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/DS2024.html

数据建模基础

- 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。

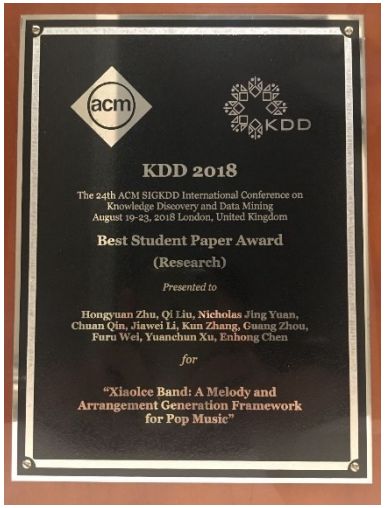
图1-2 我们的数据丰富，但信息贫乏

数据挖掘：在你的数据中搜索知识（有趣的模式）

数据建模基础

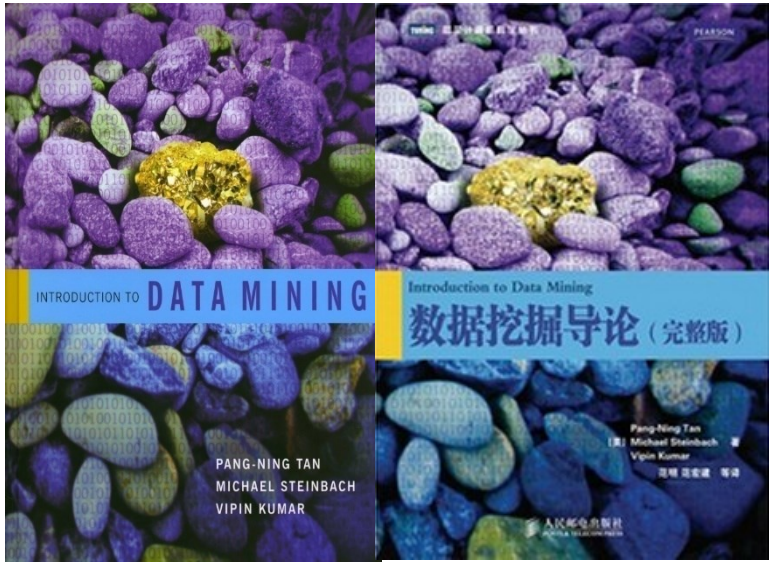
- 基本概念——数据挖掘是什么？
  - 数据挖掘的近义词
    - 从数据中挖掘知识
      - Knowledge Discovery in Data
    - 知识提炼
    - 数据/模式分析
    - 数据考古
    - 数据捕捞、信息收获、资料勘探等。

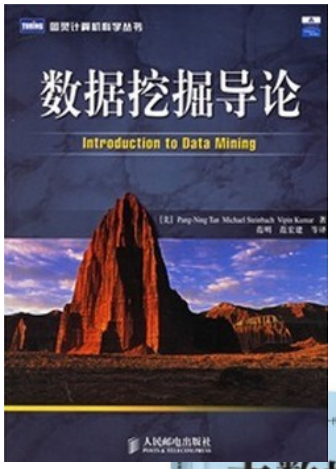
- SIGKDD: Knowledge Discovery and Data Mining

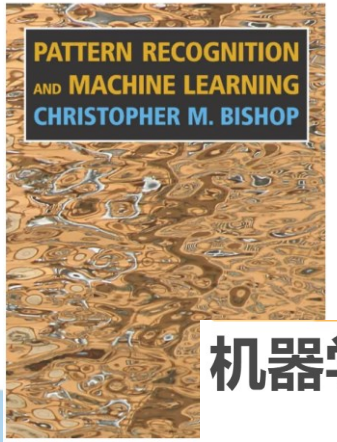


数据建模基础

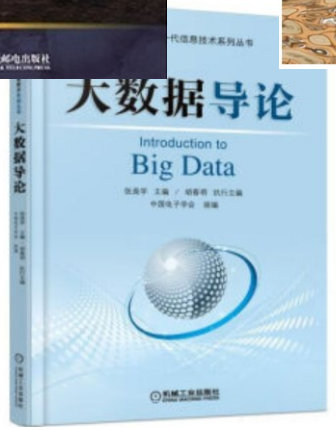
- 参考书
  - 数据挖掘导论 （Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Addison Wesley ）







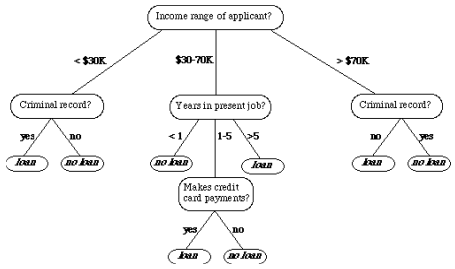
机器学习



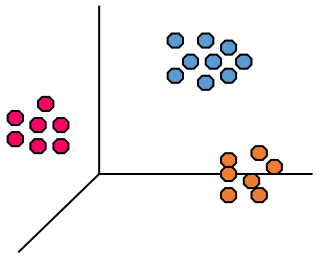
数据建模基础

- 数据挖掘有哪些典型任务?

分类与预测



聚类



数据

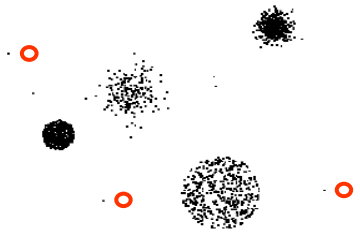
	T		H		P	
	L	H	L	H	L	H
J	-6.0	8.8	60	100	986	1044
F	-2.8	10.9	48	100	973	1025
M	-5.6	17.7	34	100	976	1037
A	-1.2	22.2	27	100	996	1036
M	-0.8	27.8	25	100	1003	1034
J	5.2	29.1	26	100	998	1030
J	9.8	30.6	23	99	997	1027
A	5.6	26.1	31	100	992	1029
S	5.2	24.8	35	100	998	1028
O	-0.4	21.3	42	100	990	1031
N	-7.6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

table 17a

2010 monthly weather variation, Cambridge (UK)

关联分析

异常检测



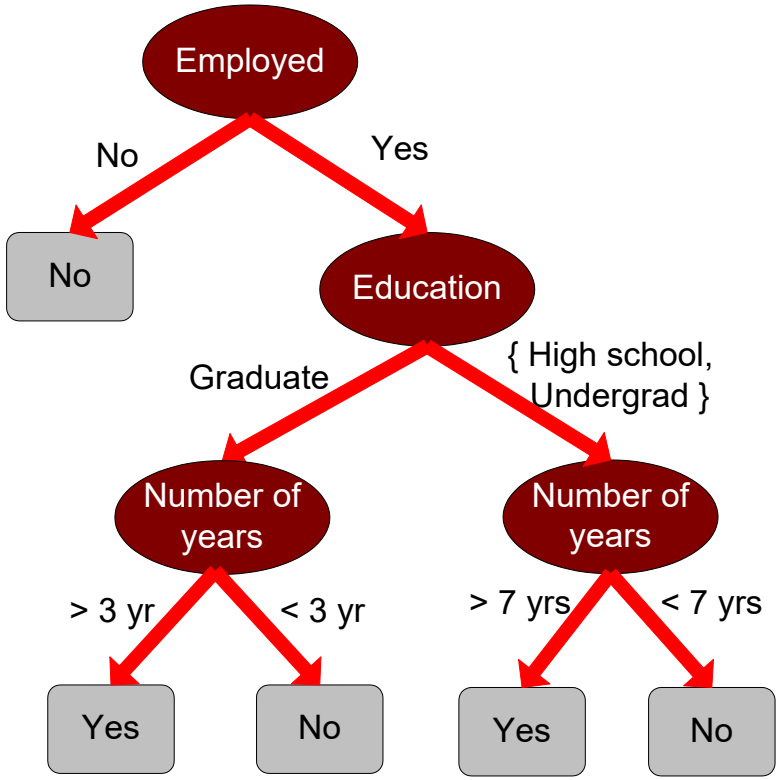
数据建模基础

- 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

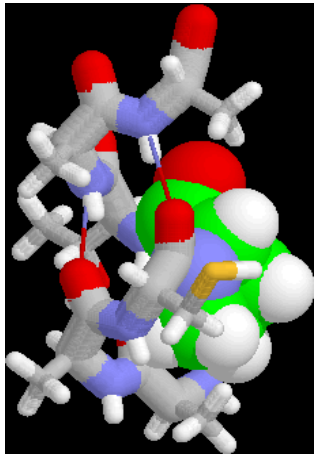
Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Model for predicting credit worthiness（信誉）



数据建模基础

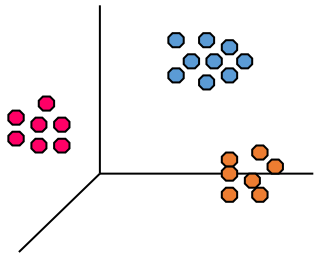
- 数据挖掘任务——分类与预测 (Classification, Prediction)
  - 邮件分类（垃圾邮件）
  - 将新闻故事分类为财经、天气、娱乐、体育等
  - 判断信用卡交易是合法的还是欺诈
  - 将蛋白质的二级结构分类为α -螺旋、β -薄片或随机螺旋
  - 预测肿瘤细胞是良性还是恶性
  - 识别网络空间的入侵者
  - 推荐系统
  - 。。。

