Lecture 5 数据分析算法（III）
数据决策

教学目标

- 认识常用的数据决策方法的原理，并了解不同方法之间的优缺点
- 掌握ID3、C4.5、CART等算法的原理，掌握信息熵、信息增益、信息增益率和基尼指数的概念和计算

电子科技大学
University of Electronic Science and Technology of China

内容概述

数据分析算法

- 数据关系：TF-IDF，余弦相似, Apriori, PageRank
- 分类与聚类：Bayes, AdaBoost, SVM, KNN, K-Means, EM
- 决策：ID3、C4.5、CART

第7讲 数据决策

- 决策树是一个预测模型，代表对象属性与对象值之间的一种映射关系。
- 决策树经常用于数据挖掘中的数据分析和预测。
- 决策树是一种特殊的树结构，由决策图和可能的结果组成，用来创建到达目标的规划。

生活中的例子

来一段生活情景对话（相亲决策树）：

母亲：女儿，你也不小了，还没对象！妈很揪心啊，这不托人给你找了个对象，明儿去见个面吧！

女儿：年纪多大了？

母亲：25

女儿：长的帅不帅？

母亲：挺帅的！

女儿：收入高不高？有没有上进心？

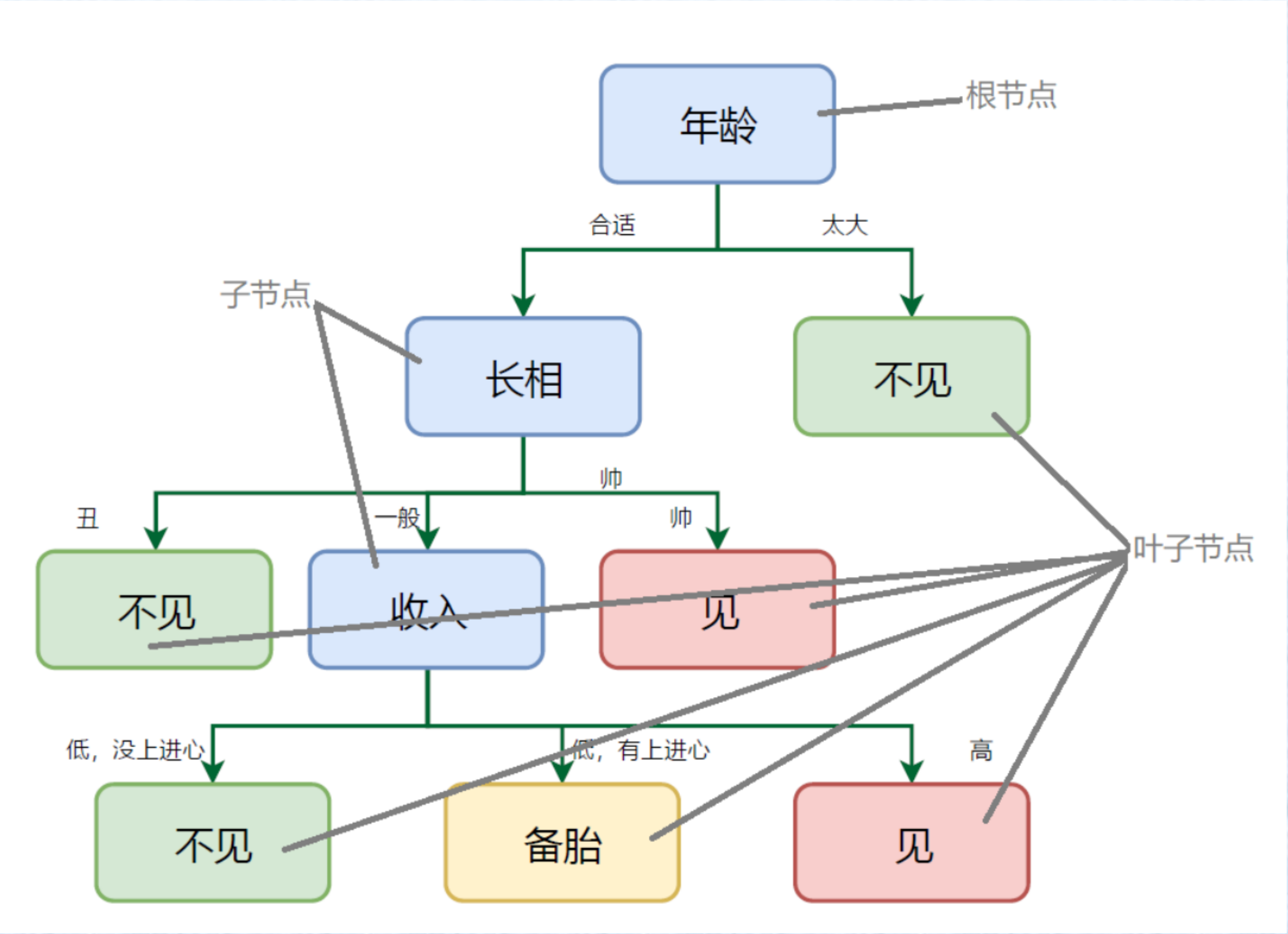
母亲：收入还行，蛮有上进心！

。

。

就这样女儿建立了一棵决策树

相亲决策树



决策树构建过程

1、收集样本

没有要决策的样本，一切都是扯淡。就是例子中母亲托人找对象的过程。

2、选择特征-----构建节点

根据特征的重要度，来构建子节点，越重要的特征越靠近根节点。也就是女儿觉得那些条件最重要，当最重要的条件不满足，就没必要继续了。

3、特征的分裂方式-----分裂节点

根据特征的分裂变量，来划分数据集，也就是根据条件区别对待。就是年纪太大的压根就不予考虑，年龄合适的才进一步考察。

其实在实际构建树模型的时候，2和3是通过遍历的方式同时进行的。

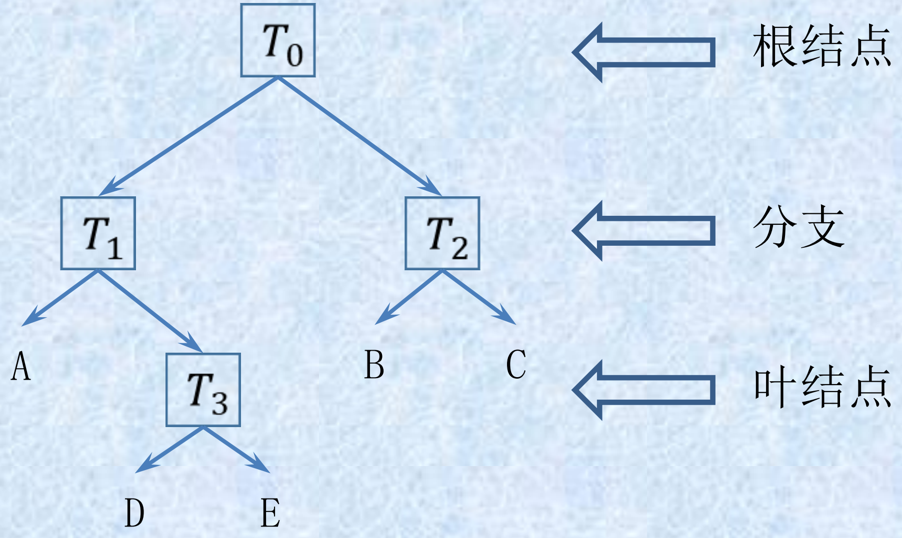
问题的提出

- 根据气候做出是否打篮球的决策，数据决策就是试图从数据中挖掘特征与结果之间的关系。

天气	温度	湿度	刮风	是否打篮球
晴天	高	低	否	打篮球
晴天	高	低	是	打篮球
晴天	中	低	否	打篮球
晴天	中	中	否	打篮球
晴天	中	低	是	打篮球
小雨	低	高	否	不打篮球
小雨	低	高	是	不打篮球
小雨	中	高	否	不打篮球
