机器学习

李成龙

安徽大学人工智能学院
“多模态认知计算”安徽省重点实验室
合肥综合性国家科学中心人工智能研究院

内容安排

- 什么是机器学习
- 机器如何学习
- 如何让机器学习的更好
- 为什么机器能学习

内容安排

- **机器如何学习**
  - 有监督学习
    - 感知机
    - 支持向量机
    - 朴素贝叶斯分类
    - 决策树
    - 集成学习（Bagging算法与随机森林、Boosting算法）
    - 线性回归
    - 逻辑回归
    - Softmax回归
    - 神经网络与深度学习
  - 无监督学习
    - 聚类
    - 主成分分析

本节目录

本节目录

- **背景知识**
- 模型结构
- 学习算法
- 剪枝处理
- 特殊属性处理

背景知识

- **归纳学习**
  - 归纳是从特殊到一般的过程
  - 归纳推理从若干个事实中表征出的特征、特性和属性中，通过比较、总结、概括而得出一个规律性的结论
  - 归纳推理试图从对象的一部分或整体的特定的观察中获得一个完备且正确的描述。即从特殊事实到普遍性规律的结论
  - 归纳对于认识的发展和完善具有重要的意义。人类知识的增长主要来源于归纳学习

背景知识

- **归纳学习**
  - 归纳学习由于依赖于检验数据，因此又称为检验学习
  - 归纳学习存在一个基本的假设
    - 任一假设如果能够在足够大的训练样本集中很好的逼近目标函数，则它也能在未见样本中很好地逼近目标函数。该假定是归纳学习的有效性的前提条件
  - 归纳过程就是在描述空间中进行搜索的过程
  - 归纳可分为自顶向下，自底向上和双向搜索三种方式
    - 自底向上法一次处理一个输入对象，将描述逐步一般化，直到最终的一般化描述
    - 自顶向下法对可能的一般性描述集进行搜索，试图找到一些满足一定要求的最优的描述

背景知识

- **决策树**
  - 决策树技术发现数据模式和规则的核心是归纳学习
  - 决策树是一种典型的分类方法
    - 对数据进行处理，利用归纳学习生成可读的规则和决策树
    - 使用决策对新数据进行分析
  - 本质上决策树是通过一系列规则对数据进行分类的过程
  - 决策树的优点
    - 推理过程容易理解，决策推理过程可以表示成If Then形式
    - 推理过程完全依赖于属性变量的取值特点
    - 可自动忽略目标变量没有贡献的属性变量，也为判断属性变量的重要性，减少变量的数目提供参考

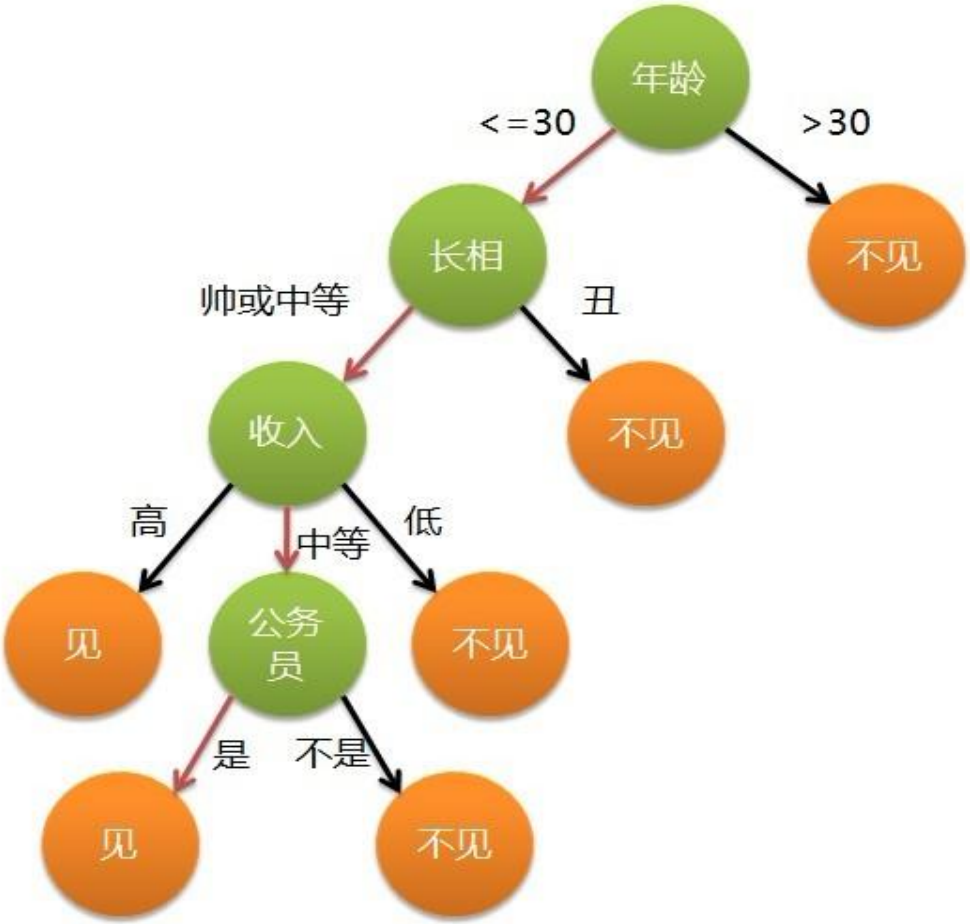
本节目录

模型结构

- 例子
  - 套用俗语，决策树分类的思想类似于找对象。现想象一个女孩的母亲要给这个女孩介绍男朋友，于是有了下面的对话：
    - 女儿：多大年纪了？
      母亲：26。
      女儿：长的帅不帅？
      母亲：挺帅的。
      女儿：收入高不？
      母亲：不算很高，中等情况。
      女儿：是公务员不？
      母亲：是，在税务局上班呢。
      女儿：那好，我去见见。

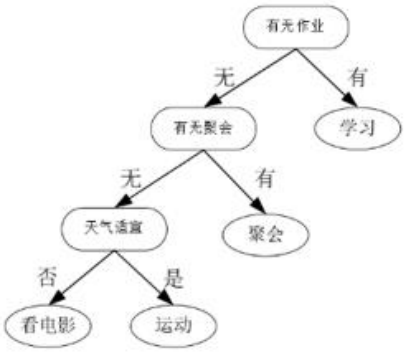
模型结构

- 例子

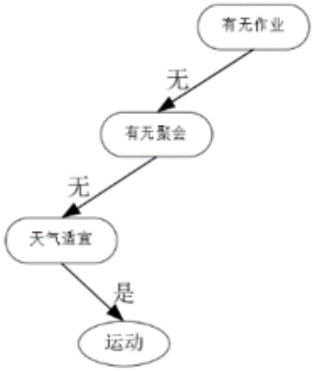


模型结构

- **决策树是解决分类问题的一般方法**
  - 基本思想：模拟人类进行级联选择或决策的过程，按照属性的某个优先级依次对数据的全部属性进行判别，从而得到输入数据所对应的预测输出
  - 模型结构：一个根结点、若干内部结点和叶结点
    - 中叶结点表示决策的结果
    - 内部结点表示对样本某一属性判别
  - 测试序列：从根结点到某一叶子结点的路径



决策树模型



测试序列

模型结构

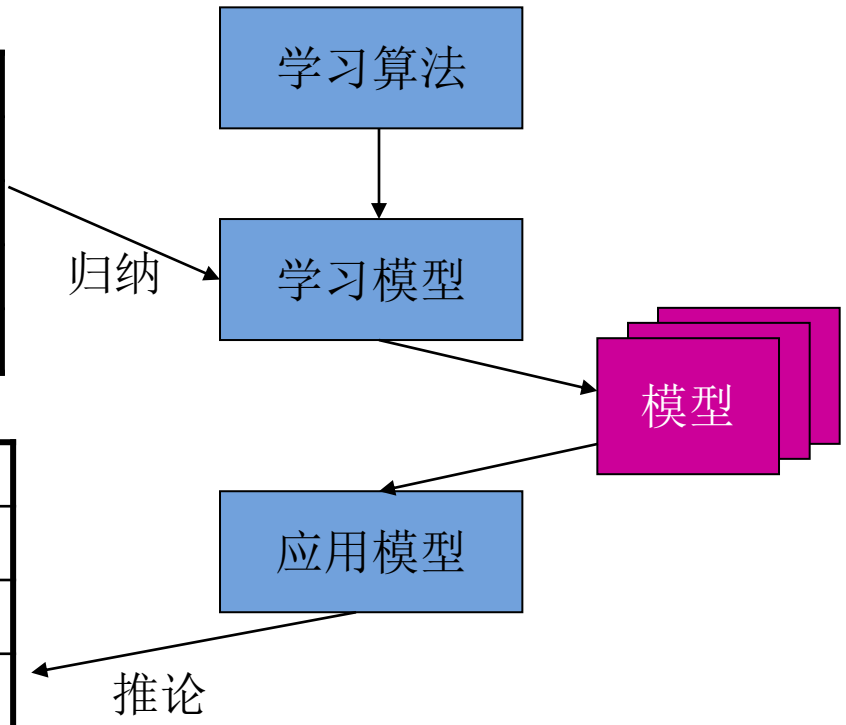
- **决策树是解决分类问题的一般方法**
  - 决策树解决分类问题一般包含两个步骤
    - 模型构建（归纳）：通过对训练集合的归纳，建立分类模型
    - 预测应用（推论）：根据分类模型，对测试集合进行测试

