决策树:

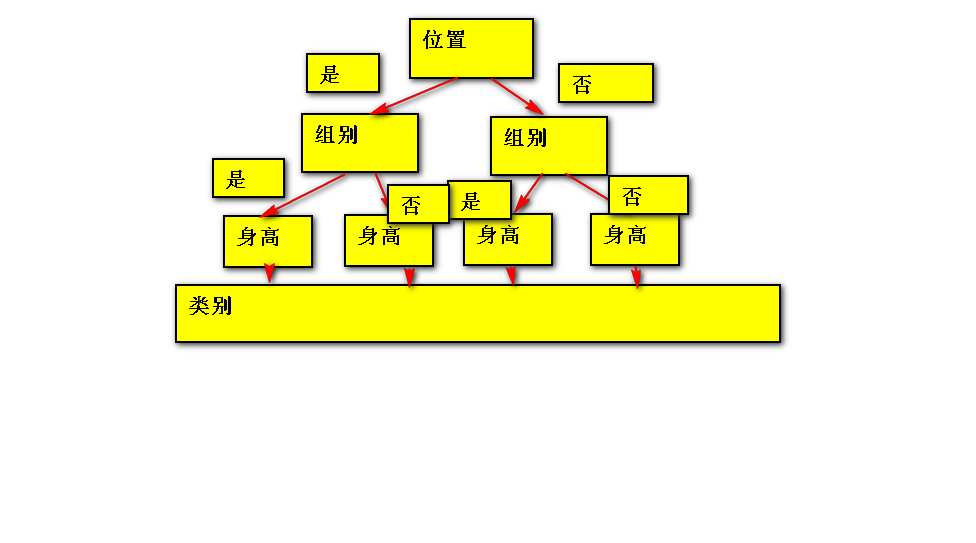
栗子：

组别 身高 性别 行数 列数 位置 结果

1. 是不是第五组？ 是
2. 是不是在175以上 是
3. 是不是女生 不是
4. 是不是高晨晨
5. 是不是第一组？ 不是
6. 是不是第二组？不是
7. 是不是第四组？不是
8. 是不是女生？ 是
9. 是不是高丹 是

