参考 : https://blog.csdn.net/MY578719543/article/details/88123667

## 决策树

决策树是一种基本的分类与回归方法，它是利用训练数据，根据损失最小化的原则建立决策树模型。决策树通常包括3个方法：特征的选择，决策树的生成，以及决策树的修建。

决策树的特征选择是基于信息增益，或信息增益比，或基尼系数对特征进行分类的，因此分别对应于ID3算法，C4.5算法，以及CART算法。

信息熵的定义

1. 球队获胜信息的价值值多少钱的问题，32支球队，按照对半猜分的思想，每次选出一半作为猜想的假设，通过5次猜想就能知道最终哪支球队获胜。那么这个信息的价值就是5比特！

2. 同样的对于64支球队的话，它的冠军球队获胜的信息价值就是6比特！

3. 我们可以发现，信息的比特数是与log函数相关的。5 = log32 6 = log64

4. 但是对于球队来说，又存在一个各个球队获胜的概率是不一样的，那么我们将选出那些概率较大的队伍进行最终冠军球队的预测！这样的话，最终的信息量就会降低，

5. 那么对于一件事的信息的价值是多少，我们可以通过它的期望值来评判：

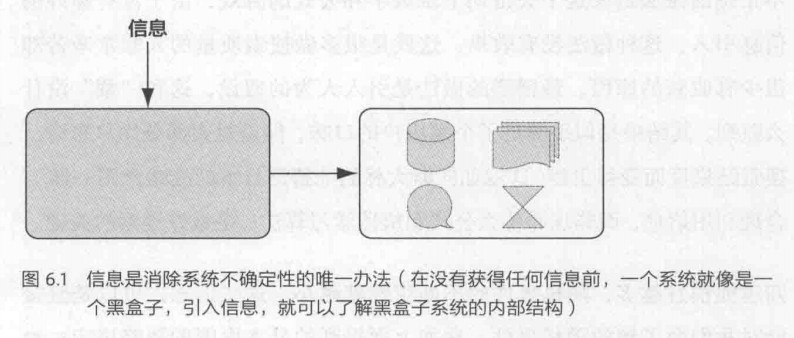
对于32支球队，当它们获胜的概率都是相同的时候，那么它们的信息熵就可以算出来就是5。

6. 对于熵而言，变量的不确定性越大，熵也就越大，要想把一件事弄清楚，所需要的信息熵也就越大！

7. 那么既然提出了熵，常用的汉子有7000字，加入每个字等概率的话，那么大约需要13比特左右的信息，2^13=8000多，2^12=4000多。但是对于每个字使用到的概率是不相同的，以及上下文的关系，因此，信息量就会降低，以至于每个汉字的信息熵只有5Bit左右。对于一本50万字的中文书平均有多少信息量，信息量大约是250W比特。采用较好的压缩算法进行压缩，那么整本书可以存成一个320KB的文件，如果按照两字节的国际编码存储这本书，大约需要1MB的大小。而320KB和1MB之间的差值大小，就被称为了冗余度！

8. 再举个例子，不同语言之间的冗余度是不同的，对于英文书翻译成中文书而言，此时的中译本要薄的的多，这就体现了冗余度之间的差异！

9. 1比特的信息低过，千军万马！苏联和日本的故事，1比特的消息指令，做出正确的战略方针！

10. 

11. 对于日本做出的战略决策具有不确定性U， 当我们引入了信息I之后，U的不确定性就会降低，这正是信息发挥了作用。另外我们进一步把它归纳到自然语言处理、信息与信号处理的应用都将是一个消除不确定的过程！

12. 网页搜索的概念就是搜索最佳的匹配网页，来消除相应的不确定性，找到与搜索关键词最相关的那个网页就是我们最终的目的!

13. 知道的信息越多，随机事件的不确定性也就越小，这就是信息！

14. 条件熵，我们知道了信息Y取不同值的前提下X的概率分布，在数学书称为了条件概率分布，定义在Y的条件下的条件熵为：



此时，在已知了Y的前提下，关于X的不确定性就降低了。上述模型成为2元模型，另外的H（X|Y，Z）的值要比二元模型H（X|Y）小，也就是说不确定性就提高了。H（X|Y，Z）小于等于H（X|Y），当等号成立时，此时增量的信息Z对X，是没有影响的，说明它们毫无关系！

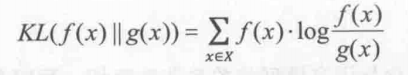
15 信息的作用就是在于消除不确定性，自然语言处理的大量问题就是寻找相关的信息！

16. 互信息，是指两个事件之前的相关性，比如，一个随机事件“北京下雨”与过去的另一个变量“北京过去24小时的湿度”是相关的，对于成都下雨与北京过去24小时的湿度是不太相关的。因此提出了互信息（Mutual information）来描述两个随机变量之间的相关性！



将互信息的定义与二元模型的条件熵进行对比发现，说明了这个互信息的概念就是 随机事件X的不确定性在知道随机事件Y条件下的不确定性！或者说是条件熵之间的差异！即I(X，Y) = H（X）— H（X,Y），就是在了解其中一个Y的前提下，对消除另一个X不确定性所提供的信息量。对于互信息，当两个随机事件完全相关时，它的取值为1，当两者不相关时，它的取值为0！

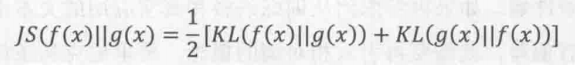
17. 相对熵用来衡量两个取值为正数的函数间的相关性。



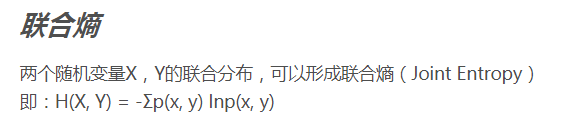
相对熵，是不对称的。



一种新的相对熵的计算方法，就可以实现对称算法！



最早的，相对熵用来描述两个随机信号的相关性，如果他们的相对熵越小，说明这两个随机信号之间越接近，否则信号的差异就越大！说白了，就是衡量两段信息的相似程度！



信息增益: 在概率中定义为：待分类的集合的熵和选定某个特征的条件熵之差（这里只的是经验熵或经验条件熵，由于真正的熵并不知道，是根据样本计算出来的） 信息增益在决策树算法中是用来选择特征的指标，信息增益越大，则这个特征的选择性越好。

https://img-blog.csdnimg.cn/20190305200806456.png

基尼不纯度: 从一个数据集中随机选取子项，度量其被错误的划分到其他组里的概率。

https://img-blog.csdnimg.cn/20190305201332128.png

结论：基尼不纯度可以作为 衡量系统混乱程度的 标准；基尼不纯度越小，纯度越高， 集合的有序程度越高，分类的效果越好；

基尼不纯度为 0 时，表示集合类别一致；

在决策树中，比较基尼不纯度的大小可以选择更好的决策条件（子节点）。

