



本课件仅用于教学使用。未经许可，任何单位、组织和个人不得将课件用于该课程教学之外的用途(包括但不限于盈利等)，也不得上传至可公开访问的网络环境

1

数据科学导论

Introduction to Data Science

第四章 数据挖掘基础

黄振亚，陈恩红

Email: huangzhy@ustc.edu.cn, cheneh@ustc.edu.cn

课程主页:

<http://staff.ustc.edu.cn/~huangzhy/Course/DS2025.html>



数据建模基础

2

□ 基本概念——数据挖掘是什么？

- **数据挖掘**：从大量的数据中挖掘哪些令人感兴趣的、有用的、隐含的、先前未知的和可能有用的**模式或知识**，并据此更好的服务人们的生活。

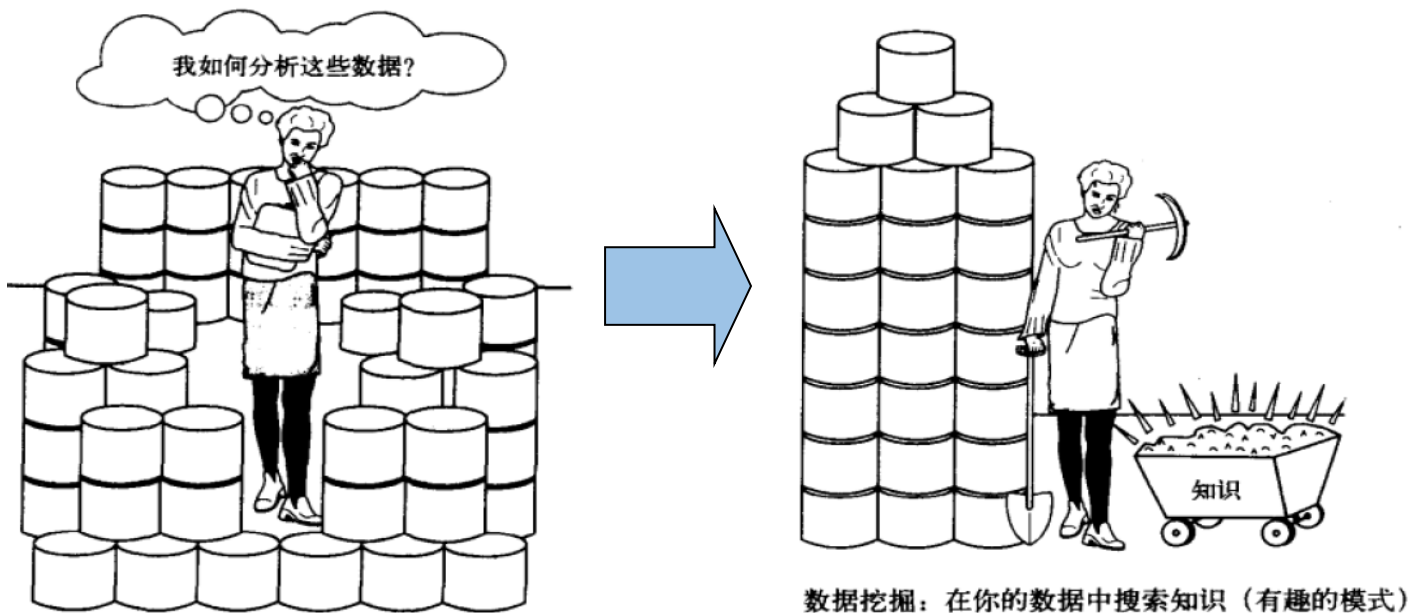


图1-2 我们的数据丰富，但信息贫乏



数据建模基础

3

□ 基本概念——数据挖掘是什么？

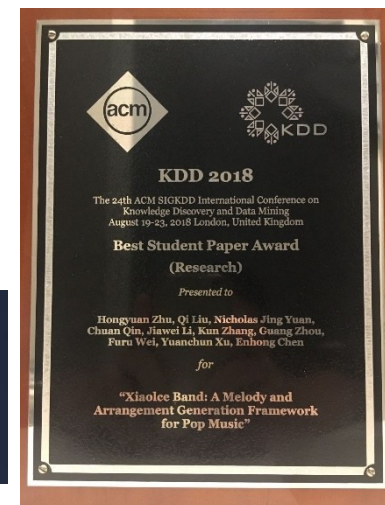
□ 数据挖掘的近义词

- 从数据中挖掘知识
 - Knowledge Discovery in Data
- 知识提炼
- 数据/模式分析
- 数据考古
- 数据捕捞、信息收获、资料勘探等。



□ SIGKDD: Knowledge Discovery and Data Mining

25TH ACM SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING



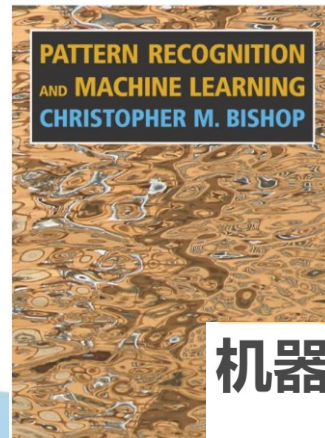
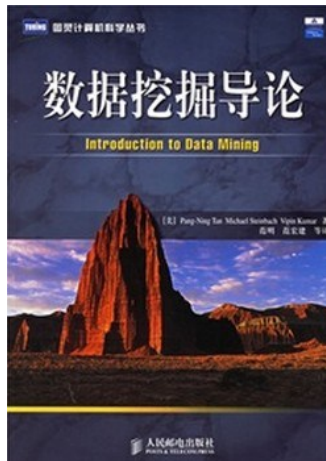
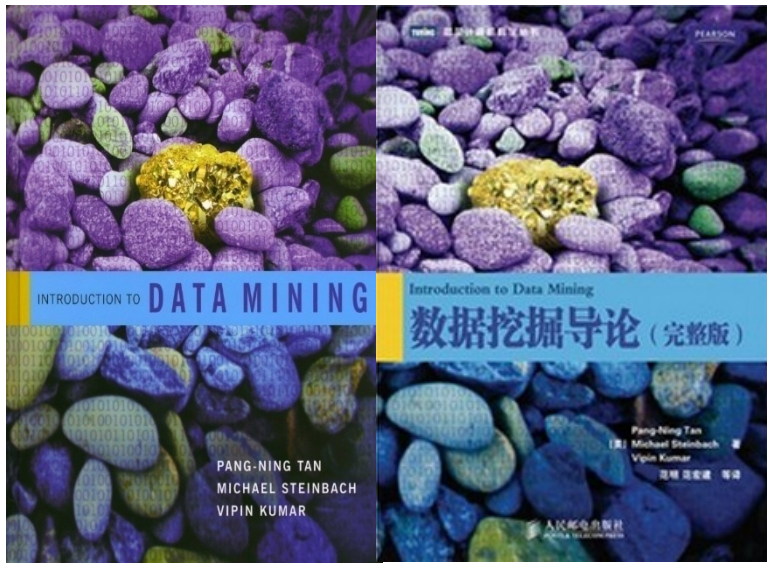


数据建模基础

4

参考书

- 数据挖掘导论 (Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Addison Wesley)



