1.[决策树] 基于信息熵，对西瓜数据集进行决策树构建，描述过程

西瓜有色泽，根蒂，敲声，纹理，脐部，触感6个特征

① 数据集包含17个训练样本，正例8，反例9， 计算得到根节点的信息熵：

$Ent(D) = -\frac{8}{17}log_2(\frac{8}{17}) - \frac{9}{17}log_2(\frac{0}{17}) = 0.998$

对于色泽，青绿(6， 正例3), 乌黑(6， 正例4), 浅白(5，正例1),信息熵为

$-\frac{6}{17}(\frac{3}{6}log_2(\frac{3}{6}) + \frac{3}{6}log_2(\frac{3}{6})) - \frac{6}{17}(\frac{4}{6}log_2(\frac{4}{6}) + \frac{2}{6}log_2(\frac{2}{6})) - \frac{5}{17}(\frac{1}{5}log_2(\frac{1}{5}) + \frac{4}{5}log_2(\frac{4}{5})) = 0.889$

信息增益为0.109

对于根蒂，蜷曲(8， 正例5), 稍蜷(7， 正例4), 浅白(2，正例0),信息熵为：

$-\frac{8}{17}(\frac{5}{8}log_2(\frac{5}{8}) + \frac{3}{8}log_2(\frac{3}{8})) - \frac{7}{17}(\frac{4}{7}log_2(\frac{4}{7}) + \frac{3}{7}log_2(\frac{3}{7})) - \frac{2}{17}(\frac{2}{2}log_2(\frac{2}{2}) + = 0.855$

信息增益为0143

类比可以得到触感的信息增益为0.006， 敲声为-0.141， 脐部为-0.289， 纹理为0.381, 所以根节点**选取纹理**

②纹理清晰有9个清晰(7个正例)， 5个烧糊(1个正例)， 3个模糊(全是反例)，则**模糊不需要继续分类了**

对于3个稍糊，根蒂全是稍蜷，色泽对应于3个青绿(2反例)浅白(1反例)乌黑(1正例1反例)，敲声对应于浊响(1反例1正例)沉闷(3反例)， 脐部对应于稍凹(1正例2反例)凹陷(2反例)， 触感对应于软粘(1正例)硬滑(4反例)，由于触感对应的信息熵为0，对应于信息增量肯定是最多的，所以**稍糊选取触感分类**

对于条理清晰9个(7个正例)：

色泽：青绿4(正例3)， 乌黑4(正例3)， 浅白1(正例1)， 根据上面计算公式得到信息熵为0.361

根蒂：蜷缩5(正例5)， 稍蜷3(正例2)， 硬挺1(反例1)，信息熵为0.306

敲声：浊响6(正例5)， 沉闷2(正例2)， 清脆1(反例1)，信息熵为0.433

脐部：凹陷5(正例5)， 稍凹3(正例2)， 平坦1(反例1)，信息熵为0.306

触感：硬滑6(正例6)， 软粘3(正例1) ，信息熵为0.306

所以可以选取根蒂，脐部，触感都可以获得最大的信息增益，不妨如图对于**条理清晰的选取根蒂**

③根蒂蜷缩5(正例5)， 稍蜷3(正例2)， 硬挺1(反例1)， **蜷缩和硬挺都不需要继续分类了**

对于稍蜷3个(2正例)

色泽：青绿1(正例1)， 乌黑2(正例1)， 信息熵为0.667

敲声：浊响3(正例2)， 信息熵为0.918

脐部：稍凹2(正例1)， 平坦1(反例1)，信息熵为0.667

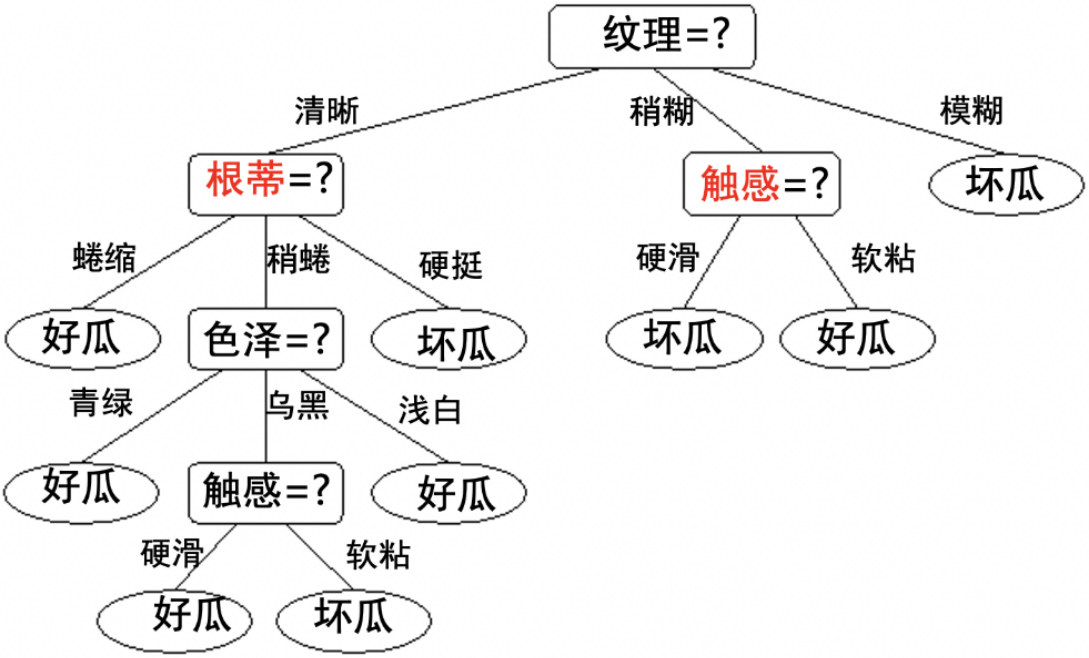
触感：软粘3(正例1)，信息熵为0.918

所以选取色泽和脐部都可以获得最大的信息增益，不妨如图对于**稍蜷的选择色泽**

④色泽的青绿1(正例1)， 乌黑2(正例1)，**色泽青绿的不需要继续分类**

对于乌黑2个(正例1)， 敲声是浊响1正1负，脐部是稍凹1正1负，触感为硬滑1正例软粘1反例，所以可以**使用触感继续分类**

通过以上的计算和分析，可以得到以下PPT所示的决策树：



2.[线性分类] 推导下述logit function和logistic function等价：

$$P(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} = 1 - \frac{1}{1+e^{\beta_0+\beta_1 X}}$$

$$1 - P(X) = \frac{1}{1+e^{\beta_0+\beta_1 X}}$$

$$\frac{P(X)}{1-P(X)} = \frac{(1+e^{\beta_0+\beta_1 X)(e^{\beta_0+\beta_1 X})}}{1+e^{\beta_0+\beta_1 X}} = e^{\beta_0+\beta_1 X}$$

由于每一步都是等价变换，所以logit function和logistic function是等价的