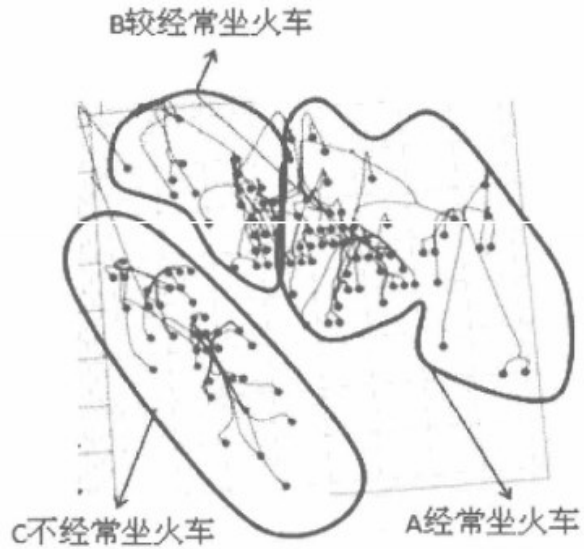


数据建模基础

- 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

- 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

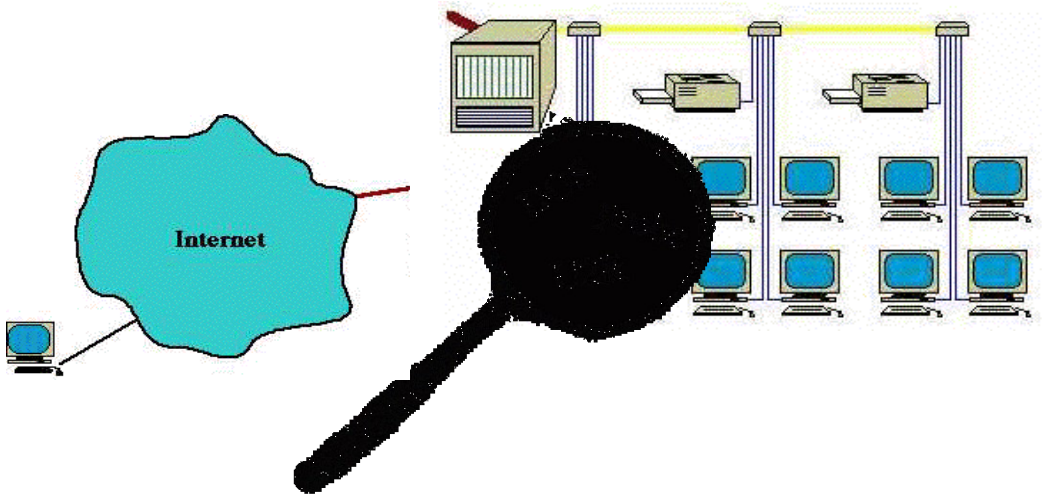
- 数据挖掘任务——关联分析(Association Analysis)
  - 例如： “啤酒与尿布”
  - 在一次圣诞节的顾客消费行为分析中，沃尔玛意外发现跟尿布一起购买最多的商品竟然是啤酒。经过深入分析后，卖场立即对两类商品的空间距离与价格都进行了调整，结果尿布与啤酒销量双双大增。

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

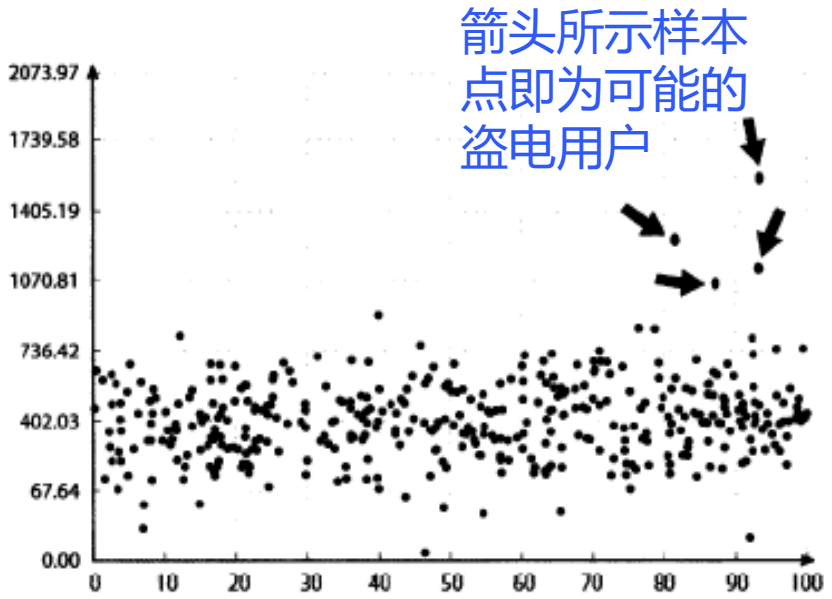
数据建模基础

- 数据挖掘任务——异常检测 (Anomaly Detection)
  - 检测非正常行为
  - 应用:
    - 信用卡诈骗检测
    - 网络入侵检测



数据建模基础

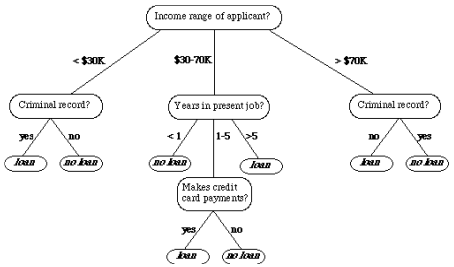
- 数据挖掘任务——异常检测 (Anomaly Detection)
  - 例：电力行业--盗电检测
    - 盗电用户行为特征与普通用户不同（电费与人员、产值、税收等形成反差）
    - 通过对用户用电数据的聚类分析，反窃电的业务人员就能对锁定的目标重点侦察，既能提高窃电客户的识别率，还能节省电力部门人力资源，为反窃电提供了另一种思路。



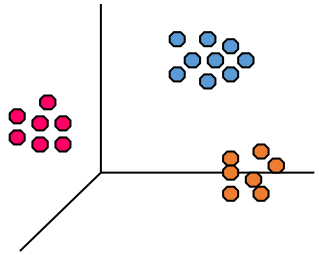
数据建模基础

- 数据挖掘——四个任务有哪些常用方法?

分类与预测



聚类



数据

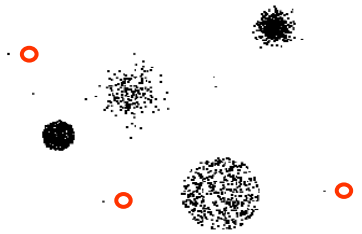
	T		H		P	
	L	H	L	H	L	H
J	-6.0	8.8	60	100	986	1044
F	-2.8	10.9	48	100	973	1025
M	-5.6	17.7	34	100	976	1037
A	-1.2	22.2	27	100	996	1036
M	-0.8	27.8	25	100	1003	1034
J	5.2	29.1	26	100	998	1030
J	9.8	30.6	23	99	997	1027
A	5.6	26.1	31	100	992	1029
S	5.2	24.8	35	100	998	1028
O	-0.4	21.3	42	100	990	1031
N	-7.6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

table 17a

2010 monthly weather variation, Cambridge (UK)

