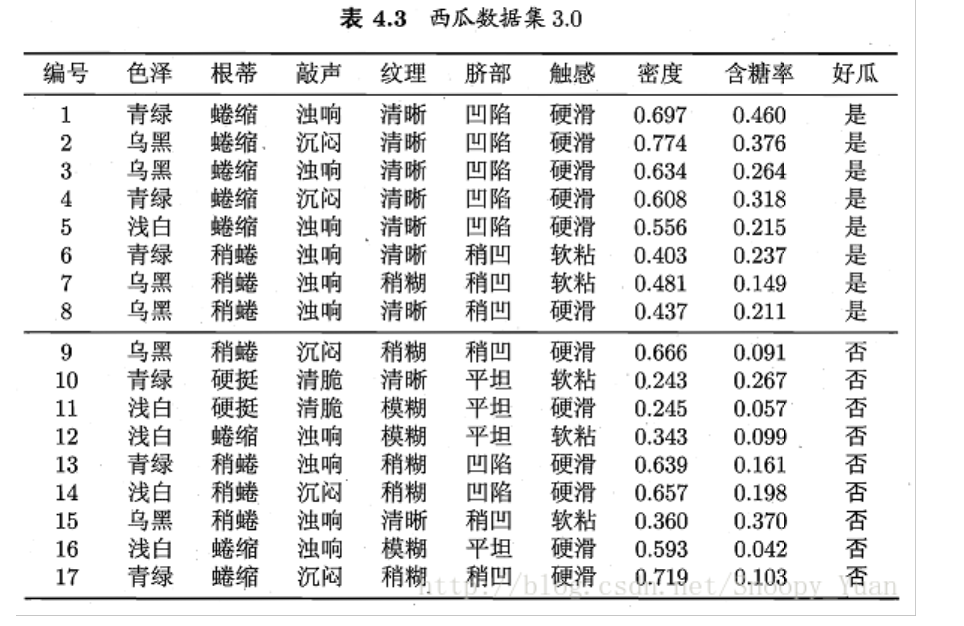


参考链接：<https://blog.csdn.net/zealfory/article/details/82866491>

参考数据：



个人理解：

对于输入的训练集D和属性集A

（集为对于训练集D中的每条数据，都存在着A中的属性），伪算法描述中用C来表示类别

决策树的创建算法分为三种情形：  
1.如果D中的训练集样本都属于同一类别C

因此可以直接将node标记未C类叶结点。

如：上述17条数据若全为好瓜，则可直接将此node结点标记为好瓜

2.如果A中已经没有可再划分的属性集，或者，D样本在属性集A上的类别取值相同

（此处可以理解为，此时A的属性取值已经影响不到D，即D类别取值相同）

将Node结点标记为此时D中样本数最多的类

此处注意：对于“样本数最多的类”的理解

由于第二种情况中包含两个分支：  
 ①如果A中已经没有可再划分的属性集：

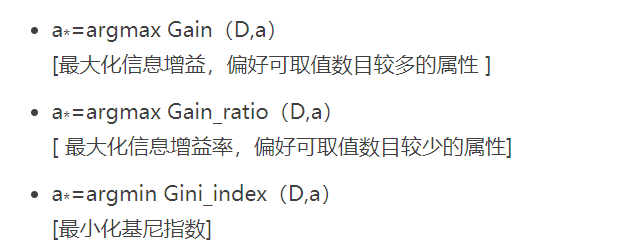
则此时Node结点标记为D中样本数最多的那一类

②D样本在属性集A上的类别取值相同

则此时Node标记为此时的那个类别取值（由于相同，因此所有的样本都是为一个类别）

综上所述，因此可表述为“样本数最多的类”

3.如果都不是上述两种情况，则就是需要进行属性划分：  
 首先选取最优属性，有三种方法：



排除了上面情况1，2后，若根据最优属性划分后样本集为空，因此可以选择此时样本中最多的类作为标记；不为空，则可进行递归。

