1

1.1 信息量

信息量作为信息的度量，可以用来衡量熵的定义，设 $p(x_i)$ 表示x_i 发生的概率，则信息量可以表示为：

$$h(x_i) = `log_a p(x_i) = log_a \frac{1}{p(x_i)} \tag{1}$$

a值常取2，表示比特，即非0即1，由此可知，信息量与概率成反比.

1.2 熵

熵在热力学熵用来描述物质的混乱程度，用来衡量不确定性，物质越混乱，不确定性越大，熵值越大。到信息论中，事件发生的不确定行越大，则熵越大。例如：掷骰子，六个面机会均等，因此投一次得到的点数不确定性最大（因为每个点数的概率都是六分之一），因此此时熵最大；熵是信息量的期望，公式如下：

$$H(X) = -\sum_{i=1}^{n} p(x_i) log_a p(x_i) \tag{2}$$

表示信息量的期望，反应不确定性。

1.3 联合熵

联合熵可以表示为两个事件的熵的并集：

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{n} p(x_i,y_i) log_a p(x_i,y_j) \tag{3}$$

其中 $max(H(X),H(Y)) <= H(X,Y) <= H(X) + H(Y)$

1.4 互信息

互信息是用来表示变量间相互以来的程度，常用在特征选择和特征关联性等方面，公式如下：

$$I(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{n} p(x_i,y_i) log_a \frac{p(x_i,y_j)}{p(x_i)p(y_j)} \tag{4}$$

互信息与相关性 ρ 相关， ρ 用来描述线性相关性，互信息用来描述非线性相关性，其中：

$$\rho = \frac{cov(x,y)}{\sqrt{x}\sqrt{y}} \tag{5}$$

1.5 条件熵

条件熵实际上是联合熵与熵的差集，也可表示为熵与互信息的差集，具体如下：

$$H(X|Y) = H(X,Y) - H(Y) = H(X) - I(X,Y) \tag{6}$$

1.6 相对熵

相对熵用来描述两个分布之间的差异。

$$KL(p||q) = \sum_{i=1}^{n} p(x_i) log \frac{p(x_i)}{q(x_i)} \tag{7}$$

其中，p，q表示两个分布，其中： $KL(p||q)$‘$KL(q||p)$ KL散度越大，两个分布间的差异越明显，并且: $KL(p||q)e0$

1.7 交叉熵

主要用于度量两个概率分布间的差异性信息, 在分类任务中常用做目标函数为什么KL散度用来衡量两个分布的相似程度,交叉熵也用来衡量,请往后看,交叉熵的公式为:

$$H(p,q) = \sum_{i=1}^{n} p(x_i) log(\frac{1}{q(x_i)}) \tag{8}$$