关联分析

异常检测



分类与预测

- 数据挖掘任务 — 分类与预测

这张照片是哪里？

该图片有关联

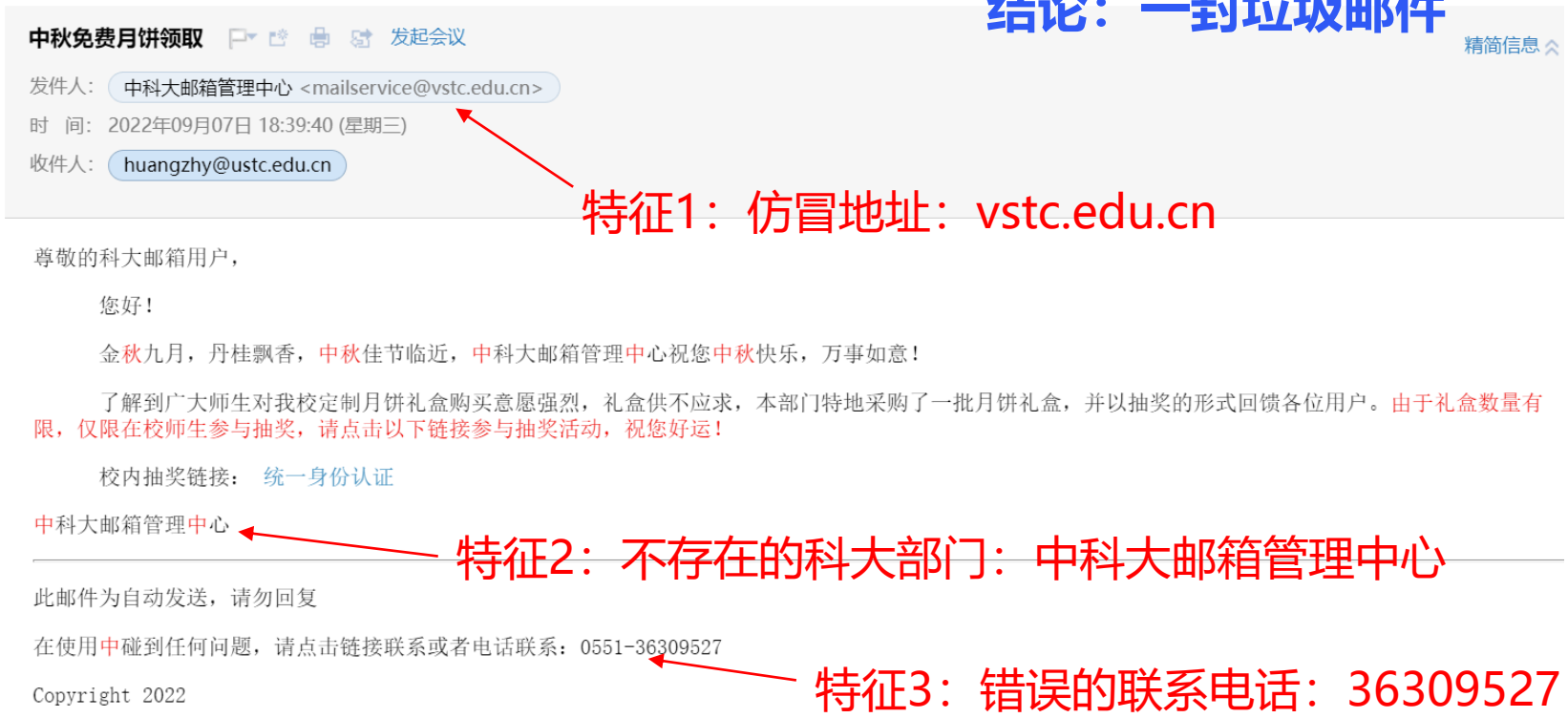
该图片是科大

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

- 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



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

- 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

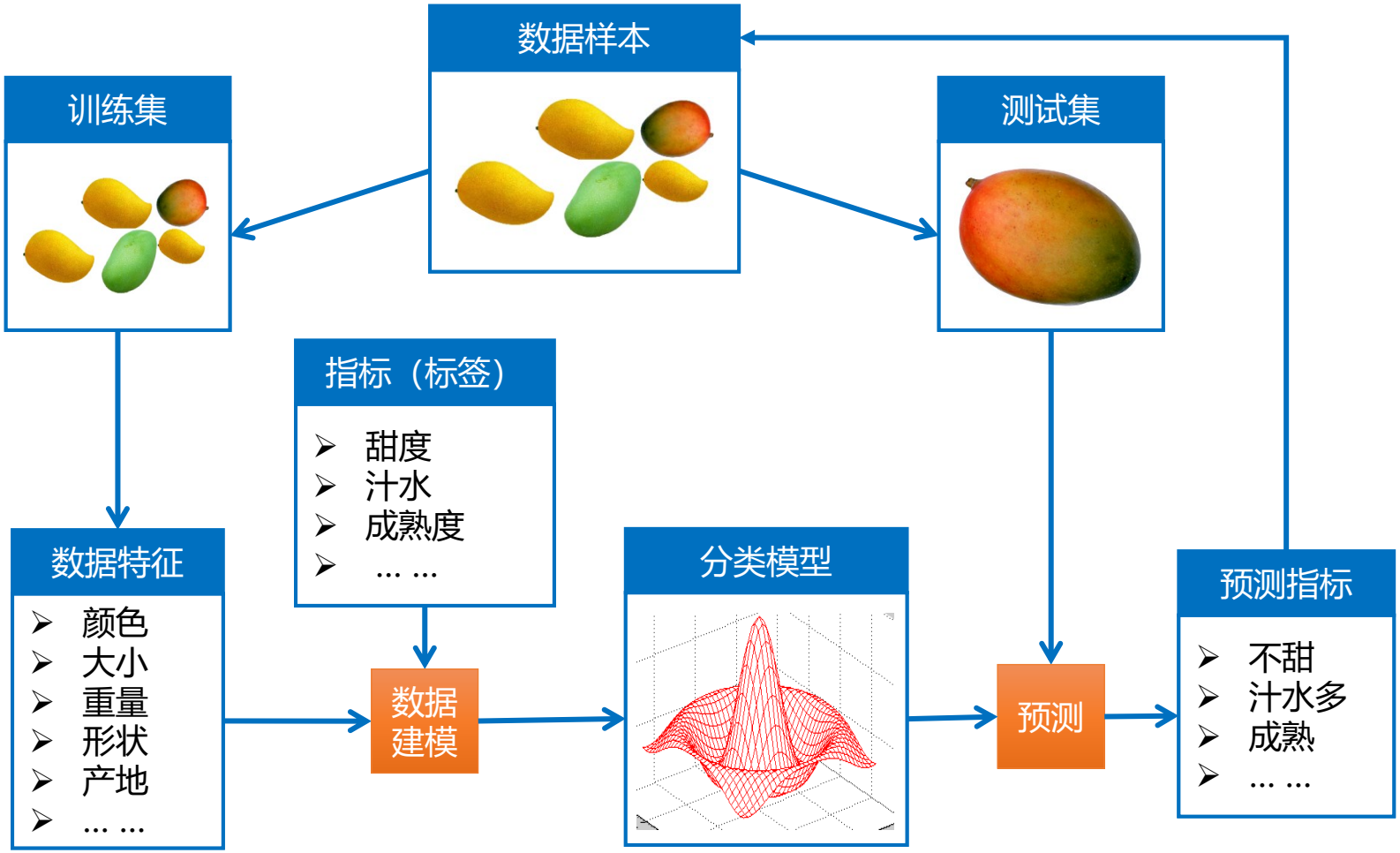
特征2：喜欢喜剧

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

- 生活中的案例：直观理解买芒果



分类与预测

- 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

- 分类与预测 — 回顾前例

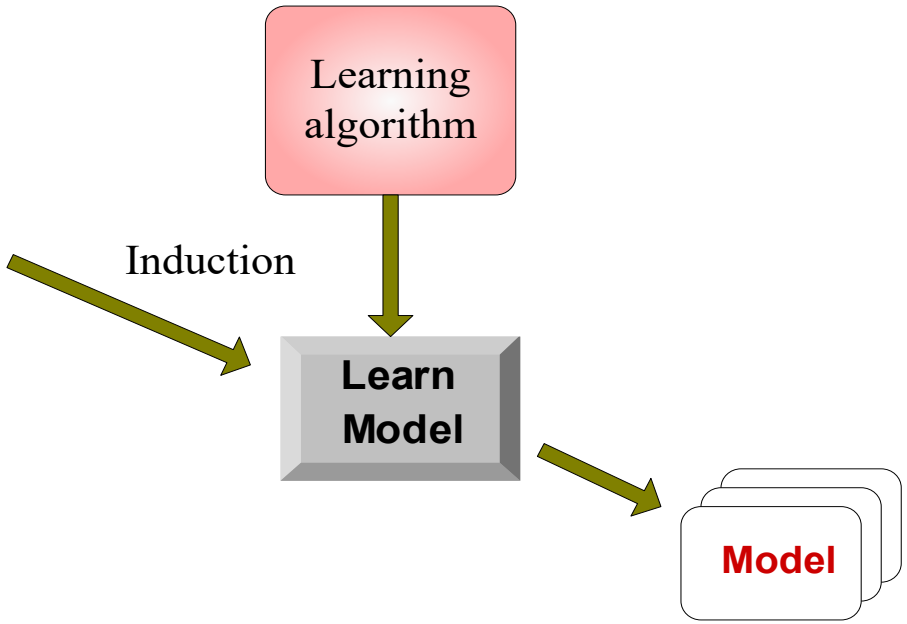
任务	特征x	类别y
垃圾邮件分类	收件人、邮箱名、邮件内容等	是否时垃圾邮件 **离散型**
电影评分预测	用户在其他电影的评分 电影的演员，类型等	实值评分[0,5] **数值型**
芒果好坏预测	芒果的颜色、大小、重量、形状、产地等	芒果的甜度、水分、成熟与否 **离散型 或 数值型**