训练集（类标记已知）

TID	A1	A2	A3	类
1	Y	100	L	N
2	N	125	S	N
3	Y	400	L	Y
4	N	415	M	N

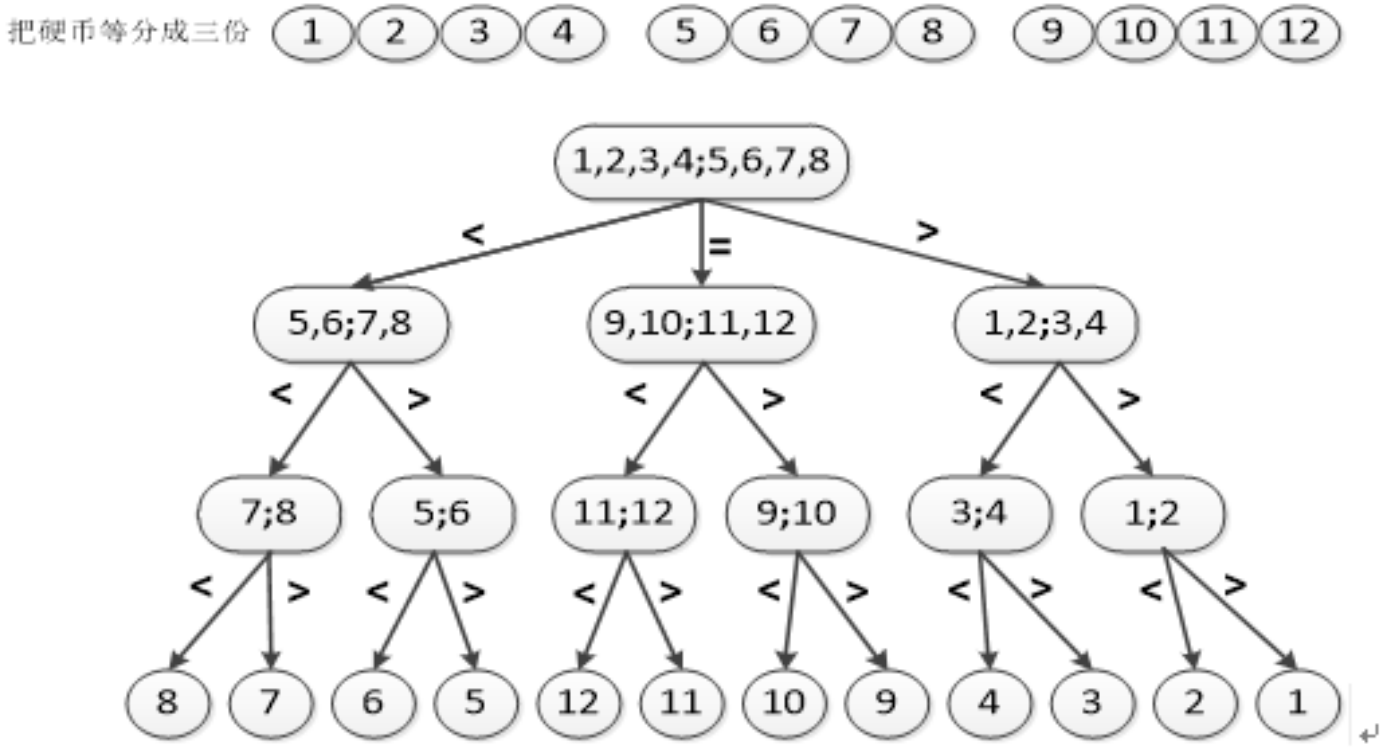
检验集（类标记未知）

TID	A1	A2	A3	类
1	Y	100	L	?
2	N	125	S	?
3	Y	400	L	?
4	N	415	M	?



模型结构

- **例题**：现有12枚外观相同的硬币，其中有1枚是假的且比真币重。如何使用一个无砝码天平把假币找出来，要求不超过三次称重



本节目录

学习算法

- **判别标准**
  - 构造决策树的关键：如何选择最优划分属性。合理选择其内部结点所对应的样本属性，使得结点所对应样本子集中的样本尽可能多地属于同一类别，即具有尽可能高的纯度
    - 如果结点对应数据子集中的样本基本属于同一个类别，则无需对结点的数据子集做进一步划分，否则就要对该结点的数据子集做进一步划分，生成新的判别标准
    - 如果新判别标准能够基本上把结点上不同类别的数据分离开，使得每个子结点都是类别比较单一的数据，那么该判别标准就是一个好规则，否则需重新选取判别标准

学习算法

- **ID3算法**
  - “信息熵”是度量样本集合纯度最常用的一种指标，假定当前样本集合D中第k类样本所占的比为$p_k(K=1,2,...,|\mathcal{Y}|)$则$D$的信息熵定义为

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

$\mathrm{Ent}(D)$的值越小，则D的纯度越高

  - 计算信息熵时约定：若$p=0$，则$p\log_2 p=0$
  - $\mathrm{Ent}(D)$的最小值为0，最大值为$\log_2|\mathcal{Y}|$

学习算法

- **ID3算法**
  - 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”：

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

为分支结点权重，样本数越多的分支结点的影响越大

  - 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
  - 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

学习算法

- **ID3算法**
  - 例题

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

该数据集包含17个训练样本，$|\mathcal{Y}| = 2$，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

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

学习算法

- **ID3算法**
  - 以属性“色泽”为例，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$p_2 = \frac{3}{6}$， D^2 、 D^3同理，3个结点的信息熵为

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

$$\mathrm{Ent}(D^1) = -(\tfrac{3}{6}\log_2\tfrac{3}{6} + \tfrac{3}{6}\log_2\tfrac{3}{6}) = 1.000 \qquad \mathrm{Ent}(D^2) = -(\tfrac{4}{6}\log_2\tfrac{4}{6} + \tfrac{2}{6}\log_2\tfrac{2}{6}) = 0.918$$