机器学习



周志华
MACHINE LEARNING
机器学习



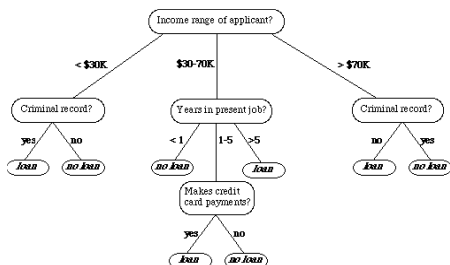


数据建模基础

5

数据挖掘有哪些典型任务？

分类与预测



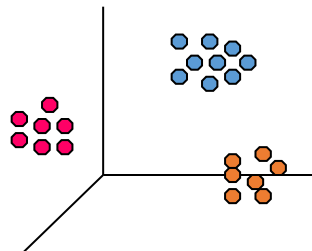
数据

| | T | | H | | P | |
|---|-------|------|----|-----|------|------|
| | L | H | L | H | L | H |
| J | -6.0 | 8.8 | 60 | 100 | 986 | 1044 |
| F | -2.8 | 10.9 | 48 | 100 | 973 | 1025 |
| M | -5.6 | 17.7 | 34 | 100 | 976 | 1037 |
| A | -1.2 | 22.2 | 27 | 100 | 996 | 1036 |
| M | -0.8 | 27.8 | 25 | 100 | 1003 | 1034 |
| J | 5.2 | 29.1 | 26 | 100 | 998 | 1030 |
| J | 9.8 | 30.6 | 23 | 99 | 997 | 1027 |
| A | 5.6 | 26.1 | 31 | 100 | 992 | 1029 |
| S | 5.2 | 24.8 | 35 | 100 | 998 | 1028 |
| O | -0.4 | 21.3 | 42 | 100 | 990 | 1031 |
| N | -7.6 | 17.3 | 55 | 100 | 963 | 1023 |
| D | -10.4 | 9.2 | 53 | 100 | 987 | 1039 |

table 17a

2010 monthly weather variation, Cambridge (UK)

聚类



关联分析





数据建模基础

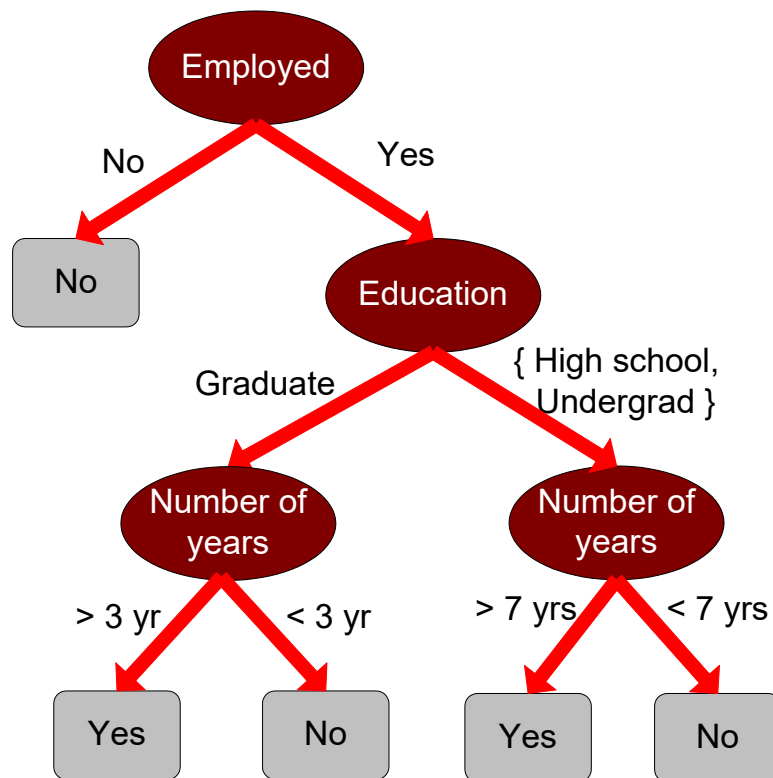
6

- 数据挖掘任务——分类与预测 (Classification, Prediction)
 - 预测性建模 (监督学习)
 - 寻找一个模型：特征→类别的函数

类别

| Tid | Employed | Level of Education | # years at present address | Credit Worthy |
|-----|----------|--------------------|----------------------------|---------------|
| 1 | Yes | Graduate | 5 | Yes |
| 2 | Yes | High School | 2 | No |
| 3 | No | Undergrad | 1 | No |
| 4 | Yes | High School | 10 | Yes |
| ... | ... | ... | ... | ... |

Model for predicting credit worthiness (信誉)



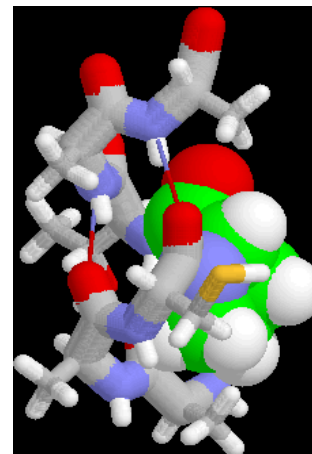


数据建模基础

7

数据挖掘任务——分类与预测 (Classification, Prediction)

- 邮件分类 (垃圾邮件)
- 将新闻故事分类为财经、天气、娱乐、体育等
- 判断信用卡交易是合法的还是欺诈
- 将蛋白质的二级结构分类为 α -螺旋、 β -薄片或随机螺旋
- 预测肿瘤细胞是良性还是恶性
- 识别网络空间的入侵者
- 推荐系统
- . . .





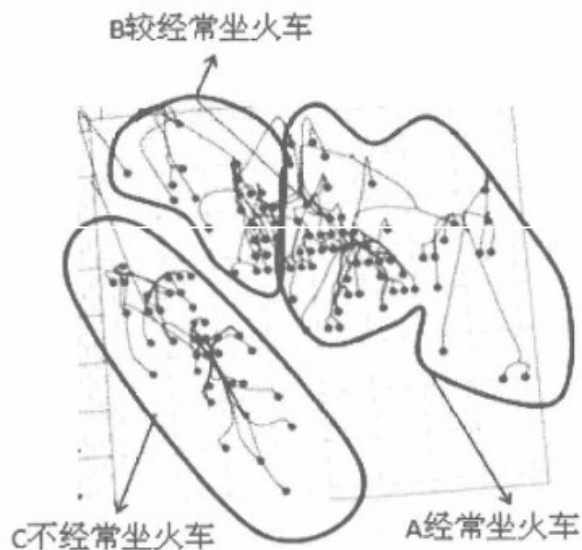
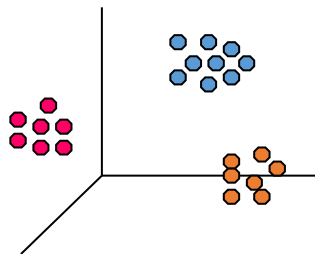
数据建模基础

8

数据挖掘任务——聚类(Clustering, unsupervised learning)

- 例如：铁路票价制定
- 问：如何制定合适的票价提高上座率？
- 方案：将旅客进行聚类分析，根据旅客乘坐高铁的频率 提供不同的优惠政策。合适的定价是提高高铁上座率的保障

聚类





数据建模基础

9

数据挖掘任务——聚类(Clustering)

- 例：搜索词条聚类(Query clustering)
- “USTC”，“中科大”，“中国科大”，“中国科学技术大学”
- “长城”，“颐和园”，“故宫”





