。



熵（Entropy）

- 定义 一个离散随机变量 X 的熵 $H(X)$ 定义为 $$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x)$$
- 熵的量纲根据对数 log 的底来定义
  - ✓ 对数取2为底，对应的熵的单位是比特(bit)；
  - ✓ 取e为底（自然对数），熵的单位为奈特(nat)；
  - ✓ 取10为底（常用对数），熵的单位为哈特(hart)
- 各单位间的换算：$H_b(X) = (\log_b a)H_a(X)$

熵与信息的关系

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

- 消息是信息的载体。信息是抽象的，消息是具体的。
- 一个人获得消息→消除不确定性→获得信息。
- 信息的度量（信息量）和不确定性消除的程度有关，消除的不确定性＝获得的信息量；
- 熵是随机变量平均不确定度的度量，同时它也代表了消除随机变量不确定度所需获得的信息量。

熵和不确定度

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

- 熵是随机变量平均不确定度的度量，是平均信息量的度量

例 两个随机变量X、Y，其取值和概率分布分别为

$$\begin{bmatrix} X \\ p(X) \end{bmatrix} = \begin{Bmatrix} 0 & 1 \\ 0.99 & 0.01 \end{Bmatrix}, \quad H(X) = 0.08\text{比特}$$

$$\begin{bmatrix} Y \\ p(Y) \end{bmatrix} = \begin{Bmatrix} 0 & 1 \\ 0.5 & 0.5 \end{Bmatrix}, \quad H(Y) = 1\text{比特}$$

Y的平均不确定度大于X的平均不确定度，也就是说我们确定Y所需要的平均信息量要大于确定X所需要的平均信息量。

零概率事件对熵的影响

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

- 当某一事件x出现的概率p(x)为零时，我们规定$0\log 0 = 0$，也就是说，增加一些零概率的项不会改变熵的值，同样，也不会影响信息量的大小。

例

$$\begin{bmatrix} X \\ p(X) \end{bmatrix} = \begin{Bmatrix} 0 & 1 \\ 0.99 & 0.01 \end{Bmatrix}, \quad H(X) = 0.08\text{比特}$$

$$\begin{bmatrix} Y \\ p(Y) \end{bmatrix} = \begin{Bmatrix} 0 & 1 & 2 & 3 \\ 0.99 & 0.01 & 0 & 0 \end{Bmatrix}, \quad H(Y) = 0.08\text{比特}$$

熵与期望

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

- 随机变量X的熵可以解释为随机变量 $-\log p(X)$ 的期望值：

$$H(X) = E_p\{-\log p(X)\}$$

- 信息熵H(X)是各离散消息自信息量的数学期望，表示了每个消息提供的平均信息量。

熵的性质

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

- 由于H(X)的表达式和热力学中熵的表达式相似，且在概念上也有相似之处，因此借用“熵”这个词，把H(X)称为信息“熵”；
- **非负性** $H(X)\geq 0$ **；当且仅当X是一确知量时取等号。**
- 熵是在平均意义上来表征随机变量的总体特性的，对于给定概率分布的随机变量，熵是一个确定的值。
- 对于离散随机变量，熵的值是有限的。

熵的性质

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

- 熵不依赖于随机变量的实际取值，而仅依赖于其概率分布，且与概率分布的顺序无关。

例

$$\begin{bmatrix} X \\ p(X) \end{bmatrix} = \begin{Bmatrix} \text{晴} & \text{阴} & \text{雨} & \text{雪} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{Bmatrix}, \quad H(X) = 1.75 \text{ 比特}$$