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

  - 属性“色泽”的信息增益为

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

学习算法

- **ID3算法**
  - 其他属性的信息增益为

$$\text{Gain}(D, \text{根蒂}) = 0.143$$

$$\text{Gain}(D, \text{敲声}) = 0.141$$

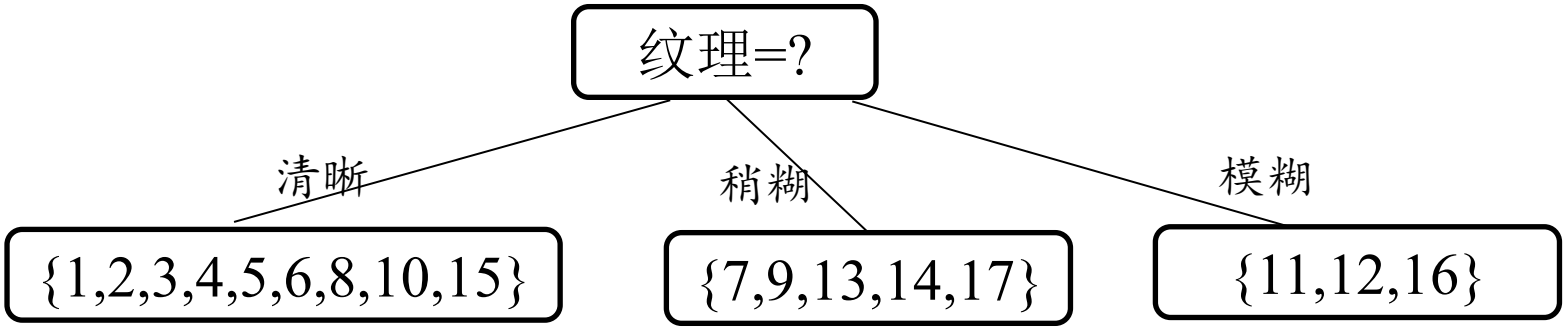
$$\text{Gain}(D, \text{脐部}) = 0.289$$

$$\text{Gain}(D, \text{纹理}) = 0.381$$

$$\text{Gain}(D, \text{触感}) = 0.006$$

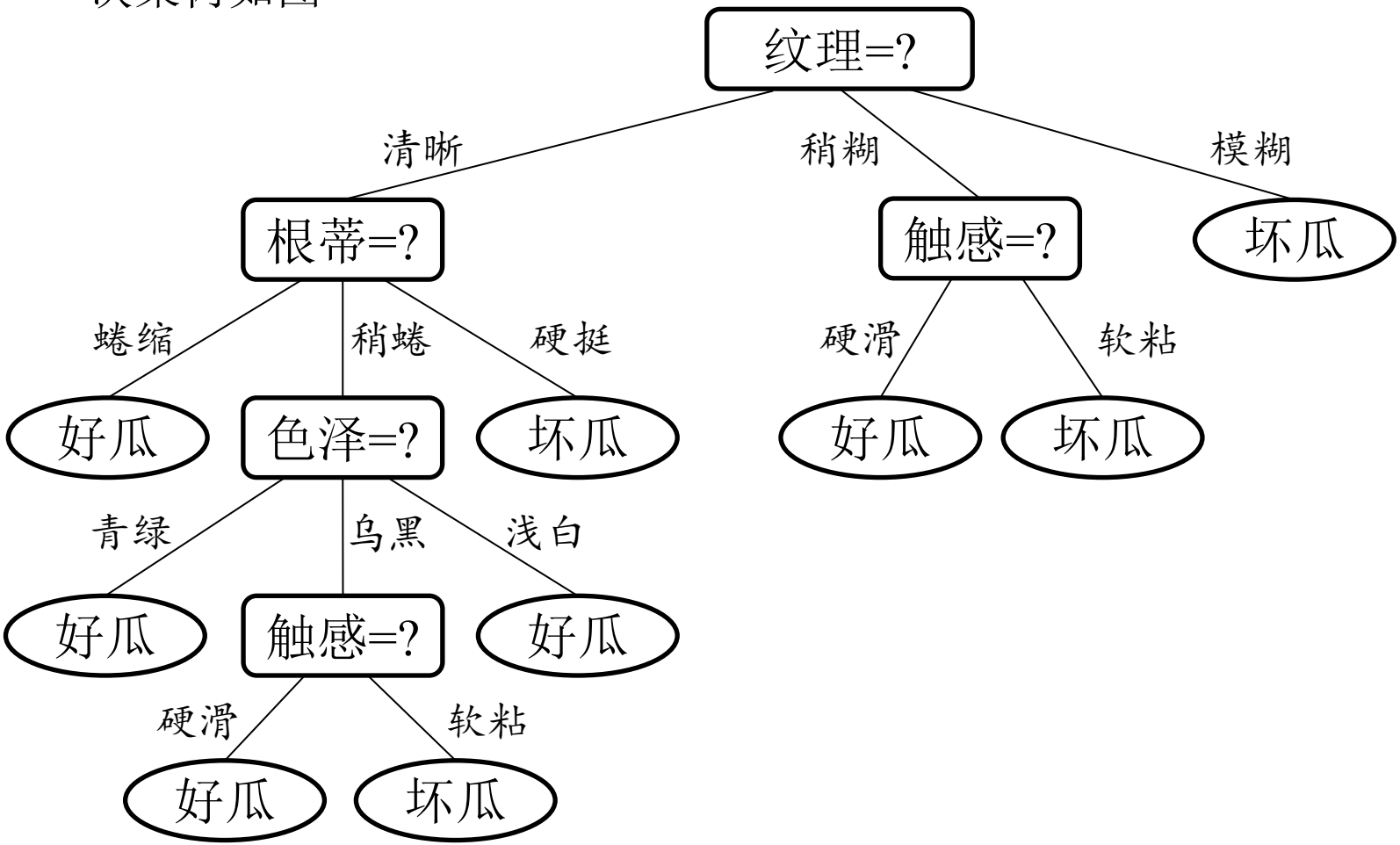
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	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

  - 显然，属性“纹理”的信息增益最大，其被选为划分属性



学习算法

- **ID3算法**
  - ID3决策树学习算法将对每个分支结点做进一步划分，最终得到的决策树如图



学习算法

- **ID3算法**
  - 存在问题
    - 若把“编号”也作为一个候选划分属性，则其信息增益一般远大于其他属性。显然，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

学习算法

- **C4.5算法**
  - 增益率

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

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 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

学习算法

- **CART算法**
  - 数据集 的纯度可用“基尼值”来度量

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

D的基尼值越小，数据集D的纯度越高

  - 属性a的基尼指数定义为

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

  - 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

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

  - CART [Breiman et al., 1984]采用“基尼指数”来选择划分属性

本节目录

剪枝处理

- 为什么要剪枝
  - “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

- 数据集

训练集

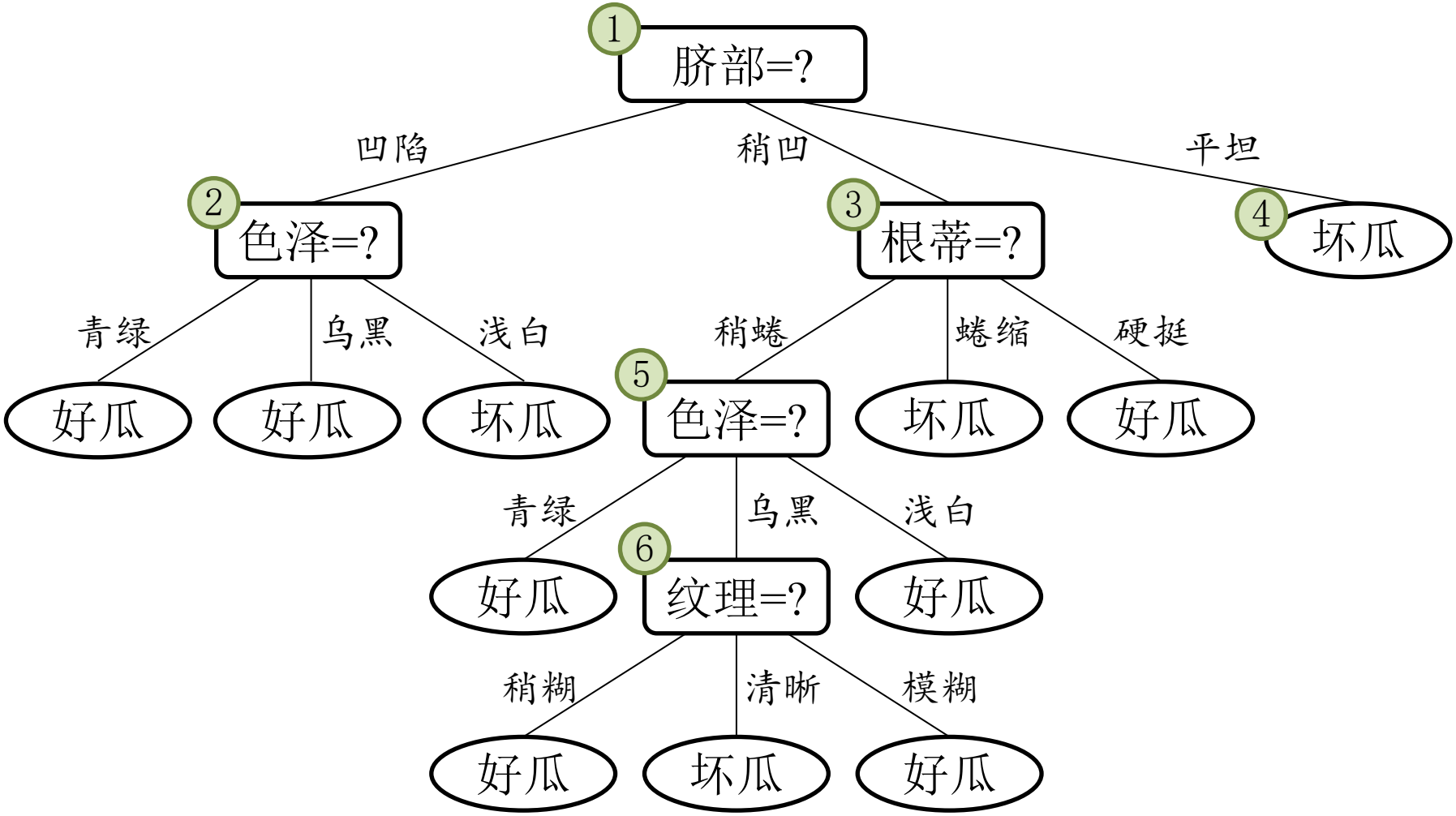
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