阴天	低	高	否	不打篮球
阴天	中	高	否	打篮球
阴天	中	中	是	打篮球

决策树构成

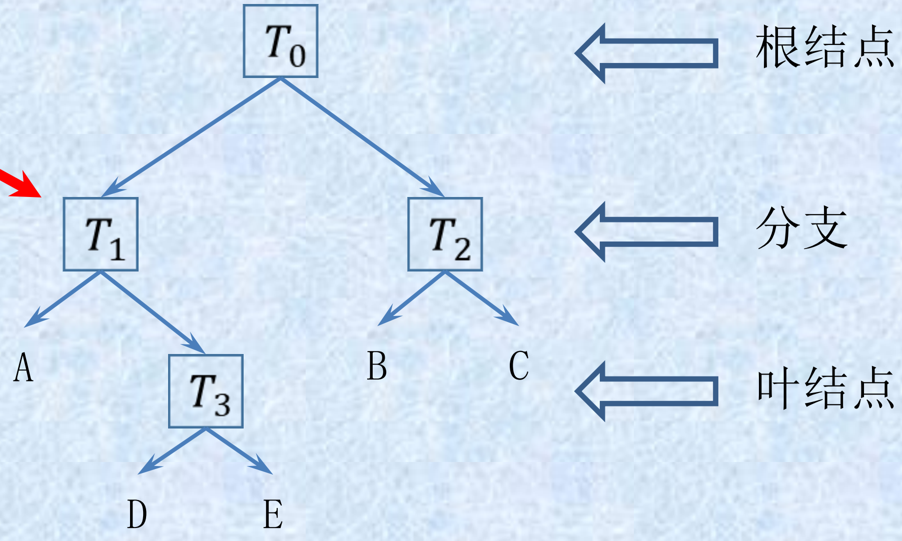
- 一棵决策树通常由结点和有向边组成，结点包括**根结点、内部结点和叶节点**
  - 根结点和内部节点表示一个特征或者属性
  - 叶节点表示一个具体分类。



ID3算法

- ID3算法是以信息论为基础，以信息熵和信息增益为衡量标准，从而实现对数据的归纳分类
  – 根据信息增益运用自顶向下的贪心策略是ID3建立决策树的主要方法
- 决策树关键问题：树分支的裂变依据，属性选择

依据什么裂变？



电子科技大学
University of Electronic Science and Technology of China

熵的概念

熵(entropy)是表征一个系统的混乱程度和不确定性的量。该系统越混乱越不确定，它的熵越大；系统越整齐越稳定，它的熵越小。

$$\mathbf{Entropy}(x) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

看下面两个集合，相比于 B 集合，A 集合就比较混乱、不确定性比较大。 B 集合中大部分都是 ①，比较整齐，确定性较强，所以 B 集合的熵就比较小， A 集合混乱则熵比较大。

A = { ①②⑤⑥⑦①②④ }

B = { ①①①①⑦①①④ }

$$E(A) = -\left(2*\frac{2}{8}\log_2\frac{2}{8} + 4*\frac{1}{8}\log_2\frac{1}{8} \right) = 6.82$$

$$E(B) = -\left(\frac{6}{8}\log_2\frac{6}{8} + 2*\frac{1}{8}\log_2\frac{1}{8} \right) = 1.06$$

算法概念

- 信息熵是接收信息量的平均值，用于度量信息的不确定程度，是随机变量的均值。
  - 信息熵的处理信息是一个让信息的熵减少的过程。
  - 假设X是一个离散的随机变量，且它的取值有限范围$R = \{x_1, x_2, \cdots, x_n\}$，设$p_i = P\{X = x_i\}$，则$X$的熵

$$\mathbf{Entropy}(\boldsymbol{x}) = -\sum_{i=1}^{n} \boldsymbol{p}_i \log_2 \boldsymbol{p}_i$$