数据建模基础

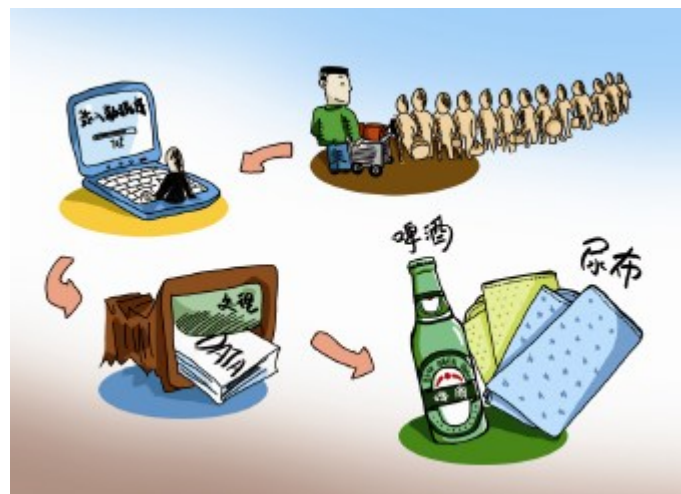
10

数据挖掘任务——关联分析(Association Analysis)

- 例如：“啤酒与尿布”
- 在一次圣诞节的顾客消费行为分析中，沃尔玛意外发现跟尿布一起购买最多的商品竟然是啤酒。经过深入分析后，卖场立即对两类商品的空间距离与价格都进行了调整，结果尿布与啤酒销量双双大增。



萨姆·沃尔顿
沃尔玛公司创始人



轰动一时的啤酒与尿布关联规则

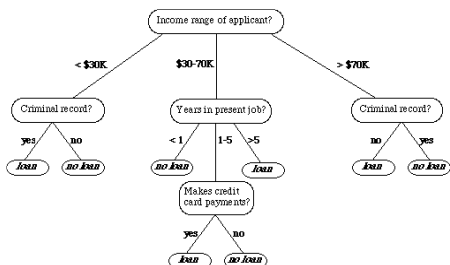


数据建模基础

11

数据挖掘——四个任务有哪些常用方法？

分类与预测



关联分析



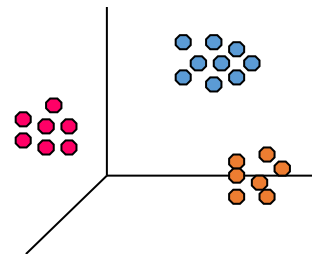
数据

| | T | | H | | P | |
|---|-------|------|----|-----|------|------|
| | L | H | L | H | L | H |
| J | -6.0 | 8.8 | 60 | 100 | 986 | 1044 |
| F | -2.8 | 10.9 | 48 | 100 | 973 | 1025 |
| M | -5.6 | 17.7 | 34 | 100 | 976 | 1037 |
| A | -1.2 | 22.2 | 27 | 100 | 996 | 1036 |
| M | -0.8 | 27.8 | 25 | 100 | 1003 | 1034 |
| J | 5.2 | 29.1 | 26 | 100 | 998 | 1030 |
| J | 9.8 | 30.6 | 23 | 99 | 997 | 1027 |
| A | 5.6 | 26.1 | 31 | 100 | 992 | 1029 |
| S | 5.2 | 24.8 | 35 | 100 | 998 | 1028 |
| O | -0.4 | 21.3 | 42 | 100 | 990 | 1031 |
| N | -7.6 | 17.3 | 55 | 100 | 963 | 1023 |
| D | -10.4 | 9.2 | 53 | 100 | 987 | 1039 |

table 17a

2010 monthly weather variation, Cambridge (UK)

聚类





分类与预测

12

数据挖掘任务 — 分类与预测

这张照片是哪里？



该图片有关联



该图片是科大



如果数据有标签，即已知图片是科大，则可以预测新图片的类



分类与预测

13

案例一：垃圾邮件分类 — 中科大安全演练

判断：下面这封邮件是垃圾邮件吗？

结论：一封垃圾邮件

中秋免费月饼领取 发起会议

发件人：中科大邮箱管理中心 <mailservice@vstc.edu.cn>

时间：2022年09月07日 18:39:40 (星期三)

收件人：huangzhy@ustc.edu.cn

特征1：仿冒地址：vstc.edu.cn

尊敬的科大邮箱用户，

您好！

金秋九月，丹桂飘香，中秋佳节临近，中科大邮箱管理中心祝您中秋快乐，万事如意！

了解到广大师生对我校定制月饼礼盒购买意愿强烈，礼盒供不应求，本部门特地采购了一批月饼礼盒，并以抽奖的形式回馈各位用户。由于礼盒数量有限，仅限在校师生参与抽奖，请点击以下链接参与抽奖活动，祝您好运！

