数据挖掘

第3章 分类-基本概念与决策树

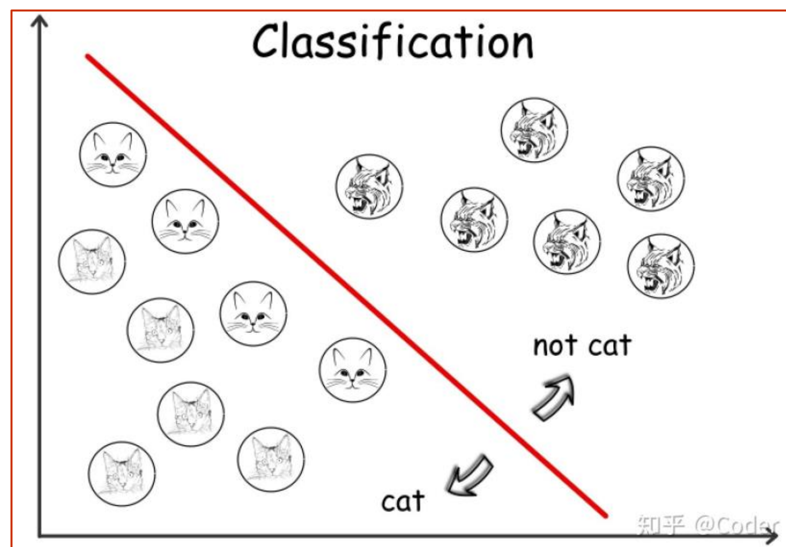
教师：王东京

学院：计算机学院

邮箱：dongjing.wang@hdu.edu.cn

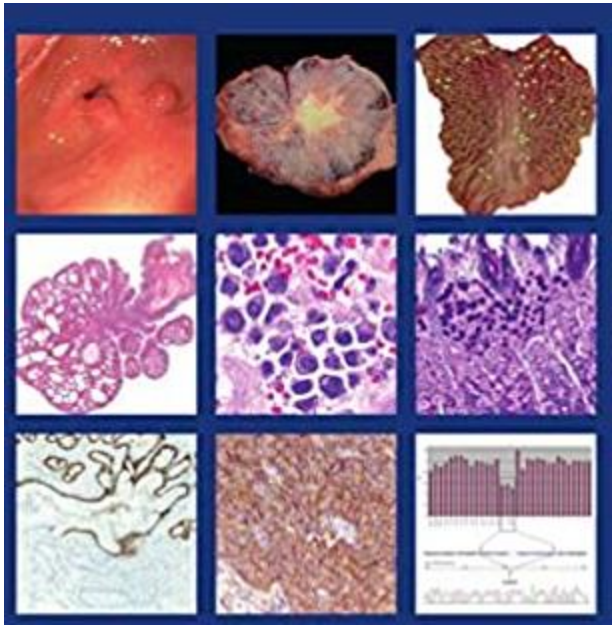
分类Classification: 定义

- 给定记录的集合（训练集 training set）
  - 每条记录表示为元组 (x,y), x 是属性（attribute）集合， y 是类别标签（class label）
    - ◆ x：属性，预测变量，自变量，输入
    - ◆ y：类别，响应，因变量，输出
- 任务（task）：
  - 学习将每个属性集 x 映射到预定义的类别标签 y 中的模型



Examples of Classification Task

任务	属性集合, x	类别标签, y
邮件/信息分类	从电子邮件/信息的标题和内容中提取的特征	“垃圾邮件/信息”或者“非垃圾邮件/信息”
识别肿瘤细胞	从X射线或核磁共振成像扫描中提取的特征	恶性或良性细胞
星系编目	从望远镜图像中提取的特征	椭圆形，螺旋形或不规则形状的星系



Examples of Classification Task

还有哪些分类任务?

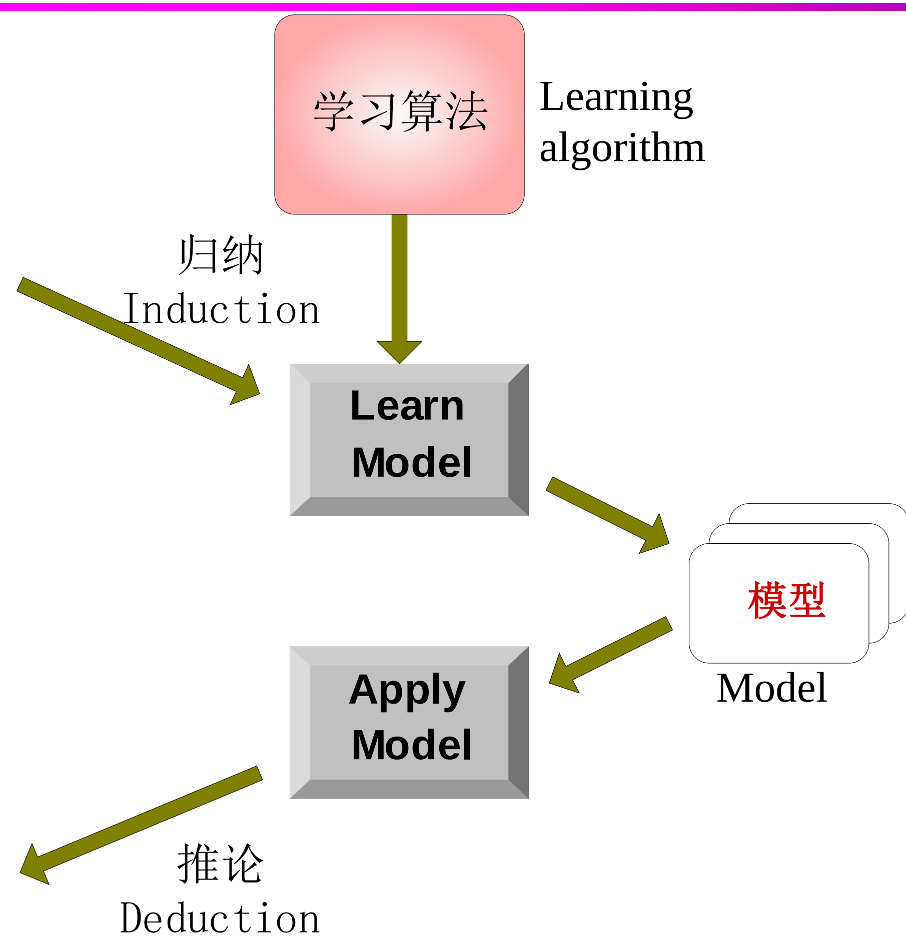
构建分类模型的通用手段

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

训练集
Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

测试集
Test Set



分类技术 Classification Techniques

基本分类器 Base Classifiers

- 基于决策树的方法 Decision Tree based Methods
- 基于规则的方法 Rule-based Methods
- 最近邻 Nearest-neighbor
- 神经网络 Neural Networks
- 深度学习 Deep Learning
- 朴素贝叶斯和贝叶斯信念网络 Naïve Bayes and Bayesian Belief Networks
- 支持向量机Support Vector Machines

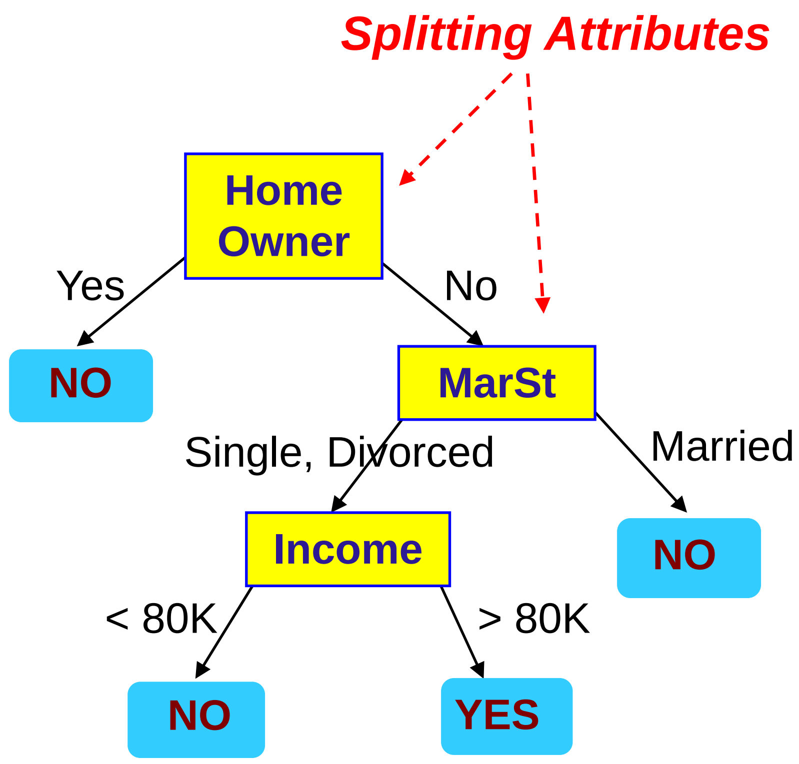
集成分类器 Ensemble Classifiers

- Boosting, Bagging, 随机森林 Random Forests

决策树示例：借款人违约

ID	户主 Home Owner	婚姻状况 Marital Status	年收入 Annual Income	拖欠贷款 Defaulted Borrower
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