- 信息增益用于度量属性A对降低样本集合X的熵的贡献大小，也就是度量A对使信息有序的贡献。

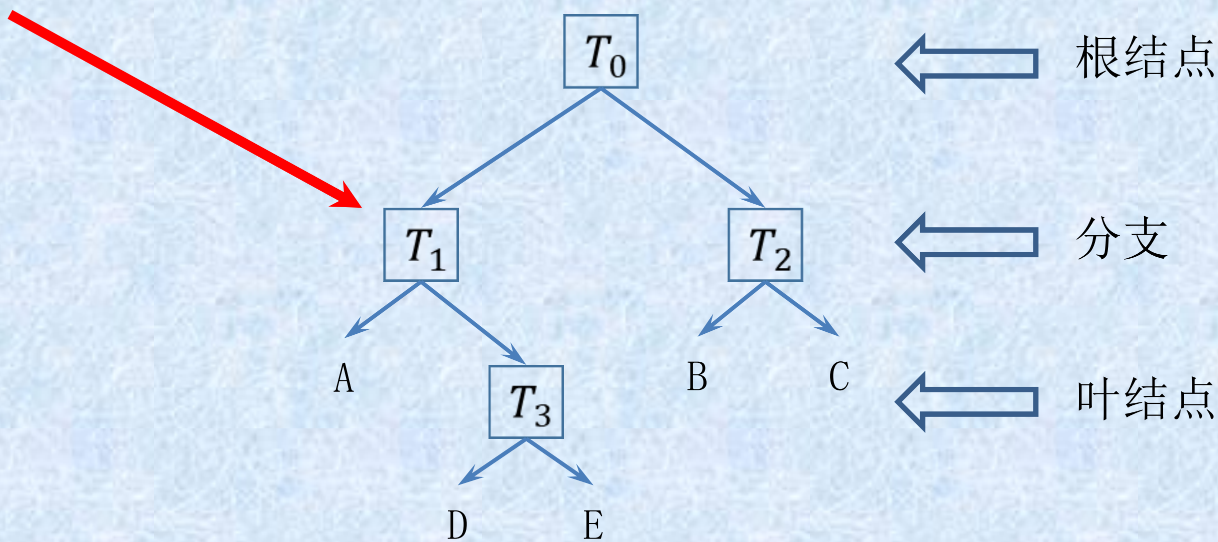
$$\mathbf{Gain}(\boldsymbol{A}, \boldsymbol{X}) = \mathbf{Entropy}(\boldsymbol{X}) - \sum_{X_v} \left(\frac{|\boldsymbol{X}_v|}{|\boldsymbol{X}|} \times \mathbf{Entropy}(\boldsymbol{X}_v) \right)$$

电子科技大学
University of Electronic Science and Technology of China

ID3算法流程

1. 对当前样本集合计算所有属性的信息增益。
2. 选择信息增益最大的属性裂变。
3. 若子样本集的类别属性只含有单个属性，则分支为叶子节点，判断其属性值并标上相应的符号，然后返回调用处；否则对子样本集递归调用本算法。

依据什么裂变？



电子科技大学 University of Electronic Science and Technology of China

算法实例

- 通过不同的因素判定学科是否可能通过。

考试成绩	作业完成情况	出勤率	是否能够通过总评
优	优	高	是
优	良	高	是
良	优	高	是
良	良	高	是
及格	良	高	是
及格	及格	高	是
及格	不及格	低	否
及格	不及格	高	是
不及格	及格	低	否
不及格	不及格	低	否

- 样本集合的信息熵：$-\frac{7}{10}\log_2\frac{7}{10}-\frac{3}{10}\log_2\frac{3}{10}=0.881$

考试成绩	作业完成情况	出勤率	是否能够通过总评
优	优	高	是
优	良	高	是
良	优	高	是
良	良	高	是
及格	良	高	是
及格	及格	高	是
及格	不及格	低	否
及格	不及格	高	是
不及格	及格	低	否
不及格	不及格	低	否

- 样本集合的信息熵：$-\frac{7}{10}\log_2\frac{7}{10}-\frac{3}{10}\log_2\frac{3}{10}=0.881$
- 考试成绩为及格时的信息熵：$-\frac{3}{4}\log_2\frac{3}{4}-\frac{1}{4}\log_2\frac{1}{4}=0.811$
- 考试成绩为优、良、不及格时的信息熵：0
- 属性考试成绩的信息增益

$$\mathrm{Gain}(\text{考试成绩})=0.881-\left(\frac{2}{10}\times 0+\frac{2}{10}\times 0+\frac{2}{10}\times 0+\frac{4}{10}\times 0.811\right)=0.5388$$

考试成绩	作业完成情况	出勤率	是否能够通过总评
优	优	高	是
优	良	高	是
良	优	高	是
良	良	高	是
及格	良	高	是
及格	及格	高	是
及格	不及格	低	否
及格	不及格	高	是
不及格	及格	低	否
不及格	不及格	低	否

- Gain(考试成绩) = 0.5388
- Gain(作业完成情况) = 0.4056
- Gain(出勤率) = 0.881

- 根据出勤率把样本分为两个子集（高、低），递归形成决策树

算法实例

出勤率=高 子集

考试成绩	作业完成情况	是否能够通过总评
优	优	是
优	良	是
良	优	是
良	良	是
及格	良	是
及格	及格	是
及格	不及格	是

出勤率=低 子集

考试成绩	作业完成情况	是否能够通过总评
及格	不及格	否
不及格	及格	否
不及格	不及格	否

注：这里不需要再递归，因为出勤率已经可以确定总评通过与否，满足迭代停止条件。

决策树算法pseudo-code

决策树分类算法

输入:

训练集: $D=(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)$

属性集: $A=a_1,a_2,\cdots,a_n$

过程:

函数$TreeGenerate(D,A)$

输出:

以node为根节点的决策树

算法:

生成结点根node
if D中样本全属于同一类别C_k **then**
  将node标记为C_k类叶结点
  return
end if
if $A=\varnothing$ **OR** D中样本在A上取值相同 **then**
  将node标记为叶结点，其类别标记为D中样本数最多的类
  return
end if
从A中选择最优划分属性a_*
for a_* 的每一个值a_*^v **do**
  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
  if D_v为空 **then**
    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
  else
    以$TreeGenerate(D_v,A-\{a_*\})$为分支结点
  end if
end for

ID3算例

吃瓜群众最爱---用决策树算法选瓜（训练集如下）

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

ID3算例

各个属性的值域：

色泽 = {青绿，乌黑，浅白}

根蒂 = {卷缩，稍卷，硬挺}

敲声 = {浊响，沉闷，清脆}

纹理 = {清晰，稍糊，模糊}

脐部 = {凹陷，稍凹，平坦}

触感 = {硬滑，软粘}

ID3算例

训练集的正例(好瓜)占 8/17，反例占 9/17 ，训练集的信息熵为

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

依次计算当前属性集合 {色泽，根蒂，敲声，纹理，脐部，触感} 中每个属性的信息增益。

色泽属性有3个可能的取值：{青绿，乌黑，浅白}

D^1 (色泽=青绿) = {1, 4, 6, 10, 13, 17}，正例 3/6，反例 3/6

D^2 (色泽=乌黑) = {2, 3, 7, 8, 9, 15}，正例 4/6，反例 2/6

D^3 (色泽=浅白) = {5, 11, 12, 14, 16}，正例 1/5，反例 4/5

ID3算例

3 个分支结点的信息熵

$$\text{Ent}(D^1) = -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right) = 1.000$$

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

$$\text{Ent}(D^3) = -\left(\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right) = 0.722$$

计算属性色泽的信息增益：

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

ID3算例

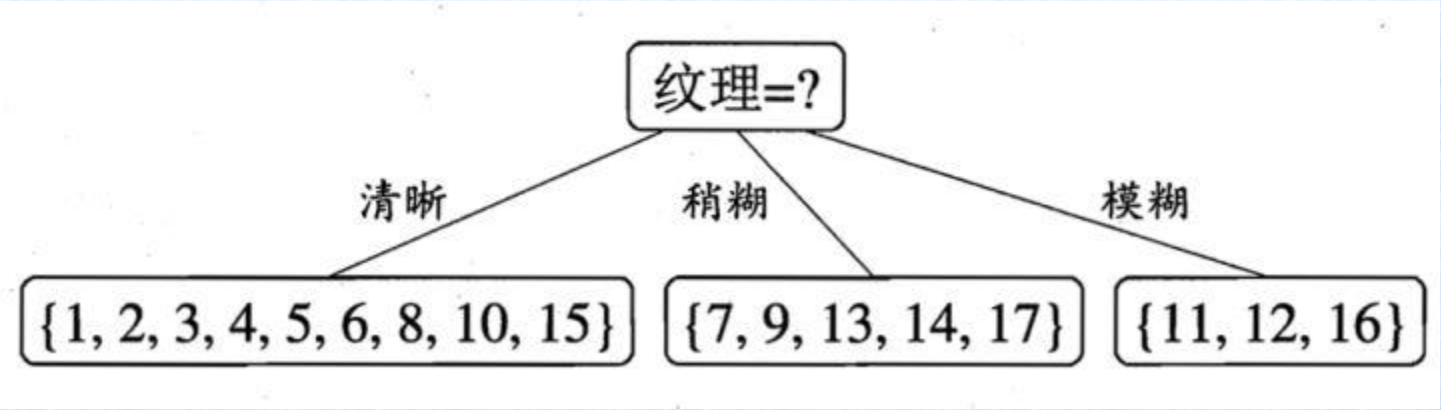
同理可以求出其它属性的信息增益：

$$\mathrm{Gain}(D, \text{根蒂}) = 0.143; \quad \mathrm{Gain}(D, \text{敲声}) = 0.141;$$

$$\mathrm{Gain}(D, \text{纹理}) = 0.381; \quad \mathrm{Gain}(D, \text{脐部}) = 0.289;$$

$$\mathrm{Gain}(D, \text{触感}) = 0.006.$$

于是我们找到信息增益最大的属性纹理，Gain(D, 纹理) = 0.381



ID3算例

对于这3个子节点，我们可以递归方法寻找信息增益最大的特征属性。

如：D^1 (纹理=清晰) = {1, 2, 3, 4, 5, 6, 8, 10, 15}，第一个分支结点可用属性集合{色泽、根蒂、敲声、脐部、触感}，基于 D^1各属性的信息增益求得如下：

$$\mathrm{Gain}(D^1,\text{色泽}) = 0.043;\quad \mathrm{Gain}(D^1,\text{根蒂}) = 0.458;$$

$$\mathrm{Gain}(D^1,\text{敲声}) = 0.331;\quad \mathrm{Gain}(D^1,\text{脐部}) = 0.458;$$

