

图解机器学习 | 决策树模型详解

韩信子 2022-03-08 25250 1 人工智能 机器学习 决策树

作者: 韩信子@ShowMeAI
教程地址: <https://www.showmeai.tech/tutorials/34>
本文地址: <https://www.showmeai.tech/article-detail/190>
声明: 版权所有, 转载请联系平台与作者并注明出处

引言

决策树 (Decision Tree) 是机器学习中一种经典的分类与回归算法。在本篇中我们讨论用于分类的决策树的原理知识。决策树模型呈树形结构, 在分类问题中, 一颗决策树可以视作 if-then 规则的集合。模型具有可读性, 分类速度快的特点, 在各种实际业务建模过程中广泛使用。

(本篇内容会涉及到不少机器学习基础知识, 没有先序知识储备的宝宝可以查看ShowMeAI的文章 [图解机器学习 | 机器学习基础知识](#)。

1.决策树算法核心思想

1) 决策树结构与核心思想

决策树 (Decision tree) 是基于已知各种情况 (特征取值) 的基础上, 通过构建树型决策结构来进行分析的一种方式, 是常用的有监督的分类算法。

决策树模型 (Decision Tree model) 模拟人类决策过程。以买衣服为例, 一个顾客在商店买裤子, 于是有了下面的对话:



顾客: 什么材料?
售货员: 牛仔

顾客: 裤型修身吗?

材料

牛仔
裤型

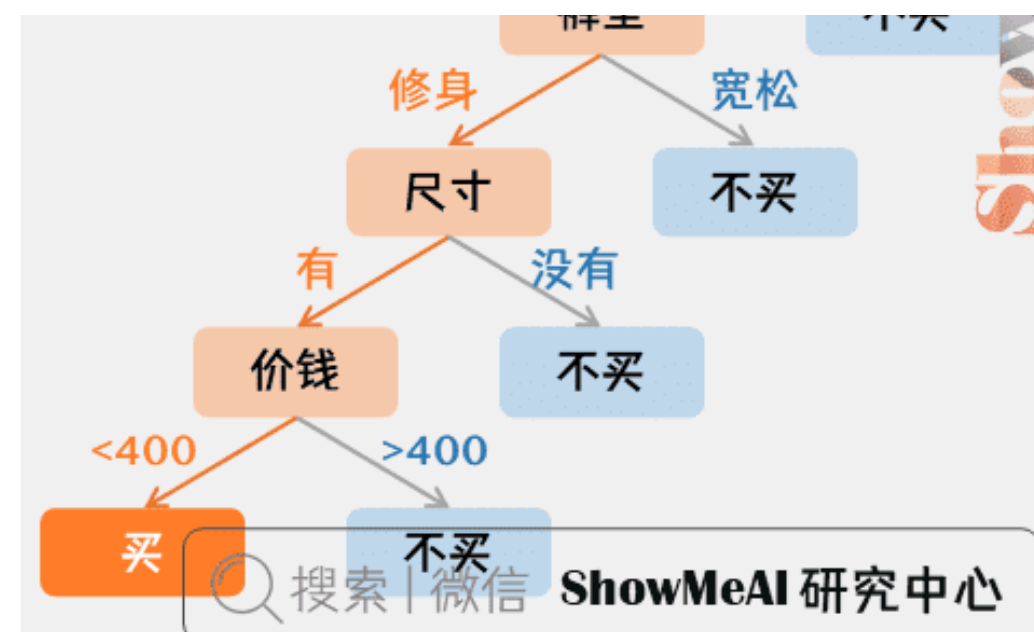
非牛仔
不买

决策树算法详解

算法核心思想

结构

<http://www.showmeai.tech/>



决策树是一种预测模型，代表的是对象属性与对象值之间的映射关系。决策树是一种树形结构，其中：

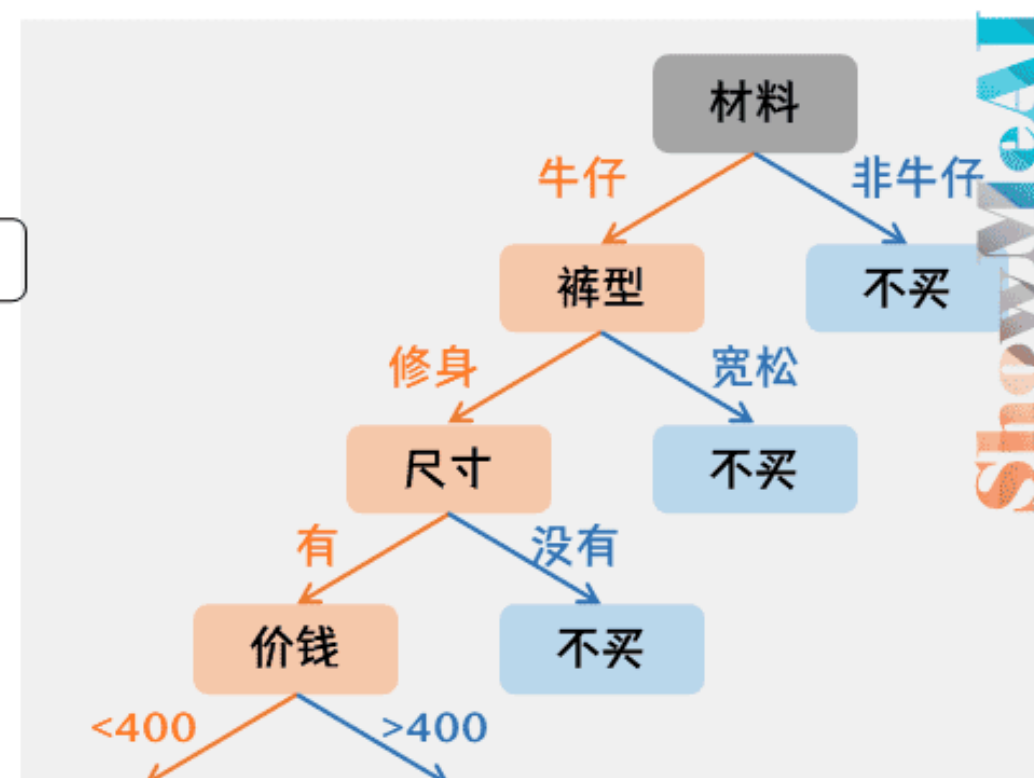
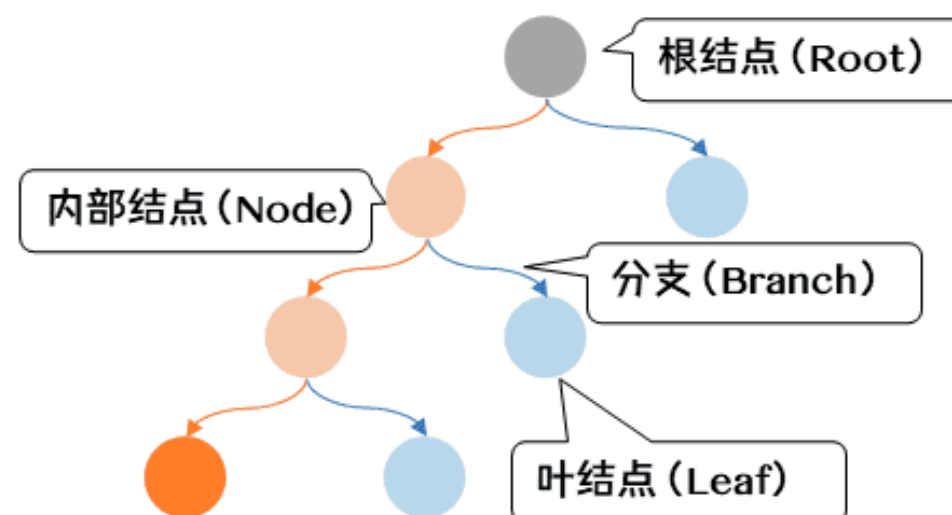
- 每个内部结点表示一个属性的测试
- 每个分支表示一个测试输出
- 每个叶结点代表一种类别



