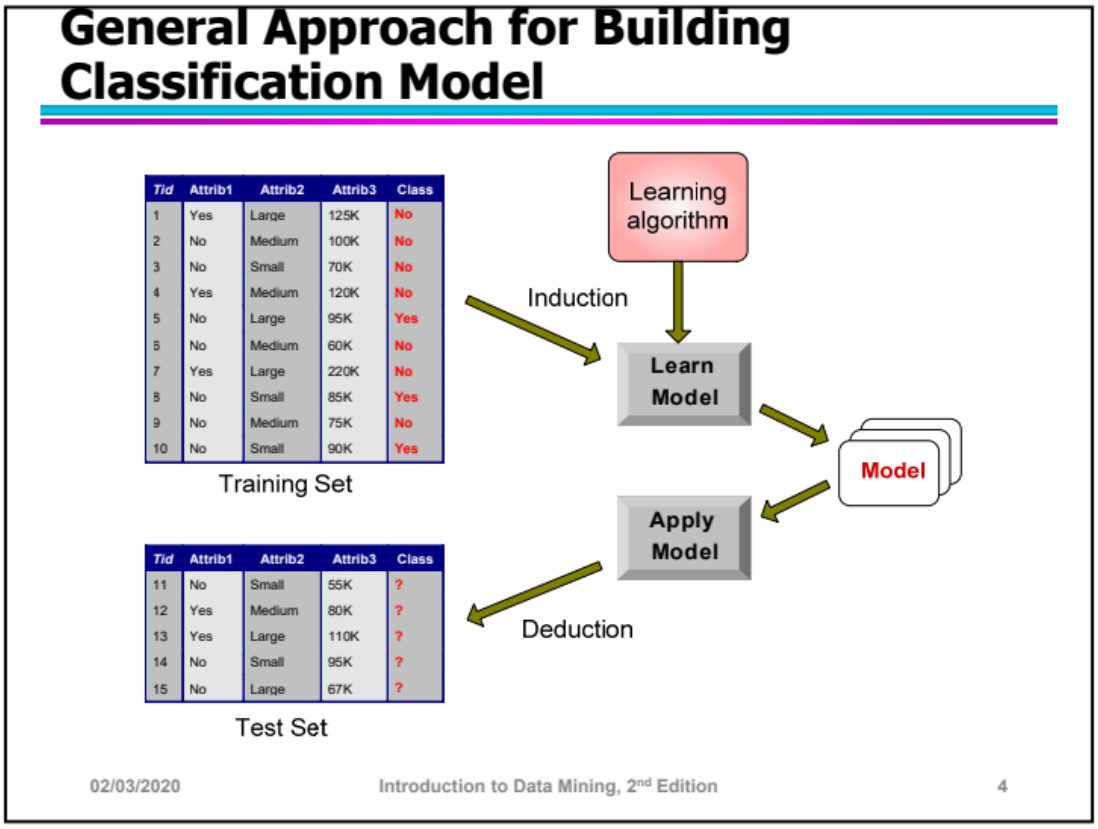
数据挖掘技术
分类:基本概念和技术
第三章的课堂笔记
数据挖掘导论，第二版
经过
谭、斯坦贝克、卡帕特内、库马尔
02/03/2020 数据挖掘导论，第 2 版 1

1

分类:定义
给定一组记录(训练集)
每个记录由一个元组(x，y)表征，其中 x 是属性集，y 是类标签
x:属性，预测，独立变量，输入⊠ y:类，响应，因变量，输出
任务:
学习将每个属性集 x 映射到一个预定义的类标签 y 的模型
02/03/2020 数据挖掘导论，第 2 版 2

分类任务示例
任务属性集，x 类标签，y
电子邮件分类
信息
从电子邮件标题和内容中提取的特征
垃圾邮件或非垃圾邮件
识别肿瘤细胞
从 x 光或核磁共振扫描中提取的特征
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
02/03/2020 数据挖掘导论，第 2 版 3

3



4

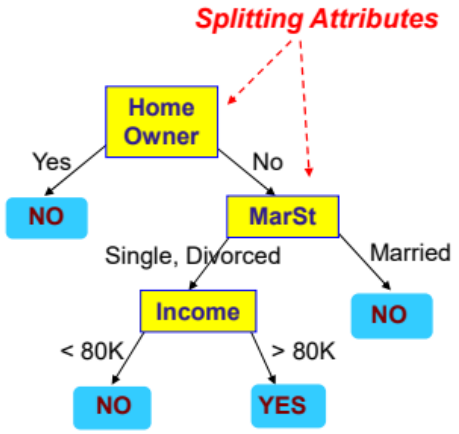
分类技术
基础分类器
基于决策树的方法基于规则的方法最近邻神经网络深度学习
朴素贝叶斯和贝叶斯信念网络支持向量机
集成分类器
助推，装袋，随机森林 02/03/2020 数据挖掘导论，第 2 版 5

5

Example of a Decision Tree

ID	Home Owner	Marital Status	Annual Income	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

Training Data

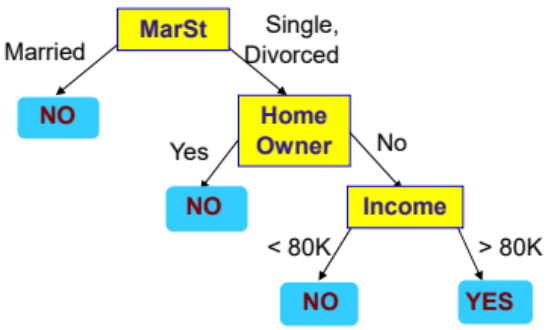


Model: Decision Tree

6

Another Example of Decision Tree

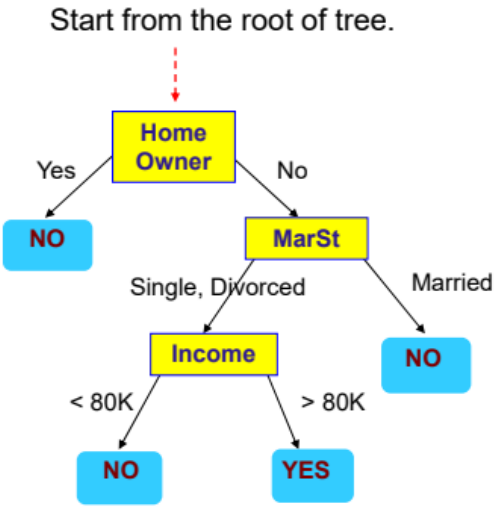
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
		categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single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



There could be more than one tree that fits the same data!

