智能软件开发方向基础

第五章 决策树
Decision Tree

第1部分 理解认识决策树

张朝晖

2022~2023学年第二学期

河北师范大学软件学院
Software College of Hebei Normal University

序号	内容
1	概述
2	机器学习的基本概念
3	模型的选择与性能评价
4	数据的获取、探索与准备
5	近邻模型------分类、回归
6	决策树模型------分类、回归
7	集成学习------分类、回归
8	(朴素)贝叶斯模型------分类
9	聚类
10	特征降维及低维可视化(PCA, t-SNE)
11	总复习

本课件主要内容及有关例子，主要参考了

1. 周志华，《机器学习》
2. 李　航，《统计学习方法》

特此感谢！

河北师范大学软件学院
Software College of Hebei Normal University

思考题

1. 什么是决策树？
   决策树模型的叶子节点与特征空间、训练样本集存在什么对应关系？

2. 如何利用到达决策树某节点处的训练集度量该节点的不纯度？
   (三种典型的节点不纯度度量方式)

3. ID3,C4.5,CART三种典型决策树的算法实现步骤？

4. 三种决策树模型中，非叶子节点所用的特征是采用何种规则进行选择的？给出具体的选择方式.
   以根节点处特征选择为例，描述原理。

5. 哪种决策树模型还可用于实值函数回归？若用于回归，如何生成预测结果？

6. 给定一棵初步构建的决策树，如何对其进行剪枝？

河北师范大学软件学院
Software College of Hebei Normal University

主要内容

决策树

基于树形结构的决策模型—决策树

包括：决策树构建方法；决策树的剪枝；决策树的使用

1 非度量特征(*nonmetric features*)

2 初步认识决策树

3.决策树的构建

4.过学习与决策树的剪枝

样本的特征描述

(1)**度量型特征** (***metric features***)

(2)**非度量型特征**(***nonmetric features***)

如： 名义特征/标称数据(***nominal features***)

序数特征(***ordinal features***)

区间特征(***interval features***)

非度量型特征描述的样本分类，处理方式：

方式1

非度量型特征 —编码→ 度量型特征 —基于度量型特征的样本分类→ 分类结果

编码可能会
- 造成信息损失
- 引入人为信息

方式2

基于非度量型特征的样本直接分类

决策树可直接面向非度量型、度量型特征描述的样本.

主要内容

决策树

基于树形结构的决策模型—决策树

包括：决策树构建方法；决策树的剪枝；决策树的使用

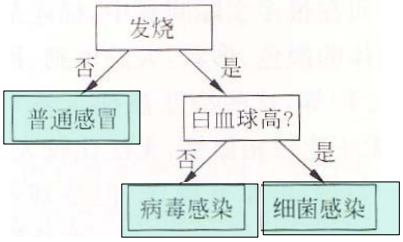
1 非度量特征(*nonmetric features*)