训练集 Training Data

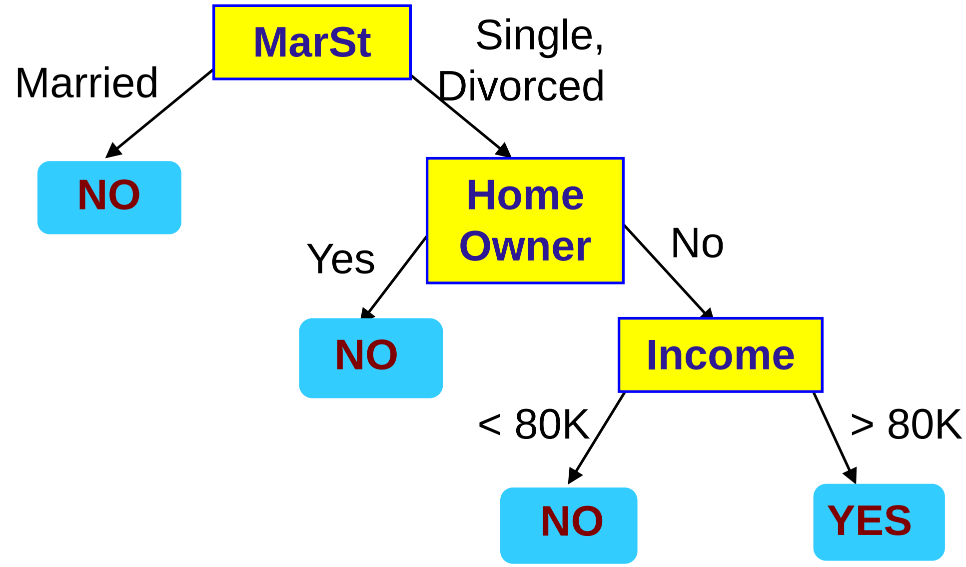


Model: Decision Tree

决策树示例2：借款人违约

categorical categorical continuous class

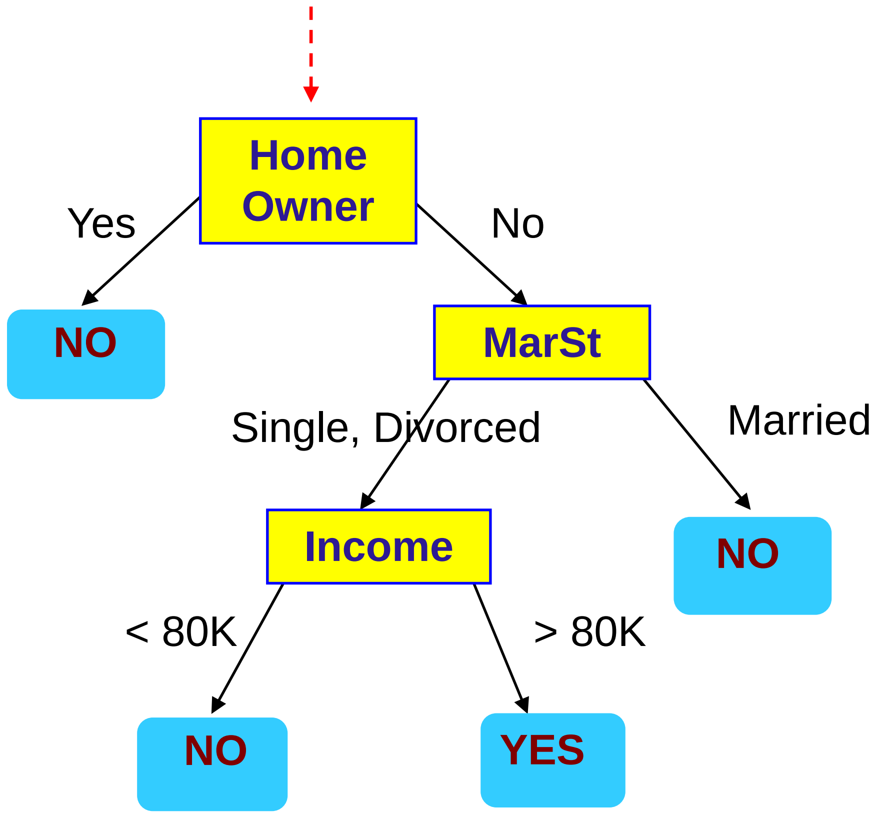
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



对于相同的数据，可能存在不止一棵适合的决策树！

应用到测试集 Apply Model to Test Data

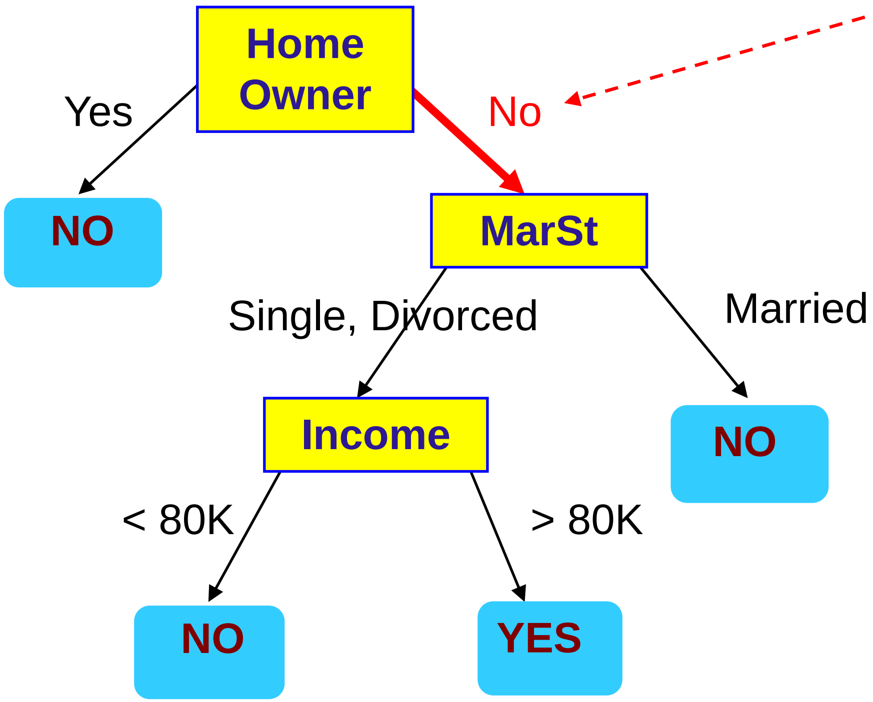
Start from the root of tree.



测试数据 Test Data

Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?

Apply Model to Test Data



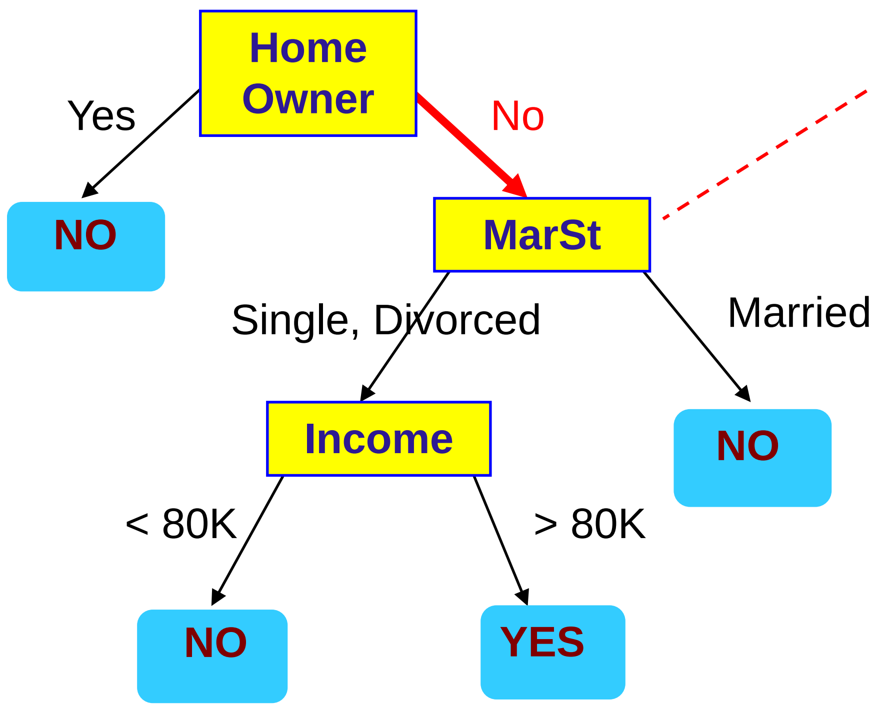
Test Data

Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?

Apply Model to Test Data

Test Data

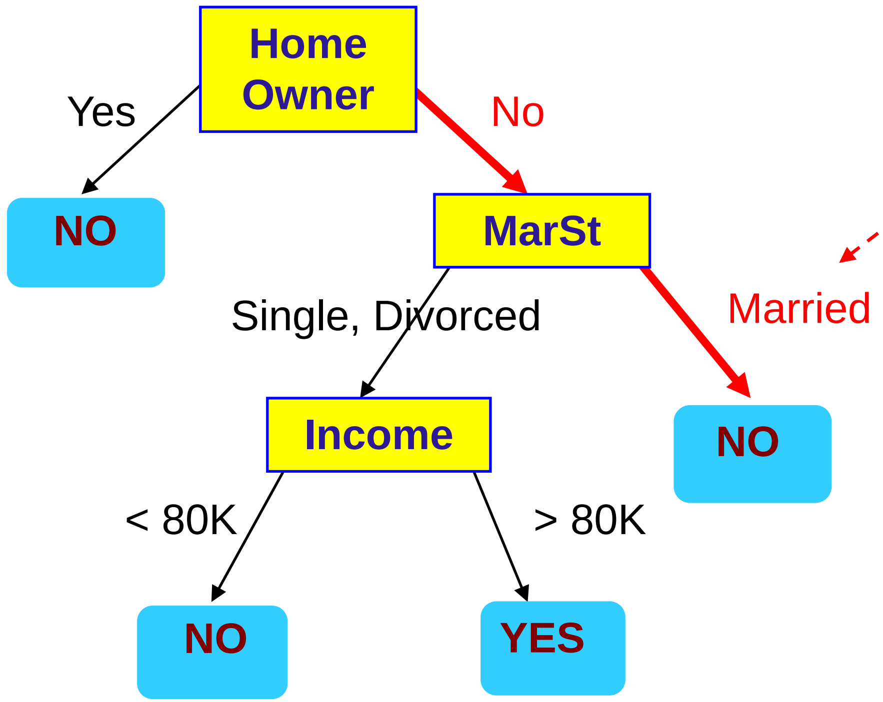
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



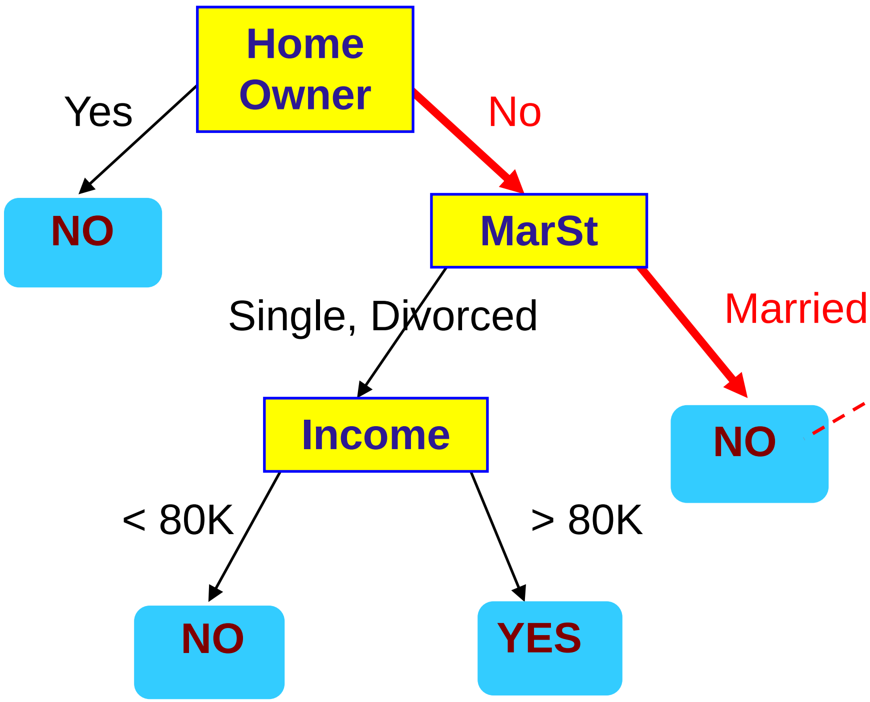
Apply Model to Test Data

Test Data

Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



Apply Model to Test Data



Test Data

Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?