决策树算法详解

算法核心思想

结构



如上图买衣服的例子，第一个「内部结点」对应于属性「材料」上的测试，两个分支分别是该属性取值为「牛仔」和「非牛仔」两种可能结果。当取值为「牛仔」时，则对下个属性「裤型」进行测试；若取值为「非牛仔」时，则对应于「叶结点」——「不买」。

决策树模型核心是下面几部分：

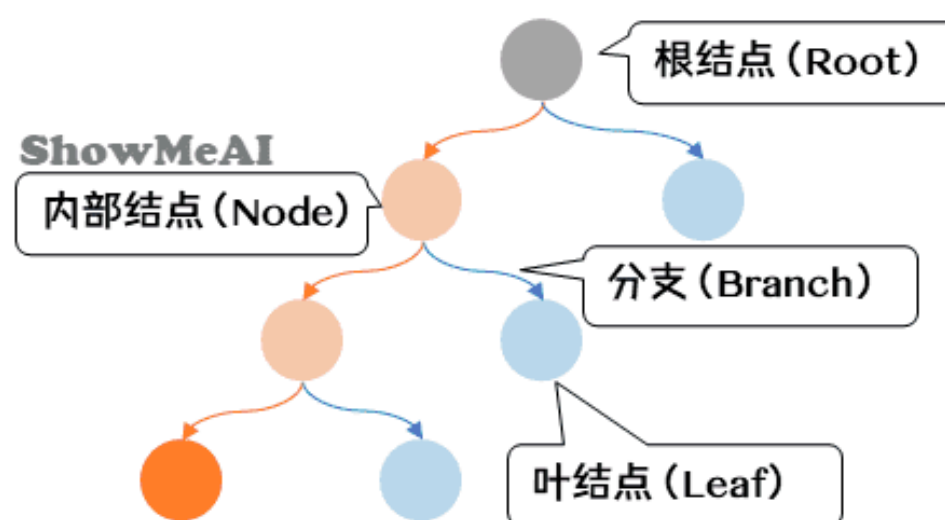
- 结点和有向边组成。
- 结点有内部结点和叶结点两种类型。
- 内部结点表示一个特征，叶结点表示一个类。



决策树算法详解

算法核心思想

结构



学习过程：


通过对训练样本的分析来确定“划分属性”
(即内部结点所对应的属性)

预测过程：

将测试示例从根结点开始，沿着划分属性所
构成的“判定测试序列”下行，直到叶结点

2) 决策树的发展史

决策树在发展过程中，有过很多不同类型的模型，典型的模型如ID3、C4.5和CART等，下面我们来简单介绍一下发展史中不同的模型。




决策树算法详解

算法核心思想

决策树的发展史

<http://www.showmeai.tech/>



CLS

ID3

C4.5

CART

RF

CLS (Concept Learning System) 是最早的决策树算法。

ID3是主流决策树算法之一，基于信息增益进行特征选择和树的生长。

C4.5是ID3的改进，基于信息增益率来选择最优属性。

CART是可用于分类与回归任务的二叉决策树，广泛使用。

RF (Random Forest, 随机森林) 是把决策树并行集成的组合算法。

GBDT/XGBoost/LightGBM/catboost: boosting系列模型，可以基于树模型做串行集成。

搜索 | 微信 ShowMeAI 研究中心

2.决策树生长与最优属性的选择

上面介绍的决策树发展史里，大家对于不同的决策树模型有一个基础的理解了，下面一部分，我们来一起看一下决策树是如何生长构成的。

1) 决策树生长流程

决策树的决策过程就是从根结点开始，测试待分类项中对应的特征属性，并按照其值选择输出分支，直到叶子结点，将叶子结点的存放的类别作为决策结果。简单说来，决策树的总体流程是自根至叶的递归过程，在每个中间结点寻找一个「划分」(split or test) 属性。

如下图的伪代码，是详细的决策树生长（构建）流程。大家可以特别注意图中3类终止条件和返回的结果，而整个流程中，有非常核心的一步是「最优划分属性的选择」。



决策树算法详解

决策树生长流程

<http://www.showmeai.tech/>

输入 训练集 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$
属性集 $A = \{a_1, a_2, \dots, a_d\}$
过程 函数 $TreeGenerate(D, A)$

- 1: 生成结点 $node$;
- 2: if D 中样本全属于同一类别 C then
- 3: 将 $node$ 标记为 C 类叶结点; return
- 4: end if
- 5: if $A = \emptyset$ OR D 中样本在 A 上取值相同 then
- 6: 将 $node$ 标记为叶结点, 其类别标记为 D 中样本数最多的类; return
- 7: end if
- 8: 从 A 中选择最优划分属性 a_* ;
- 9: for a_* 的每一个值 a_*^v do
- 10: 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;
- 11: if D_v 为空 then
- 12: 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
- 13: else
- 14: 以 $TreeGenerate(D, A \setminus \{a_*\})$ 为分支结点
- 15: end if
- 16: end for

输出 以 $node$ 为根结点的一棵决策树

停止条件(1)情形
递归返回

停止条件(2)情形
递归返回

停止条件(3)情形
递归返回

利用当前结点的
后验分布

决策树算法的核心

将父结点的样本分布作为
当前结点的先验分布
ShowMeAI 研究中心

搜索 | 微信

决策树停止生长的三个条件:



决策树算法

停止条件(1)

当前结点包含的样全属于同一类别。无需划分。

停止条件(2)

样本的属性取值都相同或属性集为空。不能划分。

ShowMeAI

2) 最优属性选择

下面我们来看看，决策树的最优划分属性选择，是怎么做的。

(1) 信息熵

要解决决策树的「最优属性」选择，我们需要先了解一个信息论的概念「**信息熵 (entropy)**」(相关知识可以参考ShowMeAI文章 [图解AI数学基础 | 信息论](#))，它是消除不确定性所需信息量的度量，也是未知事件可能含有的信息量，可以度量样本集合「纯度」。

对应到机器学习中，假定当前数据集 D 中有 y 类，其中第 k 类样本占比为 p_k ，则信息熵的计算公式如下：

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



决策树算法详解

最优属性选择

D的信息熵

假定当前样本集合 D 中共有 y 类，其中第 k 类样本所占的比例为 p_k

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

计算信息熵时约定：

- 若 $p = 0$ ，则 $p \log_2 p = 0$
- $Ent(D)$ 的最小值为0，最大值为 $\log_2 |y|$

但 p_k 取值为 1 的时候，信息熵为 0（很显然这时候概率 1 表示确定事件，没有任何不确定性）；而当 p_k 是均匀分布的时候，信息熵取最大值 $\log(|y|)$ （此时所有候选同等概率，不确定性最大）。

