第五讲 决策树

授课老师：郭 迟 教授

guochi@whu.edu.cn

武汉大学测绘学院

2021.11

线性模型

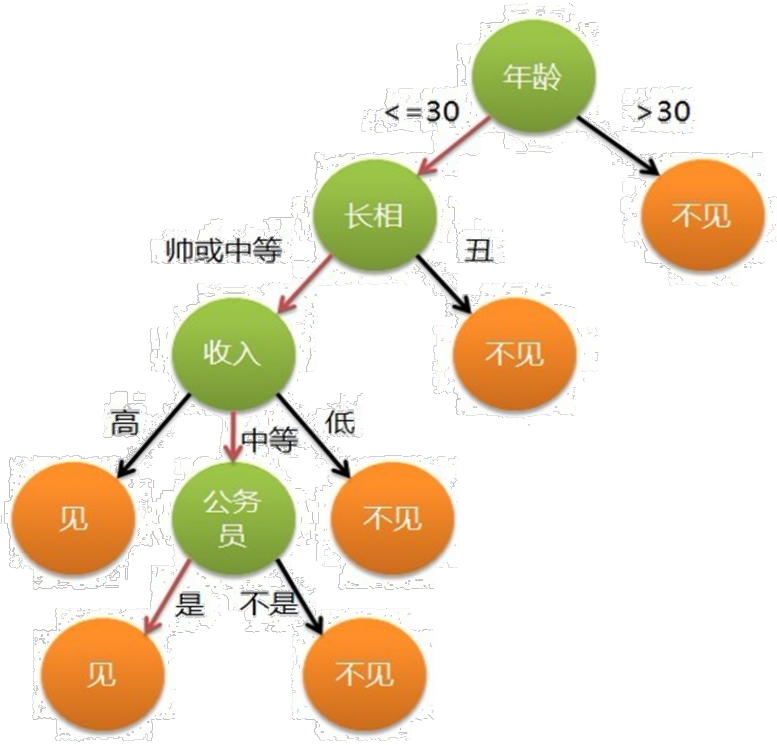
1. 决策树的基本概念
2. 最佳划分的选择
3. 决策树剪枝 (选学)
4. 决策树相关问题的讨论 (选学)

第五讲 决策树

决策树的基本概念

- 基于规则的思维模式：人类在进行某些取舍决策或事物判断时，一般会基于一些规则，而且这些规则往往是潜在的，只在遇到具体的判断条件时，才会显式地、具体地表现出来

- **决策树（Decision Tree）模型**：一通过对一系列问题进行*if-then* 的推导而实现最终决策的ML模型。由于很多判断规则多数时候是潜在的，我们能了解的往往是据以判断的属性取值和最终的判断结果。因此，使用计算机建立决策树模型，与建立其他分类器模型一样，需要研究如何从这些训练数据中，提取潜在判断规则的方法



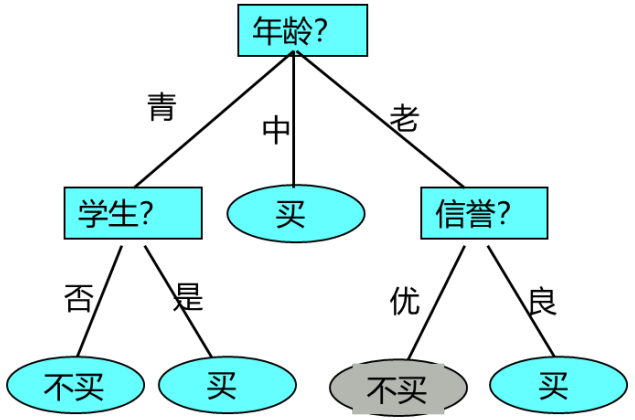
“是否同意相亲”的决策示意图

第五讲 决策树

决策树的基本概念

1. 决策树（Decision Tree）的生成过程

- 决策树中最上面的结点称为根结点，是整个决策树的开始。每个分支是一个新的内部结点（internal node）也叫决策结点，代表一个问题或者决策，通常对应待分类对象的属性。每个叶结点（leaf node）代表一种可能的分类结果
- 在沿着决策树从上到下的遍历过程中，对每个结点上问题的不同测试输出导致不同的分支，最后会达到一个叶子结点。这一过程就是利用决策树进行分类的过程



计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
128	青	中	否	良	不买
64	青	低	是	良	买
132	老	中	是	良	买
64	青	中	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

决策树的基本概念

- 递归实现
- 深度优先

1. 决策树（Decision Tree）的生成过程

- 给定数据集 $D=\{(x_1,y_1),(x_2,y_2),\ldots(x_m,y_m)\}$ 和属性集$A=\{a_1,a_2,\ldots,a_d\}$

  决策树生成函数记为 TreeGenerate(D, A)

输入： 训练集 $D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\ldots,(\boldsymbol{x}_m,y_m)\}$;
　　属性集 $A=\{a_1,a_2,\ldots,a_d\}$.
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** $A=\varnothing$ **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_v, $A\setminus\{a_*\}$)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

(A) 递归出口
(B)
(C)

- 贪心法（Greedy algorithm）思想：不断划分

决策树的基本概念

1. 决策树（Decision Tree）的生成过程

TreeGenerate(D, A) {

1. 从A中选出最优划分属性“年龄”

2. 依照“年龄”的属性值划分了3个子数据集

}

1

年龄

D

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
128	青	中	否	良	不买
64	青	低	是	良	买
132	老	中	是	良	买
64	青	中	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

第五讲 决策树

决策树的基本概念

1. 决策树（Decision Tree）的生成过程

TreeGenerate(D, A) {　1

1. 从A中选出最优划分属性“年龄”

2. 依照“年龄”的属性值划分了3个子数据集{

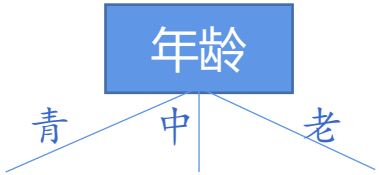
TreeGenerate($D_{年龄1}$, A\{年龄}) ; 1.1

TreeGenerate($D_{年龄2}$, A\{年龄}) ; 1.2

TreeGenerate($D_{年龄3}$, A\{年龄}) ; 1.3

}

}



$D_{年龄1}$

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	青	中	否	良	不买
64	青	低	是	良	买
64	青	中	是	优	买

$D_{年龄2}$

计数	年龄	收入	学生	信誉	归类：买计算机？
128	中	高	否	良	买
64	中	低	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买

$D_{年龄3}$

计数	年龄	收入	学生	信誉	归类：买计算机？
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
132	老	中	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

第五讲 决策树

决策树的基本概念

1. 决策树（Decision Tree）的生成过程

TreeGenerate($D_{年龄1}$, A\{年龄}) {

1.1

1. 从{A\{年龄}}中选出最优划分属性 “学生”

2. 依照 “学生” 的属性值划分了2个子数据集{

1.1.1

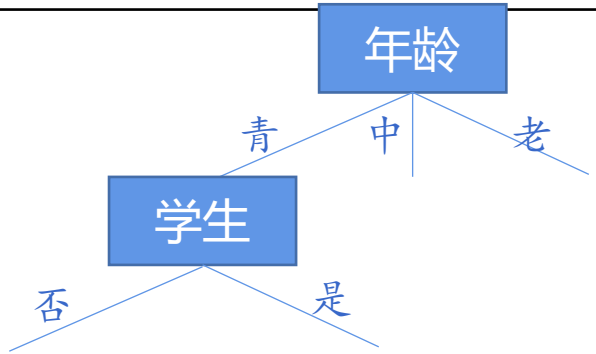
TreeGenerate($D_{年龄1\text{-}学生1}$, A\{年龄、学生}) ;

1.1.2

TreeGenerate($D_{年龄1\text{-}学生2}$, A\{年龄、学生}) ;

}

}



$D_{年龄1\text{-}学生1}$

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	青	中	否	良	不买

$D_{年龄1\text{-}学生2}$

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	低	是	良	买
64	青	中	是	优	买

决策树的基本概念

1. 决策树（Decision Tree）的生成过程

TreeGenerate($D_{年龄1\text{-}学生1}$, A\{年龄、学生}) { 1.1.1

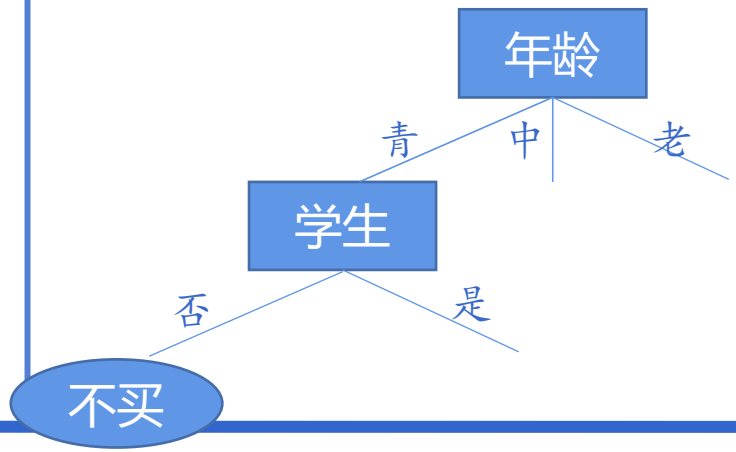
1. $D_{年龄1\text{-}学生1}$ 中的样本属于同一类“不买”