Assign Defaulted to "No"

动物分类例子

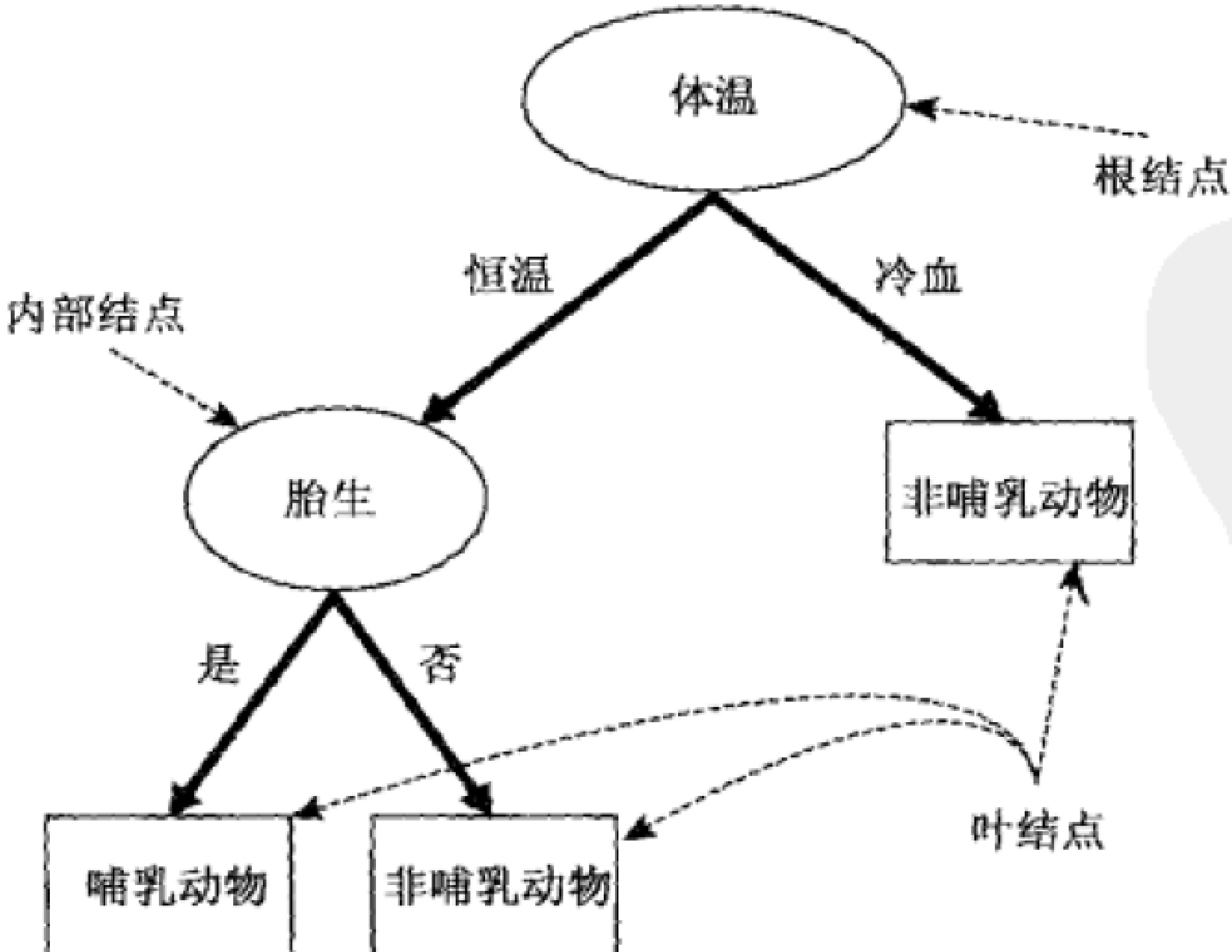


图 4-4　哺乳动物分类问题的决策树

动物分类例子

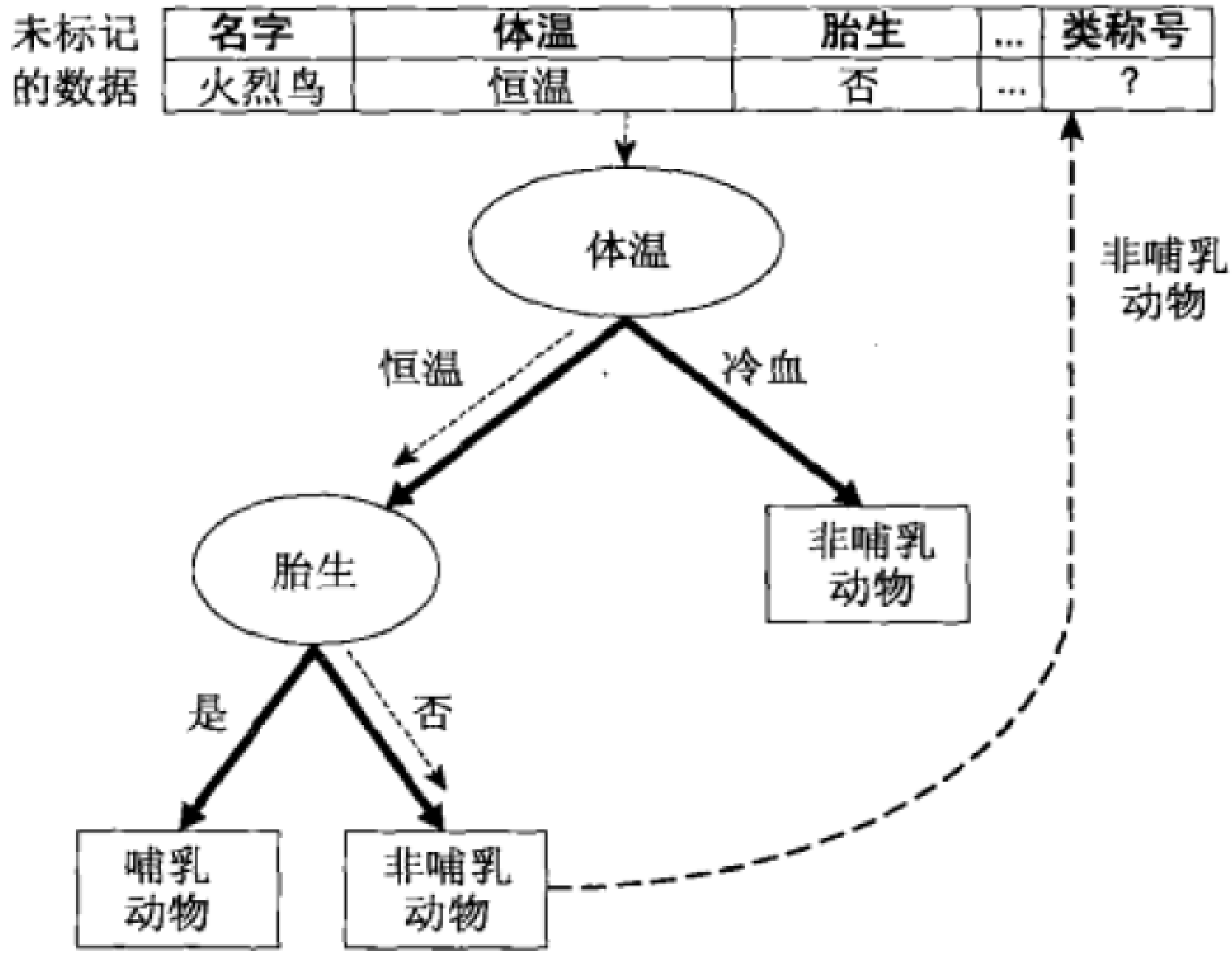


图 4-5　对一种未标记的脊椎动物分类。虚线表示在未标记的脊椎动物上使用各种属性测试条件的结果。该脊椎动物最终被指派到非哺乳动物类

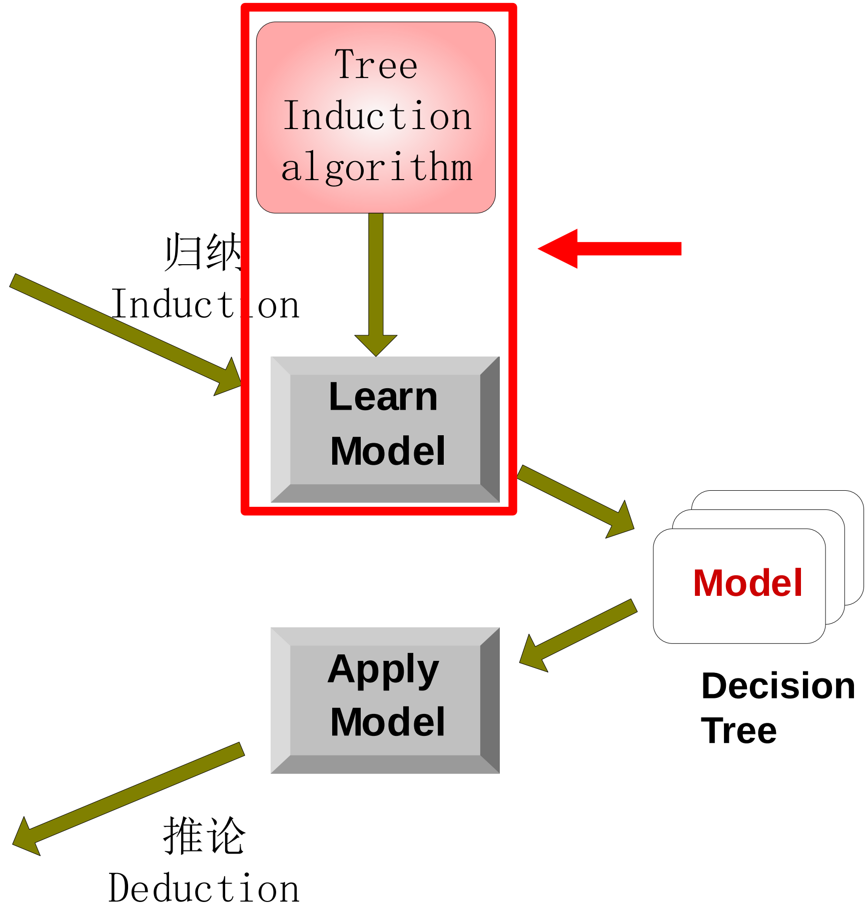
决策树分类任务

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

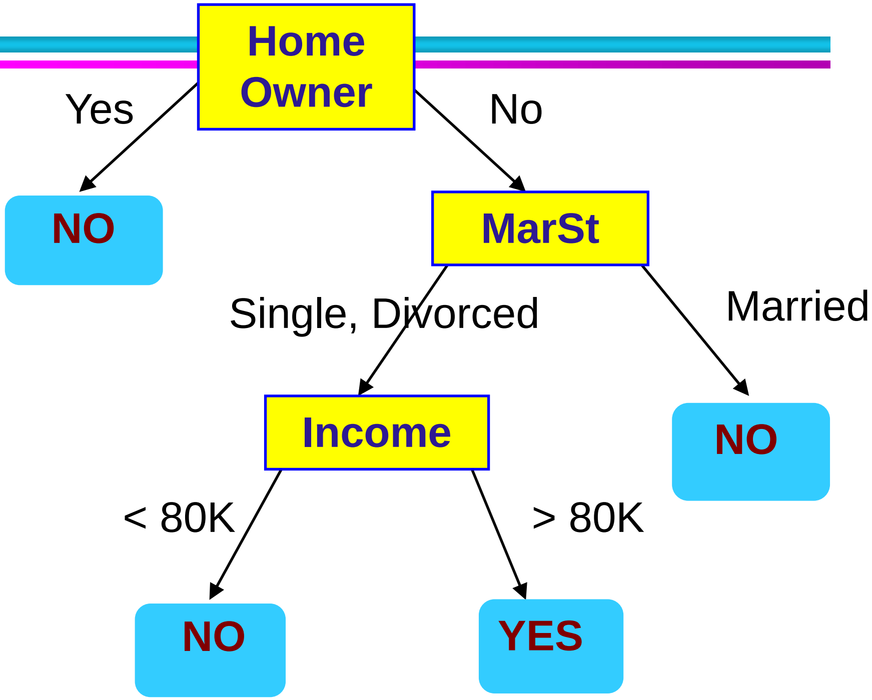
Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



决策树归纳 Decision Tree Induction

如何建立决策树?



多种算法:

– **Hunt’s Algorithm (one of the earliest)**
– CART
– ID3, C4.5
– SLIQ,SPRINT

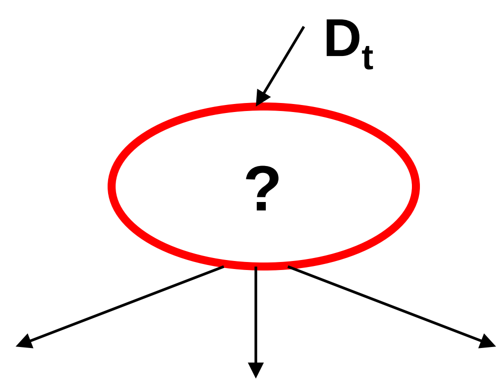
Hunt 算法的一般结构

General Structure of Hunt’s Algorithm

- D_t 表示到达节点 t 的训练集（set of training records）

- 一般过程 General Procedure:
  - 如果 D_t 只包含属于相同类别 y_t 的记录，则 t 是标记为 y_t 的叶节点
  - 如果 D_t 包含属于多个类的记录，则使用属性测试将数据拆分为较小的子集。将该过程递归地应用于每个子集。

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Hunt 算法，Hunt’s Algorithm

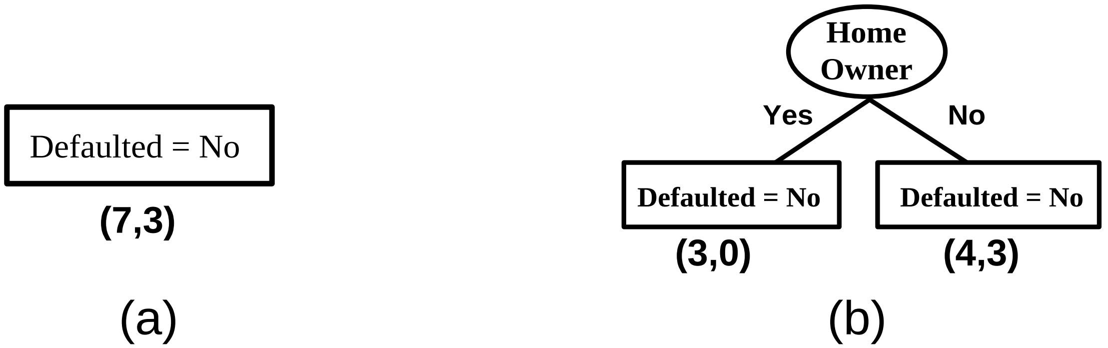
Defaulted = No

(7,3)

(a)

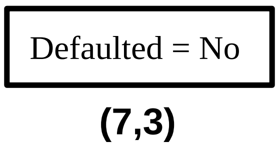
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Hunt 算法

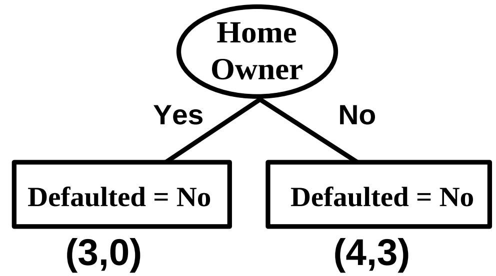


ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Hunt 算法

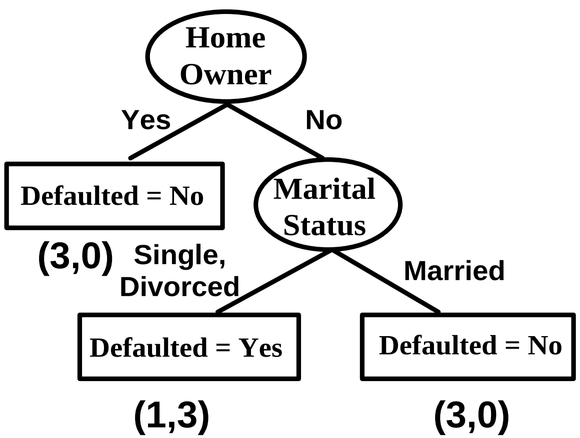


(a)



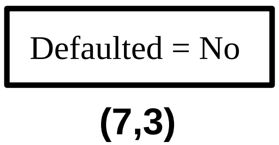
(b)

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

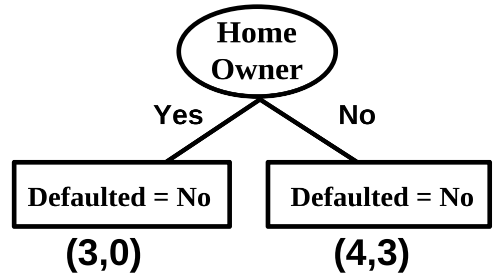


(c)

Hunt 算法

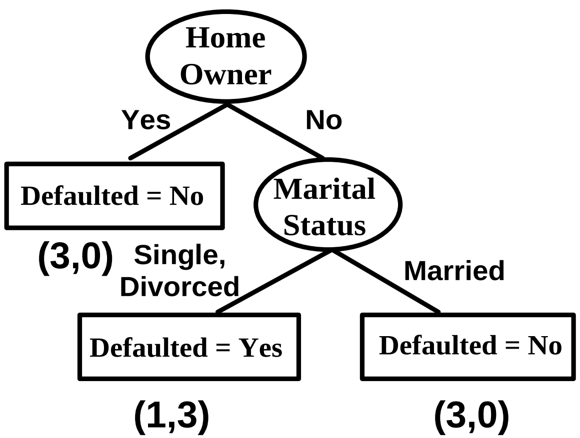


(a)

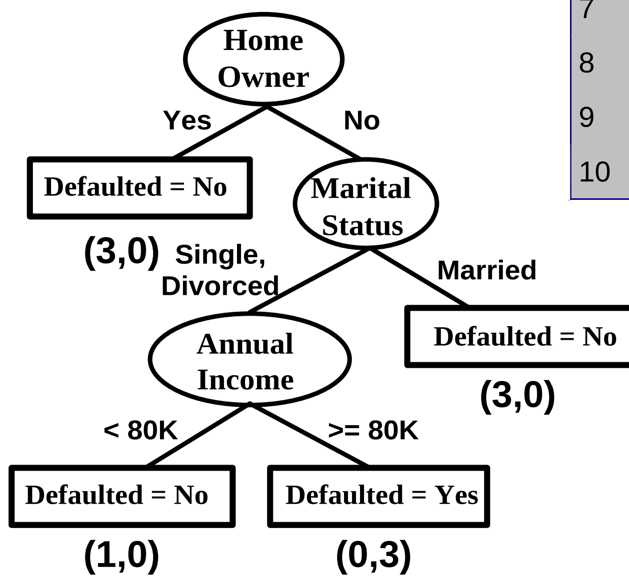


(b)

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



(c)



(d)

决策树归纳的设计问题
Design Issues of Decision Tree Induction

- 如何分裂训练记录应如何拆分?
  - 指定测试条件 (test condition) 的方法
    - 取决于属性类型 (attribute types)
  - 评估测试条件是否良好的措施

- 如何停止分裂过程 (splitting procedure)?
  - 如果所有记录属于同一类或具有相同的属性值，则停止拆分
  - 提前终止 (Early termination)

测试条件表示方法 Expressing Test Conditions

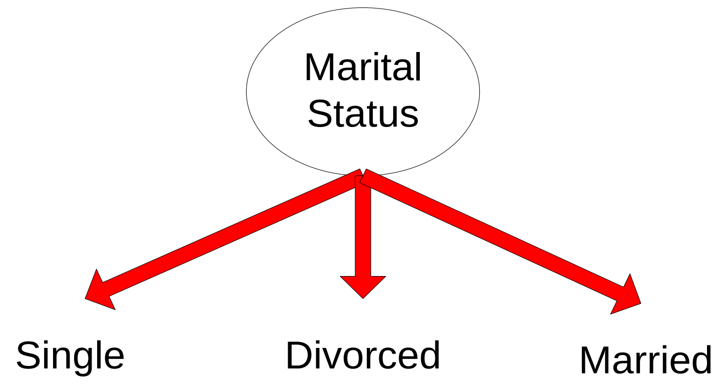
- 取决于属性类型
  - 二元 Binary
  - 标称 Nominal
  - 序数 Ordinal
  - 连续值 Continuous

- 取决于分裂个数 Depends on number of ways to split
  - 2路划分
  - 多路划分 Multi-way split

标称属性测试条件
Test Condition for Nominal Attributes

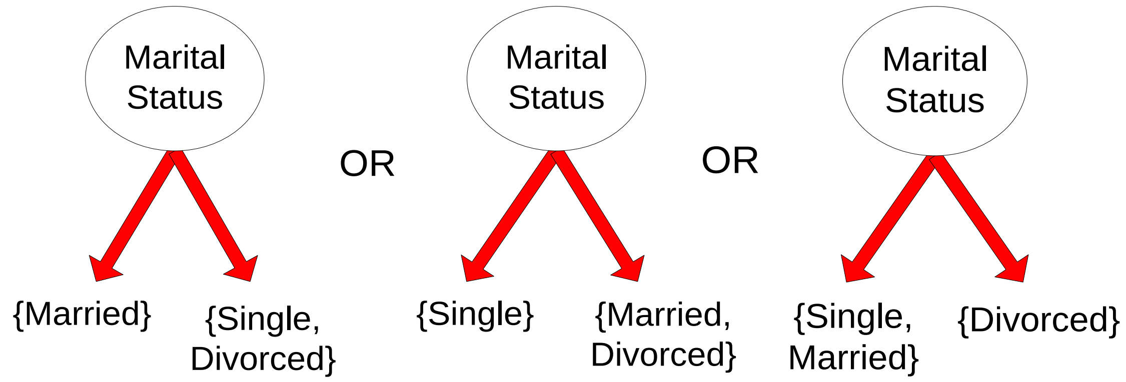
多路划分Multi-way split:

– Use as many partitions as distinct values.



