信息量

熵在信息学中的概念就是一个系统“内在的混乱程度”，也可以理解为系统中所含的信息量大小。比如说一句话：“杨倩在东京奥运会中夺得了金牌”，如果在东京奥运会之前说这句话，那么这句话包含的信息量就很大，因为在赛前杨倩能不能拿金牌是一件非常不确定的事；但是当奥运会结束后，这句话包含的信息量就非常小了，因为杨倩已经确定拿了金牌。所以信息量也可以定义为：事件又不确定变为确定的困难程度。假设将信息量定义为：f(x) := 信息量。

那么我们可以得到 f（杨倩赢得金牌）= f（杨倩进了决赛）+ f（杨倩赢了决赛）。根据概率论的方法，还可以得到：p（杨倩赢得金牌）= p（杨倩进了决赛）× f（杨倩赢了决赛）。这两个式子应该是等价的，那么 f(x)的表达式中一定含有 log，才可以让相乘变相加。且符号为负，因为概率越大，信息量越小。即：

$$f(x) := -log_2 x$$

底数为 2，就可以使计算结果的单位是比特。

熵

然后看信息量在一个系统中表现：假设在一场篮球比赛中，A 队获胜的概率是 50%，B 队获胜的概率是 50%，那这两个事件的信息量分别是：

$$f(A) = -log_2 / frac12 = 1 \quad f(B) = -log_2 / frac12 = 1$$

而在另一场篮球比赛中，C 队获胜的概率是 99%，D 对获胜的概率是 1%，那这两个时间的信息量分别是：

$$f(C) = -log_2 / frac99100 / approx0.0145$$

$$f(D) = -log_2 / frac1100 / approx6.6439$$

从直观上看，C 和 D 的比赛信息量应该更小，因为 C 获胜的概率非常大；A 和 B 的比赛信息量应该更大，因为二者谁能赢很难说。但是这两场比赛的信息量该如何计算呢？如果简单的将系统的所有事件的信息量相加，那从上面这个例子来看肯定是不合适的。所以系统整体的信息量应该是每个事件的概率乘上它的信息量，再累加起来。这样计算下来，A 和 B 的比赛的信息量是 1，C 和 D 的比赛的信息量大约是 0.08079。也就是说 A 和 B 的系统的熵为 1，C 和 D 的系统的熵为 0.08079。

从这个例子中可以知道熵实质上求的是系统内所有时间信息量的期望和。那么熵的公式就是：

$$H(p) = \sum_{i=1}^{n} p_i \cdot f(p_i) = \sum_{i=1}^{n} p_i \cdot (-log_2 p_i) = -\sum_{i=1}^{n} p_i \cdot log_2 p_i$$

相对熵（KL 散度）

现在已经知道了如何求一个系统的熵，那我们就可以用熵来求两个系统的差别，即相对熵。假设有 P 和 Q 两个系统，现在以 P 为基准，考虑 Q 与 P 差多少。相对熵公式如下：

$$\begin{aligned} & D_{KL}(P||Q) \\ := & \sum_{i=1}^{n} p_i(f_Q(q_i) - f_P(p_i)) \\ := & \sum_{i=1}^{n} p_i((-log_2 q_i) - (-log_2 p_i)) \\ := & \sum_{i=1}^{n} p_i \cdot (-log_2 q_i) - \sum_{i=1}^{n} p_i \cdot (-log_2 p_i) \end{aligned}$$

现在我们的计算是以 P 系统为基准的，那么式子最后一行减号后的那一项是固定不变的，现在要想让 Q 和 P 最接近，那么减号前一项就要越小越好，而这一项就是交叉熵。

$$H(P,Q) = \sum_{i=1}^{n} p_i \cdot (-log_2 q_i)$$

神经网络中的交叉熵

在神经网络中，我们要比较的两个模型分别是神经网络所拟合出的模型和神经网络要逼近的那个真实概率模型。那么 H(P,Q)中的 Q 就是神经网络所拟合出的模型，P 就是要逼近的那个真实概率模型，现在我们要以神经网络要逼近的那个概率模型为基准，看看当下神经网络所拟合出的模型与真实概率模型的差别。神经网络输入的是一个长长的序列，也就是真实概率模型的真实数据，输出的是每个样本对应的输出 ，那么交叉熵公式中的 就是 ， 就是 。如下所示：

$$H(P,Q) = \sum_{i=1}^{n} p_i \cdot (-log_2 q_i) = \sum_{i=1}^{n} x_i \cdot (-log_2 y_i)$$

在二分类问题中，如果为 1 表示是正样本，为 0 表示是负样本，而 表示的是一个概率，那么二者就是不同的事件，所以当 为 1 时， 要表示是正样本的概率；为 0 时，要表示不是正样本的概率。所以交叉熵最后的公式就是：

$$H(P,Q) = -\sum_{i=1}^{n} (x_i \cdot log_2 y_i + (1 - x_i) \cdot log_2 (1 - y_i))$$

在多分类问题中，只需要计算每个分类的熵的期望即可，此时交叉熵的公式就是：

$$H(P,Q) = -\sum_{i=1}^{n} x_i \cdot log_2 y_i$$