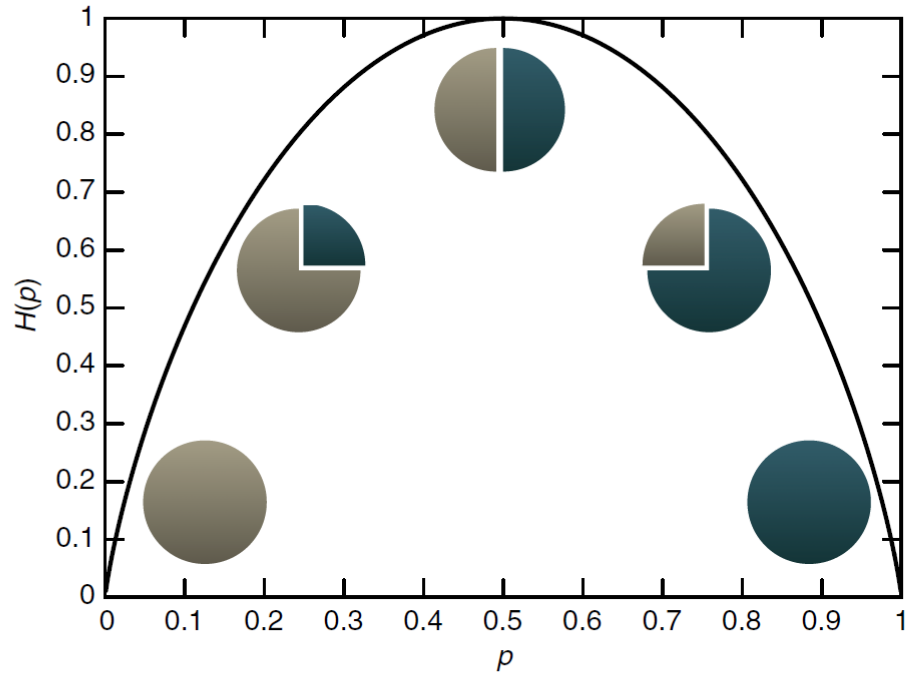
$$\begin{bmatrix} Y \\ p(Y) \end{bmatrix} = \begin{Bmatrix} \spadesuit & \heartsuit & \diamondsuit & \clubsuit \\ \frac{1}{4} & \frac{1}{8} & \frac{1}{2} & \frac{1}{8} \end{Bmatrix}, \quad H(Y) = 1.75 \text{ 比特}$$

Bernoulli分布的熵

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

例

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



$$H(X) = -p\log p - (1-p)\log(1-p) \triangleq H(p)$$

联合熵（Joint Entropy）

- 定义 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)定义为

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

- 联合熵的定义是单个变量的熵定义的简单推广。

条件熵（Conditional Entropy）

- 条件熵是用来度量在已知一个随机变量的情况下，另一个随机变量还存在的不确定度。
- 定义 对于服从联合分布为p(x, y)的一对离散随机变量(X, Y)，其条件熵H(Y|X)定义为