2. 生成“不买”类的叶子结点;

3. return; (A出口)
}

$D_{年龄1\text{-}学生1}$

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	青	中	否	良	不买



第五讲 决策树

决策树的基本概念

1. 决策树（Decision Tree）的生成过程

TreeGenerate($D_{年龄1\text{-}学生2}$, A\{年龄、学生}) { 1.1.2

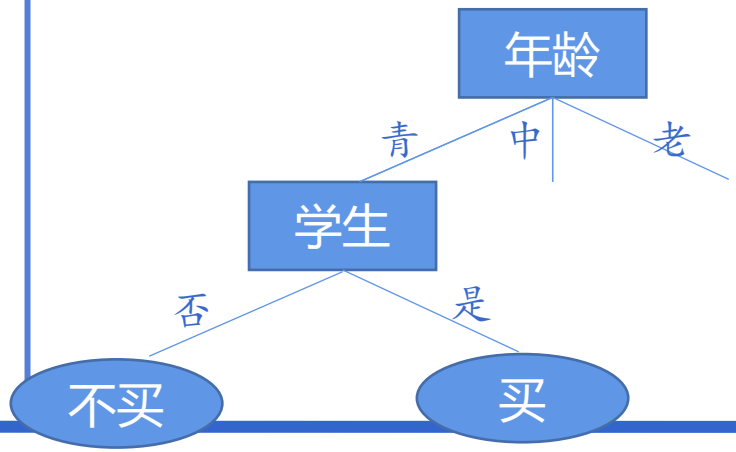
1. $D_{年龄1\text{-}学生2}$ 中的样本属于同一类“买”

2. 生成“买”类的叶子结点;

3. return; (A出口)
}

$D_{年龄1\text{-}学生2}$

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	低	是	良	买
64	青	中	是	优	买



第五讲 决策树

决策树的基本概念

1. 决策树（Decision Tree）的生成过程

TreeGenerate($D_{年龄2}$, $A\backslash\{年龄\}$) {

1. $D_{年龄2}$ 中的样本属于同一类“买”

2. 生成“买”类的叶子结点;

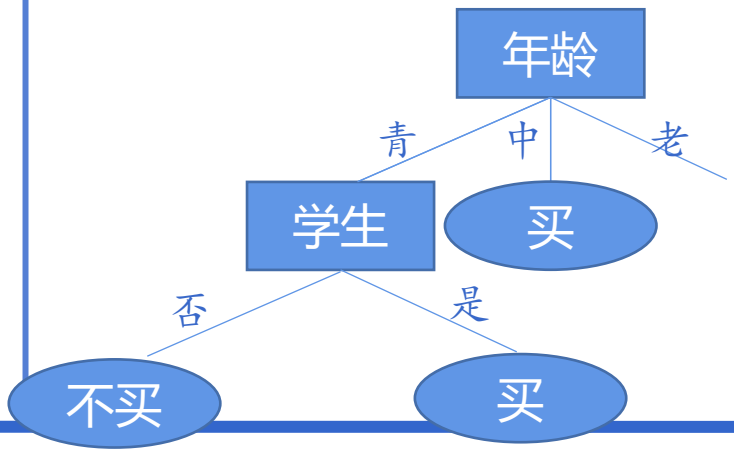
3. return; (A出口)

}

1.2

$D_{年龄2}$

计数	年龄	收入	学生	信誉	归类：买计算机？
128	中	高	否	良	买
64	中	低	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买



第五讲 决策树

决策树的基本概念

1. 决策树（Decision Tree）的生成过程

TreeGenerate($D_{年龄3}$, A\{年龄}) {

1.3

1. 从{A\{年龄}}中选出最优划分属性 “信誉”

2. 依照 “信誉” 的属性值划分了2个子数据集{

TreeGenerate($D_{年龄3\text{-}信誉1}$, A\{年龄、信誉}) ;

1.3.1

TreeGenerate($D_{年龄3\text{-}信誉2}$, A\{年龄、信誉})

1.3.2

}

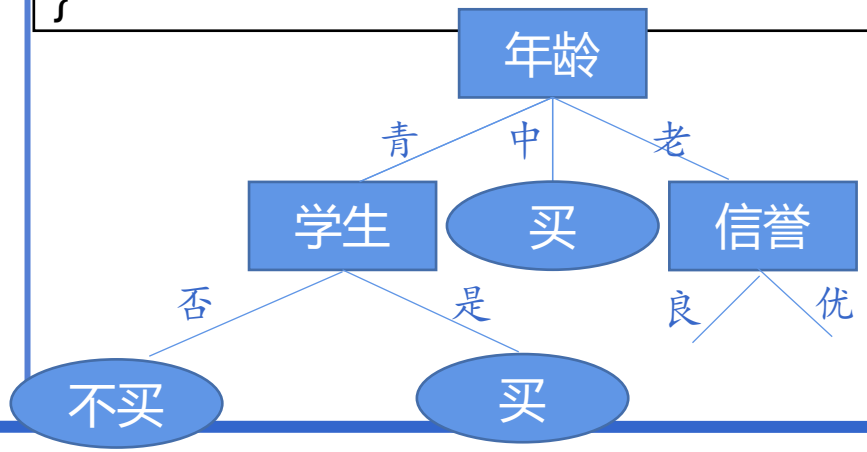
}

$D_{年龄3\text{-}信誉1}$

计数	年龄	收入	学生	信誉	归类：买计算机？
60	老	中	否	良	买
64	老	低	是	良	买
132	老	中	是	良	买

$D_{年龄3\text{-}信誉2}$

计数	年龄	收入	学生	信誉	归类：买计算机？
64	老	低	是	优	不买
63	老	中	否	优	不买
1	老	中	否	优	买



第五讲 决策树

决策树的基本概念

1. 决策树（Decision Tree）的生成过程

TreeGenerate($D_{年龄3\text{-}信誉1}$, A\{年龄、信誉}) { 1.3.1

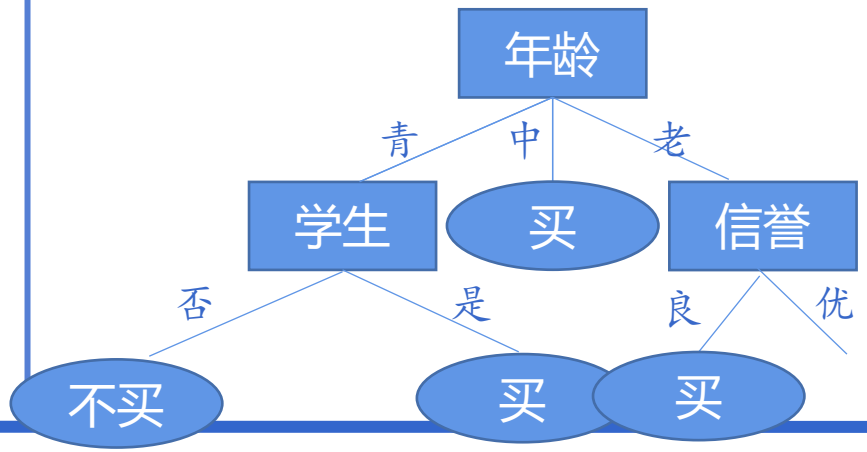
1. $D_{年龄3\text{-}信誉1}$ 中的样本属于同一类“买”

2. 生成“买”类的叶子结点;

3. return; (A出口)
}

$D_{年龄3\text{-}信誉1}$

计数	年龄	收入	学生	信誉	归类：买计算机？
60	老	中	否	良	买
64	老	低	是	良	买
132	老	中	是	良	买



第五讲 决策树

决策树的基本概念

1. 决策树（Decision Tree）的生成过程

TreeGenerate($D_{年龄3\text{-}信誉2}$, A\{年龄、信誉}) { 1.3.2

1. 从{A\{年龄、信誉}}中选出属性“学生”

