

# 数据预处理

张冬松

信阳学院  
大数据与人工智能学院

[dszhang@nudt.edu.cn](mailto:dszhang@nudt.edu.cn)



# 目录 content



## 第一节

## 选择数据

## 第二节

## 选择属性

## 第三节

## 改变属性



# 第一节 选择数据

- 聚集
- 抽样





## 1.1 聚集

- **定义**：将两个或多个数据对象合并成单个数据对象
  - 例如，按商店及日期对数据聚集
  - 定量属性通过求和或均值进行聚集
  - 定性属性可忽略或聚集成一个集合

表 包含顾客购买信息的数据集

事务ID	商品	商店位置	日期	价格	...
...	...	...	...	...	
101123	Watch	Chicago	09/06/04	\$25.99	...
101124	Battery	Chicago	09/06/04	\$5.99	...
101125	Shoes	Minneapolis	09/06/04	\$75.00	...
...	...	...	...	...	



## 1.1 聚集

- 请对下列数据集按商店位置及日期对数据聚集

事务ID	商品	商店位置	日期	价格
101122	Watch	Chicago	09/06/03	\$25.99
101123	Watch	Chicago	09/06/04	\$25.99
101124	Battery	Chicago	09/06/04	\$5.99
101125	Shoes	Minneapolis	09/06/04	\$75.00
101126	Shoes	Minneapolis	09/06/05	\$75.00
101127	Hat	Minneapolis	09/06/05	\$16.00
101128	Battery	Chicago	09/06/06	\$5.99



## 1.1 聚集

- 请对下列数据集按商店位置及日期对数据聚集

事务ID	商品	商店位置	日期	价格
001	Watch	Chicago	09/06/03	\$25.99
002	{Watch, Battery}	Chicago	09/06/04	\$31.98
003	Shoes	Minneapolis	09/06/04	\$75.00
004	{Shoes, Hat}	Minneapolis	09/06/05	\$91.00
005	Battery	Chicago	09/06/06	\$5.99

# 1.1 聚集

## • 目的

- 数据归约, 减小数据规模
- 范围或标度转换, 将低层数据视图转换为高层数据视图
- 对象或属性群体的行为通常比单个对象或属性的行为稳定

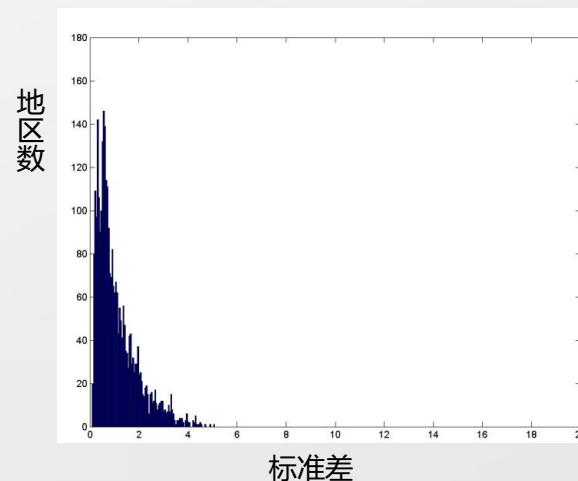
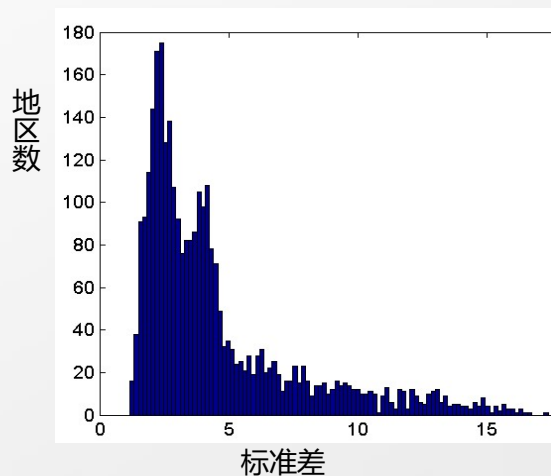