(2) 信息增益

大家对信息熵有了解后，我们就可以进一步了解信息增益（Information Gain），它衡量的是我们**选择某个属性进行划分时信息熵的变化**（可以理解为基于这个规则划分，不确定性降低的程度）。

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



决策树算法详解

最优属性选择

信息增益-ID3

信息增益 (ID3中使用)

离散属性 a 的取值 $\{a^1, a^2, a^3, \dots, a^v\}$; D^v : 样本集合D中在 a 上取值= a^v 的样本集合

以属性 a 对数据D进行划分
所获得的信息增益

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

第v个分支的权重
样本越来越重要

划分前的信息熵

划分后的信息熵

信息增益描述了一个特征带来的信息量的多少。在决策树分类问题中，信息增益就是决策树在进行属性选择划分前和划分后的信息差值。典型的决策树算法ID3就是基于信息增益来挑选每一节点分支用于划分的属性（特征）的。

这里以西瓜数据集为例。

- 数据集分为好瓜、坏瓜，所以 $|y| = 2$ 。
- 根结点包含 17 个训练样例，其中好瓜共计 8 个样例，所占比例为 $8/17$ 。
- 坏瓜共计 9 个样例，所占比例为 $9/17$ 。

将数据带入信息熵公式，即可得到根节点的信息熵。



决策树算法详解

最优属性选择

信息增益示例

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

根节点的信息熵

17个训练样例，结果2个类别

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


$$|y| = 2, p_1 = 8/17, p_2 = 9/17$$

$$Ent(D) = - \left(\frac{8}{17} \log_2 \frac{8}{17} + \frac{9}{17} \log_2 \frac{9}{17} \right)$$

$$= 0.998$$

以属性「色泽」为例，其对应的 3 个数据子集：

- D_1 (色泽 = 青绿)，包含 {1, 4, 6, 10, 13, 17}，6 个样例，其中好瓜样例为 {1, 4, 6}，比例为 3/6，坏瓜样例为 {10, 13, 17}，比例为 3/6。将数据带入信息熵计算公式即可得到该结点的信息熵。
- D_2 (色泽 = 乌黑)，包含 {2, 3, 7, 8, 9, 15}，6 个样例，其中好瓜样例为 {2, 3, 7, 8}，比例为 4/6，坏瓜样例为 {9, 15}，比例为 2/6。将数据带入信息熵计算公式即可得到该结点的信息熵。
- D_3 (色泽 = 浅白)，包含 {5, 11, 12, 14, 16}，5 个样例，其中好瓜样例为 {5}，比例为 1/5，坏瓜样例为 {11, 12, 14, 16}，比例为 4/5。将数据带入信息熵计算公式即可得到该结点的信息熵。



决策树算法详解

最优属性选择

信息增益示例

<http://www.showmeai.tech/>

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

色泽属性的信息熵

17个训练样例，结果3个类别

$$Ent(D) = - \sum_{k=1}^{|y|} p_k \log_2 p_k$$
$$Ent(D^1) = - \left(\frac{3}{6} \log_2 \frac{3}{6} + \frac{3}{6} \log_2 \frac{3}{6} \right) = 1.000$$
$$Ent(D^2) = - \left(\frac{4}{6} \log_2 \frac{4}{6} + \frac{2}{6} \log_2 \frac{2}{6} \right) = 0.918$$
$$Ent(D^3) = - \left(\frac{1}{5} \log_2 \frac{1}{5} + \frac{4}{5} \log_2 \frac{4}{5} \right) = 0.722$$

搜索 | 微信 ShowMeAI 研究中心

案例来源：周志华老师《机器学习》西瓜数据集

色泽属性的信息增益为：



决策树算法详解

最优属性选择

信息增益示例

<http://www.showmeai.tech/>

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

色泽属性的信息增益

17个训练样例，结果3个类别

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

搜索 | 微信 ShowMeAI 研究中心

案例来源：周志华老师《机器学习》西瓜数据集

同样的方法，计算其他属性的信息增益为：



决策树算法详解

最优属性选择

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

其他属性的信息增益

17个训练样例，结果3个类别

$$\text{Gain}(D, \text{色泽}) = 0.109$$

$$\text{Gain}(D, \text{根蒂}) = 0.143$$

$$\text{Gain}(D, \text{敲声}) = 0.141$$

$$\text{Gain}(D, \text{纹理}) = 0.281$$

信息增益最大
被选为划分属性

信息增益示例

<http://www.showmeai.tech/>

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

$$Gain(D, 纹理) = 0.381$$

$$Gain(D, 脐部) = 0.289$$

$$Gain(D, 触感) = 0.006$$

搜索 | 微信 ShowMeAI 研究中心

案例来源: 周志华老师《机器学习》西瓜数据集

对比不同属性, 我们发现「纹理」信息增益最大, 其被选为划分属性: 清晰 {1, 2, 3, 4, 5, 6, 8, 10, 15}、稍糊 {7, 9, 13, 14, 17}、模糊 {11, 12, 16}。

再往下一步, 我们看看「纹理」=「清晰」的节点分支, 该节点包含的样例集合 D_1 中有编号为 {1, 2, 3, 4, 5, 6, 8, 10, 15} 共计 9 个样例, 可用属性集合为 {色泽, 根蒂, 敲声, 脐部, 触感} (此时「纹理」不再作为划分属性), 我们同样的方式再计算各属性的信息增益为:



决策树算法详解

最优属性选择

信息增益示例

<http://www.showmeai.tech/>

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

纹理 = 清晰

各属性的信息增益

$$Gain(D, 色泽) = 0.043$$

$$Gain(D, 根蒂) = 0.458$$

$$Gain(D, 敲声) = 0.331$$

$$Gain(D, 脐部) = 0.458$$

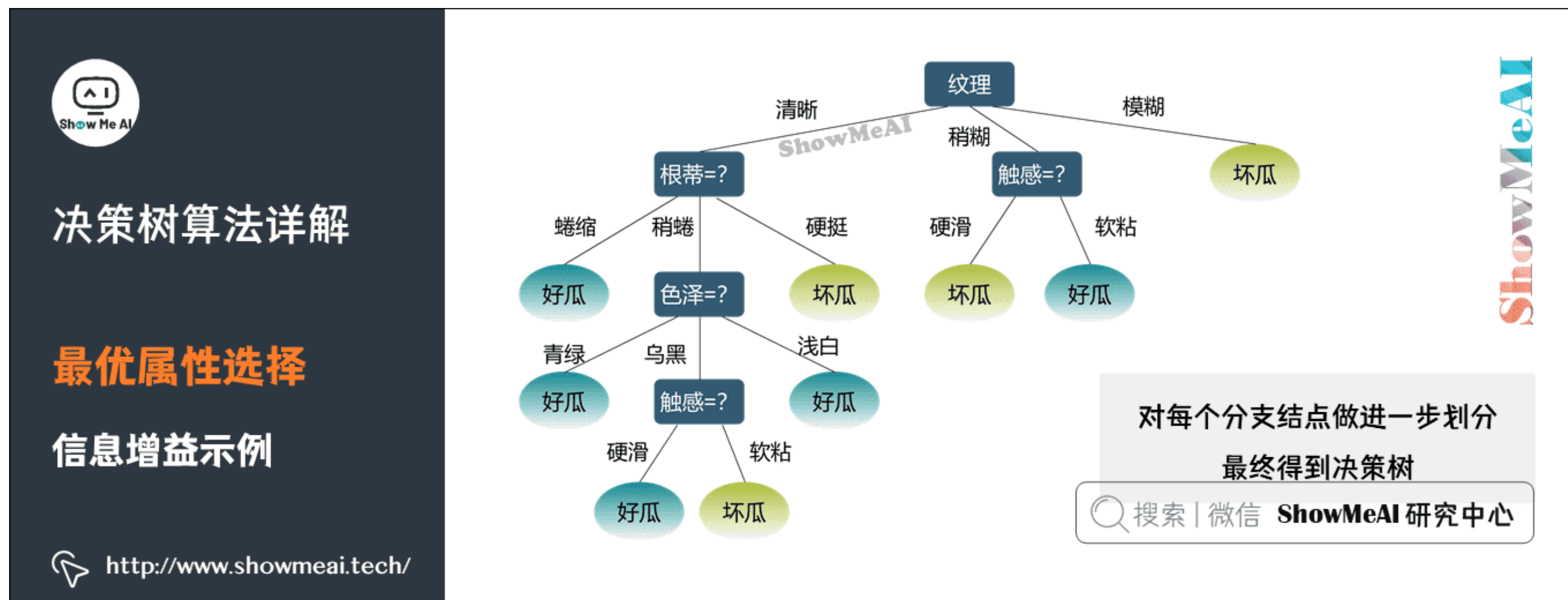
$$Gain(D, 触感) = 0.458$$

此时“纹理”不再
作为划分属性

搜索 | 微信 ShowMeAI 研究中心

案例来源: 周志华老师《机器学习》西瓜数据集

从上图可以看出「根蒂」、「脐部」、「触感」3个属性均取得了最大的信息增益，可用任选其一作为划分属性。同理，对每个分支结点进行类似操作，即可得到最终的决策树。



(3) 信息增益率 (Gain Ratio)

大家已经了解了信息增益作为特征选择的方法，但信息增益有一个问题，它偏向取值较多的特征。原因是，当特征的取值较多时，根据此特征划分更容易得到纯度更高的子集，因此划分之后的熵更低，由于划分前的熵是一定的。因此信息增益更大，因此信息增益比较偏向取值较多的特征。

那有没有解决这个小问题的方法呢？有的，这就是我们要提到信息增益率 (Gain Ratio)，信息增益率相比信息增益，多了一个衡量本身属性的分散程度的部分作为分母，而著名的决策树算法C4.5就是使用它作为划分属性挑选的原则。

信息增益率的计算细节如下所示：

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

$$IV(a)$$

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



决策树算法详解

最优属性选择

信息增益率-C4.5

<http://www.showmeai.tech/>

信息增益率 (C4.5中使用)

离散属性 a 的取值 $\{a^1, a^2, a^3, \dots, a^v\}$; D^v : 样本集合 D 中在 a 上取值 $=a^v$ 的样本集合

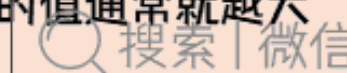
$$Gain_ratio(D, a) = \frac{Gain(D, a)}{IV(a)}$$



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

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, 再从中选取增益率最高的。

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



微信 ShowMeAI 研究中心

数学上用于信息量 (或者纯度) 衡量的不止有上述的熵相关的定义, 我们还可以使用基尼指数来表示数据集的不纯度。基尼指数越大, 表示数据集越不纯。

基尼指数 (Gini Index) 的详细计算方式如下所示:

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



决策树算法详解

最优属性选择

基尼系数-CART

<http://www.showmeai.tech/>

基尼指数 (CART 中使用)

离散属性 a 的取值 $\{a^1, a^2, a^3, \dots, a^v\}$; D^v : 样本集合 D 中在 a 上取值 $=a^v$ 的样本集合

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

- 反映了从 D 中随机抽取两个样例，其类别标记不一致的概率
- $Gini(D)$ 越小，数据集 D 的纯度越高

属性 a 的基尼指数 $Gini_index(D, a) = \sum_{v=1}^v \frac{|D^v|}{|D|} Gini(D^v)$

在候选属性集合中，选取那个使划分后基尼指数最小的属性

🔍 搜索 | 微信 ShowMeAI 研究中心

其中， p_k 表示第 k 类的数据占总数据的比例，著名的决策树算法CART就是使用基尼指数来进行划分属性的挑选（当然，CART本身是二叉树结构，这一点和上述的 ID3 和 C4.5 不太一样）。

对于基尼指数的一种理解方式是，之所以它可以用作纯度的度量，大家可以想象在一个漆黑的袋里摸球，有不同颜色的球，其中第 k 类占比记作 p_k ，那两次摸到的球都是第 k 类的概率就是 p_k^2 ，那两次摸到的球颜色不一致的概率就是 $1 - \sum p_k^2$ ，它的取值越小，两次摸球颜色不一致的概率就越小，纯度就越高。

3.过拟合与剪枝

如果我们让决策树一直生长，最后得到的决策树可能很庞大，而且因为对原始数据学习得过于充分会有过拟合的问题。缓解决策树过拟合可以通过剪枝操作完成。而剪枝方式又可以分为：预剪枝和后剪枝。

1) 决策树与剪枝操作

为了尽可能正确分类训练样本，有可能造成分支过多，造成过拟合。过拟合是指训练集上表现很好，但是在测试集上表现很差，泛化性能差。可以通过剪枝主动去掉一些分


支来降低过拟合的风险，并使用「留出法」进行评估剪枝前后决策树的优劣。

基本策略包含「预剪枝」和「后剪枝」两个：

- **预剪枝 (pre-pruning)**：在决策树生长过程中，对每个结点在划分前进行估计，若当前结点的划分不能带来决策树泛化性能的提升，则停止划分并将当前结点标记为叶结点。
- **后剪枝 (post-pruning)**：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决策树泛化性能的提升，则将该子树替换为叶结点。

2) 预剪枝与后剪枝案例

我们来看一个例子，下面的数据集，为了评价决策树模型的表现，会划分出一部分数据作为验证集。



决策树算法详解

预剪枝与后剪枝

<http://www.showmeai.tech/>

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
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