$$\mathrm{Gain}(D^1,\text{触感}) = 0.458.$$

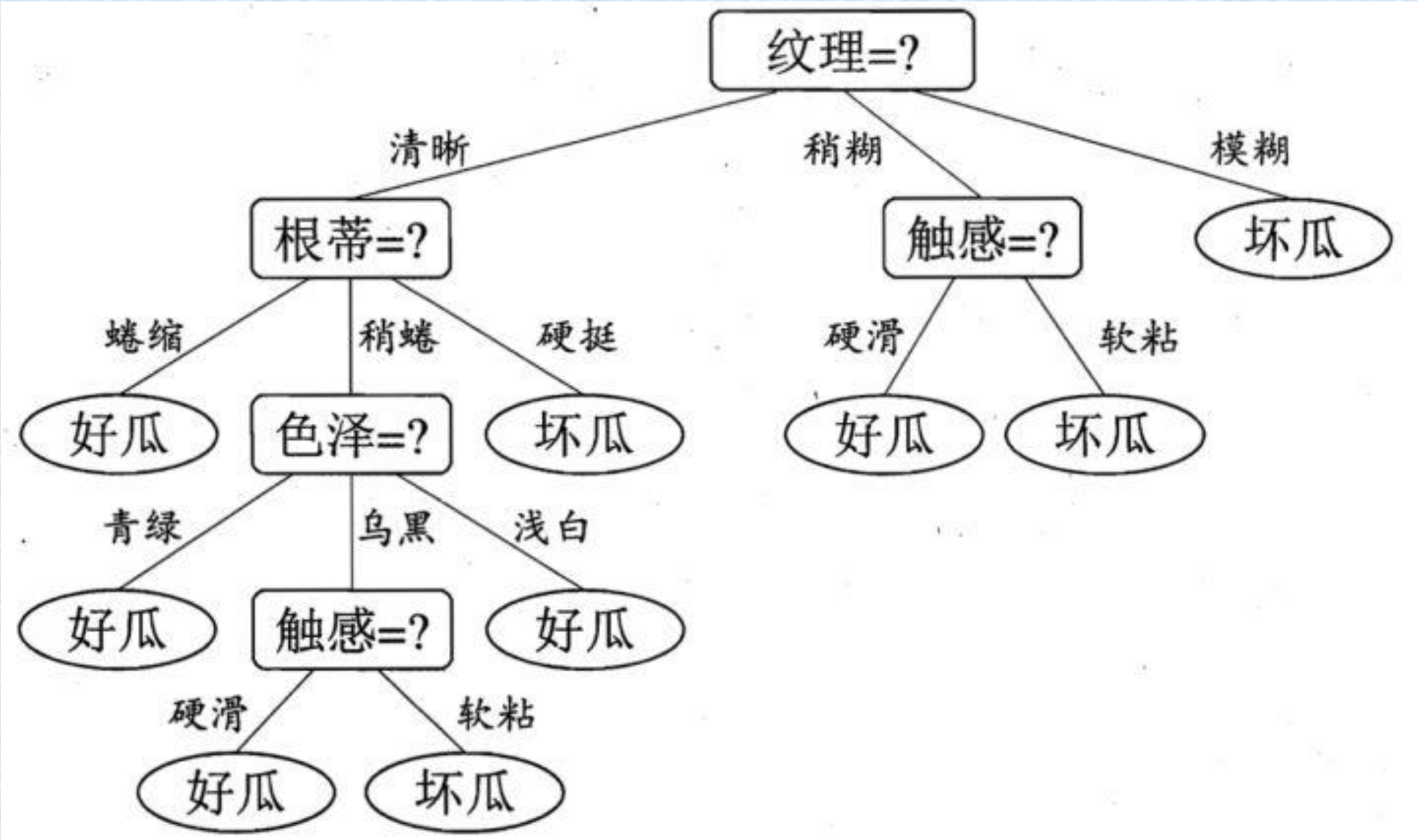
于是我们可以选择特征属性为根蒂，脐部，触感3个中任选一个（因为他们3个相等并最大），选择的特征属性为根蒂。

D^1 (纹理=稍糊) = {7, 9, 13, 14, 17}，选择的特征属性为触感。

其它子结点同理，然后得到新一层的结点。

用递归方法以此类推，可以得到各层节点，完成决策树的构建。

ID3算例



纹理清晰、根蒂稍蜷、色泽乌黑、触感硬滑 ---- 好瓜坏瓜？

算法工程师自信地告诉你：好瓜！^_^

7.2 C4.5算法

- C4.5是J. Ross Quinlan基于ID3算法改进。
  - 用信息增益率来选择属性，克服了ID3算法选择属性时偏向选择取值多的属性的不足。
  - 在决策树构造过程中支持剪枝。
  - 能够完成对连续属性的离散化处理。
  - 能够对缺失值数据进行处理。
    - 丢弃、赋予常见值
    - 概率分配：不缺失的部分中为1占60%，为0占40%；则在此属性裂变时，把缺失部分的60%分配给属性为1的分支，40%分配给属性为0的分支

电子科技大学 University of Electronic Science and Technology of China

C4.5算法流程

1. 计算出样本集合X的信息熵。
2. 计算每个属性的信息增益值$\mathrm{Gain}(V)$。
3. 计算分裂信息度量$H\ (V) = -\sum_{x_v}\left(\frac{|X_v|}{|X|}\log_2\frac{|X_v|}{|X|}\right)$。
4. 利用公式$\mathrm{IGR}(V) = \frac{\mathrm{Gain}(V)}{H(V)}$计算信息增益率。
5. 选择信息增益率最高的属性作为决策树结点进行分裂。
6. 在各个结点的子集上通过步骤2-6递归，直至满足停止条件。

信息增益率本质： 是在信息增益基础之上乘一个惩罚参数。特征个数较多时，惩罚参数较小；特征个数较少时，惩罚参数较大。

考试成绩	作业完成情况	出勤率	是否能够通过总评
优	优	高	是
优	良	高	是
良	优	高	是
良	良	高	是
及格	良	高	是
及格	及格	高	是
及格	不及格	低	否
及格	不及格	高	是
不及格	及格	低	否
不及格	不及格	低	否

1、样本集合的信息熵：$-\frac{7}{10}\log_2\frac{7}{10}-\frac{3}{10}\log_2\frac{3}{10}=0.881$