组别 身高 性别 行数 列数 位置 结果

1. 是不是男生？ 不是
2. 身高是否大于160 ？ 是
3. 是不是最后一排 ？是
4. 是不是贺春梅？



计算系统的最大信息增益所对应的一组最优的特征组合，再根据最优特征组合来构建一棵树 ，然后使用决策树做分类

如何找最优的特征组合？

在信息论中

信息熵(香农熵)：指的是系统的混乱程度

一盒火柴有序的排列，信息熵就低 ，如果火柴是随机摆放，那么信息熵比较高

信息增益：系统经过变化以后的信息熵的变化 如果信息熵变小了，则信息增益变大，反之亦然

ID3算法

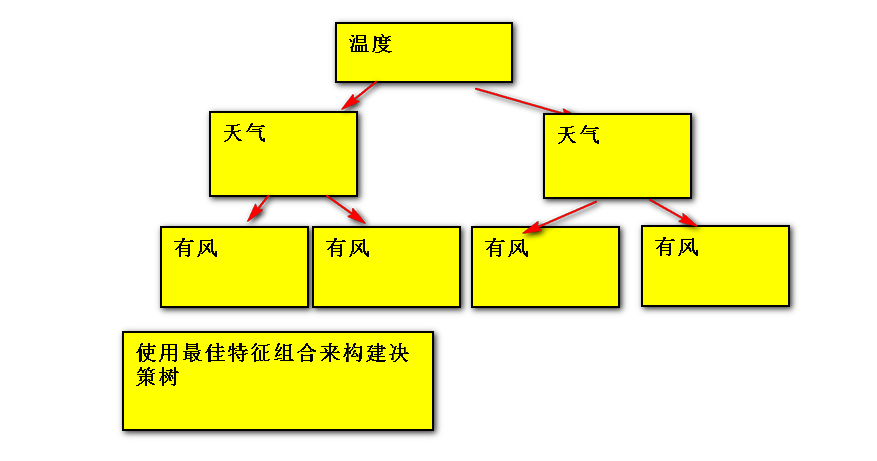
E= pi分到某个类别的概率

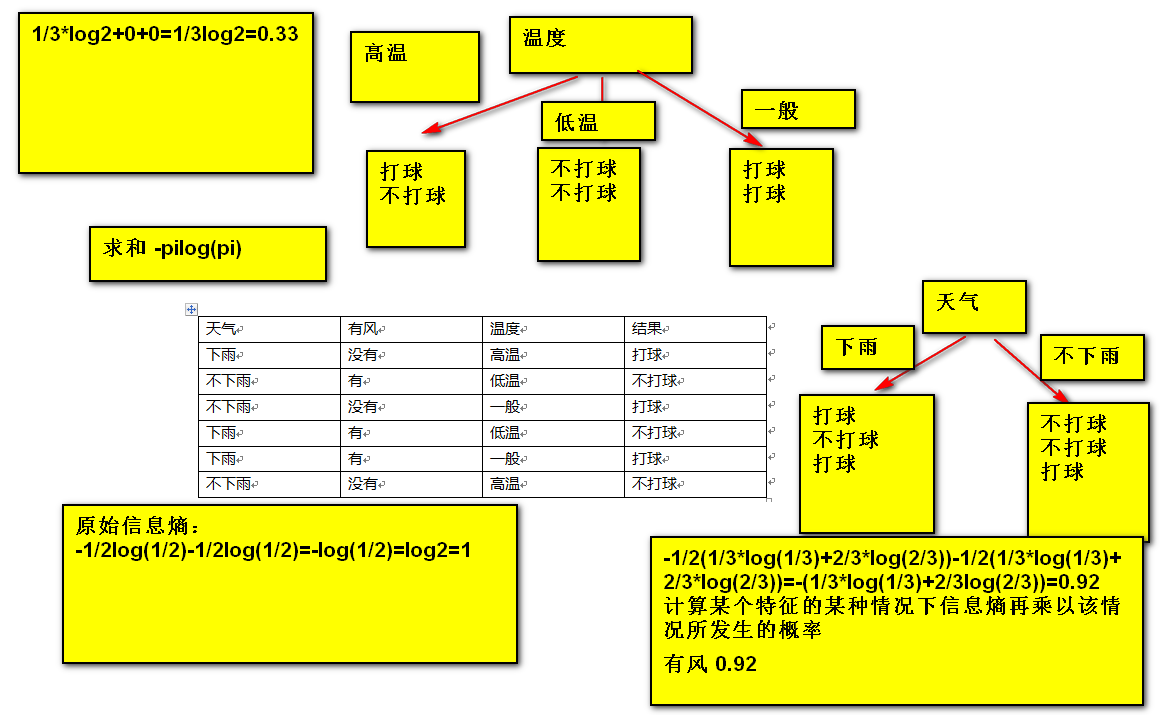
栗子：

|  |  |  |  |
| --- | --- | --- | --- |
| 天气 | 有风 | 温度 | 结果 |
| 下雨 | 没有 | 高温 | 打球 |
| 不下雨 | 有 | 低温 | 不打球 |
| 不下雨 | 没有 | 一般 | 打球 |
| 下雨 | 有 | 低温 | 不打球 |
| 下雨 | 有 | 一般 | 打球 |
| 不下雨 | 没有 | 高温 | 不打球 |

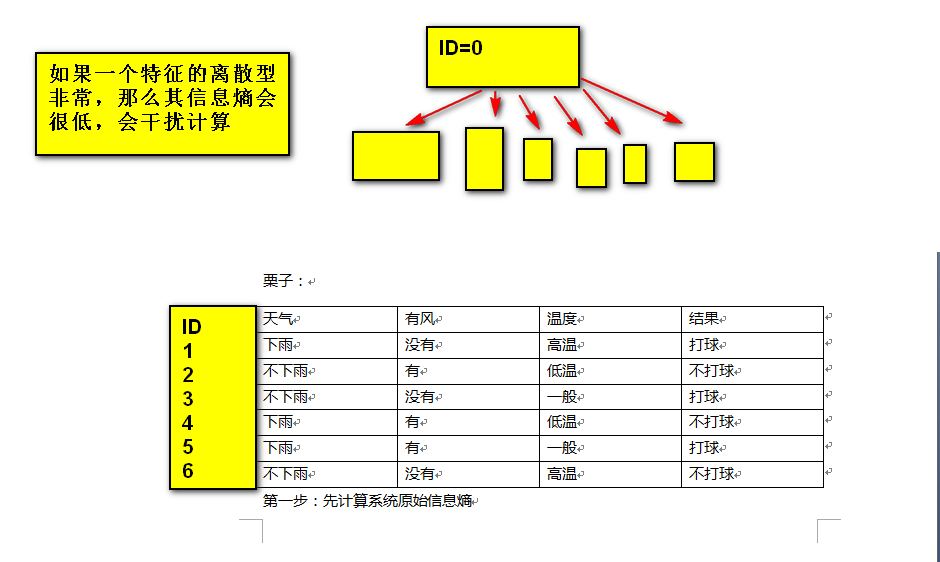
第一步：先计算系统原始信息熵

第二步：分别计算每个特征的信息熵 找到使信息增益最大的组合





ID算法所存在的问题：



C4.5算法

信息增益率：系统信息增益（原始信息熵-当前信息熵）/自身的信息熵

信息增益率越大越好

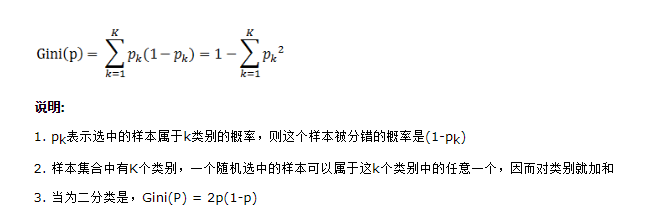
系统信息熵：系统当前信息熵

自身的信息熵：-(1/2\*log(1/2)+1/2log(1/2))

