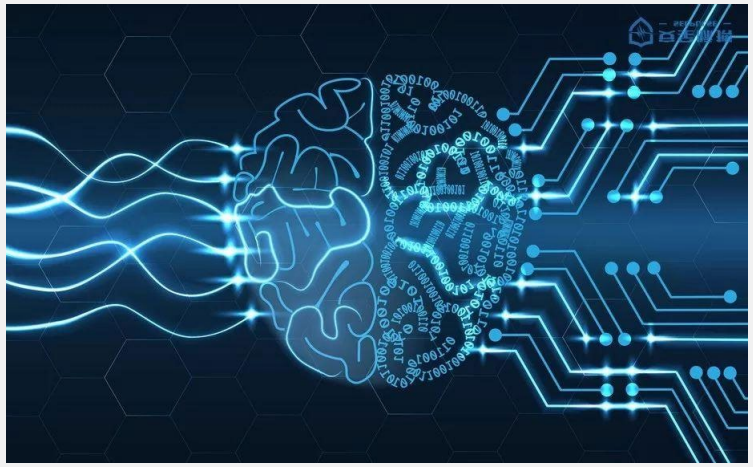


机器学习

Machine Learning

授课老师：谭毅华
电　　话：13886021197
办 公 室：科技楼1102
邮　　箱：yhtan@hust.edu.cn

第四章、决策树与集成学习

目录
CONTENTS

为人师表

1、决策树的基本概念

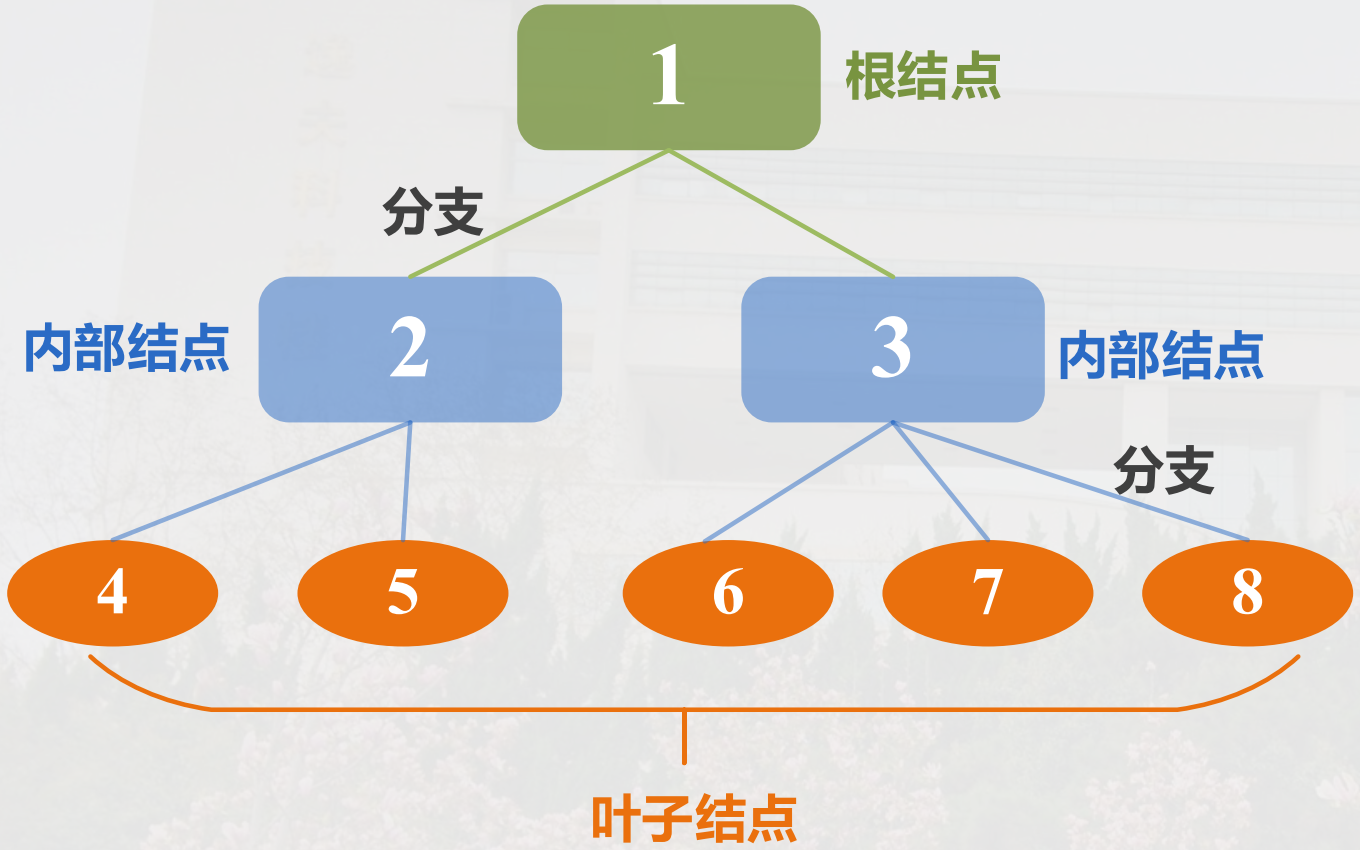
- 什么是决策树?

决策树是一种用于进行决策的树形结构。

树形结构 包括
- 根结点
- 分支
- 内部结点
- 叶子结点

1、决策树的基本概念

➢决策树结构



根结点：进行初始属性测试（该结点包含所有样本）

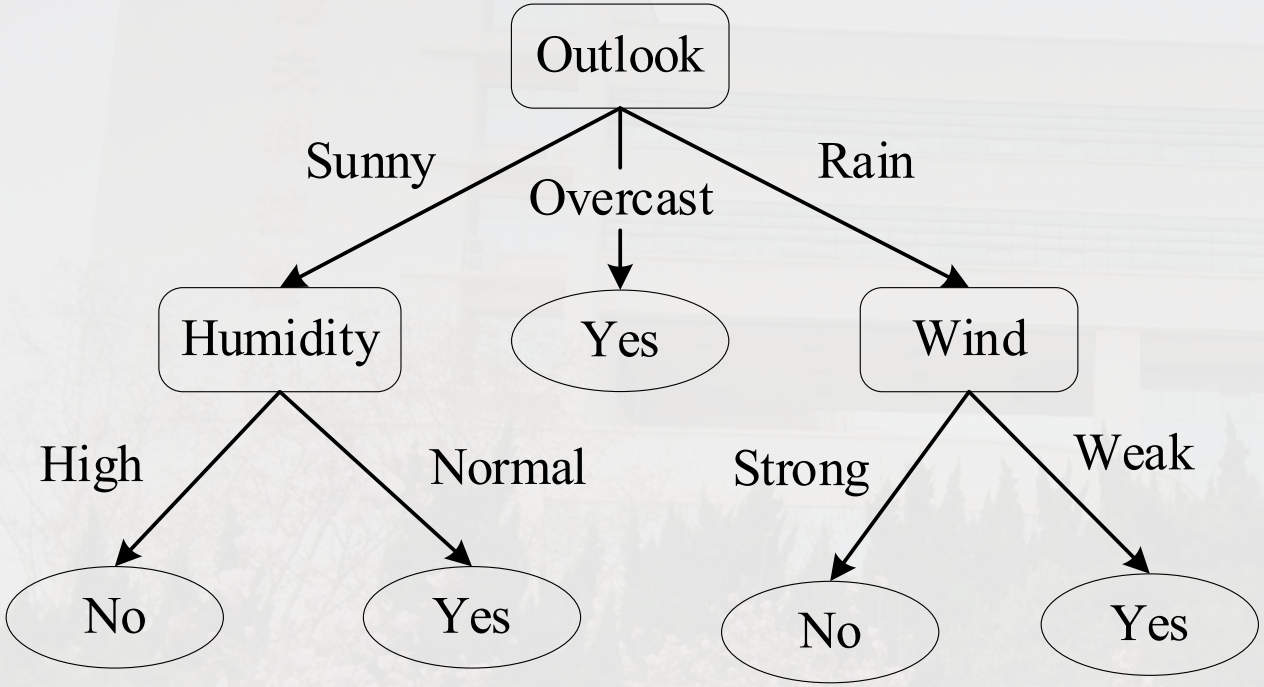
分支：代表上层结点测试结果的输出

内部结点：进行属性测试（包含部分样本）

叶子结点：表示决策结果

1、决策树的基本概念

- 举一个小例子说明决策树是如何工作的



决策规则（判定测试序列）

1. 晴天　湿度高　不去
2. 晴天　湿度正常　去
3. 阴天　去
4. 雨天　强风　不去
5. 雨天　弱风　去

周六上午是否适合打网球呢?