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

- 未剪枝决策树


1
脐部=?
凹陷
稍凹
平坦
2
色泽=?
3
根蒂=?
4
坏瓜
青绿
乌黑
浅白
稍蜷
蜷缩
硬挺
好瓜
好瓜
坏瓜
5
色泽=?
坏瓜
好瓜
青绿
乌黑
浅白
6
好瓜
纹理=?
好瓜
稍糊
清晰
模糊
好瓜
坏瓜
好瓜

剪枝处理

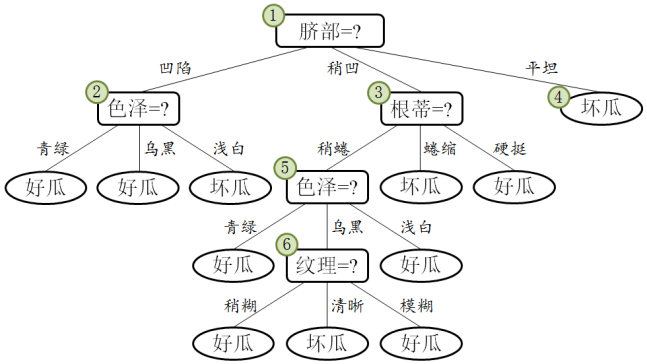
- 预剪枝
  - 决策树生成过程中，对每个结点在划分前先进行估计，若当前结点的划分不能带来决策树泛化性能提升，则停止划分并将当前结点记为叶结点，其类别标记为训练样例数最多的类别
  - 针对上述数据集，基于信息增益准则，选取属性“脐部”划分训练集。分别计算划分前（即直接将该结点作为叶结点）及划分后的验证集精度，判断是否需要划分。若划分后能提高验证集精度，则划分，对划分后的属性，执行同样判断；否则，不划分

剪枝处理

- 预剪枝

验证集

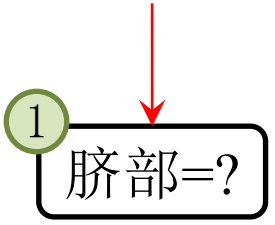
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否



结点1：若不划分，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。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

“脐部=?” 划分前：42.9%

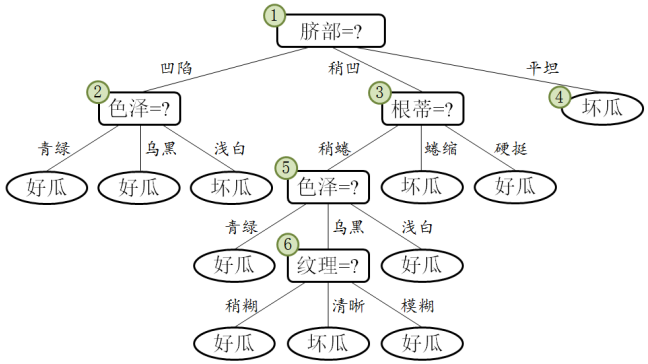


剪枝处理

- 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，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。验证集中，{4,5,8}被分类正确，得到验证集精度为 $\frac{3}{7} \times 100\% = 42.9\%$

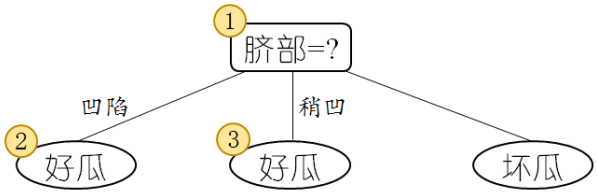
结点1：若划分，根据结点2,3,4的训练样例，将这3个结点分别标记为“好瓜”、“好瓜”、“坏瓜”。此时，验证集中编号{4,5,8,11,12}的样例被划分正确，验证集精度为 $\frac{5}{7} \times 100\% = 71.4\%$

验证集精度

“脐部=?” 划分前：42.9%

划分后：71.4%

预剪枝决策：划分

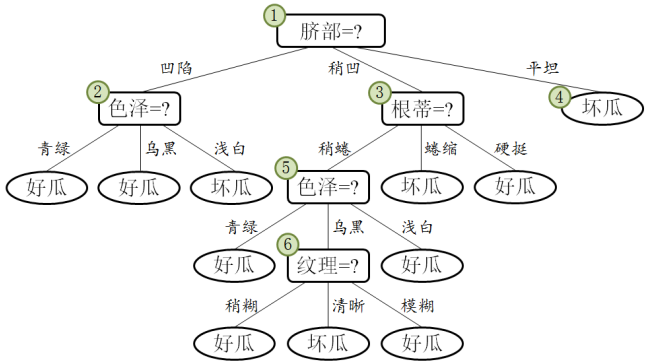


剪枝处理

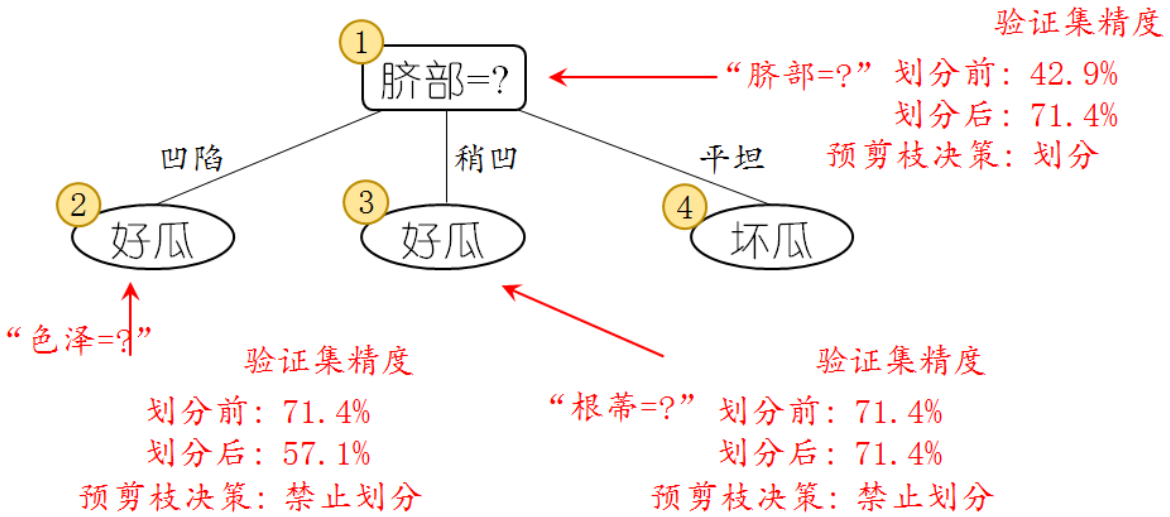
- 预剪枝

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**”