ID: -(1/6\*log(1/6)\*6)=log6

CART算法

计算gini系数:



Gini系数越小，特征组合越好

Sk-learn开发：略

集成算法：

随机森林：

多棵决策树一起做预测，把多棵树的预测结果放到一起选举，少数服从多数，最终输出的结果就是随机森林的结果

Adaboost算法：

将多个弱训练器串行做预测，通过加权强迫后续的训练器更加关注前一个训练器预测错误的数据，以提高整体的预测准确率

朴素贝叶斯算法：

朴素：特征之间的独立性比较强

用贝叶斯公式进行计算在确定类别的情况下，发生某个特征组合的概率，把概率最大的作为最后的类别

全概率公式：P(B)=P(B|A1)+P(B|A2)+..+P(B|An)

贝叶斯公式：

P(A)=P(B)\*P(A|B)/P(B|A)

P(B|A)=P(B)\*P(A|B)/P(A) A=(不帅，不好，矮，不上进) B=嫁



P(嫁|（不帅，不好，矮，不上进）)与P(不嫁|（不帅，不好，矮，不上进）)

P(嫁|（不帅，不好，矮，不上进）)=P(嫁)\*P((不帅，不好，矮，不上进)|嫁)/P(不帅，不好，矮，不上进)

=P(嫁)\*P(不帅|嫁)\*P(不好|嫁)\*P(矮|嫁)\*P(不上进|嫁)/P（（不帅，不好，矮，不上进）|嫁）+P(（不帅，不好，矮，不上进）|不嫁)

= P(嫁)\*P(不帅|嫁)\*P(不好|嫁)\*P(矮|嫁)\*P(不上进|嫁)/ P(不帅|嫁)\*P(不好|嫁)\*P(矮|嫁)\*P(不上进|嫁)/+ P(不帅|不嫁)\*P(不好|不嫁)\*P(矮|不嫁)\*P(不上进|不嫁)

=（1/2\*1/2\*1/6\*1/6\*1/6 ）/(1/2\*1/6\*1/6\*1/6+1/3\*1/2\*1\*2/3)= 1.0204081632653062%

P(不嫁|（不帅，不好，矮，不上进）)=P(不嫁)\*P((不帅，不好，矮，不上进)|不嫁)/P(不帅，不好，矮，不上进)

=P(不嫁)\*P(不帅|不嫁)\*P(不好|不嫁)\*P(矮|不嫁)\*P(不上进|不嫁)/P（（不帅，不好，矮，不上进）|嫁）+P(（不帅，不好，矮，不上进）|不嫁)

= P(不嫁)\*P(不帅|不嫁)\*P(不好|不嫁)\*P(矮|不嫁)\*P(不上进|不嫁)/ (P(不帅|嫁)\*P(不好|嫁)\*P(矮|嫁)\*P(不上进|嫁)/+ P(不帅|不嫁)\*P(不好|不嫁)\*P(矮|不嫁)\*P(不上进|不嫁))

=(1/2\*1/3\*1/2\*2/3\*1)/( 1/2\*1/6\*1/6\*1/6+1/3\*1/2\*1\*2/3)= 48.97959183673469%

因为P(不嫁|（不帅，不好，矮，不上进）)的概率大于P(嫁|（不帅，不好，矮，不上进）)的概率，所以不嫁给你

先验概率与后验概率

一个班级里面男生90% 10% 男生都穿裤子 女生一半穿裤子，一半穿裙子

先验概率： 随机选一个学生，该学生穿裤子的概率

后验概率：随机选一个学生，已知该学生穿裤子，求该学生是女生的概率

先：由因求果

后：由果求因

贝叶斯公式是由先验概率求后验概率

用朴素贝叶斯解决真实问题

垃圾邮件过滤

分析思路：

预处理数据

1. 垃圾邮件和正常邮件的内容有区别
2. 每个邮件中出现的单词次数不一样
3. 根据所有邮件中出现的前3000个常见单词构建特征矩阵
4. 样本是邮件 特征值是单词在邮件中出现的次数
5. 构建目标变量 选择模型进行训练
6. 评估模型
7. 如果模型的效果不好—优化：增大数据集 调整超参数 换模型 特征选择 检查代码逻辑

https://blog.csdn.net/liufang0001/article/details/77856255/