决策树

2020年4月24日 星期五　下午7:25

决策树构建伪算法

输入：$D=\{(x_1,y_1),(x_2,y_2)\cdots(x_m,y_m)\}$

属性集：$A=\{a_1,a_2\cdots a_d\}$

过程：函数 TreeNode(D, A)

1. node;
2. if D中样本属于同一类C
→ node 标记为C：return，endif
3. if $A=\emptyset$ or 样本在A上取值相同
→ node 标为叶结点，类标为样本中最多的类
return，endif
4. 从A中选择最优划分属性a_*；
for a_*中每一个值a_*^v do
为 node 生成分支，D_v表示D在a_*上取a_*^v的样
if $D_v=\emptyset$ →分支结点标为最多的类 return
else：TreeGenerate(D_v, $A\setminus\{a_*\}$)再生成结点
endif，end for
output：node为根结点的决策树

剪枝处理：过拟合

信息熵：$Ent(D)=-\sum_{k=1}^{|y|}p_k\cdot\log_2 p_k$　Ent(D)，D纯度越高

信息增益：$Gain(D,a)=Ent(D)-\sum_{v=1}^{V}\frac{|D^v|}{|D|}Ent(D^v)$

（越大，纯度越高）　　概率：第v个分支点取值为a^v的样本D^v

$a^*=\arg\max_{a\in A}Gain(D,a)$

以西瓜数据为例：　西瓜：好、坏 两种，为测试 8，9个

$Ent(D)=-\sum_{k=1}^{2}p_k\cdot\log_2 p_k=-(\frac{8}{17}\cdot\log_2\frac{8}{17}+\frac{9}{17}\log_2\frac{9}{17})=0.998$

色泽增益：三类 信息 $0\log_2 0=0$

$Ent(D^1)=1.000$，$Ent(D^2)=0.918$，$Ent(D^3)=0.722$

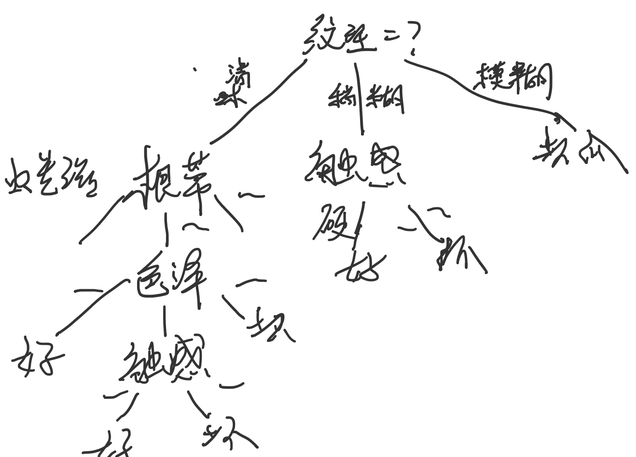
$Gain(D,色泽)=Ent(D)-\sum_{v=1}^{3}\frac{|D^v|}{|D|}Ent(D^v)$

$=0.998-(\frac{6}{17}\times1.0+\frac{6}{17}\times0.918+\frac{5}{17}\times0.722)=0.109$

同理

$Gain(D,根蒂)=0.143$ ... $Gain(D,纹理)=0.38$ ... $Gain(D,触感)=0.006$

纹理 Gain(D) max → 划分



增益率：$Gain\_ratio(D,a)=\frac{Gain(D,a)}{IV(a)}$

$IV(a)=-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\log_2\frac{|D^v|}{|D|}$

基尼指数 CART