训练集

验证集

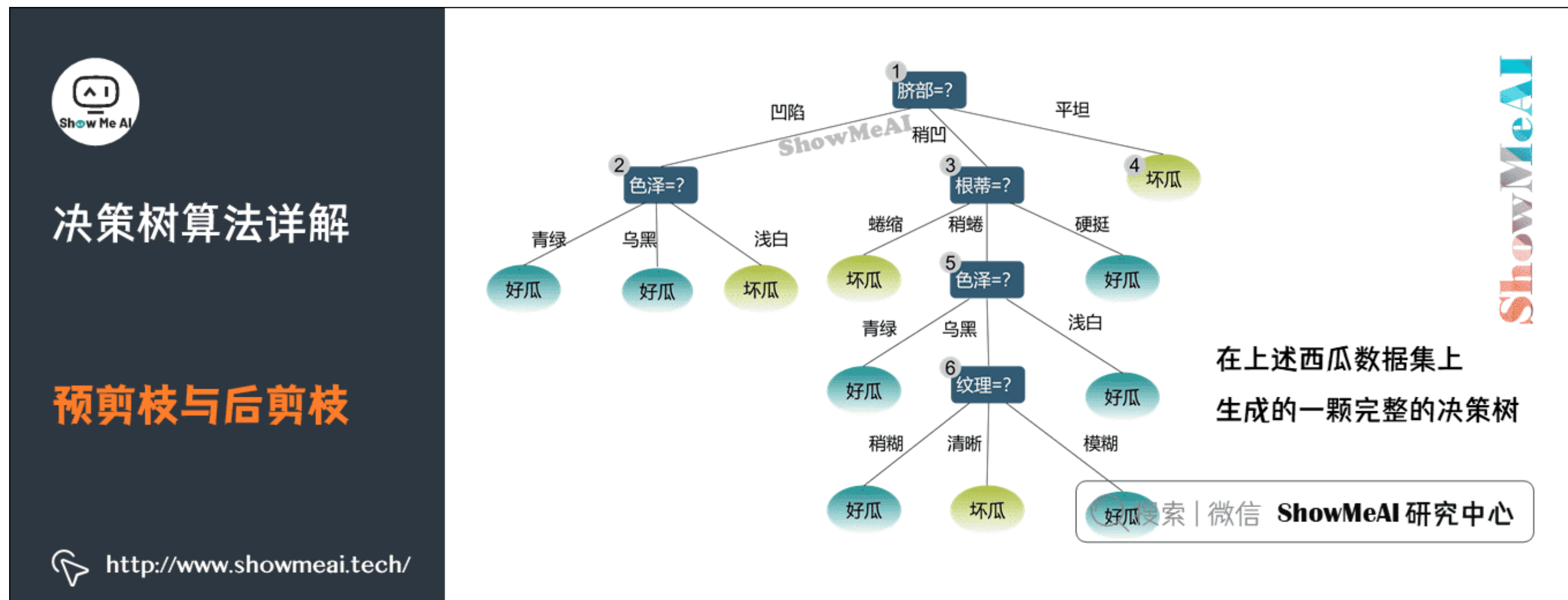
注意，
由于划分出了验证集，
生成此处决策树所使用的数据与前面生成决策树所使用的数据不同。

搜索 | 微信

ShowMeAI 研究中心

案例来源：周志华老师《机器学习》西瓜数据集

在上述西瓜数据集上生成的一颗完整的决策树，如下图所示。



(1) 预剪枝

「预剪枝」过程如下：将其标记为叶结点，类别标记为训练样例中最多的类别。

- 若选「好瓜」，验证集中 {4, 5, 8} 被分类正确，得到验证集精度为 $3/7 \times 100\% = 42.9\%$
- 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 $5/7 \times 100\%$
 - 结点2 (好瓜)：分类正确：{4, 5}，分类错误：{13}
 - 结点3 (好瓜)：分类正确：{8}，分类错误：{9}
 - 结点4 (坏瓜)：分类正确：{11, 12}



决策树算法详解

预剪枝与后剪枝

预剪枝过程

<http://www.showmeai.tech/>

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

验证集

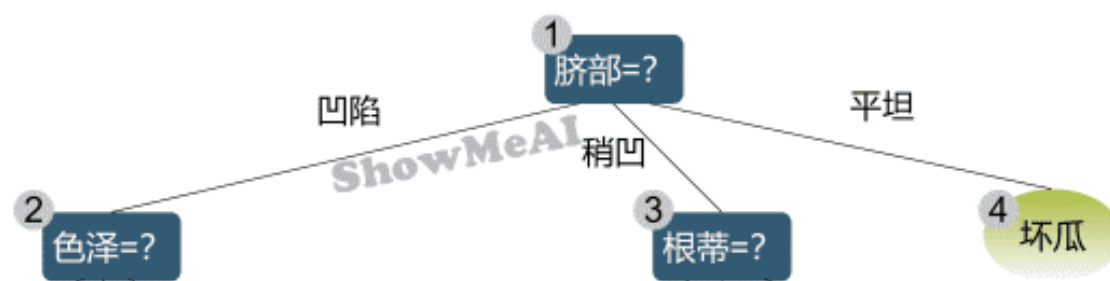
搜索 | 微信 ShowMeAI 研究中心

若划分后的验证集精度下降，则拒绝划分。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。

根据预剪枝方法，此处生成了一层决策树。这种最终得到仅有一层划分的决策树，称为「决策树桩」（decision stump）。



决策树算法详解



(脐部=?)验证集精度:
划分前: 42.9%
划分后: 71.4%
预剪枝决策: 划分

(色泽=?)验证集精度:

(根蒂=?)验证集精度:

预剪枝与后剪枝

预剪枝

<http://www.showmeai.tech/>

划分前: 71.4%
划分后: 57.1%
预剪枝决策: 禁止划分

划分前: 71.4%
划分后: 71.4%
预剪枝决策: 禁止划分

最终得到只有一层划分的决策树
称为“决策树桩”(decision stump)

搜索 | 微信 ShowMeAI 研究中心

(2) 后剪枝

我们在生成的完整决策树上进行「后剪枝」：

- 用验证集的数据对该决策树进行评估，样例 {4, 11, 12} 分类正确，而样例 {5, 8, 9, 13} 分类错误，此时的精度为 42.9。
- 当对该决策树进行后剪枝，结点⑥的标记为好瓜，此时样例 {4, 8, 11, 12} 分类正确，样例 {5, 9, 13} 分类错误，精度为 57.1。