2. 依照“学生”的属性值划分了2个子数据集{

TreeGenerate($D_{年龄3\text{-}信誉2\text{-}学生1}$, A\{年龄、信誉、学生}) ； 1.3.2.1

TreeGenerate($D_{年龄3\text{-}信誉2\text{-}学生2}$, A\{年龄、信誉、学生}) ； 1.3.2.2

}

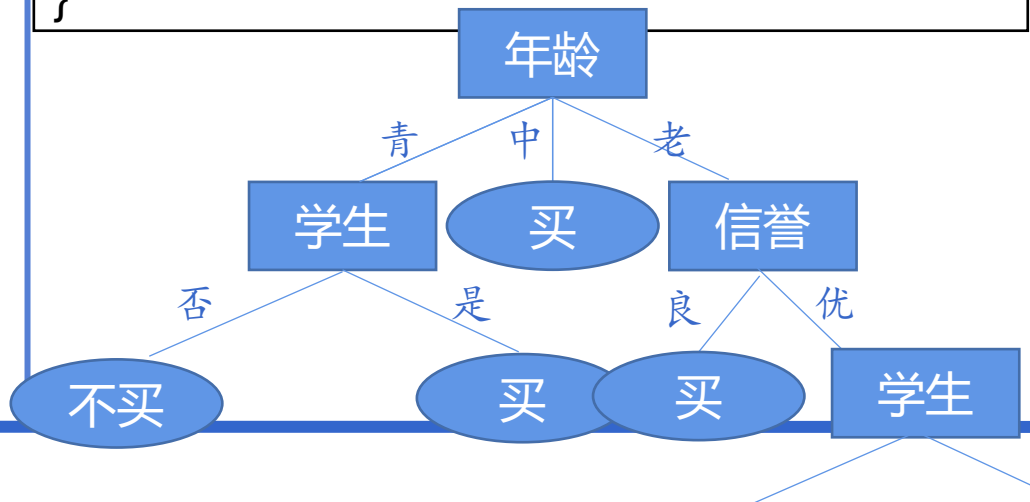
}

$D_{年龄3\text{-}信誉2\text{-}学生1}$

计数	年龄	收入	学生	信誉	归类：买计算机？
64	老	低	是	优	不买

$D_{年龄3\text{-}信誉2\text{-}学生2}$

计数	年龄	收入	学生	信誉	归类：买计算机？
63	老	中	否	优	不买
1	老	中	否	优	买



第五讲 决策树

决策树的基本概念

1. 决策树（Decision Tree）的生成过程

TreeGenerate($D_{年龄3\text{-}信誉2\text{-}学生1}$, $A\backslash\{年龄、信誉、学生\}$)
{
1. $D_{年龄3\text{-}信誉2\text{-}学生1}$中的样本属于同一类“不买”

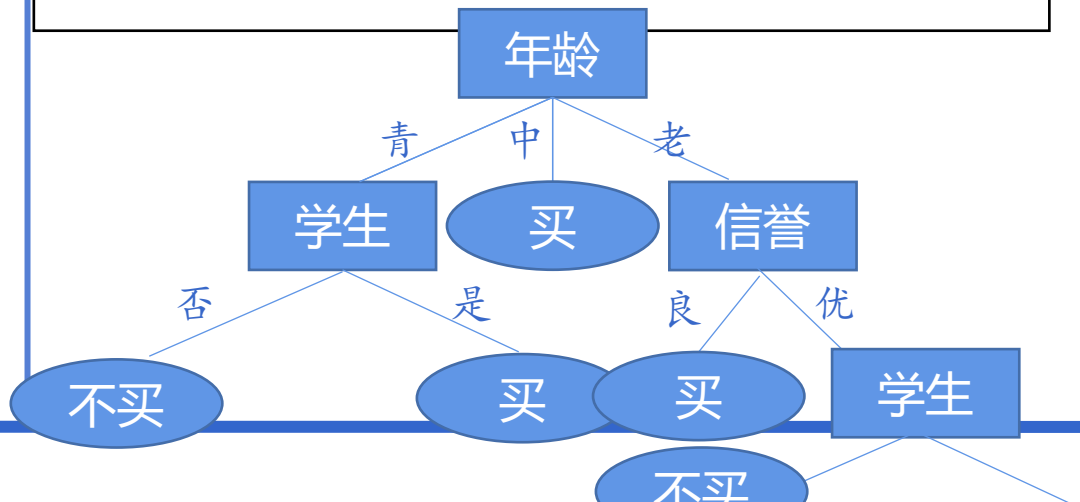
2. 生成“不买”类的叶子结点;

3. return; (A出口)
}

1.3.2.1

$D_{年龄3\text{-}信誉2\text{-}学生1}$

计数	年龄	收入	学生	信誉	归类：买计算机？
64	老	低	是	优	不买



第五讲 决策树

决策树的基本概念

1. 决策树（Decision Tree）的生成过程

TreeGenerate($D_{年龄3\text{-}信誉2\text{-}学生2}$, A\{年龄、信誉、学生}))

{

1. $D_{年龄3\text{-}信誉2\text{-}学生2}$在“收入”上的属性相同

2. 生成叶子结点;

3. 标记为$D_{年龄3\text{-}信誉2\text{-}学生2}$中多的类“不买”；

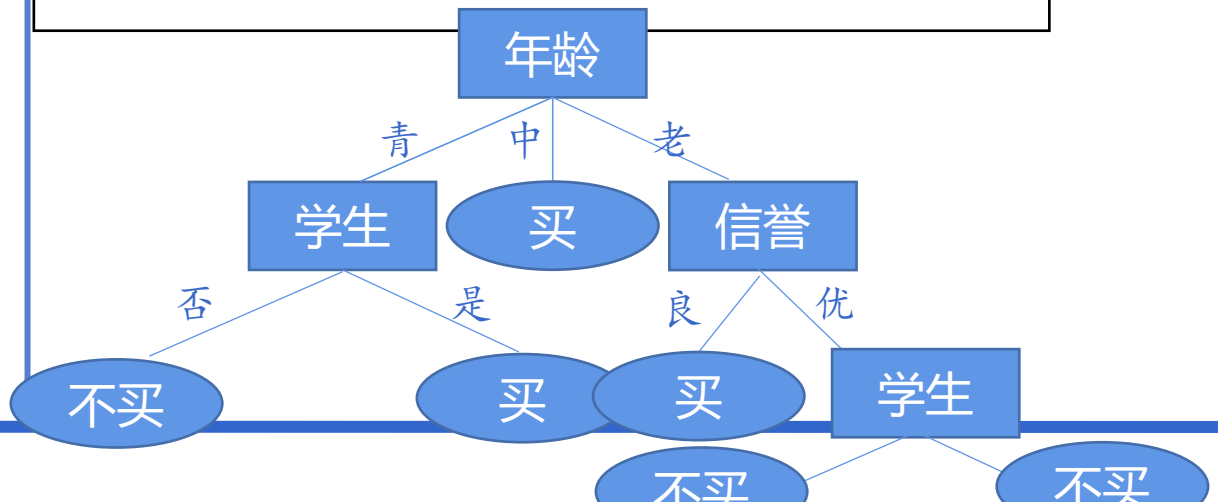
4. return; (B出口)

