# 信息论基础

信息论的由来：

信息论是应用数学的一个分支，主要研究的是对一个信号能够提供信息的多少进行量化，最初用于研究在一个含有噪声的信道上用离散的字母表来发送消息，指导最优的通信编码等

关于信息的一个基本想法

一个不太可能的事情竟然发生了要比一个非常可能的事件的发生能提供更多的信息，也就是说导致那些“异常”事件发生的背后拥有着我们更想知道的东西

## 信息熵

自信息：

一个事件所包含的信息

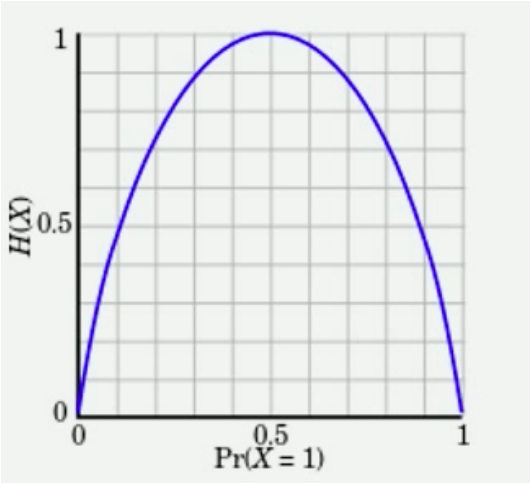
$$I(x) = -logP(x)$$

信息熵：

随机变量或整个系统的不确定性，熵越大，随机变量或系统的不确确定性就越大。它描述的是有关事件x的所有可能结果的自信息期望值。

$$H(X) = E[-logp_i] = -\sum_{i=1}^{n} p_i logp_i$$

其中，n 代表事件x的所有n种可能的取值，$p_i$代表了事件x为i时的概率




例：抛掷一枚不均匀硬币，正面的概率为1/3， 背面的概率为2/3，这一事件的信息熵为：

$$H(X) = E[-logp_i] = -\sum_{i=1}^{2} p_i logp_i$$
$$= -\frac{1}{3}log\frac{1}{3} + \left(-\frac{2}{3}log\frac{2}{3}\right)$$
$$= \frac{1}{3}log3 + \frac{2}{3}log\frac{3}{2}$$

信息熵的意义：

☛ 熵的作用计算损失(Loss function)用于调整梯度递减的步长本次熵（损失）比上次熵（损失）大，说明步长太大了

☛ 用于决策树熵越大，说明特征(feature)的划分数据能力越强

联合熵

作用：

度量二维随机变量的不确定性

例如，对于随机变量（X, Y）其联合分布为$P(x_i, y_j)$，则联合熵为：

$$H(X,Y) = -\sum_i \sum_j P(x_i, y_j) logP(x_i, y_j)$$

条件熵

定义：

X给定条件下，Y的条件概率分布的熵对X的数学期望（平均不确定性）

例如，对于随机变量Y，在X的条件下其条件熵为：

$$H(Y|X) = \sum_{i=1}^{n} P(X = x_i) H(Y|X = x_i)$$
$$= -\sum_i \sum_j P(x_i, y_j) logP(y_j|x_i)$$

信息熵，联合熵，条件熵的相互关系和结果推导

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

$$H(X,Y) = -\sum_i \sum_j P(x_i, y_j) log P(x_i, y_j)$$

$$= -\sum_i \sum_j P(x_i, y_j) log P(y_j|x_i) + (-\sum_i \left( \sum_j P(x_i, y_j) \right) log P(x_i))$$

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

**相对熵**

相对熵又称为KL散度，信息散度，信息增益。

作用：

主要用来衡量两个分布的相似度，假设连续随机变量x，真实的概率分布为p（x），模型得到的近似分布为q(x)

例如，对于分布p, q 则相对熵为：

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

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

在贝叶斯推理中，相对熵衡量当你修改了从先验分布 q 到后验分布 p 的信念之后带来的信息增益

**互信息**

作用：

用来衡量两个相同的一维分布变量之间的独立性
I (X,Y) 是衡量联合分布p(x,y)和p(x)p(y)分布之间的关系，即他们之间的相关系数

互信息

设两个随机变量（X, Y）的联合分布为p(x,y)，边际分布分别为p(x), p(y)，则互信息 I (X,Y)：

$$I(X;Y) = \sum_{x \epsilon X} \sum_{y \epsilon Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$$

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

从公式中可以看出互信息是满足对称性的，其在特性选择、分布的距离评估中应用非常广泛

互信息在信息论和机器学习中非常重要，其可以评价两个分布之间的距离，这主要归因于其对称性，假设互信息不具备对称性，那么就不能作为距离度量，例如相对熵，由于不满足对称性，故通常说相对熵是评价分布的相似程度，而不会说距离。互信息的定义为：一个随机变量由于已知另一个随机变量而减少的不确定性，或者说从贝叶斯角度考虑，由于新的观测数据y到来而导致x分布的不确定性下降程度。

信息熵，条件熵，互信息的相互关系和结果推导

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
&= H(X) + H(Y) - H(X,Y) \\
&= \sum_{x} p(x) log \frac{1}{p(x)} + \sum_{y} p(y) log \frac{1}{p(y)} - \sum_{x,y} p(x,y) log \frac{1}{p(x,y)} \\
&= \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}
\end{aligned}
$$

信息增益

假设系统原有的熵为H(Y)，后来引入了特征T，在固定特征T的情况下，系统的混乱度减小，熵减小为 H(Y|T)，那么特征 T给系统带来的信息增益为: