第4章：分类
-决策树

1. 周志华，《机器学习》，清华大学出版社，2016，P$_{73}$
2. 陈封能等，《数据挖掘导论》第2版，机械工业出版社，2019，P$_{69}$
3. 李航著，《统计学习方法》第2版，清华大学出版社，2019，P$_{67}$
4. 《机器学习实战：基于Scikit-Learn、Keras和TensorFlow(原书第2版)》，Aurelien Geron著，王静源等译，机械工业出版社，2020， 第6章

大纲

- 基本概念
- 基本流程
- 划分选择
- 剪枝处理
- 连续值
- 多变量决策树

基本概念

- 在前面的学习中，我们提到过机器学习的含义：对于当前数据集

$$\boldsymbol{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m)\}$$

通过设计算法，得到某一个模型 f，使得对于一个新的数据样本 $\boldsymbol{x}$，模型能够得到一个近似的预测 $f(\boldsymbol{x})$。

- 我们已经讨论了线性模型，即 $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ 。今天我们讨论另一类模型：决策树
- **决策树**（decision tree）模型常常用来解决**分类**和**回归**问题。

决策树常用三类算法

- 决策树算法起源于E. B. Hunt等人于1966年发表的论文“Experiments in Induction”，但真正让决策树成为机器学习主流算法的还是 Quinlan（罗斯·昆兰）（2011年获得了数据挖掘领域最高奖KDD创新奖）
- 昆兰在1979年提出了ID3(Iterative Dichotomiser 3, 迭代二分器)算法，掀起了决策树研究的高潮。
- 现在最常用的决策树算法C4.5是昆兰在1993年提出的。(关于为什么叫C4.5，有个轶事：因为昆兰提出ID3后，掀起了决策树研究的高潮，然后ID4，ID5等名字就被占用了，因此昆兰只好将自己对ID3的改进叫做C4.0(Classifier 4.0的简称)，C4.5是C4.0的改进)。现在有了商业应用新版本是C5.0。
- 分类和回归树 (简称CART, Classification and Regression Tree) 是李奥·布瑞曼(Leo Breiman) 于1984年引入。

西瓜数据集2.0的样本表示（二分类）

- 样例使用元组（$\boldsymbol{x}$, y）表示。
- 例如（$\boldsymbol{x}$, y）=（{青绿，蜷缩，浊响，清晰，凹陷，硬滑}，是）

6个属性/特征　　标记

	编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
正样本	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
负样本	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

见《机器学习》P76, 表4.1 西瓜数据集2.0

分类流程-举例

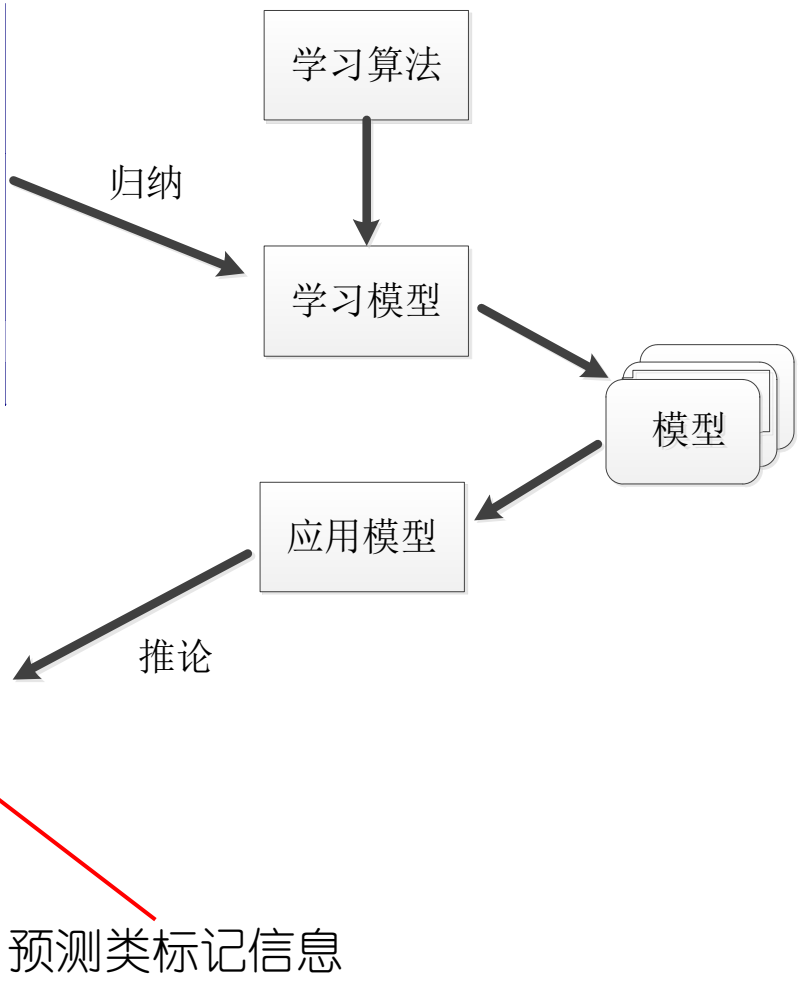
表 4.2

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

未见样本集

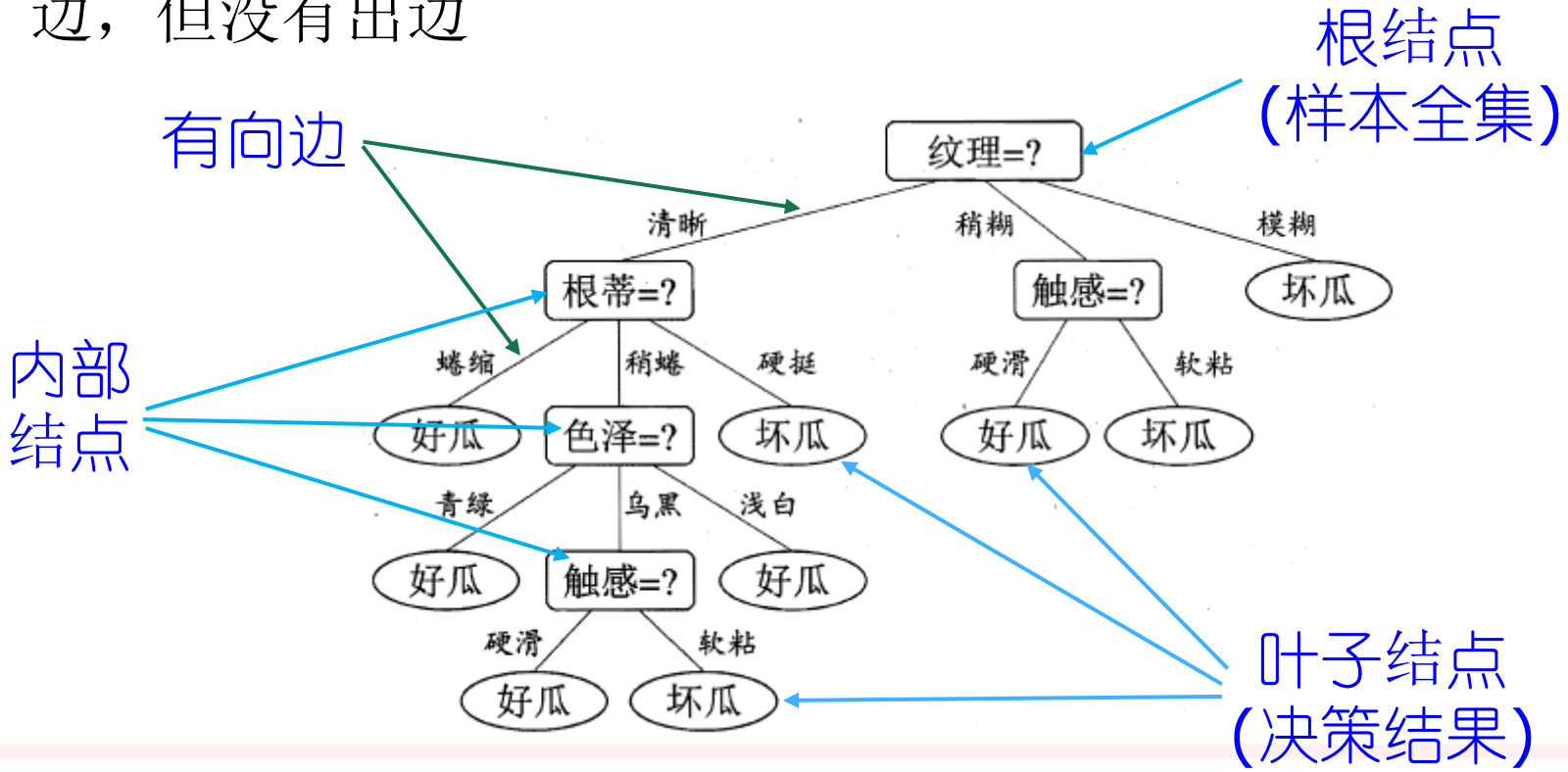
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	?
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	?
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	?
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	?
11	浅白	硬挺	清脆	模糊	平坦	硬滑	?
12	浅白	蜷缩	浊响	模糊	平坦	软粘	?
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	?



表格见 周志华，《机器学习》，清华大学出版社，2016，P_{80}

基本概念

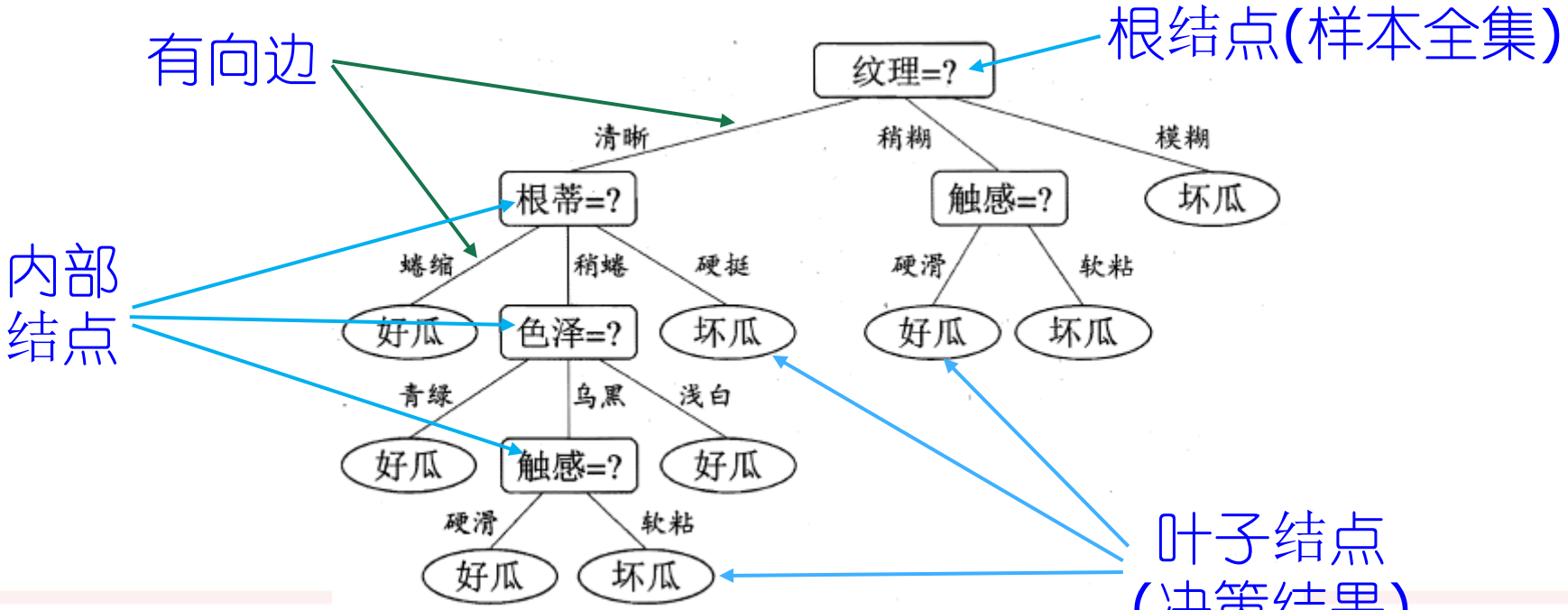
- **定义**:**决策树**是对数据样本进行分类的一种树型数据结构，包含:
- 根结点(root node)，它没有入边，但有零条或多条出边
- 内部结点(internal node)，恰有一条入边和两条或多条出边
- 叶子结点(leaf node)或终结点(terminal node)，恰有一条入边，但没有出边



西瓜问题的一棵决策树

基本概念

- 决策树一般包含结点(根结点，内部结点，叶子结点)和有向边
- 树中内部结点表示一个属性/特征
- 每个叶子结点都赋予一个类标记(决策结果)
- 从树的根结点到叶子结点的一条路径，就代表了一条决策规则



西瓜问题的一棵决策树

基本概念

判断一个瓜是否是好瓜？

色泽 —青绿→ 根蒂 —蜷缩→ 敲声 —浊响→ 好瓜

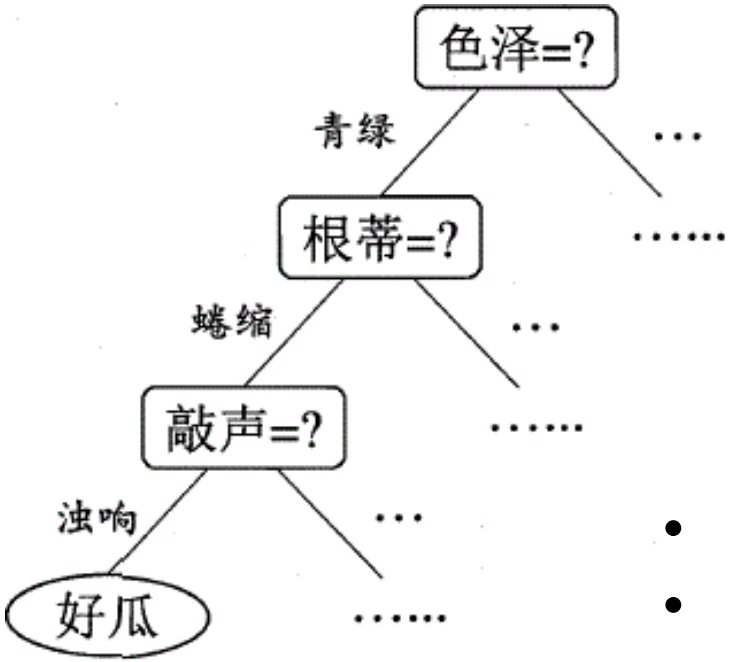


图 4.1 西瓜问题的一棵决策树

二分类学习任务
属性
属性值

- 根结点：包含全部样本
- 内部结点：对应属性测试
- 叶结点：对应决策结果 “好瓜” “坏瓜”

决策树学习的目的：为了产生一颗泛化能力强的决策树，即处理未见样本能力强。

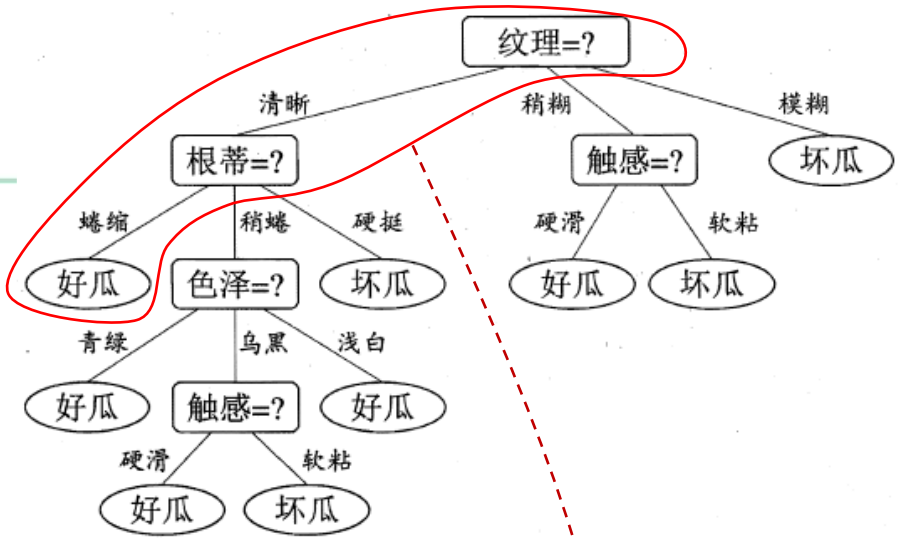
基本概念

- 由前述内容可知，决策树是以“树结构”的方式表示的一组if... then... 规则。
- 对于给定的数据集，我们可从中得出很多条if...then... 规则，决策树中所表示的那些决策规则有什么特点以及是如何得到的？

表4.1 西瓜数据集2.0

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

决策规则



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

- 如根据上表4.1，可以得到这样一条规则(由编号1)：

1. if 色泽=“青绿” and 根蒂=“蜷缩” and 敲声=“浊响” and 纹理=“清晰” and 脐部=“凹陷” and 触感=“硬滑” then 瓜=“好瓜”。

- 而观察决策树，我们也可以得到这样一条规则：

2. if 纹理=“清晰” and 根蒂=“蜷缩” then 瓜=“好瓜”。

很显然，第二条规则要更为清晰简洁（**模型更为简洁**），因此更具有泛化性能。

基本概念

- **总结：**决策树中包含的if...then... 规则应该与数据不矛盾，并且具有较强的泛化性能。
- 接下来我们讨论如何在给定数据集的情况下，构造决策树。

信息及信息熵

- 决策树的构造依赖于信息论，为此先介绍几个信息论中的重要概念。

- **信息（information, I）：**用来消除某种随机不确定性的东西，可以量化为某个事件的发生概率的对数值。
- 如果某个随机变量 X 发生的概率为 $P(x)$，它携带的信息量 $I(x)$ 为 $I(x) = -\log P(x)$
- **信息熵（information entropy，H, Ent）：**衡量随机变量的不确定性的某种度量(复杂度)。
- 一个离散型随机变量 X 的信息熵定义为 $H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$

- 熵越大，随机变量的不确定性就越大。那么如何计算不确定性？

信息及信息熵

- 不确定性是由于事物可能出现多种状态导致的，因此可以通过计算事物可能的状态及其每种状态的概率来计算。
- **信息熵**定义如下：

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

$$D : [x_1, x_2 \cdots, x_m]$$

$$P : [p_1, p_2 \cdots, p_m]$$

这里 D 代表当前样本(状态)集合，P 代表概率集合，假定当前样本集合 D 中第 k 类样本 x_k 所在比例为 $p_k\ (k = 1, 2, ..., |\mathcal{Y}|)$. $|\mathcal{Y}|$ 表示类别个数

- 信息熵是度量样本集合纯度最常用的一种指标

$\mathrm{Ent}(D)$ 的值越小，则 D 的纯度越高，样本的确定性越高

信息及信息熵

假设事物未获得某条信息之前的状态集与概率集为

$$X:[x_1,x_2\cdots,x_n],\ P:[p_1,p_2\cdots,p_n]$$

而获得了某条信息之后的状态与概率集合为

$$X':[x'_1,x'_2\cdots,x'_m],\ P':[p'_1,p'_2\cdots,p'_m]$$

则此信息的信息增益为:

$$I=H(X)-H(X')=-\sum_{i=1}^{n}p_i\log p_i-\left(-\sum_{i=1}^{m}p'_i\log p'_i\right)$$

ID3 算法

- 接下来，我们从信息论角度讨论如何构造决策树。

表4.1 西瓜数据集2.0

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

条件属性

决策属性

ID3 算法

- 对于上表来说，前面除去编号外的 6 个属性，我们称之为条件属性，最后一列称为决策属性。
- 前面的条件属性（如色泽）的作用：借助于观察西瓜的色泽是可以来帮助我们确定哪个西瓜是好瓜。换句话说，**可以帮助我们在一定程度上消除西瓜是好瓜还是坏瓜的不确定性**。
- 因此，色泽实际上是带给了我们一些信息，同样其他条件属性也带给了我们一些信息。那么很显然我们希望**找到那个带给我们最多信息的那个条件属性（即信息增益最大的属性）**。那么如何衡量每个属性的信息增益？

大纲

- 基本流程
- 划分选择
- 剪枝处理
- 连续与缺失值
- 多变量决策树

划分选择1-信息增益

- 离散属性a有V个可能的取值$\{a^1, a^2, ..., a^V\}$，用a来进行划分，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。则可计算出用属性a对样本集D进行划分所获得的“信息增益”：

属于离散属性a且取值为a^v的样本个数

$$\mathrm{Gain}(D,a) = \mathrm{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Ent}(D^v) = \underbrace{\mathrm{Ent}(D)}_{\text{划分前信息熵}} - \underbrace{\mathrm{Ent}(D|a)}_{\text{划分后信息熵}}$$

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

为分支结点权重，样本数越多的分支结点对结果影响越大

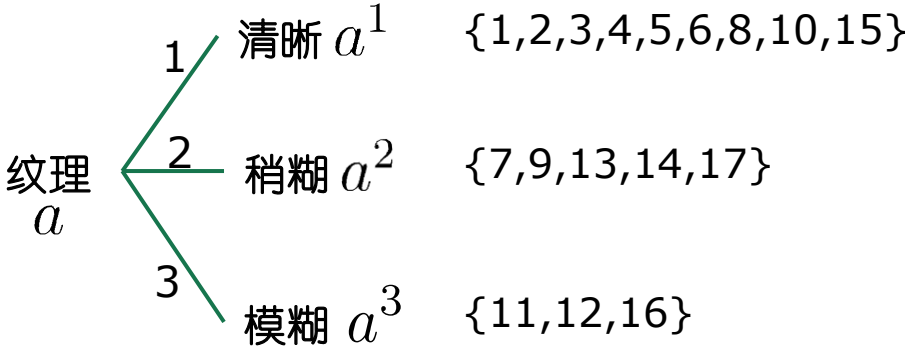
- 一般而言，信息增益越大，则意味着使用属性a来进行划分所获得的“纯度提升”越大（信息增益代表了在一个条件下，信息复杂度（不确定性）减少的程度）
- 迭代二分器（ID3）决策树学习算法[Quinlan, 1986]以**信息增益**为准则来选择划分属性

划分选择1-信息增益

- 离散属性a有V个可能的取值$\{a^1, a^2, ..., a^V\}$，用a来进行划分，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。则可计算出用属性a对样本集D进行划分所获得的“信息增益”：

属性“纹理”有3个可能取值

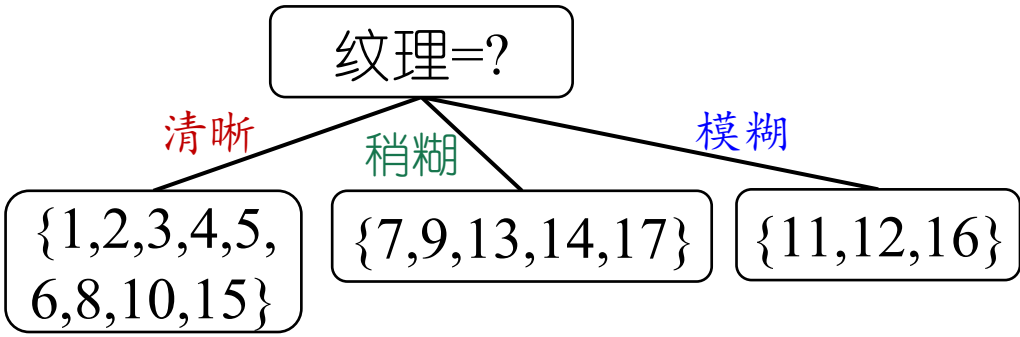
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	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



信息增益-简单示例

- 离散属性“纹理”有3个可能的取值{清晰，稍糊，模糊}，用纹理划分产生3个分支结点，其中第1个分支结点包含了D中所有在属性“纹理”上取值为清晰的样本(9个)，第2个分支结点包含了D中所有在属性“纹理”上取值为稍糊的样本(5个)，第3个分支结点包含了D中所有在属性“纹理”上取值为模糊的样本(3个)

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



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

划分选择1-信息增益实例

- 现在思考，在不考虑任何条件属性的情况下，观察西瓜好坏的分布.
- 该数据集包含17个训练样本，$|\mathcal{Y}| = 2$ ，在决策树开始学习时，根结点包含所有的样例，其中正例占$p_1 = \frac{8}{17}$，反例占 $p_2 = \frac{9}{17}$ ，计算得到根结点的信息熵 (经验熵) 为

表4.1 西瓜数据集2.0

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

$$
\begin{aligned}
\mathrm{Ent}(D) &= -\sum_{i=1}^{2} p_i \log p_i \\
&= -\left(\frac{8}{17}\log\frac{8}{17} + \frac{9}{17}\log\frac{9}{17}\right) \\
&= 0.998
\end{aligned}
$$

划分选择1-信息增益，以属性"色泽"为例

- 计算出当前属性集合{色泽，根蒂，敲声，纹理，脐部，触感}中每个属性的信息增益
- 以属性"色泽"为例，即使用"色泽"属性对D进行划分，其对应的3个数据子集分别为 D^1(色泽=青绿)，D^2(色泽=乌黑)，D^3(色泽=浅白)
- 子集D^1 包含编号为$\{1,4,6,10,13,17\}$ 的6个样例，其中正例占 $p_1=\frac{3}{6}$ ，反例占 $p_2=\frac{3}{6}$ ，第一个分支结点的信息熵为：

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

- 子集 D^2 包含编号为{2,3,7,8,9,15}的6个样例，其中正例占p_1= 4/6，反例占p_2= 2/6，第二个分支结点的信息熵为：

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

- 子集 D^3 包含编号为{5,11,12,14,16}的5个样例，其中正例占p_1= 1/5，反例占p_2= 4/5，第三个分支结点的信息熵为：

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

划分选择1-信息增益

- 属性“色泽”的信息增益为

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

- 按照相同的方法，可以计算出根蒂、敲声、纹理、脐部等的信息增益如下：

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

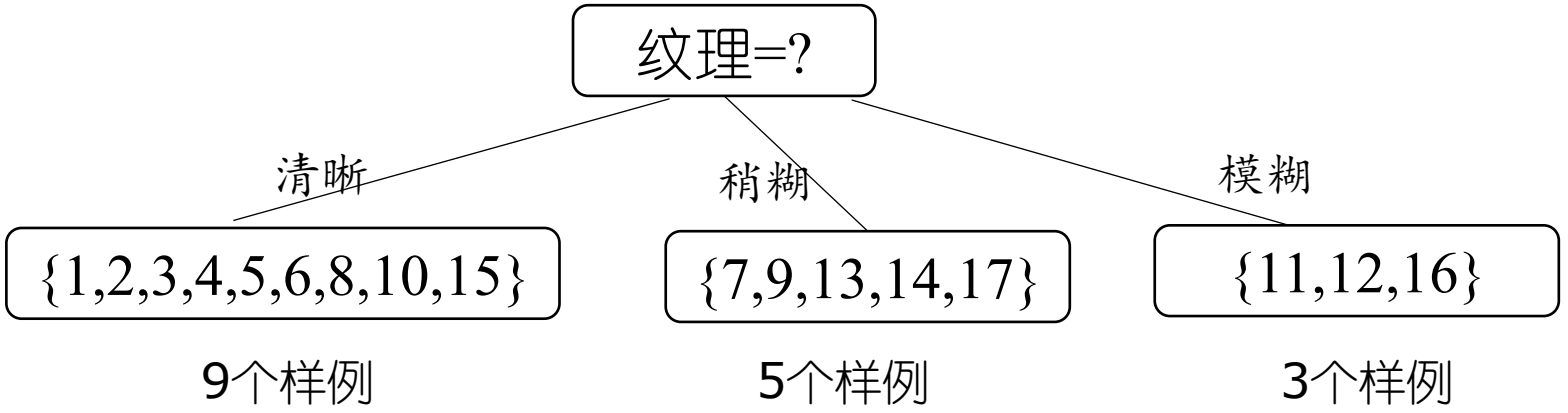
$$\text{Gain}(D, \text{纹理}) = 0.381 \qquad \text{Gain}(D, \text{脐部}) = 0.289$$

$$\text{Gain}(D, \text{触感}) = 0.006$$

- 最后可以发现，纹理的信息增益最大。因此，纹理属性帮助我们进行西瓜判断最有用，选择纹理属性作为划分属性，即作为决策树的根结点，得到如下决策树。

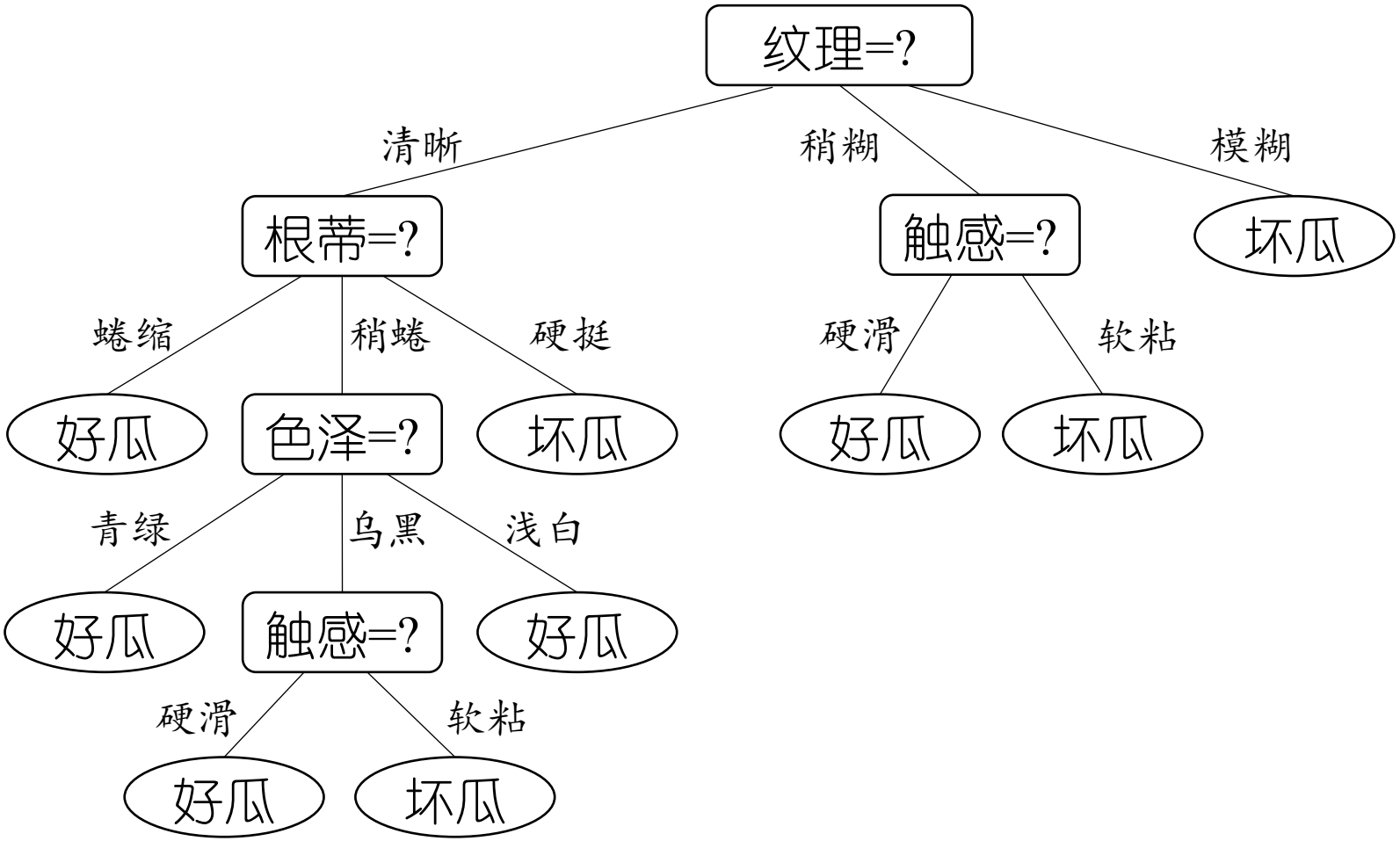
划分选择1-信息增益

- 显然，属性“纹理”的信息增益最大，其被选为划分属性(选为根结点)



划分选择1-信息增益

- 采用递归思想，类似地，决策树学习算法将对每个分支结点做进一步划分（纹理不再作为候选划分属性），最终得到的决策树如图：



ID3算法-划分选择-信息增益

算法总结（ID3 算法）：

输入： 训练集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m)\}$

属性集 $A = \{a_1, a_2, \cdots, a_d\}$

输出： 决策树 T : TreeGenerate(D, A)

STEP 1: 若 D 中所有样本都属于同一类 C, 则 T 为单结点树, 并将 C 作为该结点的类标记, 返回 T, 否则转 **STEP 2**;

STEP 2: 依据决策属性计算信息熵 Ent(D), 令 $k = 1$,

1: 选择 a_k, 假设 a_k 具有 v_k 个可能的取值, $D^{a_k^i}$ 为属性 $a_k = v_k^i$ 的样本集合, 计算条件信息熵 $\text{Ent}(D|a_k) = \sum_{i=1}^{v_k} \frac{|D^{a_k^i}|}{|D|} \text{Ent}(D^{a_k^i})$

2: 计算 a_k 属性的信息增益, $\text{Gain}(D, a_k) = \text{Ent}(D) - \text{Ent}(D|a_k)$;

3: $k = k + 1$, 若 $k < m$, 则跳转到 1;

决策树算法核心

STEP 3: 选择信息增益最大的属性 a_p 设为根结点, 根据 a_p 将数据集分成 v_p 个子集 $\{D^{a_p^1}, D^{a_p^2}, \cdots, D^{a_p^{v_p}}\}$;

STEP 4: 令 $D = D^{a_p^j}$, $A = A - a_p$, 转 **STEP 1**.

划分选择1-信息增益

存在的问题

- 若把“编号”也作为一个候选划分属性，可以计算得到编号属性的信息增益为0.998，则其信息增益远大于其他条件属性。
- 按照ID3 算法，应该将编号作为决策树的根结点，但很显然编号是无意义的属性。这样的决策树不具有泛化能力，无法对新样本进行有效预测。

信息增益对可取值数目较多的属性有所偏好

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2-增益率(C4.5算法)

- 增益率定义：

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

其中

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

称为属性a的“固有值” [Quinlan, 1993]，属性a的可能取值数目越多（即V越大），则$\text{IV}(a)$的值通常就越大

- 存在的问题

IV(触感)=0.874 (V=2)
IV(色泽)=1.580 (V=3)
IV(编号)=4.088 (V=17)

V的类别数越大，IV(a)越大

反过来，V的类别数越少，IV(a)越小，增益率越大

增益率准则对可取值数目较少的属性有所偏好

划分选择2-增益率(C4.5算法)

- 增益率定义：

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

其中

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

称为属性a的“固有值” [Quinlan, 1993]，属性a的可能取值数目越多（即V越大），则$\text{IV}(a)$的值通常就越大

- 存在的问题

> 增益率准则对可取值数目较少的属性有所偏好

- C4.5 [Quinlan, 1993]使用了一个启发式：先从候选划分属性中找出**信息增益**高于平均水平的属性，再从中选取**增益率**最高的

划分选择3-基尼指数

- 无论是ID3还是C4.5，都是基于信息论的熵模型的，这里面会涉及大量的对数运算。
- 能不能简化模型同时也不至于完全丢失熵模型的优点呢？
- CART [Breiman et al., 1984]采用“基尼指数”来选择划分属性
- CART分类树算法使用基尼系数来代替信息增益率，基尼系数代表了模型的不纯度，基尼系数越小，则不纯度越低，特征越好。这和信息增益(率)是相反的。
- CART作为分类树时，特征属性可以是连续类型也可以是离散类型，但标签属性(或者分类属性)必须是离散类型。

划分选择3-基尼指数

- 数据集 D 的纯度可用“基尼值”来度量

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

反映了从 D 中随机抽取两个样本，其类别标记不一致的概率

$p_k(k=1,2,\ldots,|\mathcal{Y}|)$ 表示选中的样本属于k类别的概率，那么这个样本被分错的概率是1-p_k

$\mathrm{Gini}(D)$ 越小，数据集越集中，D 的纯度越高

- 属性 a 的基尼指数定义为：

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

基尼指数（基尼不纯度）：表示在样本集合中一个随机选中的样本被分错的概率

- 应选择那个使划分后基尼指数最小的属性作为最优划分属性，即

$$a_*=\underset{a\in A}{\mathrm{argmin}}\ \mathrm{Gini\_index}(D,a)$$

基于基尼指数构造决策树的Python代码见学在西电“资料”栏-python编程参考

属性划分选择

- 那么到底使用基尼指数还是熵来进行属性划分选择呢？
- 其实大部分时候它们两种属性划分选择方式所生产的决策树基本相同，但Gini值的计算更快一些；
- 在机器学习标准库scikit-learn 中默认的划分方式就是Gini指数，默认的决策树是CART树；
- 但是Gini指数的划分趋向于孤立数据集中数量多的类，将它们分到一个树叶中，而熵偏向于构建一颗平衡的树，也就是数量多的类可能分散到不同的叶子中去了。

大纲

- 基本流程
- 划分选择
- 剪枝处理
- 连续与缺失值
- 多变量决策树

剪枝处理

- 为什么剪枝
  - 以上方法生成的决策树可能对训练数据有很好的分类能力，但对未知的测试数据未必有好的分类能力，即可能发生过拟合现象
  - “剪枝”是决策树学习算法对付“过拟合”的主要手段
  - 可通过“剪枝”来一定程度避免因决策分支过多，以致于把训练集自身的一些特点当做所有数据都具有的一般性质而导致的过拟合
- 剪枝的基本策略
  - 预剪枝
  - 后剪枝
- 判断决策树泛化性能是否提升的方法
  - 留出法：预留一部分数据用作“验证集”以进行性能评估

剪枝处理

- 采用**留出法**判断决策树的泛化性能
- 预留一部分数据用作“验证集”进行性能评估

表 4.2

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

未剪枝决策树前，用信息增益构造决策树，先计算出所有属性的信息增益值

对于训练集 D, 其中正例占 $p_1 = \frac{5}{10} = \frac{1}{2}$, 反例占 $p_2 = \frac{5}{10} = \frac{1}{2}$, 因此, 信息熵:

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

下面根据表 4.2 计算各个属性的信息增益值.

色泽: D^1\{色泽 = 青绿\} : $(1, 6, 10, 17)$, D^2\{色泽 = 乌黑\} : $(2, 3, 7, 15)$, D^3\{色泽 = 浅白\} : $(14, 16)$, 则

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

$$\mathrm{Ent}(D^2) = -\sum_{k=1}^{2} p_k \log_2 p_k = -\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) = 0.811,$$

$$\mathrm{Ent}(D^3) = -\sum_{k=1}^{2} p_k \log_2 p_k = -(0 + 1\log_2 1) = 0,$$

$$\mathrm{Gain}(D, \text{色泽}) = 1 - \left(\frac{4}{10} * 1 + \frac{4}{10} * 0.811 + 0\right) = 0.276.$$

根蒂: D^1\{根蒂 = 蜷缩\} : $(1, 2, 3, 16, 17)$, D^2\{根蒂 = 稍蜷\} : $(6, 7, 14, 15)$, D^3\{根蒂 = 硬挺\} : (10), 则

$$\mathrm{Ent}(D^1) = -\sum_{k=1}^{2} p_k \log_2 p_k = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0.971,$$

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

$$\mathrm{Ent}(D^3) = -\sum_{k=1}^{2} p_k \log_2 p_k = -(0 + 1\log_2 1) = 0.$$

$$\mathrm{Gain}(D, \text{根蒂}) = 1 - \left(\frac{1}{2} * 0.971 + \frac{2}{5} * 1 + \frac{1}{10} * 0\right) = 0.115.$$

敲声: D^1\{敲声 = 浊响\} : $(1, 3, 6, 7, 15, 16)$, D^2\{敲声 = 沉闷\} : $(2, 14, 17)$, D^3\{敲声 = 清脆\} : (10), 则

$$\mathrm{Ent}(D^1) = -\sum_{k=1}^{2} p_k \log_2 p_k = -\left(\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right) = 0.918,$$

$$\mathrm{Ent}(D^2) = -\sum_{k=1}^{2} p_k \log_2 p_k = -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) = 0.918,$$

$$\mathrm{Ent}(D^3) = -\sum_{k=1}^{2} p_k \log_2 p_k = -(0 + 1\log_2 1) = 0.$$

$$\mathrm{Gain}(D, \text{敲声}) = 1 - \left(\frac{6}{10} * 0.971 + \frac{3}{10} * 0.918 + \frac{1}{10} * 0\right) = 0.174.$$

纹理: D^1\{纹理 = 清晰\} : $(1, 2, 3, 6, 10, 15)$, D^2\{纹理 = 稍糊\} : $(7, 14, 17)$, D^3\{纹理 = 模糊\} : (16), 则

$$\mathrm{Ent}(D^1) = -\sum_{k=1}^{2} p_k \log_2 p_k = -\left(\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right) = 0.918,$$

$$\mathrm{Ent}(D^2) = -\sum_{k=1}^{2} p_k \log_2 p_k = -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) = 0.918,$$

$$\mathrm{Ent}(D^3) = -\sum_{k=1}^{2} p_k \log_2 p_k = -(0 + 1\log_2 1) = 0.$$

$$\mathrm{Gain}(D, \text{纹理}) = 1 - \left(\frac{6}{10} * 0.971 + \frac{3}{10} * 0.918 + \frac{1}{10} * 0\right) = 0.174.$$

脐部: D^1\{脐部 = 凹陷\} : $(1, 2, 3, 14)$, D^2\{脐部 = 稍凹\} : $(6, 7, 15, 17)$, D^3\{脐部 = 平坦\} : $(10, 16)$, 则

$$\mathrm{Ent}(D^1) = -\sum_{k=1}^{2} p_k \log_2 p_k = -\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) = 0.811,$$

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

$$\mathrm{Ent}(D^3) = -\sum_{k=1}^{2} p_k \log_2 p_k = -(0 + 1\log_2 1) = 0,$$

$$\mathrm{Gain}(D, \text{脐部}) = 1 - \left(\frac{4}{10} * 0.811 + \frac{4}{10} * 1 + \frac{1}{10} * 0\right) = 0.276.$$

触感: D^1\{触感 = 硬滑\} : $(1, 2, 3, 16, 17)$, D^2\{触感 = 软粘\} : $(6, 7, 10, 15)$, 则

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

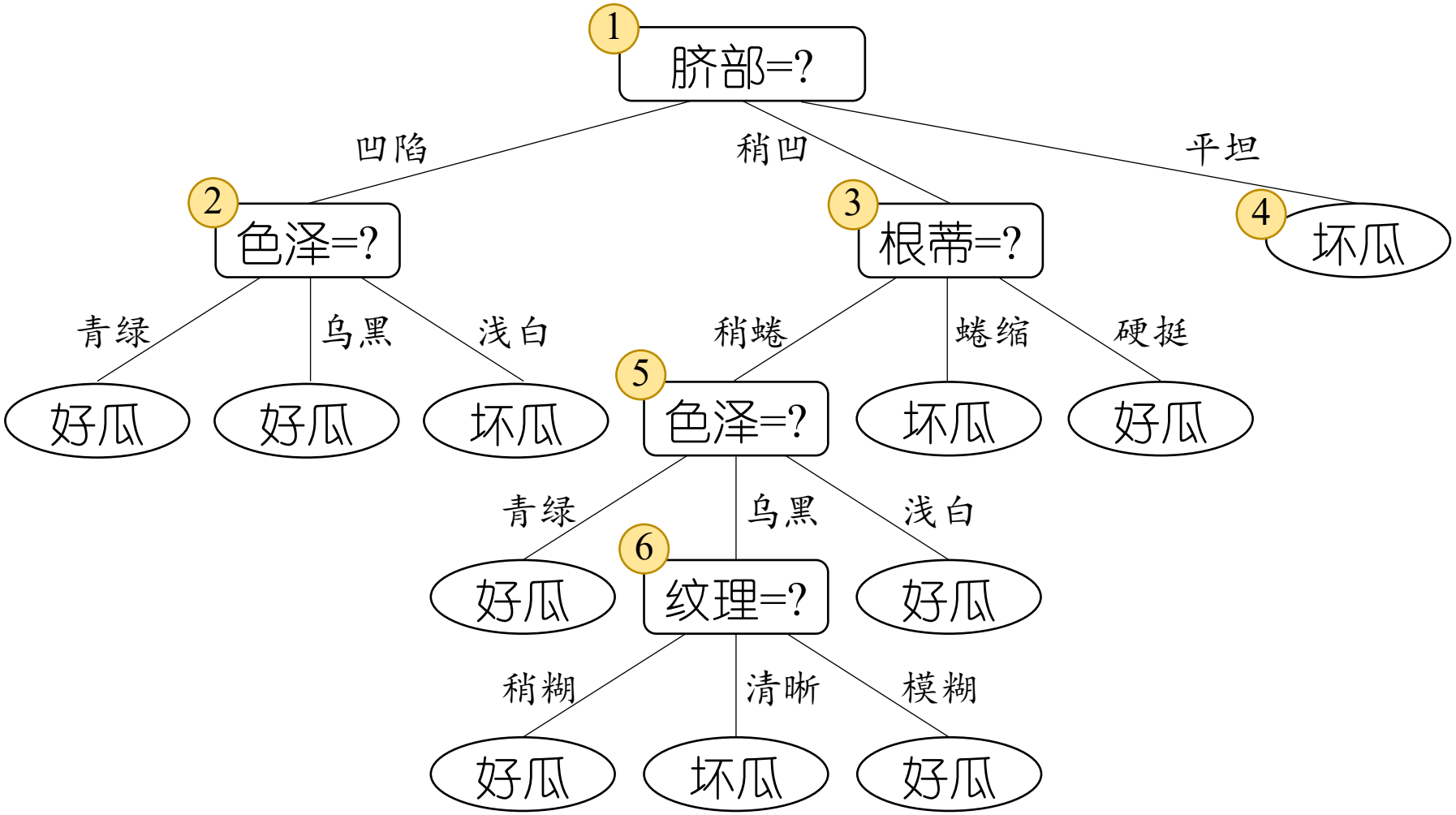
$$\mathrm{Ent}(D^2) = -\sum_{k=1}^{2} p_k \log_2 p_k = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1,$$

$$\mathrm{Ent}(D^3) = -\sum_{k=1}^{2} p_k \log_2 p_k = -(0 + 1\log_2 1) = 0,$$

$$\mathrm{Gain}(D, \text{触感}) = 1 - \left(\frac{6}{10} * 1 + \frac{4}{10} * 1\right) = 0.$$

剪枝处理

- 用信息增益作为划分属性选择，表4.2未剪枝决策树



剪枝处理-预剪枝

- 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记为叶结点，其类别标记为训练样例数最多的类别

- 针对上述数据集，基于信息增益准则，选取属性“脐部”划分训练集。分别计算划分前（即直接将该结点作为叶结点）及划分后的验证集精度，判断是否需要划分。若划分后能提高验证集精度，则划分，对划分后的每一个属性，执行同样判断；否则，不划分

剪枝处理-预剪枝

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

结点1：若不划分，就没有脐部这个根结点，则将其标记为叶结点，类别标记为训练样例中最多的类别，即好瓜，也就是将所有样例都视为正例。验证集中，$\{4,5,8\}$被分类正确，得到验证集精度为 $\frac{3}{7}\times 100\% = 42.9\%$

验证集精度

首先根据训练集计算，脐部信息增益最大，因此，脐部应该构造为决策树的根结点.

1 脐部=?

“脐部=?” 划分前: 42.9%

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

当样例最多的类不唯一时，可任选其中一类

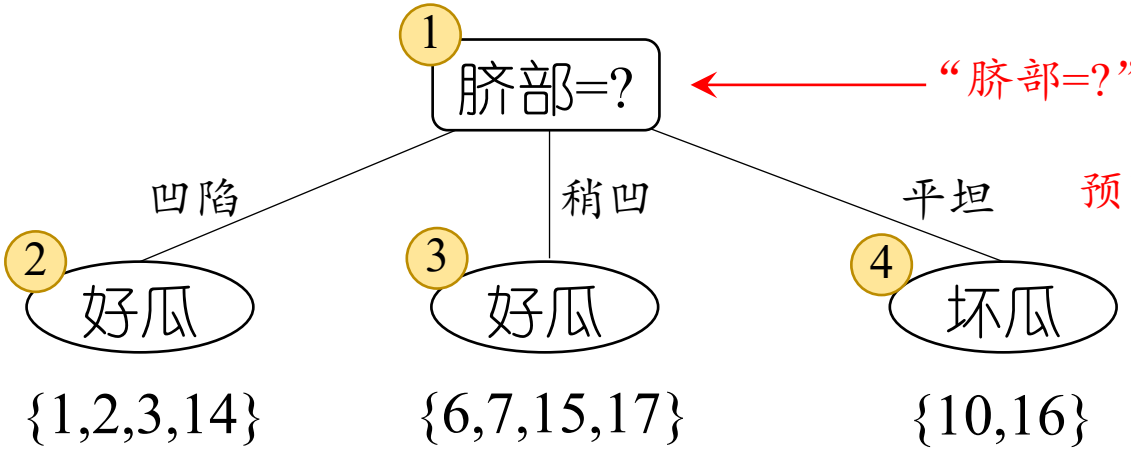
剪枝处理-预剪枝

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

结点1：若划分，根据结点②，③，④ 的训练样例，将这3 个结点分别标记为“好瓜”、“好瓜”、“坏瓜”。此时，验证集中编号为$\{4,5,8,11,12\}$的样例被划分正确，验证集精度为 $\frac{5}{7}\times 100\% = 71.4\%$

验证集精度



“脐部=?” 划分前: 42.9%
划分后: 71.4%
预剪枝决策: 划分

再用选定的属性对结点进行划分，并将划分出来的结点按照其包含最多的类别划分为叶结点（这里就是用“脐部属性”对训练样本进行划分，得到图中结点编号②，③，④ 分别包含训练样本{1，2，3，14}，{6，7，15，17}，{10，16}，并且这三个结点还被分别标记为叶结点“好瓜”，“好瓜”，“坏瓜”）

剪枝处理-预剪枝

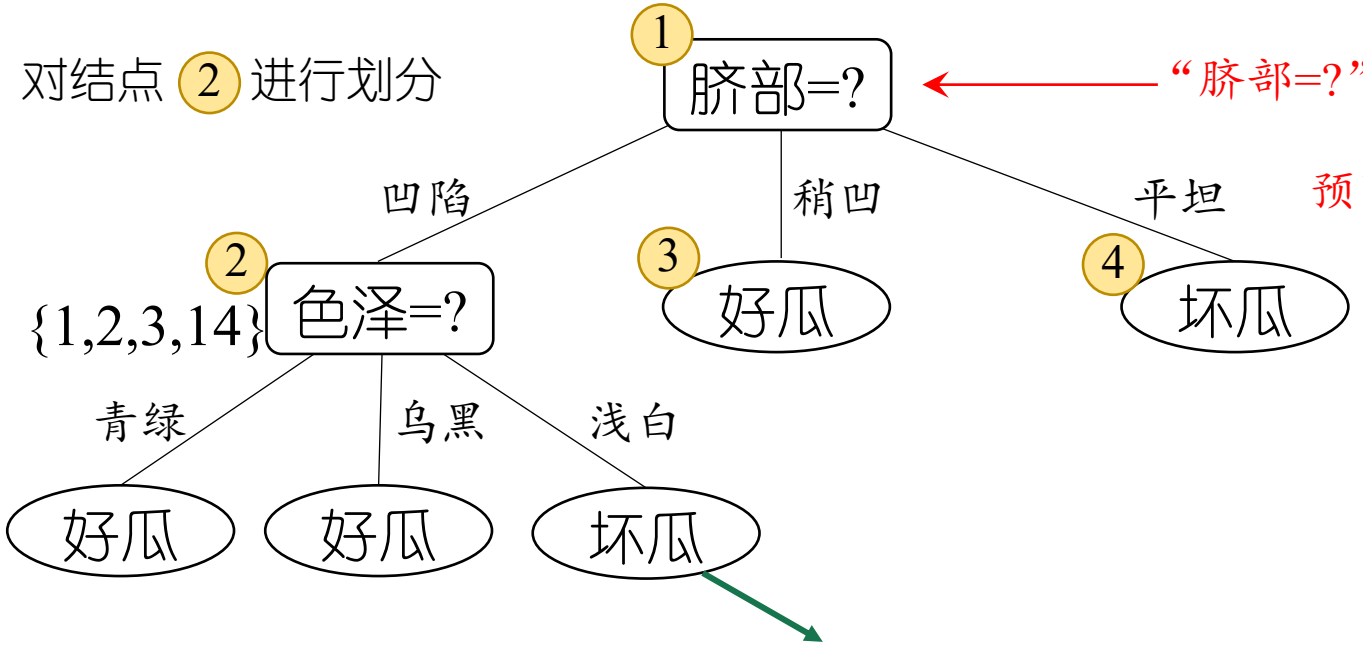
验证集

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

对结点②，③，④分别进行剪枝判断，结点②，③都禁止划分，结点④本身为叶子结点。最终得到仅有一层划分的决策树，称为**“决策树桩”**

验证集精度

对结点 ② 进行划分



“脐部=?” 划分前：42.9%
划分后：71.4%
预剪枝决策：划分

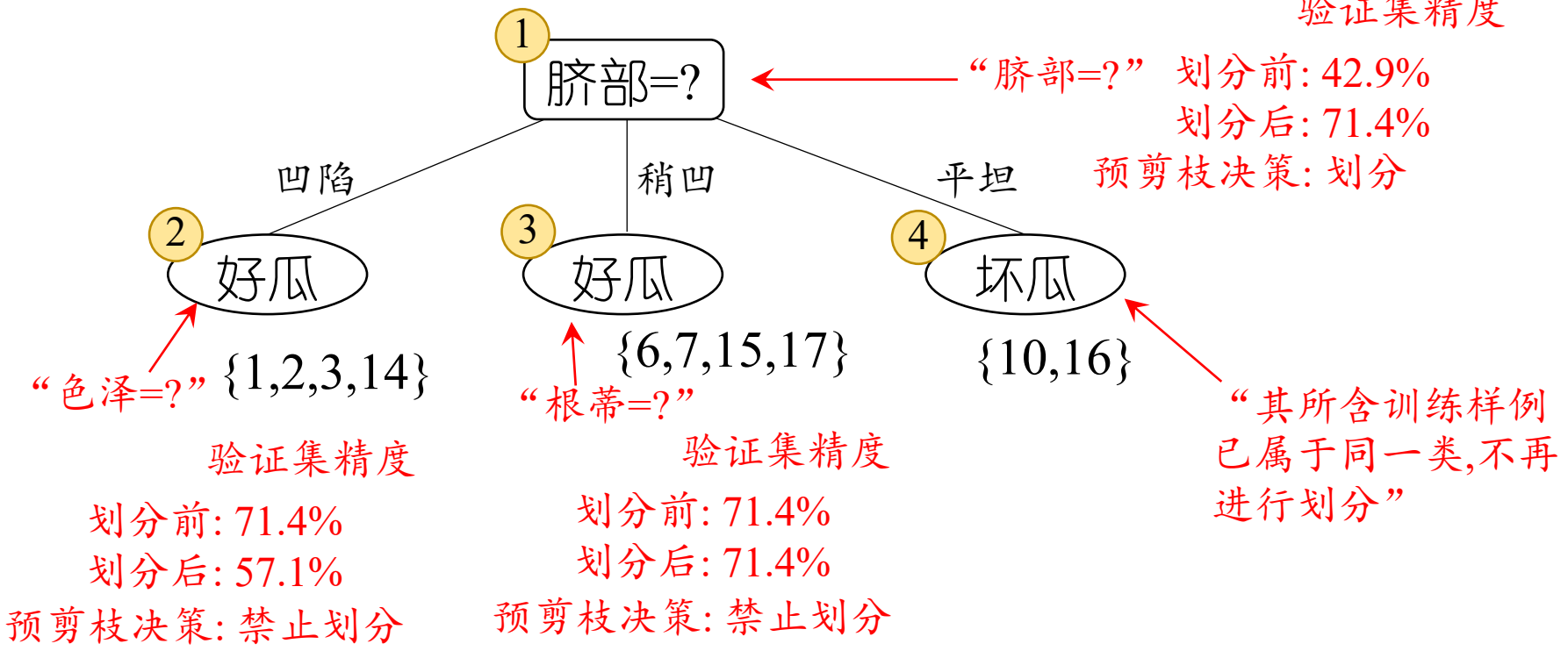
使用色泽划分后，编号为{5}的验证集样本分类结果由正确转为错误，使得验证集精度下降为 $\frac{4}{7}\times 100\%=57.1\%$

剪枝处理-预剪枝

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

对结点②，③，④分别进行剪枝判断，结点②，③都禁止划分，结点④本身为叶子结点。最终得到仅有一层划分的决策树，称为**“决策树桩”**



基于剪枝构造决策树的Python代码见学在西电“资料”栏-python编程参考

剪枝处理-预剪枝

预剪枝的优缺点

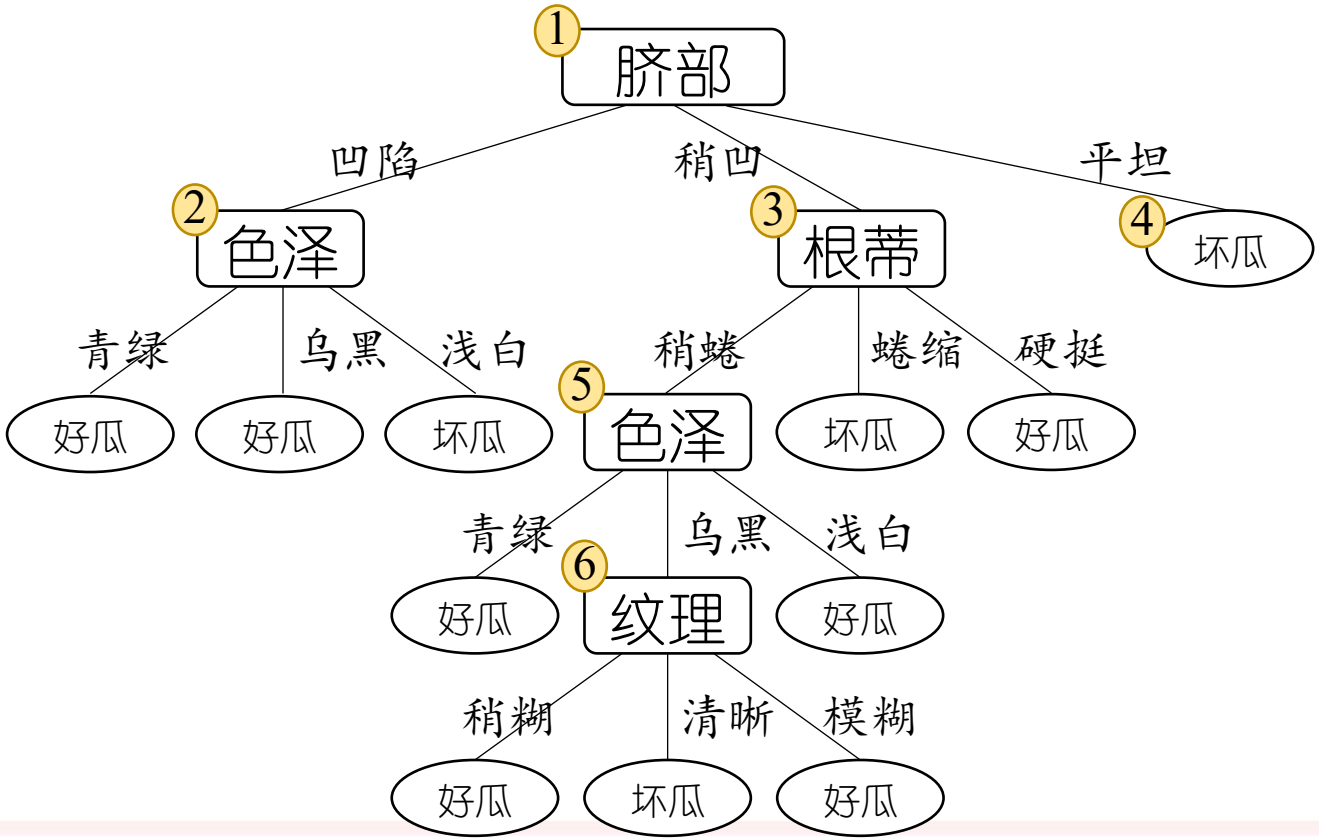
- 优点
  - 预剪枝让决策树的很多分支没有展开，降低了过拟合风险
  - 显著减少训练时间和测试时间开销

- 缺点
  - 欠拟合风险：有些分支的当前划分虽然不能提升泛化性能，但在其基础上进行的后续划分却有可能导致性能显著提高。预剪枝基于“贪心”本质禁止这些分支展开，带来了欠拟合风险

剪枝处理-后剪枝

- 先从训练集生成一棵完整的决策树，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决策树泛化性能提升，则将该子树替换为叶结点

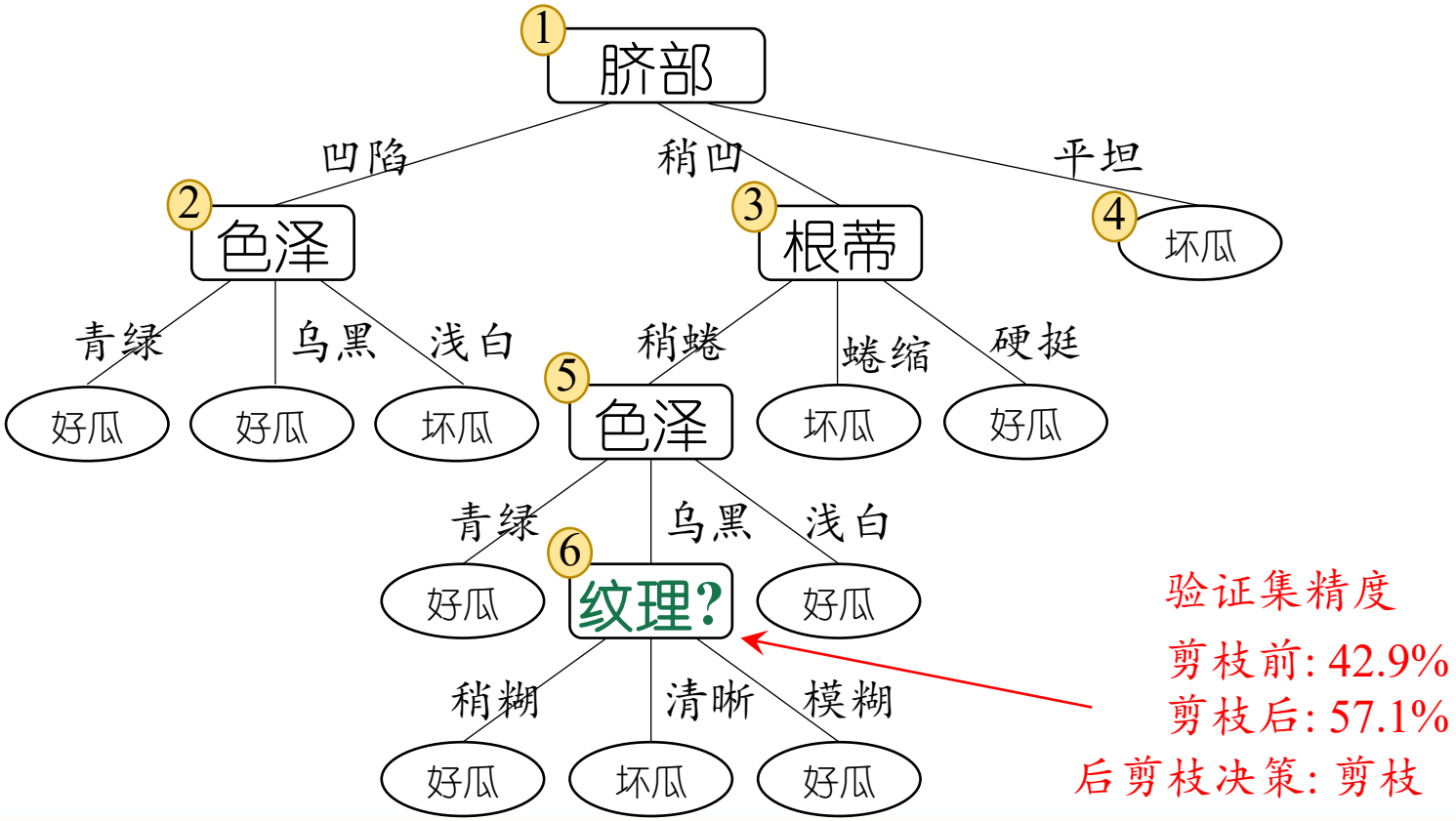
首先生成一棵完整的决策树，该决策树的验证集精度为 42.9%



基于剪枝构造决策树的Python代码见学在西电“资料”栏-python编程参考

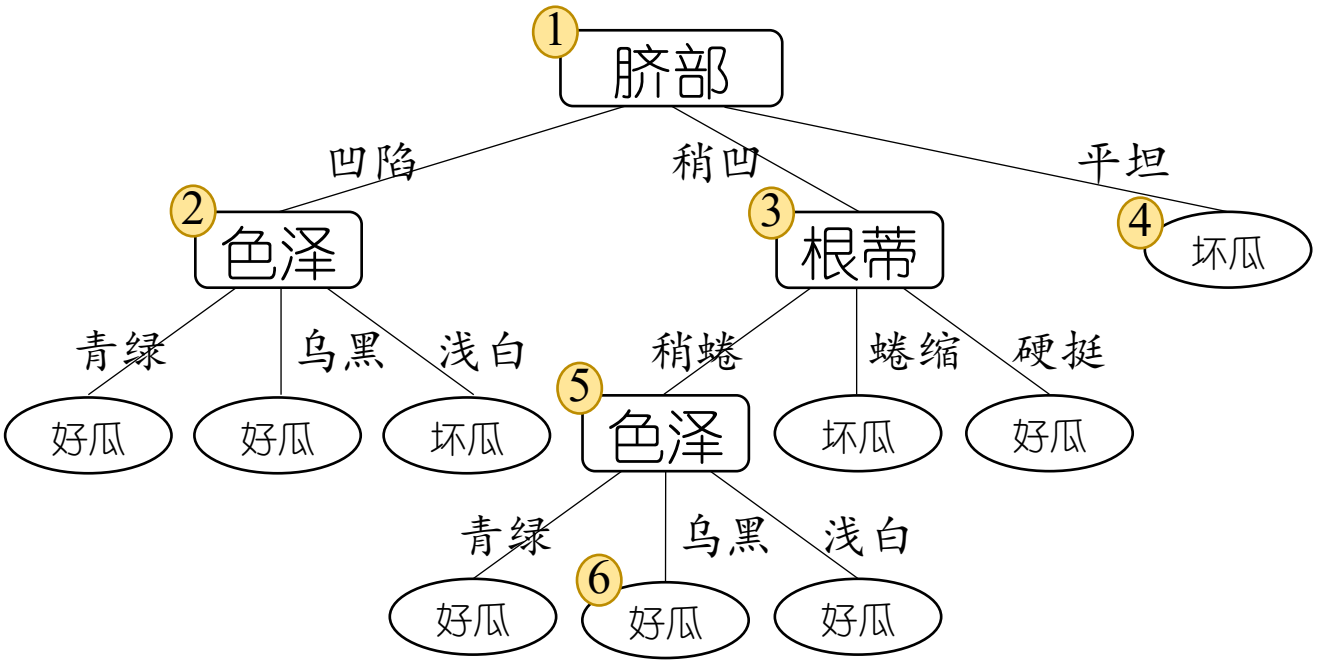
剪枝处理-后剪枝

- 首先考虑结点⑥，若将其替换为叶结点，根据落在其上的训练样本{7，15}将其标记为“好瓜”(因为这里正反类样本数量相等，所以标记为两类中任意一个类别)，得到验证集精度提高至 57.1%，则决定剪枝



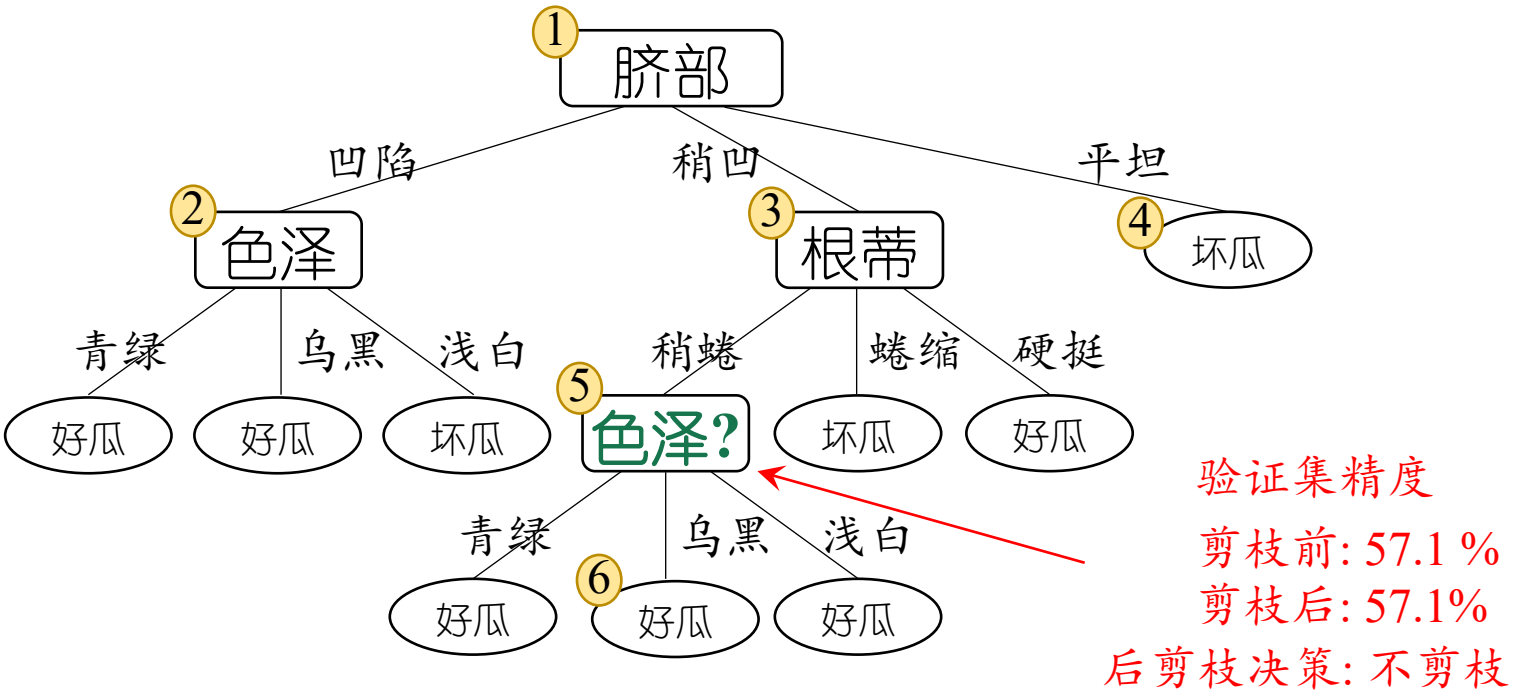
剪枝处理-后剪枝

- 首先考虑结点 ⑥，若将其替换为叶结点，则剪枝后的决策树为



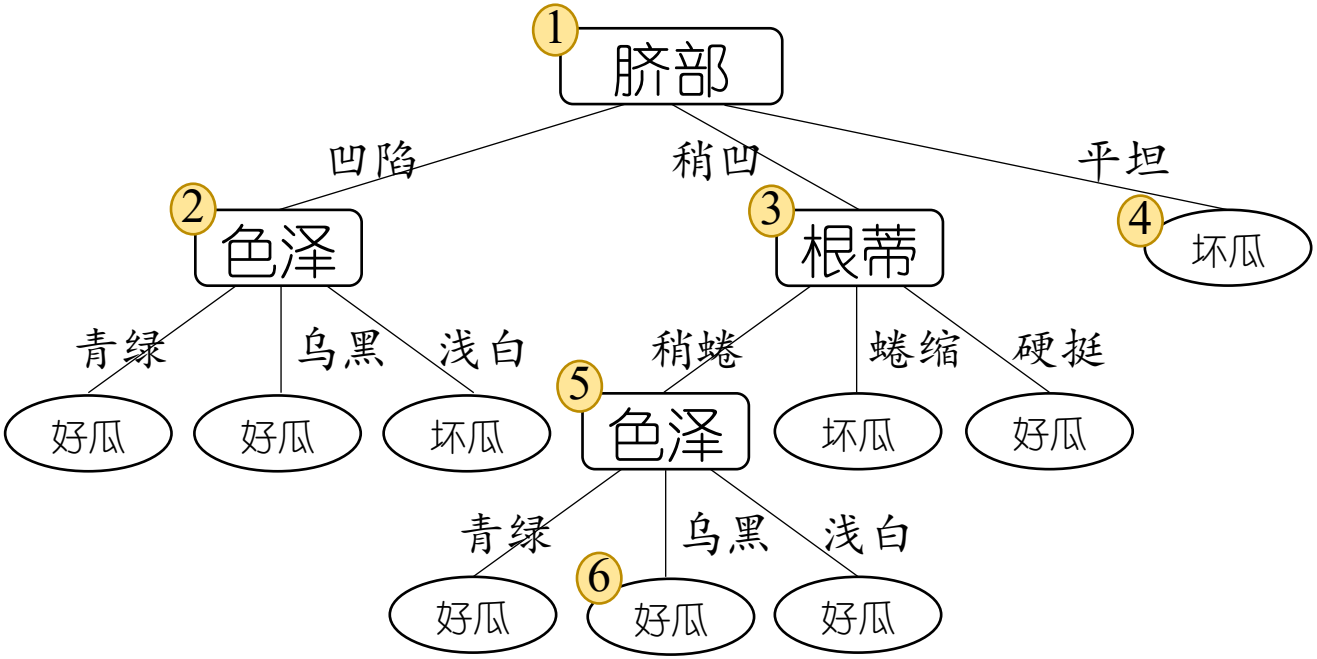
剪枝处理-后剪枝

- 然后考虑结点⑤，若将其替换为叶结点，根据落在其上的训练样本$\{6, 7, 15\}$将其标记为“好瓜”，得到验证集精度仍为57.1%，可以不进行剪枝



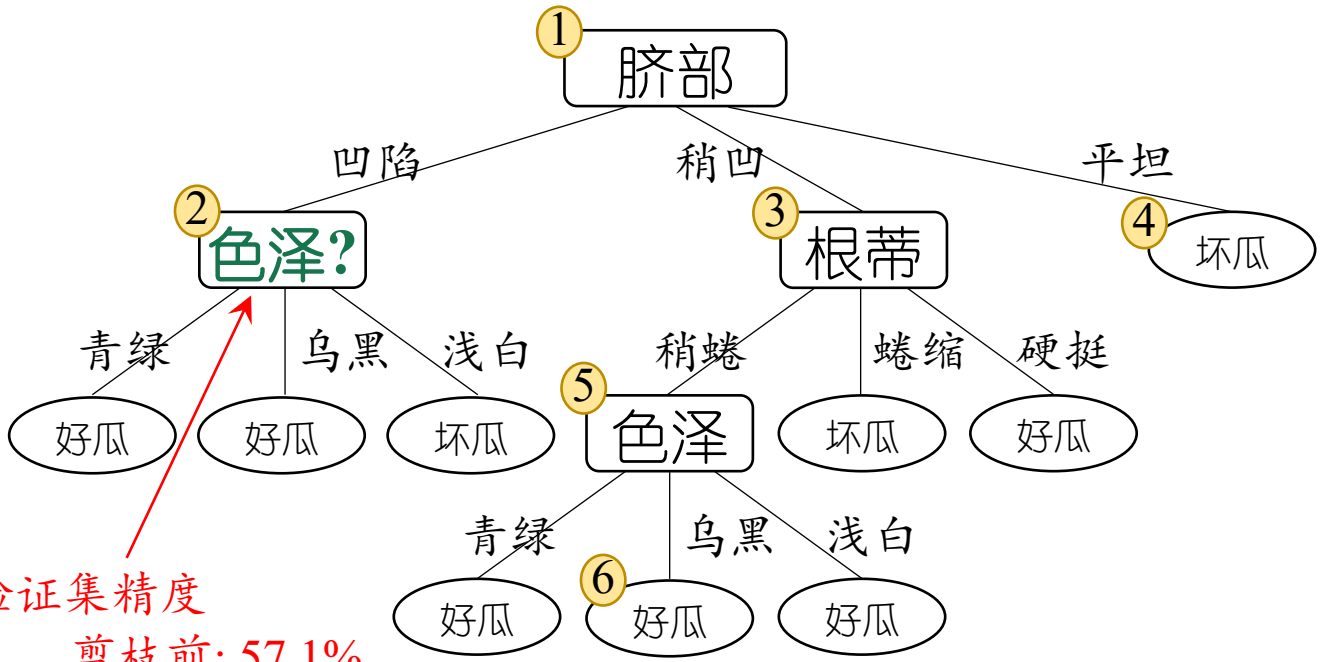
剪枝处理-后剪枝

- 然后考虑结点⑤，若将其替换为叶结点，根据落在其上的训练样本$\{6,7,15\}$将其标记为“好瓜”，得到验证集精度仍为57.1%，可以不进行剪枝



剪枝处理-后剪枝

- 对结点②，若将其替换为叶结点，根据落在其上的训练样本$\{1, 2, 3, 14\}$，将其标记为“好瓜”，得到验证集精度提升至71.4%，则决定剪枝



验证集精度

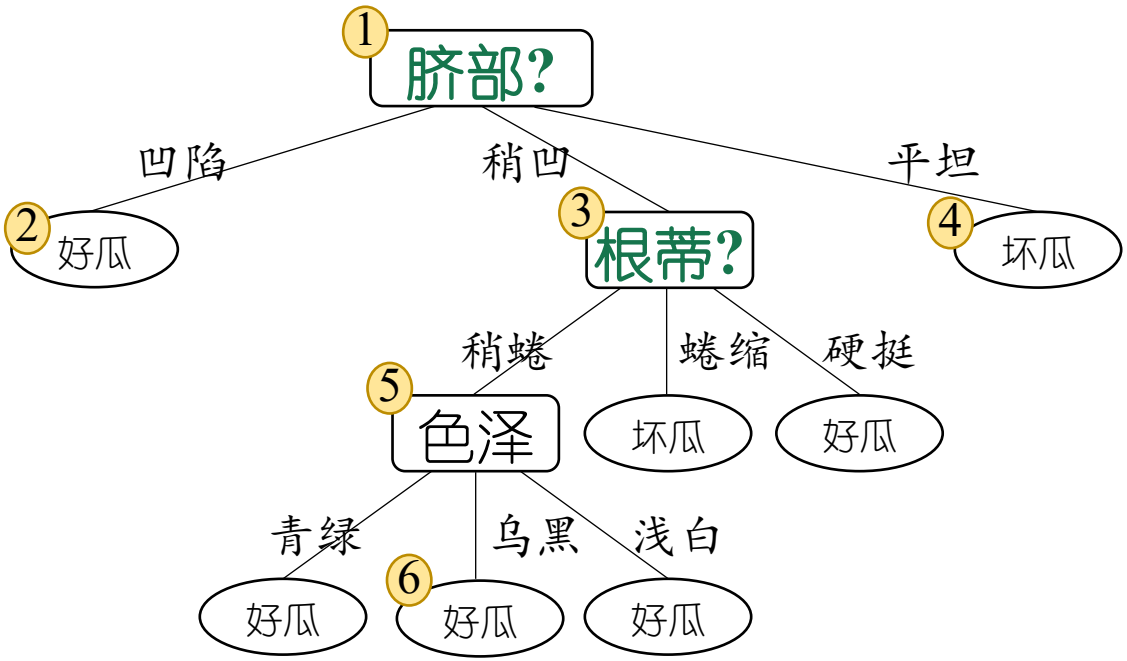
剪枝前: 57.1%

剪枝后: 71.4%

后剪枝决策: 剪枝

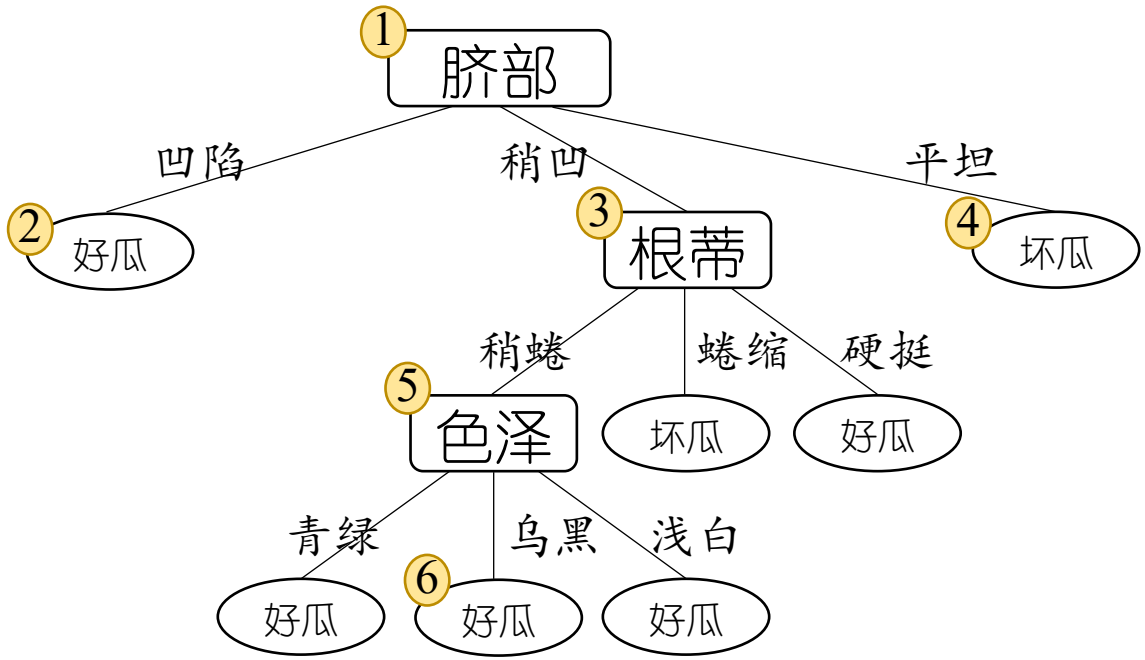
剪枝处理-后剪枝

- 对结点③和①，先后替换为叶结点，验证集精度均未提升，则分支得到保留



剪枝处理-后剪枝

- 最终基于**后剪枝**策略得到的决策树如图所示



剪枝处理-后剪枝

后剪枝的优缺点

- 优点
  - 后剪枝比预剪枝保留了更多的分支，欠拟合风险小，泛化性能往往优于预剪枝决策树

- 缺点
  - 训练时间开销大：后剪枝过程是在生成完全决策树之后进行的，需要自底向上对所有非叶结点逐一考察

大纲

- 基本流程
- 划分选择
- 剪枝处理
- 连续
- 多变量决策树

连续值处理

- 很多现实学习任务中常会遇到连续属性，连续属性的可取值数目不再有限，不能直接根据连续属性的可取值来对结点进行划分

表4.3 西瓜数据集3.0

连续属性

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

连续属性值处理

- 利用连续属性离散化技术-**二分法**，对连续属性进行处理

- 连续属性离散化(二分法)
  - **第一步**：假定连续属性 a 在样本集D上出现 n 个不同的取值，从小到大排列，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^+，其中D_t^-包含那些在属性 a 上取值不大于 t 的样本，$D_t^- = \{a^i | a^i \leq t, i = 1, 2, \cdots, n\}$，$D_t^+$包含那些在属性 a 上取值大于 t 的样本，$D_t^+ = \{a^i | a^i > t, i = 1, 2, \cdots, n\}$。
    构造包含 n-1 个元素的候选划分点集合：

$$T_a = \left\{ \frac{a^i + a^{i+1}}{2} \mid 1 \leq i \leq n-1 \right\}$$

即把区间 $[a^i, a^{i-1})$ 的中位点 $\frac{a^i + a^{i+1}}{2}$ 作为候选划分点 t

划分点 t 可设为该属性在训练集中出现的不大于中位点的最大值，从而使得划分点都在训练集中出现过

连续属性值处理

- 连续属性离散化(二分法)
  - **第二步**：采用离散属性值方法，考察这些划分点，选取最优的划分点进行样本集合的划分

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

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

连续值处理实例

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

下面开始计算 $t \in T_a$ 取不同值时的信息增益：

- 对于连续属性“密度”，决策树开始学习时，根结点包含的17个训练样本在该属性上取值均不同。先把“密度”这些值**从小到大**排序：

$$\{0.243, 0.245, 0.343, 0.360, 0.403, 0.437, 0.481, 0.556, 0.593, 0.608, 0.634, 0.639, 0.657, 0.666, 0.697, 0.719, 0.774\}$$

- 根据上面计算 T_a 的公式，可得16个候选划分点集合：

$$T_{midu} = \{0.244, 0.294, 0.351, 0.381, 0.420, 0.459, 0.518, 0.574, 0.600, 0.621, 0.636, 0.648, 0.661, 0.681, 0.708, 0.746\}$$

- 故需要计算16个信息增益的值，选取信息增益值最大时对应的划分点作为最终划分点

计算 t 取不同值时的信息增益：

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

当 $t = 0.244$ 时：

$D_t^- = \{0.243\}$，$D_t^+ = \{0.245, 0.343, 0.360, 0.403, \ldots\ldots, 0.657, 0.666, 0.697, 0.719, 0.774\}$

$$Ent(D_t^-) = -(0 \cdot \log_2 0 + 1 \cdot \log_2 1) = 0,$$

$$Ent(D_t^+) = -\left(\frac{8}{16} \cdot \log_2\frac{8}{16} + \frac{8}{16} \cdot \log_2\frac{8}{16}\right) = 1,$$

$$\therefore Gain(D, a, t) = Gain(D, 密度, 0.244) = 0.998 - \left(\frac{1}{17} * 0 + \frac{16}{17} * 1\right) = 0.057$$

当 $t = 0.294$ 时：

$D_t^- = \{0.243, 0.245\}$，$D_t^+ = \{0.343, 0.360, 0.403, \ldots\ldots, 0.657, 0.666, 0.697, 0.719, 0.774\}$

$$Ent(D_t^-) = -\left(0 \cdot \log_2 0 + \frac{2}{2} \cdot \log_2\frac{2}{2}\right) = 0,$$

$$Ent(D_t^+) = -\left(\frac{8}{15} \cdot \log_2\frac{8}{15} + \frac{7}{15} \cdot \log_2\frac{7}{15}\right) = 0.997,$$

$$\therefore Gain(D, a, t) = Gain(D, 密度, 0.294) = 0.998 - \left(\frac{2}{17} * 0 + \frac{15}{17} * 0.997\right) = 0.118$$

当 $t = 0.351$ 时：

$D_t^- = \{0.243, 0.245, 0.343\}$，$D_t^+ = \{0.360, 0.403, \ldots\ldots, 0.657, 0.666, 0.697, 0.719, 0.774\}$

$$Ent(D_t^-) = -\left(0 \cdot \log_2 0 + \frac{3}{3} \cdot \log_2\frac{3}{3}\right) = 0,$$

$$Ent(D_t^+) = -\left(\frac{8}{14} \cdot \log_2\frac{8}{14} + \frac{6}{14} \cdot \log_2\frac{6}{14}\right) = 0.985,$$

$$\therefore Gain(D, a, t) = Gain(D, 密度, 0.351) = 0.998 - \left(\frac{3}{17} * 0 + \frac{14}{17} * 0.985\right) = 0.187$$

当 $t = 0.381$ 时：

$D_t^- = \{0.243, 0.245, 0.343, 0.360\}$，$D_t^+ = \{0.403, \ldots\ldots, 0.657, 0.666, 0.697, 0.719, 0.774\}$

$$Ent(D_t^-) = -\left(0 \cdot \log_2 0 + \frac{4}{4} \cdot \log_2\frac{4}{4}\right) = 0,$$

$$Ent(D_t^+) = -\left(\frac{8}{13} \cdot \log_2\frac{8}{13} + \frac{5}{13} \cdot \log_2\frac{5}{13}\right) = 0.961,$$

$$\therefore Gain(D, a, t) = Gain(D, 密度, 0.381) = 0.998 - \left(\frac{4}{17} * 0 + \frac{13}{17} * 0.961\right) = 0.263$$

当 $t = 0.420$ 时：

$D_t^- = \{0.243, 0.245, 0.343, 0.360, 0.403\}$，$D_t^+ = \{0.437, \ldots, 0.657, 0.666, 0.697, 0.719, 0.774\}$

$$Ent(D_t^-) = -\left(\frac{1}{5} \cdot \log_2\frac{1}{5} + \frac{4}{5} \cdot \log_2\frac{4}{5}\right) = 0.722,$$

$$Ent(D_t^+) = -\left(\frac{7}{12} \cdot \log_2\frac{7}{12} + \frac{5}{12} \cdot \log_2\frac{5}{12}\right) = 0.980,$$

$$\therefore Gain(D, a, t) = Gain(D, 密度, 0.420) = 0.998 - \left(\frac{5}{17} * 0.722 + \frac{12}{17} * 0.980\right) = 0.094$$

当 $t = 0.459$ 时：

$D_t^- = \{0.243, 0.245, 0.343, 0.360, 0.403, 0.437\}$，$D_t^+ = \{0.481, \ldots, 0.666, 0.697, 0.719, 0.774\}$

$$Ent(D_t^-) = -\left(\frac{2}{6} \cdot \log_2\frac{2}{6} + \frac{4}{6} \cdot \log_2\frac{4}{6}\right) = 0.918,$$

$$Ent(D_t^+) = -\left(\frac{6}{11} \cdot \log_2\frac{6}{11} + \frac{5}{11} \cdot \log_2\frac{5}{11}\right) = 0.994,$$

$$\therefore Gain(D, a, t) = Gain(D, 密度, 0.459) = 0.998 - \left(\frac{6}{17} * 0.918 + \frac{11}{17} * 0.994\right) = 0.03$$

当 $t = 0.518$ 时：

$D_t^- = \{0.243, 0.245, 0.343, 0.360, 0.403, 0.437, 0.481\}$，$D_t^+ = \{0.556, \ldots, 0.697, 0.719, 0.774\}$

$$Ent(D_t^-) = -\left(\frac{3}{7} \cdot \log_2\frac{3}{7} + \frac{4}{7} \cdot \log_2\frac{4}{7}\right) = 0.985,$$

$$Ent(D_t^+) = -\left(\frac{5}{10} \cdot \log_2\frac{5}{10} + \frac{5}{10} \cdot \log_2\frac{5}{10}\right) = 1,$$

$$\therefore Gain(D, a, t) = Gain(D, 密度, 0.518) = 0.998 - \left(\frac{7}{17} * 0.985 + \frac{10}{17} * 1\right) = 0.004$$

当 $t = 0.574$ 时：

$D_t^- = \{0.243, 0.245, 0.343, 0.360, 0.403, 0.437, 0.481, 0.574\}$，$D_t^+ = \{0.593, \ldots, , 0.719, 0.774\}$

$$Ent(D_t^-) = -\left(\frac{4}{8} \cdot \log_2\frac{4}{8} + \frac{4}{8} \cdot \log_2\frac{4}{8}\right) = 1,$$

$$Ent(D_t^+) = -\left(\frac{4}{9} \cdot \log_2\frac{4}{9} + \frac{5}{9} \cdot \log_2\frac{5}{9}\right) = 0.991,$$

$$\therefore Gain(D, a, t) = Gain(D, 密度, 0.574) = 0.998 - \left(\frac{8}{17} * 1 + \frac{9}{17} * 0.991\right) = 0.002$$

当 $t = 0.600$，$t = 0.621$，$t = 0.636$，$t = 0.648$，……就不在展示详细的计算过程了。

比较能够发现，当 $t = 0.381$ 时，$Gain(D, a, t)$ 最大为 0.263。因此选择该划分点。

对于属性“含糖率”，按照同样的方法能够计算出，$t = 0.126$，$Gain(D, a, t) = 0.349$。

连续值处理

- 对属性“**密度**”，其候选划分点集合包含 16 个候选值：

$$T_{\text{密度}} = \{0.244, 0.294, 0.351, 0.381, 0.420, 0.459, 0.518, 0.574, 0.600, 0.621, 0.636, 0.648, 0.661, 0.681, 0.708, 0.746\}$$

计算其最大信息增益为 0.262，对应划分点为 t = 0.381

- 对属性“**含糖量**”进行同样处理，最大信息增益为 0.349，对应划分点 t = 0.126

表4.3 西瓜数据集3.0

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

- 表4.3的数据上的各属性的信息增益为

$$\text{Gain}(D, \text{色泽}) = 0.109; \quad \text{Gain}(D, \text{根蒂}) = 0.143;$$
$$\text{Gain}(D, \text{敲声}) = 0.141; \quad \text{Gain}(D, \text{纹理}) = 0.381;$$
$$\text{Gain}(D, \text{脐部}) = 0.289; \quad \text{Gain}(D, \text{触感}) = 0.006;$$
$$\text{Gain}(D, \text{密度}) = 0.262; \quad \text{Gain}(D, \text{含糖率}) = 0.349.$$

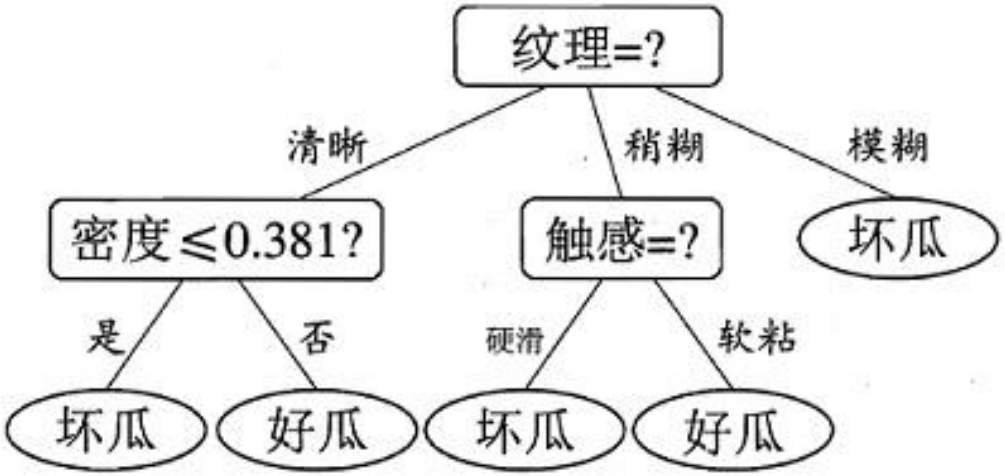
“纹理”被选为根结点的划分属性，此后结点划分过程递归进行

连续值处理

表4.3的数据上的各属性的信息增益为

$$\text{Gain}(D, \text{色泽}) = 0.109; \quad \text{Gain}(D, \text{根蒂}) = 0.143;$$
$$\text{Gain}(D, \text{敲声}) = 0.141; \quad \text{Gain}(D, \text{纹理}) = 0.381;$$
$$\text{Gain}(D, \text{脐部}) = 0.289; \quad \text{Gain}(D, \text{触感}) = 0.006;$$
$$\text{Gain}(D, \text{密度}) = 0.262; \quad \text{Gain}(D, \text{含糖率}) = 0.349.$$

“纹理”被选为根结点的划分属性，此后结点划分过程递归重复进行。最终生成的决策树如下：

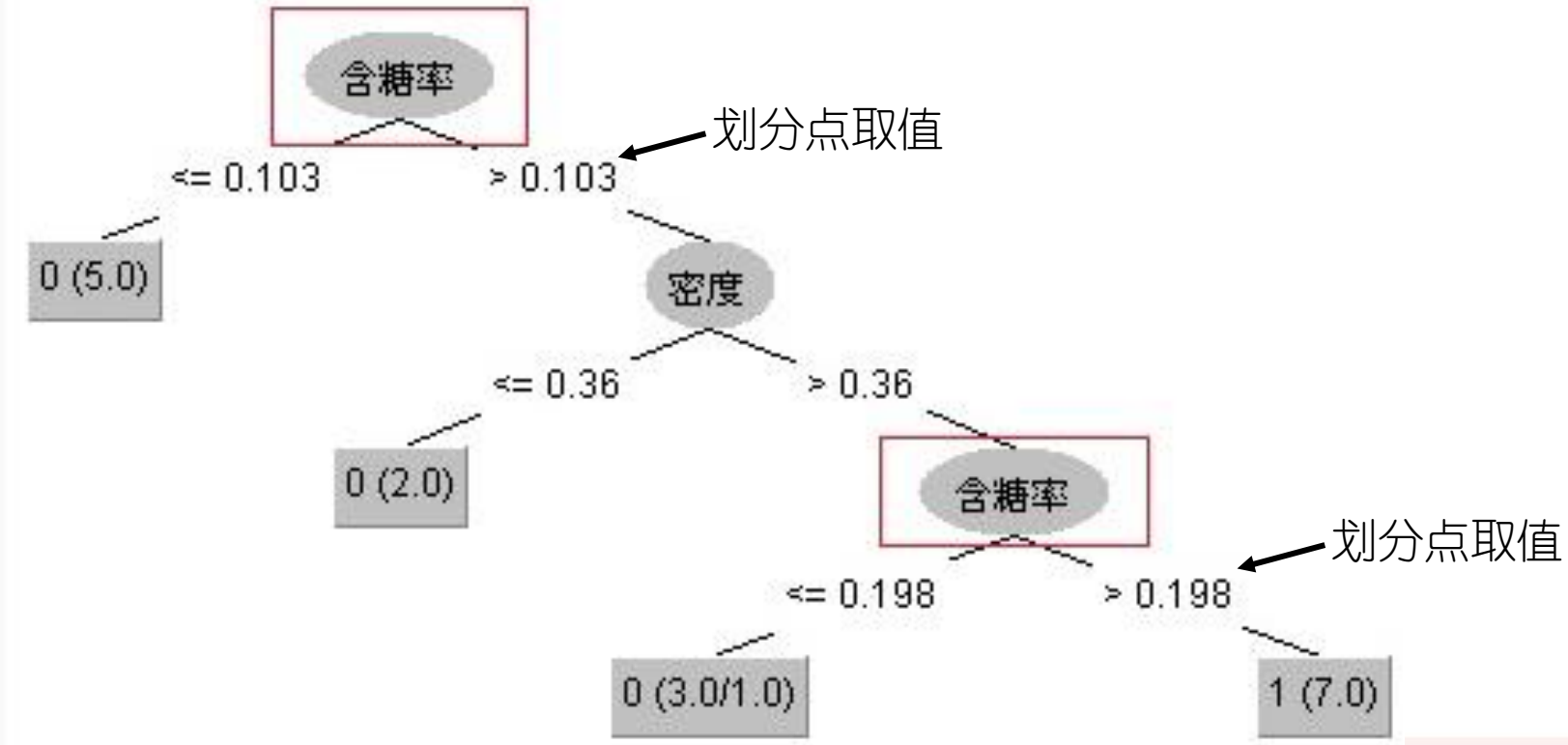


含连续值的属性构造决策树的Python代码见学在西电“资料”栏-python编程参考

连续值处理

- 与离散属性不同，**若当前结点划分属性为连续属性，该属性还可作为其后代结点的划分属性。**

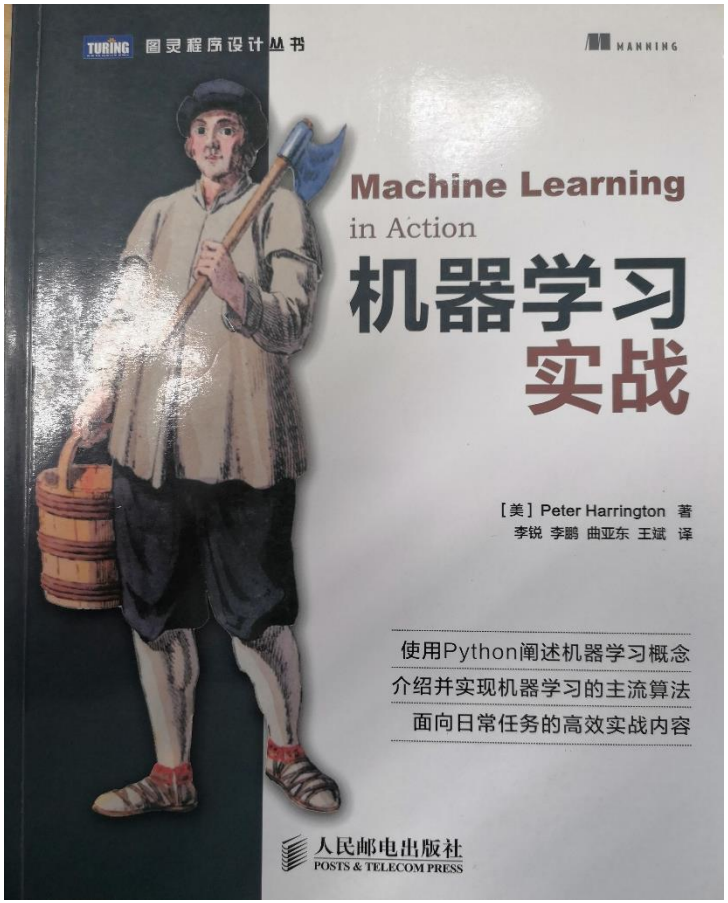
如下图所示的一颗决策树，“含糖率”这个属性在根结点用了一次，后代结点也用了一次，只是两次划分点取值不同。

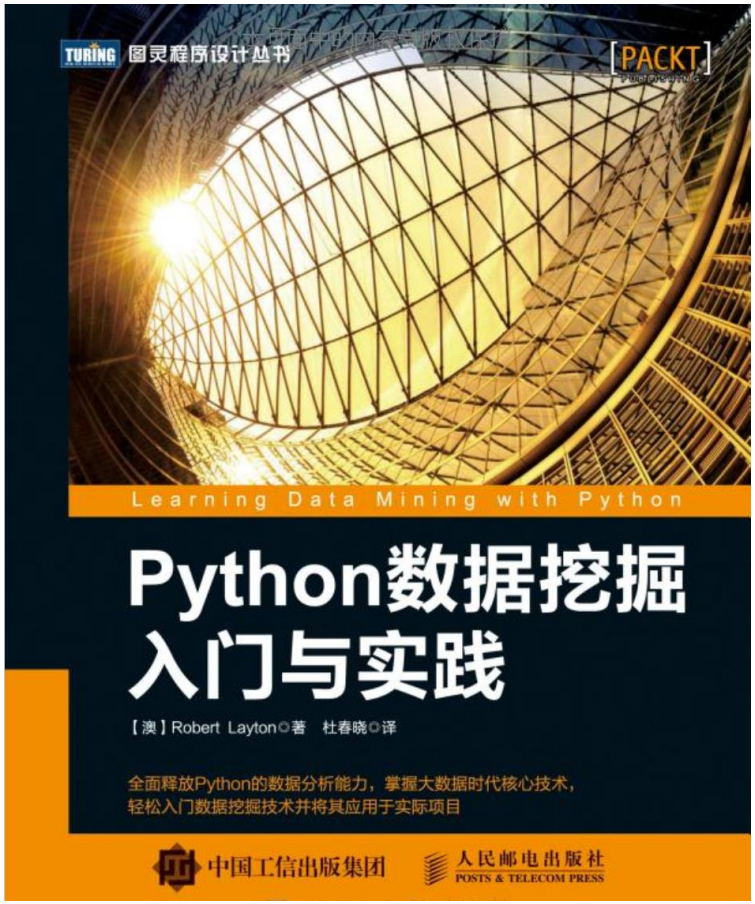


作业三

- 见“学在西电”资料-作业要求

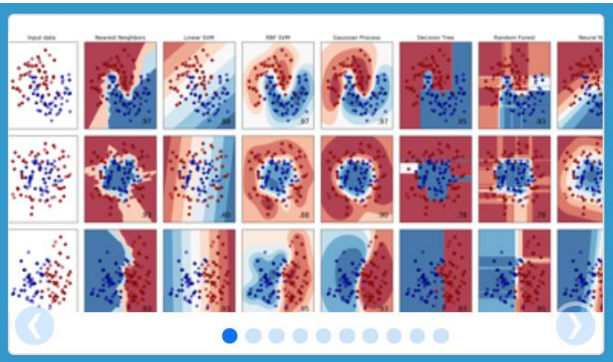
Python编程实践


TURING 图灵程序设计丛书
MANNING
Machine Learning
in Action
机器学习
实战
[美] Peter Harrington 著
李锐 李鹏 曲亚东 王斌 译
使用Python阐述机器学习概念
介绍并实现机器学习的主流算法
面向日常任务的高效实战内容
人民邮电出版社
POSTS & TELECOM PRESS


TURING 图灵程序设计丛书
[PACKT]
Learning Data Mining with Python
Python数据挖掘
入门与实践
[澳] Robert Layton 著 杜春晓 译
全面释放Python的数据分析能力，掌握大数据时代核心技术，
轻松入门数据挖掘技术并将其应用于实际项目
中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

Python编程实践

https://scikit-learn.org/stable/



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, … — Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, … — Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, … — Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: PCA, feature selection, non-negative matrix factorization. — Examples

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning
Modules: grid search, cross validation, metrics. — Examples

Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.
Modules: preprocessing, feature extraction. — Examples

下次课-《神经网络》