二元划分 Binary split:

– Divides values into two subsets



序数属性测试条件
Test Condition for Ordinal Attributes

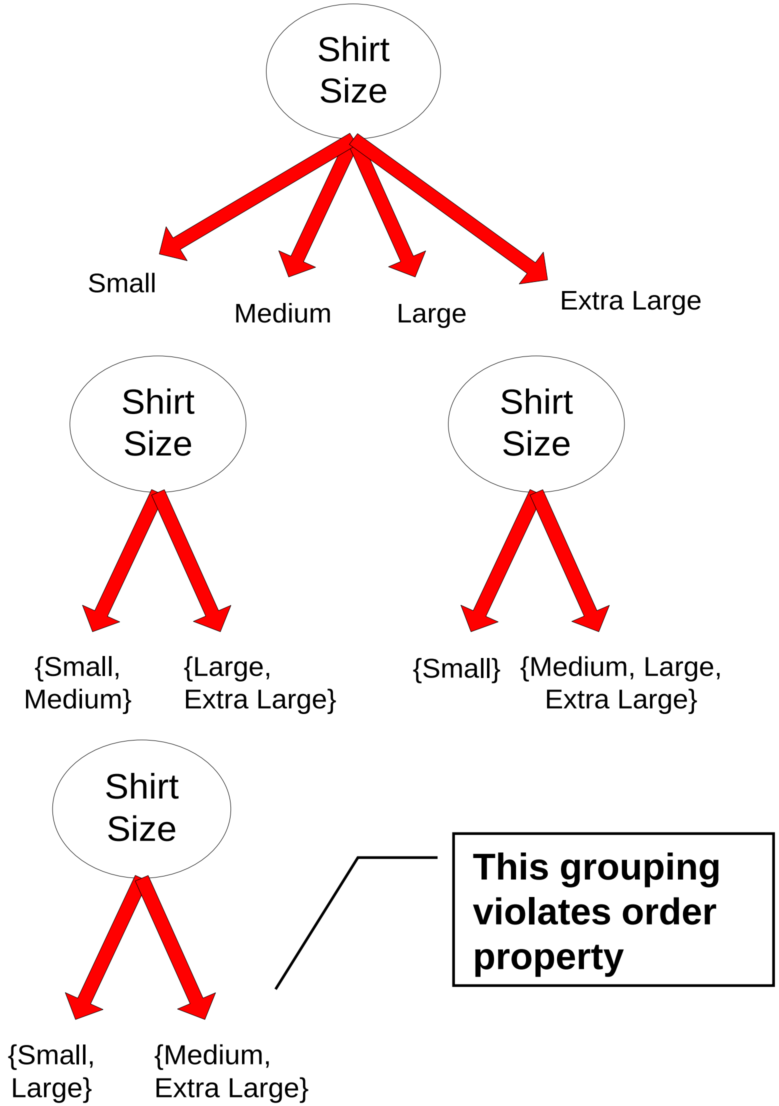
多路划分Multi-way split:

– Use as many partitions as distinct values

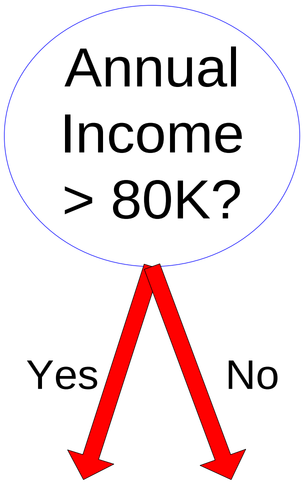
二元划分 Binary split:

– Divides values into two subsets

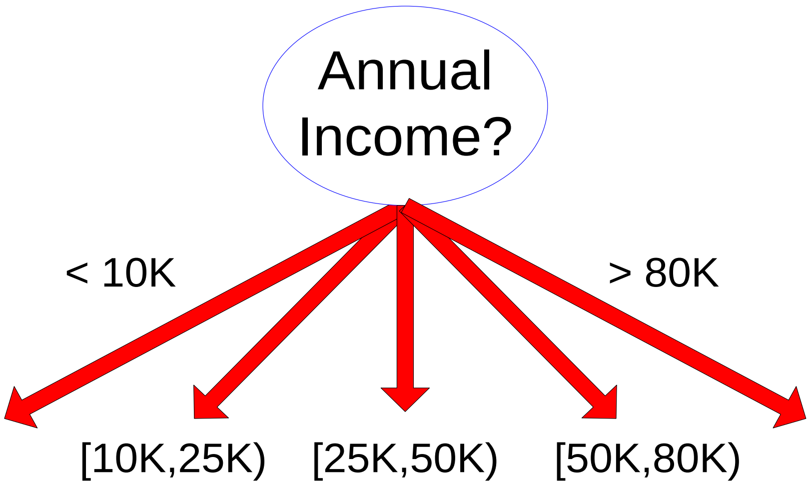
– Preserve order property among attribute values



连续属性测试条件
Test Condition for Continuous Attributes



(i) Binary split



(ii) Multi-way split

基于连续属性的划分
Splitting Based on Continuous Attributes

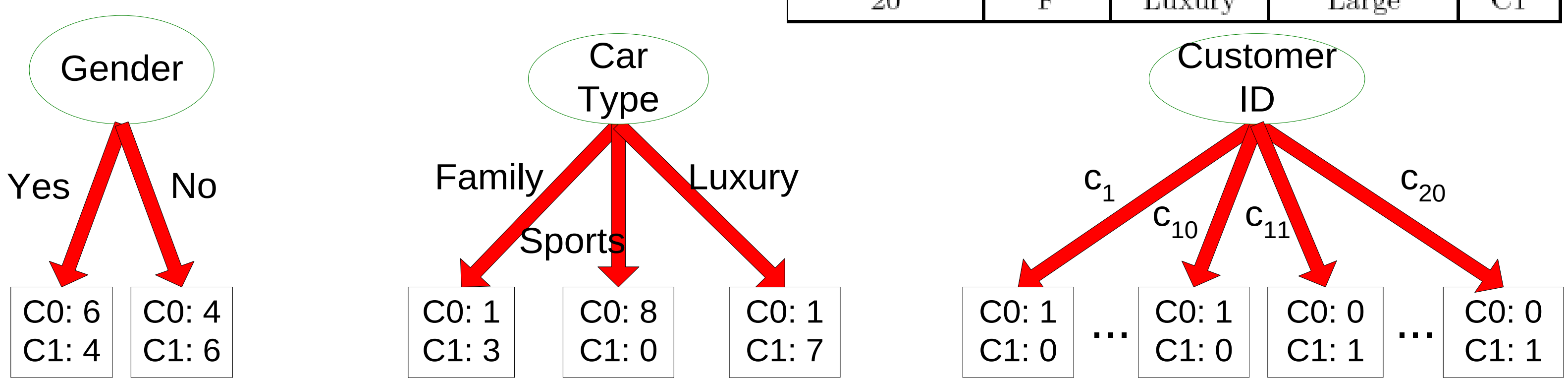
不同的处理方式

- 离散化(Discretization)以形成序数分类（ordinal categorical）属性：可以通过等间隔、等频率时段（百分位数）或聚类来找到范围。
  - 静态(Static)离散化 –一次离散化
  - 动态(Dynamic)离散化 – 在每个节点重复
- 二元决策(Binary Decision)： （A <v）或（A ≥ v)
  - 考虑所有可能的划分并找到最佳分割(best cut)
  - 一般需要更多的计算量 (more compute intensive)

选择最佳划分的度量
How to determine the Best Split

Customer Id	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

划分前: 类别 C0 有10个 records,
类别 C1 有10个 records



哪一种测试条件是最好的?