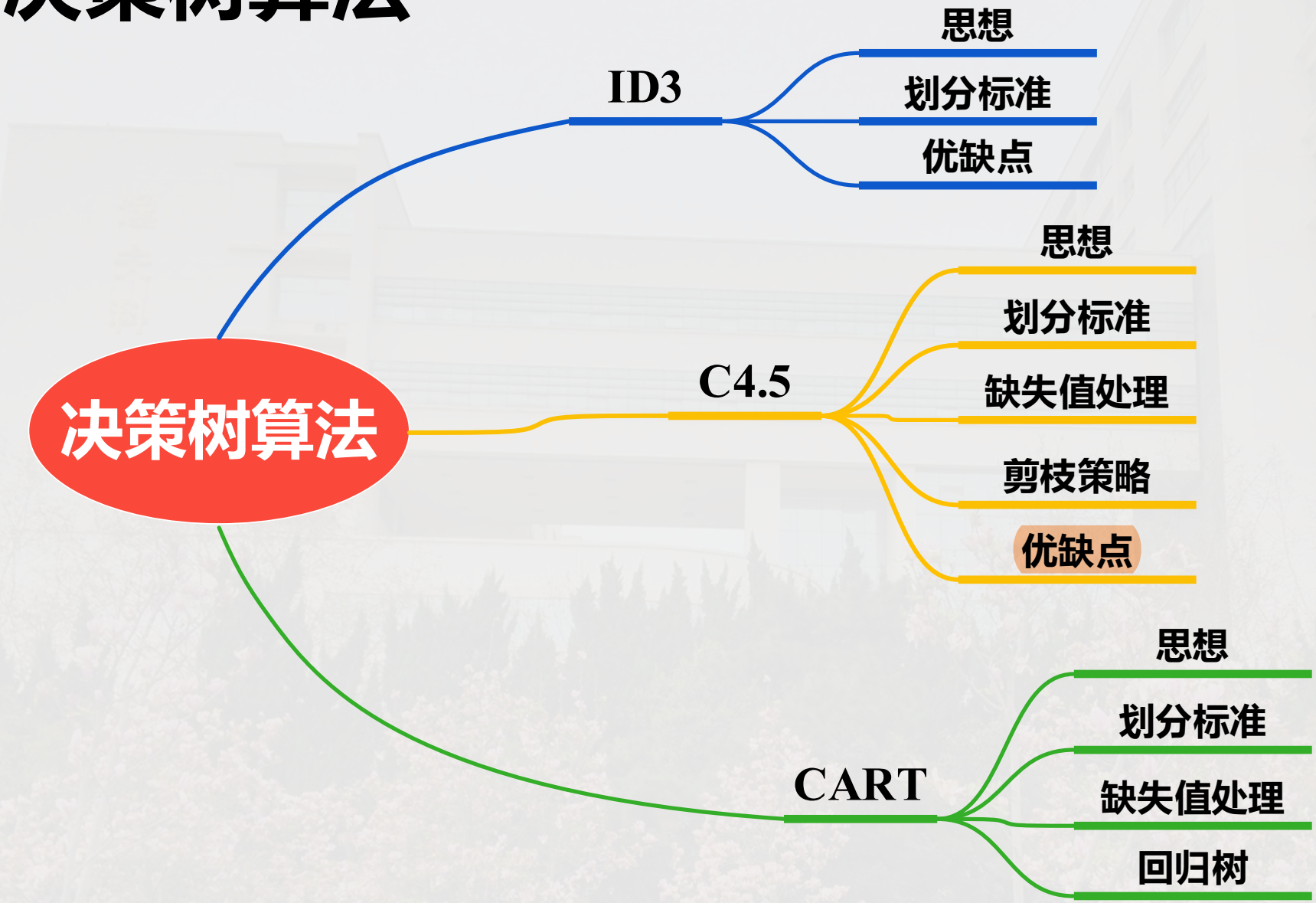
- 优点：可解释性强（应用商业/医疗）、简单、效率高

1、决策树的基本概念

- 通过举例可知，决策树学习本质上就是从训练数据集中归纳出一组分类规则。

- 如何归纳，或者说如何通过训练数据集构建一棵决策树呢?

2、决策树算法


决策树算法
ID3
思想
划分标准
优缺点
C4.5
思想
划分标准
缺失值处理
剪枝策略
优缺点
CART
思想
划分标准
缺失值处理
回归树

2.1、ID3算法

➢ 思想

✓ 决策树学习关键：如何选择最优划分属性

❑ 考虑如果现在属性集合里包括以下两个属性，如何选择：

编号	色泽	脐部	好瓜
1	乌黑	凹陷	是
2	青绿	凹陷	是
3	青绿	平坦	否
4	乌黑	平坦	否

色泽 or 脐部？

2.1、ID3算法

➢ 思想

✓ 决策树学习关键：**如何选择最优划分属性?**

❑ 考虑如果现在属性集合里包括以下两个属性，如何选择：

编号	色泽	脐部	好瓜
1	乌黑	凹陷	是
2	青绿	凹陷	是
3	青绿	平坦	否
4	乌黑	平坦	否

色泽 or 脐部?

脐部√

结论：我们希望随着划分过程的不断进行，决策树分支结点所包含的样本尽可能属于同一类别，即结点的“**纯度**”越来越高。

纯度如何度量?

2.1、ID3算法

- 从信息论的知识可知：**信息熵**越大，样本纯度越低。因此可以使用**信息熵**来度量结点的纯度。

假定样本集合D中第k类样本所占的比例为p_k，则该样本的信息熵可定义为

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

约定当 $p = 0$时， $plog_2 p = 0$

2.1、ID3算法

➢ 信息增益

属性 A 对训练数据集 D 的信息增益定义为集合 D 的**经验熵**$H(D)$（即前面提到的信息熵）与给定 A 条件下的**经验条件熵** $H(D|A)$ 之差，记为 $Gain(D,A)$ 。

$$Gain(D,A) = H(D) - H(D|A)$$

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

$$H(D|A) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v)$$

其中D^v代表数据集合 D 中所有在属性 A 上取值为A^v的样本

2.1、ID3算法

- 核心思想：根据**信息增益**来选择划分属性
- 具体方法：

- 从根结点开始，计算所有属性的信息增益
- 选择信息增益最大的属性作为结点划分属性
- 由该属性的不同取值进行分支
- 重复以上步骤

2.1、ID3算法

- 例：当新客户提出申请贷款时,根据申请人的特征利用决策树决定是否批准申请贷款.

ID	年龄	有工作	有自己的房子	信贷情况	类别
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否

2.1、ID3算法

针对前面提到的贷款申请的例子，使用ID3算法来决定是否批准贷款申请。

类别	ID	数量
是	3,4,8,9,10,11,12,13,14	9
否	1,2,5,6,7,15	6

计算根结点信息熵：

$$Ent(D) = -\sum_{k=1}^{y} p_k \, log_2 p_k = -\left(\frac{9}{15} log_2 \frac{9}{15} + \frac{6}{15} log_2 \frac{6}{15}\right) = 0.971$$

2.1、ID3算法

属性集集合A = {年龄，工作，房子，信贷情况}，分别用A_1, A_2, A_3, A_4表示集合中的属性元素。

$$H(D|A_1 = \text{青年}) = -\left(\frac{2}{5}log_2\frac{2}{5} + \frac{3}{5}log_2\frac{3}{5}\right) = 0.971$$

$$H(D|A_1 = \text{中年}) = -\left(\frac{3}{5}log_2\frac{3}{5} + \frac{2}{5}log_2\frac{2}{5}\right) = 0.971$$

$$H(D|A_1 = \text{老年}) = -\left(\frac{4}{5}log_2\frac{4}{5} + \frac{1}{5}log_2\frac{1}{5}\right) = 0.722$$

2.1、ID3算法

$$H(D|A_1) = \frac{5}{15} * 0.971 + \frac{5}{15} * 0.971 + \frac{5}{15} * 0.722 = 0.888$$

在A_1条件下，数据集的经验条件熵

$$Gain(D, A_1) = H(D) - H(D|A_1) = 0.971 - 0.888 = 0.083$$

同理可得

$$Gain(D, A_2) = 0.324$$

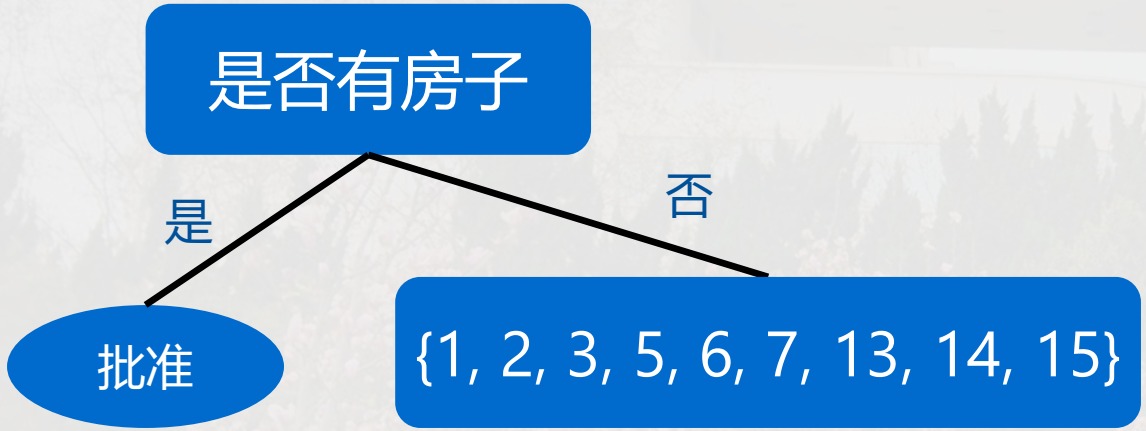
$$Gain(D, A_3) = 0.420$$

$$Gain(D, A_4) = 0.363$$

A_3(即是否有自己的房子)的信息增益最大，因此选择A_3为根结点的划分属性，接下来对其他结点递归调用上述方法，建立决策树

2.1、ID3算法

- 测验：上例采用是否有房子作为根结点属性后得到下面左图的决策树桩，请确定接下来选择哪个属性。



ID	年龄	有工作	信贷情况	类别
1	青年	否	一般	否
2	青年	否	好	否
3	青年	是	好	是
5	青年	否	一般	否
6	中年	否	一般	否
7	中年	否	好	否
13	老年	是	好	是
14	老年	是	非常好	是
15	老年	否	一般	否