7

Apply Model to Test Data



Test Data

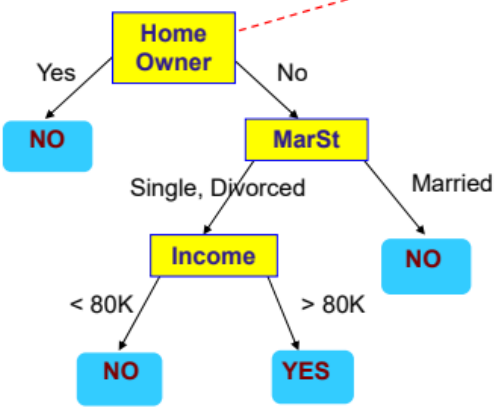
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?

8

Apply Model to Test Data

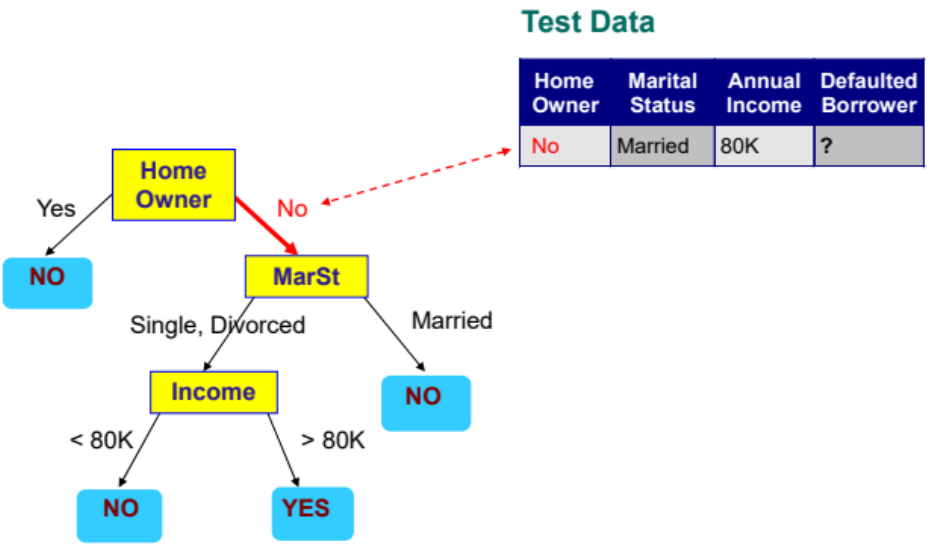
Test Data

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



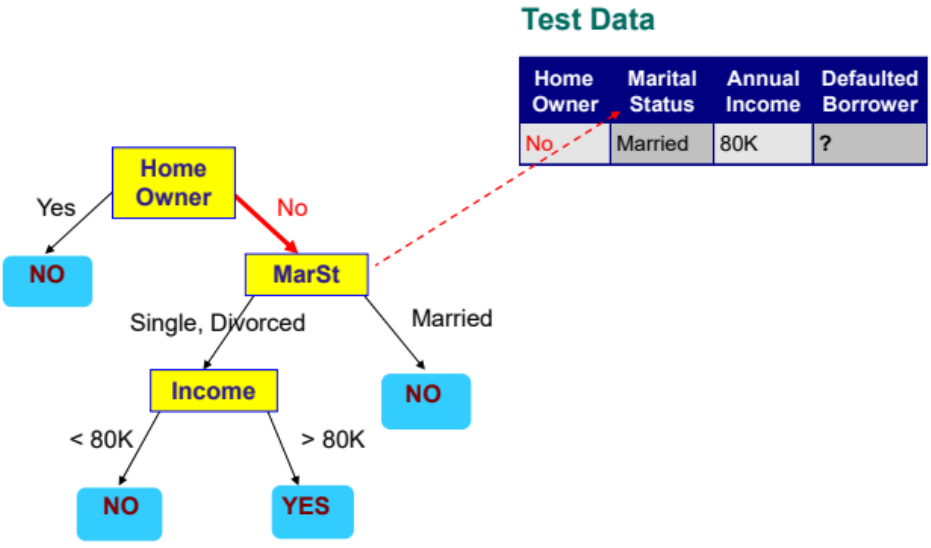
9

Apply Model to Test Data



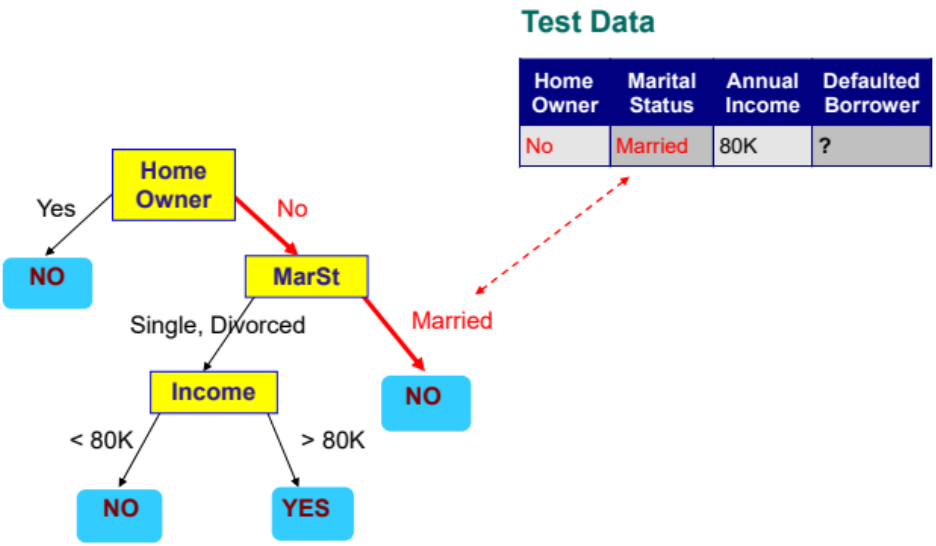
10

Apply Model to Test Data



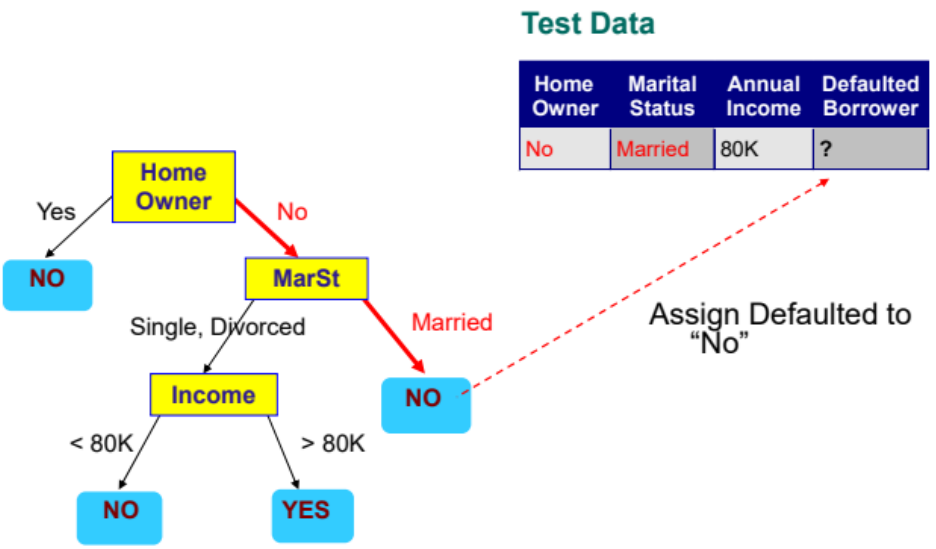
11

Apply Model to Test Data



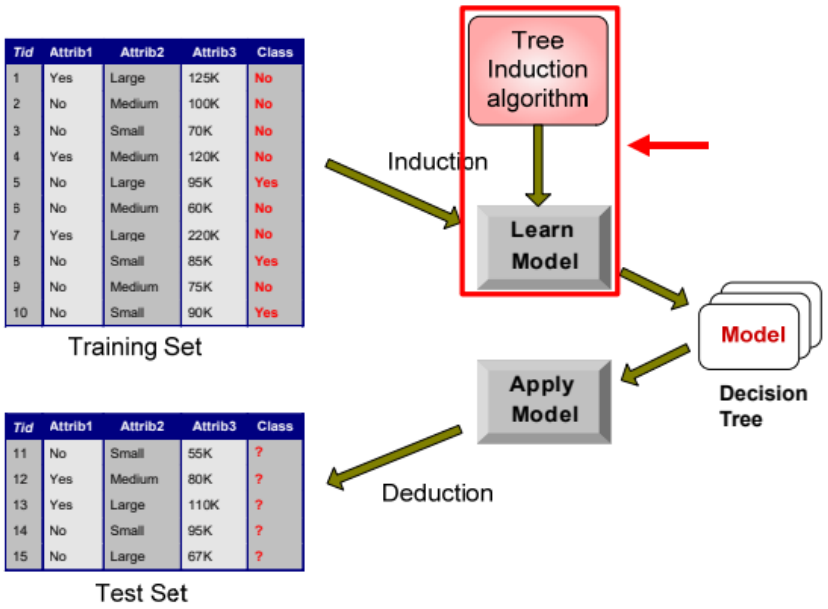
12

Apply Model to Test Data



13

Decision Tree Classification Task

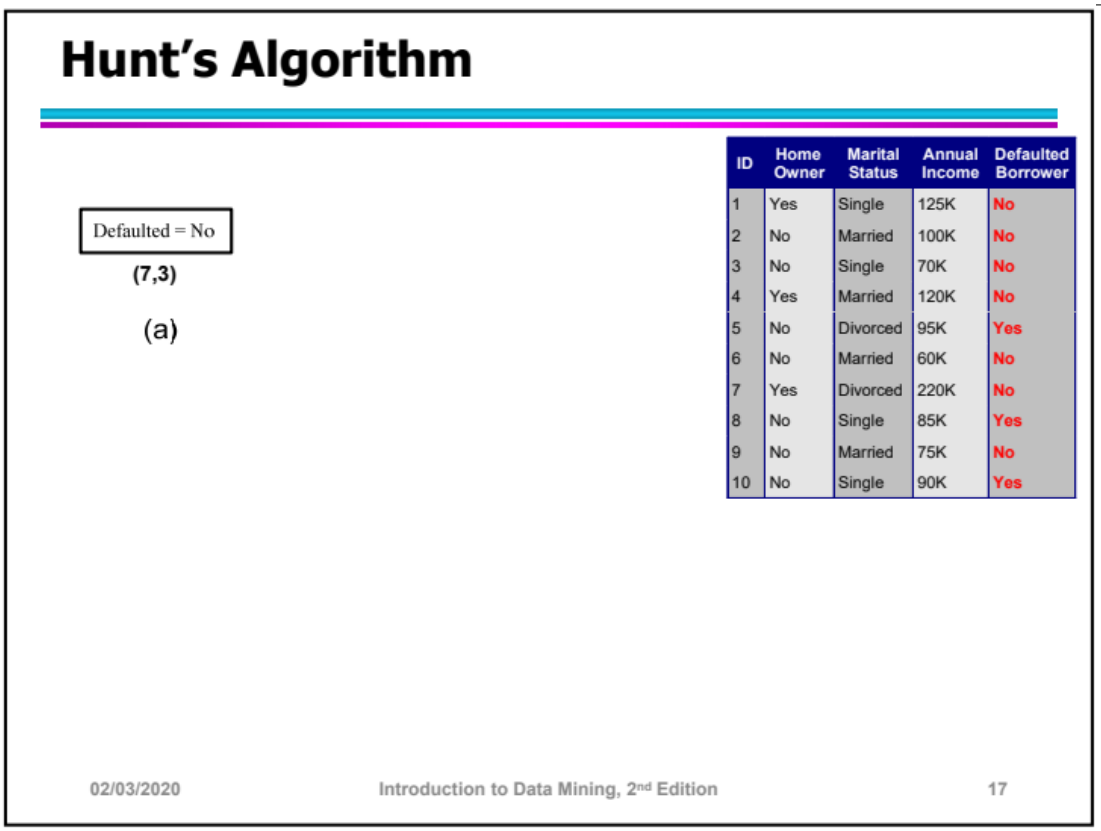