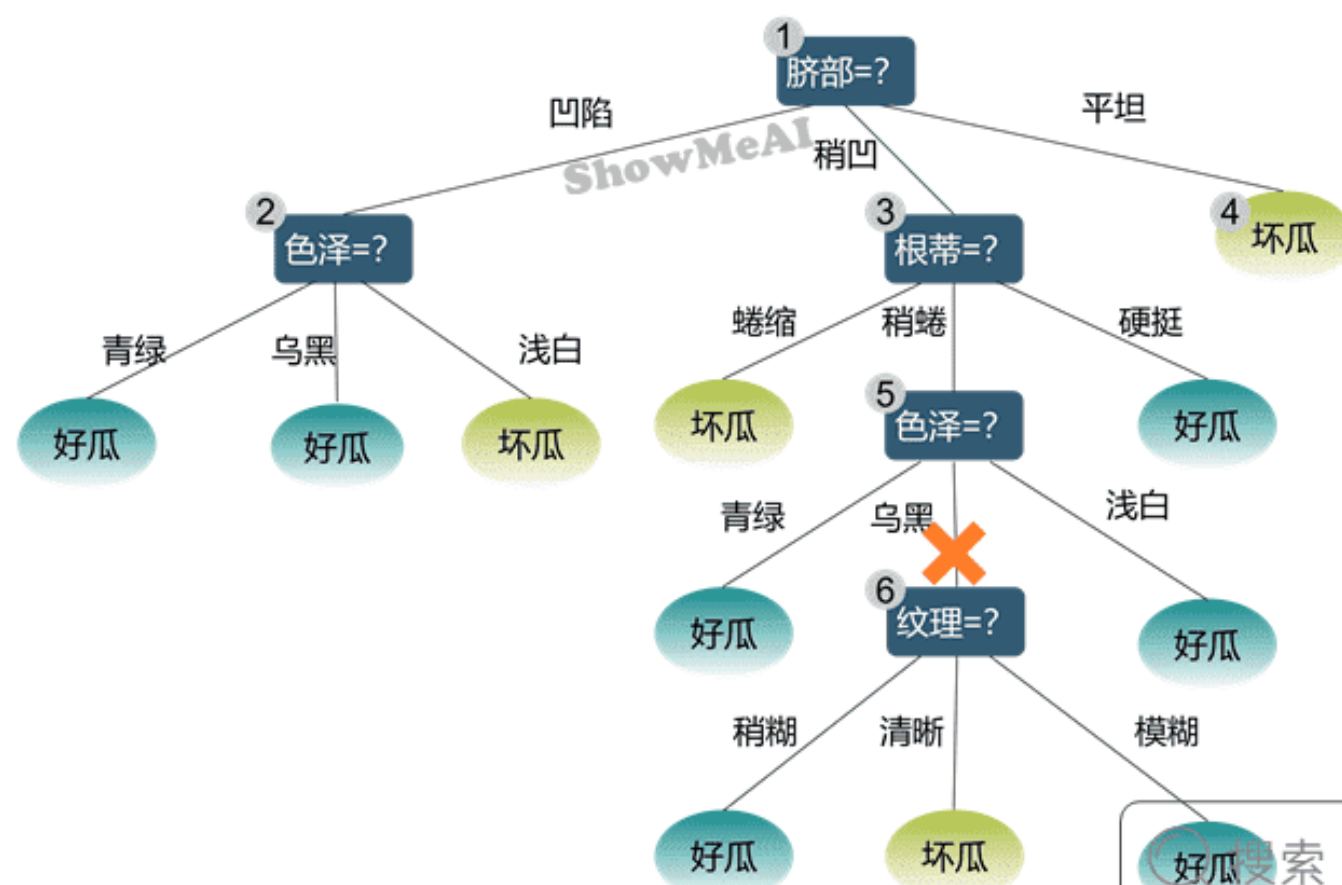
剪枝后的精度提升了，因此该决策树需要在结点 ⑥ 处进行剪枝。



决策树算法详解

预剪枝与后剪枝

后剪枝



(纹理=?) 验证集精度:
剪枝前: 42.9%
剪枝后: 57.1%
后剪枝决策: 剪枝

搜索 | 微信 ShowMeAI 研究中心

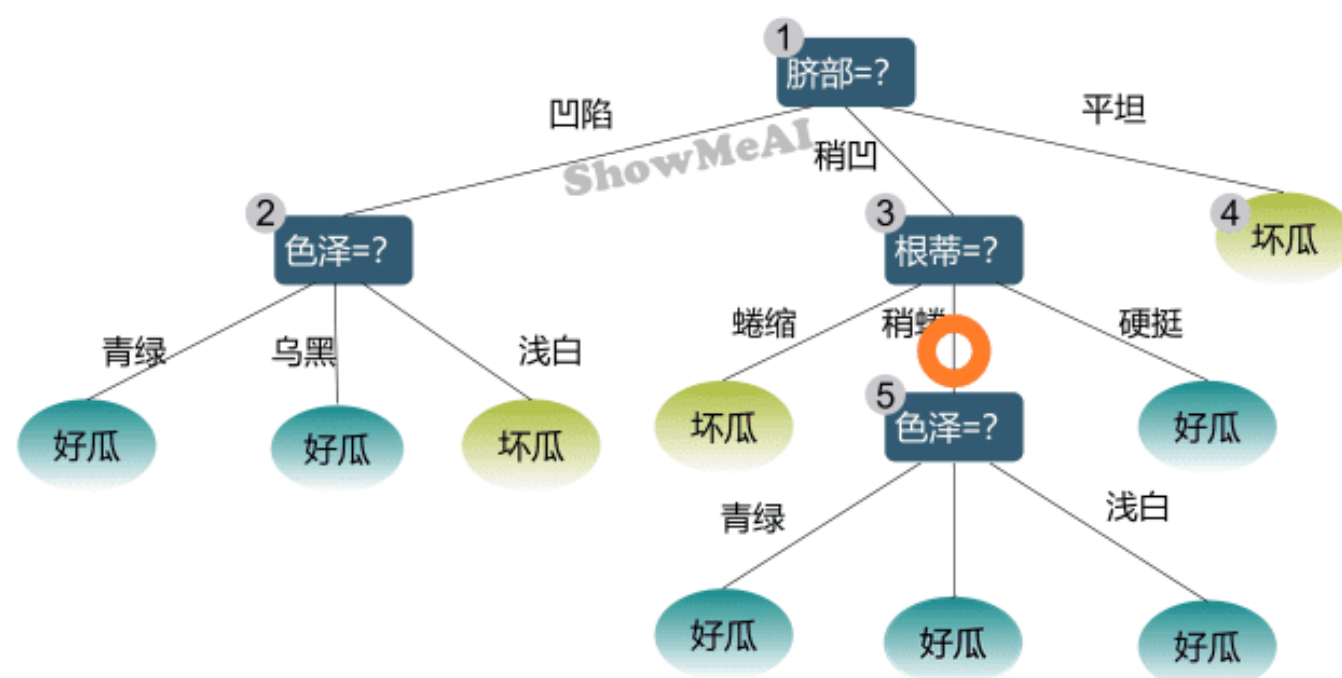
考虑结点⑤，若将其替换为叶结点，根据落在其上的训练样例 {6,7,15} 将其标记为「好瓜」，测得验证集精度仍为 57.1，可以不剪枝。



决策树算法详解

预剪枝与后剪枝

后剪枝



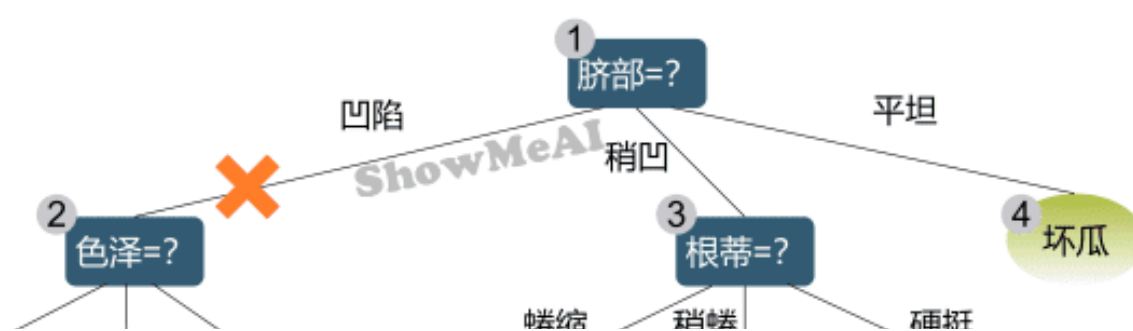
(色泽=?) 验证集精度:
剪枝前: 57.1%
剪枝后: 57.1%
后剪枝决策: 不剪枝