2.1、ID3算法

✓ 优点

- 假设空间包含所有的决策树，搜索空间完整
- 健壮性好，特征噪声影响较小
- 方法简单，理论清晰

2.1、ID3算法

✓ 缺点

- 只考虑离散属性，没有考虑连续属性
- 对缺失值状况没有进行考虑
- 没有考虑过拟合情况
- 采用信息增益作为划分标准，但信息增益倾向于取值较多的属性。

2.1、ID3算法

✓ 缺点

- 采用信息增益作为划分标准，但信息增益倾向于取值较多的属性。

例如我们选择前面贷款的例子中的第一列ID(使用A_5来表示)作为划分属性：

$$H(D|A_5 = i) = -(1 * log_2 1 + 0 * log_2 0) = 0 \quad (i = 1,2, \dots 15)$$

$$H(D|A_5) = \sum_{i=1}^{15} \frac{1}{15} * H(D|A_5 = i) = 0$$

$$Gain(D, A_5) = H(D) - H(D|A_5) = 0.971 - 0 = 0.971$$

由此可以看出当选择ID作为划分属性时，子结点的信息熵会直接降为0，此时的信息增益也远远大于其他属性

2.1、ID3算法

➢测验：

下面关于ID3算法中说法错误的是（）

A. ID3算法要求特征必须离散化

B. 采用信息增益作为划分准则

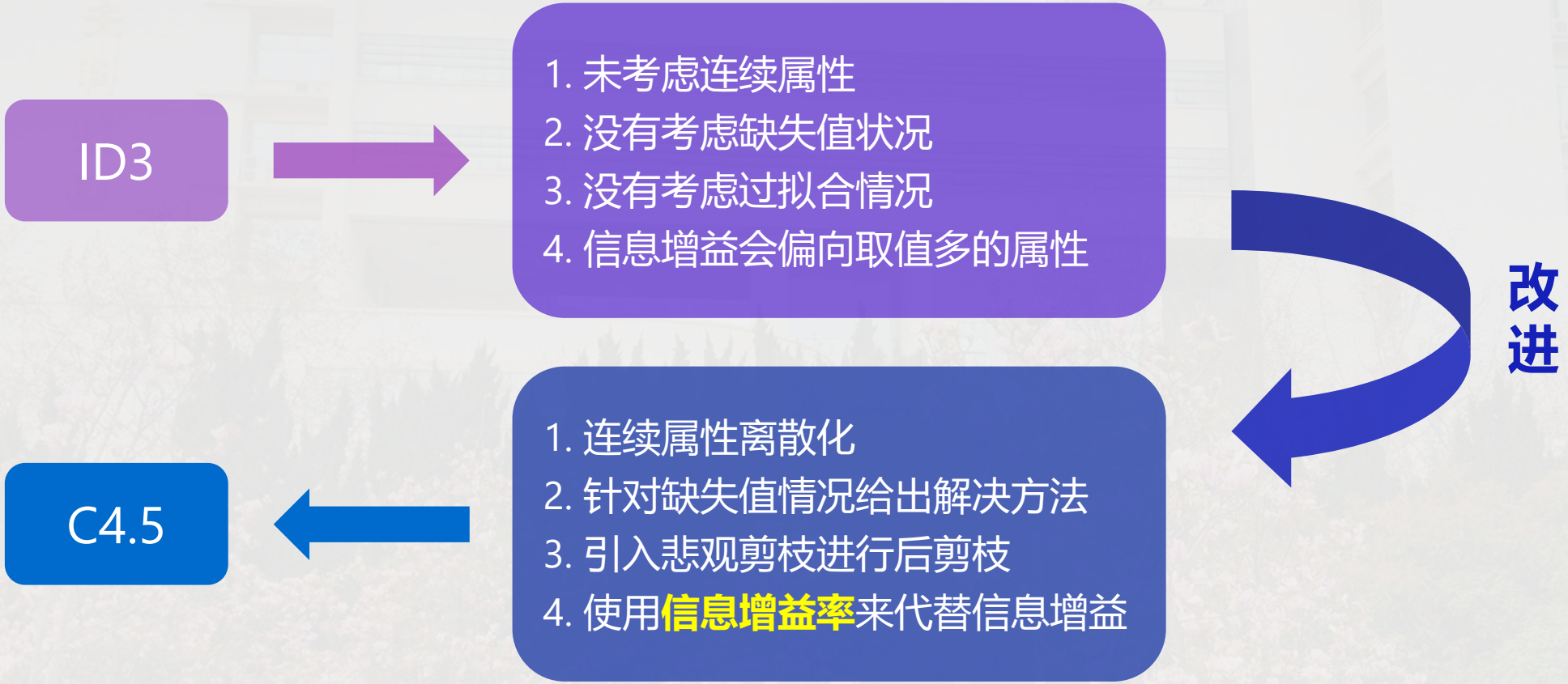
C. 选取信息增益最大的特征，作为树的根节点

✓ D. ID3算法是一个二叉树模型

2.2、C4.5算法

➢ 思想

可将C4.5算法看成ID3算法的改进算法。



2.2、C4.5算法

➢ 划分标准

利用信息增益率可以克服信息增益中对取值较多的属性偏袒这一缺点。

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

其中$IV(A) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} log_2 \frac{|D^v|}{|D|}$，称为属性 A 的“固有值”，属性 A 的取值数目越多（即 V 越大），$IV(A)$的值通常越大。

- 属性A取值数目越少 —> IV(A)越小 —>对取值少的属性有所偏袒

2.2、C4.5算法

➢ 划分标准

利用信息增益率可以克服信息增益中对取值较多的属性偏袒这一缺点。

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

其中$IV(A) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} log_2 \frac{|D^v|}{|D|}$，称为属性 A 的“固有值”，属性 A 的取值数目越多（即 V 越大），$IV(A)$的值通常越大。

- 采用启发式算法：先从候选划分特征中找到**信息增益**高于平均值的特征，再从中选择**增益率**最高的。

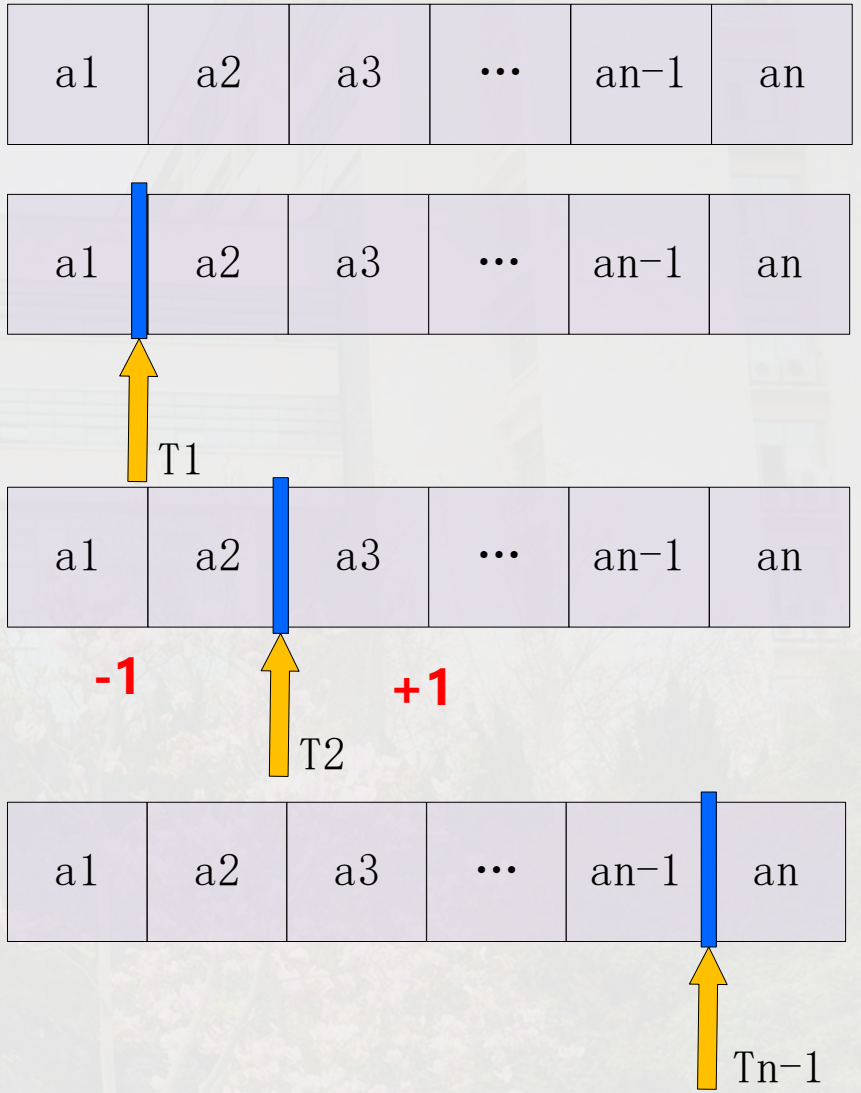
2.2、C4.5算法

➢ 连续属性处理

✓ 连续离散化

1. 将连续属性在数据集D上的取值按从小到大排列，记为$\{a_1, a_2, \ldots, a_n\}$
2. 此时会产生n-1个划分点（划分点取两侧值平均数）例如$T_1 = \frac{a_1+a_2}{2}$，得到一个划分点集合$T_a = \{\frac{a_i+a_{i+1}}{2} | 1 \le i \le n-1\}$
3. 将其看为离散属性值来考察这些划分点（例如使用信息增益进行划分）

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



2.2、C4.5算法

➢ 缺失值处理

仅选用无缺失值样本?

{4,7,14,16}

这样显然浪费了大量数据。

西瓜数据集 2.0α

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

2.2、C4.5算法

➢ 缺失值处理

考虑两个问题：

1. 属性值缺失情况下，该如何选择划分属性?

2. 如果现在已经给定了划分属性，样本在该属性上的值缺失，如何划分该样本?

2.2、C4.5算法

➢ 缺失值处理

问题1：取在属性a上无缺失值的样本子集$\tilde{D}$，使用$\tilde{D}$来判断属性a的优劣。

以右侧数据为例，我们取“色泽”这一属性，计算其信息增益。

西瓜数据集 2.0α

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

2.2、C4.5算法

➢ 缺失值处理

色泽 $\tilde{D} = \{2,3,4,6,7,8,9,10,11,12,14,15,16,17\}$

好瓜	坏瓜
2,3,4,6,7,8	9,10,11,12,14,15,16,17

$$Ent(\tilde{D}) = -\sum_{k=1}^{2} \tilde{p}_k log_2 \tilde{p}_k$$

$$= -\left(\frac{6}{14} log_2 \frac{6}{14} + \frac{8}{14} log_2 \frac{8}{14}\right)$$

$$= 0.985$$

西瓜数据集 2.0α

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

2.2、C4.5算法

➢ 缺失值处理

色泽	好瓜	坏瓜
乌黑	2,3,7,8	9,15
青绿	4,6	10,17
浅白	\	11,12,14,16

$$Ent(\tilde{D}^1 = 乌黑) = -\left(\frac{4}{6}log_2\frac{4}{6} + \frac{2}{6}log_2\frac{2}{6}\right) = 0.918$$

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

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

$$Gain(\tilde{D}, 色泽)$$

$$= Ent(\tilde{D}) - \sum_{v=1}^{3} Ent(\tilde{D}^v)$$

$$= 0.985 - \left(\frac{6}{14} * 0.918 + \frac{4}{14} * 1.000 + \frac{4}{14} * 0\right)$$