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

链式法则

- 定理 对于服从联合分布为p(x, y)的一对离散随机变量(X, Y)，

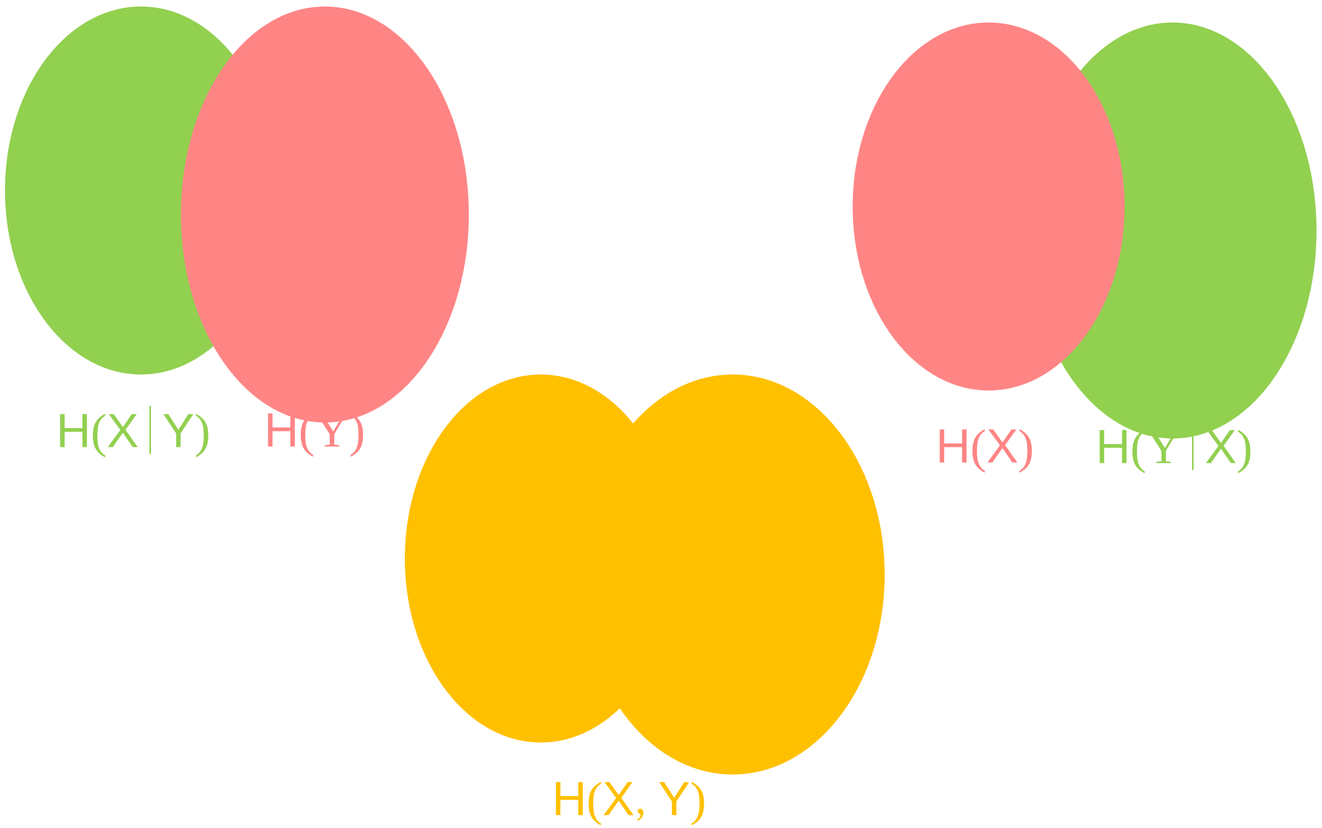
$$H(X,Y)=H(X)+H(Y|X)$$

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

- 推论 对于服从联合分布为p(x, y, z)的三个随机变量(X, Y, Z)，

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

链式法则的文氏图表示

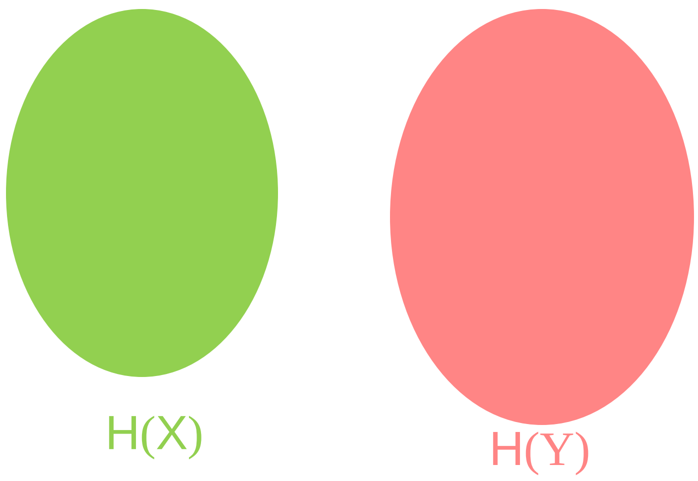


X和Y统计独立时

- 当X和Y相互独立时，

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

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



例子

例

$Y \backslash X$	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

$$\begin{bmatrix} X \\ p(X) \end{bmatrix} = \left\{ \begin{matrix} 1 & 2 & 3 & 4 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{matrix} \right\}$$