}

1.3.2.2

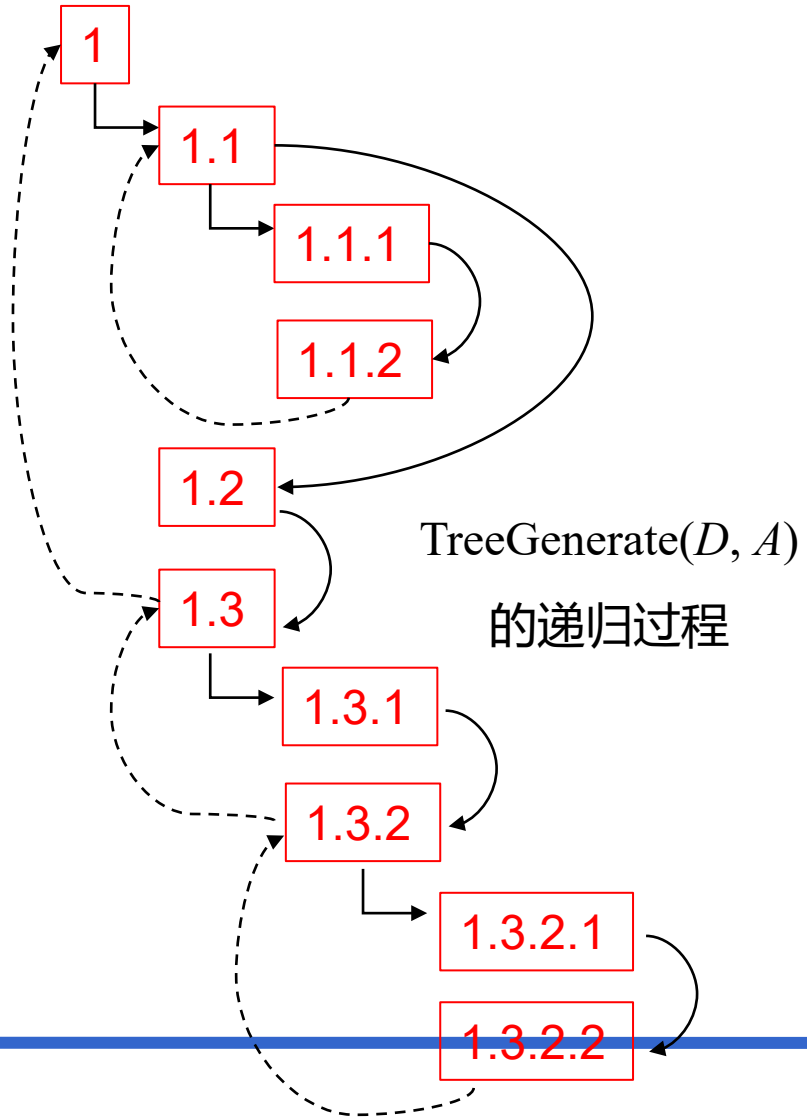
$D_{年龄3\text{-}信誉2\text{-}学生2}$

计数	年龄	收入	学生	信誉	归类：买计算机？
63	老	中	否	优	不买
1	老	中	否	优	买



决策树的基本概念

1. 决策树（Decision Tree）的生成过程



- 三种递归出口分析：
  - 当前结点包含的样本属于同一类，不需要继续划分，生成该类的叶子结点；
  - 当前属性集为空，或样本在所有属性集上取值相同，按样本最多的类生成该类的叶子结点；（以当前结点的后验概率决定叶子的类）
  - 下一个划分的结点的样本集为空，按当前结点样本最多的类生成该类的叶子结点；（把当前结点的样本分布当作下一个划分的结点的先验）

2. 决策树的形态

- 决策树可以是多叉树，也可以是二叉树
- 决策树可以实现多分类
- 决策树是一个非参数、非线性的分类器
- 每次递归的划分属性不同，决策树的形态不同

提问：生成最优决策树的关键是什么？

线性模型

1 决策树的基本概念

2 最佳划分的选择

3 决策树剪枝 (选学)

4 决策树相关问题的讨论 (选学)

最佳划分的选择

1. 信息熵与结点纯度

- 在信息论与概率统计中，熵（Entropy）是表示随机变量不确定性的度量。设X是一个取有限个值的离散随机变量，其概率分布

$$P(X = X_i) = P_i\ ,\ i = 1,\ 2,\ \ldots\ ,\ n$$

则随机变量X 的熵定义为

$$H(X) = -\sum_{i=1}^{n} p_i \log\ p_i$$

对应到ML里，假设样本集D中第k类样本所占的比例是$p_k(k=1,2,\ldots,|y|)$，则样本D的熵为：

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

教材P25

- 样本D的熵越小，则可认为其纯度（purity）越高，是度量纯度的代表性指标之一。用不纯性度量（impurity measure）可以评估分类树划分的优劣

2. ID3决策树

- ID3算法是一种经典的决策树学习算法，由Quinlan于1979年在悉尼大学提出。ID3 算法的核心是在决策树各个结点上使用信息增益（information gain）进行不纯性度量，递归地构建决策树
- 特征a对样本集D的信息增益定义为样本集D的信息熵与给定特征a条件下D的条件熵之差，记为：

$$\begin{aligned}\text{Gain}(D,a) &= \text{Ent}(D) - \text{Ent}(D|a) \\ &= \text{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|}\text{Ent}(D^v)\end{aligned}$$

假定离散属性a有V个可能的取值$\{a^1, a^2, \ldots, a^V\}$,若使用$a$来对样本集$D$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$D^v$

- ID3算法最优划分属性选择依据：$a_* = \arg\max_{a \in A} \text{Gain}(D, a)$

最佳划分的选择

2. ID3决策树

e.g. 前面决策树生成的例子中，为什么首先选择“年龄”属性进行划分

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
128	青	中	否	良	不买
64	青	低	是	良	买
132	老	中	是	良	买
64	青	中	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

$$\text{Ent}(D) = -\frac{641}{1024}log_2\frac{641}{1024} - \frac{383}{1024}log_2\frac{383}{1024}$$

$$= 0.287$$

$$\text{Gain}(D, 年龄) = 0.287 - (\frac{384}{1024}\text{x}0.276 + 0 + \frac{384}{1024}\text{x}0.276)$$
$$= 0.08$$

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	青	中	否	良	不买
64	青	低	是	良	买
64	青	中	是	优	买

$D^v_{年龄1}=384$ $\text{Ent}(D_{年龄1})=0.276$

计数	年龄	收入	学生	信誉	归类：买计算机？
128	中	高	否	良	买
64	中	低	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买

$D^v_{年龄2}=256$ $\text{Ent}(D_{年龄2})=0$

计数	年龄	收入	学生	信誉	归类：买计算机？
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
132	老	中	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

$D^v_{年龄3}=384$ $\text{Ent}(D_{年龄3})=0.276$

最佳划分的选择

2. ID3决策树

e.g. 前面决策树生成的例子中，为什么首先选择“年龄”属性进行划分

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
128	青	中	否	良	不买
64	青	低	是	良	买
132	老	中	是	良	买
64	青	中	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

$$\text{Ent}(D)= -\frac{641}{1024}log_2\frac{641}{1024}-\frac{383}{1024}log_2\frac{383}{1024}$$

$$= 0.287$$

$$\text{Gain}(D,\text{学生})=0.287-(\frac{484}{1024}\text{x}0.1697+\frac{540}{1024}\text{x}0.3874)$$
$$=0.0025$$

计数	年龄	收入	学生	信誉	归类：买计算机？
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
64	青	低	是	良	买
132	老	中	是	良	买
64	青	中	是	优	买
32	中	高	是	良	买

$D^v_{\text{学生}1}=484$　$\text{Ent}(D_{\text{学生}1})=0.1697$

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
128	青	中	否	良	不买
32	中	中	否	优	买
63	老	中	否	优	不买
1	老	中	否	优	买

$D^v_{\text{学生}2}=540$　$\text{Ent}(D_{\text{学生}2})=0.3874$

最佳划分的选择

3. C4.5决策树

- 信息增益准则对可取值数目较多（分枝较多）的属性有所偏好，为减少这种偏好可能带来的不利影响， Quinlan在1993年提出了著名的C4.5决策。C4.5不直接使用信息增益，而是使用“增益率”(gain ratio)来选择最优划分属性：

$$\text{Gain\_ratio}(D,a)=\frac{\text{Gain}(D,a)}{\text{IV}(a)}, \qquad \text{IV}(a)=-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\log_2\frac{|D^v|}{|D|}$$

其中IV(a)称为属性a的固有值（intrinsic value）。属性a可能取值数目越大，一般IV(a)也会越大，作为分母平衡一下信息增益

- C4.5算法最优划分属性选择依据：$a_*=\underset{a\in A}{\arg\max}\ \text{Gain\_ratio}(D,a)$

4. CART决策树

- 分类与回归树 (classification and regression tree, CART) 模型由Breiman 等人于1984 年提出，是一种应用广泛的决策树学习方法。CART由特征选择、树的生成及剪枝组成，既可以用于分类也可以用于回归
- CART 决策树是二叉树，内部结点特征的取值为“是”和“否”，左分支是取值为“是”的分支，右分支是“否”分支。这样的决策树等价于递归地二分每个特征，将输入空间即特征空间划分为有限个单元，并在这些单元上确定预测的概率分布，也就是在给定的输入特征条件下，输出的条件概率分布
- CART算法最优划分属性选择的依据是最小化基尼系数：

$$a_* = \arg\min_{a \in A} \text{Gini\_index}(D, a)$$

最佳划分的选择

4. CART决策树

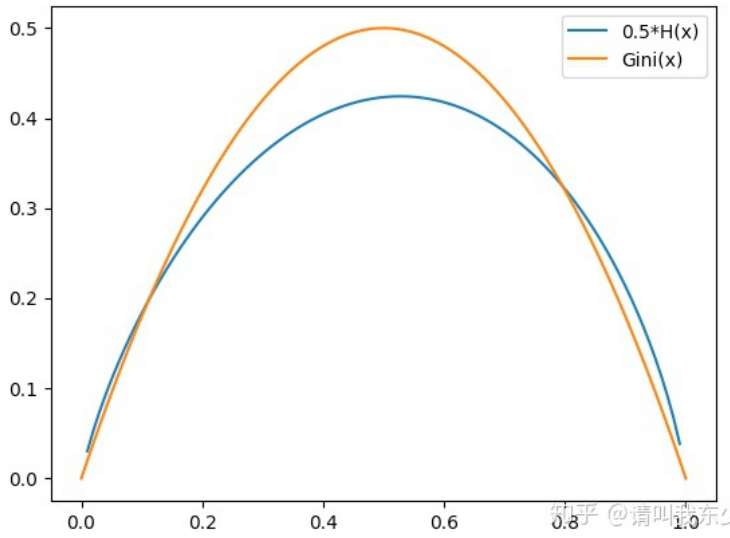
- 与熵的定义类似，基尼系数（Gini index）也可以用来表示数据纯度。假设样本集D中第k类样本所占的比例是$p_k(k=1,2,\ldots,|\mathcal{Y}|)$，则样本$D$的基尼系数为：

