1、信息熵（Information entropy）

熵（entropy)源于热力学热熵。Shannon在研究随机变量不确定性量度时所得的式在数学上与热熵完全类似，所以就把它称作熵，一般称其为Shannon熵（Shannon entropy)或信息熵（information entropy)。

确定性很小的事件(或变量),一般需要的信息越多才能搞清楚，用信息论的语言就是熵越大。确定性很高的事件（或变量）需要引入信息就也少，信息熵就越小。

如何来量化度量信息量？

处理思路：

（1）以离散随机变量为例，若X的概率质量函数为$P(X=x)=p(x)$，我们用一个函数$I(x)$来表示X的信息量，显然$I(x)$是$p(x)$的单减函数，由于独立事件（或变量）同时发生时的信息，应该为各自发生时的信息的和，故一般可取

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

并称之为X的信息量，也可以取$I(x)=(p(x))^{\frac{1}{n}}$等。

注1. 当$X=1_{\{X=x\}}$时，$I(x)=\log\frac{1}{p(x)}$，这里$p(x)=P(X=x)$

注2. 当对数底为2是，信息量的单位为比特（bit），在机器学习中，对数底一般取e，此时的单位为奈特（nat）。

定义 1：离散型随机变量 X 的（信息）熵为

$H(X)=-\sum_{x} p(x)\log p(x)$（这里 $0\log 0 \triangleq 0$）

即：$H(X)=E(\log\frac{1}{p(X)})$

在信息论中，熵代表着根据信息的概率分布对信息编码所需要的最短平均编码长度。

显然，当 X 为 n 个值的离散均匀分布时，所知道的信息最少，即熵最大，事实上，我们有

性质 1：$0\le H(X)\le \log n$

证明：此问题即为给定 n 时，求解

$$\max \quad -\sum_{i=1}^{n} p_i \log p_i$$

$$s.t. \quad \sum_{i=1}^{n} p_i = 1$$

用 Lagrange 乘子法易的，$p_i=\frac{1}{n}(i=1,2,\cdots,n)$

注：$H(X)=0$ 当且仅当 X 为确定性的

定义 2：离散型随机变量 (X,Y) 的联合熵定义为为

$H(X,Y)=-\sum_{x,y} p(x,y)\log p(x,y)=-\sum_{i}\sum_{j} p(x_i,y_j)\log p(x_i,y_j)$

同样有：$H(X,Y)=-E(\log p(X,Y))$

2、条件熵

设 $(X,Y)\sim p(x,y), Y\mid X=x\sim p(y\mid x)$

定义 2：称给定 X 的条件下，Y 的条件熵为

$$H(Y|X)=\sum_x H(Y|X=x)p(x)$$

这里，$H(Y|X=x)=-\sum_y p(y|x)\log p(y|x)$

性质 2：$H(Y|X)\le H(Y)$

性质 3：$H(Y|X)=-\sum_{x,y} p(x,y)\log p(y|x)=-E(\log p(Y|X))$

证明：

$$H(Y|X)=\sum_x H(Y|X=x)p(x)$$
$$=-\sum_x\sum_y p(y|x)\log p(y|x)p(x)$$
$$=-\sum_{x,y} p(x,y)\log p(y|x)$$

性质 4：（链式法则）$H(X,Y)=H(X)+H(Y|X)$，从而有

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

证明：

$$H(X,Y)=-\sum_{x,y} p(x,y)\log p(x,y)$$
$$=-\sum_{x,y} p(x,y)\log[p(y|x)p(x)]$$
$$=-\sum_{x,y} p(x,y)\log p(x)-\sum_{x,y} p(x,y)\log p(y|x)$$
$$=-\sum_x p(x)\log p(x)-\sum_{x,y} p(x,y)\log p(y|x)$$
$$=H(X)+H(Y|X)$$