2、计算每个属性的信息增益：

$$\text{Gain}(考试成绩)=0.5388$$

$$\text{Gain}(作业完成情况)=0.4056$$

$$\text{Gain}(出勤率)=0.881$$

考试成绩	作业完成情况	出勤率	是否能够通过总评
优	优	高	是
优	良	高	是
良	优	高	是
良	良	高	是
及格	良	高	是
及格	及格	高	是
及格	不及格	低	否
及格	不及格	高	是
不及格	及格	低	否
不及格	不及格	低	否

3、计算每个属性的$H\ (V) = -\sum_{x_v}\left(\frac{|X_v|}{|X|}\log_2\frac{|X_v|}{|X|}\right)$

属性考试成绩有4个取值，其中优有2个样本，良有2个样本，及格有4个样本，不及格有2个样本，则:

$$H(考试成绩) = -\frac{2}{10}\log_2\frac{2}{10} - \frac{2}{10}\log_2\frac{2}{10} - \frac{4}{10}\log_2\frac{4}{10} - \frac{2}{10}\log_2\frac{2}{10} = 1.9219$$

$$H(作业完成情况) = 1.9709$$

$$H(出勤率) = 0.881$$

考试成绩	作业完成情况	出勤率	是否能够通过总评
优	优	高	是
优	良	高	是
良	优	高	是
良	良	高	是
及格	良	高	是
及格	及格	高	是
及格	不及格	低	否
及格	不及格	高	是
不及格	及格	低	否
不及格	不及格	低	否

4、计算每个属性的信息增益率$IGR(V) = \frac{\mathrm{Gain}(V)}{H(V)}$

$\mathrm{IGR}(\text{考试成绩}) = \mathrm{Gain}(\text{考试成绩}) \div H(\text{考试成绩}) = 0.5388 \div 1.9219 = 0.2803$

$\mathrm{IGR}(\text{作业完成情况}) = 0.2058$

$\mathrm{IGR}(\text{出勤率}) = 1$

7.3 CART算法

- 决策树主要有两种类型：分类树和回归树。
  - 分类树的输出是样本的类标，回归树的输出是一个实数
- 分类和回归树，即CART（classification and regression tree），最先由Breiman等提出。
  - 决策树生成：基于训练数据集生成决策树，生成的决策树尽量大
  - 决策树剪枝：用验证数据集对已生成的树进行剪枝并选择最优子树，这时用损失函数最小作为剪枝的标准
- 本节只关注分类树

与ID3的区别

- 用于选择变量的度量不同。ID3使用的度量是信息增益，CART使用的不纯度量是GINI指数
- 对于连续的目标变量，在CART算法中，预测目标变量的方法是找出一组基于树的回归方程
- 对于具有两个以上类别的多类问题，CART算法可能考虑将目标类别合并成两个超类别（双化）
- CART算法的决策树是个二叉树（运算速度较多叉树快得多），属性可重复出现（运用剪枝方法）

GINI指数

- 有M个类，样本属于第i类的概率为p_i，则概率分布的GINI指数定义为$\mathrm{GINI}(p) = \sum_{i=1}^{M} p_i \times (1 - p_i)$
- 对于给定的样本集合D，其GINI指数为

$$\mathrm{GINI}(D) = 1 - \sum_{i=1}^{M} \left(\frac{|C_i|}{|D|}\right)^2$$

  – C_i是D中属于第i类的样本子集。

- 样本集D被特征$A = \alpha$划分成D_1和D_2两部分，则在特征值A的条件下，集合D的Gini指数定义为

$$\mathrm{GINI}(D, A) = \frac{|D_1|}{|D|}\mathrm{GINI}(D_1) + \frac{|D_2|}{|D|}\mathrm{GINI}(D_2)$$

GINI指数

- 基尼指数$\mathrm{GINI}(D)$表示集合D的不确定性，基尼指数$\mathrm{GINI}(D,A)$表示经$A=\alpha$分割后集合D的不确定性。
  - 基尼指数值越大，样本集合的不确定性（不纯度）也越大

Gini指数越小表示集合中被选中的样本被分错的概率越小，也就是说集合的纯度越高，反之，集合越不纯。

示例

- 银行客户贷款申请

ID	A_1年龄	A_2有工作	A_3有自己的房子	A_4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

CART算法流程

- 计算各特征的基尼指数，选择最优特征以及其最优切分点。
- 以计算$\mathrm{GINI}(D, A_1 = \text{青年})$为例

$A_1 = \text{青年}$，$D_1 = \{\text{青年}\}, D_2 = \{\text{中年,老年}\}$

$|D_1| = 5, |D_2| = 10, |D| = 15$。

在集合D_1中，放贷类别$C_1 = \{\text{是}\}$的数量是2，$C_2 = \{\text{否}\}$的数量是3，则$\mathrm{GINI}(D_1) = 1 - \left(\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right) = 0.48$。同理，$\mathrm{GINI}(D_2) = 1 - \left(\left(\frac{7}{10}\right)^2 + \left(\frac{3}{10}\right)^2\right) = 0.42$。由此可计算出

$$\mathrm{GINI}(D, A_1 = \text{青年}) = 5/15 \times 0.48 + 10/15 \times 0.42 = 0.44$$

- $\mathrm{GINI}(D, A_1 = 青年) = 0.44$，最优切分点
- $\mathrm{GINI}(D, A_1 = 中年) = 0.48$
- $\mathrm{GINI}(D, A_1 = 老年) = 0.44$，最优切分点
- $\mathrm{GINI}(D, A_2 = 是) = 0.32$，最优切分点
- $\boxed{\mathrm{GINI}(D, A_3 = 是) = 0.27\text{，最优切分点}}$ 最优特征
- $\mathrm{GINI}(D, A_4 = 非常好) = 0.36$
- $\mathrm{GINI}(D, A_4 = 好) = 0.47$
- $\mathrm{GINI}(D, A_4 = 一般) = 0.32$，最优切分点