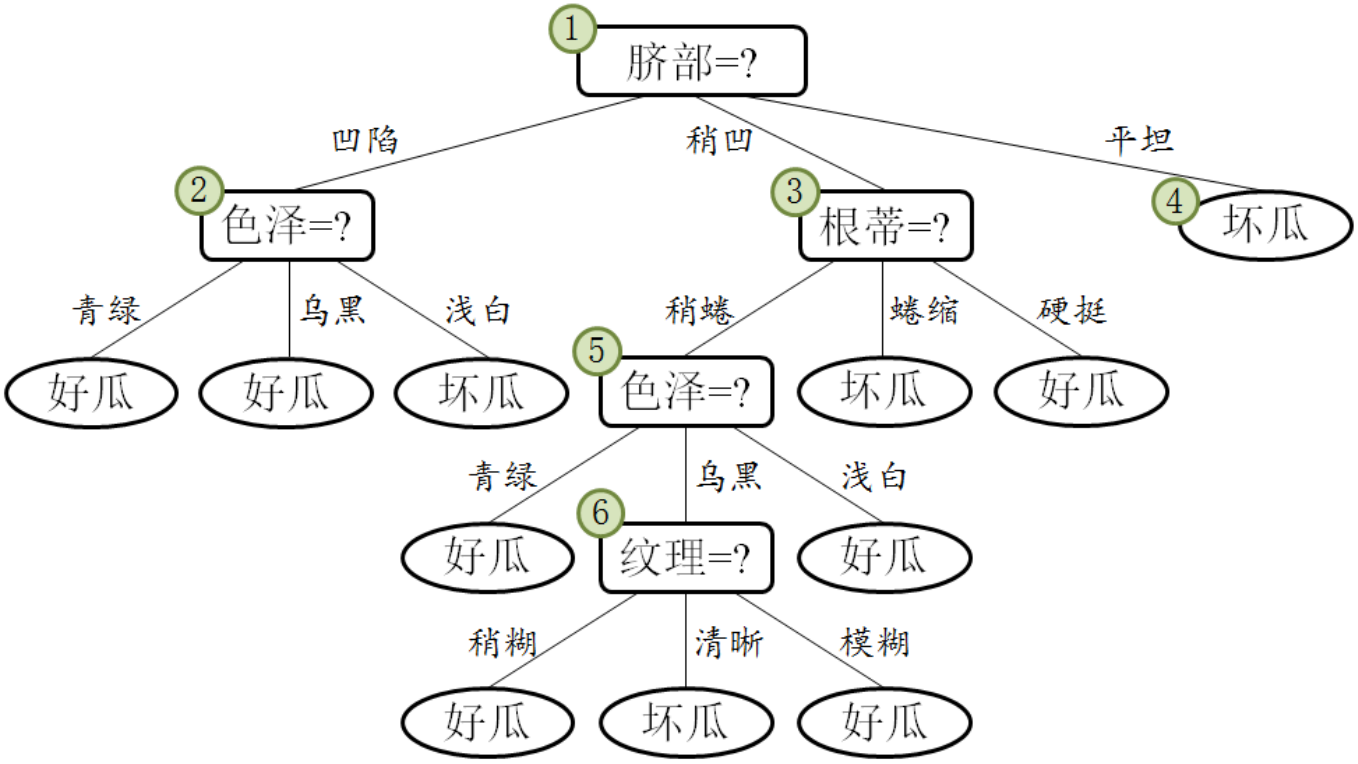


剪枝处理

- 预剪枝
  - 优点
    - 降低过拟合风险
    - 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

