信息论基础

李　莹
liying2009@ecust.edu.cn

第二章：信息的度量

一、自信息和互信息

二、平均自信息

1. 平均自信息的概念
2. 熵函数的性质
3. 联合熵与条件熵

三、平均互信息

1. 平均自信息的概念

自信息是一个随机变量：

自信息是指信源发出的某一消息所含有的信息量。不同的消息，它们所含有的信息量也就不同。

平均自信息（信息熵/信源熵/香农熵/无条件熵/熵函数/熵）：

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

信息熵的单位，取决于对数选取的底：

- 以2为底，单位为比特/符号
- 以e为底，单位为奈特/符号
- 以10为底，单位为哈特莱/符号。

信息熵的意义：

信源的信息熵是从整个信源的统计特性来考虑的。它是从平均意义上来表征信源的总体特性的。对于某特定的信源，其信息熵只有一个。不同的信源因统计特性不同，其信息熵也不同。

信息熵是从平均意义上表征随机变量总体特性的一个量，其含义体现在如下几方面：

(1) 在事件发生后，表示平均每个事件（或符号）所提供的信息量；

(2) 在事件发生前，表示随机变量取值的平均不确定性；

(3) 表示随机变量随机性大小，熵大的，随机性大；

(4) 当事件发生后，其不确定性就被解除，熵是解除随机变量不确定性平均所需信息量。

离散随机变量的概率空间为：

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \dots & x_n \\ p(x_1) & p(x_2) & \dots & p(x_n) \end{bmatrix} \quad 0 \le p(x_i) \le 1, \quad \sum_{i=1}^{n} p(x_i) = 1$$

通常把一个随机变量的样本空间和样本空间中的元素对应的概率称为概率空间。

记 $p_i = p(x_i)$，则

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

$$= H(p_1, p_2, \dots, p_n) = H(\mathbf{p})$$

由于概率的完备性，即 $\sum_{i=1}^{n} p_i = 1$，所以 $H(\mathbf{p})$ 实际上是 $(n-1)$ 元函数。

当 $n=2$ **时，** $H(\mathbf{p}) = H(p, 1-p) = H(p)$

例1　掷一个六面均匀的骰子，每次出现朝上一面的点数是随机的，以朝上一面的点数作为随机试验的结果，并把试验结果看作一个信源的输出，试建立数学模型。

解：　信源的输出：离散随机变量X

X：{1，2，3，4，5，6}——样本空间

$P(X)$：$\{P(X=1)=1/6,\ P(X=2)=1/6,\ \ldots,\ P(X=6)=1/6\}$

$$[X\bullet P]=\begin{cases} X: & 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \\ P(X): & 1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6 \end{cases}$$

——概率空间

$$0\le p(x_i)\le 1\ (i=1,2\cdots,6) \qquad \sum_{i=1}^{6} p(x_i)=1$$

$$H(X)=\log 6 \quad \text{bit/symbol}$$

例2：一信源有6种输出符号，概率分别为$P(A)$=0.5，$P(B)$=0.25，$P(C)$=0.125，$P(D)=P(E)$=0.05，$P(F)$=0.025。

1)计算$H(X)$。

2)求符号序列*ABABBA*和*FDDFDF*的信息量，并将之与6位符号的信息量期望值相比较。

解： 1)由信息熵定义，该信源输出的信息熵为

$$H(X)=\sum_{i=1}^{6}p(x_i)\log\frac{1}{p(x_i)}$$

$$=0.5\log 2+0.25\log 4+0.125\log 8+2\times 0.05\log 20+0.025\log 40$$

$$=1.94 \quad \text{bit/symbol}$$

符号序列$ABABBA$所含的信息量为

$$I_1 = 3I_A + 3I_B = 3\left(-\log P_A - \log P_B\right) = 3\left(\log 2 + \log 4\right) = 9 \ \text{bit}$$

符号序列$FDDFDF$所含的信息量为

$$I_2 = 3I_D + 3I_F = 3\left(-\log P_D - \log P_F\right) = 3\left(\log 20 + \log 40\right) = 28.932 \ \text{bit}$$