校内抽奖链接：统一身份认证

中科大邮箱管理中心

特征2：不存在的科大部门：中科大邮箱管理中心

此邮件为自动发送，请勿回复

在使用中碰到任何问题，请点击链接联系或者电话联系：0551-36309527

特征3：错误的联系电话：36309527

Copyright 2022

基于一些特征与规则，我们可以将垃圾邮件的判别视作一个分类问题



分类与预测

14

案例二：电影评分预测

预测：用户对电影《功夫》的评分是多少？

已知：他对4部电影的评分分别为：5.0, 4.8, 4.9, 4.5

特征1：喜欢周星驰



特征2：喜欢喜剧



结论：预测评分5分



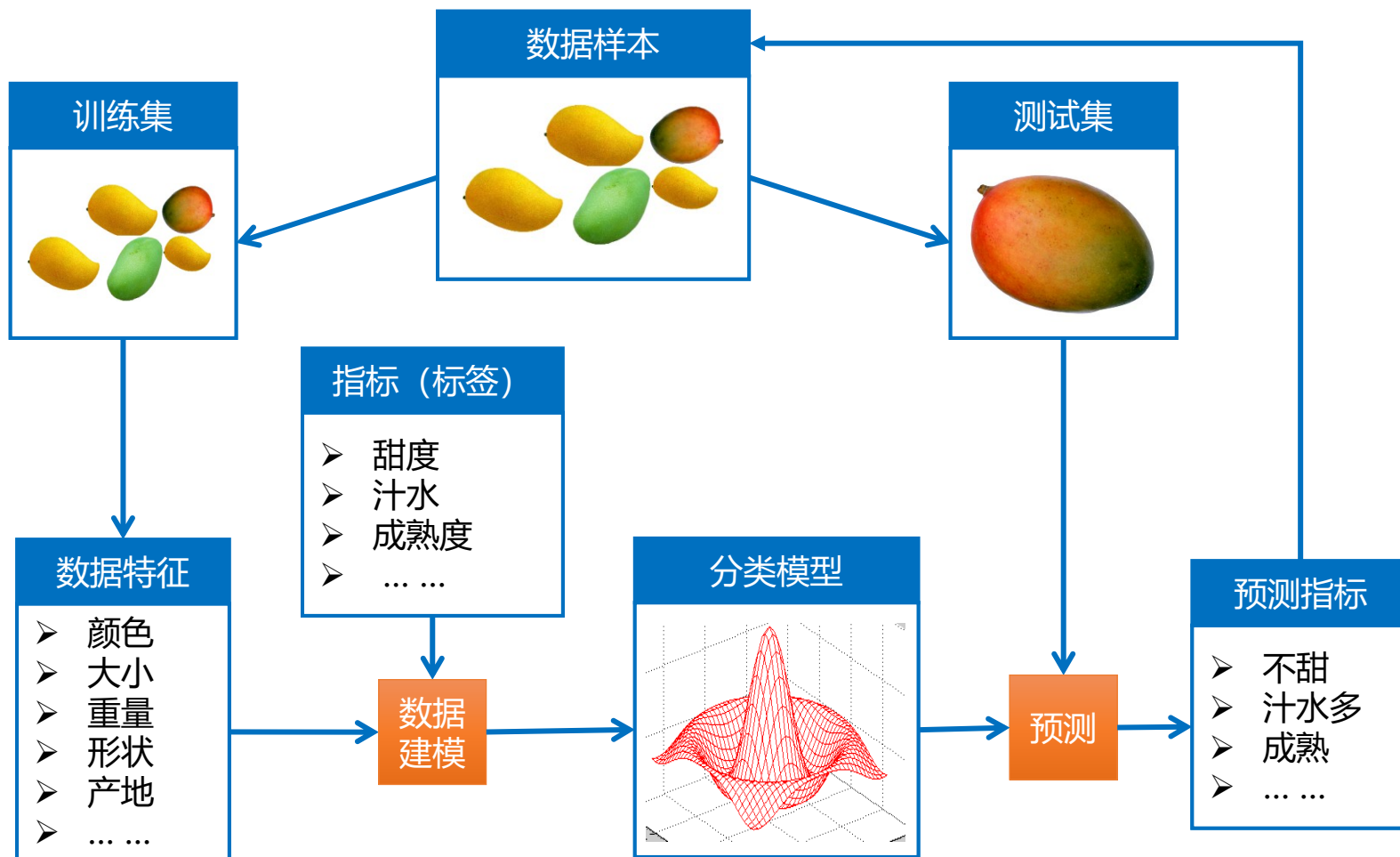
基于一些特征与规则，我们可以将电影评分(连续值)估计视作一个预测问题



分类与预测

15

案例三：生活中买芒果





分类与预测

16

□ 分类与预测 — 有监督学习

- 已知：一组数据（训练集） (X, Y)
- 如右图，每一条记录表示为 (x, y)

- x ：数据特征/属性（如收入）
- y ：类别标记（是否有借款）

□ 任务：

- 学习一个模型，利用每一条记录的特征 x 去预测它对应的类别 y

即：输入未标记的数据（含特征 x ），
预测数据的类别 y

**分类 / 数值预测 取决于 类别标签是
离散型 / 数值型**

3个特征：

- 是否有住房
- 婚姻状态
- 年收入

类别：

是否拖欠贷款

| 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 |



分类与预测

17

□ 分类与预测 — 回顾前例

| 任务 | 特征 x | 类别 y |
|--------|-------------------------|-----------------------------------|
| 垃圾邮件分类 | 收件人、邮箱名、邮件内容等 | 是否垃圾邮件 离散型 |
| 电影评分预测 | 用户在其他电影的评分 电影的演员，类型等 | 实值评分[0,5] 数值型 |
| 芒果好坏预测 | 芒果的颜色、大小、重量、形状、产地等 | 芒果的甜度、水分、成熟与否 离散型 或 数值型 |



分类与预测

18

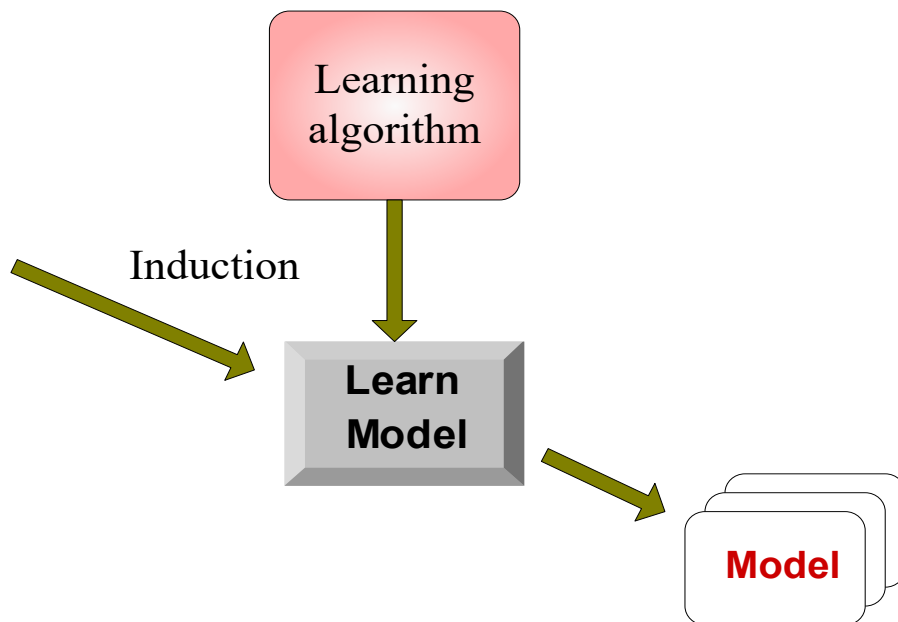
如何建立分类与预测模型？

- 一般流程：有监督学习
- 通常包括两个阶段：模型训练、模型预测
 - 模型训练：目标是利用训练数据，学习一个分类或预测模型

| 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

训练集有类别标签





分类与预测

19

□ 如何建立分类与预测模型？

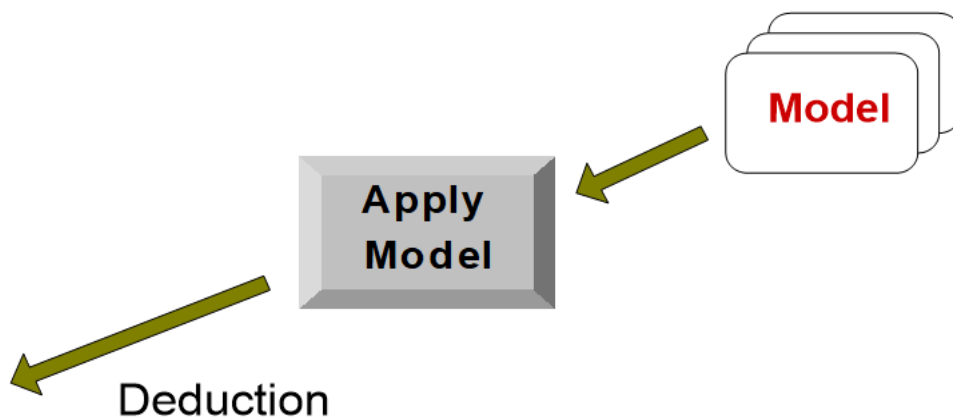
- 一般流程：有监督学习
- 通常包括两个阶段：模型训练、模型预测
 - 模型预测：目标是利用学习的模型，预测测试数据的标签

| 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



测试集无类别标签，需要预测





分类与预测

20

- 有监督学习：分类与预测
- 常用方法
 - 规则方法
 - 决策树
 - 最近邻方法
 - 支持向量机 (SVM)
 - 集成方法
- 分类的评价指标



分类：规则方法

21

□ 规则方法

- 基于规则的分类器 (Rule-based Classifier) 就是使用一组 if-then 的模式来进行分类
- 基本形式：Condition \rightarrow y (标签)
 - 其中，Condition是一组属性的组合，也被称作规则的前提
- 例如：
 - (胎生= 否) \wedge (飞行动物= 是) \rightarrow 鸟类
 - (胎生= 是) \wedge (体温= 恒温) \rightarrow 哺乳类
- 最基础的获得规则的方法：人工制定规则进行分类