选择最佳划分的度量
How to determine the Best Split

- 贪婪方法 Greedy approach:
  - 具有更纯净（purer）类别分布的节点是首选

- 需要针对节点进行不纯度（impurity）度量:

C0: 5
C1: 5

不纯度高

High degree of impurity

C0: 9
C1: 1

不纯度低

Low degree of impurity

节点不纯度（impurity）度量

- 基尼指数 Gini Index

$$Gini\ Index = 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

其中 $\boldsymbol{p_i(t)}$ 是节点t上类别 $\boldsymbol{i}$ 的比例，$\boldsymbol{c}$ 是类别的总数

- 熵 Entropy

$$Entropy = -\sum_{i=0}^{c-1} p_i(t) log_2 p_i(t)$$

- 误分类错误 Misclassification error

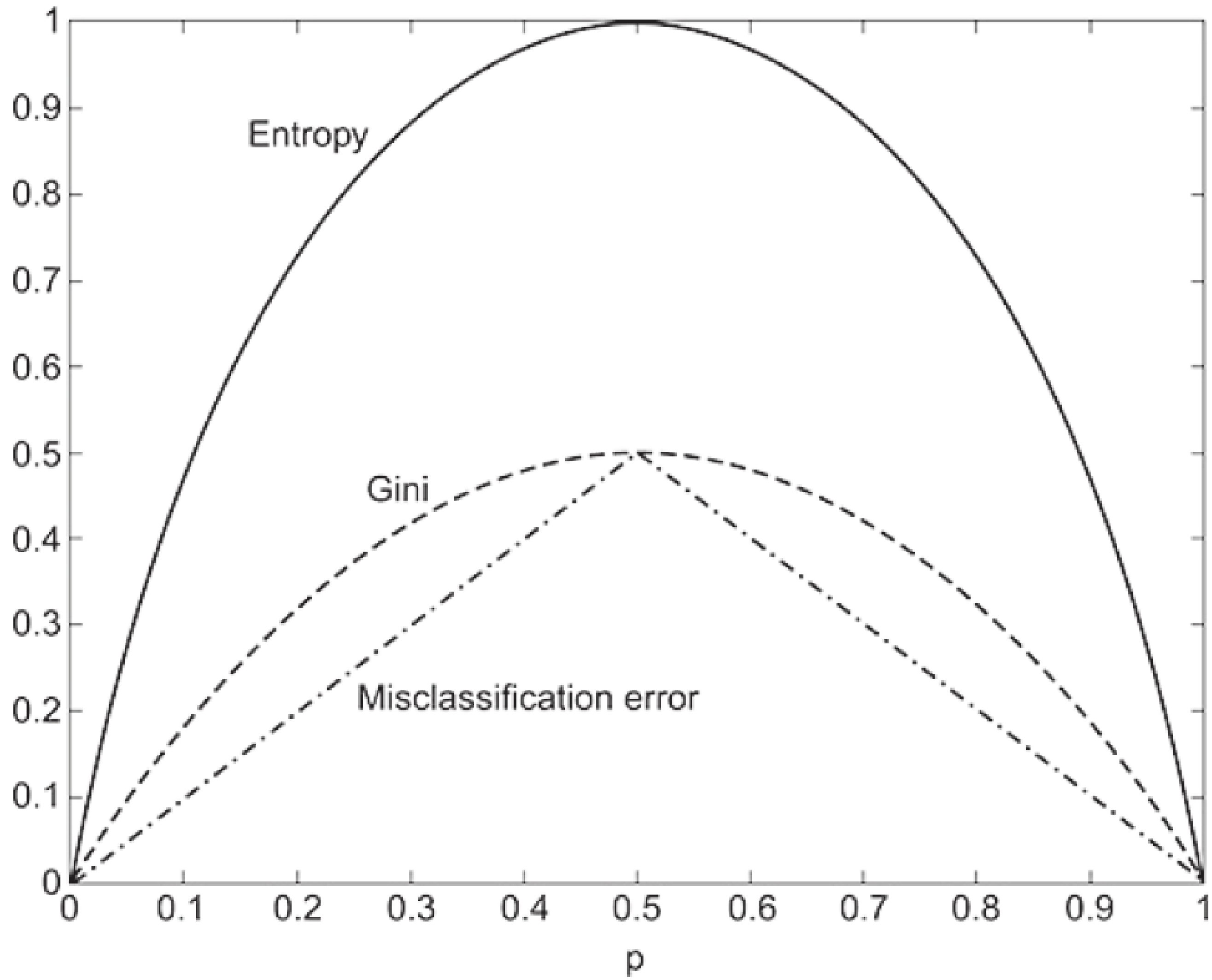
$$Classification\ error = 1 - \max[p_i(t)]$$

二元分类问题不纯性度量之间的比较

Comparison among Impurity Measures

For a 2-class problem

二分类问题



找到最佳划分 Finding the Best Split

1. 计算分裂前的不纯度（impurity）度量 P
2. 计算分裂后的不纯度（impurity）度量 M
   - Compute impurity measure of each child node
   - M is the weighted impurity of child nodes
3. 选择能够获得最大收益（gain）的测试条件

Gain = P - M

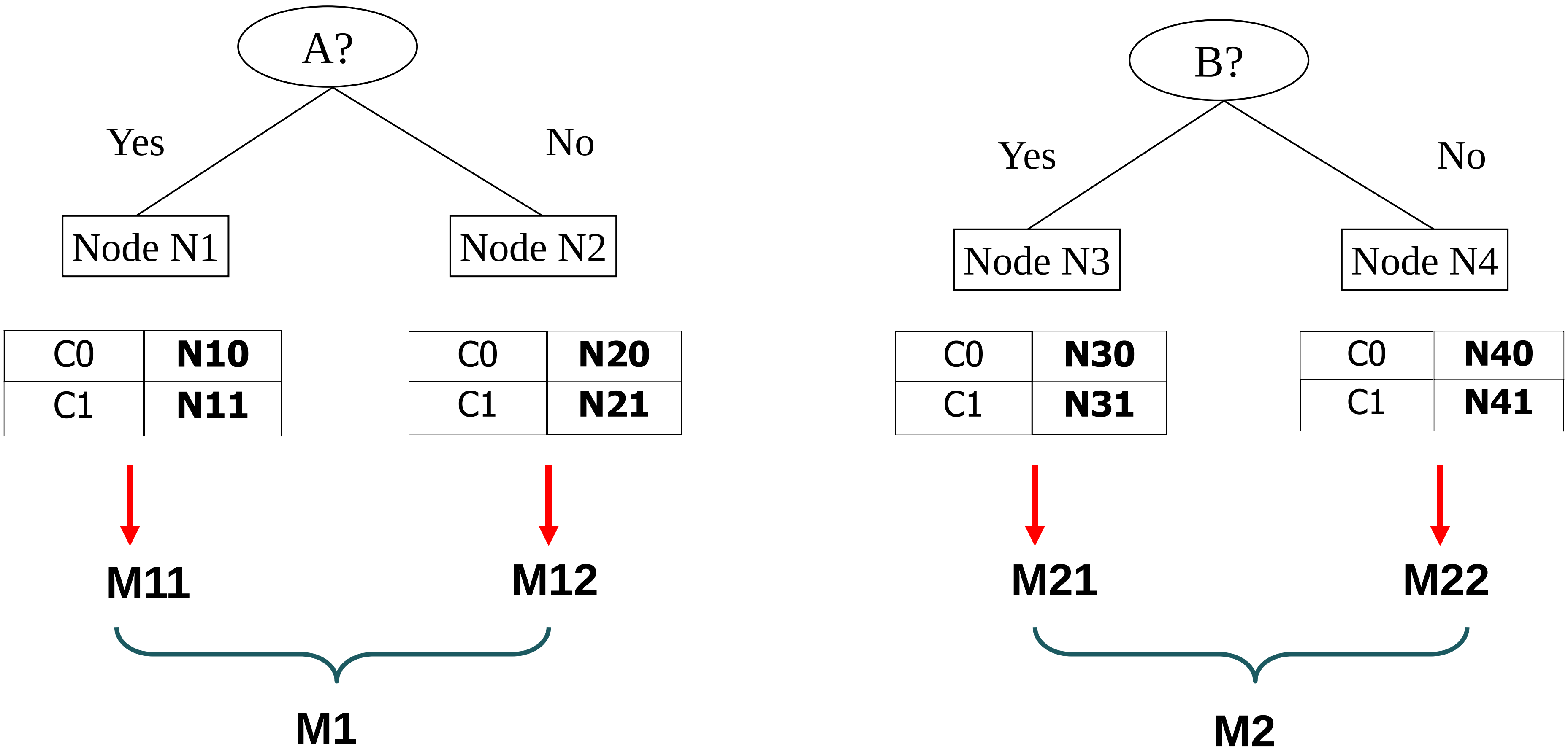
或者选择使得“分裂后不纯度M”最低的测试条件（等价）

找到最佳划分 Finding the Best Split

划分前

C0	**N00**
C1	**N01**

→ **P**



收益 Gain = P – M1 vs P – M2

Impurity 度量: GINI

Gini Index for a given node $\boldsymbol{t}$

$$Gini\ Index = 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

其中 $\boldsymbol{p_i(t)}$ 是节点t上类别 $\boldsymbol{i}$ 的比例， $\boldsymbol{c}$ 是类别的总数

– 当记录在所有类别中平均分配时，取到最大值，为1−1 $/c$，这意味着分类的最不利情况

– 当所有记录都属于一个类别时，取到最小值0，这意味着最有利于分类的情况

计算单个节点的基尼指标

Gini Index for a given node t :

$$Gini\ Index = 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

– 对于2分类问题 (p, 1 – p):
  - GINI = 1 – p^2 – (1 – p)^2 = 2p (1-p)

C1	**0**
C2	**6**

C1	**1**
C2	**5**

C1	**2**
C2	**4**

C1	**3**
C2	**3**

计算基尼指标?

计算单个节点的基尼指标

$$Gini\ Index = 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0　　P(C2) = 6/6 = 1

Gini = 1 – P(C1)^2 – P(C2)^2 = 1 – 0 – 1 = 0

C1	**1**
C2	**5**

P(C1) = 1/6　　P(C2) = 5/6

Gini = 1 – (1/6)^2 – (5/6)^2 = 0.278

C1	**2**
C2	**4**

P(C1) = 2/6　　P(C2) = 4/6

Gini = 1 – (2/6)^2 – (4/6)^2 = 0.444

右边例子的基尼指标是多少?

$$Gini\ Index = 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

C1	3
C2	3

A 0

B 0.3

C 0.5

D 1

提交

计算多个节点的基尼指标

- 当节点 p 划分为 k 个分区 partitions (子节点)

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

其中 n_i 为子节点 i 上的记录数目，
n 为父节点 p 上的记录数目。

- 选择使子节点的加权平均基尼指标最小的属性

- 基尼指标用于多种决策树算法，例如CART，SLIQ，SPRINT

二元属性的基尼指标

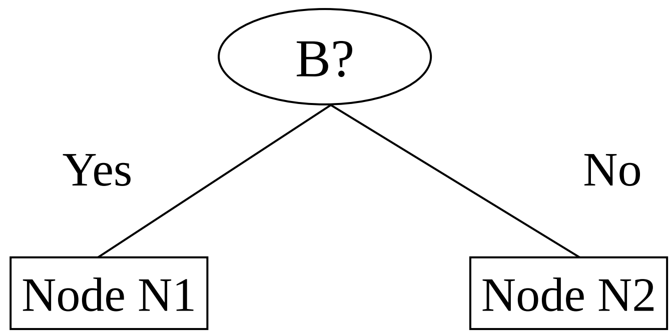
$$Gini\ Index = 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

分为两个分区（子节点）

衡量分区的效果：

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

– 寻求更高的纯度的分区（purer partitions）



	Parent
C1	**7**
C2	**5**
Gini = 0.486	

Gini(N1)
= 1 – $(5/6)^2$ – $(1/6)^2$
= 0.278

Gini(N2)
= 1 – $(2/6)^2$ – $(4/6)^2$
= 0.444

	N1	N2
C1	**5**	**2**
C2	**1**	**4**
Gini=0.361		

Weighted Gini of N1 N2
= 6/12 * 0.278 +
6/12 * 0.444
= 0.361

Gain = 0.486 – 0.361 = 0.125

类别（Categorical）属性的基尼指标

- 对于每个不同的类别属性值，获取数据集对应的每个类的计数
- 使用计数矩阵（count matrix）进行决策

Multi-way split

	CarType		
	Family	Sports	Luxury
C1	1	8	1
C2	3	0	7
Gini	0.163		

Two-way split
(find best partition of values)

	CarType	
	{Sports, Luxury}	{Family}
C1	9	1
C2	7	3
Gini	0.468	

	CarType	
	{Sports}	{Family, Luxury}
C1	8	2
C2	0	10
Gini	0.167	

哪一种是最佳分类

连续属性的基尼指标计算

- 根据一个值使用二元决策
- 属性值划分有多种选择
  - 可能的划分值数量=不同值的数量
- 每个划分值都有一个与之关联的计数矩阵
  - 每种划分中的类数，A≤v和A> v
- 选择最佳候选划分点 v 的简单方法
  - 对于每个v，扫描数据库以获取计数矩阵并计算其基尼指标
  - 该方法计算效率低下！ 重复计算。

ID	Home Owner	Marital Status	Annual Income	Defaulted
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Annual Income ?