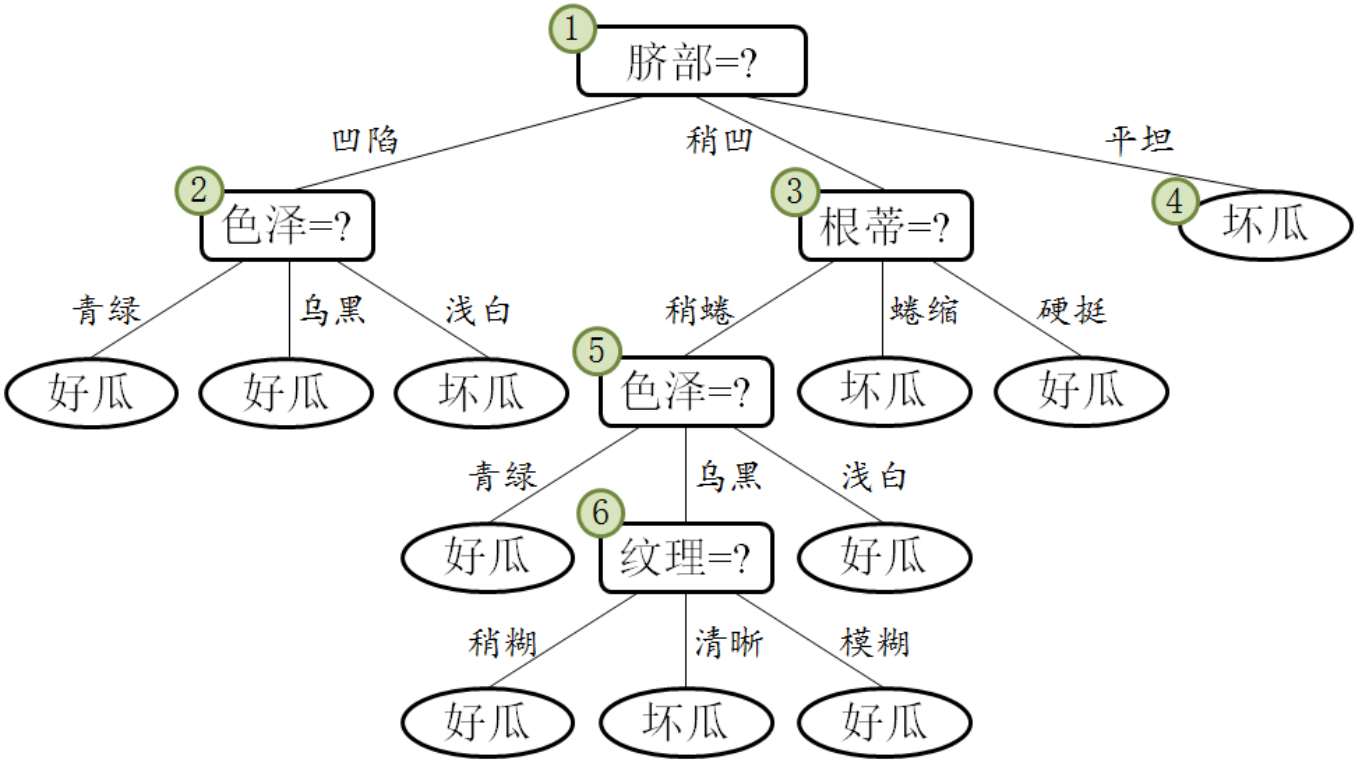
剪枝处理

- 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



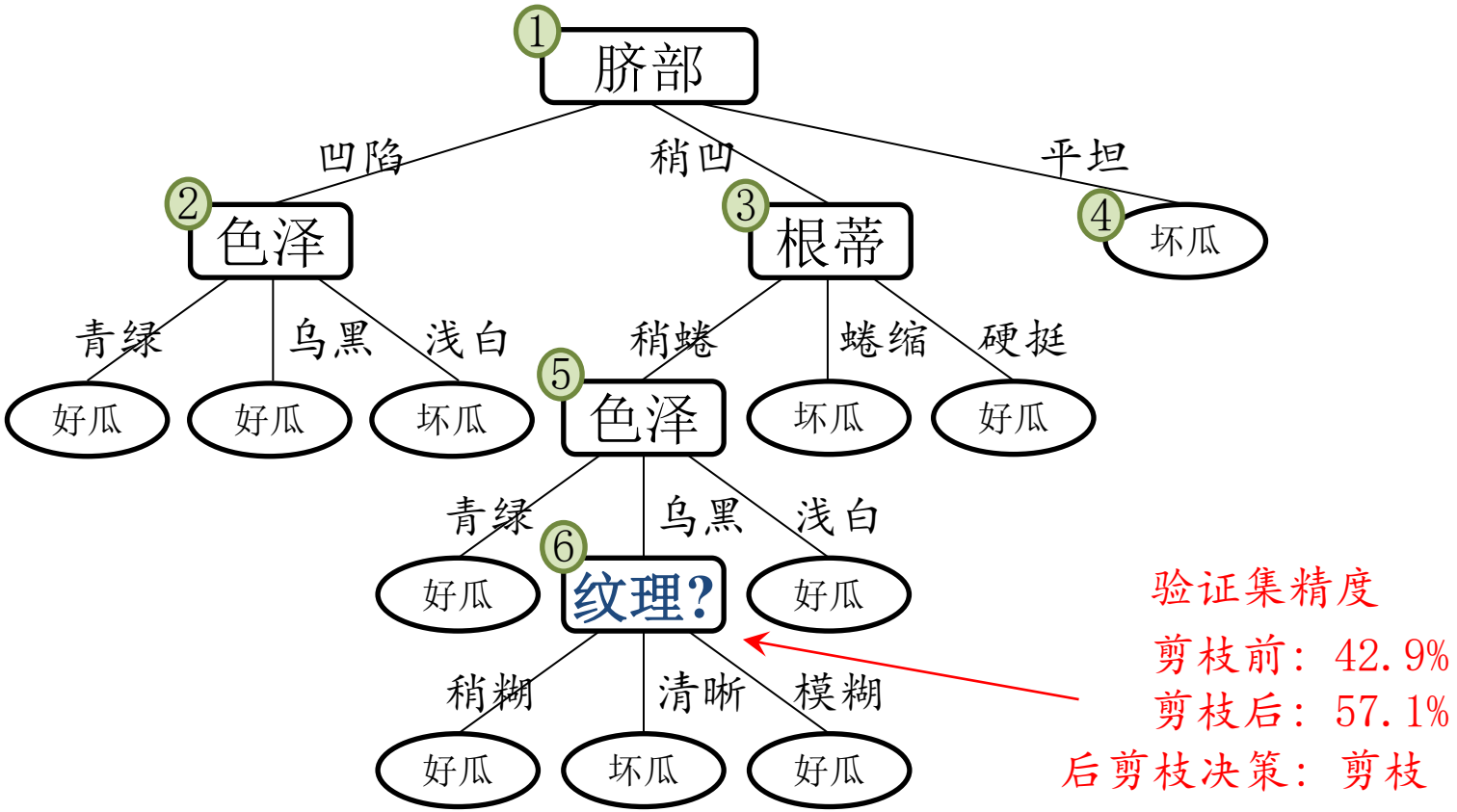
剪枝处理

- 后剪枝
  - 首先生成一棵完整的决策树，该决策树的验证集精度为0.429



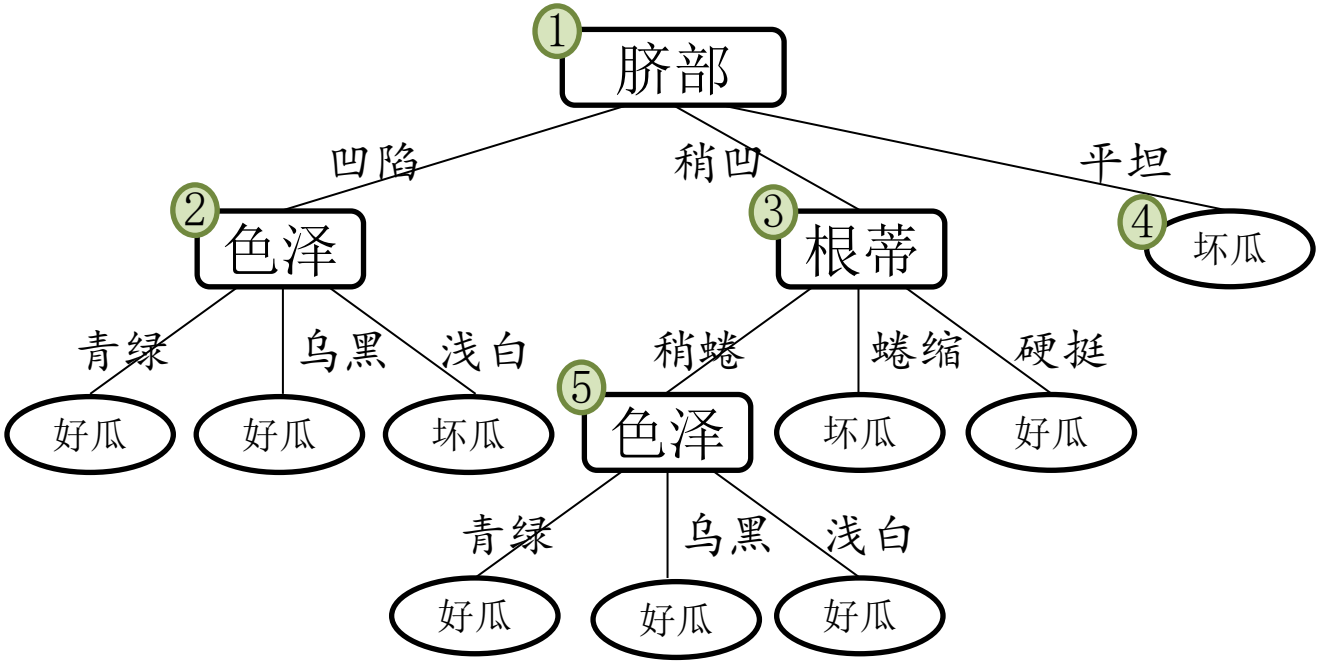
剪枝处理

- 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”，得到验证集精度提高至0.571，则决定剪枝



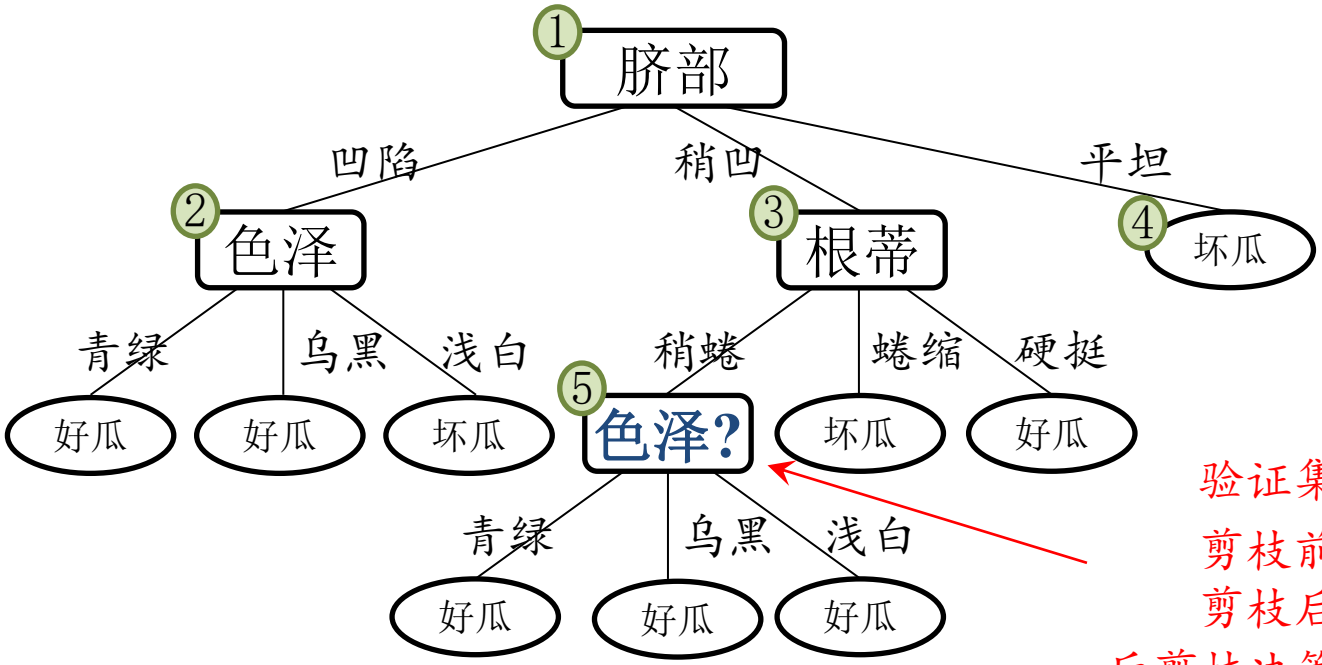
剪枝处理

- 后剪枝
  - 首先考虑结点6，若将其替换为叶结点，根据落在其上的训练样本{7,15}将其标记为“好瓜”，得到验证集精度提高至0.571，则决定剪枝



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0.571，可以不进行剪枝



验证集精度

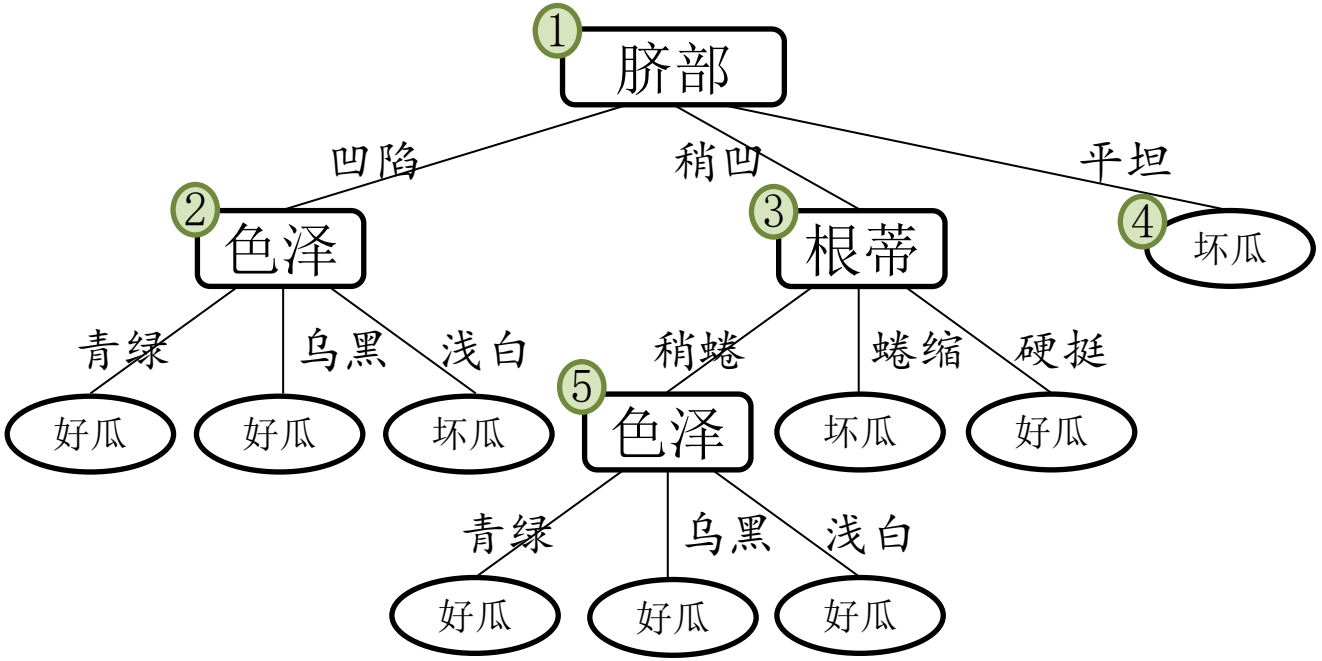
剪枝前：57.1 %

剪枝后：57.1%

后剪枝决策：不剪枝

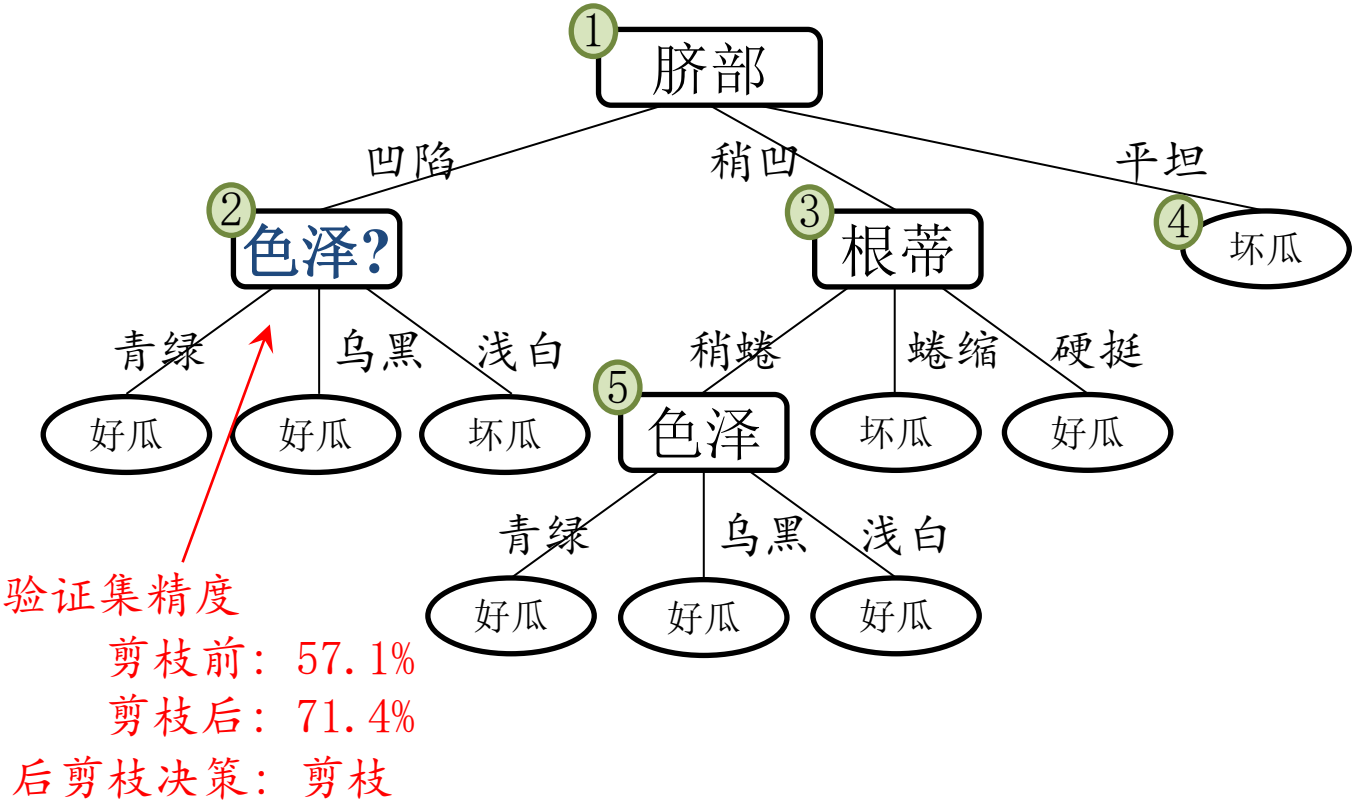
剪枝处理

- 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0.571，可以不进行剪枝



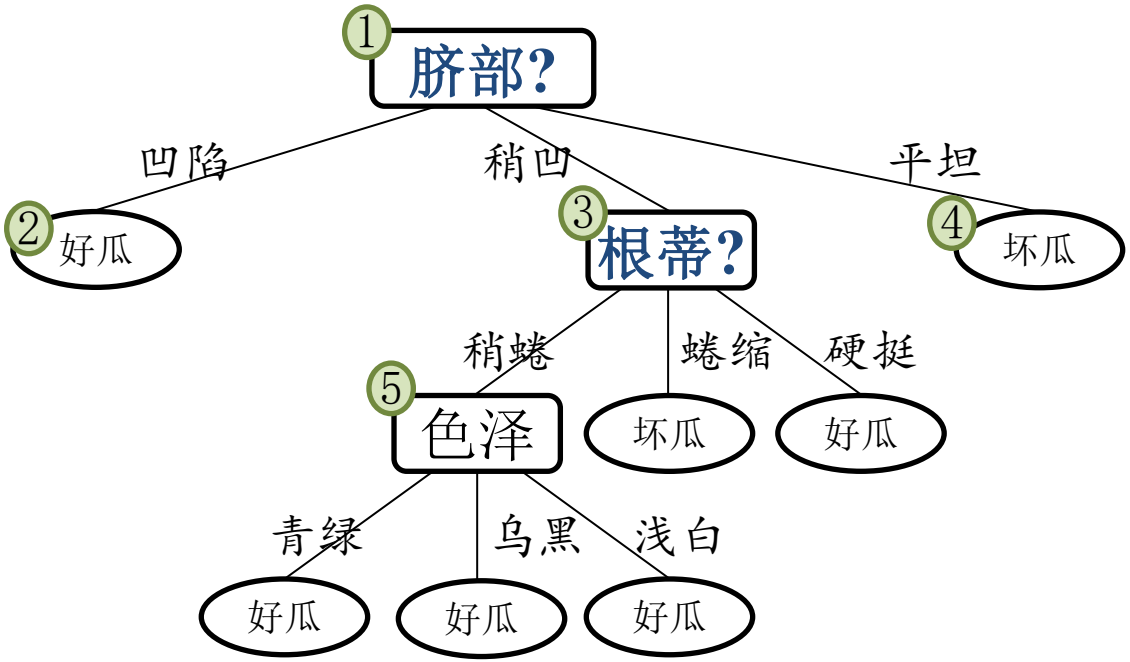
剪枝处理

- 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0.714，则决定剪枝



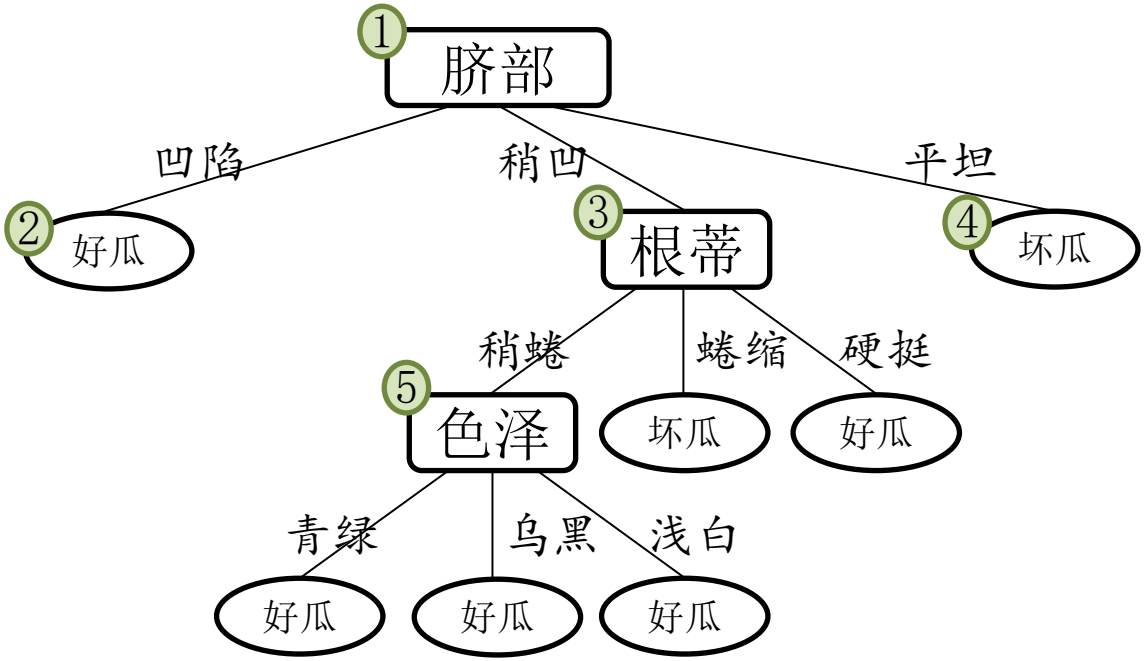
剪枝处理

- 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
  - 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

本节目录

特殊属性处理

- 连续值
  - 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^+包含那些在属性a上取值大于t的样本。考虑包含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

特殊属性处理

- 连续值
  - 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

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

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

特殊属性处理

- 连续值
  - 连续属性离散化(二分法)

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

对属性“密度”，其候选划分点集合包含16个候选值：

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

可计算其信息增益为0.262

对应划分点为0.381

对属性“含糖量”进行同样处理

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