$$= 0.306$$

$$Gain(D, 色泽)$$

$$= \rho * Gain(\tilde{D}, 色泽)$$

$$= \frac{14}{17} * 0.306 = 0.252$$

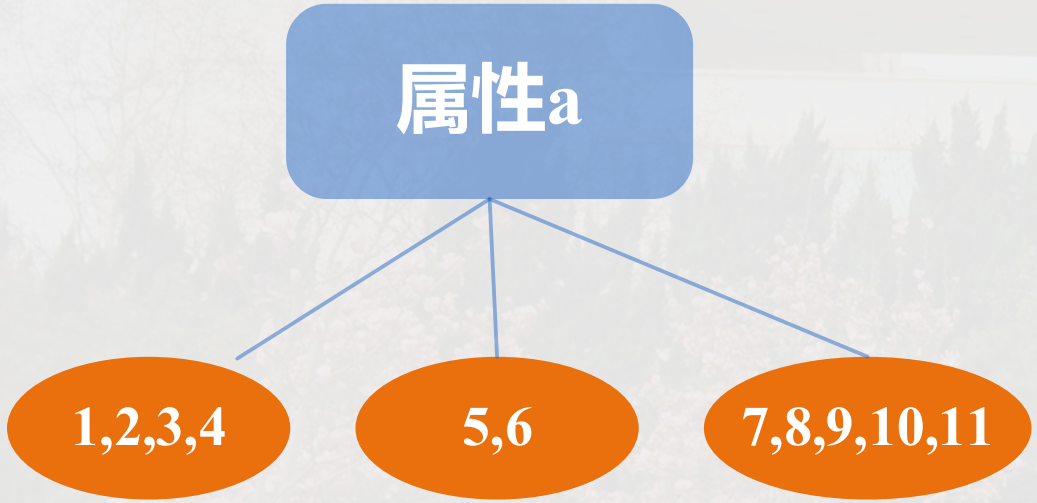
ρ 代表对于属性a来说，无缺失值样本所占比例

2.2、C4.5算法

➢ 缺失值处理

问题2：**如果现在已经给定了划分属性，样本在该属性上的值缺失，如何划分该样本?**

将样本以**不同概率**划分到所有的属性中去。



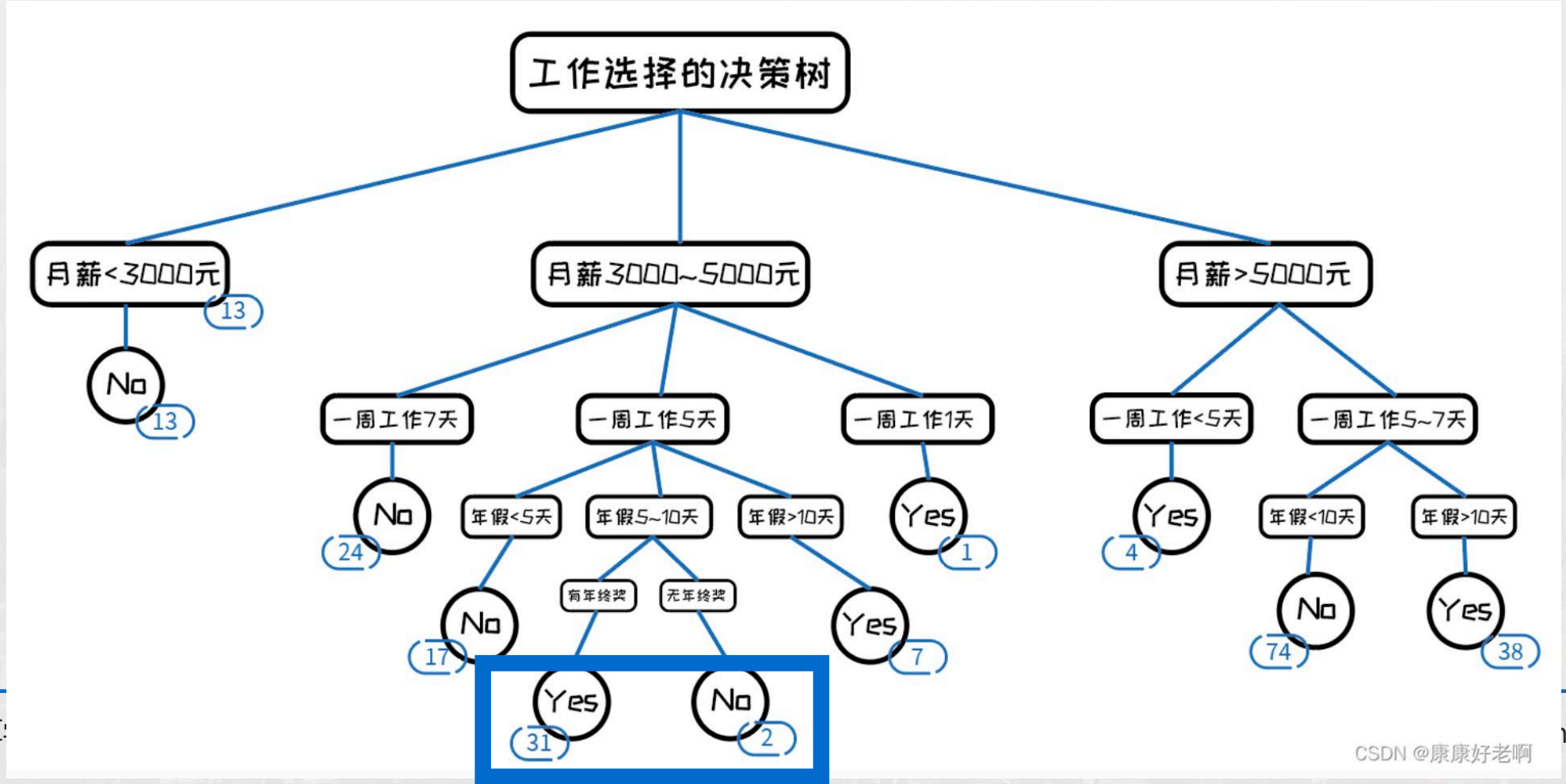
例如选择属性a进行样本划分，三种取值的比例分别为$\frac{4}{11},\frac{2}{11},\frac{5}{11}$。现在有另一个样本其在属性a上取值未知，将其同时划分到三个分支中，并将权重分别调整为$\frac{4}{11},\frac{2}{11},\frac{5}{11}$。

样本属于某类的概率：所有分支的类别概率再乘上对应权重后求和。

2.2、C4.5算法

➢ 剪枝策略

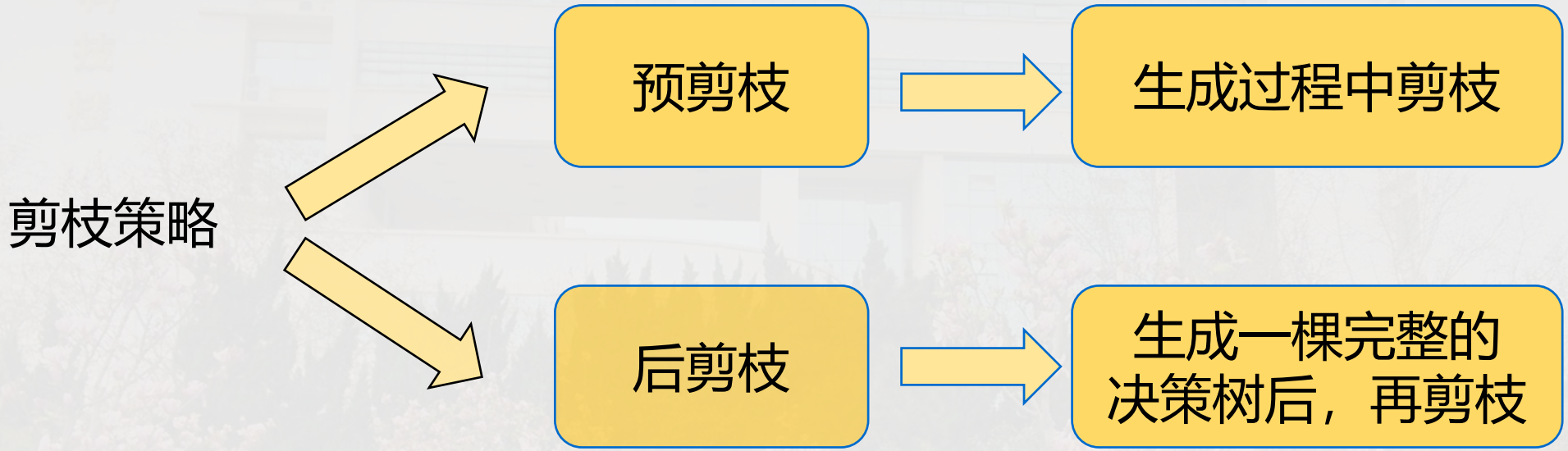
剪枝策略是在决策树算法中用来对付**“过拟合”**的一种方法。



2.2、C4.5算法

➢ 剪枝策略

剪枝策略是在决策树算法中用来对付**“过拟合”**的一种方法。



2.2、C4.5算法

➢ 预剪枝

在结点划分之**前**来确定是否继续划分。

判断划分是否会使决策树**泛化性能提升**，若不能，就停止划分。

➢ 后剪枝

在**已经生成的决策树**上进行剪枝

自底而上，递归判断将非叶子结点置换为叶结点，**泛化能力是否提升**，若是，则将该节点对应的子树替换为叶结点。

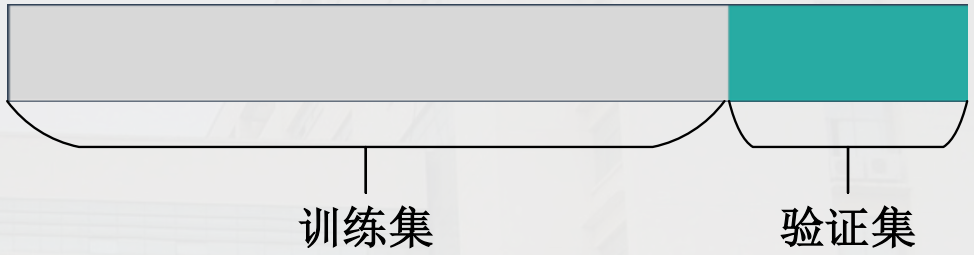
预剪枝

西瓜数据集 2.0 划分出的训练集(双线上部)与验证集(双线下部)

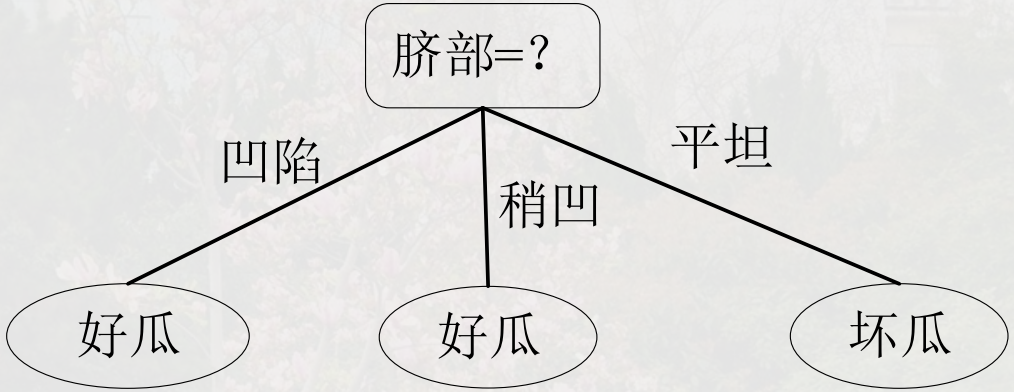
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

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

留出法



选取“脐部”作为根结点进行属性划分（信息增益准则）



预剪枝

训练集

编号	脐部	好瓜
1	凹陷	是
2	凹陷	是
3	凹陷	是
6	稍凹	是
7	稍凹	是
10	平坦	否
14	凹陷	否
15	稍凹	否
16	平坦	否
17	稍凹	否

测试集

编号	脐部	好瓜
4	凹陷	是
5	凹陷	是
8	稍凹	是
9	稍凹	否
11	平坦	否
12	平坦	否
13	凹陷	否

判断是否要进行划分—>使用验证集进行评估

1. 不划分 训练集：好瓜坏瓜数相等，随机设该叶结点标记为“好瓜”