图 澳大利亚从1982年到1993年月和年降水量标准差的直方图



# 第一节 选择数据

- 聚集
- 抽样







## 1.2 抽样

■ **定义：**抽样是一种选择数据对象子集进行分析的常用方法

- **统计学：**避免**获取**整个数据集的高代价
- **数据挖掘：**避免**处理**整个数据集的高代价

■ **基本原理：**

- 如果样本是有代表性的，则使用样本与使用整个数据集的效果几乎是一样的
- **样本是有代表性的**，如果它近似地具有与原数据集相同的（感兴趣的）性质



## 1.2 抽样

### ■ 抽样方法

➤ **简单随机抽样：**每个对象以相同的概率被选择

□ **无放回抽样：**对象被选中时立即从总体中删除

□ **有放回抽样：**对象被选中时不从总体中删除

□ 当样本与数据集相比较小时，两种方法产生的样本差别不大，但放回抽样较为简单



## 1.2 抽样

### ■ 抽样方法

➤ **分层抽样：**现将数据进行划分，然后针对不同的集合进行随机采样。

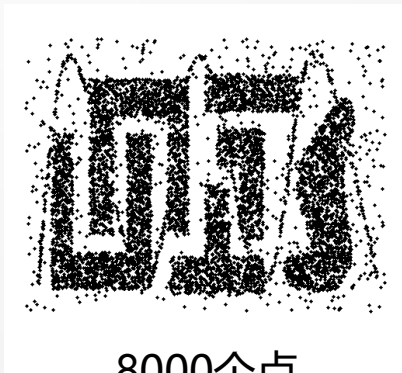
□ **适用情况：**总体由不同类型的数据对象组成，每种类型的对象数量差别很大时