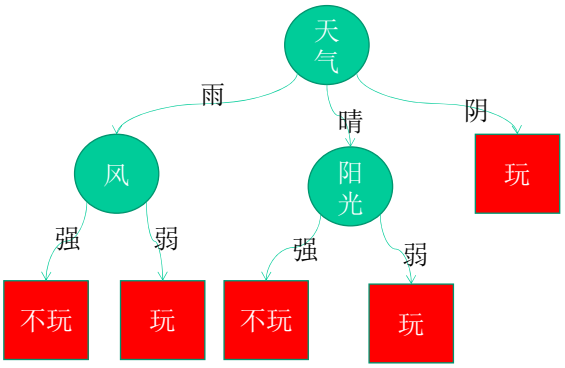
2 初步认识决策树

3.决策树的构建

4.过学习与决策树的剪枝

(1)什么是决策树？



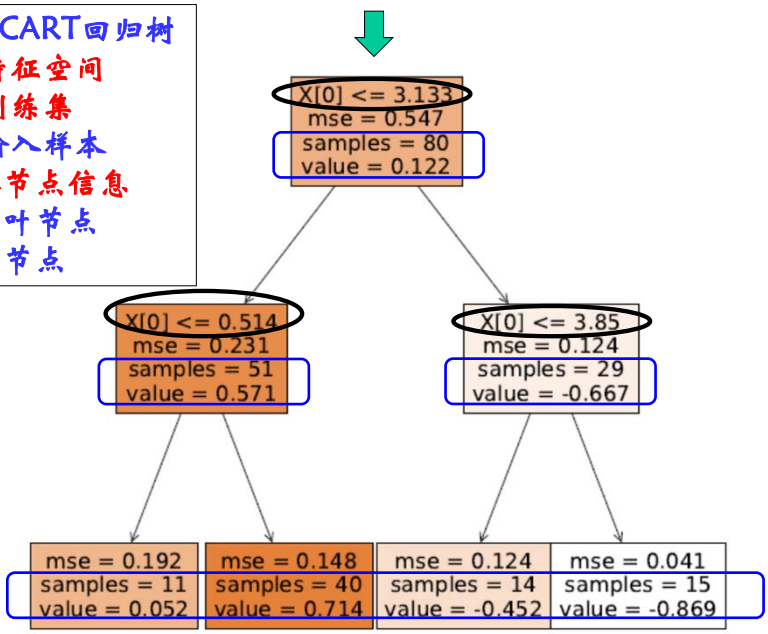


决策树是关于if-then规则的集合

规则互斥、完备

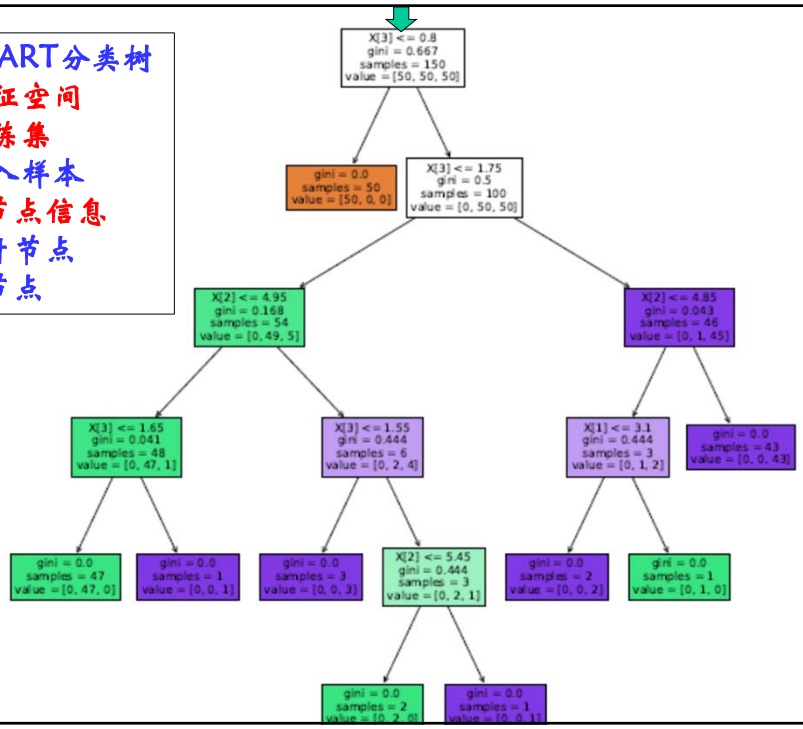
例：CART回归树

- 特征空间
- 训练集
- 输入样本
- 各节点信息
- 非叶节点
- 叶节点



例：CART分类树

- 特征空间
- 训练集
- 输入样本
- 各节点信息
- 非叶节点
- 叶节点



决策树是一种以倒立树形结构描述的决策规则集合。

由一个根节点、若干内部节点、若干叶节点组成。

每个非叶节点代表关于输入样本的一个特征测试(查询)，该节点的每个分枝表示测试的一个结果；每个叶节点代表关于输入样本的一个决策结果。

- 若为分类树，则决策结果为预测类别(或关于所有类别的预测概率)；
- 若为回归树，则决策结果为实数值。

从根节点通向叶节点的一条路径对应一条决策规则。

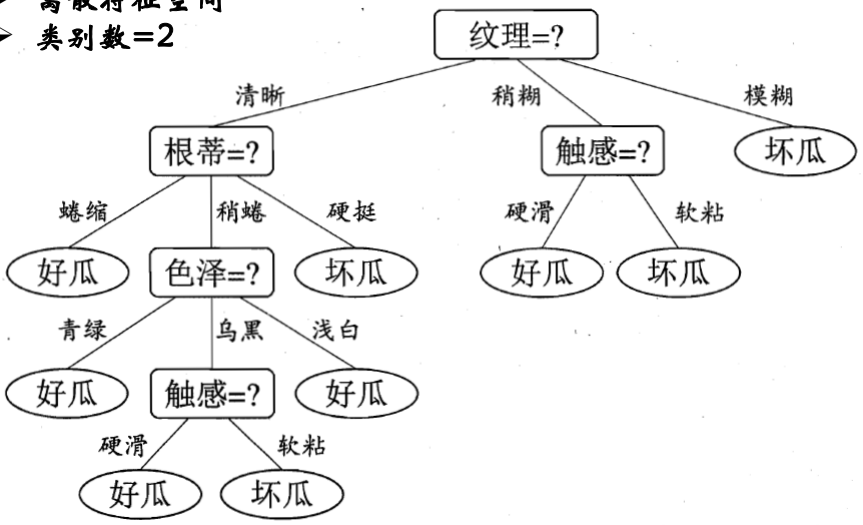
决策树是应用最广的归纳推理方法之一，模型直观。

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

西瓜数据集2.0

基于绝对信息增益的决策树生成--ID3分类树

- 离散特征空间
- 类别数=2

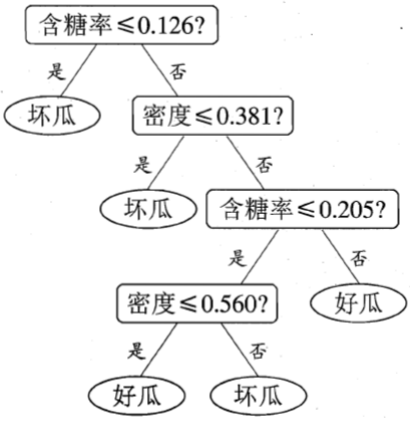


西瓜数据集3.0a

编号	密度	含糖率	好瓜
1	0.697	0.460	是
2	0.774	0.376	是
3	0.634	0.264	是
4	0.608	0.318	是
5	0.556	0.215	是
6	0.403	0.237	是
7	0.481	0.149	是
8	0.437	0.211	是
9	0.666	0.091	否
10	0.243	0.267	否
11	0.245	0.057	否
