第14章 决策树

北京交通大学《机器学习》课程组

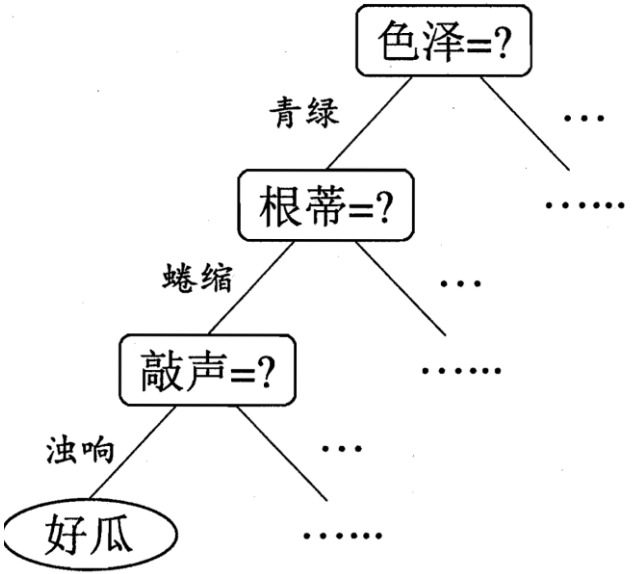
分而治之。

——俞樾《群经平议 周官二》

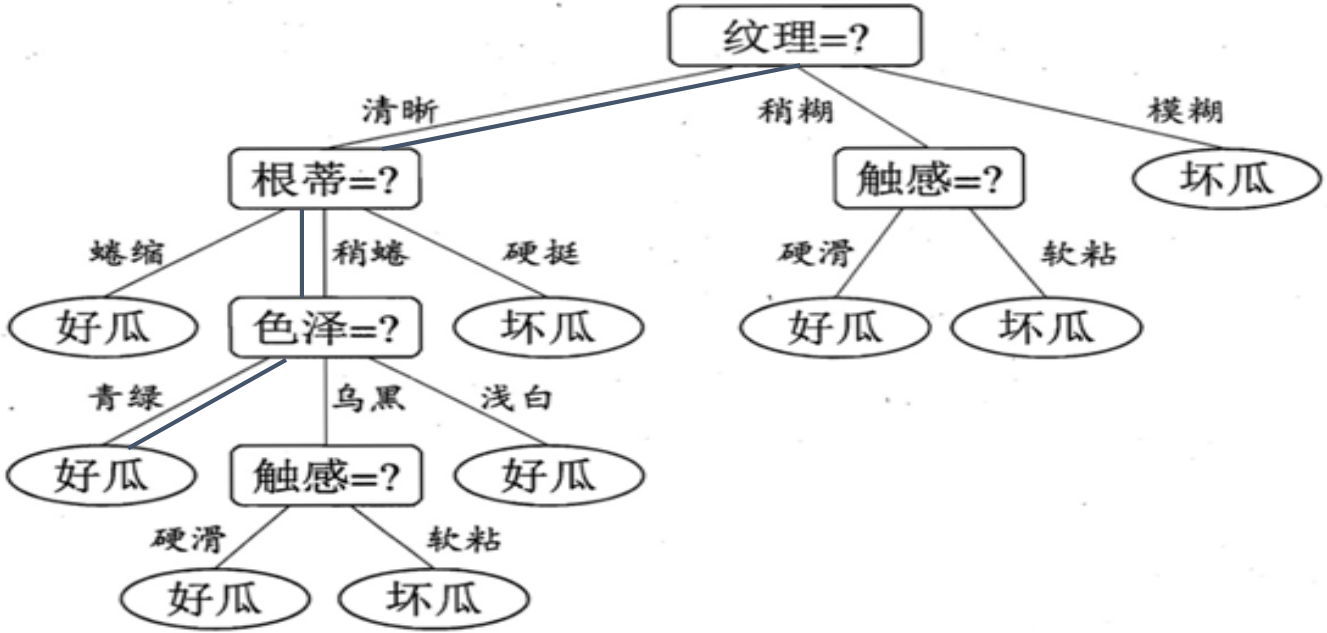
目录

决策树

- 决策树是基于树结构来进行决策。
- 一棵决策树包含一个根结点、若干个内部结点和若干个叶结点。
- 叶结点对应于决策结果，其他每个结点则对应于一个属性测试；
- 每个结点包含的样本集合根据属性测试的结果被划分到子结点中；
- 根结点包含样本全集。
- 从根结点到每个结点的路径对应了一个判定测试序列。



决策树



纹理清晰 根蒂稍蜷 色泽青绿 好瓜

决策树

目的：产生一颗泛化能力强的决策树，采用“分而治之”策略。

输入:训练集$D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$；属性集$A = \{a_1, a_2, \cdots, a_N\}$

过程：函数 TreeGenerate(*D*, *A*)

生成结点node；

if *D* 中样本全部属于同一类别 *C* then

将node标记为 *C* 类叶结点；return

end if

所有样本属于同一类：
叶结点无须进一步细分

无属性可用或者所有样本所有属性取值相同：
无法细分

if $A = \emptyset$ OR *D* 中样本在*A*上取值相同 then　利用当前结点的后验分布

将node标记为叶结点，其类别标记为 *D* 中样本数量最多的类；return

end if

决策树

从 A 中选择最优划分属性 a_*；

关键

仍有样本混杂 & 有属性可用
找出最优属性
依照该属性划分结点

for a_* 的每一个值 a_*^v do

为node生成一个分支；令D_v 表示 D 中在 a_* 上取值为a_*^v 的样本子集；

if D_v 为空 then

将父结点的样本分布作为当前结点的先验分布

将分支结点标记为叶结点，其类别标记为 D 中样本最多的类；return

else

表示从A中去除a_*

以TreeGenerate(D_v, $A\setminus\{a_*\}$)为分支结点

end if

end for

输出：以node为根结点的一颗决策树

目录

目录

划分选择—信息增益

决策树学习的关键: 如何选择最优划分属性?

希望节点的“纯度” 越来越高

信息熵

信息熵是度量样本集合纯度最常用的一种指标。假定当样本集D中第k类样本所占的比例为$p_k(k=1,2,\cdots,|y|)$，则D的信息熵定义为

$$\mathrm{Ent}(D)=-\sum_{k=1}^{|y|}p_k log_2\, p_k$$

如果p=0，则$plog_2p=0$

Ent(D)的值越小，则D的纯度越高（类标签越一致）。

划分选择—信息增益

■信息增益

- 假设离散属性a有V个可能的取值$\{a^1, a^2, \ldots, a^V\}$，若使用$a$来对样本集$D$进行划分，则会产生$V$个分支结点，
- 其中第v个分支结点包含了D中所有在属性a上取值为a^v的样本，记为：D^v。
- 考虑到不同的分支结点所包含的样本数不同，给分支结点赋予权重$\frac{|D^v|}{|D|}$，

■信息增益定义为：

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

属性a取值a^v的样本总数

样本总数

■选择信息增益值最大的属性进行划分 ➡ $a_* = \arg\max_{a \in A} \text{Gain}(D,a)$

■信息增益越大，使用属性a进行划分所获得的“纯度提升”越大。

■著名的ID3决策树学习算法(1986)就是以信息增益为准则来选择划分属性。

划分选择—信息增益

■例子

- 一个西瓜数据集
- 17个训练样例，用以学习一棵能预测没刨开的西瓜是不是好瓜的决策树。
- 正例占$\frac{8}{17}$，反例占$\frac{9}{17}$

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

划分选择—信息增益

1. 根结点的信息熵为：

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

2. 计算当前属性集合中每个属性的信息增益， 这里以“色泽”为例：

青绿(3/6好)、乌黑（4/6好）、浅白（1/5好）

$$\mathrm{Ent}(D^1) = -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right) = 1.000$$

$$\mathrm{Ent}(D^2) = -\left(\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right) = 0.918$$

$$\mathrm{Ent}(D^3) = -\left(\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right) = 0.722$$

划分选择—信息增益

属性“色泽”的信息增益为：

$$\begin{aligned}\text{Gain}(D, \text{色泽}) &= \text{Ent}(D) - \sum_{v=1}^{3} \frac{|D^v|}{|D|}\text{Ent}(D^v) \\ &= 0.998 - \left(\frac{6}{17} \times 1.000 + \frac{6}{17} \times 0.918 + \frac{5}{17} \times 0.722\right) \\ &= 0.109\ .\end{aligned}$$

其他属性的信息增益依次为：

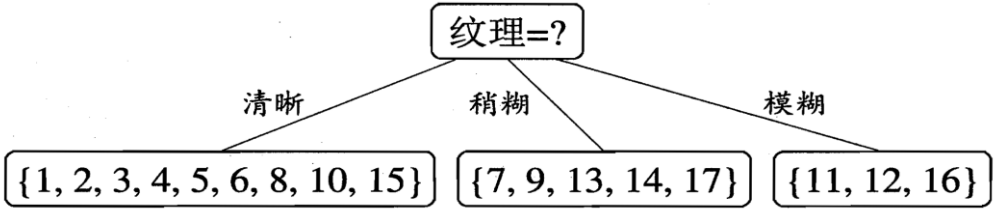
$$\text{Gain}(D, \text{根蒂}) = 0.143; \quad \text{Gain}(D, \text{敲声}) = 0.141;$$

$$\boxed{\text{Gain}(D, \text{纹理}) = 0.381;} \quad \text{Gain}(D, \text{脐部}) = 0.289;$$

$$\text{Gain}(D, \text{触感}) = 0.006.$$

划分选择—信息增益

3. “纹理”被选为划分属性对根结点进行划分：



4. 接着对每个分支结点做进一步划分：

以“纹理=清晰”为例，该结点包含的样例集合D^1中有9个样例，可用属性集合为：{色泽，根蒂，敲声，脐部，触感}，基于D^1计算出各属性的信息增益：

$$\text{Gain}(D^1, \text{色泽}) = 0.043; \quad \text{Gain}(D^1, \text{根蒂}) = 0.458;$$

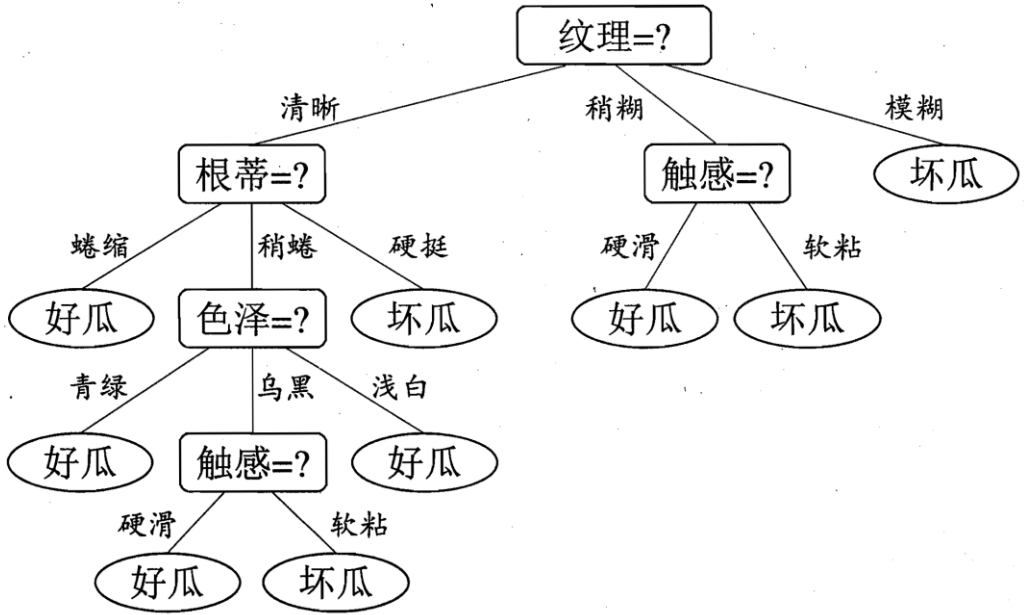
$$\text{Gain}(D^1, \text{敲声}) = 0.331; \quad \text{Gain}(D^1, \text{脐部}) = 0.458;$$

$$\text{Gain}(D^1, \text{触感}) = 0.458.$$

划分选择—信息增益

5. 有3个属性的信息增益最大，任选其一，这里选择“根蒂”

类似地，对每个分支结点进行上述操作，最终的决策树为：



目录

划分选择—信息增益率

信息增益准则对**可取值数目较多的属性有所偏好**，为减少这种偏好带来的不利影响，著名的**C4.5算法（1993）**用“增益率”来选择最优划分属性。

信息增益率

$$\text{Gain\_ratio}(D,a)=\frac{\text{Gain}(D,a)}{\text{IV}(a)}$$

其中IV(a)称为属性a的“固有值”，定义为：

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

与类标签y无关

信息增益率准则对**可取值数目较少的属性有所偏好**，C4.5算法不是直接采用增益率最大的属性划分，而是先从划分属性中**找出信息增益高于平均水平的属性，再从中选择增益率最高的属性**。

划分选择—信息增益率

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

a的可能取值越多，IV(a)的值越大

IV（触感）=0.874（V=2）

IV（色泽）=1.580（V=3）

$$\mathrm{Gain\_ratio}(D, a) = \frac{\mathrm{Gain}(D, a)}{\mathrm{IV}(a)}$$

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

连续值处理

当属性是连续值时，采用连续属性离散化技术，以二分法为例：

- 给定数据集D和连续属性a，假设a在D上出现了n个不同的取值，将这些值从小到大排序后记为$\{a^1, a^2, \ldots, a^n\}$。
- 基于划分点t可将D分为子集D_t^-和D_t^+，D_t^-包含在属性a上取值不大于t的样本，D_t^+则包含大于t的样本。
- 对两个相邻的属性取值a^i和a^{i+1}来说，t在区间$[a^i, a^{i+1})$中取任意值所产生的划分结果相同
- 对于连续属性a，将两个相邻属性取值的中位点作为划分点（共n-1个）

$$T_a = \left\{ \frac{a^i + a^{i+1}}{2} \,|\, 1 \le i \le n-1 \right\}$$

连续值处理

对于连续属性 a，就可以像离散属性一样来考察这些划分，选取最优的划分点来进行样本集合的划分：

$$\mathrm{Gain}(D,a)=\max_{t\in T_a}\mathrm{Gain}(D,a,t)=\max_{t\in T_a}\mathrm{Ent}(D)-\sum_{\lambda\in\{-,+\}}\frac{|D_t^{\lambda}|}{|D|}\mathrm{Ent}(D_t^{\lambda})$$

T_a为对连续属性排序后相邻两个属性取值的中位数

$\mathrm{Gain}(D,a,t)$是样本集 D 基于划分点 t 二分后的信息增益，选择使 $\mathrm{Gain}(D,a,t)$最大化的划分点

连续值处理

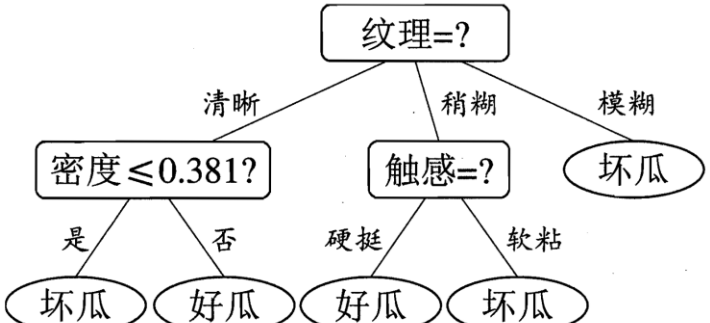
例子：西瓜数据集3.0 增加密度和含糖率两个属性

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

以密度为例，该属性的候选划分点集合包含16个候选值：

$T_{密度}$ ={0.244,0.294,0.351,0.381,0.420,0.459,0.581,0.574,0.600,0.621,0.636,0.648,0.661,0.681,0.708,0.746}，

由上述公式可计算出属性“密度”的信息增益最大为0.262，对应划分点为0.381。



目录

划分选择—基尼指数

CART决策树算法(1984)使用“基尼指数”来选择划分属性。

基尼值　Gini(D)反映从数据集D中随机抽取两个样本，其类别标记不一致的概率

$$\text{Gini}(D) = \sum_{k=1}^{|y|} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{|y|} {p_k}^2$$

样本集D中第k类样本所占的比例为p_k

样本类别属性清晰p_1=1，p_2=0，Gini(D)=0
样本类别属性清晰p_1=p_2=1/2，Gini(D)=1/2

Gini(D)越小，数据集D纯度越高

基尼指数

$$\text{Gini\_index}(D, a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} \text{Gini}(D^v)$$

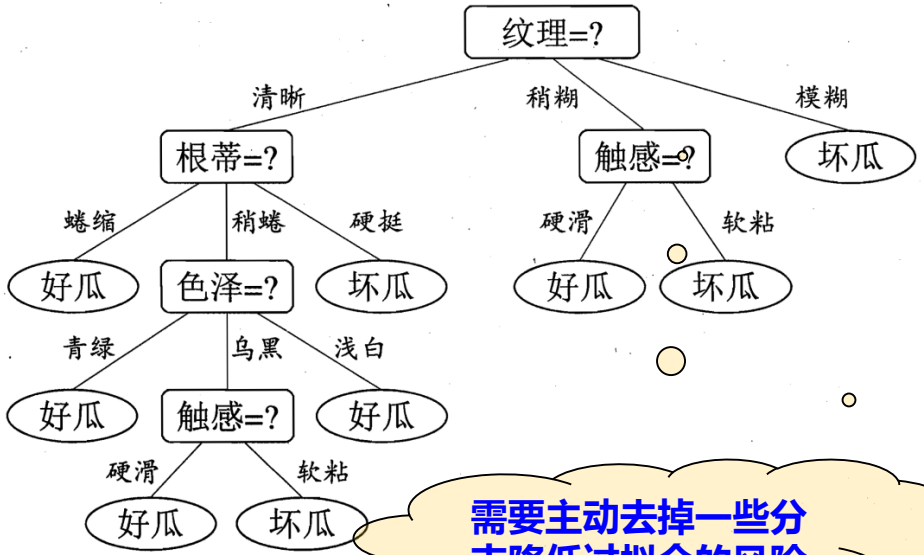
选择基尼指数最小的属性进行划分 → $a_* = \arg\min_{a \in A} \text{Gini\_index}(D, a)$

目录

剪枝处理

■剪枝是决策树学习算法对付“过拟合”的主要手段。

■所有训练样本都被正确分类，造成决策树分支过多，可能存在过拟合。



需要主动去掉一些分支降低过拟合的风险

剪枝处理

■剪枝的基本策略有“预剪枝”和“后剪枝”。

预剪枝：在决策树生成过程中，对每个结点在划分前先进行估计，若当前结点的划分不能带来决策树泛化性能提升，则停止划分并将当前结点标记为叶结点。

判断泛化性能是否提升
- 留出法
- 交叉验证法
- 自助法

这里以“留出法”为例，即：预留一部分数据用作“验证集”以进行性能评估。

剪枝处理—预剪枝

例子：西瓜数据集2.0

训练集

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

验证集

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否

剪枝处理—预剪枝

基于训练集，根据信息增益准则，选择“脐部”对训练集进行划分，并产生3个分支。然而，**是否应该进行这个划分呢？对划分前后的泛化性能进行估计。**

剪枝处理—预剪枝

如果不进行划分：

样例数一样多时，任选一类

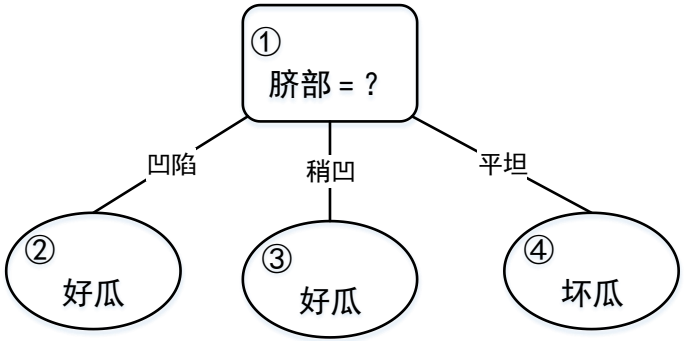
- 划分前所有样例集中在根结点，将该结点标记为叶结点，其类别标记为训练样例数最多的类别，假设将这个叶结点标记为“好瓜”。
- 用验证集对该单结点决策树进行评估，编号{4,5,8}的样例被分类正确，编号{9,11,12,13}的样例分类错误，**验证集精度**为：$3/7 \times 100\% = 42.9\%$。

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否

剪枝处理—预剪枝

■如果进行划分：

- 在用属性“脐部”划分之后，图中的结点②、③、④分别包含编号为：{1,2,3,14}、{6,7,15，17}、{10,16}的训练样例，这三个结点分别被标记为叶结点“好瓜”、“好瓜”、“坏瓜”
- 用验证集对该单结点决策树进行评估，编号{4,5,8,11,12}的样例被分类正确，编号{9,13}的样例分类错误，验证集精度为：5/7×100% = 71.4% > 42.9%，
- 应该选择用“脐部”进行划分。



编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否

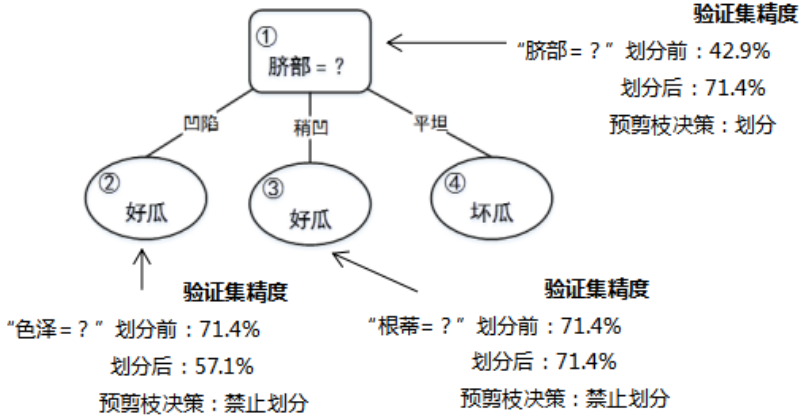
剪枝处理—预剪枝

“凹陷”结点要不要划分?

【不划分】5/7，71.4%样本正确分类

【划分】“凹陷青绿”为好瓜，“凹陷乌黑”为好瓜，“凹陷浅白”为坏瓜。应用于验证集4/7=57.1%样本正确分类

决定不划分



训练集

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

验证集

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否

剪枝处理—预剪枝

预剪枝总结：

优点

使得决策树很多节点都没有“展开”，这不仅降低了过拟合的风险，还显著减少了决策树的训练时间开销和测试时间开销。

缺点

有些分支的当前划分虽然不能提升泛化性，甚至导致泛化性能暂时下降，但在其基础上进行的后续划分却有可能导致性能的显著提高，因此剪枝后存在欠拟合的风险。

剪枝处理—后剪枝

■**后剪枝：**

- 先从训练集生成一颗完整的决策树，然后自底向上地对非叶结点进行考察
- 若将该结点对应的子树替换为叶结点能带来泛化性能的提升，则将孩子树替换为叶结点。

训练集

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

验证集

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否

剪枝处理—后剪枝

例子：先从训练集生成一颗完整的决策树，如右图。该决策树的验证集中编号为{4,11,12}的样例被分类正确，编号为{5,8,9,13}的样例被分类错误，所以该决策树的验证集精度为：3/7 ×100% = 42.9%。

剪枝处理—后剪枝

例子：先考察结点⑥，若将其分支剪除，则相当于把⑥替换为叶结点。替换后的叶结点包含编号为{7,15}的样例，将该叶结点的类别标记为“好瓜”，此时：验证集中编号为{4,8,11,12}的样例被分类正确，验证集精度为 $4/7 \times 100\% = 57.1\%$ ，于是决定剪枝。

剪枝处理—后剪枝

后剪枝总结：

优点

- 后剪枝的决策树往往比预剪枝的决策树保留了更多的分支。
- 一般地，后剪枝决策树的欠拟合风险减小，泛化性能往往优于预剪枝决策树。

缺点

- 剪枝过程是在生成完全决策树之后进行的，并且要自底向上地对树中的所有非叶结点进行逐一考察，因此其训练时间开销比预剪枝决策树要大得多。

目录

决策树的缺陷

决策树的两大缺陷

1） 决策树会因为异常值的影响导致预测不准确

2） 决策树是采用了所有的特征及样本，容易出现过拟合

可行解决方案：三个臭皮匠，顶个诸葛亮

集成学习

- **集成学习：**将多个 “单个学习器（Individual Learner）” 用某种策略来结合起来，组成一个“学习委员会（committee）”，使得整体的泛化性能得到大大提高。

- **同质集成 vs 异质集成：**如果所有的单个学习器都是同类的，例如都是决策树，或者都是神经网络，那么这个集成就叫做同质（Homogeneous）；反之，如果既有决策树又有神经网络，那么集成就叫做异质（heterogeneous）的。

- 集成的泛化能力是远大于单个学习器的泛化能力的，但也需关注两点：

1) 准确性：个体学习器不能太差，要有一定的准确度

2) 多样性：个体学习器之间的输出要具有差异性，各有所长

- **Bagging算法过程：**

① 从原始样本集中抽取训练集.每轮从原始样本集中使用Bootstraping的方法抽取n个训练样本（在训练集中，有些样本可能被多次抽取到，而有些样本可能一次都没有被抽中）.共进行k轮抽取，得到k个训练集.（k个训练集相互独立）

② 每次使用一个训练集得到一个模型，k个训练集共得到k个模型.（注：根据具体问题采用不同的分类或回归方法，如决策树、神经网络等）

③ 对分类问题：将上步得到的k个模型采用投票的方式得到分类结果；对回归问题，计算上述模型的均值作为最后的结果.

■**Boosting的两个核心思路：**

1）通过提高那些在前一轮被弱分类器分错样例的权值，减小前一轮分对样本的权值，而误分的样本在后续受到更多的关注.

2）通过加法模型将弱分类器进行线性组合，比如AdaBoost通过加权多数表决的方式，即增大错误率小的分类器的权值，同时减小错误率较大的分类器的权值；而提升树通过拟合残差的方式逐步减小残差，将每一步生成的模型叠加得到最终模型。

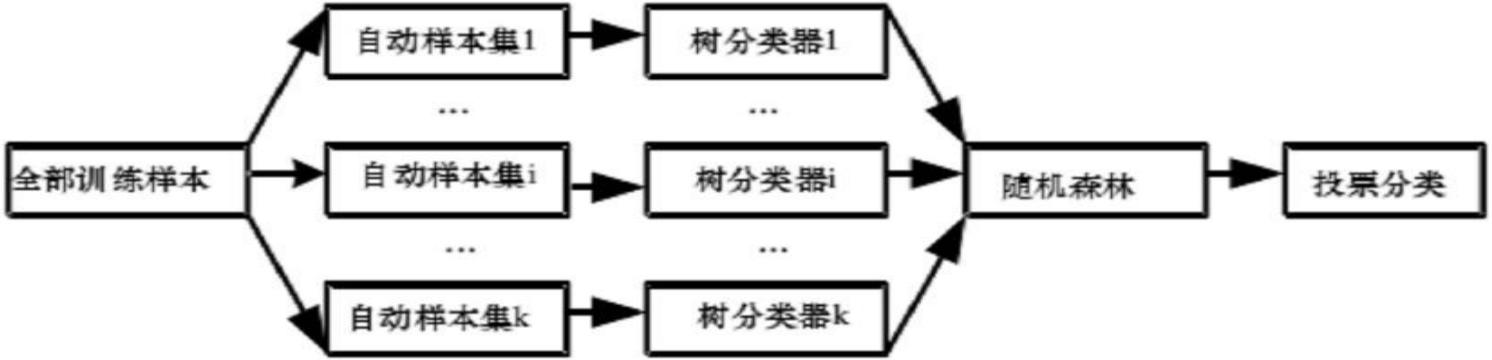
随机森林 vs. Adaboost

随机森林 vs. Adaboost

- ✓ 随机森林是采用有放回抽样得到n个训练集，每个训练集都会有重复的样本，然后对每个训练集生成一个决策树。
- ✓ 决策的时候综合每棵决策树的评分，最终得出一个总的评分。
- ✓ Adaboost每次训练的时候用的是同一个数据集。
- ✓ 后面的决策树会利用了前面决策树学习的结果，不断的优化。

拓展1：随机森林

随机森林是一个包含多个决策树的分类器，并且其输出的类别是单个决策树输出类别投票而定。



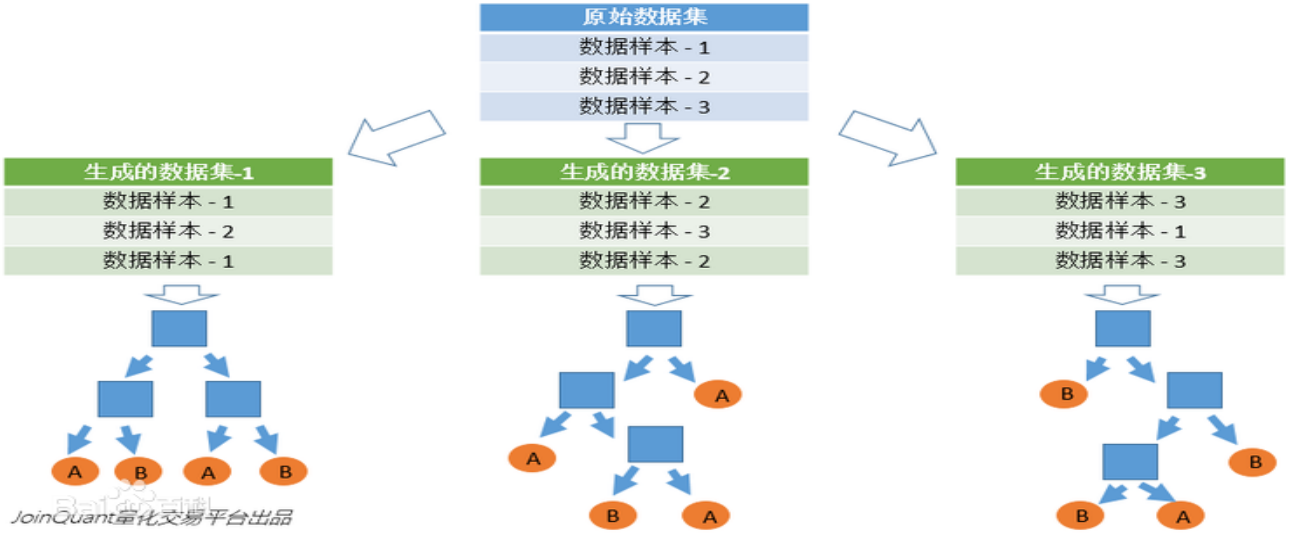
拓展1：随机森林

- 数据的随机性选取
- 待选特征的随机选取

拓展1： 随机森林

数据的随机选取

- 从原始的数据集中采取**有放回的抽样**，构造子数据集，**子数据集的数据量是和原始数据集相同的**。
- 不同子数据集的元素可以重复，同一个子数据集中的元素也可以重复。



拓展1：随机森林

待选特征的随机选取

- 从所有的待选特征中**随机选取一定数量的特征**，基于这些特征构建决策树。
- 这样的随机选取能够使得随机森林中的**多棵决策树彼此不同，提升系统的多样性**，从而提升分类性能。
- 从M个特征中，选择m个（m<<M）

拓展1：随机森林

随机森林的优势

- 在数据集上表现良好
- 能够处理很高维度数据，且不用做特征选择
- 在训练完后，能够给出哪些feature比较重要
- 训练速度快
- 在训练过程中，能够检测到feature间的互相影响
- 容易做成并行化方法
- 实现比较简单

拓展1：随机森林

Matlab自带的随机森林工具包TreeBagger

用法：

B= TreeBagger(Ntrees,x,y,'Method', ‘classification');

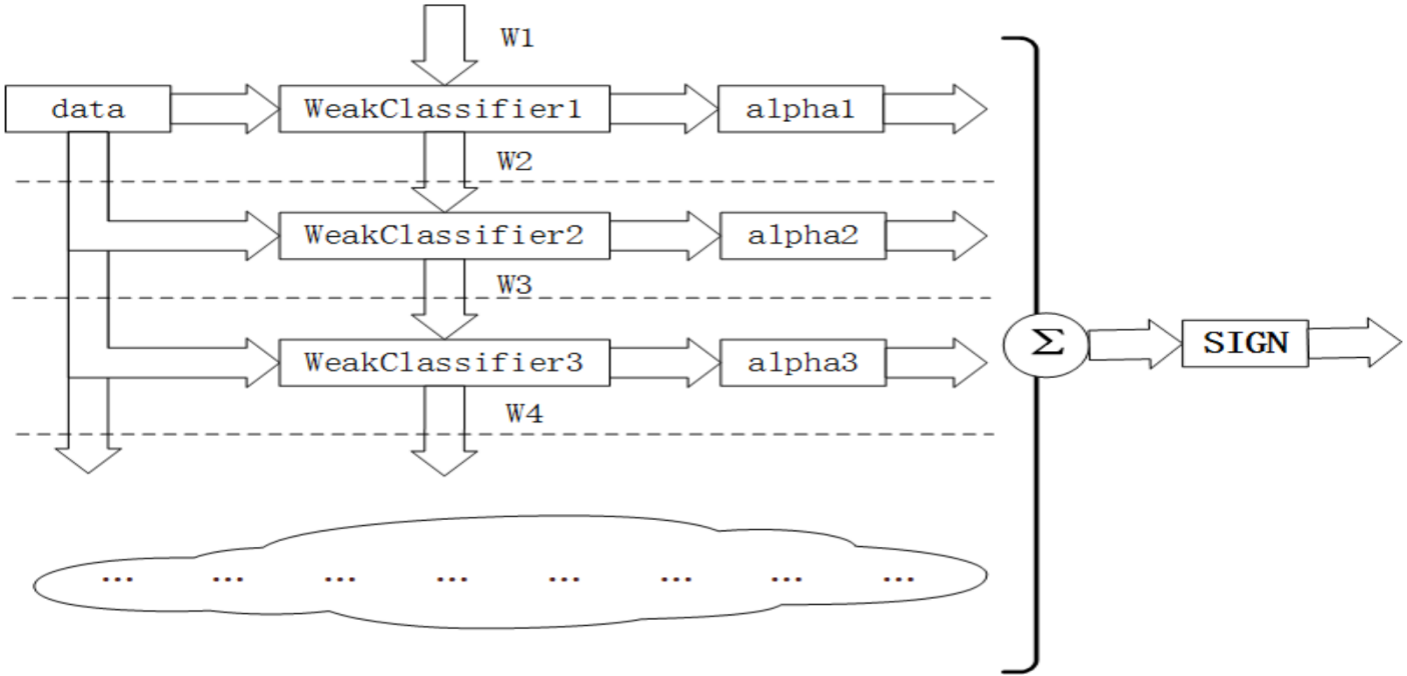
Y3 = B.predict(x3);

Ntrees为树的数量；

x为训练集；

y为训练集标签；

拓展2：Adaboost



拓展2：Adaboost

- AdaBoost is for **binary classification:** $\mathcal{Y} = \{-1, 1\}$
- **Base hypothesis space** $\mathcal{H} = \{h : \mathcal{X} \to \{-1, 1\}\}$.
  - **Note**: not producing a score, but an actual class label.
  - we'll call it a **base learner**
  - (when base learner satisfies certain conditions, it's called a "weak learner")
- Typical base hypothesis spaces:
  - **Decision stumps** (tree with a single split)
  - Trees with few terminal nodes
  - Linear decision functions

拓展2：Adaboost

Given training set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.

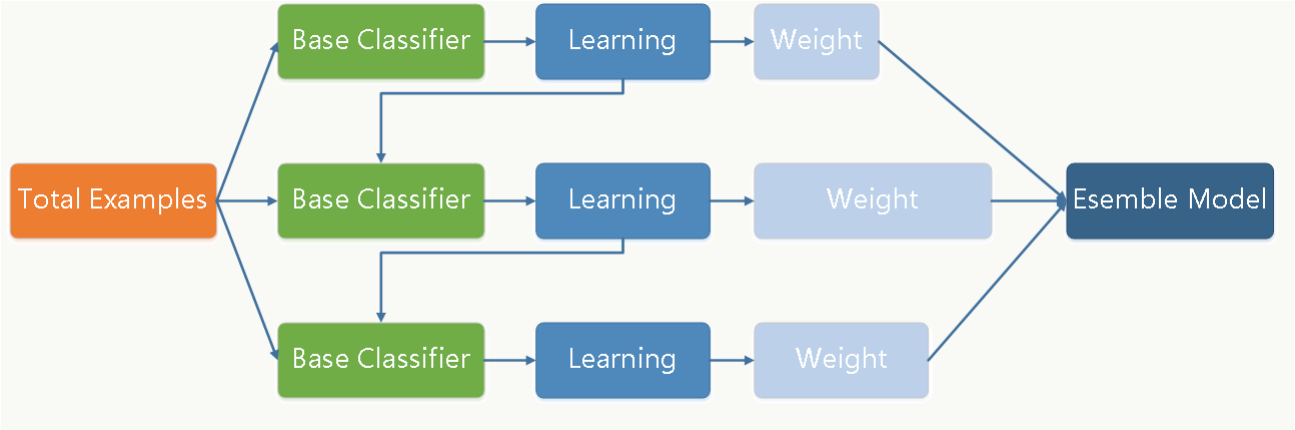
1. Initialize observation weights $w_i = 1,\ i = 1, 2, \ldots, n$.
2. For $m = 1$ to M:
   1. Base learner fits weighted training data and returns $G_m(x)$
   2. Compute **weighted empirical 0-1 risk**:

$$\text{err}_m = \frac{1}{W} \sum_{i=1}^{n} w_i 1(y_i \neq G_m(x_i)) \quad \text{where } W = \sum_{i=1}^{n} w_i.$$

   3. Compute $\alpha_m = \ln\left(\frac{1-\text{err}_m}{\text{err}_m}\right)$ **[classifier weight]**
   4. Set $w_i \leftarrow w_i \cdot \exp\left[\alpha_m 1(y_i \neq G_m(x_i))\right], \quad i = 1, 2, \ldots, n$ **[example weight adjustment]**
3. Ouptut $G(x) = \text{sign}\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right]$.

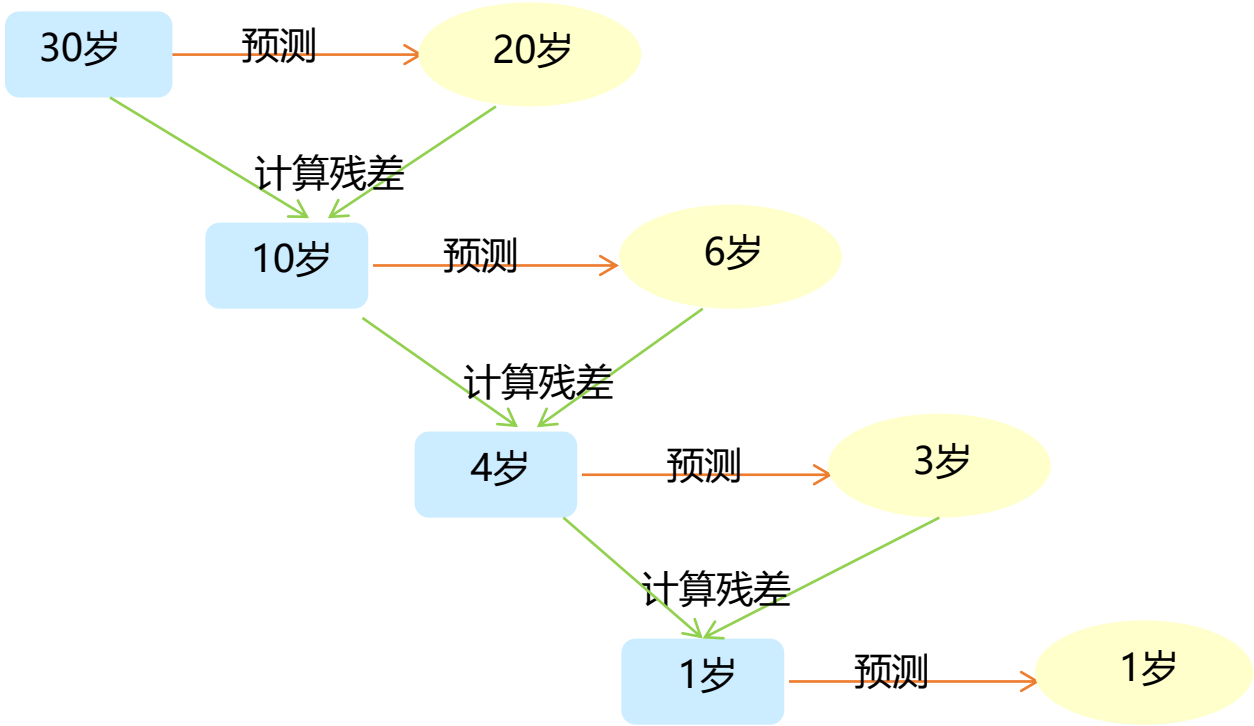
拓展3：GBDT

GBDT通过多轮迭代，每轮迭代产生一个弱分类器，每个分类器在上一轮分类器的残差基础上进行训练。



拓展3：GBDT

直观理解



拓展3：GBDT

优点

- ✓ 预测阶段的计算速度快，树与树之间可并行化计算。
- ✓ 在分布稠密的数据集上，泛化能力和表达能力都很好，这使得GBDT在Kaggle的众多竞赛中，经常名列榜首。
- ✓ 采用决策树作为弱分类器使得GBDT模型具有较好的解释性和鲁棒性，能够自动发现特征间的高阶关系。

局限性

- ✓ GBDT在高维稀疏的数据集上，表现不如支持向量机或者神经网络。
- ✓ GBDT在处理文本分类特征问题上，相对其他模型的优势不如它在处理数值特征时明显。
- ✓ 训练过程需要串行训练，只能在决策树内部采用一些局部并行的手段提高训练速度。

拓展3：GBDT

随机森林 & GBDT异同

相同点：

- ✓ 都是由多棵树组成，最终的结果都是由多棵树一起决定。
- ✓ RF和GBDT在使用CART树时，可以是分类树或者回归树。

不同点：

- ✓ 组成随机森林的树可以并行生成，而GBDT是串行生成
- ✓ 随机森林的结果是多数表决的，而GBDT则是多棵树累加之和
- ✓ 随机森林对异常值不敏感，而GBDT对异常值比较敏感
- ✓ 随机森林是减少模型的方差，而GBDT是减少模型的偏差
- ✓ 随机森林不需要进行特征归一化，而GBDT则需要进行特征归一化

应用1：基于随机森林的脑活动解码

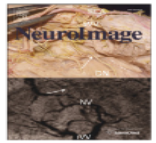
“brain reading”

NeuroImage 56 (2011) 544–553

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg



Performance comparison of machine learning algorithms and number of independent components used in fMRI decoding of belief vs. disbelief

P.K. Douglas [a,*], Sam Harris [b], Alan Yuille [c], Mark S. Cohen [a,b,d]

[a] Department of Biomedical Engineering, University of California, Los Angeles, USA
[b] Interdepartmental Neuroscience Program, University of California, Los Angeles, USA
[c] Department of Statistics, University of California, Los Angeles, USA
[d] Department of Psychiatry, Neurology, Psychology, Radiological Sciences, Biomedical Physics, University of California, Los Angeles, USA

K*
Naive Bayes
SVM
Decision Tree
Adaboost
Random Forest

应用1：基于随机森林的脑活动解码

Maximum accuracy was achieved at 92% for Random Forest, followed by 91% for AdaBoost, 89% for Naïve Bayes, 87% for a J48 decision tree, 86% for K*, and 84% for support vector machine. For real-time decoding

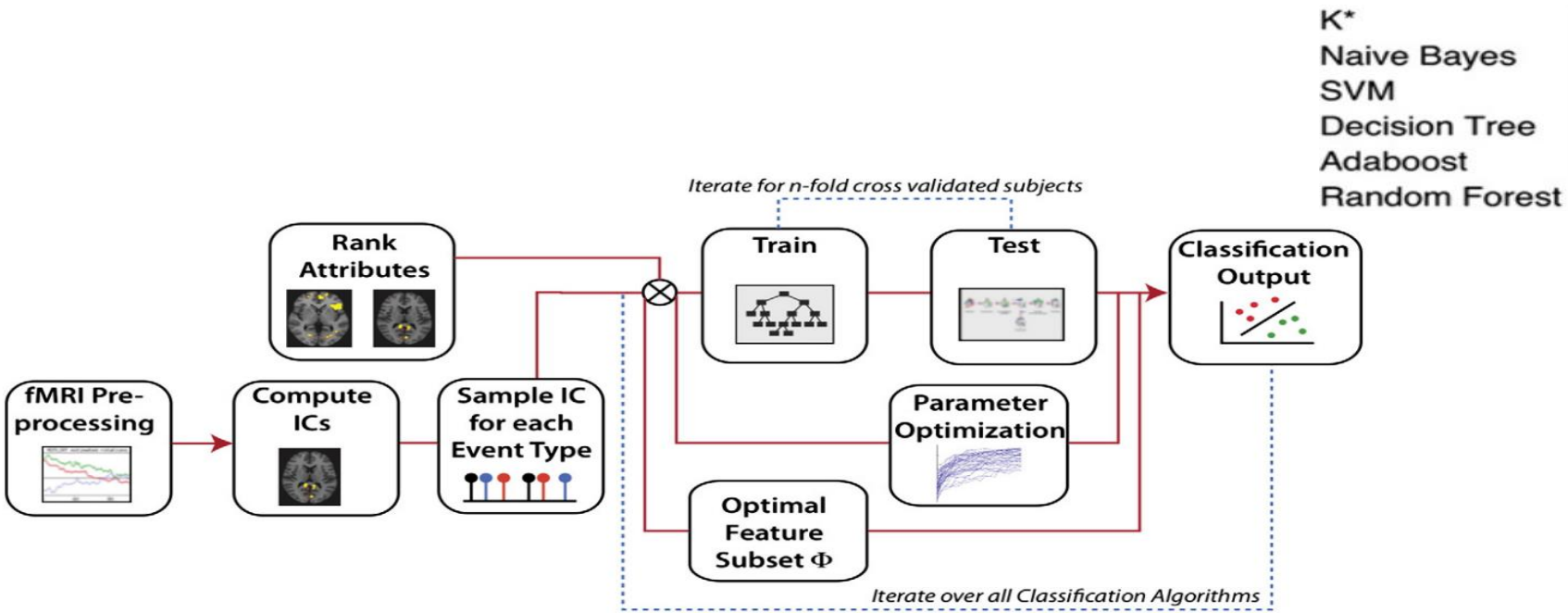


Fig. 1. Methodology flow diagram. Following preprocessing steps that included motion correction and brain extraction, independent component analysis (ICA) was performed and time courses associated with each spatial IC were sampled for "belief" and "disbelief" conditions. IC component features were ranked and then sent as inputs into machine learning for training and testing of the classifier, which proceeds over an n-fold cross-validated sample. Classifier parameters are adjusted and optimized.

应用1：基于随机森林的脑活动解码

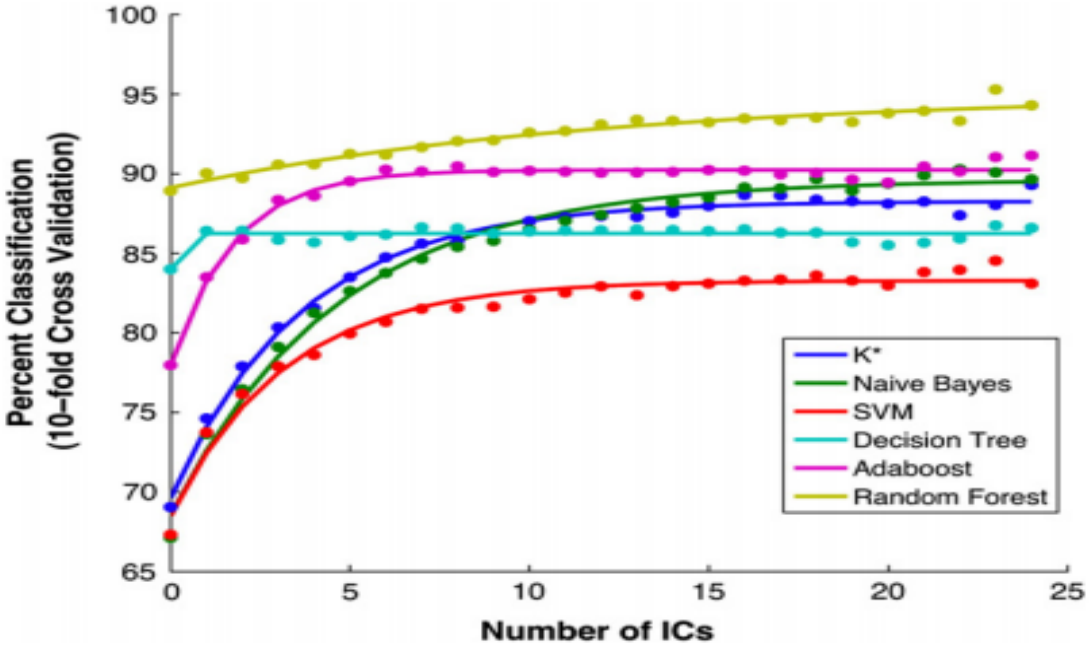


Fig. 4. Classification accuracy averaged across all subjects, shown for each of the six classifiers as a function of the number of ICs, with fits to 3-parameter first order exponential model (lines).

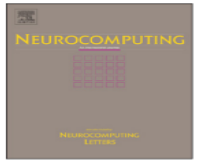
应用2：基于Adaboost的脑图像分类

Neurocomputing 177 (2016) 188–197

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Brain MR image classification using two-dimensional discrete wavelet transform and AdaBoost with random forests

Deepak Ranjan Nayak *, Ratnakar Dash, Banshidhar Majhi

Pattern Recgnition Research Lab, Department of Computer Science and Engineering, National Institute of Technology, Rourkela 769 008, India

应用2：基于Adaboost的脑图像分类

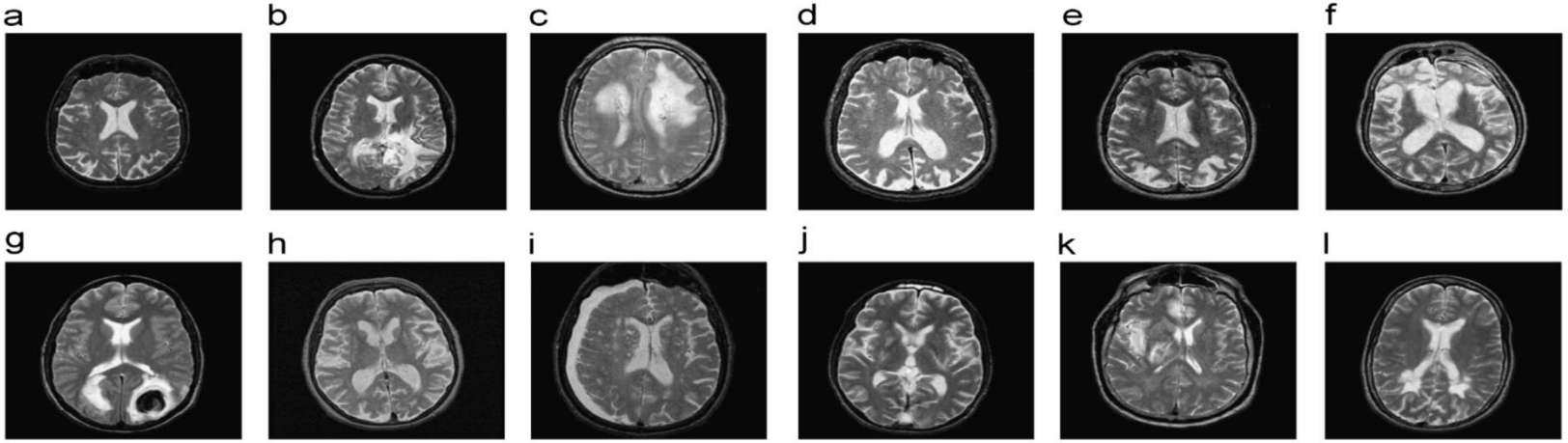


Fig. 4. Samples from different brain MR images: (a) Normal, (b) glioma, (c) meningioma, (d) Alzheimer's disease, (e) Alzheimer's disease plus visual agnosia, (f) Pick's disease, (g) sarcoma, (h) Huntington's disease, (i) chronic subdural hematoma, (j) cerebral toxoplasmosis, (k) herpes encephalitis, and (l) multiple sclerosis.

应用2：基于Adaboost的脑图像分类

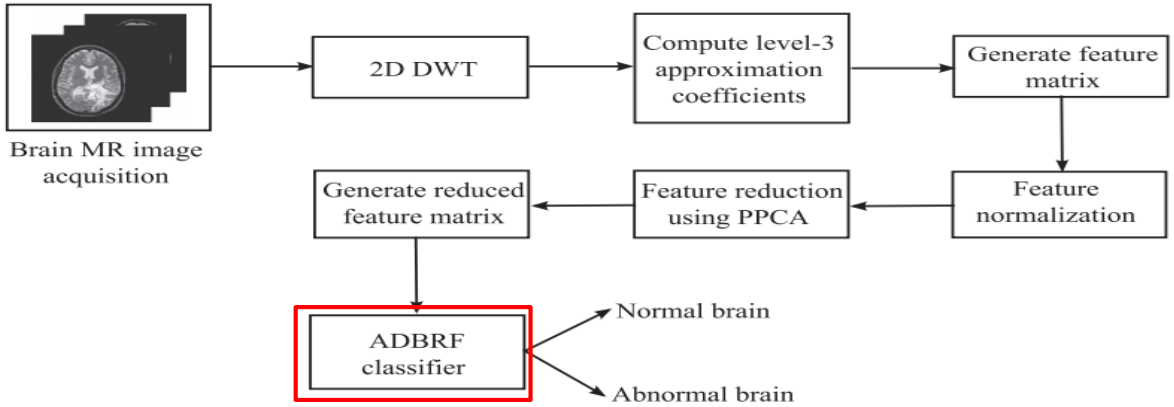


Fig. 1. Block diagram of the proposed method for brain MR image classification.

Performance metrics of different classifiers in the first run of CV procedure.

Classifier	Dataset-66			Dataset-160			Dataset-255		
	Sensitivity	Specificity	*ACC* (%)	Sensitivity	Specificity	*ACC* (%)	Sensitivity	Specificity	*ACC* (%)
k-NN	0.98	0.99	98.49	1.00	0.85	98.12	0.95	1.00	95.69
BPNN	1.00	1.00	100.00	1.00	0.90	98.75	0.95	0.94	95.29
SVM	1.00	1.00	100.00	1.00	1.00	100.00	1.00	0.91	98.82
RF	0.98	0.99	98.49	0.97	0.82	97.50	0.96	0.87	96.07
ADBRF	**1.00**	**1.00**	**100.00**	**1.00**	**1.00**	**100.00**	**0.99**	**1.00**	**99.61**

目录

作业

1. 给定如下样本数据集，计算并画出ID3决策树，判断样本属于类1还是类2。

序号	属性1	属性2	属性3	类别
1	A	E	真	类1
2	A	G	真	类2
3	A	G	假	类2
4	A	G	假	类2
5	A	E	假	类1
6	B	G	真	类1
7	B	F	假	类1
8	B	E	真	类1
9	B	F	假	类1
10	C	F	真	类2
11	C	E	真	类2
12	C	F	假	类1
13	C	F	假	类1
14	C	G	假	类1

作业

2. 数据表给出了14个基于天气（Outlook）、温度（Temp.）、湿度(Humidity) 和风力强弱（Wind）判断是否打高尔夫球（Decision）的例子。计算Outlook、Temp.、Humidity和Wind的基尼指数（Gini_index），并选择决策树的第一个划分属性。

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

作业

3. 有如下二分类问题数据集，左侧为原数据，右侧上下两个表为属性A、B的统计数据。

（1）计算按照属性A和B划分时的信息增益。决策树算法将首先会选择哪个属性?

（2）计算按照属性A和B划分时GINI指标。决策树归纳算法将首先会选择哪个属性?

A	B	类标号
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

统计A

	A=T	A=F
+	4	0
-	3	3

统计B

	B=T	B=F
+	3	1
-	1	5

上机作业

1 请基于datamelon.txt提供的西瓜数据集分别使用ID3，C4.5算法构建决策树并体现构建过程。

西瓜数据集共分为六个属性特征：色泽、根蒂、敲声、纹理、脐部、触感，根据这六种属性特征来决定是否是好瓜。

用testmelon.txt提供的数据集进行测试，并报告分类正确率。

2 请基于“iris数据.csv”文件鸢尾花数据集用CART分类算法识别鸢尾花的类别。

北京交通大学《机器学习》课程组成员

于　剑：jianyu@bjtu.edu.cn，http://faculty.bjtu.edu.cn/6463/

景丽萍：lpjing@bjtu.edu.cn，http://faculty.bjtu.edu.cn/8249/

田丽霞：lxtian@bjtu.edu.cn，http://faculty.bjtu.edu.cn/7954/

黄惠芳：hfhuang@bjtu.edu.cn，http://faculty.bjtu.edu.cn/7418/

吴　丹：wudan@bjtu.edu.cn，http://faculty.bjtu.edu.cn/8925/

万怀宇：hywan@bjtu.edu.cn，http://faculty.bjtu.edu.cn/8793/

王　晶：wj@bjtu.edu.cn，http://faculty.bjtu.edu.cn/9167/

参考答案：作业1计算

1. 给定如下样本数据集，计算并画出ID3决策树，判断样本属于类1还是类2。

序号	属性1	属性2	属性3	类别
1	A	E	真	类1
2	A	G	真	类2
3	A	G	假	类2
4	A	G	假	类2
5	A	E	假	类1
6	B	G	真	类1
7	B	F	假	类1
8	B	E	真	类1
9	B	F	假	类1
10	C	F	真	类2
11	C	E	真	类2
12	C	F	假	类1
13	C	F	假	类1
14	C	G	假	类1

参考答案：作业1计算

(1)首先计算类别的信息熵：

$$Ent(D) = -p_1\log_2(p_1) - p_2\log_2(p_2) = -0.6429\log_2(0.6429) - 0.3571\log_2(0.3571)$$
$$= 0.94$$

(2) 以‘属性 1’ 划分得到三个子集：D1（属性1=A）；D2（属性1=B）；D3(属性1=C)

以‘属性 1’ 划分之后所获得的三个子节点的信息熵分别为：

$$Ent(D1) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$$
$$Ent(D2) = -\log_2 1 - 0 = 0$$
$$Ent(D3) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$$

所以，‘属性1’的信息增益为：

$$Gain(D, 属性1) = Ent(D) - \sum_{v=1}^{3}\frac{|D^v|}{D}Ent(D^v)$$
$$= 0.94 - (\frac{5}{14}\times 0.971 + \frac{4}{14}\times 0 + \frac{5}{14}\times 0.971)$$
$$= 0.246$$

参考答案：作业1计算

(3)对于以‘属性 2’ 划分得到子集D1（属性2=E）；D2（属性2=F）；D3(属性2=G)

$$Ent(D1) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = 0.811$$

$$Ent(D2) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.722$$

$$Ent(D3) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$$

所以，‘属性2’ 的信息增益为：

$$Gain(D, 属性2) = Ent(D) - \sum_{v=1}^{3}\frac{|D^v|}{D}Ent(D^v)$$

$$= 0.94 - (\frac{4}{14}\times 0.811 + \frac{5}{14}\times 0.722 + \frac{5}{14}\times 0.971)$$

$$= 0.103$$

参考答案：作业1计算

(4)对于以‘属性 3’ 划分得到子集，同理可求出：

$$Ent(D1) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$Ent(D2) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = 0.811$$

所以： ‘属性3’ 的信息增益为:

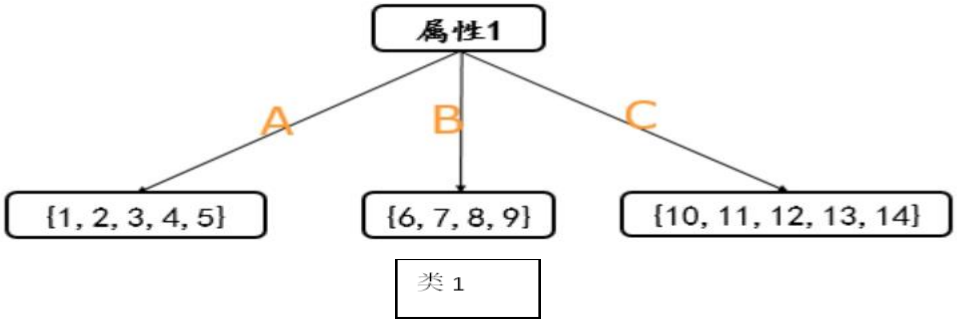
$$Gain(D, 属性3) = Ent(D) - \sum_{v=1}^{3}\frac{|D^v|}{D}Ent(D^v)$$

$$= 0.94 - (\frac{6}{14}\times 1 + \frac{8}{14}\times 0.811)$$

$$= 0.048$$

参考答案：作业1计算

(5)显然，‘属性1’的信息增益最大，所以把它选为划分属性。下图给出了基于‘属性1’对根节点进行划分的结果，各分支结点的样例子集显示在结点中：



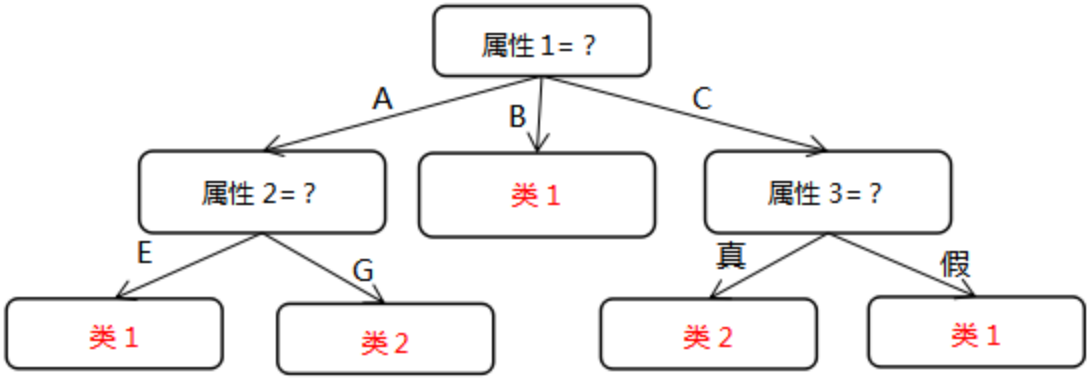
参考答案：作业1计算

(6)进一步划分

①对于分支结点C的样例集合，下一分支再用属性3判别即可。

②对于分支结点A的样例集合{1,2,3,4,5}的5个样例。用属性2的判别即可到达最大信息增益。

所以决策树如下图所示：



参考答案：作业2计算

2. 数据表给出了14个基于天气（Outlook）、温度（Temp.）、湿度(Humidity) 和风力强弱（Wind）判断是否打高尔夫球（Decision）的例子。计算Outlook、Temp.、Humidity和Wind的基尼指数（Gini_index），并选择决策树的第一个划分属性。

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

参考答案：作业2计算

（1）整理Outlook如下表

Outlook	Yes	No	Number of instances
Sunny	2	3	5
Overcast	4	0	4
Rain	3	2	5

由此可得：

$$Gini(Outlook = Sunny) = 1 - (\frac{2}{5})^2 - (\frac{3}{5})^2 = 0.48$$

$$Gini(Outlook = Overcast) = 1 - (\frac{4}{4})^2 - (\frac{0}{4})^2 = 0$$

$$Gini(Outlook = Rain) = 1 - (\frac{3}{5})^2 - (\frac{2}{5})^2 = 0.48$$

$$Gini\_index(Outlook) = \frac{5}{14} \times 0.48 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.48 = 0.342$$

参考答案：作业2计算

(2) 整理Temperature如下表

Temperature	Yes	No	Number of instances
Hot	2	2	4
Cool	3	1	4
Mild	4	2	6

由此可得：

$$Gini(Temp = Hot) = 1 - (\frac{2}{4})^2 - (\frac{2}{4})^2 = 0.5$$

$$Gini(Temp = Cool) = 1 - (\frac{3}{4})^2 - (\frac{1}{4})^2 = 0.375$$

$$Gini(Temp = Mild) = 1 - (\frac{4}{6})^2 - (\frac{2}{6})^2 = 0.445$$

$$Gini\_index(Temp) = \frac{4}{14} \times 0.5 + \frac{4}{14} \times 0.375 + \frac{6}{14} \times 0.445 = 0.439$$

参考答案：作业2计算

（3）整理Humidity如下表

Humidity	Yes	No	Number of instances
High	3	4	7
Normal	6	1	7

由此可得：

$$Gini(Humidity = High) = 1 - (\frac{3}{7})^2 - (\frac{4}{7})^2 = 0.489$$

$$Gini(Humidity = Normal) = 1 - (\frac{6}{7})^2 - (\frac{1}{7})^2 = 0.244$$

$$Gini\_index(Humidity) = \frac{7}{14} \times 0.489 + \frac{7}{14} \times 0.244 = 0.367$$

参考答案：作业2计算

（4）整理Wind如下表

Wind	Yes	No	Number of instances
Weak	6	2	8
Strong	3	3	6

由此可得：

$$Gini(Wind = Weak) = 1 - (\frac{6}{8})^2 - (\frac{2}{8})^2 = 0.375$$

$$Gini(Wind = Strong) = 1 - (\frac{3}{6})^2 - (\frac{3}{6})^2 = 0.5$$

$$Gini\_index(Wind) = \frac{8}{14} \times 0.375 + \frac{6}{14} \times 0.5 = 0.428$$

参考答案：作业2计算

（5）决策

$$Gini\_index(Outlook) < Gini\_index(Humidity) < Gini\_index(Wind) < Gini\_index(Temp)$$

所以应选择Outlook为决策树的第一个划分节点。

参考答案：作业3计算

3. 有如下二分类问题数据集，左侧为原数据，右侧上下两个表为属性A、B的统计数据。

（1）计算按照属性A和B划分时的信息增益。决策树算法将首先会选择哪个属性?

（2）计算按照属性A和B划分时GINI指标。决策树归纳算法将首先会选择哪个属性?

A	B	类标号
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

统计A

	A=T	A=F
+	4	0
-	3	3

统计B

	B=T	B=F
+	3	1
-	1	5

参考答案：作业3计算

（1）划分前样本集的信息熵：

$$Ent(D) = -0.4\log_2 0.4 - 0.6\log_2 0.6 = 0.9710$$

属性A信息熵：

$$Ent(A = T) = -\frac{4}{7}\log_2 \frac{4}{7} - \frac{3}{7}\log_2 \frac{3}{7} = 0.9852$$

$$Ent(A = \mathrm{F}) = -\frac{0}{3}\log_2 \frac{0}{3} - \frac{3}{3}\log_2 \frac{3}{3} = 0$$

属性A信息增益：

$$Gain(D, A) = 0.9710 - \frac{7}{10} \times 0.9852 - 0 = 0.2813$$

同理可得，按照B属性划分样本集的信息增益 $Gain(D, B) = 0.2565$

因此将选择A属性

参考答案：作业3计算

（2）由统计A 可得：

$$Gini(A=T)=1-(\frac{4}{7})^2-(\frac{3}{7})^2=0.4898$$

$$Gini(A=F)=1-(\frac{0}{3})^2-(\frac{3}{3})^2=0$$

$$Gini\_index(A)=\frac{7}{10}\times 0.4898+\frac{3}{10}\times 0=0.34286$$

由统计B可得：

$$Gini(A=T)=1-(\frac{3}{4})^2-(\frac{1}{4})^2=0.3750$$

$$Gini(A=F)=1-(\frac{1}{6})^2-(\frac{5}{6})^2=0.2778$$

$$Gini\_index(A)=\frac{4}{10}\times 0.3750+\frac{6}{10}\times 0.2778=0.31668$$

由于 $Gini\_index(B)<Gini\_index(A)$

所以将选择B属性