6位符号序列的信息量平均值为

$$\bar{I} = 6H(X) = 11.64 \ \text{bit}$$

三者比较为 $I_1 < \bar{I} < I_2$

2. 熵函数的性质

熵函数的数学特性包括:

(1)对称性	**(5)**连续性
(2)确定性	**(6)**递增性
(3)非负性	**(7)**上凸性
(4)扩展性	**(8)**极值性

2. 熵函数的性质-对称性

（1）对称性

当概率矢量 $\mathbf{p}=(p_1,p_2,\ldots,p_n)$ 中各分量的次序任意变更时，熵函数的值不变，即

$$
\begin{aligned}
H(p_1,\ p_2,\ \ldots,\ p_n) &= H(p_2,\ p_1,\ \ldots,\ p_n) \\
&= H(p_3,\ p_1,\ \ldots,\ p_2) \\
&= \ldots \\
&= -\sum_{i=1}^{n} p_i \log p_i
\end{aligned}
$$

该性质说明：熵只与随机变量(信源)的总体统计特性有关。如果某些信源的统计特性相同（含有的符号数和概率分布相同），那么这些信源的熵就相同。

2. 熵函数的性质-对称性

例3： 三个信源分别为：

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} x_1\,(\text{红}) & x_2\,(\text{黄}) & x_3\,(\text{蓝}) \\ 1/3 & 1/6 & 1/2 \end{bmatrix}$$

$$\begin{bmatrix} Y \\ P(Y) \end{bmatrix} = \begin{bmatrix} y_1\,(\text{红}) & y_2\,(\text{黄}) & y_3\,(\text{蓝}) \\ 1/6 & 1/2 & 1/3 \end{bmatrix}$$

$$\begin{bmatrix} Z \\ P(Z) \end{bmatrix} = \begin{bmatrix} z_1\,(\text{晴}) & z_2\,(\text{雾}) & z_3\,(\text{雨}) \\ 1/3 & 1/6 & 1/2 \end{bmatrix}$$

① X与Z信源的差别：
具体消息其含义不同；

② X与Y信源的差别：
同一消息的概率不同；

③ 但它们的信息熵是相同的。

2. 熵函数的性质-确定性

（2）确定性

$$\boldsymbol{H(1,0)=H(1,0,0)=H(1,0,0,0)=\ldots=H(1,0,\ldots,0)=0}$$

$$(\because \quad 1\log 1=0,\quad \lim_{p_i\to 0} p_i\log p_i=0)$$

在概率空间中，只要有一个事件是必然事件，那么其它事件一定是不可能事件，因此信源没有不确定性，熵必为**0**。

2. 熵函数的性质-非负性

（3）非负性

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

$$\because 0 \leq p_i \leq 1$$

$$\therefore -\log p_i \geq 0$$

$$\therefore H(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log p_i \geq 0$$

只有当随机变量是一确知量时，熵***H*(*X*)=0**。

离散信源的熵满足非负性，而连续信源的熵可能为负。

2. 熵函数的性质-扩展性

（4）扩展性

$$\lim_{\varepsilon\to 0} H_{n+1}\left(p_1, p_2, \ldots, p_n-\varepsilon, \varepsilon\right) = H_n\left(p_1, p_2, \ldots, p_n\right)$$

$$\left(\because \quad \lim_{\varepsilon\to 0} \varepsilon \log \varepsilon = 0, \quad \lim_{\varepsilon\to 0}\left(p_n-\varepsilon\right)\log\left(p_n-\varepsilon\right) = p_n \log p_n\right)$$

扩展性说明，增加一个概率接近于零的事件，信源熵保持不变。

虽然小概率事件出现后，给予收信者较多的信息，但从总体来考虑时，因为这种概率很小的事件几乎不会出现，所以它对于离散集的熵的贡献可以忽略不计。这也是熵的总体平均性的一种体现。

