<http://www.cnblogs.com/heaad/archive/2011/01/02/1924088.html>

http://xiaofeng1982.blog.163.com/blog/static/315724582011628102353740/

# 信息熵：

一个 *X* 值域为{*x*1, ..., *xn*}的随机变量的熵值 H 定义为：

H(X)  =  \operatorname{E}(I(X))，

其中，E 代表了[期望](http://zh.wikipedia.org/w/index.php?title=%E6%9C%9F%E6%9C%9B&action=edit&redlink=1)函数，而 *I*(*X*) 是 *X* 的信息量（又称为[信息本体](http://zh.wikipedia.org/wiki/%E8%B3%87%E8%A8%8A%E6%9C%AC%E9%AB%94)）。*I*(*X*) 本身是个随机变量。如果 *p* 代表了 *X* 的[机率质量函数](http://zh.wikipedia.org/wiki/%E6%A9%9F%E7%8E%87%E8%B3%AA%E9%87%8F%E5%87%BD%E6%95%B8)（probability mass function），则熵的公式可以表示为：

H(X) = \sum_{i=1}^n {p(x_i)\,I(x_i)} = -\sum_{i=1}^n {p(x_i) \log_b p(x_i)}

在这里 *b* 是[对数](http://zh.wikipedia.org/wiki/%E5%B0%8D%E6%95%B8)所使用的[底](http://zh.wikipedia.org/wiki/%E5%BA%95%E6%95%B0_%28%E5%AF%B9%E6%95%B0%29)，通常是 2, 自然常数 [e](http://zh.wikipedia.org/wiki/E_%28%E6%95%B0%E5%AD%A6%E5%B8%B8%E6%95%B0%29)，或是10。当*b* = 2，熵的单位是[bit](http://zh.wikipedia.org/wiki/%E4%BD%8D%E5%85%83)；当*b* = e，熵的单位是 [nat](http://zh.wikipedia.org/wiki/%E5%A5%88%E7%89%B9_%28%E5%8D%95%E4%BD%8D%29)；而当 *b* = 10,熵的单位是 [dit](http://zh.wikipedia.org/w/index.php?title=%E8%BF%AA%E7%89%B9&action=edit&redlink=1)。

*pi* = 0时，对于一些*i*值，对应的被加数0 log*b* 0的值将会是0，这与[极限](http://zh.wikipedia.org/w/index.php?title=%E5%87%BD%E6%95%B0%E6%9E%81%E9%99%90&action=edit&redlink=1)一致。

b=2可以认为是2进制，b=10可以认为是10进制。

如果有一个系统S内存在多个事件S = {E1,...,En}，每个事件的机率分布 P = {p1, ..., pn}，则每个事件本身的讯息（[信息本体](http://zh.wikipedia.org/wiki/%E8%B3%87%E8%A8%8A%E6%9C%AC%E9%AB%94)）为：

I_e = -\log_2 {p_i} （对数以2为底，单位是[比特](http://zh.wikipedia.org/wiki/%E4%BD%8D%E5%85%83)(bit)）

I_e = -\ln {p_i} （对数以e为底，单位是[纳特](http://zh.wikipedia.org/wiki/%E7%BA%B3%E7%89%B9)/nats）

如英语有26个字母，假如每个字母在文章中出现次数平均的话，每个字母的讯息量为：

I_e = -\log_2 {1\over 26} = 4.7 

而汉字常用的有2500个，假如每个[汉字](http://zh.wikipedia.org/wiki/%E6%B1%89%E5%AD%97)在文章中出现次数平均的话，每个汉字的信息量为：

I_e = -\log_2 {1\over 2500} = 11.3 

熵是整个系统的平均消息量

，分支越多，熵越大。信息增益，天然地偏向那些分支较多的属性。（随机事件越多）

X均匀分布时熵最大。

证明：H(x)当|X|个值为等可能时，达到最大值log|X|, |X|为X中元素的个数。

首先证明：



设，则一阶导数为，二阶导数为，二阶导数为负，存在极大值，那么x=1，就是极大值。

于是有.

证明过程：（连续情况下，正态分布，熵最大，prml）

由此可得，

# 信息增益

条件熵



表明了在已知随机变量X的条件下随机变量Y的不确定性。

信息增益：

反应了随机变量A对信息的贡献。

选择特征时，选择信息增益最大。

信息增益比



归一化

# 特征分段

利用信息增益，对连续的特征分段。例如身高。

采用类似二叉树的方式，先对训练数据的特征A的取值，设为listA，划分为两段，使用信息增益求的最大值。然后对这两段的字段再划分，直到算出的信息增益（信息增益比）小于某个域值。

分段信息增益的算法：

将特征A的分成两段，小于阀值的，全部一个值比如0，大于阀值的为另一个值1，这样得到新的特征A’，用A’替换A，得到新的训练数据，求出信息增益。使用穷举法，对A所有的分段求出相应的信息增益，其中最大的信息增益对应的分段，就为特征A的分段。