	≤ 80	> 80
Defaulted Yes	0	3
Defaulted No	3	4

连续属性的基尼指标计算

- 为了提高计算效率：对于每个属性，
  - 按值对属性进行排序
  - 线性扫描这些值，每次更新计数矩阵并计算基尼指标
  - 选择基尼指标最小的分割位置

	Cheat	No	No	No	Yes	Yes	Yes	No	No	No	No
		Annual Income									
Sorted Values →		60	70	75	85	90	95	100	120	125	220

连续属性的基尼指标计算

- 为了提高计算效率：对于每个属性，
  - 按值对属性进行排序
  - 线性扫描这些值，每次更新计数矩阵并计算基尼指标
  - 选择基尼指标最小的分割位置

Cheat	No	No	No	Yes	Yes	Yes	No	No	No	No	
	Annual Income										
Sorted Values	60	70	75	85	90	95	100	120	125	220	
Split Positions	55	65	72	80	87	92	97	110	122	172	230
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>

连续属性的基尼指标计算

- 为了提高计算效率：对于每个属性，
  - 按值对属性进行排序
  - 线性扫描这些值，每次更新计数矩阵并计算基尼指标
  - 选择基尼指标最小的分割位置

Cheat	No	No	No	Yes	Yes	Yes	No	No	No	No
Annual Income										
Sorted Values	60	70	75	85	90	95	100	120	125	220

Split Positions	55		65		72		80		87		92		97		110		122		172		230	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Yes	0	3	0	3	0	3	0	3														
No	0	7	1	6	2	5	3	4														
Gini	0.420		0.400		0.375		0.343															

连续属性的基尼指标计算

- 为了提高计算效率：对于每个属性，
  - 按值对属性进行排序
  - 线性扫描这些值，每次更新计数矩阵并计算基尼指标
  - 选择基尼指标最小的分割位置

Cheat	No	No	No	Yes	Yes	Yes	No	No	No	No
	Annual Income									
Sorted Values	60	70	75	85	90	95	100	120	125	220

Split Positions	55		65		72		80		87		92		97		110		122		172		230	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Yes	0	3	0	3	0	3	0	3	1	2												
No	0	7	1	6	2	5	3	4	3	4												
Gini	0.420		0.400		0.375		0.343		0.417													

连续属性的基尼指标计算

- 为了提高计算效率：对于每个属性，
  - 按值对属性进行排序
  - 线性扫描这些值，每次更新计数矩阵并计算基尼指标
  - 选择基尼指标最小的分割位置

Cheat	No	No	No	Yes	Yes	Yes	No	No	No	No
	Annual Income									
Sorted Values	60	70	75	85	90	95	100	120	125	220

Split Positions	55		65		72		80		87		92		97		110		122		172		230	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Yes	0	3	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0	3	0
No	0	7	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1	7	0
Gini	0.420		0.400		0.375		0.343		0.417		0.400		***0.300***		0.343		0.375		0.400		0.420	

不纯度度量：熵 Measure of Impurity: Entropy

- 给定节点 t 的熵为：

$$Entropy = -\sum_{i=0}^{c-1} p_i(t) log_2 p_i(t)$$

其中 $\boldsymbol{p_i(t)}$ 是节点t上类别 $\boldsymbol{i}$ 的比例， $\boldsymbol{c}$ 是类别的总数

- 当记录在所有类中平均分配时， $\log_2 c$ 取到最大值，这意味着分类的最不利情况
- 当所有记录都属于一个类别时，取到最小值0，这意味着最有利于分类的情况

– 基于熵的计算与GINI系数计算非常相似

计算单个节点的熵

$$Entropy = -\sum_{i=0}^{c-1} p_i(t) log_2 p_i(t)$$

C1	**0**
C2	**6**

熵分别为?

C1	**1**
C2	**5**

C1	**2**
C2	**4**

计算划分后的信息增益 Information Gain

- Information Gain:

$$Gain_{split} = Entropy(p) - \sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)$$

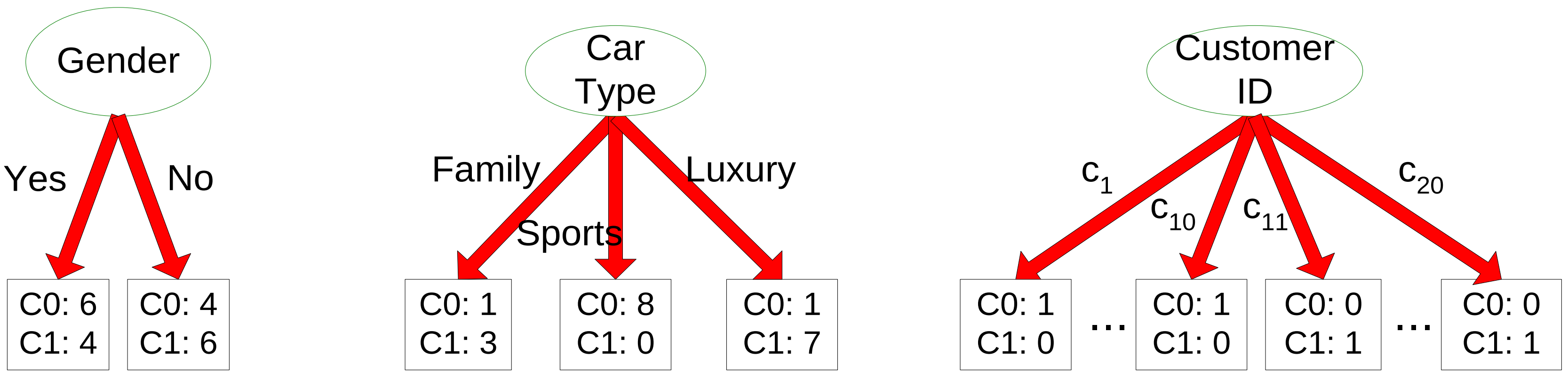
父节点 p 划分为 k 个分区 (children)

n_i 是子节点 i 中的记录数

  - 选择可获得最大减少量的划分（最大化增益)
  - 在ID3 和 C4.5 等决策树算法中使用
  - 信息增益是类别变量和划分变量（splitting variable）之间的互信息（mutual information)

Problem with large number of partitions

节点不纯度度量倾向于产生大量的分区，每个分区很小，但是纯度很高。



– 例如，顾客ID具有最高的信息增益，因为根据该属性的划分得到的所有子节点的熵均为0

增益率 Gain Ratio

- Gain Ratio:

$$Gain\ Ratio = \frac{Gain_{split}}{Split\ Info} \qquad Split\ Info = -\sum_{i=1}^{k}\frac{n_i}{n}log_2\frac{n_i}{n}$$

父节点 p 划分为 k 个分区 (children)

n_i 是子节点 i 中的记录数

  – 通过分区的熵调整信息增益($Split\ Info$).
    - 较高的熵分区（大量的小分区）会受到惩罚!
  – Used in C4.5 algorithm
  – 旨在克服信息增益的缺点

增益率 Gain Ratio

- Gain Ratio:

$$Gain\ Ratio = \frac{Gain_{split}}{Split\ Info} \qquad Split\ Info = \sum_{i=1}^{k} \frac{n_i}{n} log_2 \frac{n_i}{n}$$

父节点 p 划分为 k 个分区 (children)

n_i 是子节点 i 中的记录数

	CarType		
	Family	Sports	Luxury
C1	1	8	1
C2	3	0	7
Gini	**0.163**		

SplitINFO = 1.52

	CarType	
	{Sports, Luxury}	{Family}
C1	9	1
C2	7	3
Gini	**0.468**	

SplitINFO = 0.72

	CarType	
	{Sports}	{Family, Luxury}
C1	8	2
C2	0	10
Gini	**0.167**	

SplitINFO = 0.97

不纯度度量：分类错误 Classification Error

- 节点 t 的分类错误

$$Error(t) = 1 - \max_{i}[p_i(t)]$$

  - 当记录在所有类别之间平均分配时，最大值为1−1/c，这意味着最无趣的情况
  - 当所有记录都属于一个类别时，最小值为0，这表示我们最感兴趣的情况

单个节点的错误率

$$Error(t) = 1 - \max_i [p_i(t)]$$

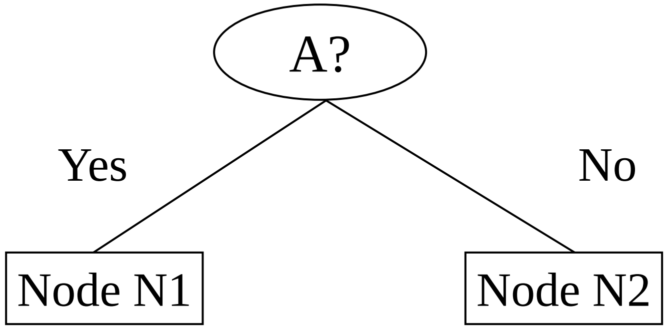
C1	**0**
C2	**6**

错误率分别为?

C1	**1**
C2	**5**

C1	**2**
C2	**4**

Misclassification Error vs Gini Index



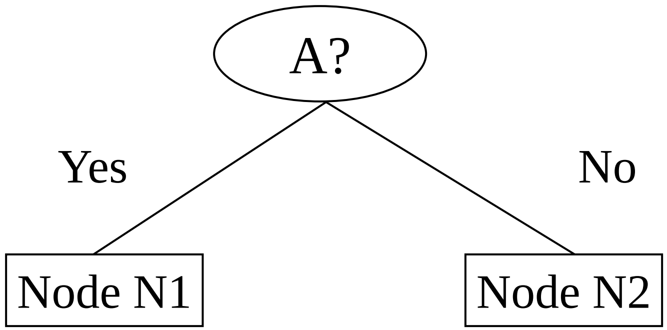
	Parent
C1	**7**
C2	**3**
Gini = 0.42	

	N1	N2
C1	**3**	**4**
C2	**0**	**3**
Gini=0.342		

	N1	N2
C1	**3**	**4**
C2	**1**	**2**
Gini=0.416		

Misclassification error for all three cases = 0.3 !