2. 熵函数的性质-连续性

（5）连续性

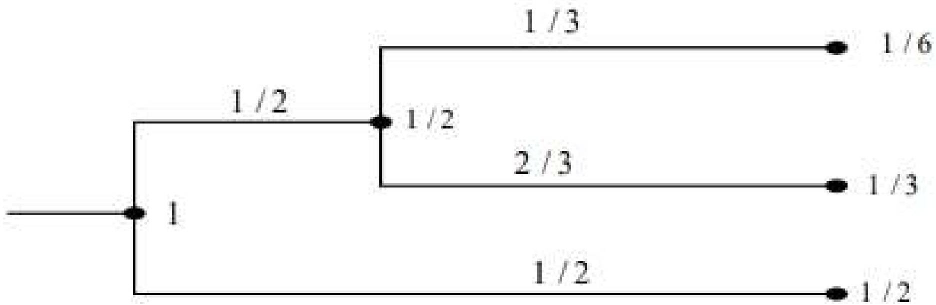
$$\lim_{\varepsilon \to 0} H(p_1,\ p_2,\cdots,p_{q-1}-\varepsilon,\ p_q+\varepsilon)=H(p_1,\ p_2,\cdots,p_q)$$

概率分量的微小波动不会引起熵的变化。

（6）可加性（递推性、递增性)

先看一个简单例子，某随机事件集合有3个事件，概率分别为：$p_1=1/2$，$p_2=1/3$，$p_3=1/6$；

这3个事件可以直接产生，也可分两步产生，即先以1/2的概率产生两个事件，选择其中之一作为输出，或者在另一事件发生条件下再以2/3和1/3的概率产生两事件，选择其中的一个作为输出。



复合事件产生例

（6）递增性（递推性)

熵的可加性首先由香农提出，含义如下：如果一个事件可以分成两步连续选择来实现（多步产生的事件也称复合事件），那么原来的熵H应为H的单独值的加权和。

注："H的单独值"是指每次选择的熵值，"权值"就是每次选择的概率。

$$H(1/2,1/3,1/6)=H(1/2,1/2)+(1/2)H(1/3,2/3)$$

上式等号右边的第1项是第1步选择的熵；由于第2步选择只有1/2的概率发生，所以第2项是第2步选择的熵与权值1/2的乘积。

（6）递推性（可加性)

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

$$=H(p_1,p_2,\cdots,p_n)+p_nH(\frac{q_1}{p_n},\frac{q_2}{p_n}\cdots,\frac{q_m}{p_n})$$

其中：$\sum_{i=1}^{n}p_i=1,\sum_{j=1}^{m}q_j=p_n$

2. 熵函数的性质-递推性

例4： 利用递推性计算熵函数 H(1/3, 1/3, 1/6, 1/6)的值。

解：$$H(\frac{1}{3},\frac{1}{3},\frac{1}{6},\frac{1}{6}) = H(\frac{1}{3},\frac{2}{3})+\frac{2}{3}H(\frac{1}{2},\frac{1}{4},\frac{1}{4})$$

$$=H(\frac{1}{3},\frac{2}{3})+\frac{2}{3}H(\frac{1}{2},\frac{1}{2})+\frac{2}{3}\cdot\frac{1}{2}H(\frac{1}{2},\frac{1}{2})$$

$$=H(\frac{1}{3},\frac{2}{3})+H(\frac{1}{2},\frac{1}{2})$$

$$=1.915$$ 比特/符号

2. 熵函数的性质-递推性

例　　设某地气象为随机变量X，符号集A={晴，多云，阴，雨，雪，雾，霾}，概率分别为0.3, 0.2, 0.2, 0.05, 0.05, 0.05, 0.15；现将多云和阴用阴代替，雨和雪用降水代替，雾和霾用雾霾代替，得到简化气象Y，符号集B ={晴，阴，降水，雾霾}，求两气象熵的差。