分类与预测

- 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

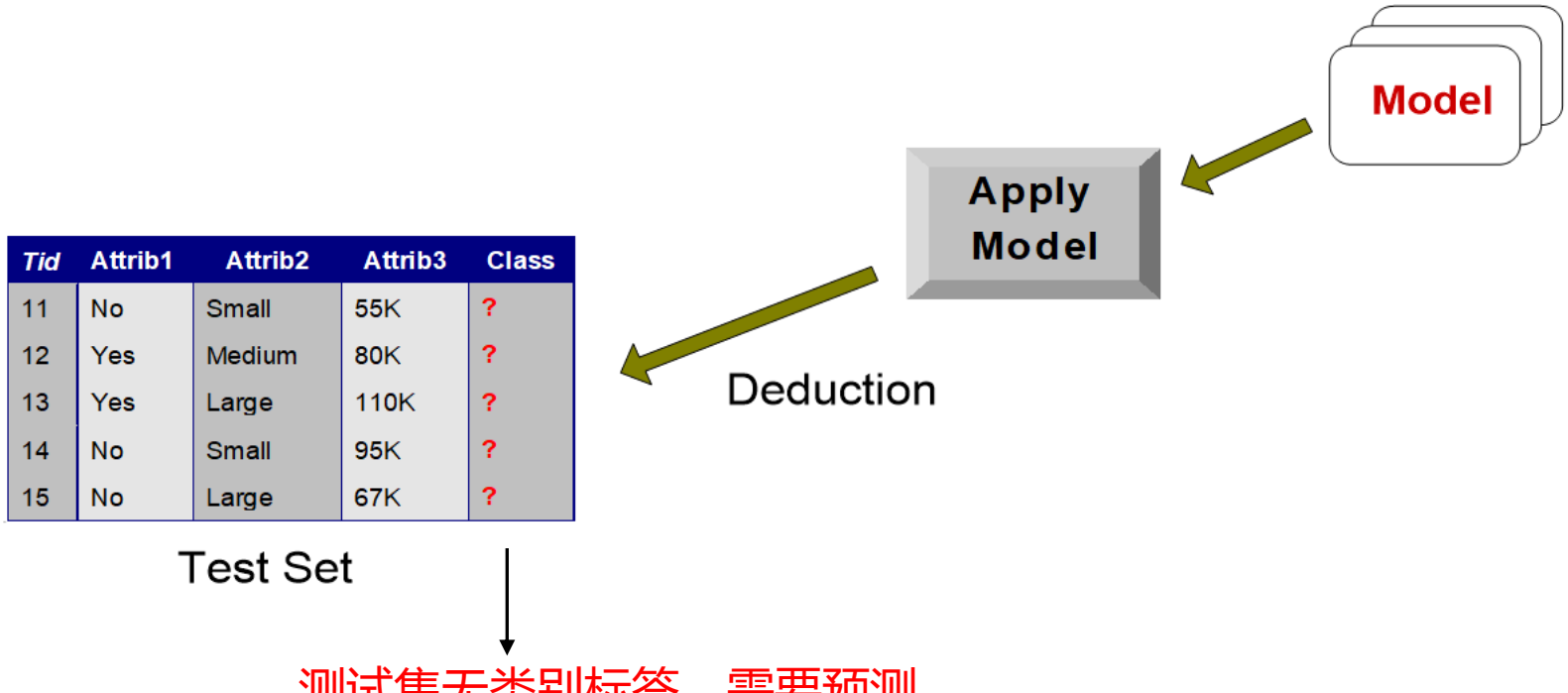
Training Set



训练集有类别标签

分类与预测

- 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



分类与预测

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

分类：规则方法

- **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 → 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) ∧ (飞行动物= 是) → 鸟类
    - (胎生= 是) ∧ (体温= 恒温) → 哺乳类
  - 最基础的获得规则的方法：人工制定规则进行分类