Decision Tree Induction

Many Algorithms:

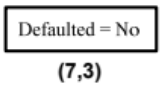
- Hunt's Algorithm (one of the earliest)
- CART
- ID3, C4.5
- SLIQ,SPRINT

亨特算法的一般结构
让 Dt 成为训练的一部分
到达节点 t 的记录
一般程序:
如果 Dt 包含属于同一类 yt 的记录，则 t 是标记为 yt 的叶节点
如果 Dt 包含属于多个类的记录，使用属性测试将数据分成更小的子集。递归地将过程应用于每个子集。
Dt
？
身份证房主婚姻状况年收入拖欠借款人 1 是单身 125，000 否 2 否已婚 100，000 否 3 否单身 70，000 否 4 是已婚 120，000 否 5 否离婚 95，000 是 6 否已婚 60，000 否 7 是离婚 220，000 否 8 否单身 85，000 是 9 否已婚 75，000 否单身 90，000 是 10
02/03/2020 数据挖掘导论，第 2 版 16



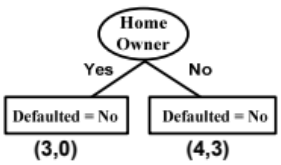
17

Hunt's Algorithm



(7,3)

(a)

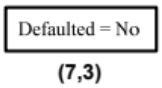


(b)