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

基尼系数越小纯度越高

- 属性a的基尼系数表示为：

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



最佳划分的选择

5. 连续与缺失值的处理*（略作了解）

西瓜书P83 4.4

- 属性取值是连续值的不能直接根据数值进行结点划分。需要使用连续属性离散化技术。C4.5决策树采用二分法(bi-partition)对连续属性进行处理

计数	年龄	收入	学生	信誉	归类：买计算机？
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
128	青	中	否	良	不买
64	青	低	是	良	买
132	老	中	是	良	买
64	青	中	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

计数	年龄	收入	学生	信誉	归类：买计算机？
64	18	高	否	良	不买
64	19	高	否	优	不买
128	25	高	否	良	买
60	45	中	否	良	买
64	50	低	是	良	买
64	55	低	是	优	不买
64	28	低	是	优	买
128	20	中	否	良	不买
64	21	低	是	良	买
132	55	中	是	良	买
64	22	中	是	优	买
32	32	中	否	优	买
32	35	高	是	良	买
63	60	中	否	优	不买
1	65	中	否	优	买

最佳划分的选择

5. 连续与缺失值的处理*（略作了解）

西瓜书P83 4.4

- 可以根据样本中属性a的连续取值$\{a^1,a^2,a^3,\ldots,a^n\}$，按照中位点$t_i=\frac{a^i+a^{i+1}}{2}$的方式设置一批候选划分点$T_a=\{t_1,t_2,\ldots\}$，再通过计算每一个划分点二分后的

计数	年龄	收入	学生	信誉	归类：买计算机？
64	18	高	否	良	不买
64	19	高	否	优	不买
128	25	高	否	良	买
60	45	中	否	良	买
64	50	低	是	良	买
64	55	低	是	优	不买
64	28	低	是	优	买
128	20	中	否	良	不买
64	21	低	是	良	买
132	55	中	是	良	买
64	22	中	是	优	买
32	32	中	否	优	买
32	35	高	是	良	买
63	60	中	否	优	不买
1	65	中	否	优	买

(25+22)/2=23.5

计数	年龄	收入	学生	信誉	归类：买计算机？
64		高	否	良	不买
64		高	否	优	不买
128		高	否	良	买
60		中	否	良	买
64		低	是	良	买
64		低	是	优	不买
64		低	是	优	买
128		中	否	良	不买
64		低	是	良	买
132		中	是	良	买
64		中	是	优	买
32		中	否	优	买
32		高	是	良	买
63		中	否	优	不买
1		中	否	优	买

5. 连续与缺失值的处理*（略作了解）

西瓜书P83 4.4

- 对于样本存在缺失值的情况，决策树也可以生成：
  - 如何在属性值缺失的情况下进行划分属性选择?

    以属性不缺失的样本计算信息增益，乘以一个无缺失样本占该属性所有样本的比例作为权重系数

  - 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?

    将该样本以一定的概率（权重系数）划分到所有子结点中。权重系数可以用各个分枝上属性值不缺失的样本之间的比例确定

线性模型

1 决策树的基本概念

2 最佳划分的选择

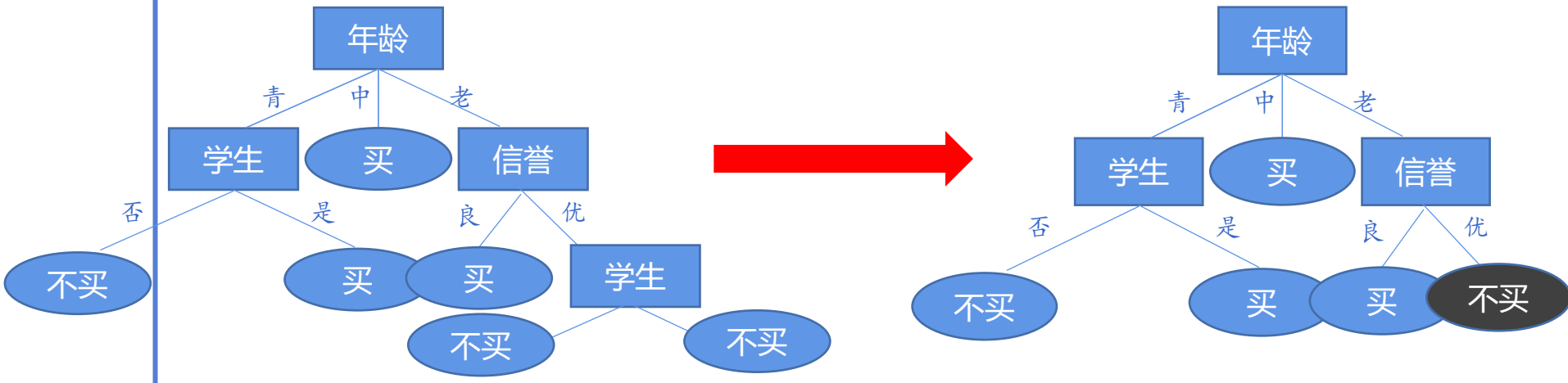
3 决策树剪枝 (选学)

4 决策树相关问题的讨论 (选学)

第五讲 决策树

决策树剪枝

- 决策树生成算法递归地产生决策树，直到不能继续下去为止。这样产生的树往往对训练数据的分类很准确，但对未知的测试数据的分类却没有那么准确，即出现过拟合现象。过拟合的原因在于学习时过多地考虑如何提高对训练数据的正确分类，从而构建出过于复杂的决策树
- 在决策树学习中将已生成的树进行简化的过程称为**剪枝(pruning)**。具体地说，剪枝从已生成的树上裁掉一些子树或叶结点，并将其根结点或父结点作为新的叶结点，从而简化分类树模型



第五讲 决策树

决策树剪枝

- 剪枝的方法有先剪枝（prepruning）和后剪枝（postpruning）。后者在实践中效果更好

教材P130

- 判定是否剪枝的依据可以是划分前后的分类精度

错误率定义为

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)\neq y_i\right)\ .$$

复习第二讲

精度则定义为

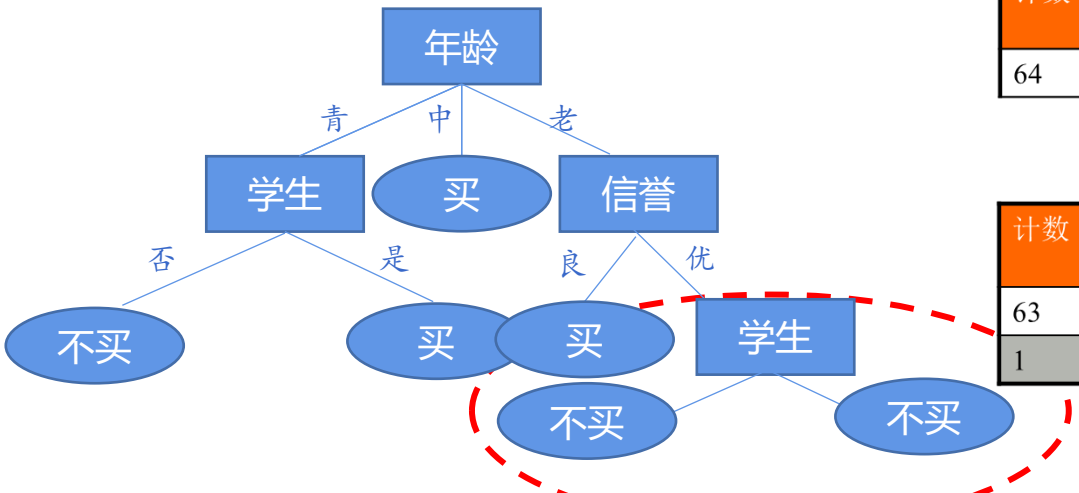
$$\begin{aligned}\mathrm{acc}(f;D) &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)=y_i\right)\\ &= 1-E(f;D)\ .\end{aligned}$$