搜索 | 微信 ShowMeAI 研究中心

考虑结点②，若将其替换为叶结点，根据落在其上的训练样例 {1,2,3,14} 将其标记为「好瓜」，测得验证集精度提升至 71.4，决定剪枝。



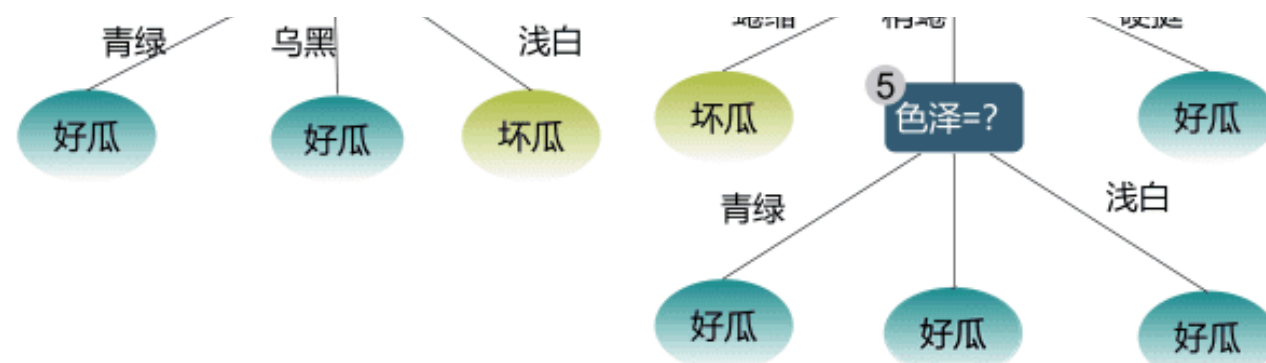
决策树算法详解



预剪枝与后剪枝

后剪枝

<http://www.showmeai.tech/>



(色泽=?) 验证集精度:

剪枝前: 57.1%

剪枝后: 71.4%

后剪枝决策: 剪枝

搜索 | 微信 ShowMeAI 研究中心

对结点 ③ 和 ①, 若将其子树替换为叶结点, 则所得决策树的验证集精度分布为 71.4 和 42.9, 均未提高, 所以不剪枝。得到最终后剪枝之后的决策树。

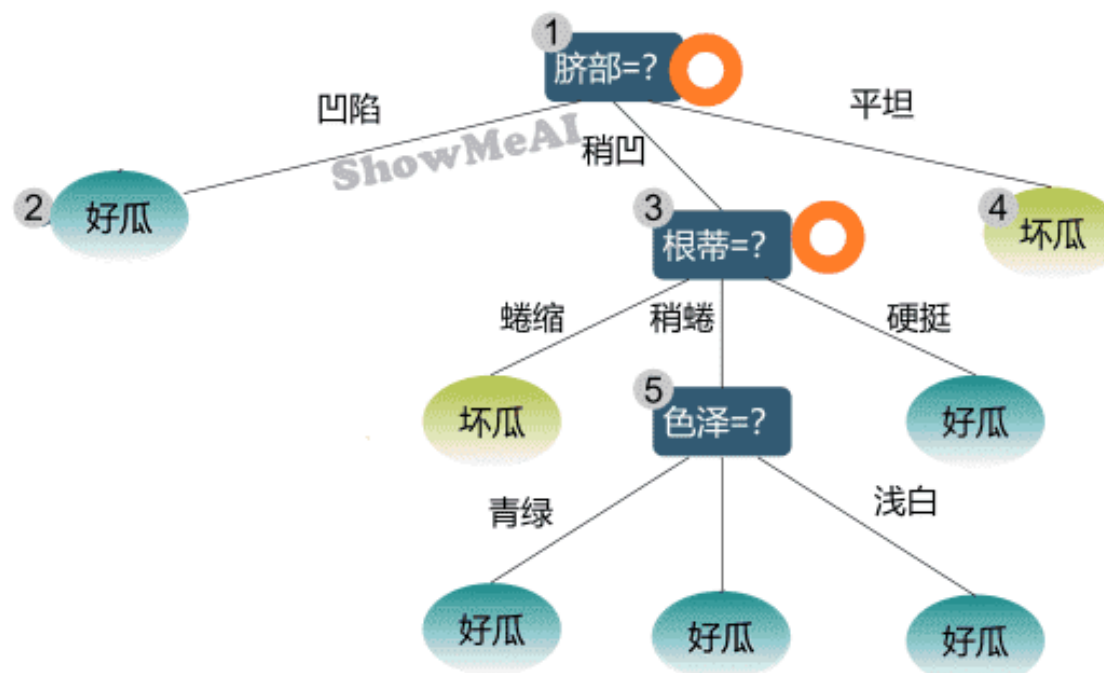


决策树算法详解

预剪枝与后剪枝

后剪枝

<http://www.showmeai.tech/>



(跟蒂=?) 验证集精度:

剪枝前: 71.4%

剪枝后: 42.9%

后剪枝决策: 不剪枝

(脐部=?) 验证集精度:

剪枝前: 71.4%

剪枝后: 71.4%

后剪枝决策: 不剪枝

搜索 | 微信 ShowMeAI 研究中心

3) 预剪枝与后剪枝的特点

时间开销：

- 预剪枝：训练时间开销降低，测试时间开销降低。
- 后剪枝：训练时间开销增加，测试时间开销降低。

过/欠拟合风险：

- 预剪枝：过拟合风险降低，欠拟合风险增加。
- 后剪枝：过拟合风险降低，欠拟合风险基本不变。

泛化性能：后剪枝通常优于预剪枝。

4.连续值与缺失值的处理

1) 连续值处理

我们用于学习的数据包含了连续值特征和离散值特征，之前的例子中使用的都是离散值属性（特征），决策树当然也能处理连续值属性，我们来看看它的处理方式。

对于**离散取值的特征**，决策树的划分方式是：选取一个最合适的特征属性，然后将集合按照这个特征属性的不同值划分为多个子集合，并且不断的重复这种操作的过程。

对于**连续值属性**，显然我们不能以这些离散值直接进行分散集合，否则每个连续值将会对应一种分类。那我们如何把连续值属性参与到决策树的建立中呢？

因为连续属性的可取值数目不再有限，因此需要连续属性离散化处理，**常用的离散化策略是二分法**，这个技术也是 C4.5 中采用的策略。

具体的二分法处理方式如下图所示：



决策树算法详解

基本思路：连续属性离散化

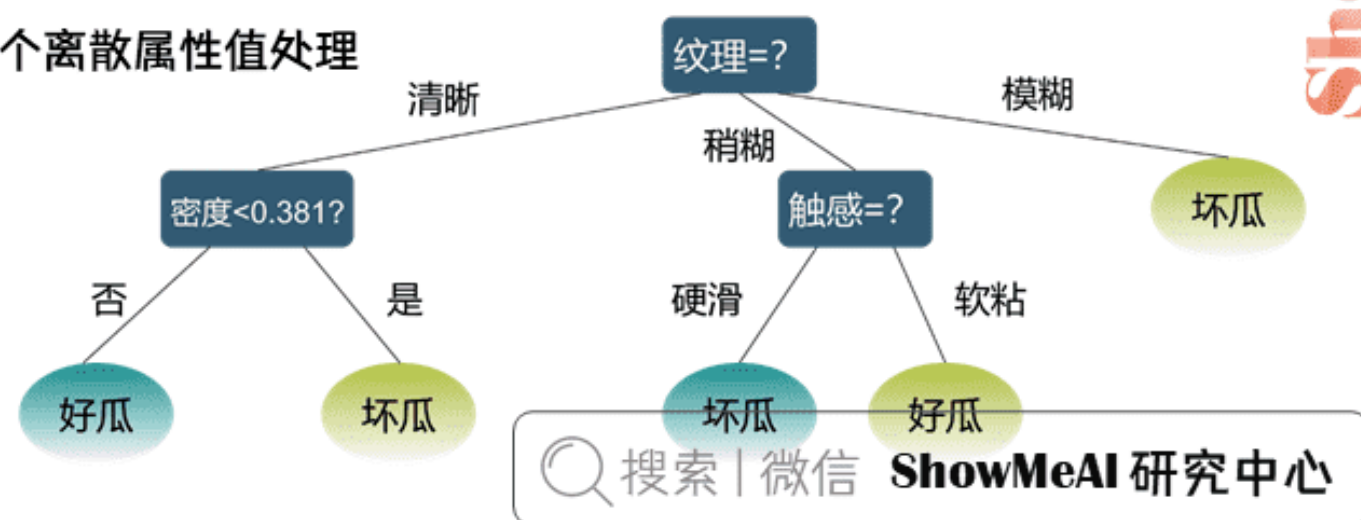
常见做法：二分法 (bi-partition)