↓

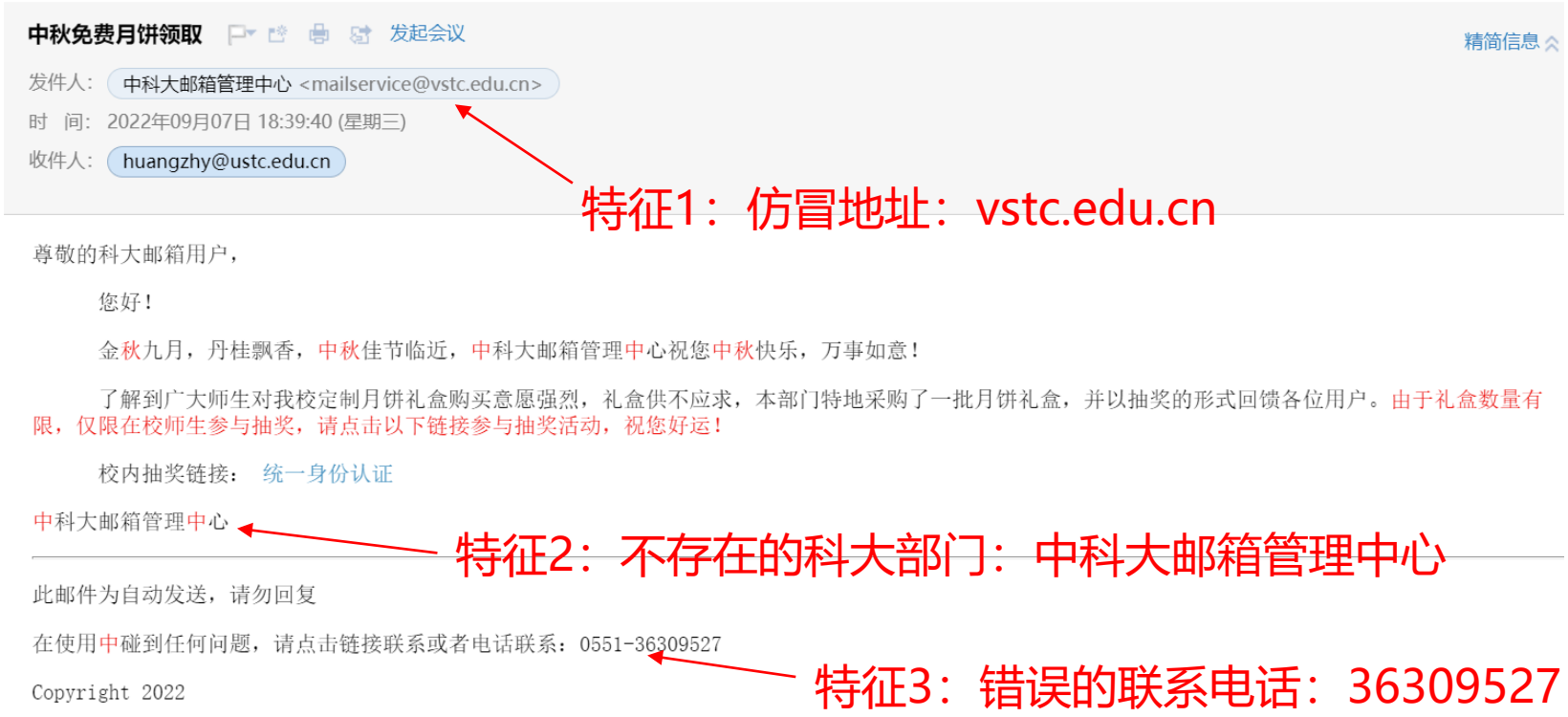
不足：人工定义规则、效率低，难以处理复杂问题

↓

自动生成规则？决策树

分类：规则方法

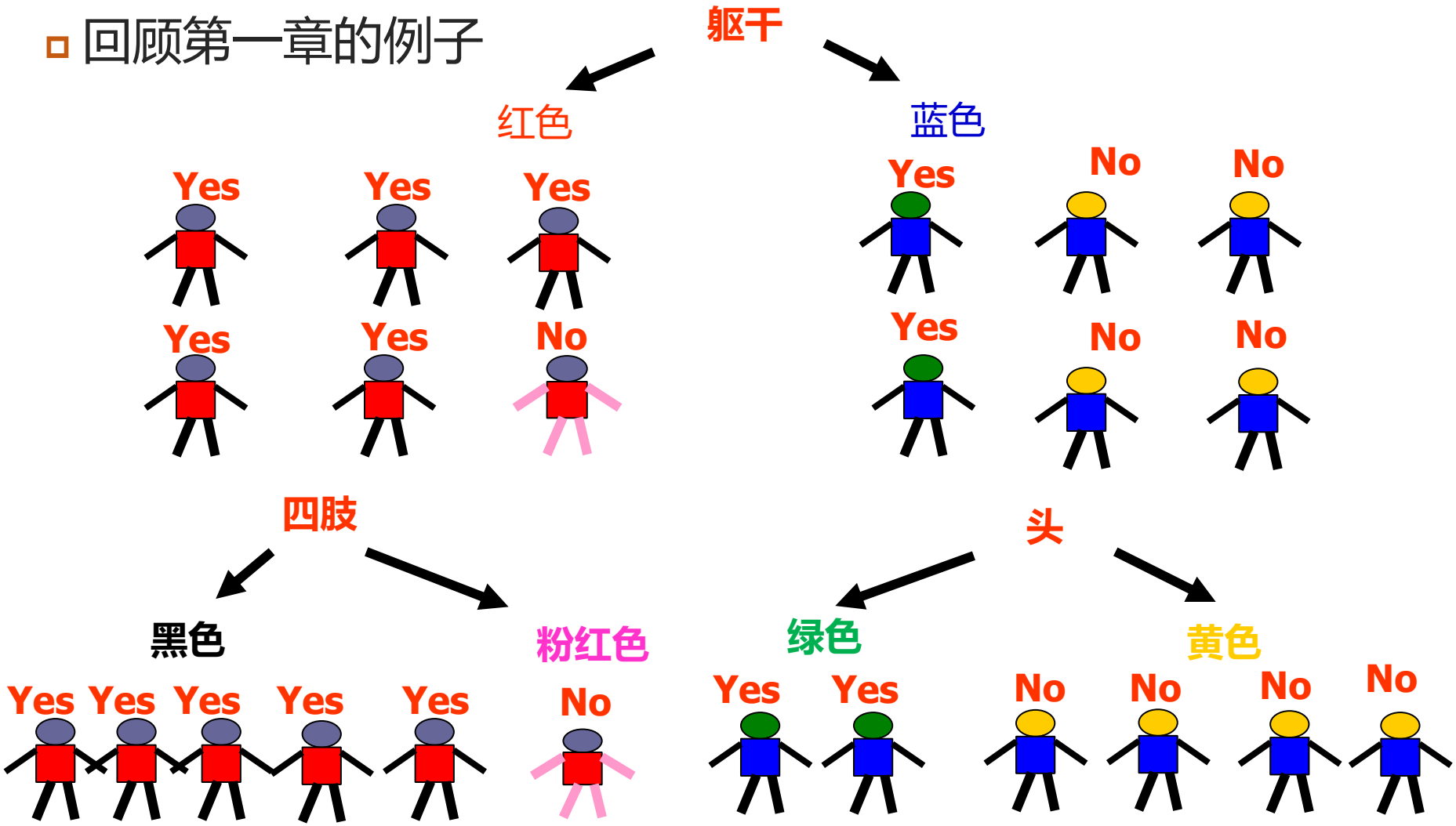
- 回顾垃圾邮件分类的例子
  - 判断：下面这封邮件是垃圾邮件吗？



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

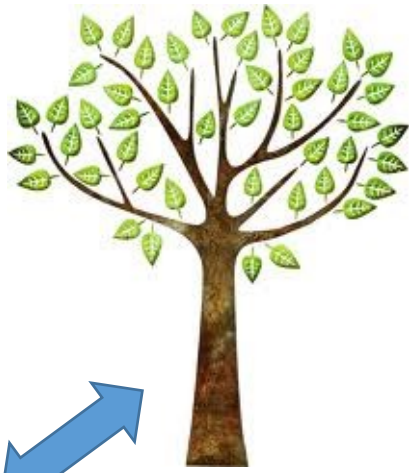
分类：决策树

- 回顾第一章的例子



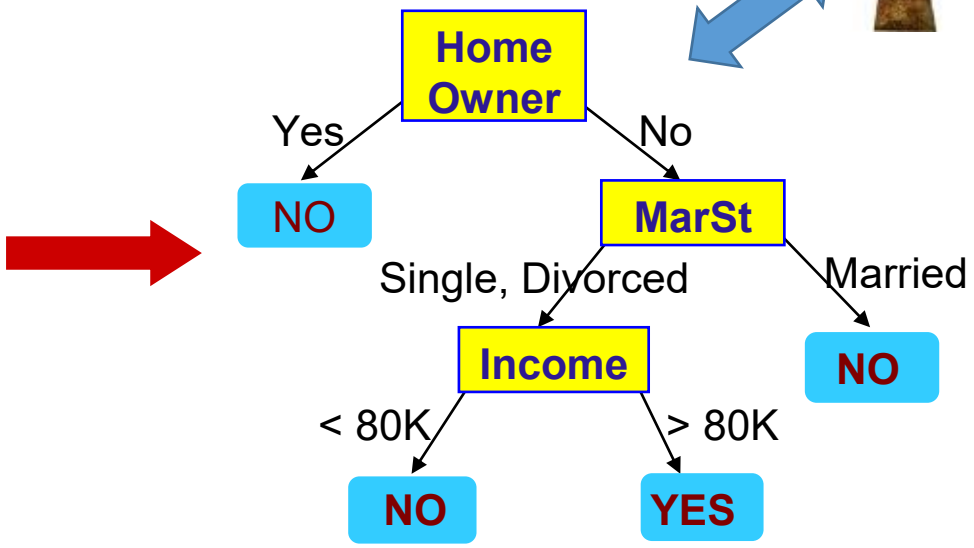
分类：决策树

- 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

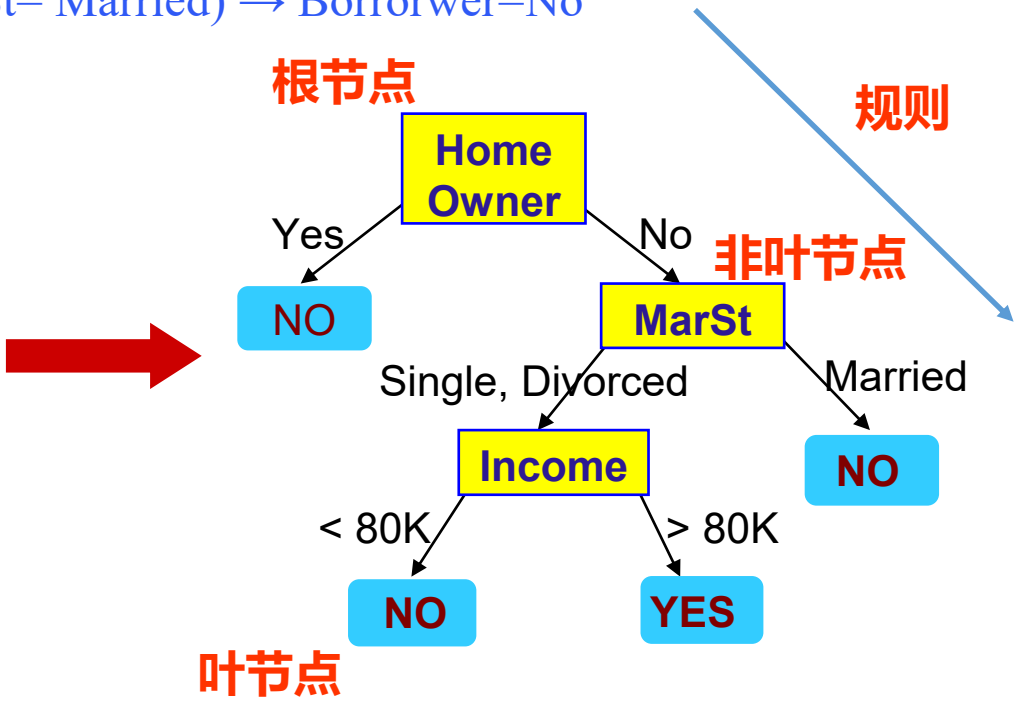


模型：决策树

分类：决策树

- 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
    - (HomOwn= No) ∧ (MarSt= Married) → Borrorwer=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