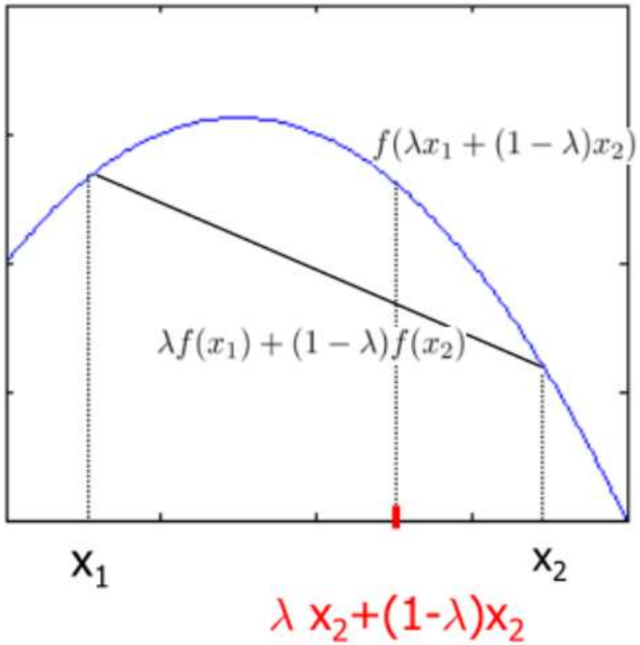
解：

$$H(X)-H(Y)=0.4\times H(1/2)+0.1\times H(1/2)+0.2\times H(1/4)=0.6623\text{比特}/\text{符号}$$

2. 熵函数的性质-上凸性

（7）上凸性

凸函数的定义



$$\forall \lambda (0 < \lambda < 1)$$

$$f[\lambda x_1 + (1-\lambda)x_2] \geq \lambda f(x_1) + (1-\lambda) f(x_2)$$

几何描述：
在上凸函数的任意两点之间画一条割线，函数曲线总在割线上面。

凸函数的定义及詹森不等式

$$f\left[\sum_{k=1}^{q} \lambda_k x_k\right] \geq \sum_{k=1}^{q} \lambda_k f(x_k)$$

2. 熵函数的性质-上凸性

引理（香农辅助定理）：

$$H(p_1,p_2,\dots,p_n)=-\sum_{i=1}^{n}p_i\log p_i\le-\sum_{i=1}^{n}p_i\log q_i$$

其中 $\sum_{i=1}^{n}p_i=1,\quad \sum_{i=1}^{n}q_i=1$

证明： $H(p_1,p_2,\dots,p_n)+\sum_{i=1}^{n}p_i\log q_i=-\sum_{i=1}^{n}p_i\log p_i+\sum_{i=1}^{n}p_i\log q_i$

$$=\sum_{i=1}^{n}p_i\log\frac{q_i}{p_i}$$

$$\le\log\sum_{i=1}^{n}\left(p_i\cdot\frac{q_i}{p_i}\right)=0$$

$\frac{q_i}{p_i}=1,\ \ i=1,2,\cdots,n$ 时等号成立。

2. 熵函数的性质-上凸性

$H(p_1,p_2,...,p_n)$是概率分布$(p_1,p_2,...,p_n)$的严格上凸函数，即

$$H\left[\alpha\mathbf{p}+(1-\alpha)\mathbf{p}'\right]>\alpha H(\mathbf{p})+(1-\alpha)H(\mathbf{p}') \qquad 0<\alpha<1$$

证明： $\alpha p_i+(1-\alpha)\ p_i'$可以被看做是一种新的概率分布。

（因为 $0\le\alpha p_i+(1-\alpha)\ p_i'\le 1$）

$$H\left[\alpha\mathbf{p}+(1-\alpha)\mathbf{p}'\right]=-\sum_{i=1}^{n}\left[\alpha p_i+(1-\alpha)\ p_i'\right]\log\left[\alpha p_i+(1-\alpha)\ p_i'\right]$$

$$=-\alpha\sum_{i=1}^{n}p_i\log\left[\alpha p_i+(1-\alpha)\ p_i'\right]-(1-\alpha)\sum_{i=1}^{n}p_i'\log\left[\alpha p_i+(1-\alpha)\ p_i'\right]$$

$$\ge-\alpha\sum_{i=1}^{n}p_i\log p_i-(1-\alpha)\sum_{i=1}^{n}p_i'\log p_i'$$

等号成立条件

$$=\alpha H(\mathbf{p})+(1-\alpha)H(\mathbf{p}')$$

2. 熵函数的性质-上凸性

当 $\dfrac{\alpha p_i + (1-\alpha)\ p_i'}{p_i} = 1$ 且 $\dfrac{\alpha p_i + (1-\alpha)\ p_i'}{p_i'} = 1$ 时等号成立。

但是

当$\mathbf{p} \neq \mathbf{p}'$时，$\dfrac{\alpha p_i + (1-\alpha)\ p_i'}{p_i} \neq 1$ 且 $\dfrac{\alpha p_i + (1-\alpha)\ p_i'}{p_i'} \neq 1$