□ 从每组抽取相同数量的数据对象或抽取数量与组大小成正比

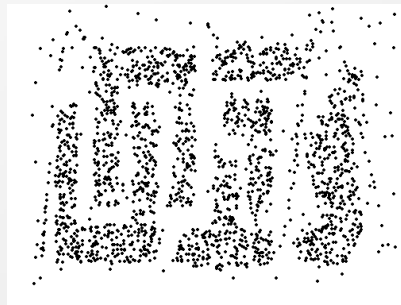
## 1.2 抽样

### ■选择多大的样本容量？

- 较大的样本容量，保留信息多，数据规模大
- 较小的样本容量，数据规模小，可能损失信息



8000个点



2000个点

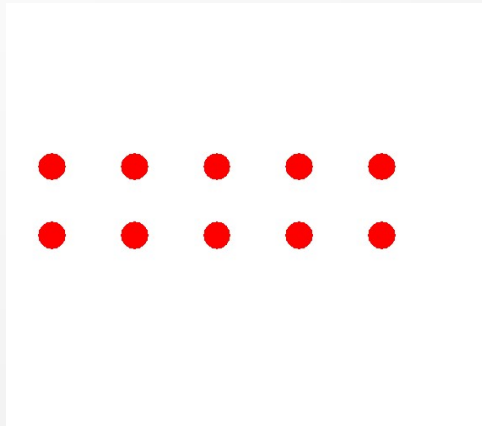


500个点

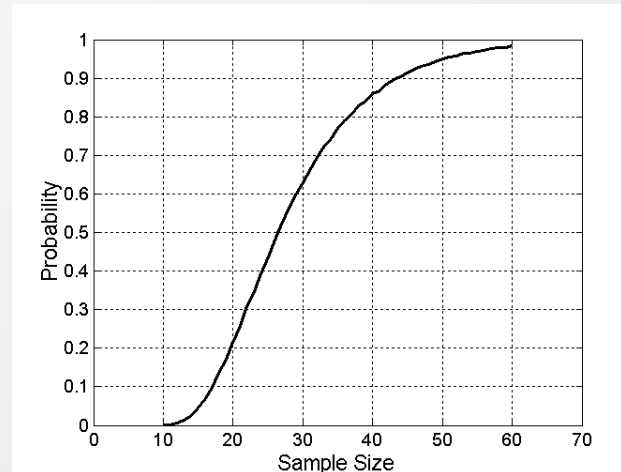
## 1.2 抽样

### ■合适的样本容量

- 数据集分为若干个大致相等的组，每个组的数据对象之间高度相似，不同组中的对象不太相似，每个组中至少一个代表点被选出



(a) 点的10个组



(b) 样本包含所有10个组中点的概率



## 1.2 抽样

### ■ 渐进抽样

- 从一个小样本开始，然后增加样本容量直至得到足够容量的样本
- **停止条件：**用抽样的样本训练模型，当模型性能的提高趋于稳定的时候，停止增加样本容量



## 1.2 抽样

### 问题

给定 $m$ 个对象的集合，这些对象划分成 $K$ 组，其中第 $i$ 组的大小为 $m_i$ 。

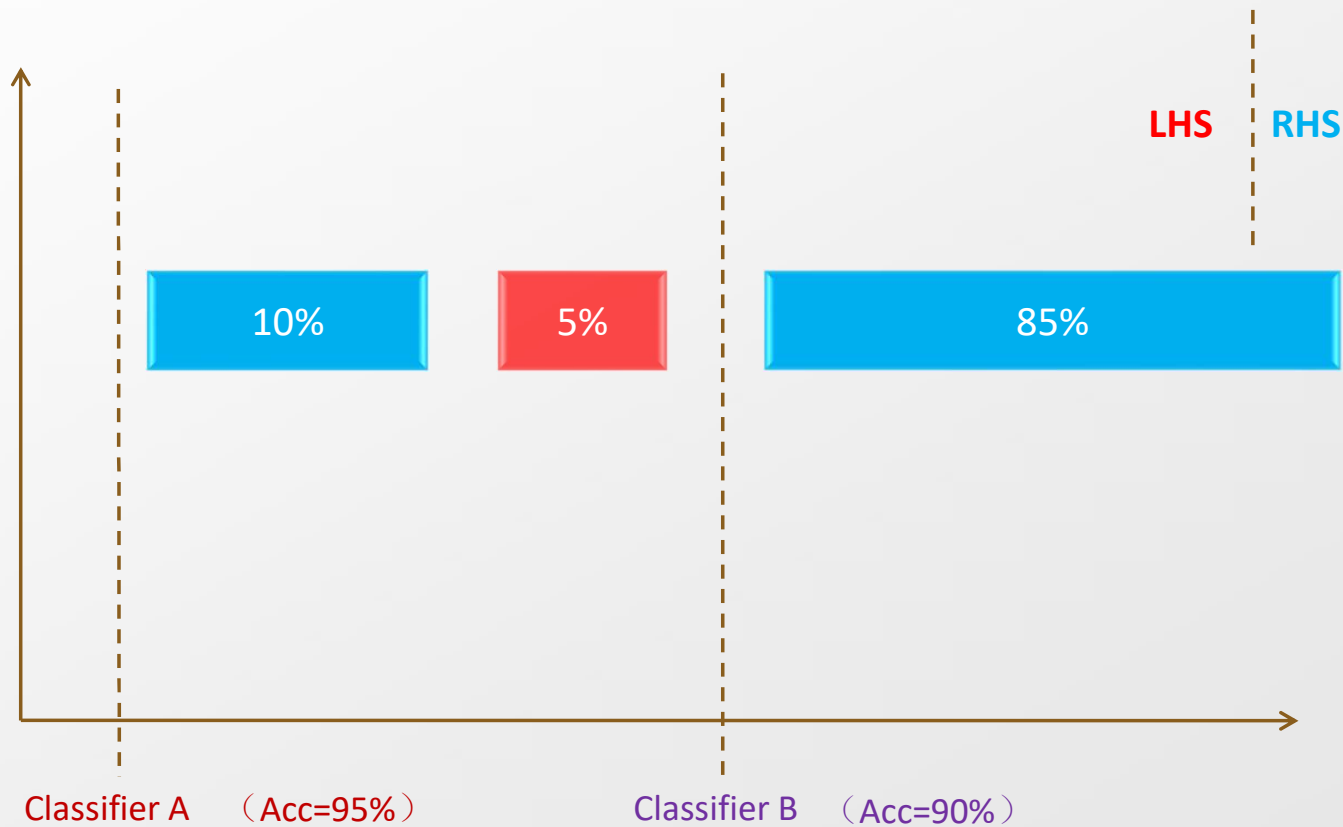
如果目标是得到容量为 $n < m$ 的样本，下面两种抽样的方案有什么区别？

- 1) 从每组随机地选择 $n \cdot m_i / m$ 个元素，而不管对象属于哪个组。
- 2) 从数据集中随机的选择 $n$ 个元素，而不管对象属于哪个组。