- 划分前父结点对应的样本取其中样本最多的类作为分类结果，会得到一个分类精度。经过属性a划分后，统计各分枝上分类精度。如果精度有提升，则保留这个划分，否则中止划分（先剪枝）或剪掉该分枝（后剪枝）

第五讲 决策树

决策树剪枝

- 通过分类精度确定后剪枝



$D_{年龄3\text{-}信誉2\text{-}学生1}$

计数	年龄	收入	学生	信誉	归类：买计算机？
64	老	低	是	优	不买

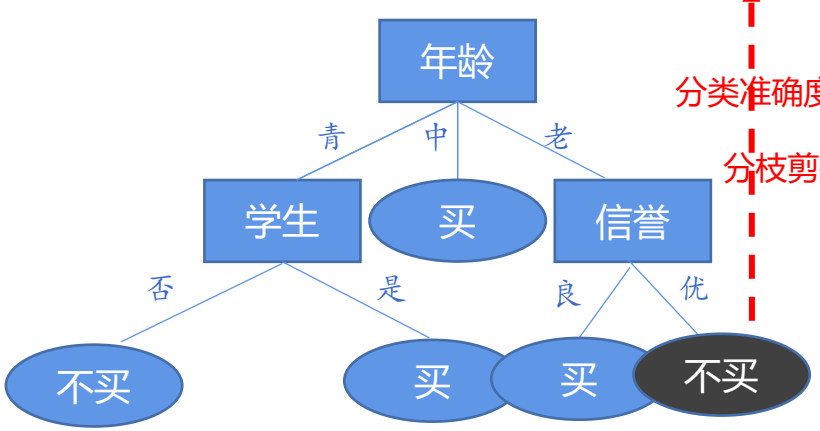
$D_{年龄3\text{-}信誉2\text{-}学生2}$

计数	年龄	收入	学生	信誉	归类：买计算机？
63	老	中	否	优	不买
1	老	中	否	优	买

准确度：127/128=0.99

分类准确度不变

分枝剪掉



$D_{年龄3\text{-}信誉2}$

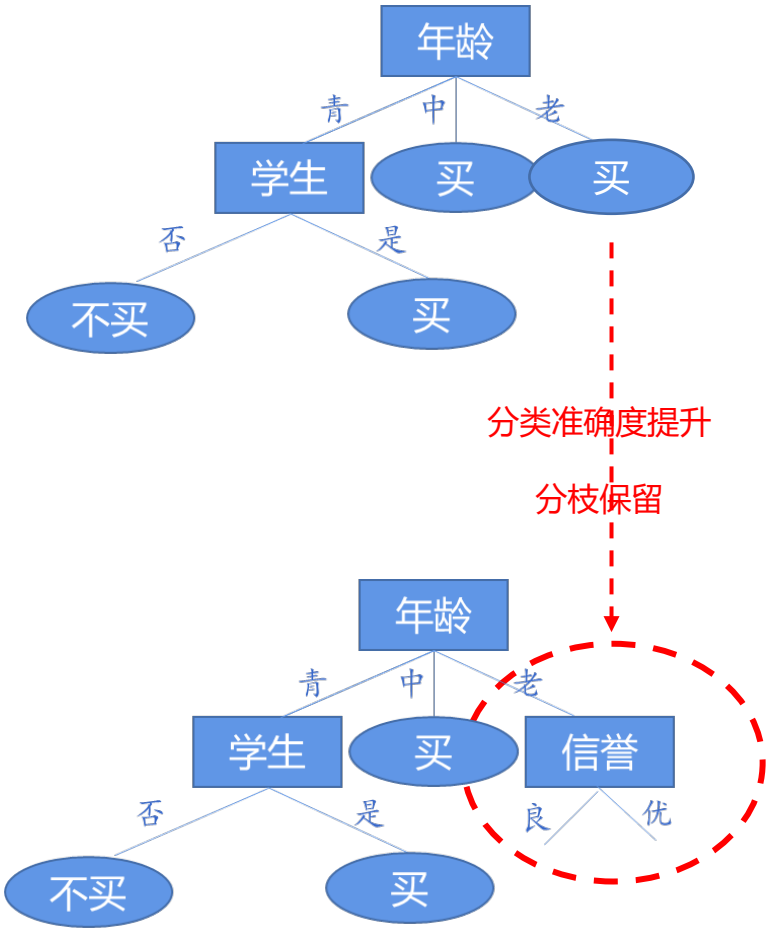
计数	年龄	收入	学生	信誉	归类：买计算机？
64	老	低	是	优	不买
63	老	中	否	优	不买
1	老	中	否	优	买

准确度：127/128=0.99

第五讲 决策树

决策树剪枝

- 通过分类精度确定后剪枝



$D_{年龄3}$

计数	年龄	收入	学生	信誉	归类：买计算机？
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
132	老	中	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

准确度：257/384=0.67

$D_{年龄3-信誉1}$

计数	年龄	收入	学生	信誉	归类：买计算机？
60	老	中	否	良	买
64	老	低	是	良	买
132	老	中	是	良	买

$D_{年龄3-信誉2}$

计数	年龄	收入	学生	信誉	归类：买计算机？
64	老	低	是	优	不买
63	老	中	否	优	不买
1	老	中	否	优	买

准确度：383/384=0.997

线性模型

1 决策树的基本概念

2 最佳划分的选择

3 决策树剪枝 (选学)

4 决策树相关问题的讨论 (选学)

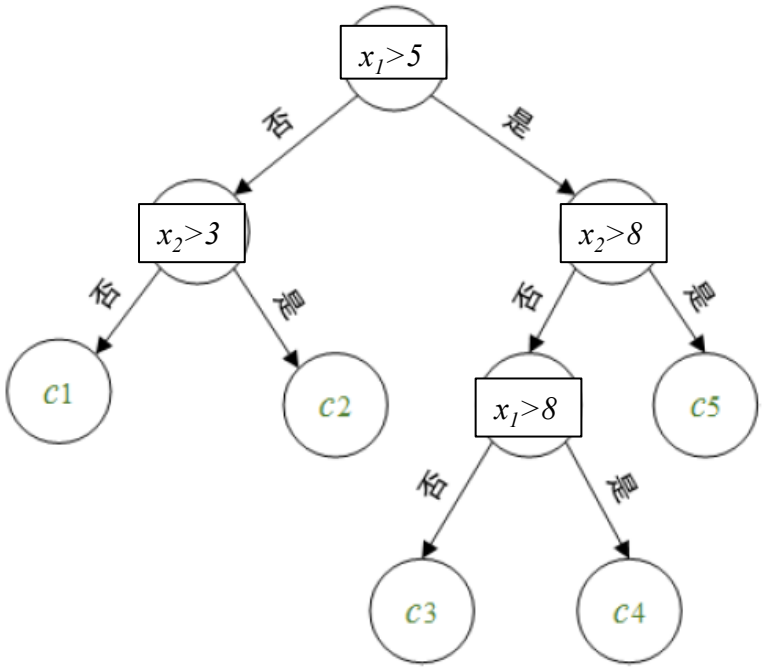
决策树相关问题的讨论

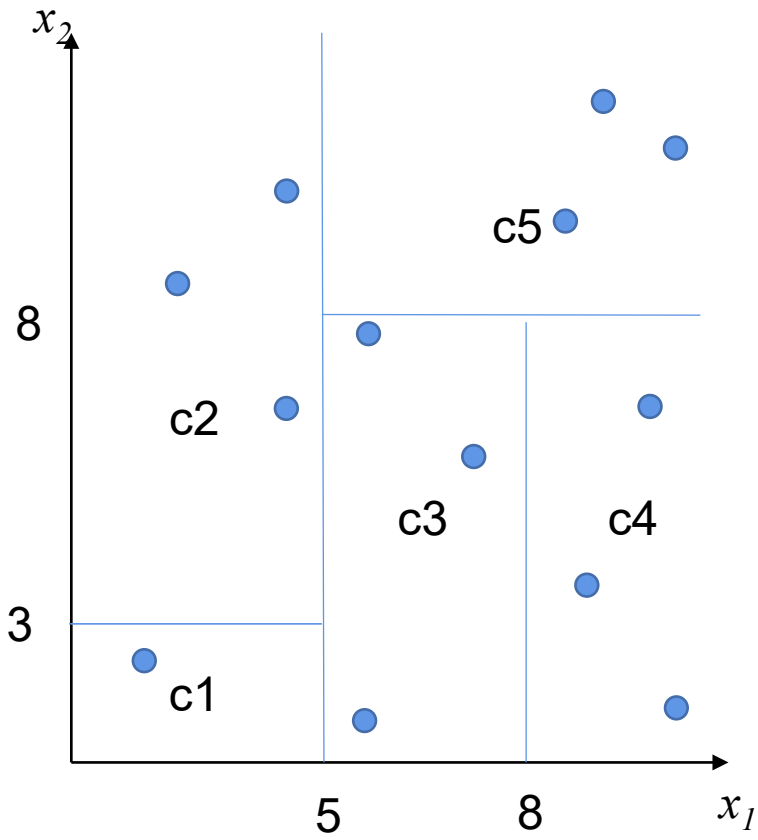
- 决策树学习本质上是从训练数据集中归纳出一组分类规则。与训练数据集不相矛盾的决策树（即能对训练数据进行正确分类的决策树）可能有多个，也可能一个也没有。我们需要的是一个与训练数据矛盾较小的决策树，同时具有很好的泛化能力