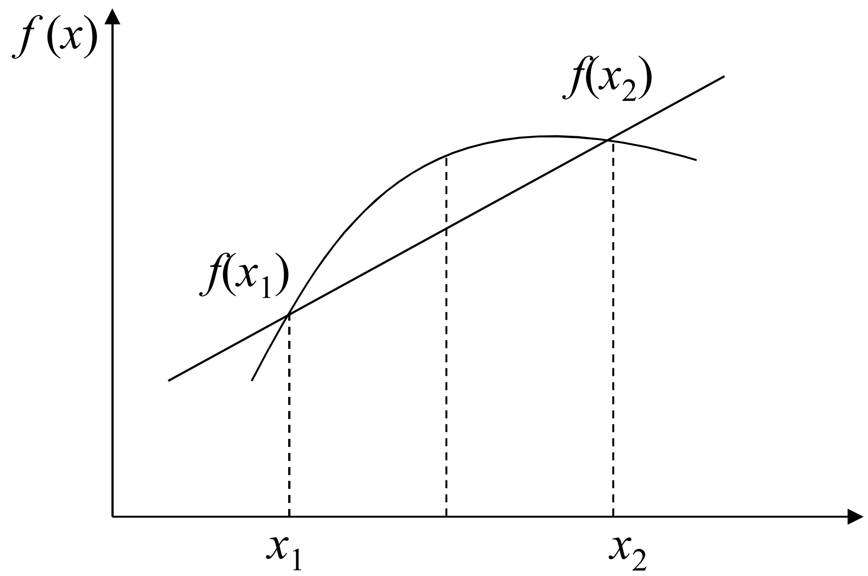
（因为$0 < \alpha < 1)$

所以等号不成立。

2. 熵函数的性质-上凸性

上凸性的几何意义：

在上凸函数的任两点之间画一条割线，函数总在割线的上方.



上凸函数在定义域内的极值必为最大值，这对求最大熵很有用。

2. 熵函数的性质-极值性

（8）极值性（最大离散熵定理）

定理： 离散无记忆信源输出$\boldsymbol{n}$个不同的信息符号，当且仅当各个符号出现概率相等时(即$p_i=\frac{1}{n}$)，熵最大，即

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

2. 熵函数的性质-极值性

$$\frac{\partial}{\partial p(x_i)}\left(\sum_{i=1}^{n} p(x_i)\right) = \frac{\partial}{\partial p(x_i)}\left[p(x_1)+\cdots+p(x_i)+\cdots p(x_n)\right] = 1$$

在约束条件$\sum_{i=1}^{n} p(x_i) = 1$下求$H(\mathbf{p}) = -\sum_{i=1}^{n} p(x_i)\log_e p(x_i)$的极值。

解：作辅助函数$F = -\sum_{i=1}^{n} p(x_i)\log_e p(x_i) - \lambda\left(\sum_{i=1}^{n} p(x_i) - 1\right)$

$$\frac{\partial F}{\partial p(x_i)} = -\log_e p(x_i) - 1 - \lambda \qquad i = 1,2,\cdots,n$$

令$\dfrac{\partial F}{\partial p(x_i)} = 0,$　　得 $p(x_i) = e^{-(\lambda+1)}$

由$\sum_{i=1}^{n} p(x_i) = 1$　　得$\mathrm{e}^{-(\lambda+1)} = \dfrac{1}{n}$　　$\therefore p(x_i) = \dfrac{1}{n}$　　$i = 1,2,\cdots,n$