Misclassification Error vs Gini Index



	Parent
C1	7
C2	3
Gini = 0.42	

Gini(N1)
$= 1 - (3/3)^2 - (0/3)^2$
$= 0$

Gini(N2)
$= 1 - (4/7)^2 - (3/7)^2$
$= 0.489$

	N1	N2
C1	3	4
C2	0	3
Gini=0.342		

Gini(Children)
$= 3/10 * 0$
$+ 7/10 * 0.489$
$= 0.342$

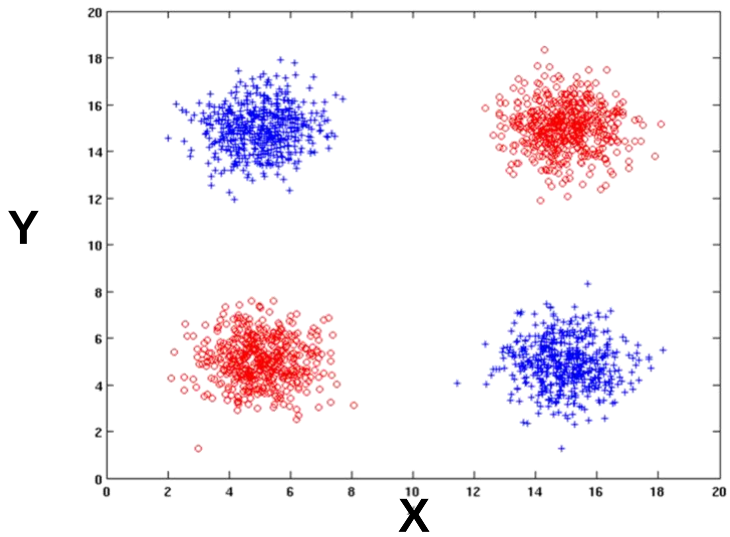
经过上述划分，基尼指数降低了，但是错误率没有发生变化！

基于决策树的分类

Decision Tree Based Classification

- 优点:
  - 构造成本低
  - 对未知记录进行分类的速度非常快
  - 小规模决策树解释性强
  - 强大的抗噪能力（尤其是在采用避免过度拟合的方法时）
  - 可以轻松处理冗余或不相关的属性（除非属性存在交互）
- 缺点:
  - 可能的决策树的空间是指数级别的。 因此无法通过遍历找到最优解，所采用的贪婪的方法通常无法找到最好的树。
  - 无法考虑属性之间的交互
  - 每个决策边界仅涉及一个属性

处理属性存在交互的情况 Handling interactions

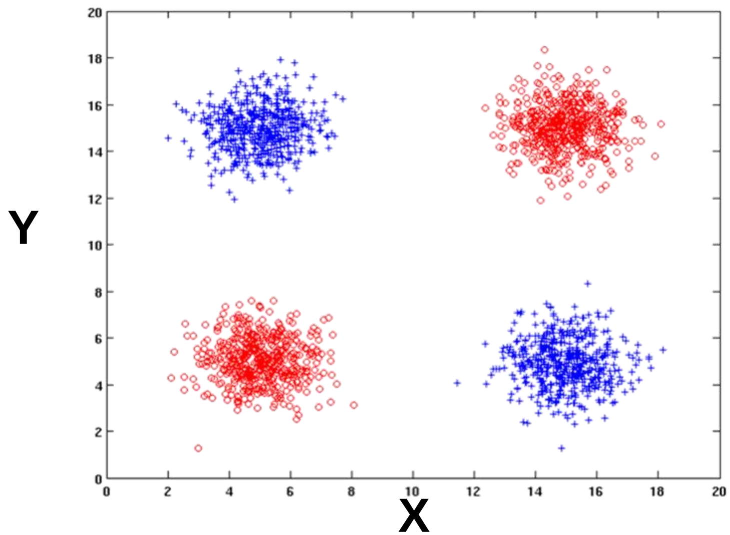


+ : 1000 instances

o : 1000 instances

Entropy (X) : 0.99
Entropy (Y) : 0.99

处理属性存在交互的情况 Handling interactions



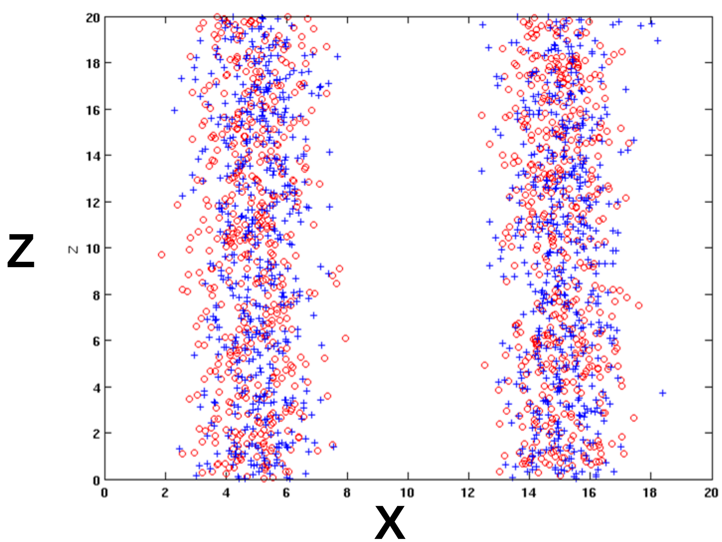
+ : 1000 instances

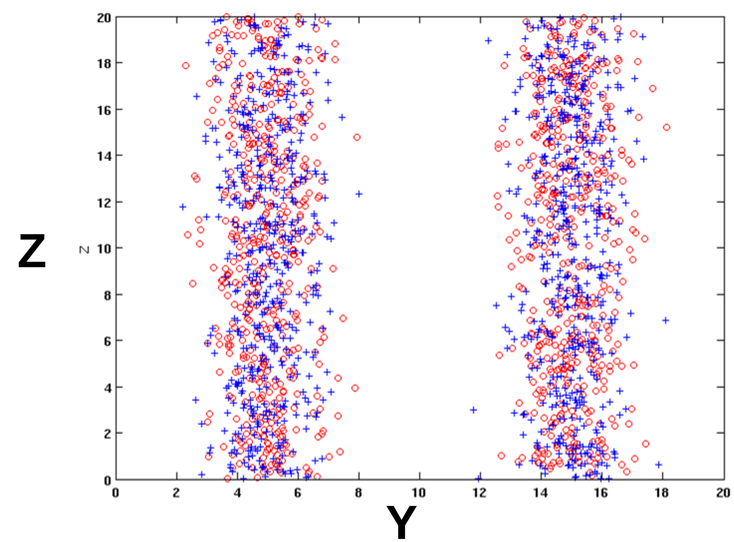
o : 1000 instances

将Z添加为从均匀分布（uniform distribution）生成的噪声属性

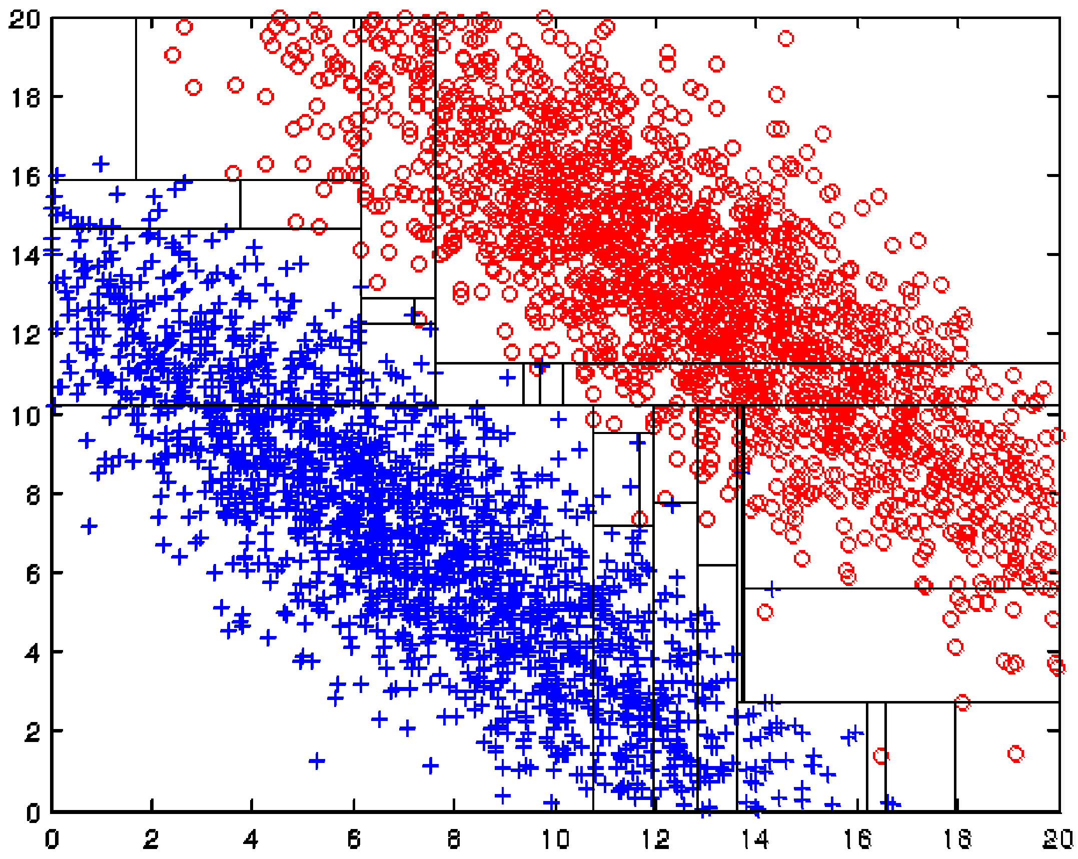
Entropy (X) : 0.99
Entropy (Y) : 0.99
Entropy (Z) : 0.98

属性Z将成为用于划分（splitting）的属性





Limitations of single attribute-based decision boundaries



Both positive (+) and negative (o) classes generated from skewed Gaussians with centers at (8,8) and (12,12) respectively.

谢谢！

数据挖掘

教师：王东京
学院：计算机学院
邮箱：dongjing.wang@hdu.edu.cn