## 1.2 抽样

### A. 不平衡数据集

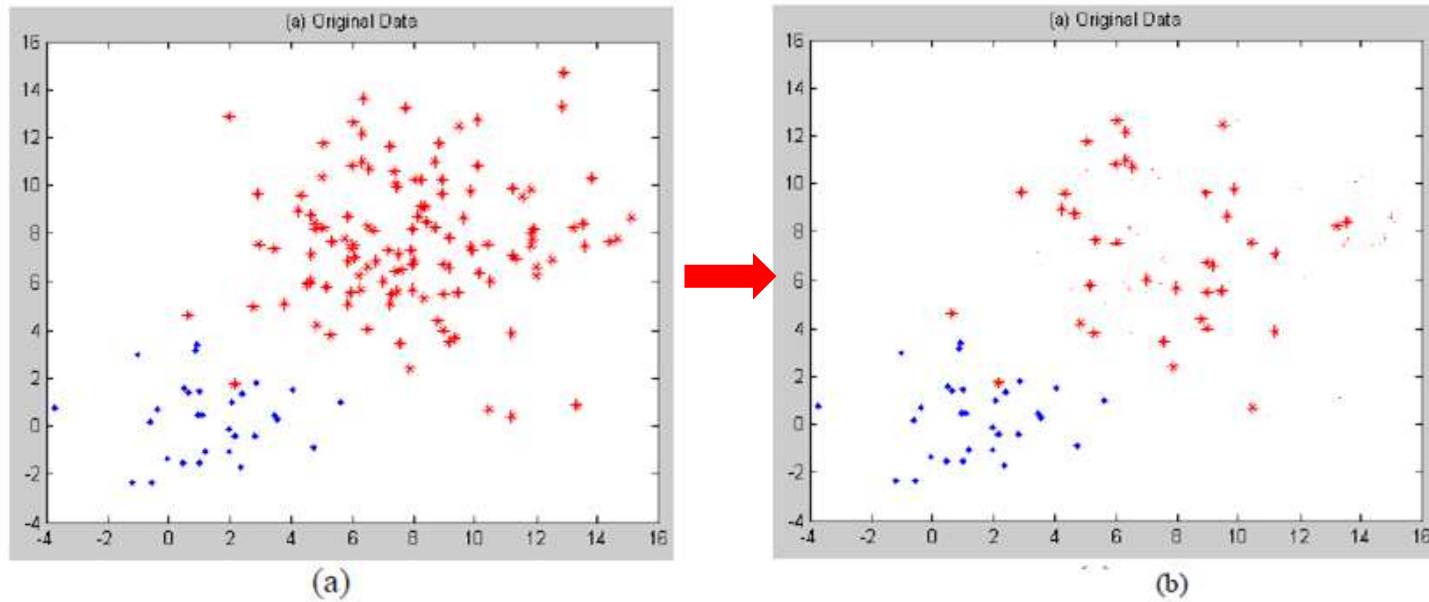
■ 不平衡数据集指不同类别的数据实例数目差别较大的数据集





## 1.2 抽样

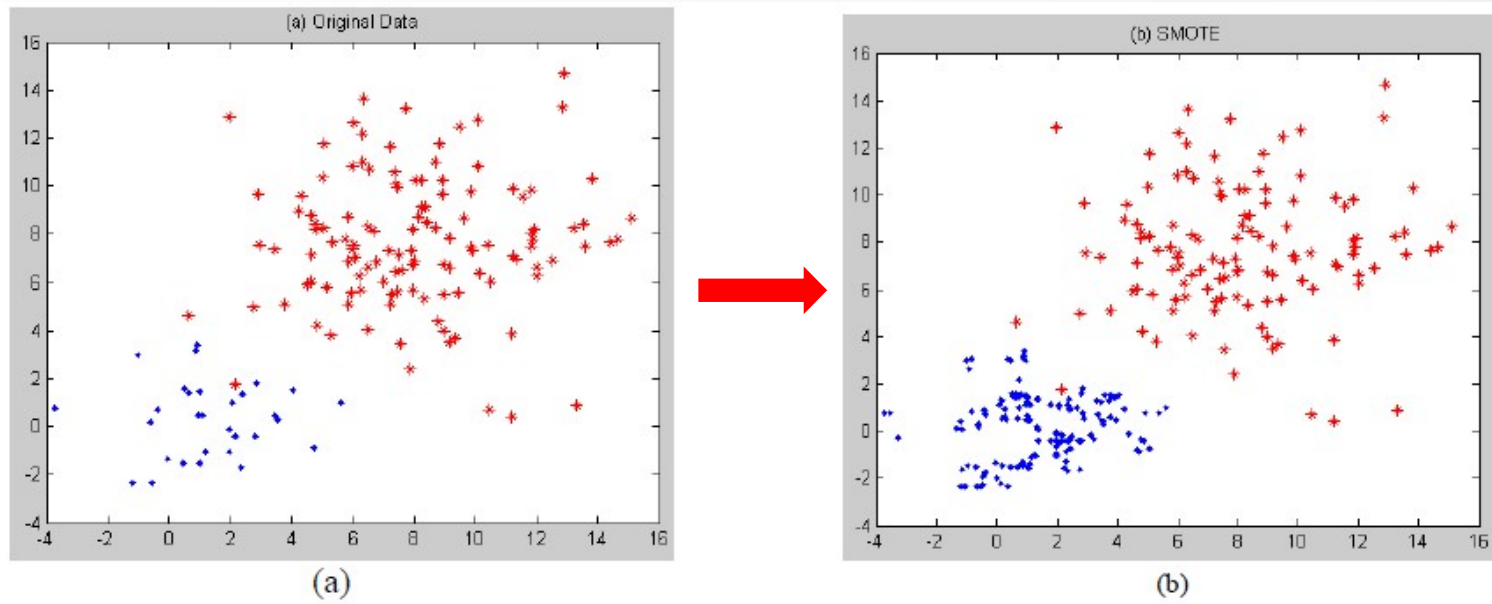
### B. 欠采样



从少数类不变，从多数类中随机选择一部分样例，使得筛选后的多数类与少数类的样例比例接近1:1