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

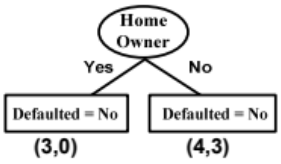
18

Hunt's Algorithm



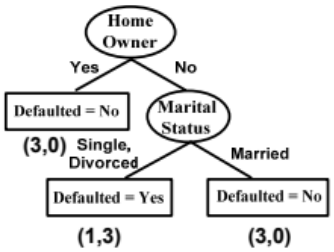
(7,3)

(a)



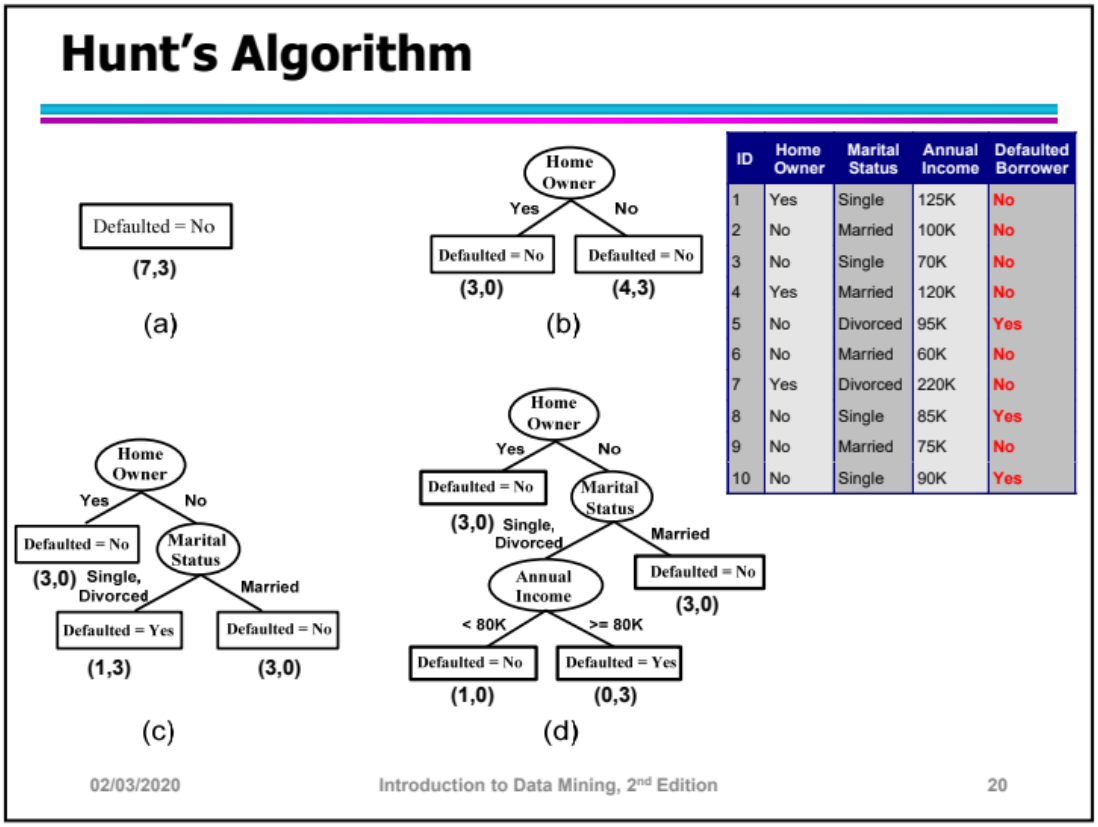
(b)

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)

19



20

决策树归纳的设计问题
培训记录应该如何分割?
指定测试条件的方法
　取决于属性类型
用于评估测试条件的良好性的措施
分裂程序应该如何停止?
如果所有记录属于同一类别或具有相同的属性值，则停止拆分提前终止
02/03/2020 数据挖掘导论，第 2 版 21

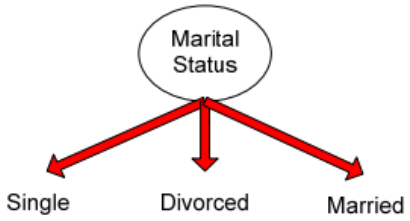
21

表达测试条件的方法
取决于二进制名义序数的属性类型
连续的
l 取决于双向分割的方式数量
02/03/2020 数据挖掘导论，第 2 版 22

Test Condition for Nominal Attributes

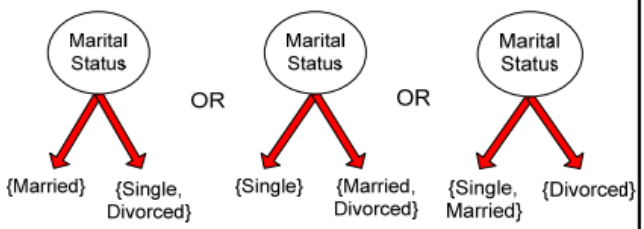
Multi-way split:

– Use as many partitions as distinct values.



Binary split:

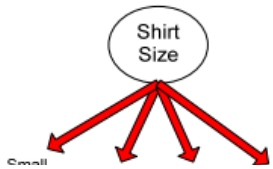
– Divides values into two subsets



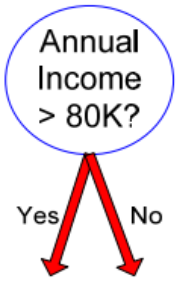
23

Test Condition for Ordinal Attributes

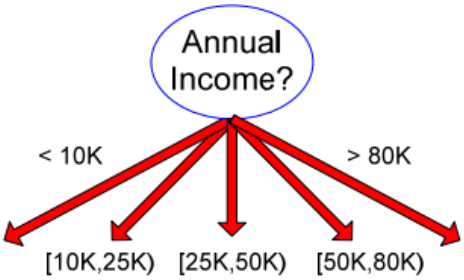
- Multi-way split:
  – Use as many partitions as distinct values



Test Condition for Continuous Attributes



(i) Binary split



(ii) Multi-way split

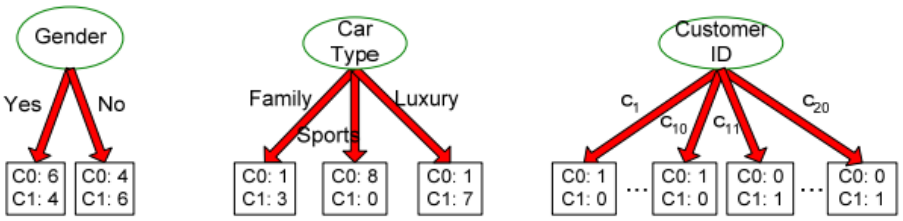
25

基于连续属性的分割
不同的处理方式
离散化以形成有序分类属性
范围可以通过等间隔时段、等频率时段(百分比)或聚类来找到。
静态-在开始时离散一次□动态-在每个节点重复
二元决策:(A < v)或(A v)
考虑了所有可能的分割，并找到了最佳分割□可以更计算密集型 02/03/2020 数据挖掘导论，第 2 版
26

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

Before Splitting: 10 records of class 0, 10 records of class 1



Which test condition is the best?