$$\begin{bmatrix} Y \\ p(Y) \end{bmatrix} = \left\{ \begin{matrix} 1 & 2 & 3 & 4 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{matrix} \right\}$$

$$H(X) = \frac{7}{4}\text{比特}; \quad H(Y) = 2\text{比特}$$

$$H(X|Y) = \frac{11}{8}\text{比特}; \quad H(Y|X) = \frac{13}{8}\text{比特}$$

$$H(X,Y) = \frac{27}{8}\text{比特}$$

$$H(Y|X) \neq H(X|Y)!!$$

相对熵（Relative Entropy）

- 相对熵是两个随机分布之间距离的度量，也称为Kullback-Leibler距离。
- 定义 两个概率密度函数为p(x)和q(x)之间的相对熵定义为

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

- 约定：$0\log\frac{0}{0}=0;\quad 0\log\frac{0}{q}=0;\quad p\log\frac{p}{0}=\infty$

相对熵的性质

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

- 相对熵总是非负的。而且，当且仅当 $p = q$ **时相对熵为零。**
- 如果存在一个$x\in\mathcal{X}$使得 $p(x) > 0,\ q(x) = 0$，则有 $D(p\|q) = \infty$ 。
- 相对熵并不是真正意义上的距离，它不对称，也不满足三角不等式。我们把相对熵看作一种“距离”是方便对以后很多概念的理解。

相对熵的例子

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

一般来说

$$D(p\|q) \neq D(q\|p)$$

例

$$\begin{bmatrix} X \\ p(X) \end{bmatrix} = \begin{Bmatrix} 0 & 1 \\ 1-r & r \end{Bmatrix}$$

$$\begin{bmatrix} X \\ q(X) \end{bmatrix} = \begin{Bmatrix} 0 & 1 \\ 1-s & s \end{Bmatrix}$$

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

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

$$r = s \Rightarrow D(p\|q) = D(q\|p) = 0$$

$$r = \frac{1}{2}, s = \frac{1}{4} \Rightarrow D(p\|q) = 0.2075\text{比特}, D(q\|p) = 0.1887\text{比特}$$

$$r = \frac{1}{2}, s = \frac{1}{3} \Rightarrow D(p\|q) = 0.085\text{比特}, D(q\|p) = 0.0817\text{比特}$$

互信息（Mutual Information）

- 互信息是一个随机变量包含另一个随机变量信息量的度量。
- 定义 两个随机变量X和Y的联合概率密度函数为p(x, y)，边缘概率密度函数分别为p(x)和p(y)。则互信息I(X;Y)定义为

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

熵与互信息的关系

- 熵表示单个随机变量的平均不确定度。
- 条件熵表示在给定一个随机变量知识的情况下，另一个随机变量还存在的平均不确定度。
- 互信息表示在给定一个随机变量知识的情况下，另一个随机变量平均不确定度的缩减量。
- 互信息是一个表征信息流通的量