电子科技大学
University of Electronic Science and Technology of China

裂变

第二层 ：A3 = “是”时数据如下

序号	年龄A1	是否工作A2	信贷情况A4	贷款审批结果
1	老年	否	很好	是
2	老年	否	很好	是
8	中年	是	好	是
9	中年	否	很好	是
10	中年	否	很好	是
14	青年	是	一般	是

第二层 ：A3 = “否”时数据如下

序号	年龄A1	是否工作A2	信贷情况A4	贷款审批结果
3	老年	是	好	是
4	老年	是	好	是
5	老年	否	一般	否
6	中年	否	一般	否
7	中年	否	好	否
11	青年	否	一般	否
12	青年	否	好	否
13	青年	是	好	是
15	青年	否	一般	否

https://blog.csdn.net/weixin_43216017

裂变

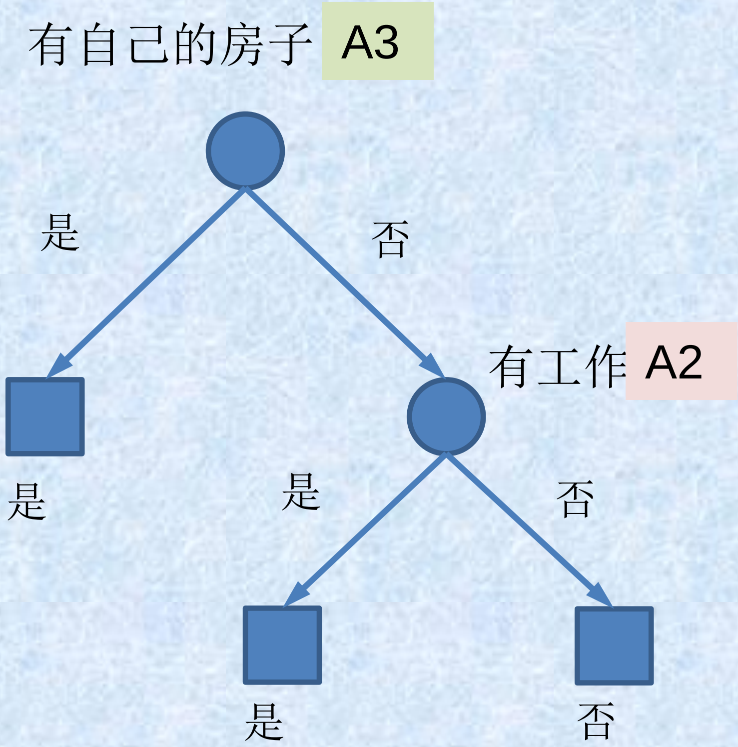
划分	ID	A_1年龄	A_2有工作	A_3有自己的房子	A_4信贷情况	类别
D_1	4	青年	是	是	一般	是
	8	中年	是	是	好	是
	9	中年	否	是	非常好	是
	10	中年	否	是	非常好	是
	11	老年	否	是	非常好	是
	12	老年	否	是	好	是
D_2	1	青年	否	否	一般	否
	2	青年	否	否	好	否
	3	青年	是	否	好	是
	5	青年	否	否	一般	否
	6	中年	否	否	一般	否
	7	中年	否	否	好	否
	13	老年	是	否	好	是
	14	老年	是	否	非常好	是
	15	老年	否	否	一般	否

裂变

此时的总体信息熵：

$$info(DA3) = 0.9183$$

$$Gain(A1) = 0.2516 \;;$$
$$Gain(A2) = 0.9183 \;;$$



至此不需要在计算Gain(A4)了，
因为A2已经将数据完全划分开了。

在本数据集中，A1和A4对于数据的划分没有作用。

电子科技大学 University of Electronic Science and Technology of China

CART连续属性处理

划分1 a_1 | $a_2 \quad a_3 \quad a_4 \quad a_5 \quad a_6 \quad a_7 \quad \cdots \quad a_{m-1} \quad a_m$

$$\frac{a_1 + a_2}{2}$$

划分2 $a_1 \quad a_2$ | $a_3 \quad a_4 \quad a_5 \quad a_6 \quad a_7 \quad \cdots \quad a_{m-1} \quad a_m$

$$\frac{a_2 + a_3}{2}$$

……

划分$m-1$ $a_1 \quad a_2 \quad a_3 \quad a_4 \quad a_5 \quad a_6 \quad a_7 \quad \cdots \quad a_{m-1}$ | a_m

$$\frac{a_{m-1} + a_m}{2}$$

CART连续属性处理

1）m个样本的连续特征A有m个值，从小到大排列 $a_1, a_2, ..., a_m$，则CART取相邻两样本值的平均数做划分点，一共有m-1个，其中第i个划分点T_i表示为：$T_i = (a_i+a_{i+1})/2$ 。

2）分别计算以这m-1个点作为二元分类点时的基尼系数。选择基尼系数最小的点为该连续特征的二元离散分类点。比如取到的基尼系数最小的点为 a_t，则小于 a_t 的值为类别1，大于 a_t 的值为类别2，这样就做到了连续特征的离散化。

剪枝

- 当分类回归树划分得太细时，会对噪声数据产生过拟合
  - 预剪枝：在每一次对结点进行划分之前，先采用验证集的数据来验证划分是否能提高结果的准确性。如果不能，就把结点标记为叶结点并退出进一步划分；如果可以就继续递归生成节点。
  - 后剪枝：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泛化性能提升，则将该子树替换为叶结点。
    - 代价复杂性剪枝、最小误差剪枝、悲观误差剪枝……

算例：天气与打篮球

CART算法：

把“天气”属性分为不同的二元组:

{小雨, (晴天, 阴天)}

{晴天, (小雨, 阴天)}

{阴天, (晴天, 小雨)}

然后分别计算三组的基尼指数，然后选取最小值作为二叉树节点。

D

天气	温度	湿度	刮风	是否打篮球
晴天	高	低	否	打篮球
晴天	高	低	是	打篮球
晴天	中	低	否	打篮球
晴天	中	中	否	打篮球
晴天	中	低	是	打篮球
小雨	低	高	否	不打篮球
小雨	低	高	是	不打篮球
小雨	中	高	否	不打篮球
阴天	低	高	否	不打篮球
阴天	中	高	否	打篮球
阴天	中	中	是	打篮球

