**[对信息熵与信息增益的初步理解](http://blog.csdn.net/minenki/article/details/8149567)**

=================

**一 信息熵**

1948年香农提出了信息熵（Entropy）的概念。

假如事件A的分类划分是（A1,A2,...,An），每部分发生的**概率**是(p1,p2,...,pn)，那信息熵定义为公式如下：

http://img.my.csdn.net/uploads/201211/05/1352105250_1611.jpg

吴军在《[数学之美系列四--怎样度量信息？](http://www.kuqin.com/math/20071204/2778.html" \t "_blank)》中认为**信息熵的大小指的是了解一件事情所需要付出的信息量是多少**，这件事的不确定性越大，要搞清它所需要的信息量也就越大，也就是它的信息熵越大。

Matrix67在《[互联网时代的社会语言学：基于SNS的文本数据挖掘](http://www.matrix67.com/blog/archives/5044" \t "_blank)》认为**信息熵衡量的是在你知道一个事件的结果后平均会给你带来多大的信息量**。如果一颗骰子的六个面都是 1 ，投掷它不会给你带来任何新信息，因为你知道它的结果肯定是1，它的信息熵为 - log(1) = 0 。（log是以2为底，lg是以10为底）

两种解释在不同的应用上可以有不同的理解。例如在《[文本分类入门（十一）特征选择方法之信息增益](http://www.blogjava.net/zhenandaci/archive/2009/03/24/261701.html)》可以看出在文本分类中对信息熵的理解是第二种。不过，两种理解其实都指出了信息熵的另一个作用，就是信息熵可以衡量事物的不确定性，这个事物**不确定性越大，信息熵也越大**。

为何信息熵会有这样的作用？为何它的公式这样表示？所以还需要理解**信息量**这一的概念。香农用“比特”这个概念来度量信息量。也即**信息量的多少就是比特的多少**。

拿吴军举的例子来讲，32只球队共有32种夺冠的可能性，用多少信息量才能包括这32个结果？按照计算机的二进制（只有1和0）表示法，我们知道2^5=32 ,也就是需要5符号的组合结果就可以完全表示这32个变化，而这里的符号通常称之为比特。既然是这样，那么当一件事的结果越不确定时，也就是变化情况越多时，那么你若想涵盖所有结果，所需要的比特就要越多，也就是，你要付出的信息量越大，也即信息熵越大。当然，每个变化出现的概率不同，因而在香农的公式中才会用概率，所以**信息熵算的是了解这件事所付出的平均信息量。**比如这个例子里假设32只球队夺冠可能性相同，即Pi=1/32 ，那么按照香农公式计算：

entropy(P1,P2,...,P32)=-(1/32)log(1/32)-(1/32)log(1/32)......-(1/32)log(1/32)

                                  =5/32+5/32...+5/32

                                  =(5\*32)/32

                                  =5

**二 信息增益（又称相对熵[relative entropy]、k-l散度[Kullback–Leibler divergence]、信息散度[information divergence]）**

在概率论和信息论中，信息增益是非对称的，用以度量两种概率分布P和Q的差异。信息增益描述了当使用Q进行编码时，再使用P进行编码的差异。通常P代表样本或观察值的分布，也有可能是精确计算的理论分布。Q代表一种理论，模型，描述或者对P的近似。

详细的应用实例可见这篇文章《[信息增益](http://blog.sina.com.cn/s/blog_67bc5aa60100qays.html)》和《[归纳决策树ID3（](http://www.cnblogs.com/zhangchaoyang/articles/2196631.html)[**Java**](http://lib.csdn.net/base/17)实现）》

信息增益的理解：

熵：表示随机变量的不确定性。  
  
条件熵：在一个条件下，随机变量的不确定性。  
  
信息增益：熵 - 条件熵  
  
在一个条件下，信息不确定性减少的程度！  
  
通俗地讲，X(明天下雨)是一个随机变量，X的熵可以算出来， Y(明天阴天)也是随机变量，在阴天情况下下雨的信息熵我们如果也知道的话（此处需要知道其联合概率分布或是通过数据估计）即是条件熵。  
  
两者相减就是信息增益！原来明天下雨例如信息熵是2，条件熵是0.01（因为如果是阴天就下雨的概率很大，信息就少了），这样相减后为1.99，在获得阴天这个信息后，下雨信息不确定性减少了1.99！是很多的！所以信息增益大！也就是说，阴天这个信息对下雨来说是很重要的！  
  
所以在特征选择的时候常常用信息增益，如果IG（信息增益大）的话那么这个特征对于分类来说很关键~~ 决策树就是这样来找特征的！