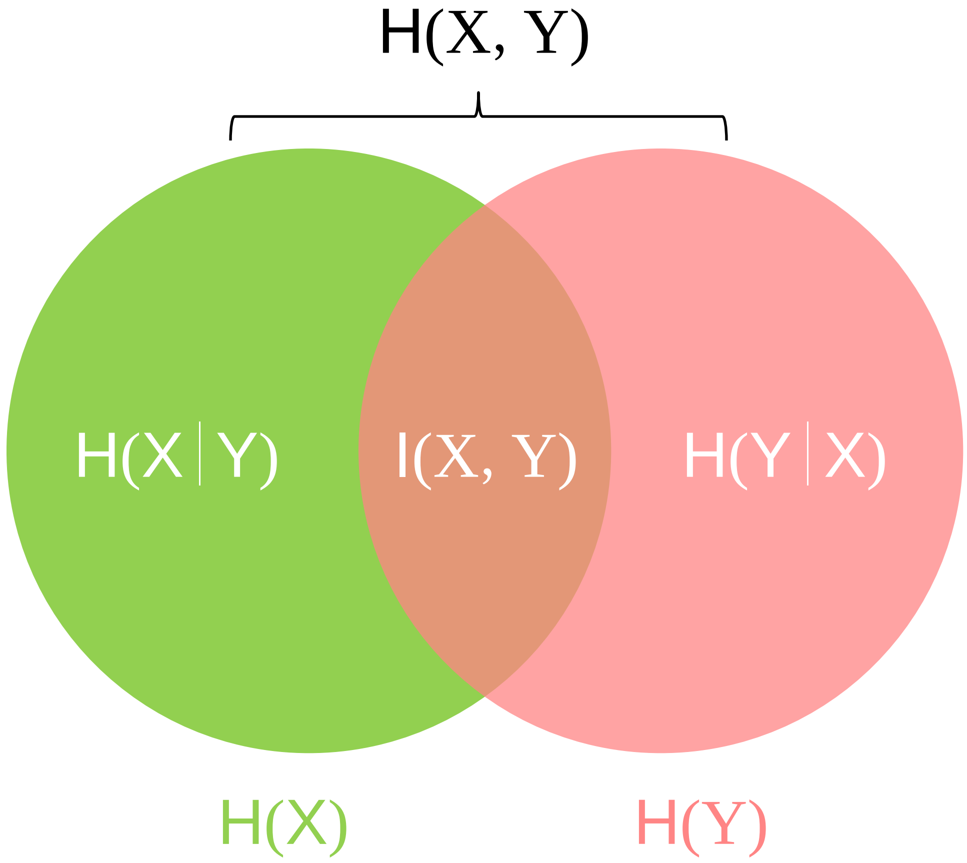
$$\boxed{\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ I(X;Y) &= H(Y) - H(Y|X) \end{aligned}}$$

熵与互信息的关系

定理

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

互信息的文氏图表示



互信息的例子

例

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

$$\begin{bmatrix} X \\ p(X) \end{bmatrix} = \begin{Bmatrix} 1 & 2 & 3 & 4 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{Bmatrix}$$

$$\begin{bmatrix} Y \\ p(Y) \end{bmatrix} = \begin{Bmatrix} 1 & 2 & 3 & 4 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{Bmatrix}$$