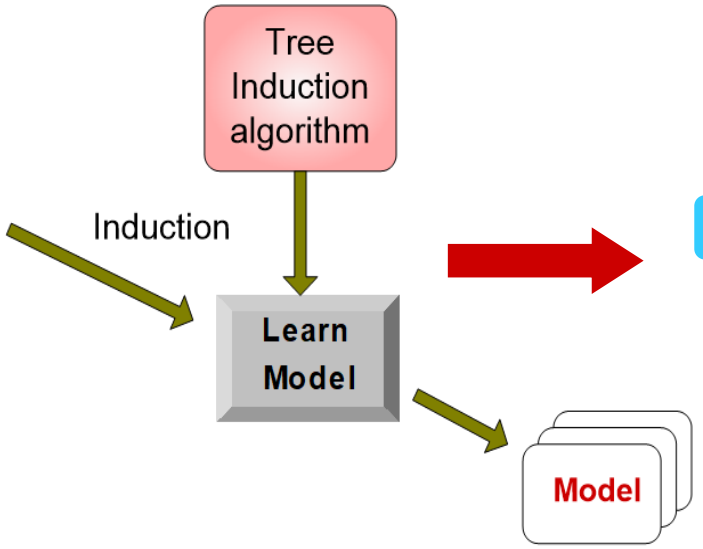


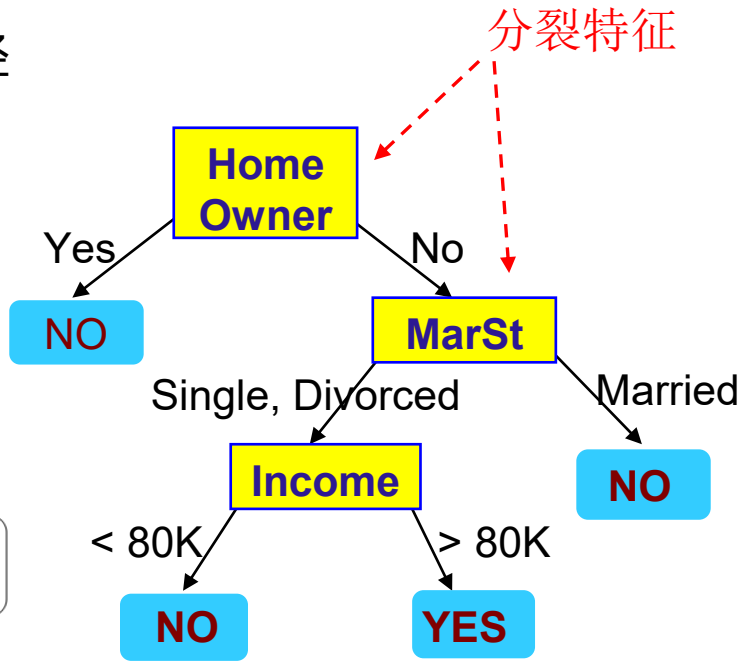
分类：决策树

- 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





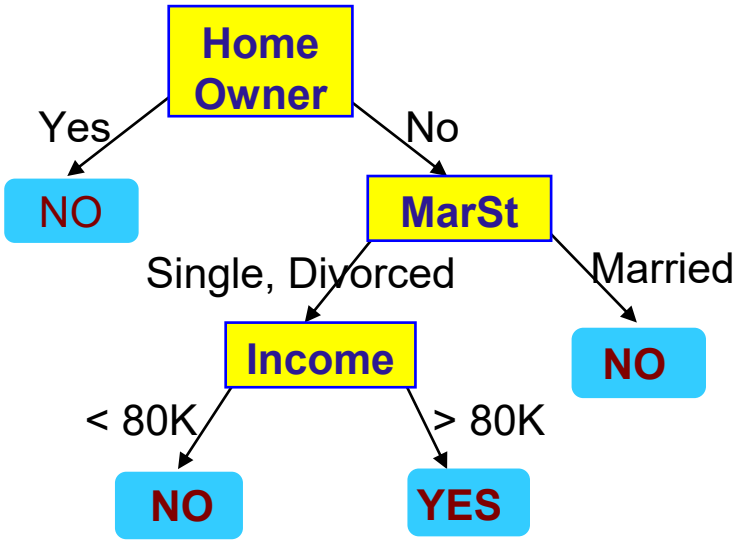
分裂特征:Home Owner, MarSt, Income

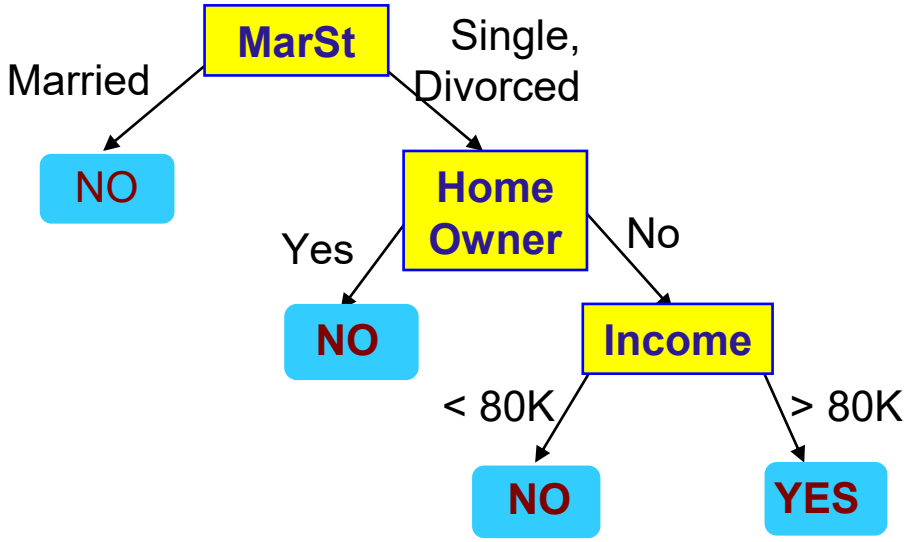
生成模型：决策树

分类：决策树

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

- 是否有其他决策树?





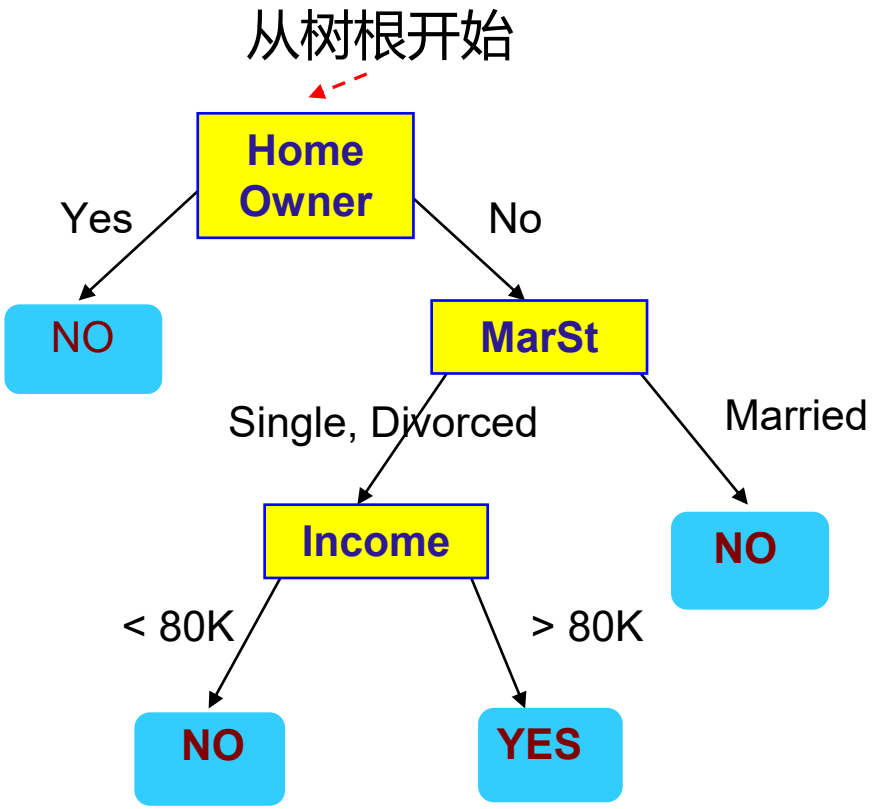
特征顺序: Home Owner, MarSt, Income

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

- 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）

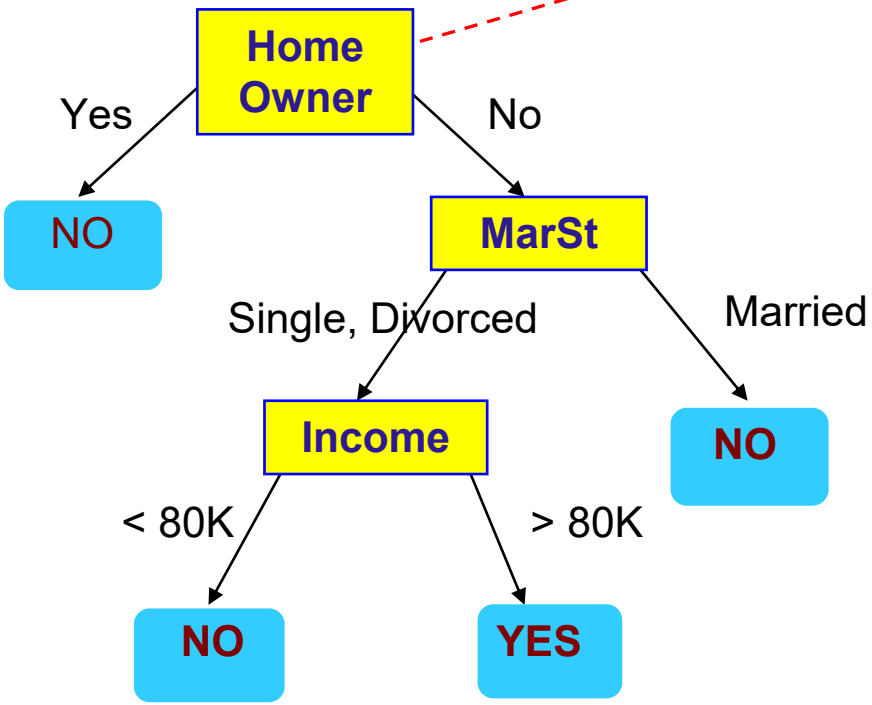
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?

分类：决策树

- 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	?

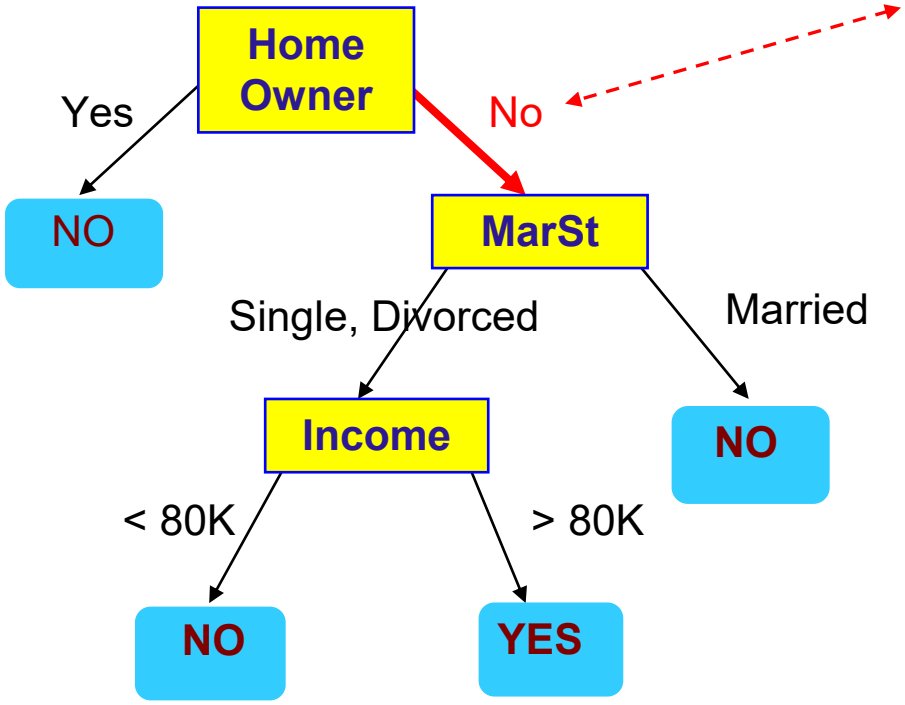


分类：决策树

- 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	?