验证集：{4,5,8}分类正确 {9,11,12,13}分类错误

验证集精度：$\frac{3}{7} = 42.9\%$

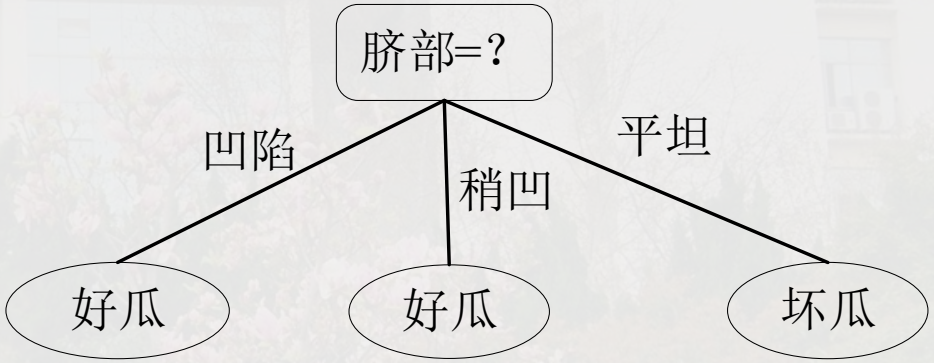
2. 选取“脐部”作为根结点进行属性划分

验证集：

{4,5,8,11,12}分类正确

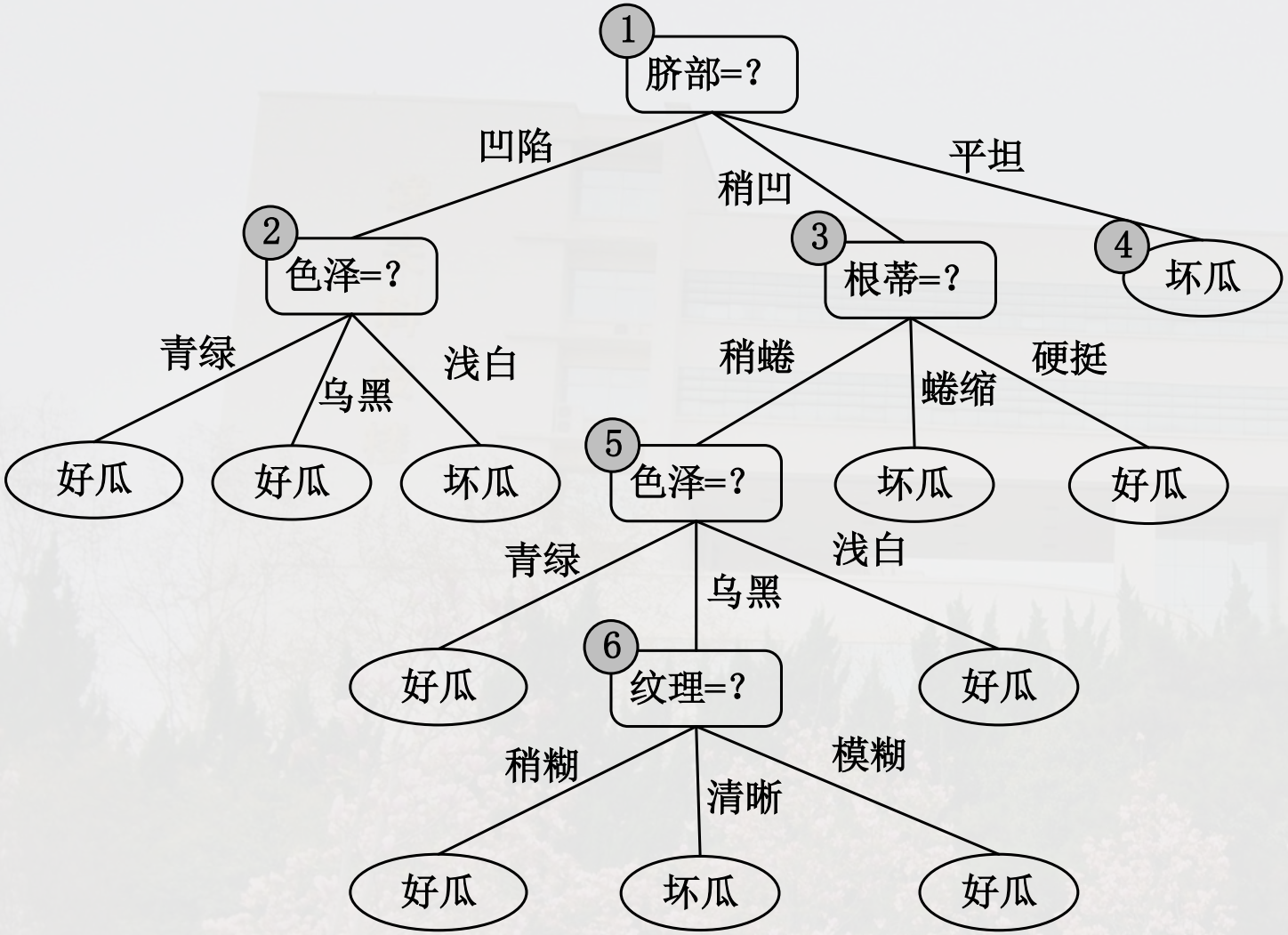
{9,13} 分类错误

验证集精度：$\frac{5}{7} = 71.4\%$



验证集精度在划分后提升，因此用“脐部”进行划分

后剪枝



自底向上：

首先考察⑥结点：

1. 若不剪枝：

划分正确：{4,11,12}

划分错误：{5,8,9,13}

验证集精度：$\frac{3}{7} = 42.9\%$

2. 将⑥替换为叶结点：

⑥结点包含训练集样本{7,15}

将⑥标记为好瓜，{4,5,8,9}

正确，此时验证集精度提升至57.1%，因此进行替换

2.2、C4.5算法

➢ 预剪枝

- ✓ 训练时间开销小
- ✓ 有欠拟合的风险

➢ 后剪枝

- ✓ 训练时间开销大
- ✓ 保留更多分支，欠拟合风险较小，泛化性能更强

2.2、C4.5算法

- 测验：

以下哪个不是C4.5算法的特点?

A.连续的特征离散化处理

B.使用信息增益比

C.通过剪枝算法解决过拟合

✓ D.对于缺失值的情况没有做考虑

2.3、CART算法

- 思想

2.3、CART算法

➢ 划分标准

CART决策树采用“**基尼指数**”来选择划分属性。由于熵模型拥有大量耗时的对数运算，基尼指数在简化模型的同时还保留了熵模型的优点。设p_k为样本属于第k类的概率，y为总类别数。

数据集D的纯度可以用基尼值来度量：

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

直观来说，基尼值反映了从数据集D中随机抽取两个样本，其类别标记不一致的概率，因此，$Gini(D)$越小，则数据集D的纯度越高。

2.3、CART算法

➢ 划分标准

属性A的基尼指数定义为

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

在候选属性集合中，选择是的划分后基尼指数最小的属性作为最优划分属性，即$A_i=\arg\min_{A_i\in A} Gini_index(D,A_i)$

2.3、CART算法

ID	年龄A1	有工作A2	有自己的房子A3	信贷情况A4	类别
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否

数据集D的基尼值定义为

$$Gini(D) = 1 - \sum_{k=1}^{y} p_k^2 = 1 - \left(\frac{6}{15}\right)^2 - \left(\frac{9}{15}\right)^2 = 0.48$$

当按照是否有房子（二分类）划分：

$$Gini(\text{A3} = \text{是}) = 1 - \left(\frac{6}{6}\right)^2 - \left(\frac{0}{6}\right)^2 = 0$$

$$Gini(\text{A3} = \text{否}) = 1 - \left(\frac{3}{9}\right)^2 - \left(\frac{6}{9}\right)^2 = 0.444$$

$$Gini\_index(D, \text{A3}) = \left(\frac{6}{15}\right) * Gini(\text{A3} = \text{是}) + \left(\frac{9}{15}\right) * Gini(\text{A3} = \text{否}) = 0.2667$$

2.3、CART算法

ID	年龄A1	类别
1	青年	否
2	青年	否
3	青年	是
4	青年	是
5	青年	否
6	中年	否
7	中年	否
8	中年	是
9	中年	是
10	中年	是
11	老年	是
12	老年	是
13	老年	是
14	老年	是
15	老年	否

当按照年龄A1（三分类）划分：

为保证二分，将数据分成以下三种情况

{青年}，{中年，老年}；{中年}，{青年，老年}；{老年}，{青年，中年}

以{青年}，{中年，老年}为例计算基尼指数

$$Gini(\text{A1} = \text{青年}) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$$

$$Gini(\text{A1} = \text{中年，老年}) = 1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2 = 0.42$$

$$Gini\_index(D, \text{A1}) = \left(\frac{5}{15}\right) * Gini(\text{A1} = \text{青年}) + \left(\frac{10}{15}\right) * Gini(\text{A1} = \text{中年老年}) = 0.44$$

计算其余两种分类情况下的基尼指数，选取基尼指数最小的分类方式，然后在所有属性中寻找基尼指数最小的属性作为根结点，并递归找到其他结点

3.1 鸢尾花分类

- （1）数据集介绍

3类鸢尾花：山鸢尾、杂色鸢尾、维吉尼亚鸢尾

山鸢尾

杂色鸢尾

维吉尼亚鸢尾

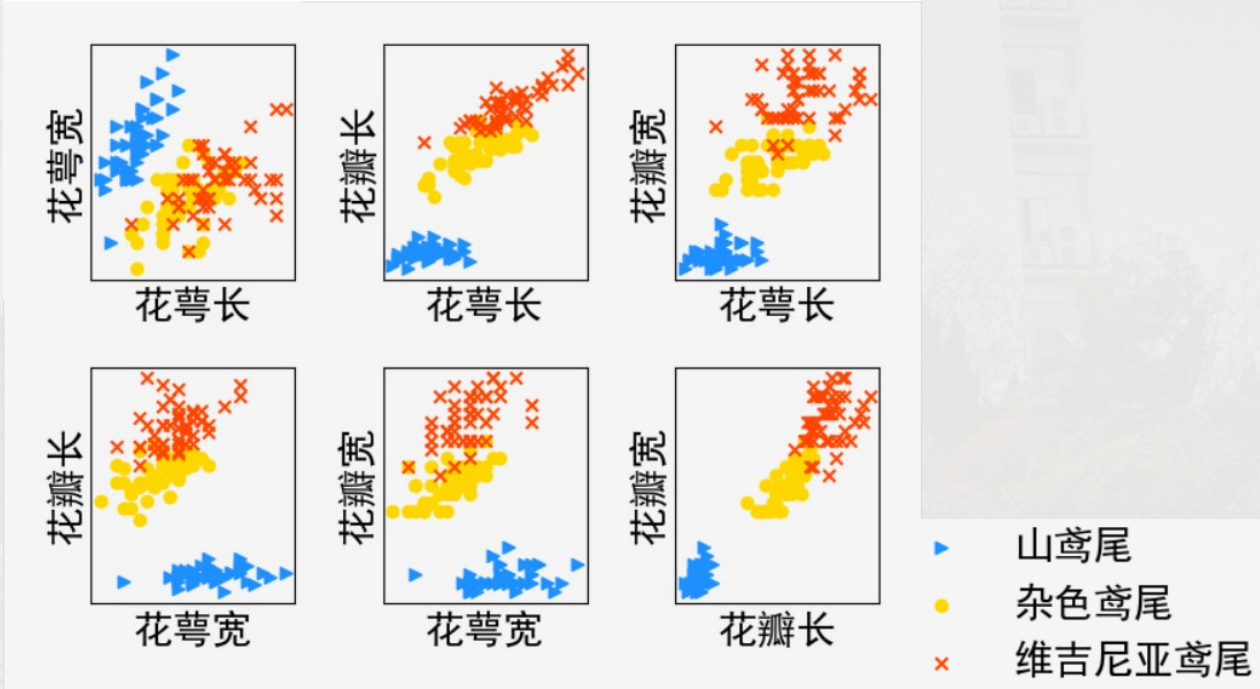
3.1 鸢尾花分类

➢ （1）数据集介绍

3类鸢尾花：山鸢尾、杂色鸢尾、维吉尼亚鸢尾

4维特征：花萼长、花萼宽、花瓣长、花瓣宽

	山鸢尾	杂色鸢尾	维吉尼亚鸢尾
花萼长-均值	5.0	5.9	6.6
花萼宽-均值	3.4	2.8	2.9
花瓣长-均值	1.5	4.3	5.6
花瓣宽-均值	0.3	1.3	2.0