$$H(X) = \frac{7}{4}\text{比特}; \quad H(Y) = 2\text{比特}$$

$$H(X|Y) = \frac{11}{8}\text{比特}; \quad H(Y|X) = \frac{13}{8}\text{比特}$$

$$H(X,Y) = \frac{27}{8}\text{比特}$$

$$I(X;Y) = \frac{3}{8}\text{比特}$$

熵的链式法则

- 一对离散随机变量(X, Y)：

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

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

- 定理 多个随机变量的熵的链式法则：

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

互信息的链式法则

- 定义 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：

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

- 定理 互信息的链式法则：

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

相对熵的链式法则

- 定义 对于联合概率密度函数$p(x,y)$和$q(x,y)$，条件相对熵定义为：

$$D(p(y|x)\|q(y|x)) = \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)}$$

- 定理 相对熵的链式法则：

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

凸函数

- **定义** 若对于任意的$x_1, x_2 \in (a,b)$, $0 \le \lambda \le 1$，满足

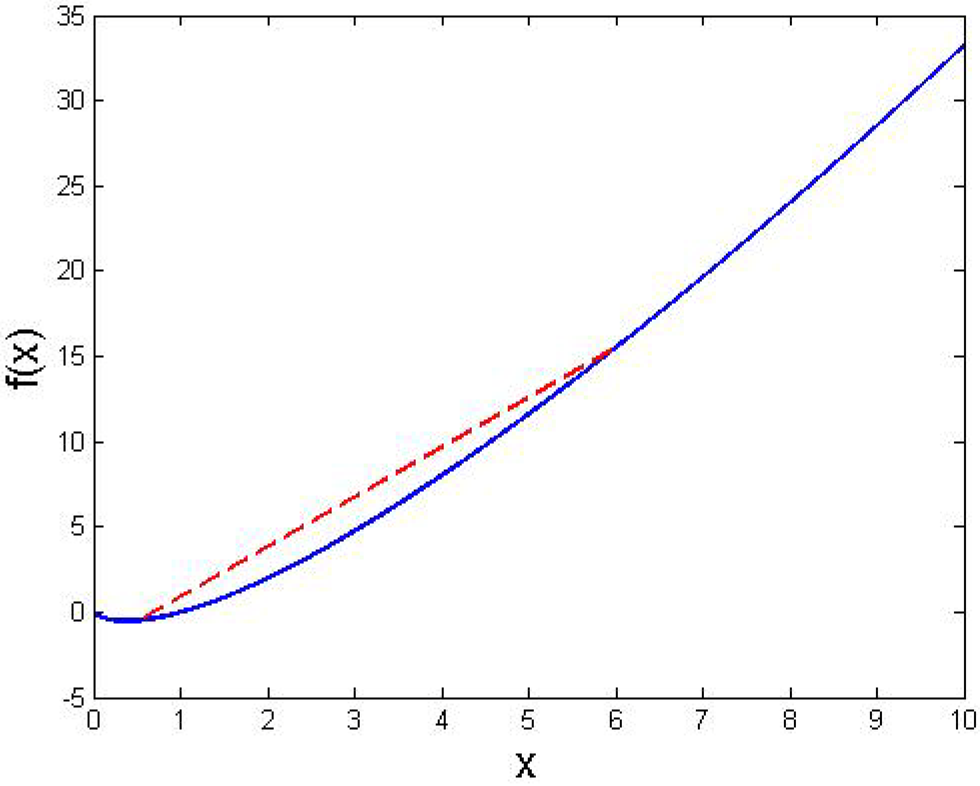
$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2)$$

则称函数f(x)在区间(a, b)上是**凸的**(convex)。如果仅当 $\lambda = 0,\ 1$，等号成立，则称函数f(x)是严格凸的(strictly convex)。