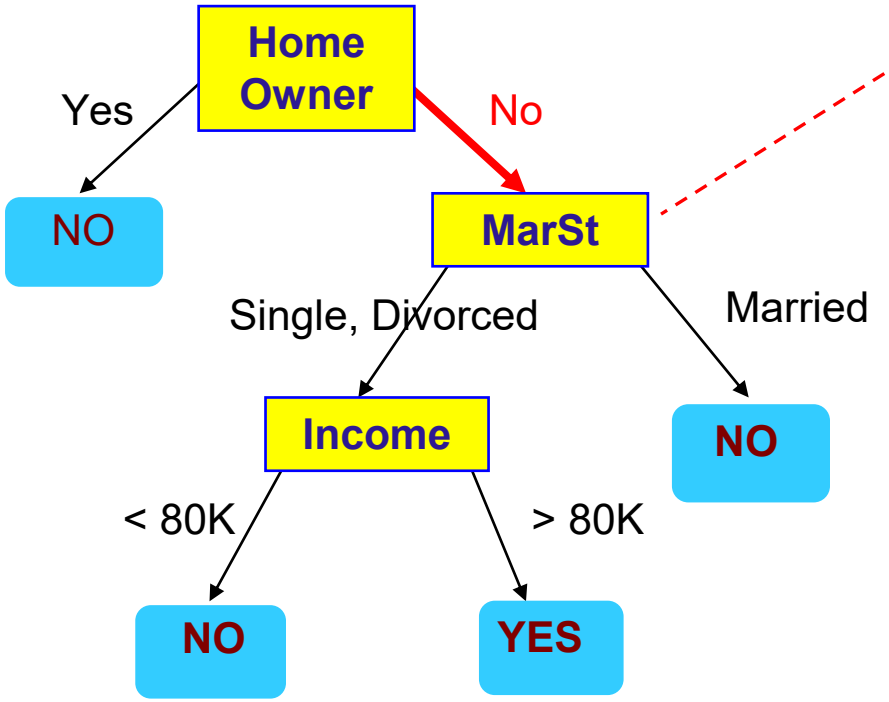


分类：决策树

- 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	?

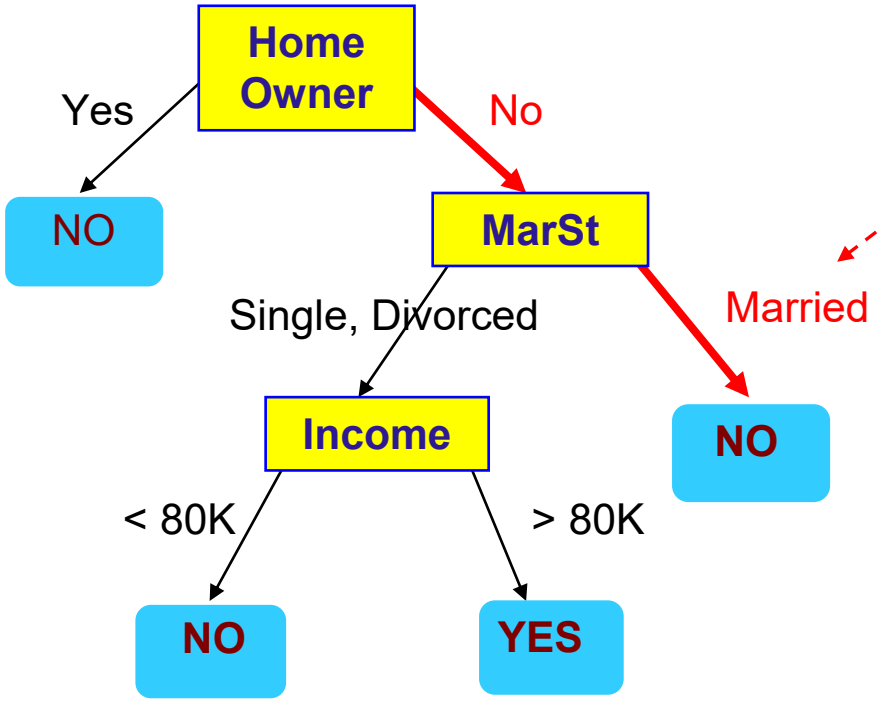


分类：决策树

- 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	?

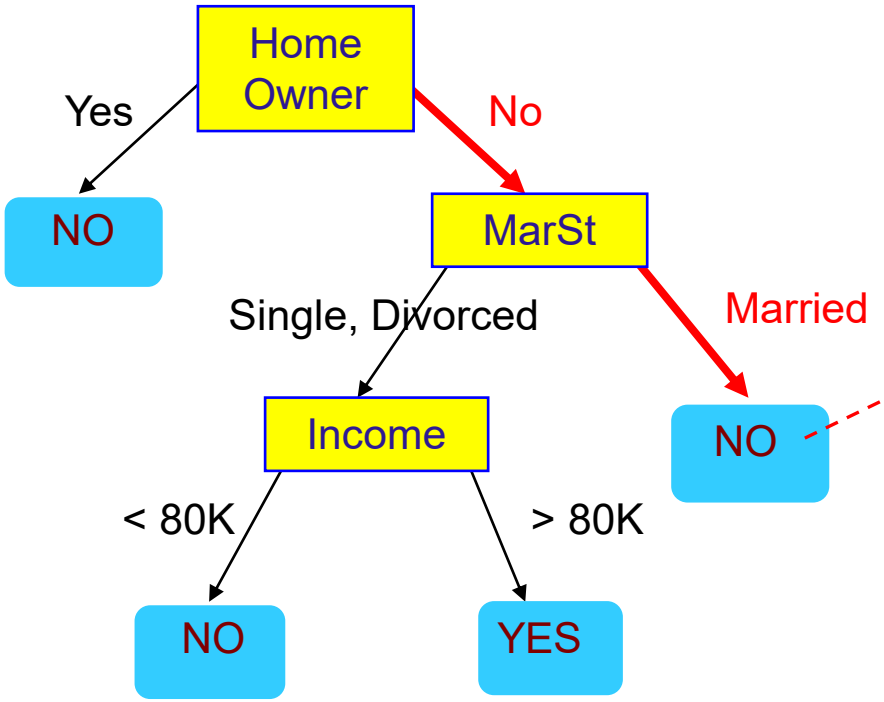


分类：决策树

- 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	?



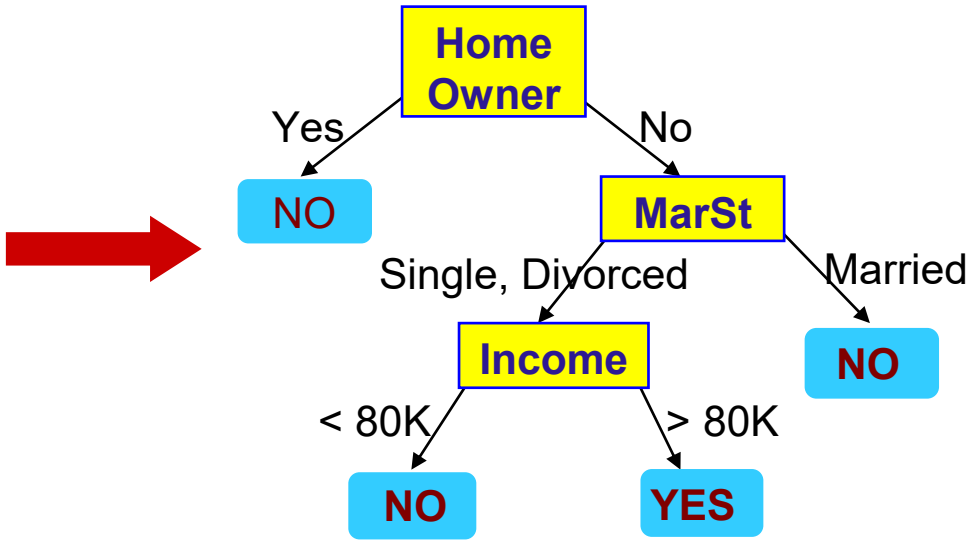
类别Defaulted Borrower 为 “No”

决策过程：未使用所有的特征/属性

分类：决策树

- 如何建立决策树?
  - 基本的决策树学习过程，可以归纳为以下三个步骤：
  - **1. 特征选择：** 选取对于训练数据有着较强区分能力的特征
  - **2. 生成决策树：** 基于选定的特征，逐步生成完整的决策树
  - **3. 决策树剪枝：** 简化部分枝干，避免过拟合因素影响

