# 信息熵

## 信息量

在观测随机变量的时候，我们需要一个变量来定义每次观测到一个样本所获得的信息量。

比如对于太阳从东方升起，我们观测到了也无法获取任何信息，而如果一个发生概率很小的事件发生了，我们所获得的信息是非常大的。

因此信息的度量应该依赖于概率值，且单调递减。

再假设，如果X Y两个随机变量相互独立，那么分别观测两者，所获得的信息是叠加的，所以其具有可加性。

显然信息不可能是负数，不存在观测一个事件后，所持有的信息反而减少的情况。

罗列其需要具备的性质

* 单调性，相对单调递减
* 非负性，值恒为正
* 累加性，当相互独立时，

我们将信息量定义为

其底数不一定是2，在机器学习中也使用自然常数e。

其负号为了保证信息恒为正数

## 信息熵

为了描述一个随机变量的离散情况（也就是熵的物理意义），定义信息熵

从公式角度也可以理解为此分布下，信息量的期望值。换言之需要多少信息才能确定这个分布，也可以认为是这个分布所蕴含的信息的描述。

利用拉格朗日乘子可证，为随机变量可能的取值

以二元信源为例，求导可得在中是凸函数且在时取得最小值。因此在所有都相等的时候取得最大值。

可见信息熵描述了概率在随机事件中的分布情况，分布越均匀，熵越大，越集中，熵越小。

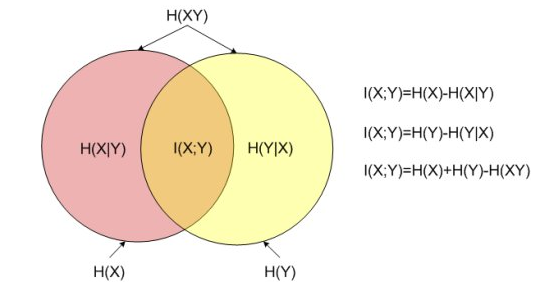
对于独立事件，信息熵满足累加性

此外，如果两个事件相互不独立，那么信息熵为

其中成为互信息

## 互信息

从中我们可以看出，的信息熵来源于各自的信息熵减去，而信息熵本质是信息量根据其概率的加权累加，如下图所示



其中称为边缘熵。表示为条件熵意为当已知时的信息熵。称为联合熵（要区别于交叉熵）。

因此，互信息的含义是在某一方的确定对另一方不确定性的减少程度。

推导：

由

根据全概公式展开

可以将展开

## 交叉熵

交叉熵是用来比较两个分布的数学工具，数学符号是，其中是两个分布（区别于联合熵中参数的两个随机变量）。其表示的是当使用编码方式来对进行编码，其信息熵的值，其显然是不对称的，编码的从属关系显然不能交换。

其公式和信息熵类似，但是因为我们使用的是编码，所以信息量来源于，而编码对象是，那么真实世界代码的分布肯定是符合的，所以概率值来源于。

## 相对熵（KL散度）

当我们使用一种不够优良的编码去编码时，虽然能正确编码，但是肯定存在许多冗余，这些冗余就体现在信息熵比原来更大了。

为了描述这种冗余的大小，我们引入相对熵也就是KL散度

KL散度的一大用途就是描述两个分布的相似程度，即冗余越小，相似程度越大。显然其也是不对称的。