12	0.343	0.099	否
13	0.639	0.161	否
14	0.657	0.198	否
15	0.360	0.370	否
16	0.593	0.042	否
17	0.719	0.103	否

CART分类树

- 连续特征空间
- 类别数=2
- 二叉树



(2)决策树的优势

- 语义可表示性

----从根节点到叶节点的一条决策规则为合取式
----利用合取式和析取式获得某个类别的明确描述

- 决策速度快

只需一系列关于待决策样本的简单查询，即可对样本的输出做出判断

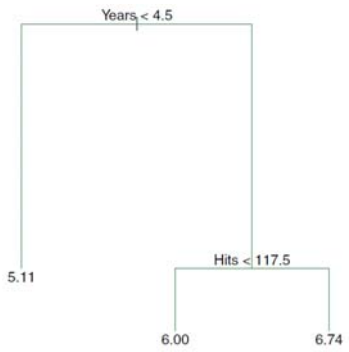
- 可以很自然的嵌入专家的先验知识

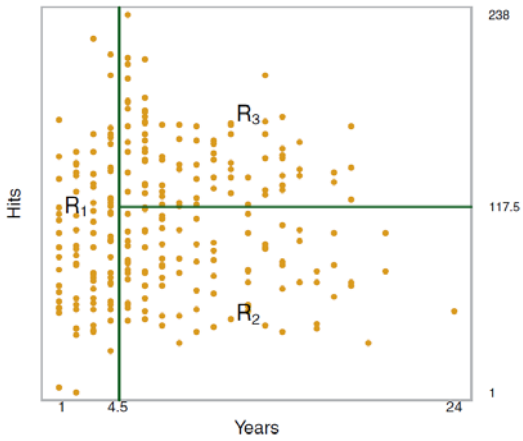
(3)决策树的叶节点与特征空间的划分及相应决策结果(决策域)

例：基于回归树，预测棒球运动员的薪金.

两种特征：

- Years --棒球运动员在大联盟中的效力时间(时间)
- Hits --棒球运动员在上一年度的击球打数(成绩)





(4)决策树模型的学习与使用

- 模型的监督式学习—决策树的构建(与剪枝)

归纳：决策规则的生成。

基于一定数量训练样本，学习决策规则，自动构造。

训练样本集的划分 ➔ 特征空间的最终划分

- 模型的使用—利用生成的规则，对观测样本进行决策推理

A. 模型的学习

- 决策树构建中的节点特征选择
  利用到达当前节点的训练样本集，从中选择最优划分特征

- 决策树的生成(模型的局部选择)
  递归生成决策树，拟合训练样本

- 决策树的剪枝(模型的全局选择)

  简化模型，使其泛化能力更好
  许多分枝反映的是训练样本中的噪声和孤立点

  为避免过学习，应控制树的规模，检测和剪枝
  预剪枝(prepruning)、后剪枝(postpruning)

B. 模型的使用

从根节点开始，对输入样本的特征取值提问

与根节点相连的不同分枝，对应于特征的不同取值

根据不同回答，转向相应的分枝

在新到达的节点处，做类似的分枝判断…

持续上述过程直到叶子节点，输出该叶子节点对应的类别标记(或函数值)。