的计算公式：



# 卡方分布：

**正态分布**的[概率密度函数](http://zh.wikipedia.org/wiki/%E6%A6%82%E7%8E%87%E5%AF%86%E5%BA%A6%E5%87%BD%E6%95%B0)均值为\mu [方差](http://zh.wikipedia.org/wiki/%E6%96%B9%E5%B7%AE)为\sigma^2 (或[标准差](http://zh.wikipedia.org/wiki/%E6%A8%99%E6%BA%96%E5%B7%AE)\sigma)是[高斯函数](http://zh.wikipedia.org/wiki/%E9%AB%98%E6%96%AF%E5%87%BD%E6%95%B8)的一个实例：


f(x;\mu,\sigma)
=
\frac{1}{\sigma\sqrt{2\pi}} \, \exp \left( -\frac{(x- \mu)^2}{2\sigma^2} \right)。

若来自正态总体的*k*个随机变量*Z_1*、……、*Z_k*相互独立，且[数学期望](http://zh.wikipedia.org/wiki/%E6%95%B0%E5%AD%A6%E6%9C%9F%E6%9C%9B)为0、[方差](http://zh.wikipedia.org/wiki/%E6%96%B9%E5%B7%AE)为1（即服从标准正态分布），则随机变量*X*

X=\sum_{i=1}^k Z_i^2

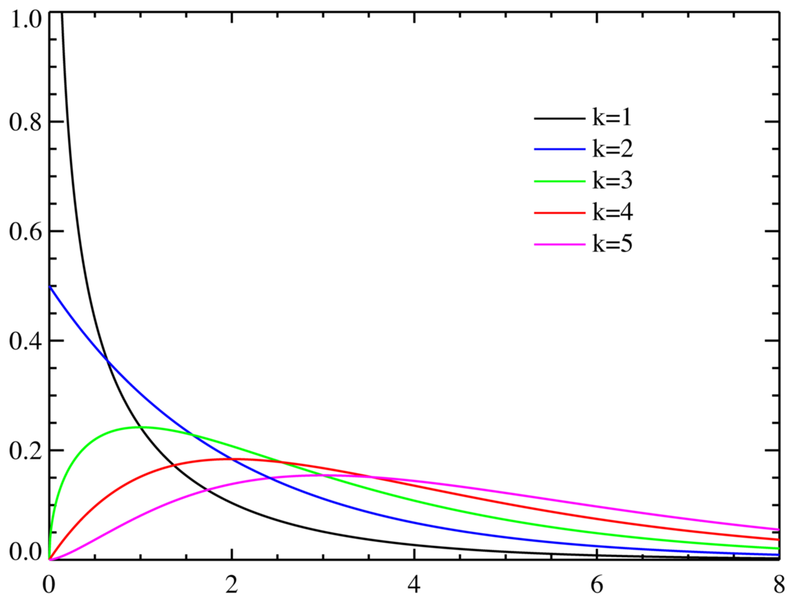
被称为服从[自由度](http://zh.wikipedia.org/wiki/%E8%87%AA%E7%94%B1%E5%BA%A6)为*k*的卡方分布，记作

    X\ \sim\ \chi^2(k). \,
  

卡方分布的[概率密度函数](http://zh.wikipedia.org/wiki/%E6%A6%82%E7%8E%87%E5%AF%86%E5%BA%A6%E5%87%BD%E6%95%B0)为：

f_k(x)=
\frac{(1/2)^{k/2}}{\Gamma(k/2)} x^{k/2 - 1} e^{-x/2}

其中x≥0，当x≤0时f_k(x)=0。这里Γ代表[Gamma函数](http://zh.wikipedia.org/wiki/%CE%93%E5%87%BD%E6%95%B0)。



用在卡方检验上，反应的一个事实是，卡方检验的值，反应了随机变量的自由度概率对应情况。图中，x轴为卡方检验的值，y轴为自由度概率，y值越小，拒绝关联的可能性越小，关联度越大。

卡方分布的[累积分布函数](http://zh.wikipedia.org/wiki/%E7%B4%AF%E7%A7%AF%E5%88%86%E5%B8%83%E5%87%BD%E6%95%B0)为：

F_k(x)=\frac{\gamma(k/2,x/2)}{\Gamma(k/2)}，

其中γ(k,z)为[不完全Gamma函数](http://zh.wikipedia.org/w/index.php?title=%E4%B8%8D%E5%AE%8C%E5%85%A8Gamma%E5%87%BD%E6%95%B0&action=edit&redlink=1)

# 卡方检验：



反应了被检验的t，c两个随机变量之间的关联程度。两个事件 A 和B 独立，是指两个事件 A、B 的概率满足 P(AB)=P(A)P(B) 或者 P( A|B)=P( A) 且P( B|A)=P( B) 。

如果检验的值，越趋近于0，则，越服从P(AB)=P(A)P(B)这个概率。

想想K维的卡方检验。

参见《信息检索导论》p192