2. 熵函数的性质

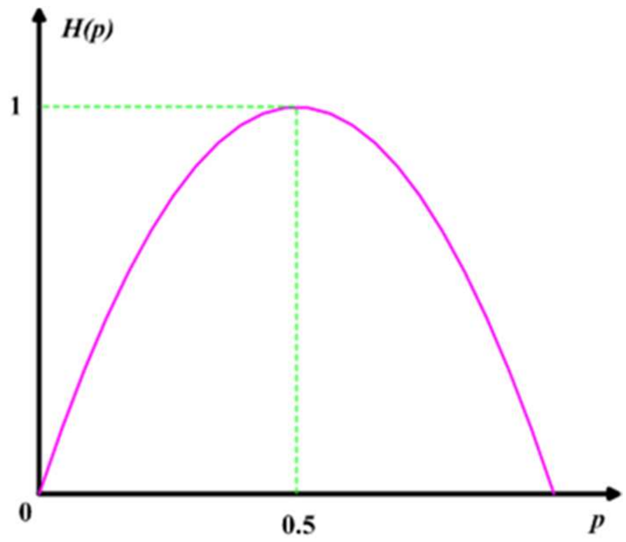
例5：

以二进制信源为例，信源的概率空间为 $\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ p & 1-p \end{bmatrix}$

二进制信源的信息熵为 $H(X) = -[p\log p + (1-p)\log(1-p)]$

这时信息熵$H(X)$是p的函数，

熵函数H(p)的曲线如图所示：



2. 熵函数的性质

从图中可以得出熵函数的一些性质：

如果二进制信源的输出是确定的(p=0或p=1)，则该信源不提供任何信息；

当二进制信源符号0和1等概率发生时，信源的熵达到最大值，等于1比特/符号；

在等概率的二进制信源输出的二进制数字序列中，每一个二元数字提供1比特的信息量。如果符号不是等概率分布，则每一个二元数字所提供的平均信息量小于1比特。

这也进一步说明了计算机术语中的“比特”与信息量单位“比特”的关系。

3. 联合熵和条件熵

定义2.4 随机变量X和Y的联合分布为$p(x_iy_j)$，则这两个随机变量的联合熵定义为：

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

联合熵表示对于二维随机变量的平均不确定性。

- 表示通信完成之后，观察者对通信系统仍然存在的平均不确定度
- 对于观察来说
  - H(X)+H(Y)称为先验不确定度
  - H(XY)称为后验不确定度

3. 联合熵和条件熵

定义2.5 随机变量X和Y的条件熵定义为：

$$H(X \mid Y) = -\sum_i \sum_j p(x_i y_j) \log p(x_i \mid y_j)$$

$$H(Y \mid X) = -\sum_i \sum_j p(x_i y_j) \log p(y_j \mid x_i)$$

条件熵表示已知一个随机变量时，对另一个随机变量的平均不确定性。

- 表示接收者收到Y后，对信源X仍然存在的平均不确定度
- 对于接收者来说，H(X)称为先验不确定度，H(X|Y)称为后验不确定度。

3. 联合熵和条件熵

$$
\begin{aligned}
H(Y|X) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i y_j)\log p(y_j|x_i) \\
&= -\sum_{i=1}^{n} p(x_i)\sum_{j=1}^{m} p(y_j|x_i)\log p(y_j|x_i) \\
&= -\sum_{i=1}^{n} p(x_i) H(Y|x_i) \\
&= E\left[H(Y|x_i)\right]
\end{aligned}
$$

$H(Y|X=x_i)$表示在已知$X=x_i$的情况下，**Y**的平均不确定性。

对于不同的 x_i，$H(Y|X=x_i)$是变化的。因此，$H(Y|X=x_i)$是一个随机变量。

3. 联合熵和条件熵

例6： 已知 联合概率分布如下，求：$H(XY)$，$H(X), H(Y)$, $H(Y|X), H(X|Y)$。

$$
\begin{array}{c}
 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ \\
\end{array}
\begin{array}{c}
\begin{array}{cccc} y_1 & y_2 & y_3 & y_4 \end{array} \\
\begin{bmatrix}
0.25 & 0 & 0 & 0 \\
0.10 & 0.30 & 0 & 0 \\
0 & 0.05 & 0.10 & 0 \\
0 & 0 & 0.05 & 0.10 \\
0 & 0 & 0.05 & 0
\end{bmatrix} \\
\begin{array}{cccc} 0.35 & 0.35 & 0.20 & 0.10 \end{array}
\end{array}
\begin{array}{c}
 \\ 0.25 \\ 0.40 \\ 0.15 \\ 0.15 \\ 0.05 \\ \\
\end{array}
$$