特殊属性处理

- 缺失值
  - 不完整样本，即样本的属性值缺失
  - 仅使用无缺失的样本进行学习?
    - 对数据信息极大的浪费
  - 使用有缺失值的样本，需要解决哪些问题？
    - Q1：如何在属性缺失的情况下进行划分属性选择？
    - Q2：给定划分属性,若样本在该属性上的值缺失，如何对样本进行划分？

特殊属性处理

- 缺失值
  - $\tilde{D}$表示D中在属性a上没有缺失值的样本子集，$\tilde{D}^v$表示$\tilde{D}$中在属性a上取值为a^v的样本子集，$\tilde{D}_k$表示 $\tilde{D}$中属于第k类的样本子集。为每个样本x赋予一个权重w_x，并定义：
    - 无缺失值样本所占的比例 $\rho = \frac{\sum_{x \in \tilde{D}} w_x}{\sum_{x \in D} w_x}$
    - 无缺失值样本中第k类所占比例

$$\tilde{p}_k = \frac{\sum_{x \in \tilde{D}_k} w_x}{\sum_{x \in \tilde{D}} w_x} \quad (1 \leq k \leq |\mathcal{Y}|)$$

    - 无缺失值样本中在属性a上取值a^v的样本所占比例

$$\tilde{r}_v = \frac{\sum_{x \in \tilde{D}^v} w_x}{\sum_{x \in \tilde{D}} w_x} \quad (1 \leq v \leq V)$$

Q1：如何在属性缺失的情况下进行划分属性选择？

特殊属性处理

- 缺失值
  - 基于上述定义，可得

$$\begin{aligned}\mathrm{Gain}(D,a) &= \rho \times \mathrm{Gain}(\tilde{D},a) \\ &= \rho \times \left(\mathrm{Ent}(\tilde{D}) - \sum_{v=1}^{V} \tilde{r}_v \mathrm{Ent}(\tilde{D}^v)\right)\end{aligned}$$

其中 $\mathrm{Ent}(\tilde{D}) = -\sum_{k=1}^{|\mathcal{Y}|} \tilde{p}_k \log_2 \tilde{p}_k$

  - 对于Q2
    - 若样本x在划分属性a上的取值已知，则将x划入与其取值对应的子结点，且样本权值在子结点中保持为w_x
    - 若样本x在划分属性a上的取值未知，则将x同时划入所有子结点，且样本权值在与属性值a^v对应的子结点中调整为 $\tilde{r}_v \cdot w_x$ (直观来看，相当于让同一个样本以不同概率划入不同的子结点中去)

特殊属性处理

- 缺失值

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

特殊属性处理

- 缺失值
  - 学习开始时，根结点包含样本集D中全部17个样例，各样例的权值均为1
  - 以属性“色泽”为例，该属性上无缺失值的样例子集 $\tilde{D}$ 包含14个样例，$\tilde{D}$ 的信息熵为

$$\begin{aligned}\mathrm{Ent}(\tilde{D}) &= -\sum_{k=1}^{2} \tilde{p}_k \log_2 \tilde{p}_k \\ &= -\left(\frac{6}{14}\log_2\frac{6}{14} + \frac{8}{14}\log_2\frac{8}{14}\right) = 0.985\end{aligned}$$