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

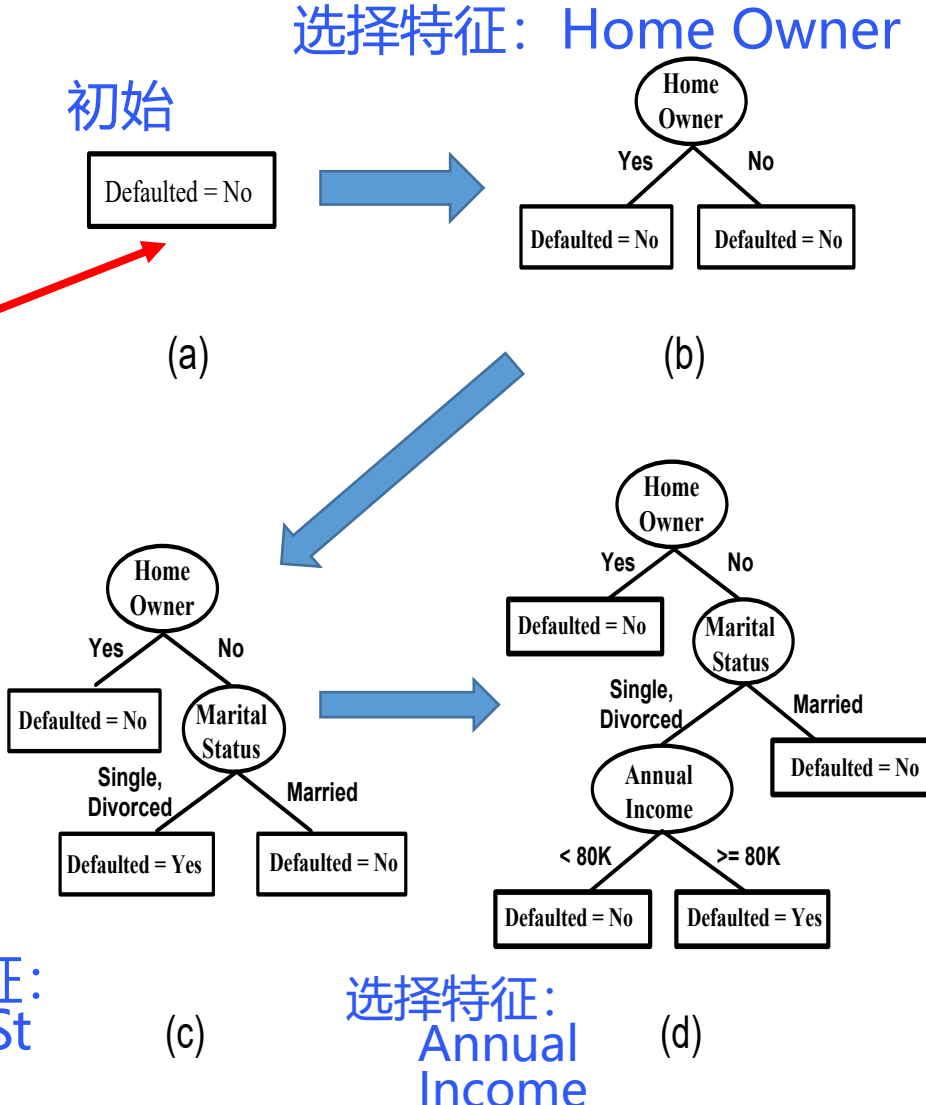


分类：决策树

1. 特征选择

- 决策树的基本思想：特征分裂
- 初始节点包含所有的数据样本，我们希望这些样本能划分到同个类里，如defaulted=No
- 但往往不成立，因此通过选择特征和取值，将样本集合不断划分

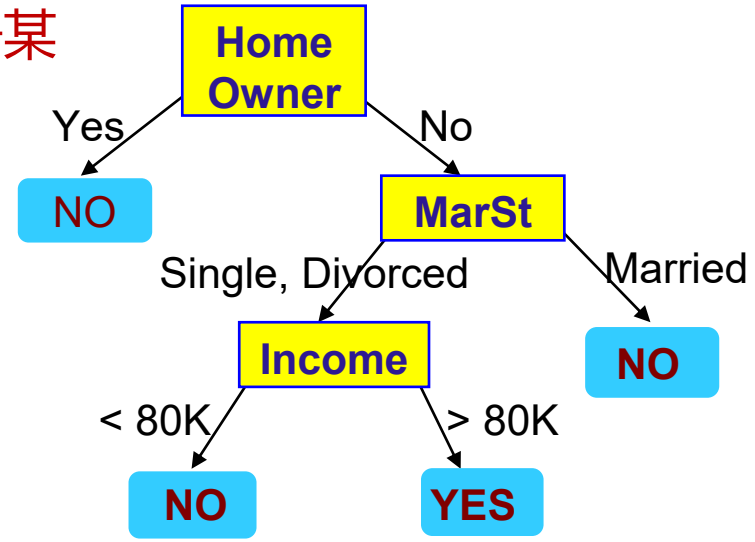
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



分类：决策树

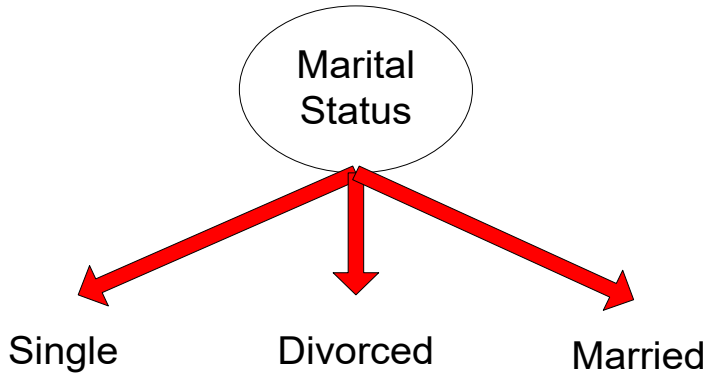
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

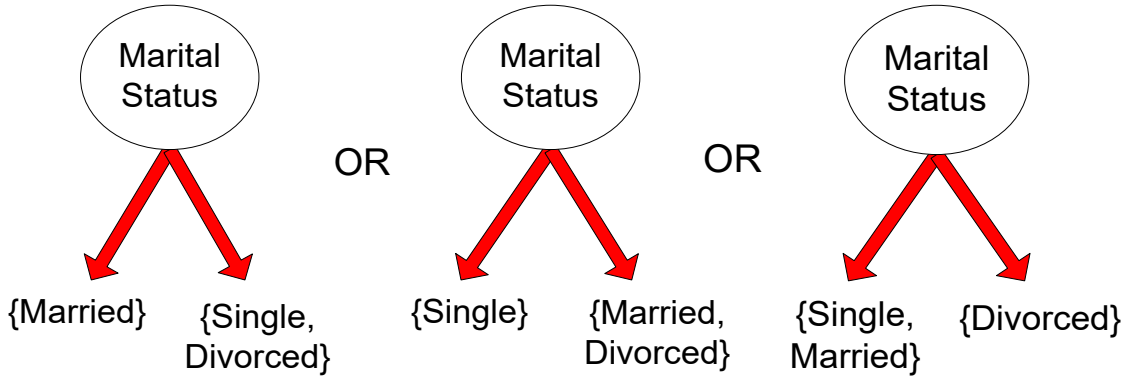
- **1. 特征选择：** 分裂思想的形式化
- Hunt贪心算法
  - D_t：为树中结点t的所有训练样本
  - 若D_t 中的样本属于同一类别y_t, 则t 作为叶子节点，标签为y_t
  - 若D_t 中的样本不属于同一类别, 则根据某特征将D_t 分为更小的样本集
  - 重复上述过程



分类：决策树

- **1. 特征选择：** 分裂过程的形式
  - 多路分裂(Multi-way split)
    - 不同取值均作为一个子集
  - 二分裂(Binary split)
    - 只划分两个子集
    - 需找到最优划分方法





分类：决策树

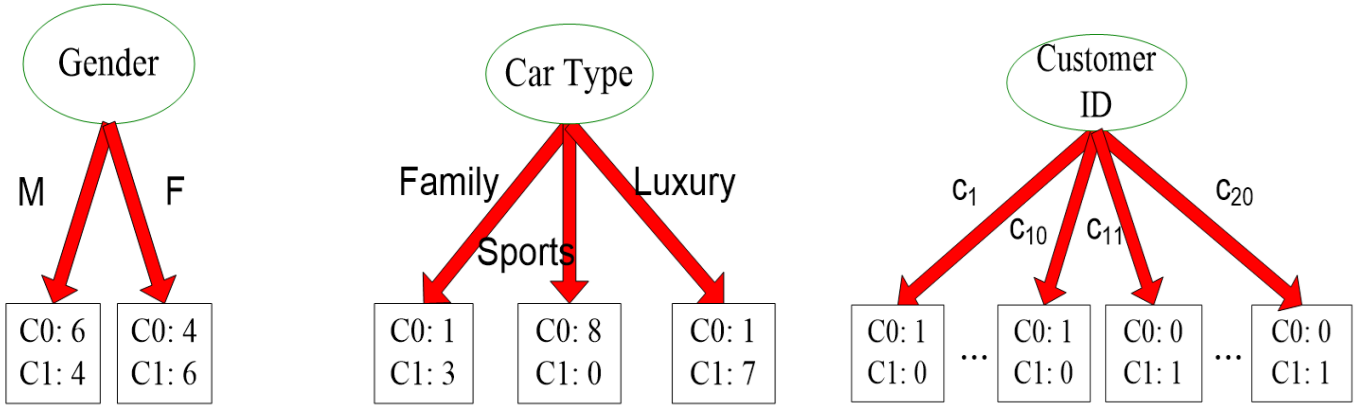
- **1. 特征选择：**分裂过程的两个问题
  - 训练样本如何分裂?
    - 选择分裂特征
    - 评价测试条件
  - 分裂过程何时停止?
    - 理想终止
      - 如果所有记录属于同一类
      - 所有数据有相同的属性值
    - 提前终止

决策树特征选择

- 问题1：如何选择分裂特征？
  - 例子：右图的数据
    - 10个记录类别为0
    - 10个记录类别为1

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

对不同特征属性进行分裂



哪一种分裂方式最优？

决策树特征选择

- 问题1：如何选择分裂特征？
  - 目标：选取对于训练数据有着**较强区分能力**的特征
    - 如果某特征分类的结果与随机结果没有很大的差别，则称这个特征是没有分类能力的，扔掉这样的特征对学习的精度影响不大
  - 常用特征选择准则
    - 信息增益 → 回顾第二章：数据离散化
    - 信息增益率
    - 基尼指数