即描述 X 和 Y 所需的信息就是描述 X 自己所需的信息,加上给定 X 的条件下具体化 Y 所需的额外信息。

一般还有：$(X_1,X_2,\cdots,X_n)\sim p(x_1,x_2,\cdots,x_n)$，则

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

性质 5：$H(X,Y\mid Z)=H(X\mid Z)+H(Y\mid X,Z)$

性质 6：$H(X)\geq H(g(X))$，且等号成立当且仅当 g 可逆。

证明：

$$H(X,g(X))=H(X)+H(g(X)\mid X)=H(g(X))+H(X\mid g(X))$$

由于 $H(g(X)\mid X)=0$

故 $H(X)-H(g(X))=H(X\mid g(X))\geq 0$

3、连续情形

$$H(X)=-E(\log f(X))=-\int_{-\infty}^{\infty}f(x)\log f(x)dx;$$

$$H(X,Y)=-E(\log f(X,Y))=-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x,y)\log f(x,y)dxdy;$$

$$H(Y\mid X)=-E(\log f(Y\mid X))=-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x,y)\log f(y\mid x)dxdy;$$

连续情形的许多性质同离散情形，但有些性质不再有，如性质 1 与性质 5.

例：不再有非负性，设 $X\sim U[a,b]$，则

$H(X)=-\int_a^b\frac{1}{b-a}\log\frac{1}{b-a}dx=\log(b-a)$，

若 $0<b-a<1$，则 $H(X)<0$。

也不一定有上界，但若 $EX^2 \le K < +\infty$，则有 $\max h(X) = \frac{1}{2}\log(2K\pi e)$，且最大值当 $X \sim N(0,K)$ 时达到（思考：用 Lagrange 乘子法）

性质 5 也不再成立，因为无非负性。

4、相对熵 (Relative entropy)与互信息（Mutual information）

相对熵也称 KL 散度 (Kullback－Leibler divergence)。

定义 3：设 $p(x), q(x)$ 为两个取相同值的概率分布，则 $p(x)$ 对 $q(x)$ 的相对熵为

$$D_{KL}(p \| q) = \sum_x p(x)\log\frac{p(x)}{q(x)} = E_p(\log\frac{p(X)}{q(X)}) = E_p(\log p(X) - \log q(X))$$

连续场合下定义为

$$D_{KL}(p \| q) = \int_{-\infty}^{\infty} f(x)\log\frac{f(x)}{g(x)}dx = E_f(\log\frac{f(X)}{g(X)})$$

相对熵可理解为用（拟合）分布 $q(x)$ 来描述真实分布 $p(x)$ 时，其信息损耗的一种度量，有时可以看作一种“距离”，但它不是真正的距离。实际上，相对熵可以用来衡量两个概率分布之间的差异，上面公式的意义就是 $p(x), q(x)$ 之间的对数差在 $p(x)$ 下的期望值。

相对熵具有如下性质：

性质 7：非对称性：$D_{KL}(p \| q) \ne D_{KL}(q \| p)$

例：$p(x) \sim \begin{pmatrix} 0 & 1 \\ 1-r & r \end{pmatrix}; q(x) \sim \begin{pmatrix} 0 & 1 \\ 1-s & s \end{pmatrix}$

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

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

性质 8：非负性（Gibbs 不等式）：$D_{KL}(p\|q)\geq 0$，等号当且仅当 $p(x)=q(x)$ 时成立。

证明：

$$D_{KL}(p\|q)=\sum_x p(x)\log\frac{p(x)}{q(x)}$$

$$=-\sum_x p(x)\log\frac{q(x)}{p(x)}$$

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

$$\geq-\log E_p(\frac{q(X)}{p(X)})(Jensen)$$

$$=-\log\sum_x p(x)\frac{q(x)}{p(x)}$$

$$=0(\sum_x q(x)=1)$$

注：若 $p(x),q(x)$ 相同，则 $D_{KL}(p\|q)=0$

注：非对称性的改进：JS（Jensen-Shannon）散度

$$JS(p\|q)=\frac{1}{2}D_{KL}(p\|\frac{p+q}{2})+\frac{1}{2}D_{KL}(q\|\frac{p+q}{2})$$