不足：人工定义规则、效率低，难以处理复杂问题



自动生成规则？决策树



分类：规则方法

22

□ 回顾垃圾邮件分类的例子

□ 判断：下面这封邮件是垃圾邮件吗？

中秋免费月饼领取     发起会议 精简信息 

发件人: 中科大邮箱管理中心 <mailservice@vstc.edu.cn>
时 间: 2022年09月07日 18:39:40 (星期三)
收件人: huangzhy@ustc.edu.cn

尊敬的科大邮箱用户，

您好！

金秋九月，丹桂飘香，中秋佳节临近，中科大邮箱管理中心祝您中秋快乐，万事如意！

了解到广大师生对我校定制月饼礼盒购买意愿强烈，礼盒供不应求，本部门特地采购了一批月饼礼盒，并以抽奖的形式回馈各位用户。由于礼盒数量有限，仅限在校师生参与抽奖，请点击以下链接参与抽奖活动，祝您好运！

校内抽奖链接： [统一身份认证](#)

中科大邮箱管理中心

此邮件为自动发送，请勿回复

在使用中碰到任何问题，请点击链接联系或者电话联系：0551-36309527

Copyright 2022

特征1：仿冒地址：vstc.edu.cn

特征2：不存在的科大部门：中科大邮箱管理中心

特征3：错误的联系电话：36309527

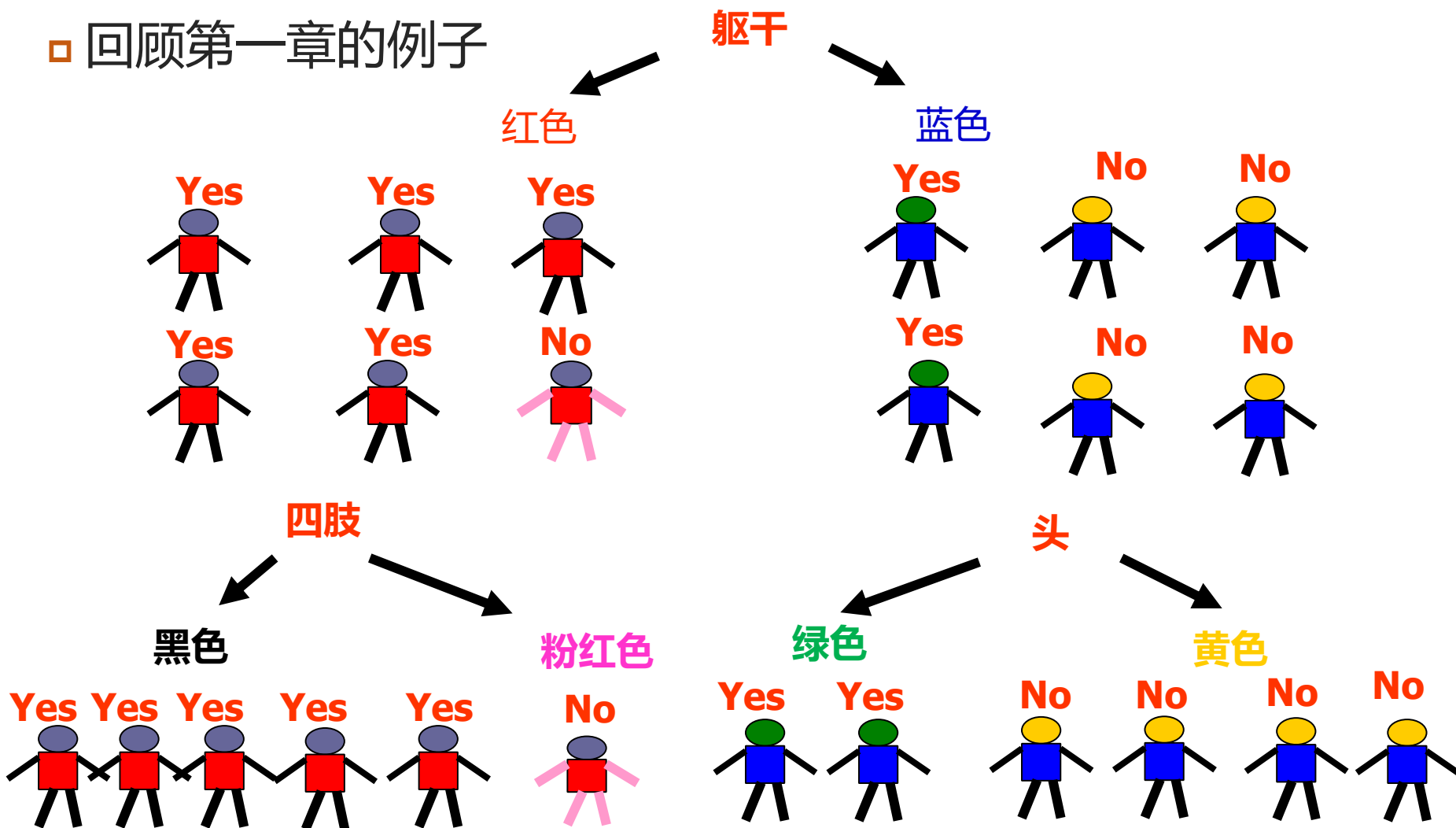
(地址= vstc.edu.cn) ^ (部门= 中科大邮箱管理中心) ^ (电话= 36309527) → 垃圾邮件