27

How to determine the Best Split

- Greedy approach:
  - Nodes with purer class distribution are preferred
- Need a measure of node impurity:

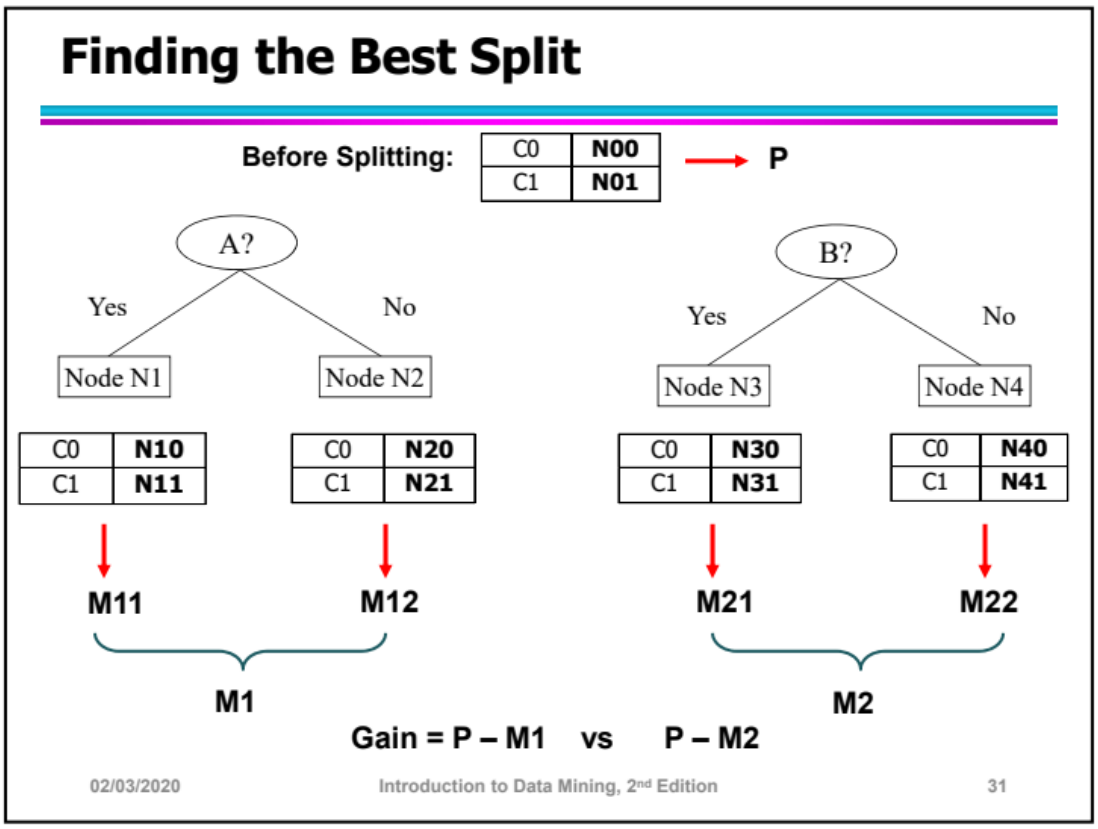
High degree of impurity　　Low degree of impurity

28

节点杂质的测量
基尼指数
熵
l 错误分类错误
02/03/2020 数据挖掘导论，第 2 版 29
$Gini\ Index = 1 - \square p\square t\square\ \square\ \square$
□□□
□□□
$Entropy = -\square p\square t\square\ log\ p\square\ \square\ \square(t)$
□□□
□□□
$Classification\ error = 1$ 最大$[p\square(t)]$
其中，$\boldsymbol{pi\ t}$是节点 t 处$\boldsymbol{i}$类的频率，$\boldsymbol{c}$是类的总数

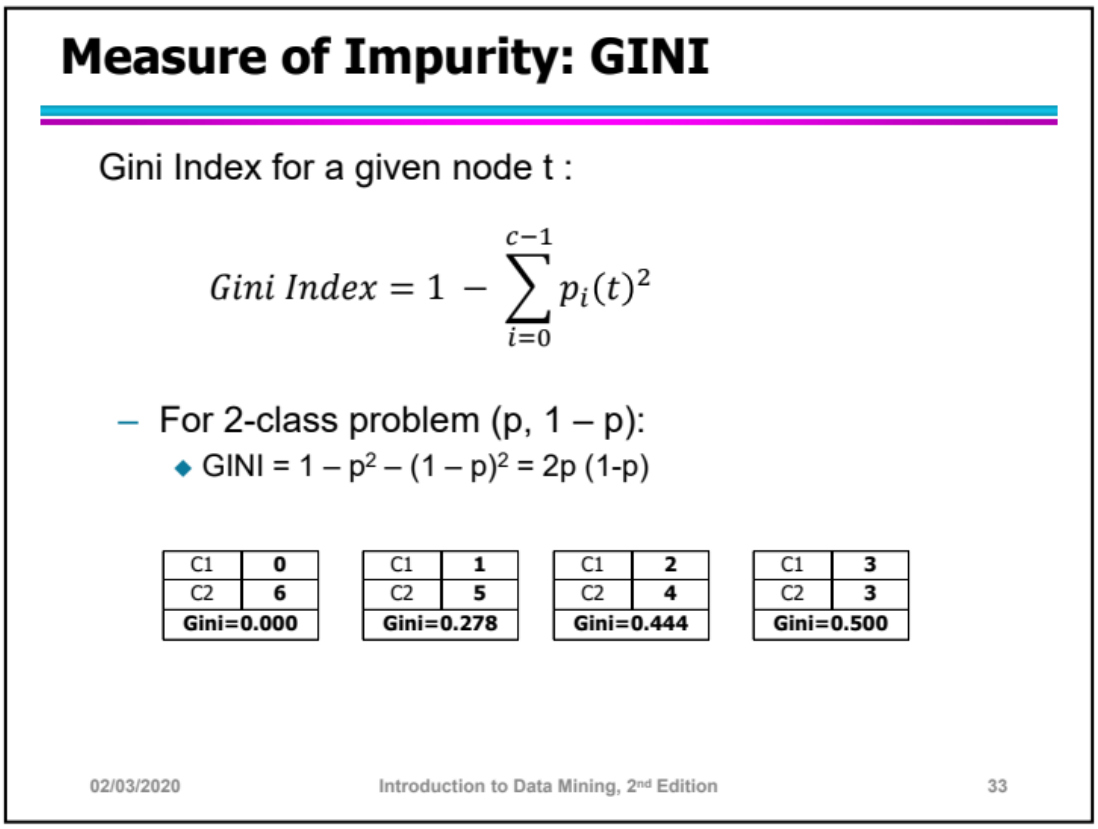
29

寻找最佳分割
1.在分离 2 之前计算杂质测量值(P)。分离后计算杂质测量值(M)
l 计算每个子节点的杂质度量 l M 是子节点的加权杂质
3.选择产生最高增益的属性测试条件
增益= P - M
或者等效地，分裂后测量最低杂质(M)
02/03/2020 数据挖掘导论，第 2 版 30