3.1 鸢尾花分类

➢ （2）决策树生成

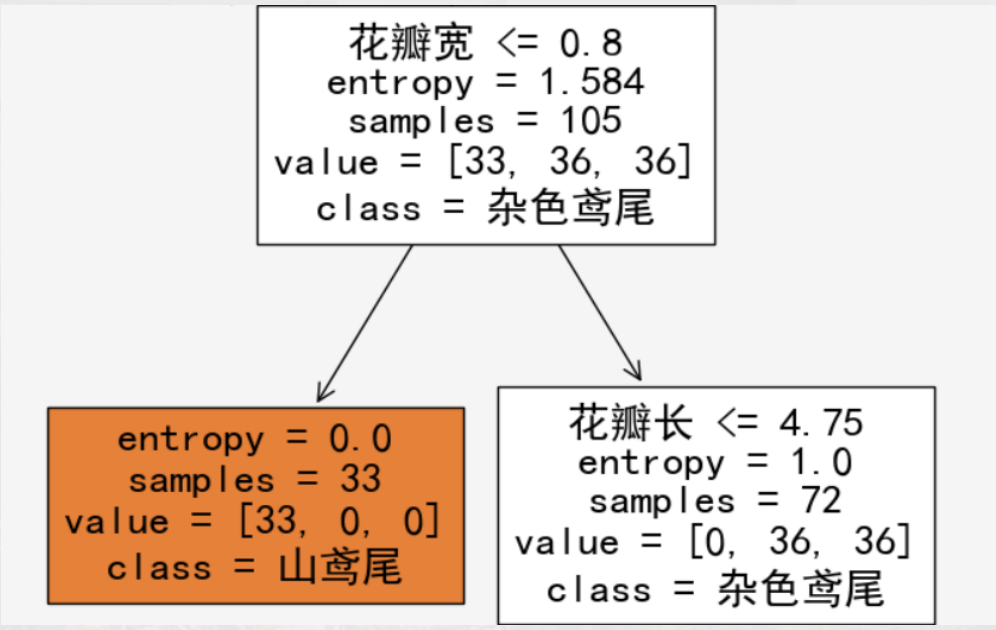
决策准则：ID3，选择最大化信息增益的特征划分准则

样本数：150，训练集样本：150*0.7=105，测试集样本：150*0.3=45

$$H(D) = -\sum \frac{C_k}{D} log_2 \frac{C_k}{D}$$
$$= -\left(\frac{33}{105} log_2 \frac{33}{105} + \frac{36}{105} log_2 \frac{36}{105} + \frac{36}{105} log_2 \frac{36}{105}\right)$$
$$= 1.584$$

$$H(D|A) = \sum \frac{D_i}{D} H(D_i)$$
$$= \frac{33}{105} * \left(-\frac{33}{33} log_2 \frac{33}{33}\right) + \frac{72}{105} * \left[-\left(\frac{36}{72} log_2 \frac{36}{72} + \frac{36}{72} log_2 \frac{36}{72}\right)\right]$$
$$= 0.686$$

$$Gain = H(D) - H(D|A) = 0.898$$



3.1 鸢尾花分类

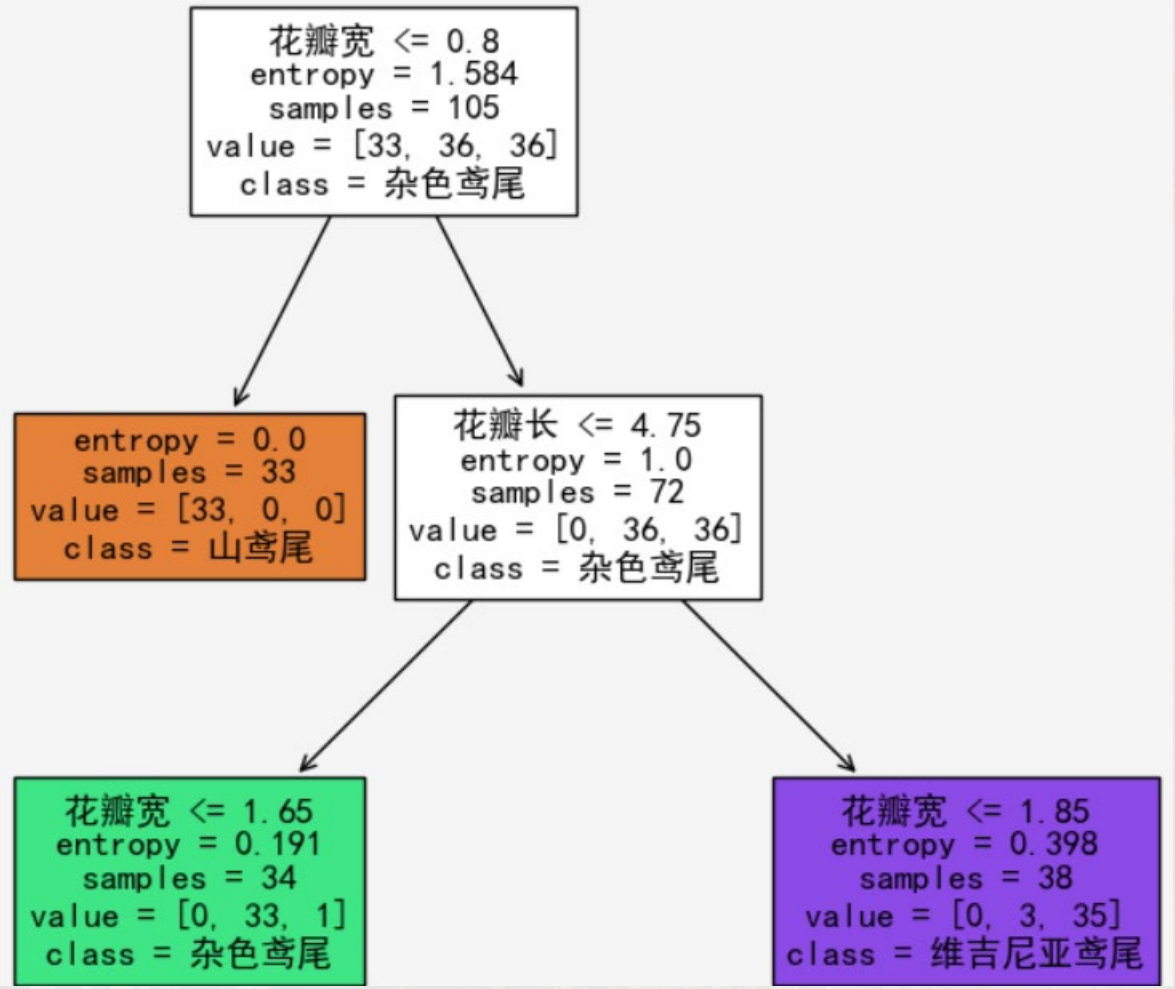
➢ （2）决策树生成

继续确定决策节点

$$H(D) = -\sum \frac{C_k}{D} log_2 \frac{C_k}{D} = 1.0$$

$$H(D|A) = \sum \frac{D_i}{D} H(D_i)$$
$$= \frac{34}{72} * 0.191 + \frac{38}{72} * 0.398$$
$$= 0.3$$