决策树相关问题的讨论

1. 决策树的决策界

- 决策树的分类边界由一系列轴平行的线段组成



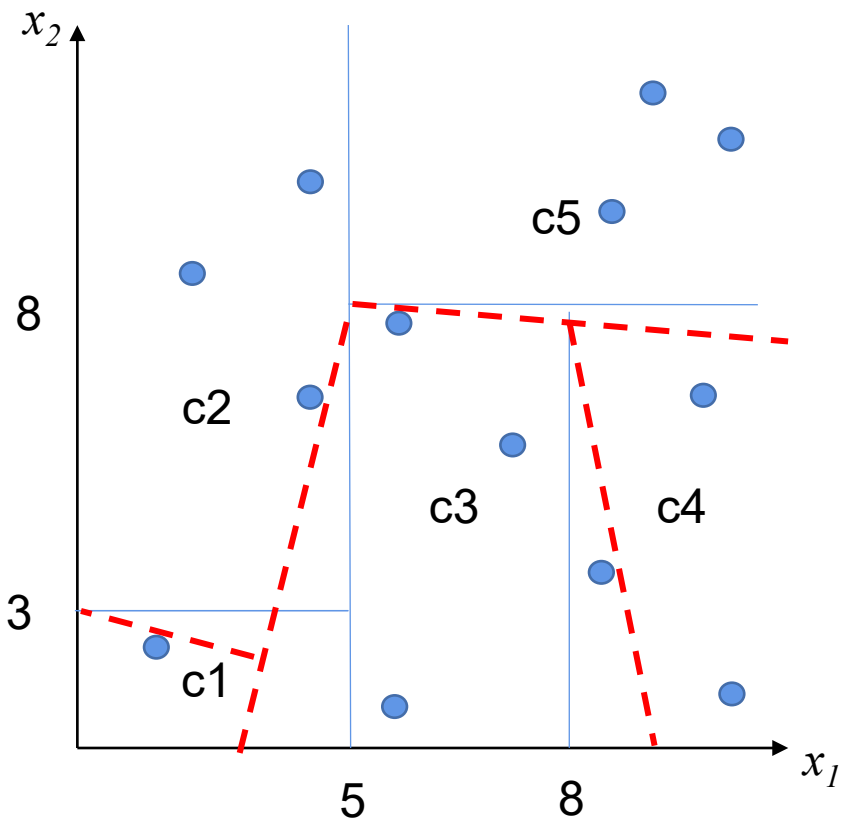


决策树相关问题的讨论

2. 多变量决策树

教材P134

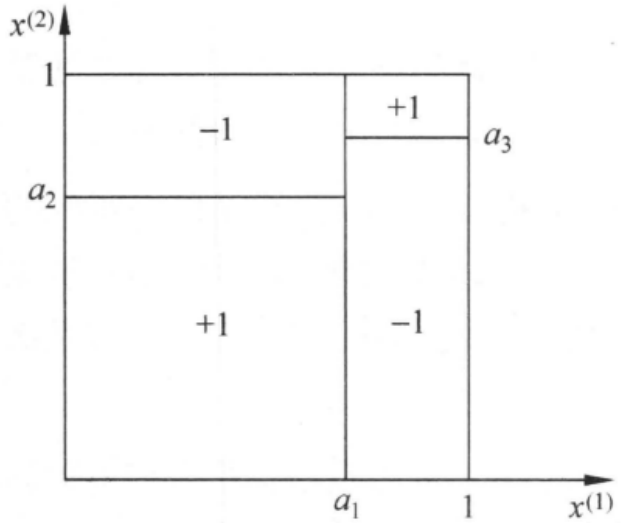
- 将“轴平行的线段”改变为“斜线段”，可以构造更加复杂的决策树。这样的决策树的内部结点不再仅根据某单个属性进行测试，而是对属性的线性组合进行测试，即每个内部结点变为了一个$\sum w_i a_i = t$的线性分类器，这种决策树称为多变量树（multivariate tree）



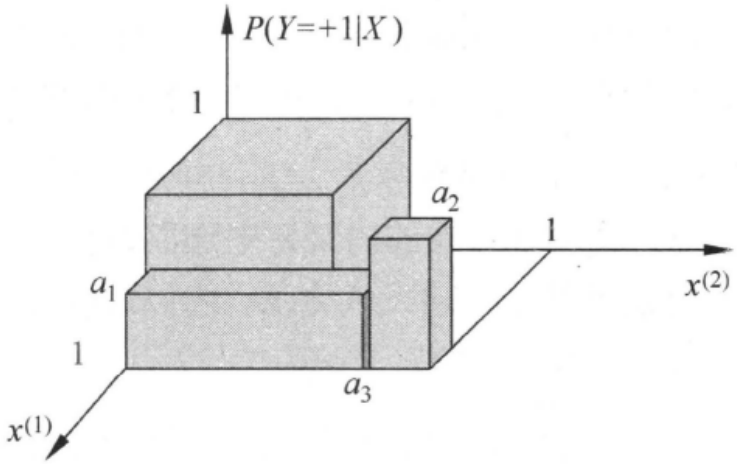
决策树相关问题的讨论

3. 决策树与贝叶斯分类

- 从另一个角度看，决策树学习是由训练数据集直接估计后验概率。基于特征空间划分的类的条件概率模型有无穷多个。我们选择的后验条件概率模型应该不仅对训练数据有很好的拟合，而且对未知数据有很好的预测



(a) 特征空间划分



(b) 条件概率分布

3. 决策树与贝叶斯分类

- 给定特征条件下类的条件概率分布，即在特征空间的一个划分(partition)上。将特征空间划分为互不相交的单元(cell)或区域(region)，并在每个单元定义一个类的概率分布就构成了一个条件概率分布
- 决策树的一条路径对应于划分中的一个单元。决策树所表示的条件概率分布由各个单元给定条件下类的条件概率分布组成
- 假设X为表示特征的随机变量，Y为表示类的随机变量，那么这个条件概率分布可表示为$P(Y|X)$。X取值于给定划分下单元的集合，Y取值于类的集合
- 各叶结点（单元）上的条件概率往往偏向某一个类，即属于某一类的概率较大。决策分类时将该结点的实例强行分到条件概率大的那一类中

决策树相关问题的讨论

4. 决策树用于回归分析（CART）

教材P128

- 假设X和Y分布为输入和输出变量，并且Y是连续变量，给定训练数据集$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中$x_i=(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)})$为输入实例（特征向量），$n$为特征个数，$i=1,2,\cdots,N$，$N$为样本容量。对特征空间的划分采用启发式方法，每次划分逐一考察当前集合中所有特征的所有取值，根据**平方误差最小化准则**选择其中最优划分点。如对训练集中第j个特征变量$x^{(j)}$和它的取值s，作为最优划分属性和化分点，并定义两个区域$R_1(j,s)=\{x|x^{(j)}\le s\}$和$R_2(j,s)=\{x|x^{(j)}>s\}$，为找出最优的$j$和$s$，对下式求解：

$$\min_{j,s}=[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2]$$

- 这种用于回归的决策树又叫做最小二乘回归树

决策树相关问题的讨论

4. 决策树用于回归分析

- c_1和c_2为划分后两个区域内固定的输出值，方括号内的两个min意为使用的是最优的c_1和c_2，也就是各自区域内平方误差最小的c_1和c_2，易知这两个最优的输出值就是各自对应区域内Y的均值，所以上式可写为

$$\min_{j,s}=[\sum_{x_i\in R_1(j,s)}(y_i-\hat{c}_1)^2+\sum_{x_i\in R_2(j,s)}(y_i-\hat{c}_2)^2]$$