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

特殊属性处理

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

- 缺失值
  - 令 $\tilde{D}^1, \tilde{D}^2, \tilde{D}^3$ 分别表示在属性“色泽”上取值为“青绿”“乌黑”以及“浅白”的样本子集，有

$$\mathrm{Ent}(\tilde{D}^1) = -\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) = 1.000$$

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

$$\mathrm{Ent}(\tilde{D}^3) = -\left(\frac{0}{4}\log_2\frac{0}{4} + \frac{4}{4}\log_2\frac{4}{4}\right) = 0.000$$

  - 样本子集 $\tilde{D}$ 上属性“色泽”的信息增益为

$$\mathrm{Gain}(\tilde{D}, \text{色泽}) = \mathrm{Ent}(\tilde{D}) - \sum_{v=1}^{3} \tilde{r}_v \mathrm{Ent}(\tilde{D}^v)$$

$$= 0.985 - \left(\frac{4}{14} \times 1.000 + \frac{6}{14} \times 0.918 + \frac{4}{14} \times 0.000\right) = 0.306$$

  - 样本集 D 上属性“色泽”的信息增益为

$$\mathrm{Gain}(D, \text{色泽}) = \rho \times \mathrm{Gain}(\tilde{D}, \text{色泽}) = \frac{14}{17} \times 0.306 = 0.252$$

特殊属性处理

- 缺失值
  - 类似地可计算出所有属性在数据集上的信息增益

$$\text{Gain}(D, \text{色泽}) = 0.252 \quad \text{Gain}(D, \text{根蒂}) = 0.171$$

$$\text{Gain}(D, \text{敲声}) = 0.145 \quad \boxed{\text{Gain}(D, \text{纹理}) = 0.424}$$

$$\text{Gain}(D, \text{脐部}) = 0.289 \quad \text{Gain}(D, \text{触感}) = 0.006$$

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

进入“纹理=清晰”分支

进入“纹理=稍糊”分支

进入“纹理=模糊”分支

样本权重在各子结点仍为1

在属性“纹理”上出现缺失值，样本8和10同时进入3个分支，调整8和10在3分分支权值分别为7/15，5/15，3/15

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

总结

- **背景知识**
  - 归纳学习
- **模型结构**
  - 基本思想
  - 一般步骤
- **学习算法**
  - ID3算法
  - C4.5算法
  - CART算法
- **剪枝处理**
  - 预剪枝、后剪枝
- **特殊属性处理**
  - 连续值、缺失值

思考题

- 试对缺失值的处理机制推广到基尼指数的计算中

练习题

- 试编程实现基于基尼指数的决策树学习算法