JS 散度具有对称性（$JS(p\|q)=JS(q\|p)$），且值域为[0,1]。进一步的改进有 **Wasserstein** 距离（俗称推土机距离（EMD））等。

定义 4：设$(X,Y)\sim p(x,y)$，其边缘分布为$X\sim p_X(x),Y\sim p_Y(y)$，则互信息$I(X,Y)$为联合分布对边缘分布乘积$p_X(x)p_Y(y)$的相对熵，即$I(X,Y)=D_{KL}(p(x,y)\|p_X(x)q_Y(y))$

显然有

$$I(X,Y)=\sum_x\sum_y p(x,y)\log\frac{p(x,y)}{p_X(x)q_Y(y)}=E_{(X,Y)}(\log\frac{p(X,Y)}{p_X(X)q_Y(Y)})$$

性质 9：

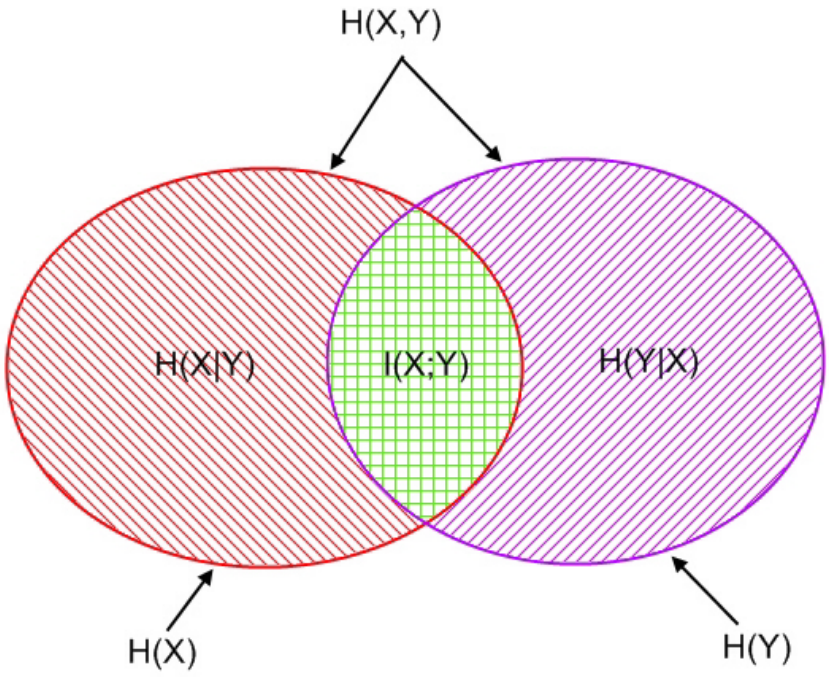
$$I(X,Y)=I(Y,X)$$
$$I(X,X)=H(X)$$
$$I(X,Y)=H(X)-H(X\mid Y)$$
$$I(X,Y)=H(X)+H(Y)-H(X,Y)$$

证明：

$$I(X,Y)=\sum_x\sum_y p(x,y)\log\frac{p(x,y)}{p_X(x)q_Y(y)}$$
$$=\sum_x\sum_y p(x,y)\log\frac{p(x\mid y)}{p_X(x)}$$
$$=-\sum_x\sum_y p(x,y)\log p_X(x)-(-\sum_x\sum_y p(x,y)\log p(x,y))$$
$$=H(X)-H(X\mid Y)$$
$$H(X,Y)=H(Y)+H(X\mid Y)$$

故$I(X,Y)=H(X)+H(Y)-H(X,Y)$

性质 10：$I(X,Y)\geq 0$，且等号当且仅当 X,Y 独立时成立。（对所有类型随机变量）



4、交叉熵(Cross entropy)