31

杂质的测量:GINI
给定节点 ☒的基尼系数
其中，p_i t是节点t的c类的频率，☒是类的总数
最大值为 1-1/c，当记录在所有类别中平均分布时，意味着对分类最不利的情况
当所有记录属于一个类别时，最小值为 0，这意味着最有利于分类
02/03/2020 数据挖掘导论，第 2 版 32
$GiniIndex = 1 - p(t)$□□
□□□
□□□



33

计算单个节点的基尼系数
C1 0
C2 6
C1 2
C2 4
C1 1
C2 5
C1 = 0/6 = 0 C2 = 6/6 = 1 基尼= 1 C1-C2 = 1-0-1 = 0
C1 = 1/6 C2 = 5/6 基尼= 1 - (1/6)- (5/6)= 0.278
C1 = 2/6 C2 = 4/6 基尼= 1 - (2/6)- (4/6)= 0.444
02/03/2020 数据挖掘导论，第 2 版 34
$GiniIndex$= 1□p[□□
□□□
□□□

计算节点集合的基尼系数
l 当节点p被分割成k分区(子节点)时
其中，n□ =儿童i的记录数，
n =父节点p.的记录数
选择最小化儿童加权平均基尼系数的属性
l 基尼指数用于决策树算法，如 CART、SLIQ、SPRINT
02/03/2020 数据挖掘导论，第 2 版 35
$GINI$□□□□□ = □n□n□GI□NI□i)
□

□□□

35

二元属性:计算 GINI 指数
分成两个分区(子节点)称重分区的效果:
寻求更大更纯的隔板
b?
是不是
节点 N1 节点 N2
父母 C1 7 C2 5 基尼= 0.486
N1·N2
C1 5 2
C2 1 4
基尼系数=0.361
基尼(N1)
= 1 - (5/6)- (1/6)= 0.278 基尼(N2)
= 1 - (2/6)- (4/6)= 0.444
N1 N2 的加权基尼系数= 6/12 * 0.278 +
6/12 * 0.444 = 0.361
增益= 0.486 - 0.361 = 0.125
02/03/2020 数据挖掘导论，第 2 版 36

分类属性:计算基尼指数
对于每个不同的值，收集数据集中每个类的计数
使用计数矩阵来做决定
卡特彼勒{体育，
豪华} {家庭} C1 9 1 C2 7 3 基尼 0.468
CarType
{体育} {家庭，奢华} C1 8 2 C2 0 10 基尼 0.167
CarType
家庭体育豪华 C1 1 8 1 C2 3 0 7 基尼 0.163
多向分裂双向分裂
(查找最佳值分区)
这些中哪一个是最好的?
02/03/2020 数据挖掘导论，第 2 版 37

37

连续属性:计算基尼指数
l 使用基于一个值的二元决策
分裂值的几种选择
可能的拆分值的数量=不同值的数量
每个拆分值都有一个与之关联的计数矩阵
每个分区中的类计数，A ≤ v 和 A > v l 选择最佳 v 的简单方法
对于每个 v，扫描数据库以收集计数矩阵并计算其基尼指数
计算效率低下！重复工作。

身份证房主婚姻状况年收入默认 1 是单身
125，000，000，000，000，000，000，000，000，000，000，000，000，000，000，000，000，000，000，000，000，000，000，000，0000，000，000，000，000，000，000，000，000，0
≤ 80 > 80
默认是 0 3
默认值 3 4
年收入?
02/03/2020 数据挖掘导论，第 2 版 38

欺骗不不不是是是是否否否否否
岁入
60 70 75 85 90 95 100 120 125 220
55 65 72 80 87 92 97 110 122 172 230
< = > < = > < = > < = > < = > < = > < = > < = > < = > < = > < = > < = >
是 0 3 0 3 0 3 0 3 1 2 1 3 0 3 0 3 0 3 0
不 0 7 1 6 2 5 3 4 3 4 3 4 4 4 3 5 2 6 1 7 0
基尼系数 0.420 0.400 0.375 0.343 0.417 0.400 0.300 0.343 0.375 0.400 0.420
连续属性:计算基尼指数...
为了高效计算:对于每个属性，根据值对属性进行排序
线性扫描这些值，每次更新计数矩阵和计算基尼指数
选择基尼系数最小的分割位置
排序值
02/03/2020 数据挖掘导论，第 2 版 39

39

欺骗不不不是是是是否否否否否
岁入
60 70 75 85 90 95 100 120 125 220
55 65 72 80 87 92 97 110 122 172 230
< = > < = > < = > < = > < = > < = > < = > < = > < = > < = > < = > < = >
是 0 3 0 3 0 3 0 3 1 2 1 3 0 3 0 3 0 3 0
不 0 7 1 6 2 5 3 4 3 4 3 4 4 4 3 5 2 6 1 7 0
基尼系数 0.420 0.400 0.375 0.343 0.417 0.400 0.300 0.343 0.375 0.400 0.420
连续属性:计算基尼指数...
为了高效计算:对于每个属性，根据值对属性进行排序
线性扫描这些值，每次更新计数矩阵和计算基尼指数
选择基尼系数最小的分割位置
分割位置
排序值
02/03/2020 数据挖掘导论，第二版 40

欺骗不不不是是是是否否否否否
岁入
60 70 75 85 90 95 100 120 125 220
55 65 72 80 87 92 97 110 122 172 230
< = > < = > < = > < = > < = > < = > < = > < = > < = > < = > < = > < = >
是 0 3 0 3 0 3 0 3 1 2 1 3 0 3 0 3 0 3 0

不 0 7 1 6 2 5 3 4 3 4 3 4 4 4 3 5 2 6 1 7 0
基尼系数 0.420 0.400 0.375 0.343 0.417 0.400 0.300 0.343 0.375 0.400 0.420
连续属性:计算基尼指数...
为了高效计算:对于每个属性，根据值对属性进行排序
线性扫描这些值，每次更新计数矩阵和计算基尼指数
选择基尼系数最小的分割位置
分割位置
排序值
02/03/2020 数据挖掘导论，第 2 版 41