3. 联合熵和条件熵

解： 1) $$H(XY) = -\sum_i \sum_j p(x_i y_j)\log p(x_i y_j)$$

$$= -0.25\log 0.25 - 0.10\log 0.10 - \ldots$$

$$= 2.665$$

2) $$\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{P} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_3 & \boldsymbol{x}_4 & \boldsymbol{x}_5 \\ 0.25 & 0.40 & 0.15 & 0.15 & 0.05 \end{bmatrix}$$ $H(X)=2.066$

3) $$\begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{P} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \boldsymbol{y}_3 & \boldsymbol{y}_4 \\ 0.35 & 0.35 & 0.20 & 0.10 \end{bmatrix}$$ $H(Y)=1.856$

3. 联合熵和条件熵

4) $$H(Y|X) = -\sum_i \sum_j p(x_i y_j) \log p(y_j | x_i)$$

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

$$\begin{array}{c|c|c}
 & \begin{matrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \boldsymbol{y}_3 & \boldsymbol{y}_4 \end{matrix} & \\
\begin{matrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \boldsymbol{x}_3 \\ \boldsymbol{x}_4 \\ \boldsymbol{x}_5 \end{matrix} & \begin{bmatrix} 0.25 & 0 & 0 & 0 \\ 0.10 & 0.30 & 0 & 0 \\ 0 & 0.05 & 0.10 & 0 \\ 0 & 0 & 0.05 & 0.10 \\ 0 & 0 & 0.05 & 0 \end{bmatrix} & \begin{matrix} 0.25 \\ 0.40 \\ 0.15 \\ 0.15 \\ 0.05 \end{matrix}
\end{array}$$

$$\begin{array}{cc}
p(y_j | x_i) & \begin{matrix} y_1 & y_2 & y_3 & y_4 \end{matrix} \\
\begin{matrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{matrix} & \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{3}{4} & 0 & 0 \\ 0 & \frac{1}{3} & \frac{2}{3} & 0 \\ 0 & 0 & \frac{1}{3} & \frac{2}{3} \\ 0 & 0 & 1 & 0 \end{bmatrix}
\end{array}$$

$$H(Y|X) = -0.25 \log 1 - 0.10 \log \frac{1}{4} - 0.30 \log \frac{3}{4} - \ldots$$

$$= 0.600$$

3. 联合熵和条件熵

5) $$H(X|Y) = -\sum_i \sum_j p(x_i y_j) \log p(x_i | y_j)$$

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

$$\begin{array}{c c} & \begin{array}{cccc} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \boldsymbol{y}_3 & \boldsymbol{y}_4 \end{array} \\ \begin{array}{c} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \boldsymbol{x}_3 \\ \boldsymbol{x}_4 \\ \boldsymbol{x}_5 \end{array} & \begin{bmatrix} 0.25 & 0 & 0 & 0 \\ 0.10 & 0.30 & 0 & 0 \\ 0 & 0.05 & 0.10 & 0 \\ 0 & 0 & 0.05 & 0.10 \\ 0 & 0 & 0.05 & 0 \end{bmatrix} \\ & \begin{array}{cccc} 0.35 & 0.35 & 0.20 & 0.10 \end{array} \end{array}$$

$$\begin{array}{c c} p(x_i | y_j) & \begin{array}{cccc} y_1 & y_2 & y_3 & y_4 \end{array} \\ \begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array} & \begin{bmatrix} \frac{5}{7} & 0 & 0 & 0 \\ \frac{2}{7} & \frac{6}{7} & 0 & 0 \\ 0 & \frac{1}{7} & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{4} & 1 \\ 0 & 0 & \frac{1}{4} & 0 \end{bmatrix} \end{array}$$

$$H(X|Y) = -0.25 \log \frac{5}{7} - 0.10 \log \frac{2}{7} - 0.30 \log \frac{6}{7} - \ldots$$