{'晴天'}、{'小雨','阴天'}

D1

天气	温度	湿度	刮风	是否打篮球
晴天	高	低	否	打篮球
晴天	高	低	是	打篮球
晴天	中	低	否	打篮球
晴天	中	中	否	打篮球
晴天	中	低	是	打篮球

D2

天气	温度	湿度	刮风	是否打篮球
小雨	低	高	否	不打篮球
小雨	低	高	是	不打篮球
小雨	中	高	否	不打篮球
阴天	低	高	否	不打篮球
阴天	中	高	否	打篮球
阴天	中	中	是	打篮球

$$Gini(D1) = 1 - \sum_{i=1}^{i} p_i^2 = 1 - (\frac{5}{5})^2 = 0$$

$$Gini(D2) = 1 - \sum_{i=1}^{i} p_i^2 = 1 - ((\frac{4}{6})^2 + (\frac{2}{6})^2) = 0.44$$

基尼指数 $Gini\_index(D, weather) = \sum_{i=1}^{i} \frac{|D^i|}{|D|} Gini(D^i) = \frac{5}{11} Gini(D1) + \frac{6}{11} Gini(D2) = 0.24$

算例：天气与打篮球

D

天气	温度	湿度	刮风	是否打篮球
晴天	高	低	否	打篮球
晴天	高	低	是	打篮球
晴天	中	低	否	打篮球
晴天	中	中	否	打篮球
晴天	中	低	是	打篮球
小雨	低	高	否	不打篮球
小雨	低	高	是	不打篮球
小雨	中	高	否	不打篮球
阴天	低	高	否	不打篮球
阴天	中	高	否	打篮球
阴天	中	中	是	打篮球

{'小雨'}、{'晴天','阴天'}

D1

天气	温度	湿度	刮风	是否打篮球
小雨	低	高	否	不打篮球
小雨	低	高	是	不打篮球
小雨	中	高	否	不打篮球

$$Gini(D1) = 1 - \sum_{i=1}^{i} p_i^2 = 1 - (\frac{3}{3})^2 = 0$$

D2

天气	温度	湿度	刮风	是否打篮球
晴天	高	低	否	打篮球
晴天	高	低	是	打篮球
晴天	中	低	否	打篮球
晴天	中	中	否	打篮球
晴天	中	低	是	打篮球
阴天	低	高	否	不打篮球
阴天	中	高	否	打篮球
阴天	中	中	是	打篮球

$$Gini(D2) = 1 - \sum_{i=1}^{i} p_i^2 = 1 - ((\frac{7}{8})^2 + (\frac{1}{8})^2) = 0.22$$

基尼指数 $Gini\_index(D, weather) = \sum_{i=1}^{i} \frac{|D^i|}{|D|} Gini(D^i) = \frac{3}{11} Gini(D1) + \frac{8}{11} Gini(D2) = 0.16$

算例：天气与打篮球

D

天气	温度	湿度	刮风	是否打篮球
晴天	高	低	否	打篮球
晴天	高	低	是	打篮球
晴天	中	低	否	打篮球
晴天	中	中	否	打篮球
晴天	中	低	是	打篮球
小雨	低	高	否	不打篮球
小雨	低	高	是	不打篮球
小雨	中	高	否	不打篮球
阴天	低	高	否	不打篮球
阴天	中	高	否	打篮球
阴天	中	中	是	打篮球

{'阴天'}、{'晴天','小雨'}

D1

天气	温度	湿度	刮风	是否打篮球
阴天	低	高	否	不打篮球
阴天	中	高	否	打篮球
阴天	中	中	是	打篮球

$$Gini(D1) = 1 - \sum_{i=1}^{i} p_i^2 = 1 - ((\frac{1}{3})^2 + (\frac{2}{3})^2) = 0.44$$

D2

天气	温度	湿度	刮风	是否打篮球
晴天	高	低	否	打篮球
晴天	高	低	是	打篮球
晴天	中	低	否	打篮球
晴天	中	中	否	打篮球
晴天	中	低	是	打篮球
小雨	低	高	否	不打篮球
小雨	低	高	是	不打篮球
小雨	中	高	否	不打篮球

$$Gini(D2) = 1 - \sum_{i=1}^{i} p_i^2 = 1 - ((\frac{5}{8})^2 + (\frac{3}{8})^2) = 0.47$$

基尼指数 $Gini\_index(D, weather) = \sum_{i=1}^{i} \frac{|D^i|}{|D|} Gini(D^i) = \frac{3}{11} Gini(D1) + \frac{8}{11} Gini(D2) = 0.46$

算例：天气与打篮球

通过计算，我们发现{小雨}、{晴天、阴天}的划分方法，其基尼指数值最小，所以如果我们以天气属性作为划分，那么就选择{小雨}、{晴天、阴天}的分类作为第一级分裂点。

以此类推，进行第二级、第三极、。。。的构建，直到结束。

ID3，C4.5，CART比较

	支持	裂变指标	特征类型	缺失值	过拟合
ID3	分类	信息增益	离散值	无法处理	无法处理
C4.5	分类	信息增益率	离散值、连续值	可以处理	预剪枝、后剪枝
CART	分类（二叉树）、回归	Gini指数、均方差	离散值、连续值	可以处理	预剪枝、后剪枝

不过他们都存在一个问题就是，都是以单个特征进行划分的。有时候如果是按着一个特征线性组合来进行划分说不定会更好。这个叫做多变量决策树。
还有就是引出随机森林的原因：如果样本有一点点改变，树结构可能会发生比较大的变化。

ID3，C4.5，CART比较