41

欺骗不不不是是是是否否否否否
岁入
60 70 75 85 90 95 100 120 125 220
55 65 72 80 87 92 97 110 122 172 230
< = > < = > < = > < = > < = > < = > < = > < = > < = > < = > < = > < = > < = >
是 0 3 0 3 0 3 0 3 1 2 1 3 0 3 0 3 0 3 0
不 0 7 1 6 2 5 3 4 3 4 3 4 4 4 3 5 2 6 1 7 0
基尼系数 0.420 0.400 0.375 0.343 0.417 0.400 0.300 0.343 0.375 0.400 0.420
连续属性:计算基尼指数...
为了高效计算:对于每个属性，根据值对属性进行排序
线性扫描这些值，每次更新计数矩阵和计算基尼指数
选择基尼系数最小的分割位置
分割位置
排序值
02/03/2020 数据挖掘导论，第 2 版 42

欺骗不不不是是是是否否否否否
岁入
60 70 75 85 90 95 100 120 125 220
55 65 72 80 87 92 97 110 122 172 230
< = > < = > < = > < = > < = > < = > < = > < = > < = > < = > < = > < = > < = >
是 0 3 0 3 0 3 0 3 1 2 1 3 0 3 0 3 0 3 0
不 0 7 1 6 2 5 3 4 3 4 3 4 4 4 3 5 2 6 1 7 0
基尼系数 0.420 0.400 0.375 0.343 0.417 0.400 0.300 0.343 0.375 0.400 0.420
连续属性:计算基尼指数...
为了高效计算:对于每个属性，根据值对属性进行排序
线性扫描这些值，每次更新计数矩阵和计算基尼指数
选择基尼系数最小的分割位置
分割位置
排序值
02/03/2020 数据挖掘导论，第 2 版 43

43

杂质的度量:熵

给定节点□的熵
其中，$p(i|t)$是节点t的c类的频率，□是类的总数
log □c的□最大值，记录平均分布在所有类别中，这意味着对分类最不利的情况
当所有记录属于一个类别时，□最小值为 0，这意味着最有利于分类
基于熵的计算非常类似于 GINI 指数的计算
02/03/2020 数据挖掘导论，第 2 版 44
$Entropy = -\sum p(i|t)\log p(i|t)$
□□□
□□□

计算单个节点的熵
C1 0
C2 6
C1 2
C2 4
C1 1
C2 5
P(C1) = 0/6 = 0 P(C2) = 6/6 = 1 熵= - 0 log 0 - 1 log 1 = - 0 - 0 = 0
P(C1) = 1/6 P(C2) = 5/6
熵=-(1/6)log2(1/6)-(5/6)log2(1/6)= 0.65
P(C1) = 2/6 P(C2) = 4/6
熵=-(2/6)log2(2/6)-(4/6)log2(4/6)= 0.92
02/03/2020 数据挖掘导论，第 2 版 45
$Entropy = -\sum p(i|t)\log p(i|t)$
□□□
□□□

45

计算分裂后的信息增益
信息增益:
父节点，p被分割成k分区(子节点)n□是子节点i中的记录数
选择实现最大缩减的分割(最大化增益)
用于 ID3 和 C4.5 决策树算法
信息增益是类变量和分裂变量之间的相互信息
02/03/2020 数据挖掘导论，第 2 版 46
$Gain = Entropy(p) - \sum Entropy(i)$
□
□□□

大量分区的问题
节点杂质度量往往倾向于导致大量分区的分裂，每个分区都很小但很纯
客户标识具有最高的信息增益，因为所有子代的熵为零
02/03/2020 数据挖掘导论，第 2 版 47

47

增益比

增益比:
父节点，p被分割成k分区(子节点)n□是子节点i中的记录数
通过分区的熵来调整信息增益($Split\ Info$).
更高熵分区(大量小分区)是不利的!
用于 4.5 算法
旨在克服信息增益的缺点
02/03/2020 数据挖掘导论，第 2 版 48
$GainRatio=Gain$□□□□□$SplitInfo$ $SplitInfo=-$□nn□□□□□□
□
□□□
n□ n

Gain Ratio

- Gain Ratio:

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

Parent Node, p is split into k partitions (children)
n_i is number of records in child node i

	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

49

杂质测量:分类错误
t节点的分类错误
最大值为 $11/c$，当记录在所有类中平均分布时，意味着最不有趣的情况
当所有记录属于一个类时，最小值为 0，这意味着最有趣的情况
02/03/2020 数据挖掘导论，第二版 50
$Error\ t$ = 1 最大值
□·[·p□t]

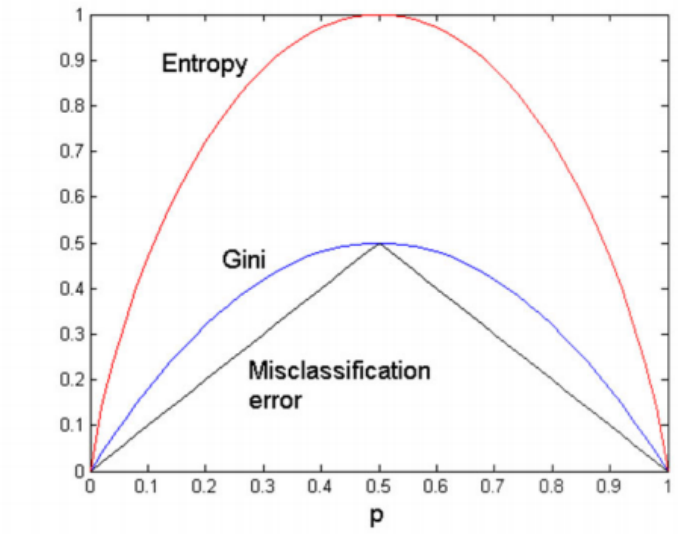
单个节点的计算误差
C1 0
C2 6

C1 2
C2 4
C1 1
C2 5
P(C1) = 0/6 = 0 P(C2) = 6/6 = 1 误差= 1 -最大值(0，1) = 1 - 1 = 0
P(C1) = 1/6 P(C2) = 5/6 误差= 1 -最大值(1/6，5/6) = 1 - 5/6 = 1/6
P(C1) = 2/6 P(C2) = 4/6 误差= 1 -最大值(2/6，4/6) = 1 - 4/6 = 1/3
02/03/2020 数据挖掘导论，第 2 版 51
$Error\ t$ = 1 最大值
□·[·p□t]

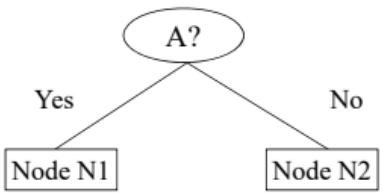
51

Comparison among Impurity Measures

For a 2-class problem:



Misclassification Error vs Gini Index



	Parent
C1	7
C2	3
Gini = 0.42	

Gini(N1)
= 1 – (3/3)² – (0/3)²
= 0

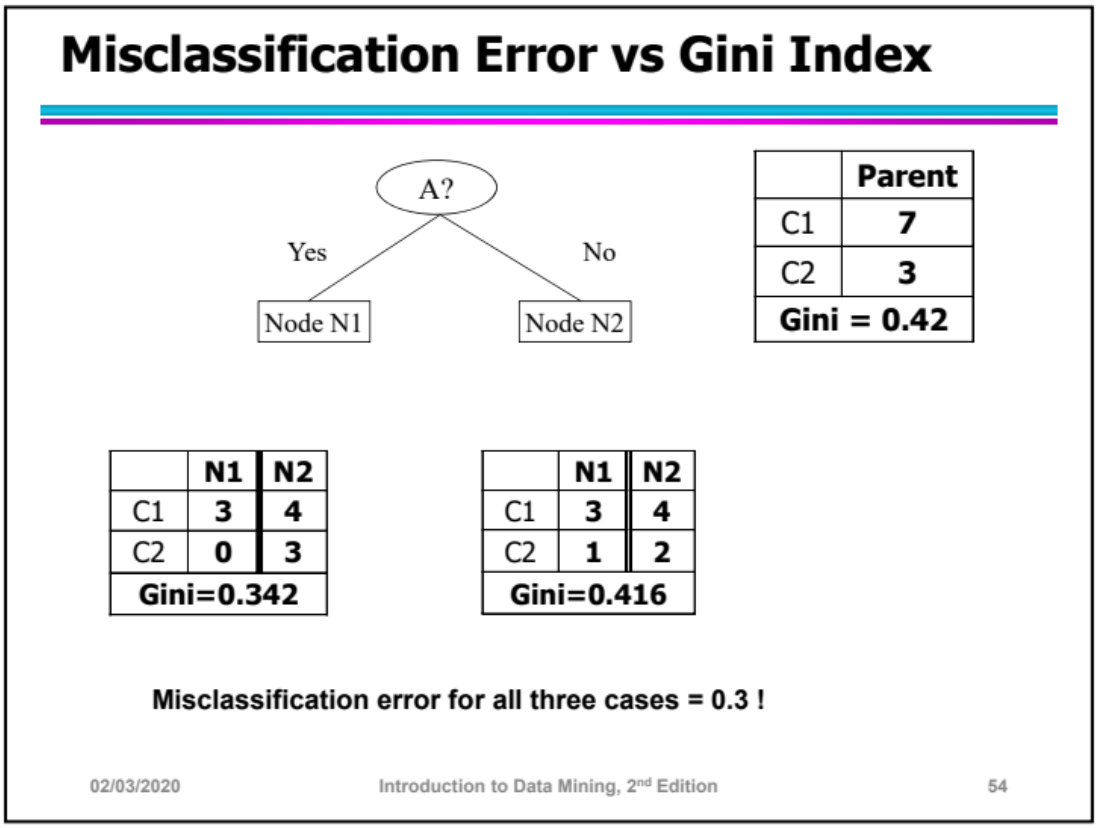
Gini(N2)
= 1 – (4/7)² – (3/7)²
= 0.489

	N1	N2
C1	3	4
C2	0	3
Gini=0.342		

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

Gini improves but error remains the same!!

53

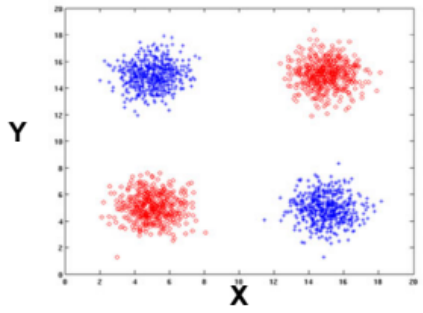


54

基于决策树的分类
优势:
建造便宜
在对未知记录进行分类方面速度极快易于对小型树进行解释对噪声具有鲁棒性(特别是在需要避免的方法时
采用过拟合)
可以轻松处理冗余或不相关的属性(除非这些属性相互作用)
缺点:
可能的决策树的空间是指数大的。贪婪的方法往往找不到最好的树。不考虑属性之间的相互作用每个决策边界只涉及一个属性
02/03/2020 数据挖掘导论，第 2 版 55

55

Handling interactions

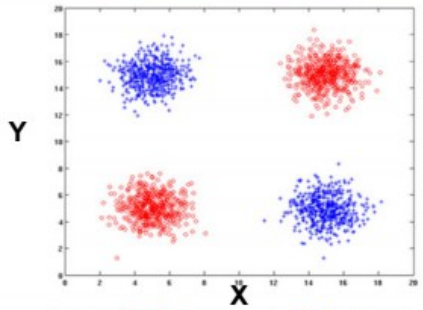


+ : 1000 instances

o : 1000 instances

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

Handling interactions



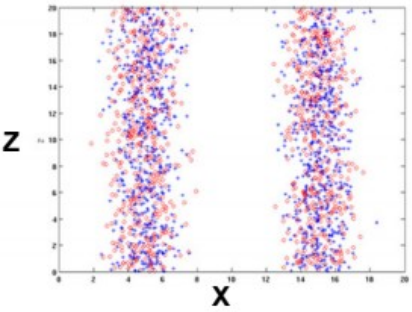
+ : 1000 instances

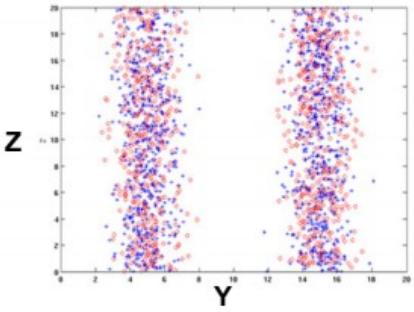
o : 1000 instances

Adding Z as a noisy attribute generated from a uniform distribution

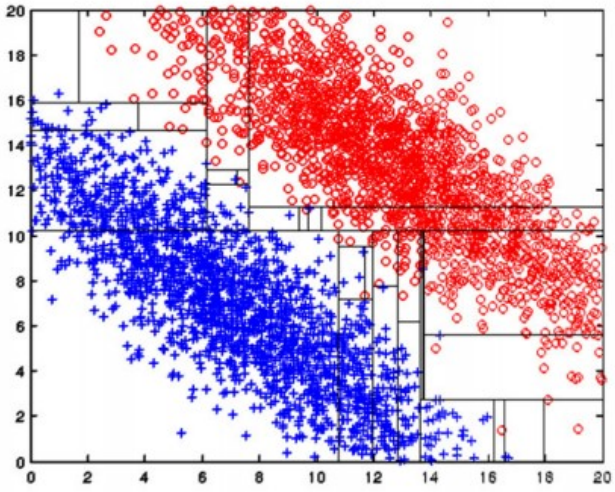
Entropy (X) : 0.99
Entropy (Y) : 0.99
Entropy (Z) : 0.98

Attribute Z will be chosen for splitting!





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.