分类：决策树

23

□ 回顾第一章的例子





分类：决策树

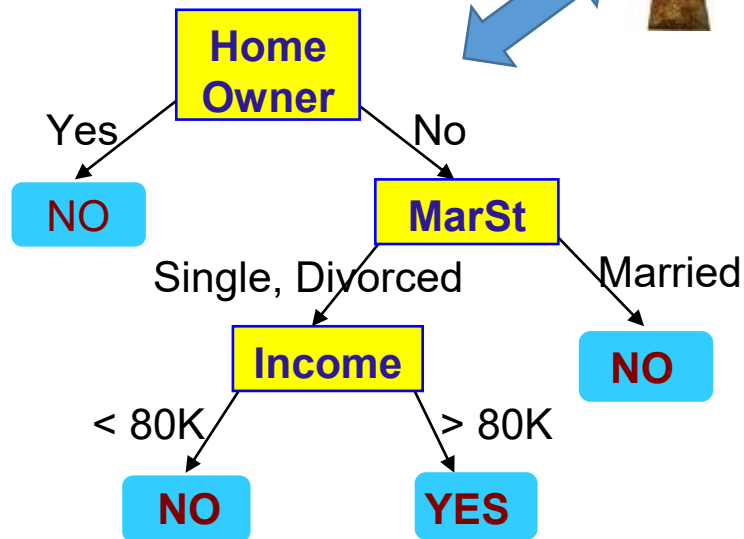
24

□ 什么是决策树

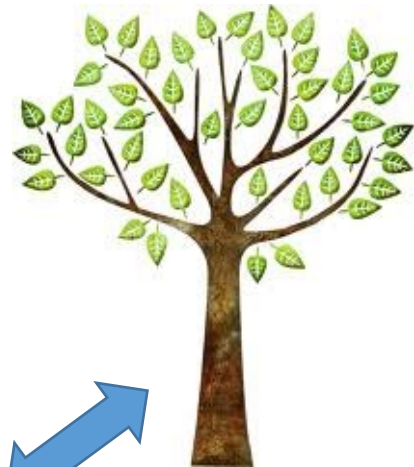
- 对数据进行处理，利用归纳算法生成可读的规则
- 模型以树状形式呈现出来

| 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 |

训练数据



模型：决策树





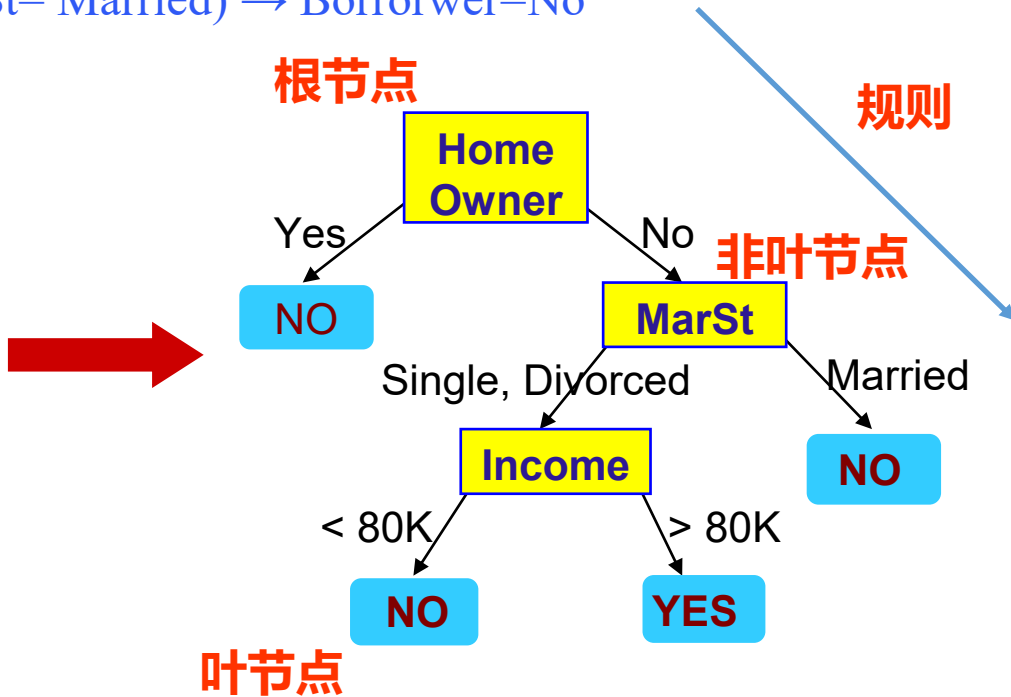
分类：决策树

25

什么是决策树 —— 基本概念

- 非叶节点：一个属性上的测试，每个分枝代表该测试的输出
- 叶节点：存放一个类标记
- 规则：从根节点到叶节点的一条属性取值路径
 - $(\text{HomOwn} = \text{No}) \wedge (\text{MarSt} = \text{Married}) \rightarrow \text{Borrower} = \text{No}$

| 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 |





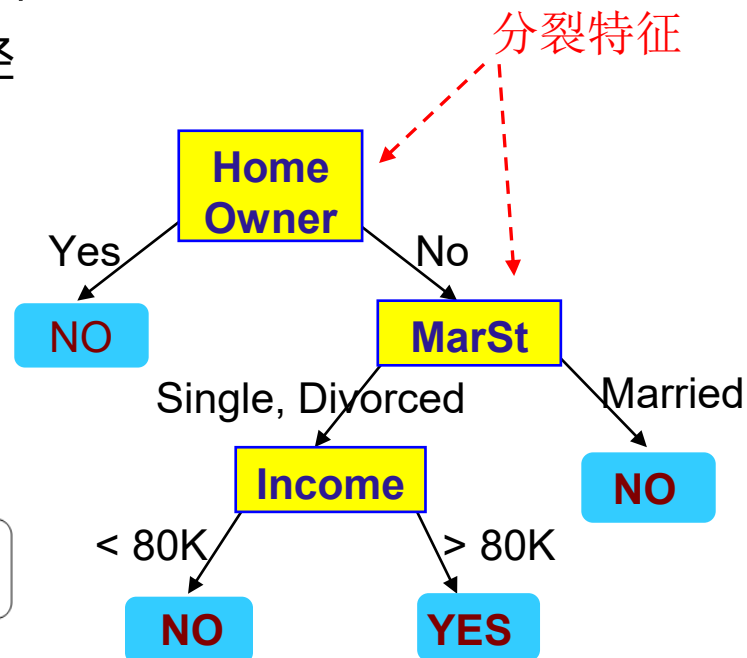
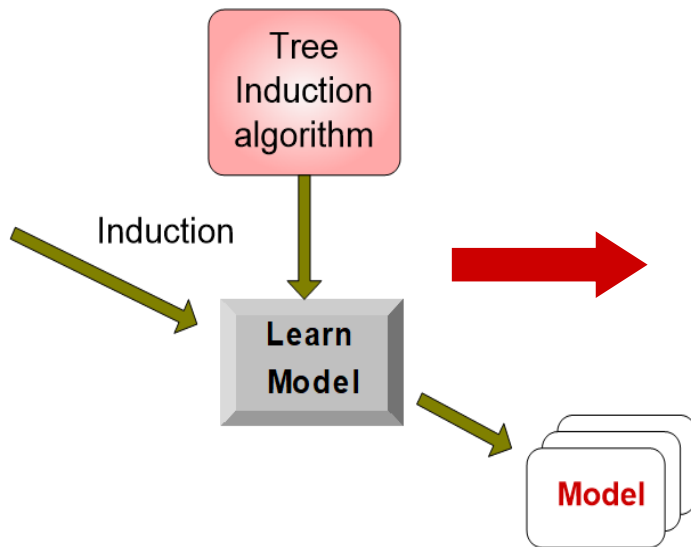
分类：决策树

26

- 建立决策树分类模型的流程
 - 模型训练：从已有数据中生成一棵决策树
 - 分裂数据的特征，寻找决策类别的路径

| 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



分裂特征: Home Owner, MarSt, Income

生成模型：决策树

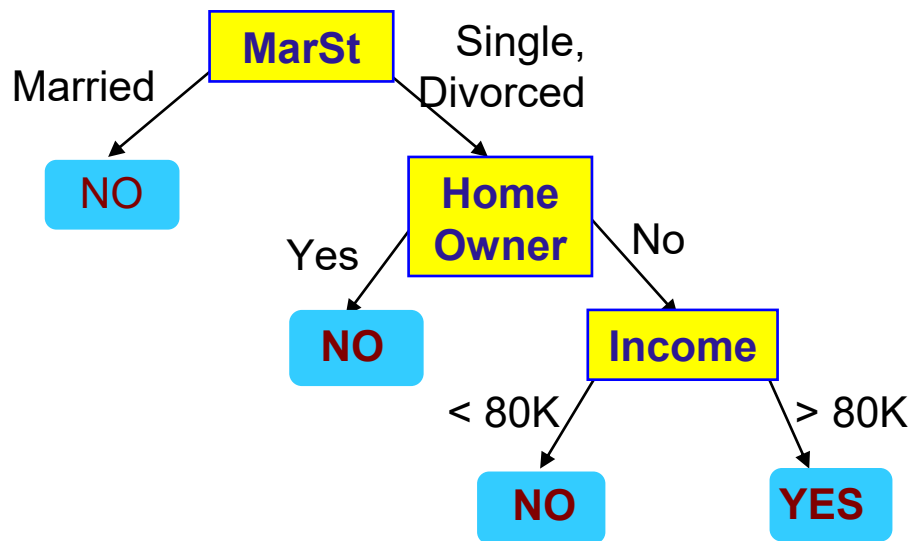
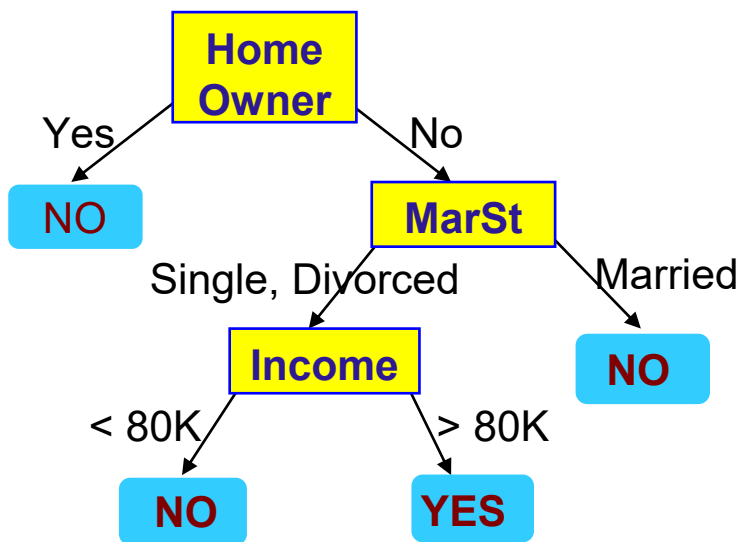