凸函数的例子

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2)$$

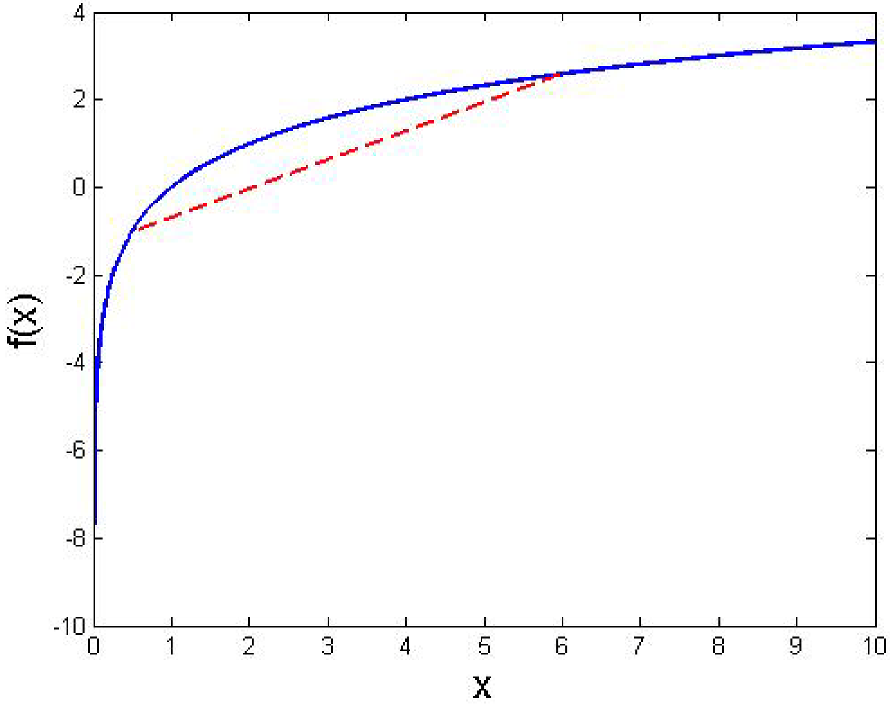
$$f(x) = x\log(x)$$



凹函数 $f(\lambda x_1 + (1-\lambda)x_2) \geq \lambda f(x_1) + (1-\lambda)f(x_2)$

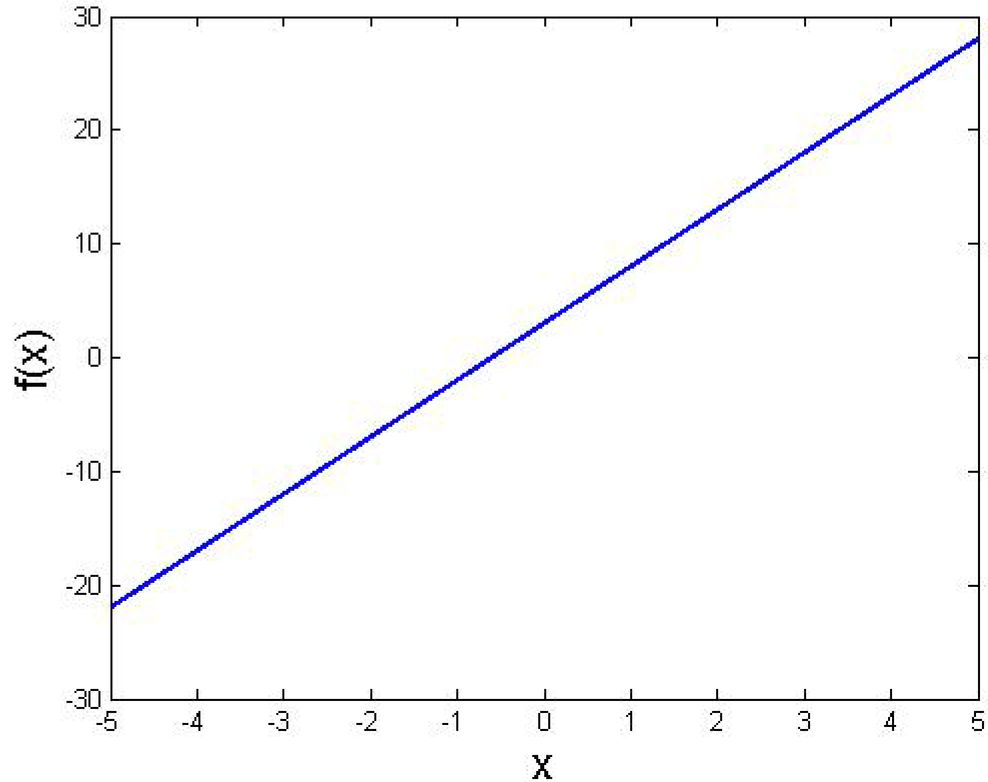
- 定义 如果-f(x)为凸函数，则称f(x)是**凹的（concave）**

$$f(x) = \log(x)$$



一个特殊的函数

$$f(x) = ax + b$$



Jensen不等式

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2)$$

- 定理 如果函数f在某个区间上的二阶导数总是非负（正）的，则f为该区间的凸函数（严格凸函数）。
- 定理 Jensen不等式：若给定凸函数f和一个随机变量X，则

$$Ef(X) \ge f(EX)$$

若f是严格凸的，则当且仅当X为常量的时候，等式成立。

相对熵的非负性

- **定理** 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

- **推论**

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

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

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

熵的性质

- **极值性**：$H(X) \leq \log|\mathcal{X}|$ ，其中$|\mathcal{X}|$表示X的字母表中元素的个数，当且仅当X服从均匀分布时，等号成立。
- 条件作用使熵减少：