$$Gain = H(D) - H(D|A) = 0.7$$



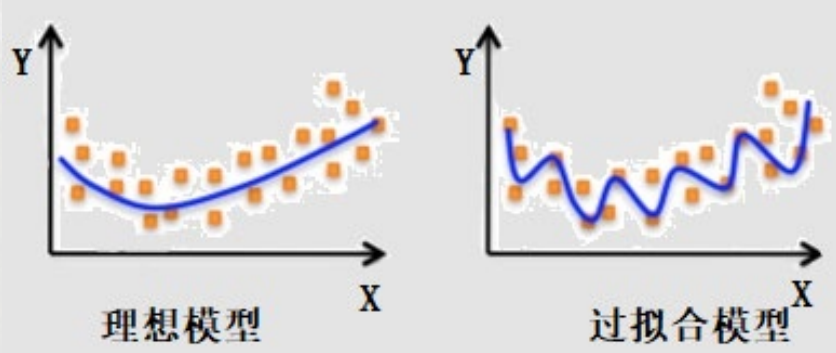
3.1 鸢尾花分类

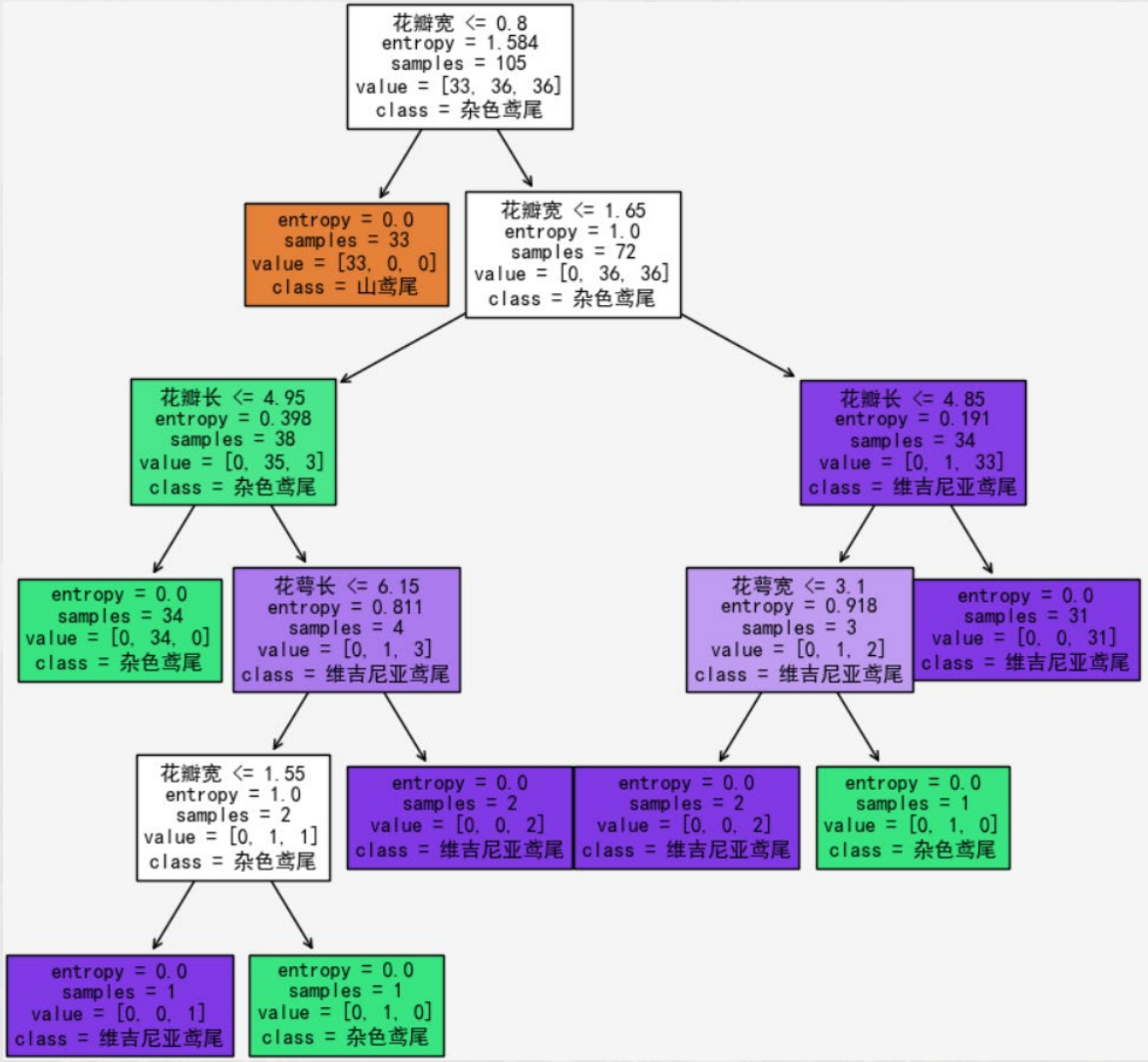
- （2）决策树生成

迭代生成决策树

训练集准确率：100%
测试集准确率：93.3%

- 存在过拟合问题





3.1 鸢尾花分类

（3）避免过拟合

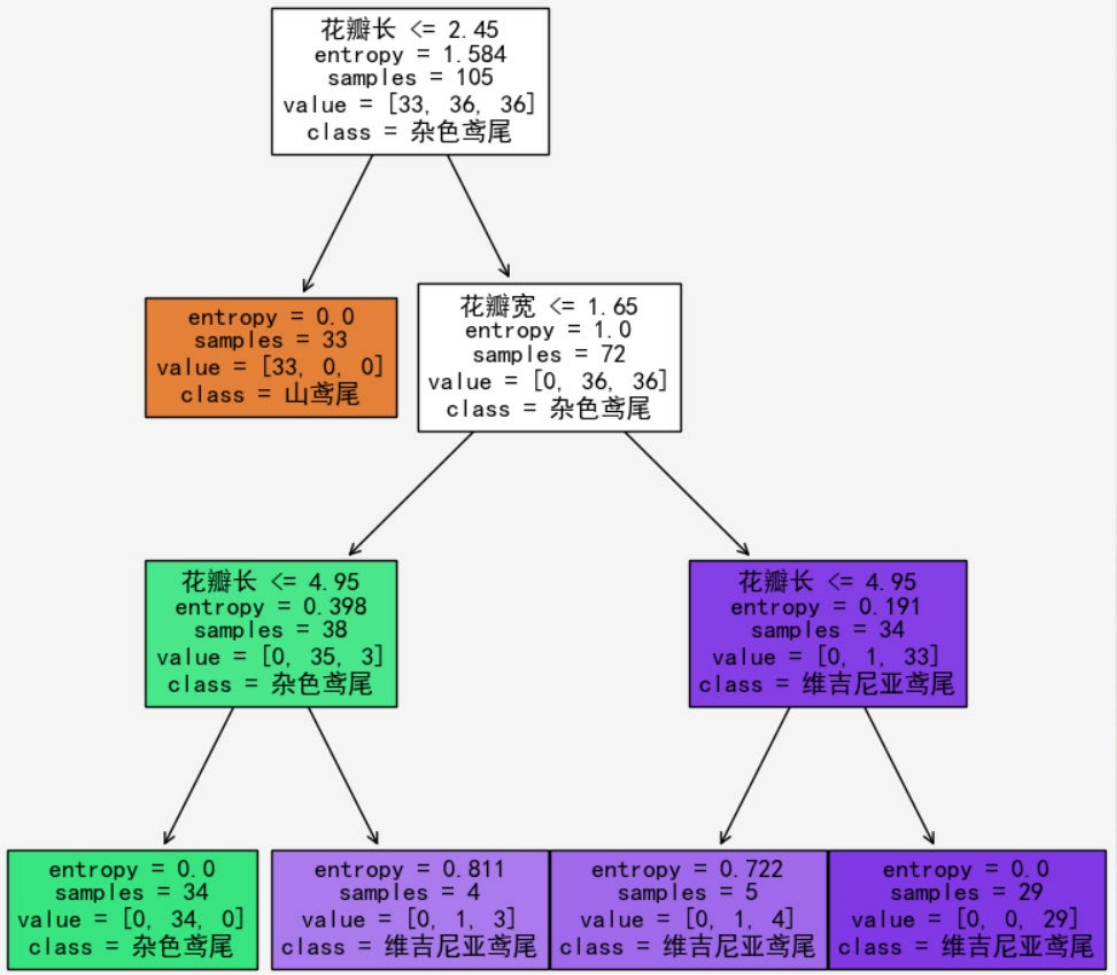
预剪枝：

通过样本数控制决策树深度

样本数≤ 4，停止分裂

训练集准确率：98.1%

测试集准确率：97.8%

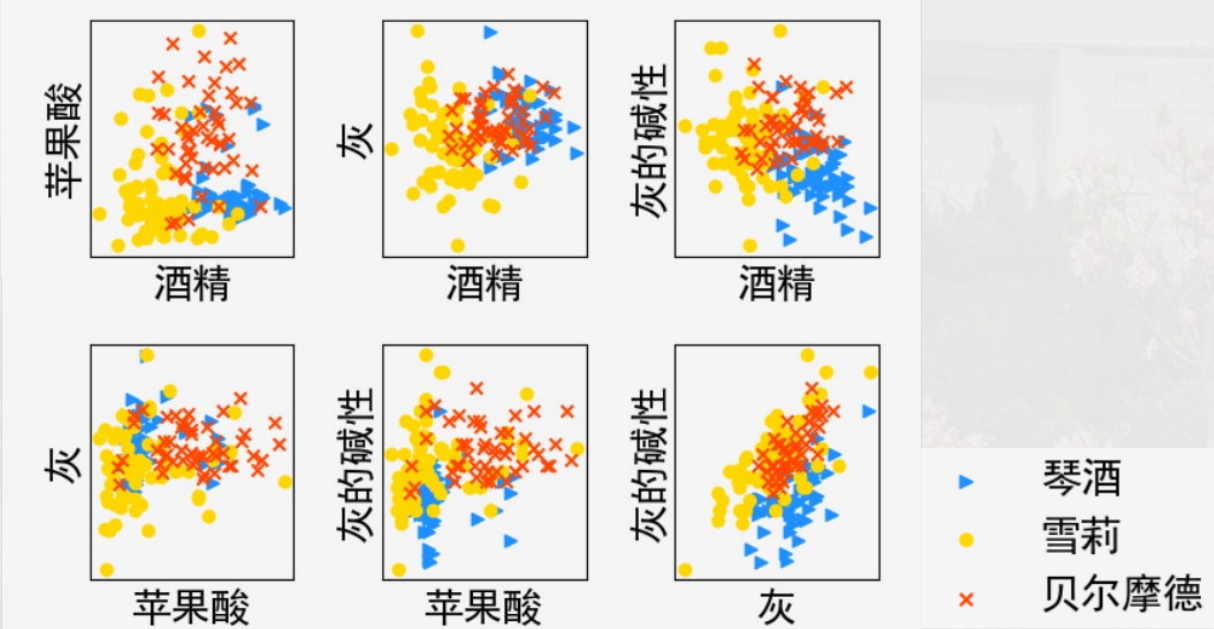


3.2 红酒分类

➢ （1）数据集介绍

3类红酒：琴酒，雪莉，贝尔摩德

13维特征：'酒精','苹果酸','灰','灰的碱性','镁','总酚','类黄酮','非黄烷类酚类',
'花青素','颜色强度','色调','od280/od315稀释葡萄酒','脯氨酸 ‘



3.2 红酒分类

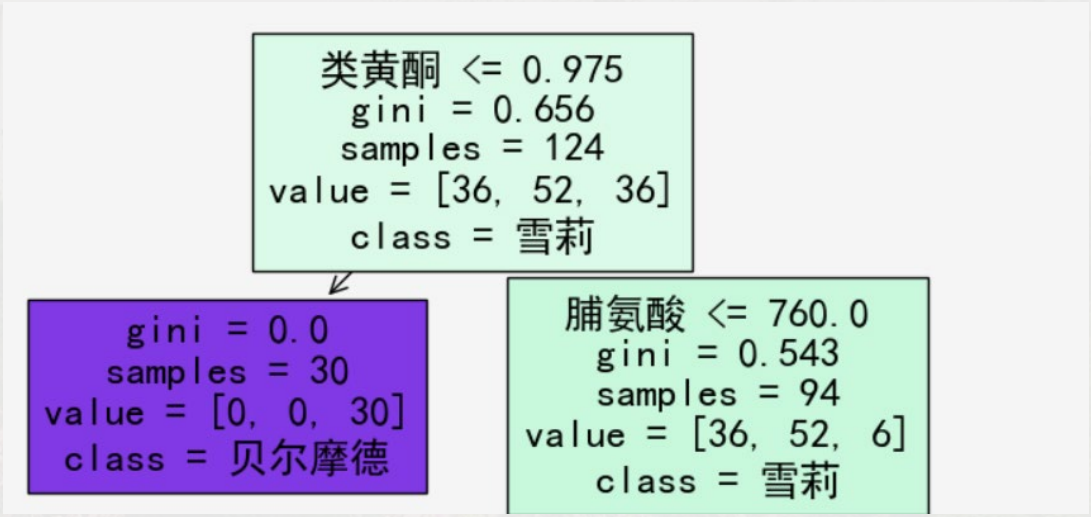
➢ （2）决策树生成

决策准则：GINI，选择最小化GINI系数的特征划分准则

样本数：178，训练集样本：178*0.7=124，测试集样本：178*0.3=54

$$\mathrm{Gini(D)} = 1 - \sum \mathrm{p}_k^2$$
$$= 1 - \left(\frac{36}{124}^2 + \frac{52}{124}^2 + \frac{36}{124}^2\right)$$
$$= 0.656$$

$$Gini\_index(D) = \sum \frac{D_i}{D} Gini(D_i)$$
$$= \frac{30}{124} * 0.0 + \frac{94}{124} * 0.543$$
$$= 0.412$$