分类：决策树

27

□ 是否有其他决策树？

| 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 |



特征顺序: Home Owner, MarSt, Income

特征顺序: MarSt, Home Owner, Income

相同的数据，根据不同的特征顺序，可以建立多种决策树

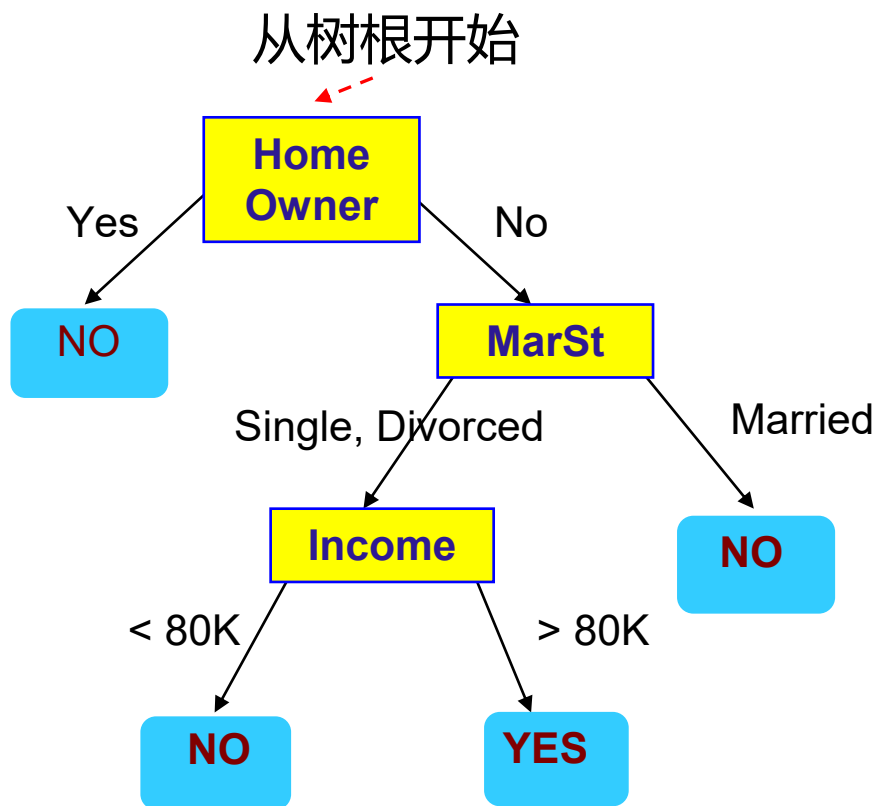


分类：决策树

28

决策树分类模型的测试过程

- 模型测试：根据规则将样本分类到某个叶子节点



测试数据（预测标签）

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



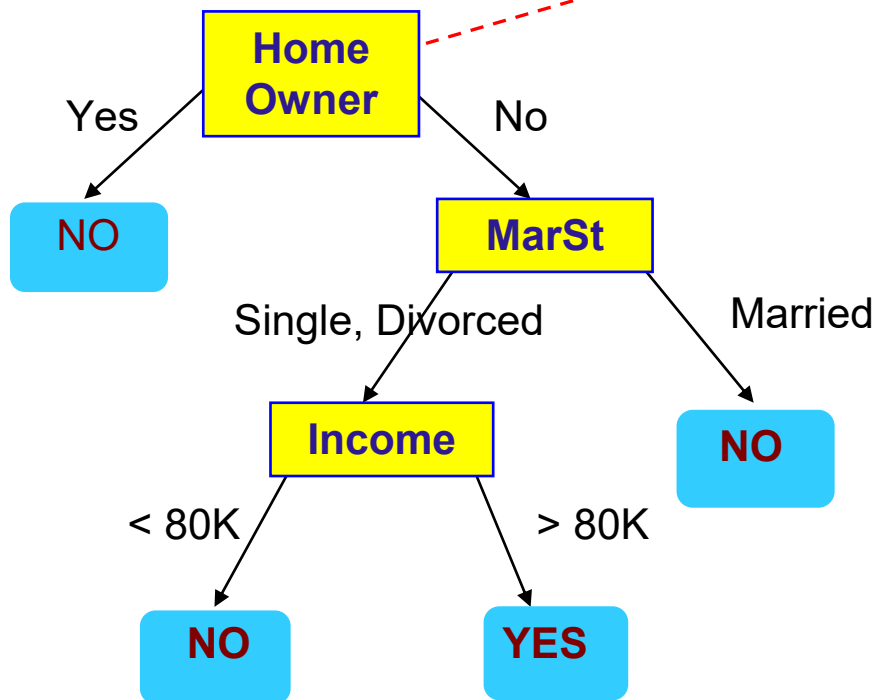
分类：决策树

29

决策树分类模型的测试过程

测试数据（预测标签）

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





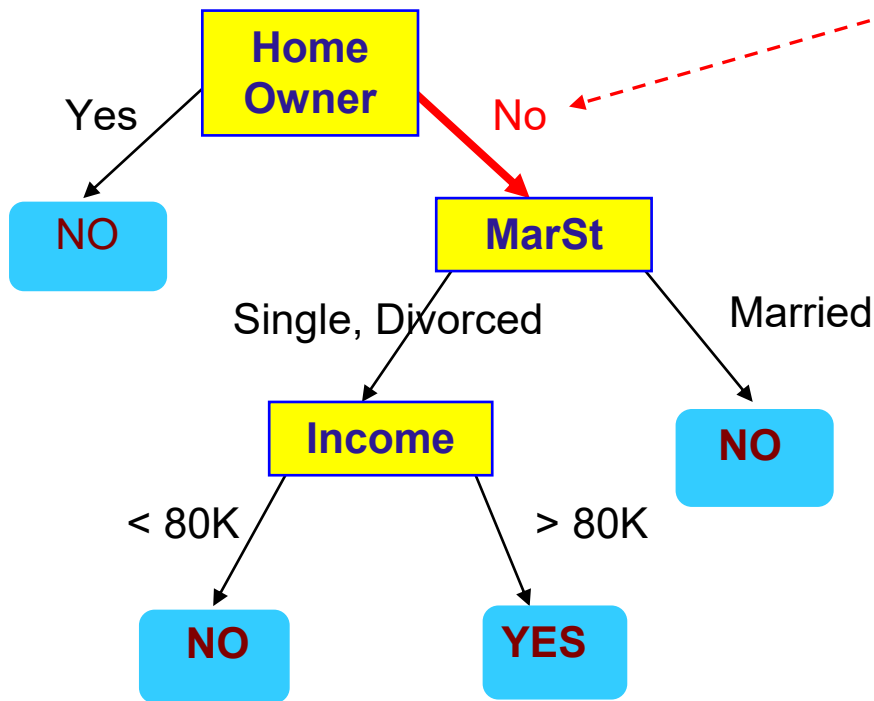
分类：决策树

30