其中$\hat{c}_1=\frac{1}{N_1}\Sigma_{x_i\in R_1(j,s)}y_i$，$\hat{c}_2=\frac{1}{N_2}\Sigma_{x_i\in R_2(j,s)}y_i$。

e.g 一维决策树回归

x	1	2	3	4	5	6	7	8	9	10
y	5.56	5.7	5.91	6.4	6.8	7.05	8.9	8.7	9	9.05

4. 决策树用于回归分析

- 一维数据回归，最优划分属性就是x。下面计算9个划分点的损失函数：

s	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5
c_1	5.56	5.63	5.72	5.89	6.07	6.24	6.62	6.88	7.11
c_2	7.5	7.73	7.99	8.25	8.54	8.91	8.92	9.03	9.05
L(s)	15.72	12.07	8.36	5.78	3.91	1.93	8.01	11.73	15.74

损失函数最小，最优划分

划分区域为：$R_1=\{1,2,3,4,5,6\}, R_2=\{7,8,9,10\}$

对应输出值：$c_1=6.24$, $c_2=8.91$

- 下面继续对R_1区域划分：

s	1.5	2.5	3.5	4.5	5.5
c_1	5.56	5.63	5.72	5.89	6.07
c_2	6.37	6.54	6.75	6.93	7.05

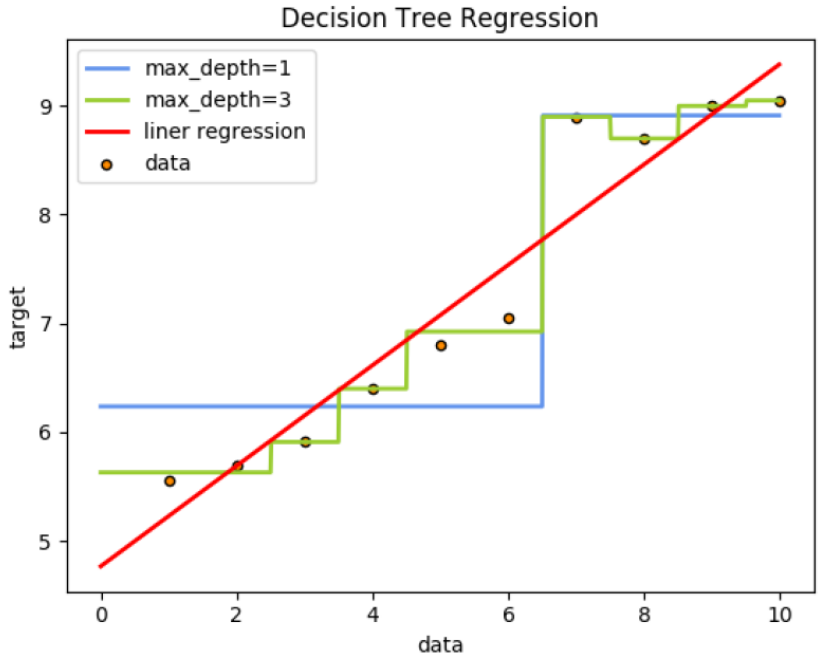
4. 决策树用于回归分析

- 得到第二个最优划分点3.5

s	1.5	2.5	3.5	4.5	5.5
L(s)	1.3087	0.754	0.2771	0.4368	1.0644

- 得到最终的回归树：

$$T = \begin{cases} 5.72, & x \le 3.5 \\ 6.75, & 3.5 < x \le 6.5 \\ 8.91, & x > 6.5 \end{cases}$$



决策树相关问题的讨论

练习题 表 5.1 是一个由 15 个样本组成的贷款申请训练数据。数据包括贷款申请人的 4 个特征（属性）：第 1 个特征是年龄，有 3 个可能值：青年，中年，老年；第 2 个特征是有工作，有 2 个可能值：是，否；第 3 个特征是有自己的房子，有 2 个可能值：是，否；第 4 个特征是信贷情况，有 3 个可能值：非常好，好，一般。表的最后一列是类别，是否同意贷款，取 2 个值：是，否。

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

决策树相关问题的讨论

练习题 对 表5.1 所给的训练数据集 D，根据信息增益准则选择最优特征。

解 首先计算经验熵 $H(D)$。

$$H(D) = -\frac{9}{15}\log_2\frac{9}{15} - \frac{6}{15}\log_2\frac{6}{15} = 0.971$$

然后计算各特征对数据集 D 的信息增益。分别以 A_1，A_2，A_3，A_4 表示年龄、有工作、有自己的房子和信贷情况 4 个特征，则

（1）

$$\begin{aligned} g(D, A_1) &= H(D) - \left[\frac{5}{15}H(D_1) + \frac{5}{15}H(D_2) + \frac{5}{15}H(D_3)\right] \\ &= 0.971 - \left[\frac{5}{15}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) + \right. \\ &\quad \left. \frac{5}{15}\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) + \frac{5}{15}\left(-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}\right)\right] \\ &= 0.971 - 0.888 = 0.083 \end{aligned}$$

这里 D_1，D_2，D_3 分别是 D 中 A_1（年龄）取值为青年、中年和老年的样本子集。类似地，

决策树相关问题的讨论

练习题

(2)

$$g(D, A_2) = H(D) - \left[\frac{5}{15}H(D_1) + \frac{10}{15}H(D_2)\right]$$

$$= 0.971 - \left[\frac{5}{15} \times 0 + \frac{10}{15}\left(-\frac{4}{10}\log_2\frac{4}{10} - \frac{6}{10}\log_2\frac{6}{10}\right)\right] = 0.324$$

(3)

$$g(D, A_3) = 0.971 - \left[\frac{6}{15} \times 0 + \frac{9}{15}\left(-\frac{3}{9}\log_2\frac{3}{9} - \frac{6}{9}\log_2\frac{6}{9}\right)\right]$$

$$= 0.971 - 0.551 = 0.420$$

(4)

$$g(D, A_4) = 0.971 - 0.608 = 0.363$$

最后，比较各特征的信息增益值。由于特征 A_3（有自己的房子）的信息增益值最大，所以选择特征 A_3 作为最优特征。

决策树相关问题的讨论

练习题 应用CART 算法生成决策树

解 首先计算各特征的基尼指数，选择最优特征以及其最优切分点。仍采用例 5.2 的记号，分别以 A_1，A_2，A_3，A_4 表示年龄、有工作、有自己的房子和信贷情况 4 个特征，并以 1，2，3 表示年龄的值为青年、中年和老年，以 1，2 表示有工作和有自己的房子的值为是和否，以 1，2，3 表示信贷情况的值为非常好、好和一般。

求特征 A_1 的基尼指数:

$$\mathrm{Gini}(D, A_1 = 1) = \frac{5}{15}\left(2 \times \frac{2}{5} \times \left(1 - \frac{2}{5}\right)\right) + \frac{10}{15}\left(2 \times \frac{7}{10} \times \left(1 - \frac{7}{10}\right)\right) = 0.44$$

$$\mathrm{Gini}(D, A_1 = 2) = 0.48$$

$$\mathrm{Gini}(D, A_1 = 3) = 0.44$$

由于 $\mathrm{Gini}(D, A_1 = 1)$ 和 $\mathrm{Gini}(D, A_1 = 3)$ 相等，且最小，所以 $A_1 = 1$ 和 $A_1 = 3$ 都可以选作 A_1 的最优切分点。

求特征 A_2 和 A_3 的基尼指数：

$$\mathrm{Gini}(D, A_2 = 1) = 0.32$$

$$\mathrm{Gini}(D, A_3 = 1) = 0.27$$

由于 A_2 和 A_3 只有一个切分点，所以它们就是最优切分点。

练习题 应用CART 算法生成决策树

求特征 A_4 的基尼指数：

$$\text{Gini}(D, A_4 = 1) = 0.36$$

$$\text{Gini}(D, A_4 = 2) = 0.47$$

$$\text{Gini}(D, A_4 = 3) = 0.32$$

$\text{Gini}(D, A_4 = 3)$ 最小，所以 $A_4 = 3$ 为 A_4 的最优切分点。

在 A_1，A_2，A_3，A_4 几个特征中，$\text{Gini}(D, A_3 = 1) = 0.27$ 最小，所以选择特征 A_3 为最优特征，$A_3 = 1$ 为其最优切分点。于是根结点生成两个子结点，一个是叶结点。对另一个结点继续使用以上方法在 A_1，A_2，A_4 中选择最优特征及其最优切分点，结果是 $A_2 = 1$。依此计算得知，所得结点都是叶结点。

对于本问题，按照 CART 算法所生成的决策树与按照 ID3 算法所生成的决策树完全一致。