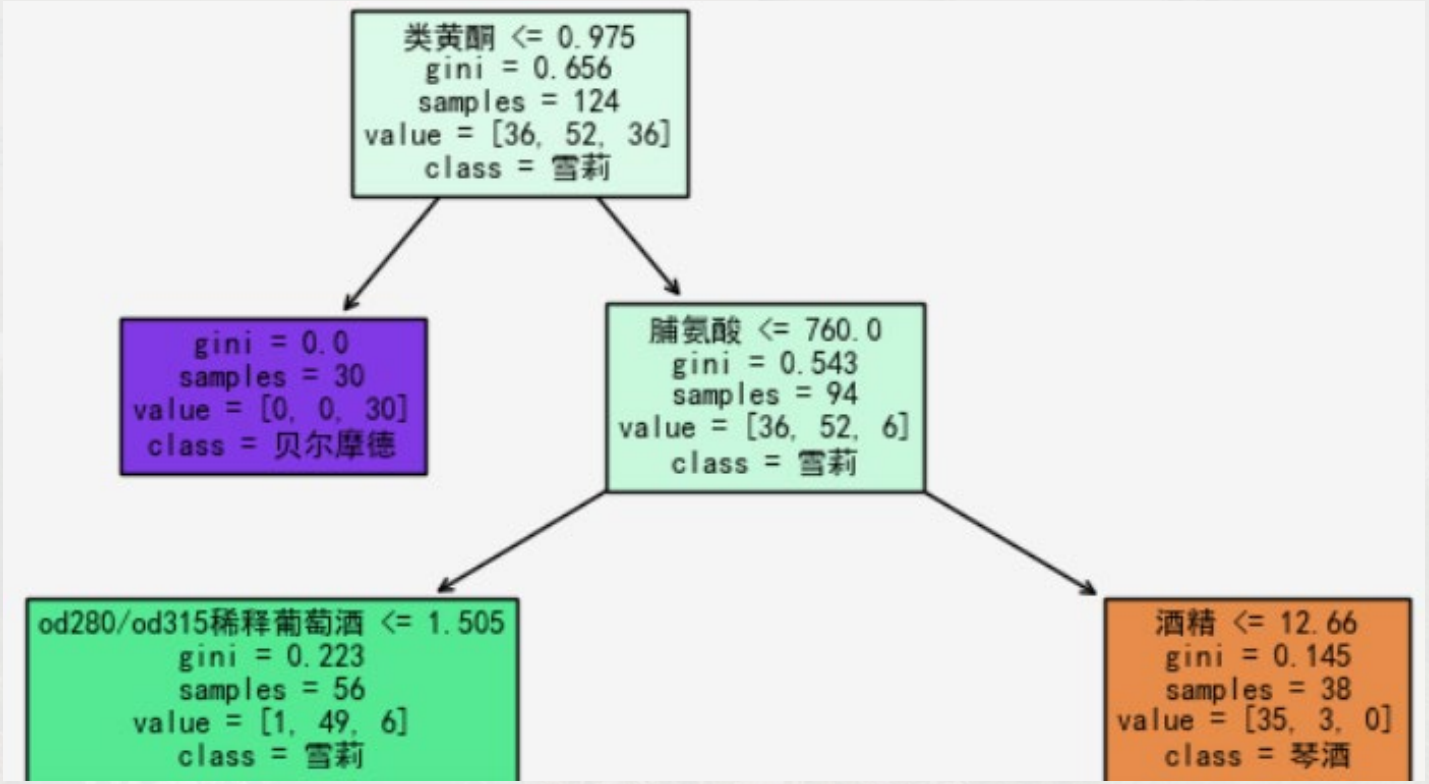
3.2 红酒分类

➢ （2）决策树生成

继续确定决策节点

$$\text{Gini(D)} = 1 - \sum \text{p}_k^2$$
$$= 0.543$$

$$Gini\_index(D) = \sum \frac{D_i}{D} Gini(D_i)$$
$$= \frac{56}{94} * 0.223 + \frac{38}{94} * 0.145$$
$$= 0.191$$



3.2 红酒分类

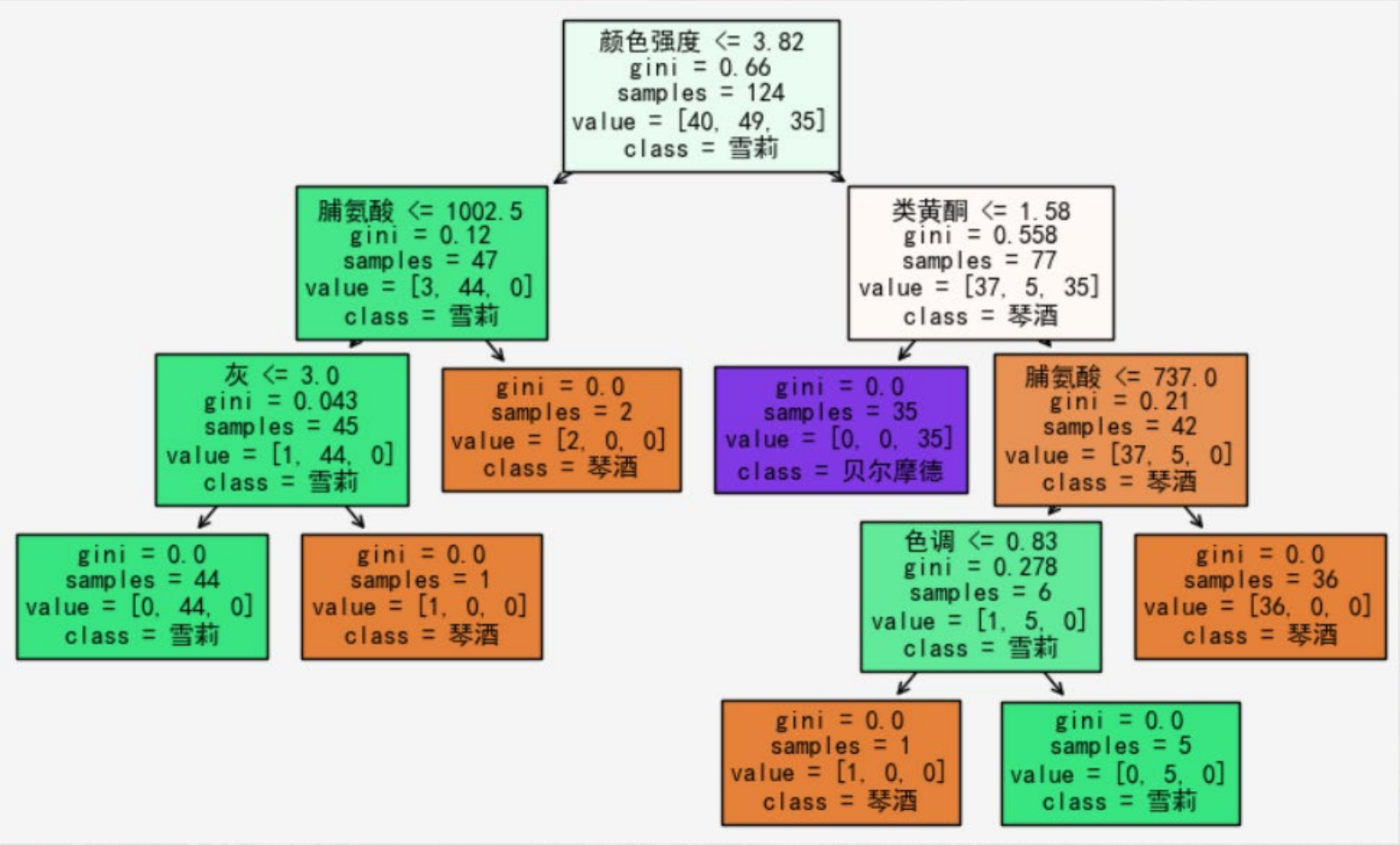
- （2）决策树生成

迭代生成决策树

训练集准确率：100%

测试集准确率：92.6%

- 存在过拟合问题



3.2 红酒分类

➢ （3）避免过拟合

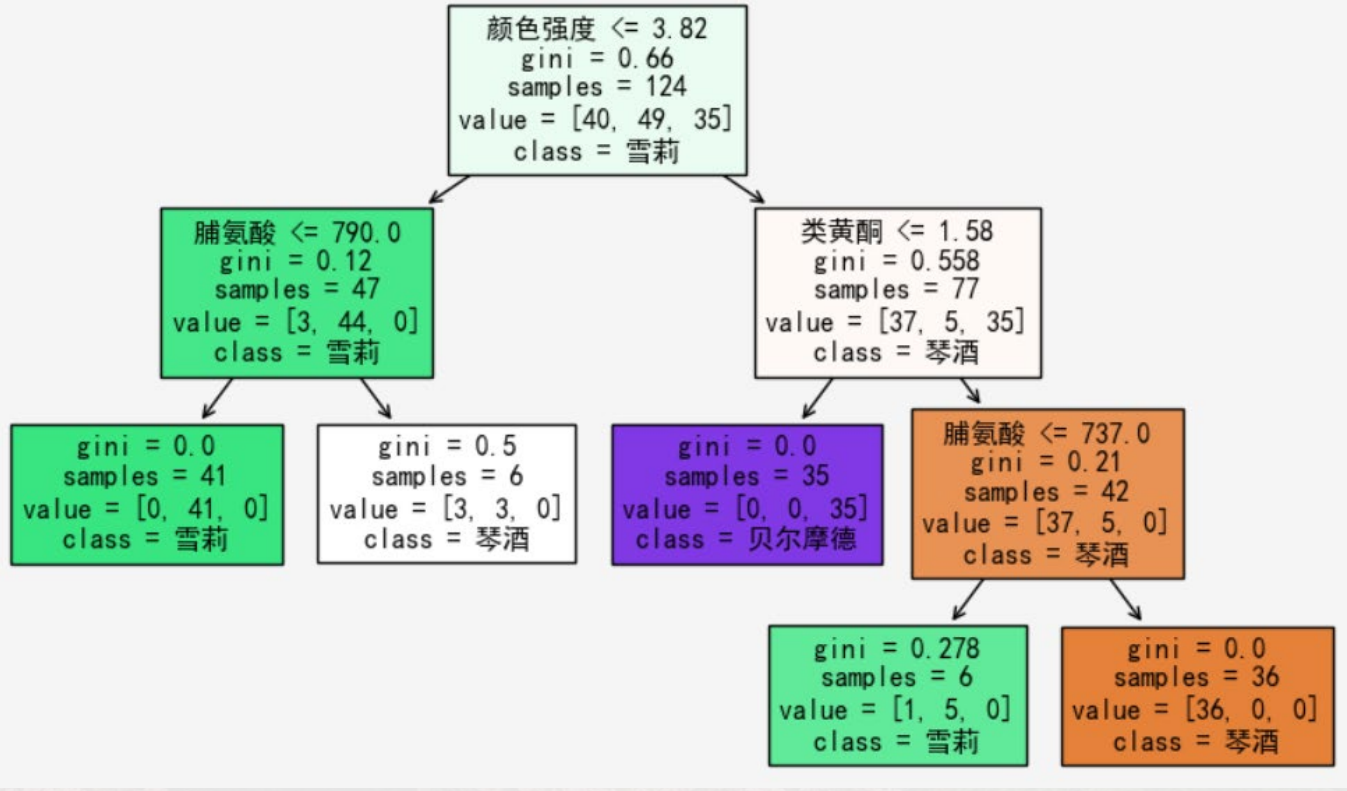
预剪枝：

通过样本数控制决策树深度

样本数≤ 5，停止分裂

训练集准确率：96.8%

测试集准确率：94.4%



本章小结

- 决策树的基本组成
- 构建决策树的三种算法：ID3，C4.5， CART
- 剪枝处理、连续值处理以及缺失值处理的常见操作方法
- 决策树实例：鸢尾花分类，红酒分类