决策树分类模型的测试过程

测试数据（预测标签）

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



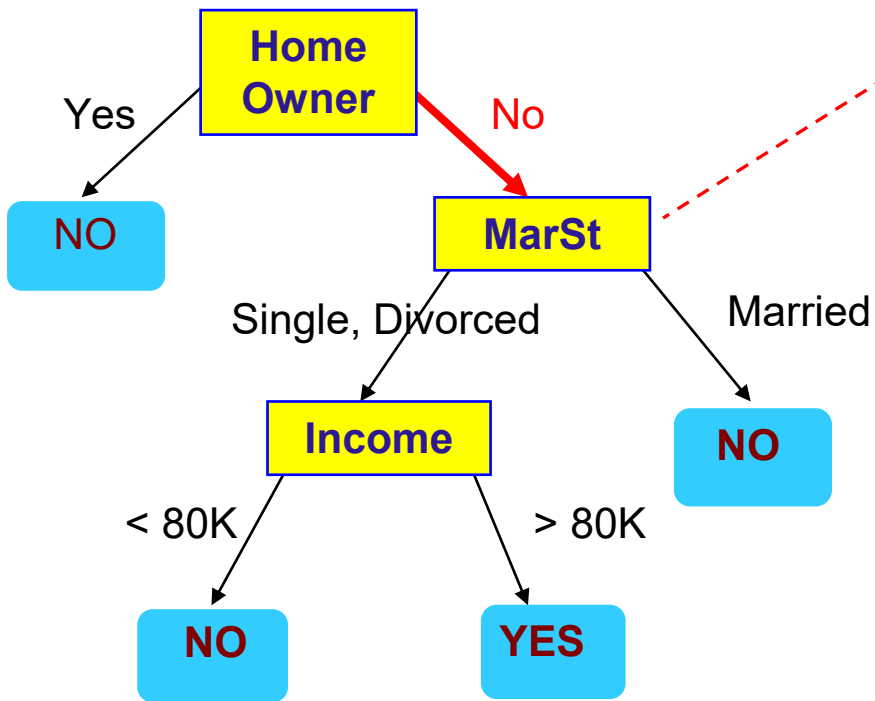


分类：决策树

31

决策树分类模型的测试过程

测试数据（预测标签）



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



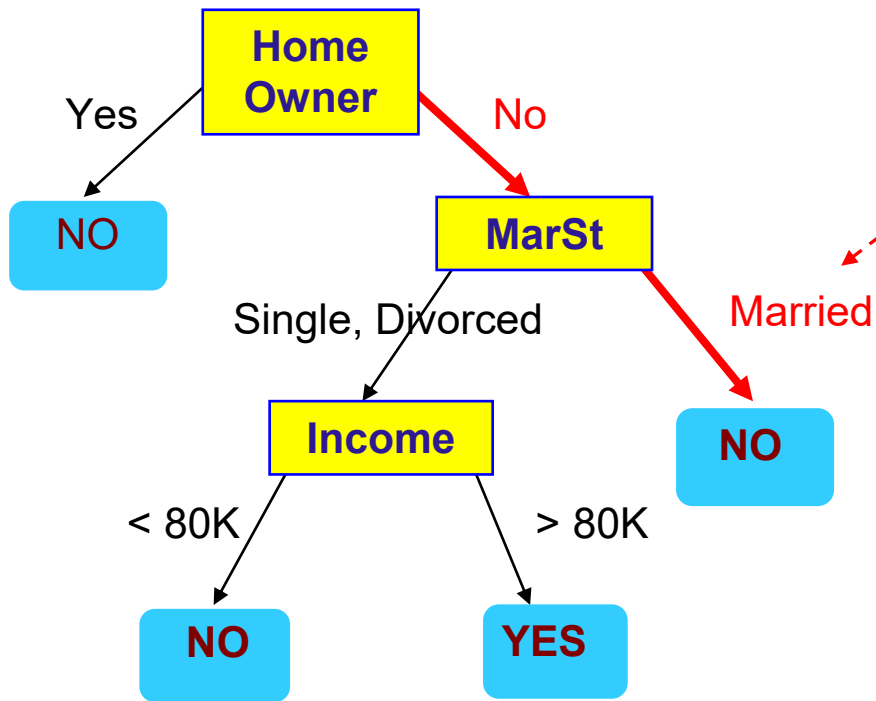
分类：决策树

32

决策树分类模型的测试过程

测试数据（预测标签）

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





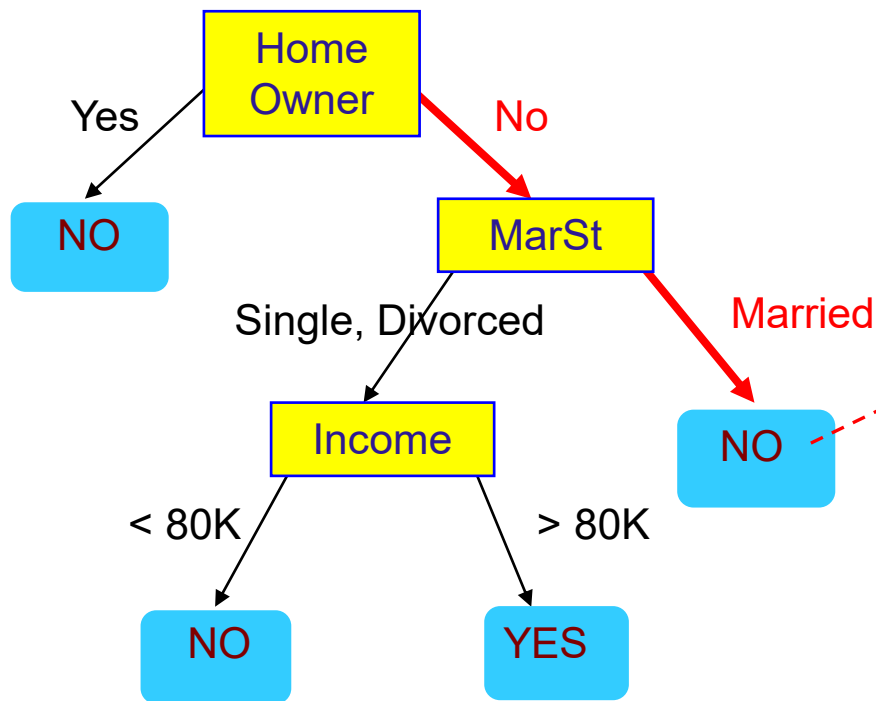
分类：决策树

33

决策树分类模型的测试过程

测试数据（预测标签）

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



类别Defaulted Borrower为“No”

决策过程：未使用所有的特征/属性