设 $p(x),q(x)$ 为两个取相同值的概率分布，其中 $p(x)$ 为真实分布，$q(x)$ 为拟合分布，如果用真实分布 $p(x)$ 来衡量识别别一个样本，所需要编码长度的期望（平均编码长度）为：

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

如果使用拟合分布 $q(x)$ 来表示来自真实分布 $p(x)$ 的平均编码长度，则是：

$$H(p,q)=\sum_{x}p(x)\log\frac{1}{q(x)}$$

（因为用$q(x)$来编码的样本来自于分布$p(x)$，所以$H(p,q)$中的概率是$p(x)$）。此时就将$H(p,q)$称之为交叉熵。

由于$D_{KL}(p\|q)=\sum_x p(x)\log\frac{p(x)}{q(x)}$，故有

$$D_{KL}(p\|q)=H(p,q)-H(p)，$$

事实上，有

$$D_{KL}(p\|q)=\sum_x p(x)\log\frac{p(x)}{q(x)}$$
$$=\sum_x p(x)\log\frac{1}{q(x)}-\sum_x p(x)\log\frac{1}{p(x)}$$
$$=H(p,q)-H(p)$$

此式表明：相对熵为用非真实分布$q(x)$得到的平均码长比用真实分布$p(x)$得到的平均码长所要多出的bits数）

由于$D_{KL}(p\|q)\geq 0$，故$H(p,q)\geq H(p)$

例：若X的真实分布为$p(x)=\left(\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{8}\right)$，非真实分布$q(x)=\left(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\right)$。

$$H(p)=\sum_x p(x)\log_2\frac{1}{p(x)}$$
$$=\frac{1}{2}\log_2 2+\frac{1}{4}\log_2 4+\frac{1}{8}\log_2 8+\frac{1}{8}\log_2 8=1.75(bits)$$

$$H(p,q)=\sum_x p(x)\log_2\frac{1}{q(x)}$$
$$=\frac{1}{2}\log_2 4+\frac{1}{4}\log_2 4+\frac{1}{8}\log_2 4+\frac{1}{8}\log_2 4=2(bits)$$

由此可以看出根据非真实分布$q(x)$得到的平均码长大于根据真实分布$p(x)$得到的平均码长。

5.条件互信息与条件相对熵

定义6.给定Z的条件下，X,Y的条件互信息定义为

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

$$=E_{(X,Y,Z)}(\log\frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)})$$

定义7.二元分布$p(x,y)$与$q(x,y)$的条件相对熵定义为

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

$$=E_{p(x,y)}[\log\frac{p(Y|X)}{q(Y|X)}]=E_{p(x,y)}[\log\frac{q(X)p(X,Y)}{p(X)q(X,Y)}]$$

性质11：

$$D_{KL}(p(x,y)\|q(x,y))=D_{KL}(p(x)\|q(x))+D_{KL}(p(y|x)\|q(y|x))$$

证明：

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

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

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

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

$$=D_{KL}(p(x)\|q(x))+D_{KL}(p(y|x)\|q(y|x))$$

在机器学习中，我们希望在训练数据上模型学到的分布 p_M 和真实数据的分布 p_R 越接近越好，一般，我们可以使其相对熵最小。但是我们没有真实数据的分布，所以只能希望模型学到的分布 p_M 和训练数据的分布 p_T 尽量相同。为此，我们只需要最小化 $D_{KL}(p_T \| p_M)$，这里 p_T 可以理解为真实分布，p_M 为非真实分布，又因为训练分布 p_T 是给定的（因为训练数据分布是固定的，$H(p_T)$ 为常数），所以最小化 $D_{KL}(p_T \| p_M)$ 等价于最小化 $H(p_T, p_M)$（因为 $D_{KL}(p_T \| p_M) = H(p_T, p_M) - H(p_T)$）所以，我们说交叉熵可以用来计算学习模型分布与训练分布之间的差异。