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

X和Y相互独立时取等号。

信息降低不确定度：仅在平均意义上成立！

例子

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

$$
\begin{aligned}
H(X) &= H\left(\frac{1}{8}, \frac{7}{8}\right) = 0.544\text{比特} \\
H(X|Y=1) &= 0\text{比特} \\
H(X|Y=2) &= 1\text{比特} \\
H(X|Y) &= \frac{3}{4}H(X|Y=1) + \frac{1}{4}H(X|Y=2) \\
&= 0.25\text{比特}
\end{aligned}
$$

熵的性质

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

- **极值性**：$H(X) \le \log|\mathcal{X}|$ ，其中$|\mathcal{X}|$表示X的字母表中元素的个数，当且仅当X服从均匀分布时，等号成立。
- 条件作用使熵减少：

$$H(X|Y) \le H(X)$$

- 熵的独立界：

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

所有随机变量相互独立时取等号。

熵和互信息的凹凸性

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

- **对数和不等式**：对于非负数 $a_i, b_i, i = 1, \cdots, n$ ，

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

  当且仅当 $a_i/b_i =$ 常数，等号成立。

- 相对熵的凸性：$D(p\|q)$ 关于(p,q)是凸的
- 熵的凹性：$H(p)$是关于p的凹函数
- 互信息的凹凸性:固定$p(y|x)$，$I(X;Y)$是$p(x)$的凹函数；固定$p(x)$，$I(X;Y)$是$p(y|x)$的凸函数。

马尔科夫链

- 定义 若随机变量X, Y, Z的联合概率密度函数满足

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

则X, Y, Z构成马尔科夫链 $\boxed{X \rightarrow Y \rightarrow Z}$

- ✓ 给定Y时，X，Z条件独立
- ✓ $X \rightarrow Y \rightarrow Z$ 蕴含 $Z \rightarrow Y \rightarrow X$
- ✓ 若 $Z=f(Y)$，则$X \rightarrow Y \rightarrow Z$

数据处理不等式

- 数据处理不等式：若 $X \to Y \to Z$ ，则有

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

  - ✓ 若 $Z = g(Y)$ ， 则 $I(X;Y) \geq I(X;g(Y))$
  - ✓ 如果 $X \to Y \to Z$， 则 $I(X;Y|Z) \leq I(X;Y)$

如果X，Y，Z不构成马尔科夫链，有可能

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

数据处理不等式

- 如果 $X \to Y \to Z$，则 $I(X;Y|Z) \le I(X;Y)$

> 如果X，Y，Z不构成马尔科夫链，有可能
> $I(X;Y|Z) \ge I(X;Y)$

$$X, Y\text{统计独立，则} I(X;Y) = 0;$$

$$\text{设} Z = X + Y,$$

$$\text{则} I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = H(X|Z) \ge 0$$

费诺（Fano）不等式

- 通过Y来估计X：$\hat{X} = g(Y)$
- **费诺不等式**：对任何满足 $X \to Y \to \hat{X}$ 的估计量 $\hat{X}$ ，设 $P_e = \Pr\{X \neq \hat{X}\}$，有

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

上述不等式可以减弱为

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

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

费诺不等式的推论 $H(P_e)+P_e\log|\mathcal{X}|\ge H(X|\hat{X})\ge H(X|Y)$

- 对于任意两个随机变量X和Y，设 $p=\Pr(X\neq Y)$

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

- 设 $P_e=\Pr\{X\neq\hat{X}\},\hat{X}:\mathcal{Y}\to\mathcal{X}$ ，

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

- 如果X和X’独立同分布，具有熵H(X)，则

$$\Pr(X=X')\ge 2^{-H(X)}$$