$$= 0.809$$

3. 联合熵和条件熵

各种熵之间的关系

熵的可加性

- $H(XY)=H(X)+H(Y|X)=H(Y)+H(X|Y)$
- $H(X|Y)\le H(X)$，$H(Y|X)\le H(Y)$
- $H(XY)\le H(X)+H(Y)$

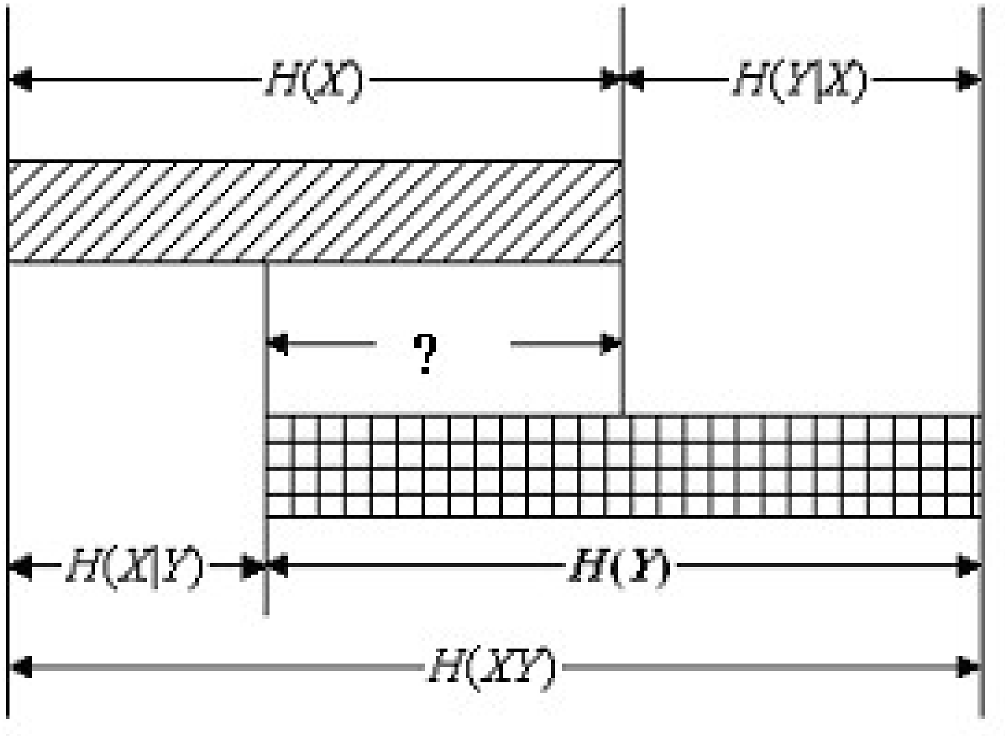
熵的不增原理

若X与Y统计独立，则$H(XY)=H(X)+H(Y)$

可推广到多个随机变量的情况：

$$H(X_1X_2 \ldots X_N) = H(X_1)+H(X_2|X_1)+H(X_3|X_1X_2)+\ldots+H(X_N|X_1X_2\ldots X_{N-1})$$

3. 联合熵和条件熵



$$? = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) = H(X) + H(Y) - H(XY)$$

3. 联合熵和条件熵

加权熵

引入事件的重量，度量事件的重要性或主观价值。

$$\begin{bmatrix} X \\ P \\ W \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \dots & x_n \\ p(x_1) & p(x_2) & \dots & p(x_n) \\ w_1 & w_2 & \dots & w_3 \end{bmatrix}$$

加权熵定义为：

$$H_W(X) = -\sum_{i=1}^{n} w_i p(x_i) \log p(x_i)$$

3. 联合熵和条件熵

例7：信源 $X=\{A,B,C\}$，信源 $Y=\{D,E,F,G\}$，已知条件概率分布和X的概率分布，求联合熵 和条件熵。

$P(Y\|X)$	D	E	F	G	$P(X)$
A	1/4	1/4	1/4	1/4	1/2
B	3/10	1/5	1/5	3/10	1/3
C	1/6	1/2	1/6	1/6	1/6

3. 联合熵和条件熵

解：

$P(XY)$	D	E	F	G
A	1/8	1/8	1/8	1/8
B	1/10	1/15	1/15	1/10
C	1/36	1/12	1/36	1/36
$P(Y)$	91/360	33/120	79/360	91/360

$H(XY) = 3.41$

$H(X) = 1.46$　　$H(Y) = 1.99$

$H(Y|X) = 1.95$　　$H(X|Y) = H(XY) - H(Y) = 1.42$