ShowMeAI

连续值处理

<http://www.showmeai.tech/>

- n 个属性值可形成 $n-1$ 个候选划分
- 然后即可将它们当做 $n-1$ 个离散属性值处理



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

2) 缺失值处理

原始数据很多时候还会出现缺失值，决策树算法也能有效的处理含有缺失值的数据。使用决策树建模时，处理缺失值需要解决2个问题：

- Q1：如何进行划分属性选择？
- Q2：给定划分属性，若样本在该属性上的值缺失，如何进行划分？

缺失值处理的基本思路是：样本赋权，权重划分。我们来通过下图这份有缺失值的西瓜数据集，看看具体处理方式。

仅通过无缺失值的样例来判断划分属性的优劣，学习开始时，根结点包含样例集 D 中全部 17 个样例，权重均为 1。

- 根结点选择「色泽」属性时，有 3 个缺失值，因此样例总数为 14。
- 此时好瓜样例为 {2, 3, 4, 6, 7, 8}，比例为 6/14，坏瓜样例为 {9, 10, 11, 12, 14, 15, 16, 17}，比例为 8/14。

将数据带入信息熵计算公式即可得到该结点的信息熵。



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

基本思路：样本赋权，权重划分

决策树算法详解

缺失值处理

<http://www.showmeai.tech/>

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

基本思路：计算熵，计算划分的

$$Ent(\bar{D}) = - \sum_{k=1}^2 \bar{p}_k \log_2 \bar{p}_k$$
$$= - \left(\frac{6}{14} \log_2 \frac{6}{14} + \frac{8}{14} \log_2 \frac{8}{14} \right)$$
$$= 0.985$$

搜索 | 微信 ShowMeAI 研究中心

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：



决策树算法详解

缺失值处理

色泽 = 青绿

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

色泽 = 乌黑

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

色泽 = 浅白

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

因此，在样本集
上属性“色泽”
的信息熵增益为

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

无缺失值样例中属性α
取值为v的占比

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

搜索 | 微信 ShowMeAI 研究中心

- \tilde{D}^1 (色泽 = 青绿), 包含 {4, 6, 10, 17}, 4 个样例, 其中好瓜样例为 {4, 6}, 比例为 2/4, 坏瓜样例为 {10, 17}, 比例为 2/4。将数据带入信息熵计算公式即可得到该结点的信息熵。
- \tilde{D}^2 (色泽 = 乌黑), 包含 {2, 3, 7, 8, 9, 15}, 6 个样例, 其中好瓜样例为 {2, 3, 7, 8}, 比例为 4/6, 坏瓜样例为 {9, 15}, 比例为 2/6。将数据带入信息熵计算公式即可得到该结点的信息熵。
- \tilde{D}^3 (色泽 = 浅白), 包含 {11, 12, 14, 16}, 4 个样例, 其中好瓜样例为 $\{\phi\}$, 比例为 0/4, 坏瓜样例为 {11, 12, 14, 16}, 比例为 4/4。将数据带入信息熵计算公式即可得到该结点的信息熵。

于是, 样本集 D 上属性「色泽」的信息增益可以计算得出, $Gain(D, 纹理) = 0.424$ 信息增益最大, 选择「纹理」作为接下来的划分属性。

于是, 样本集 D 上属性「色泽」的信息增益可以计算得出, $Gain(D, 纹理) = 0.424$ 信息增益最大, 选择「纹理」作为接下来的划分属性。



决策树算法详解

缺失值处理

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

无缺失值样例占比

于是, 样本集 D 上属性“色泽”的信息增益为

$$Gain(D, 色泽) = 0.252$$

$$Gain(D, 根蒂) = 0.171$$

$$Gain(D, 敲声) = 0.145$$

$$Gain(D, 纹理) = 0.424$$

$$Gain(D, 脐部) = 0.289$$

$$Gain(D, 触感) = 0.006$$

信息增益最大, 选择“纹理”
作为接下来的划分属性

类似地可计算出所有属性在数据集上的信息增益

更多监督学习的算法模型总结可以查看 ShowMeAI 的文章 [AI知识技能速查 | 机器学习-监督学习](#)。

视频教程

可以点击 [B站](#) 查看视频的【双语字幕】版本

<https://www.bilibili.com/video/BV1y44y187wN?p=12>

机器学习【算法】系列教程

- [图解机器学习 | 机器学习基础知识](#)
- [图解机器学习 | 模型评估方法与准则](#)
- [图解机器学习 | KNN算法及其应用](#)
- [图解机器学习 | 逻辑回归算法详解](#)
- [图解机器学习 | 朴素贝叶斯算法详解](#)
- [图解机器学习 | 决策树模型详解](#)
- [图解机器学习 | 随机森林分类模型详解](#)
- [图解机器学习 | 回归树模型详解](#)
- [图解机器学习 | GBDT模型详解](#)
- [图解机器学习 | XGBoost模型最全解析](#)
- [图解机器学习 | LightGBM模型详解](#)
- [图解机器学习 | 支持向量机模型详解](#)
- [图解机器学习 | 聚类算法详解](#)
- [图解机器学习 | PCA降维算法详解](#)

机器学习【实战】系列教程

- [机器学习实战 | Python机器学习算法应用实践](#)
- [机器学习实战 | SKLearn入门与简单应用案例](#)
- [机器学习实战 | SKLearn最全应用指南](#)

- [机器学习实战 | SKLearn机器学习应用](#)
- [机器学习实战 | XGBoost建模应用详解](#)
- [机器学习实战 | LightGBM建模应用详解](#)
- [机器学习实战 | Python机器学习综合项目-电商销量预估](#)
- [机器学习实战 | Python机器学习综合项目-电商销量预估<进阶方案>](#)
- [机器学习实战 | 机器学习特征工程最全解读](#)
- [机器学习实战 | 自动化特征工程工具Featuretools应用](#)
- [机器学习实战 | AutoML自动化机器学习建模](#)

ShowMeAI 系列教程推荐

- [大厂技术实现：推荐与广告计算解决方案](#)
- [大厂技术实现：计算机视觉解决方案](#)
- [大厂技术实现：自然语言处理行业解决方案](#)
- [图解Python编程：从入门到精通系列教程](#)
- [图解数据分析：从入门到精通系列教程](#)
- [图解AI数学基础：从入门到精通系列教程](#)
- [图解大数据技术：从入门到精通系列教程](#)
- [图解机器学习算法：从入门到精通系列教程](#)
- [机器学习实战：手把手教你玩转机器学习系列](#)
- [深度学习教程：吴恩达专项课程·全套笔记解读](#)
- [自然语言处理教程：斯坦福CS224n课程·课程带学与全套笔记解读](#)