## 1.2 抽样

### C.过采样



从多数类不变，从复制少数类中的样本直至与多数类中样本数目比例接近1：1



## 1.2 抽样

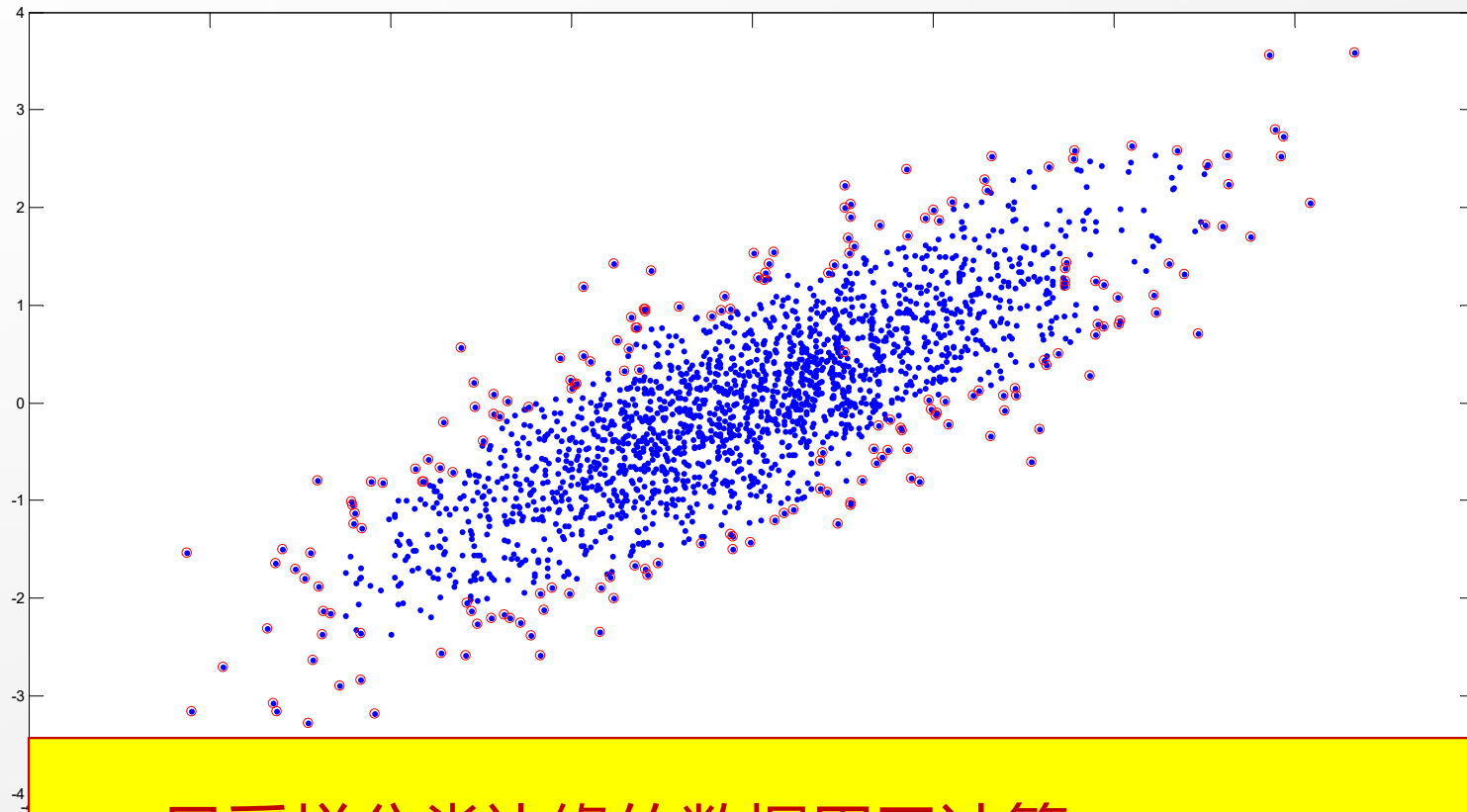
### ■ 欠采样 vs. 过采样

- 欠采样：丢失数据
- 过采样：引入冗余数据

如何改进？

## 1.2 抽样

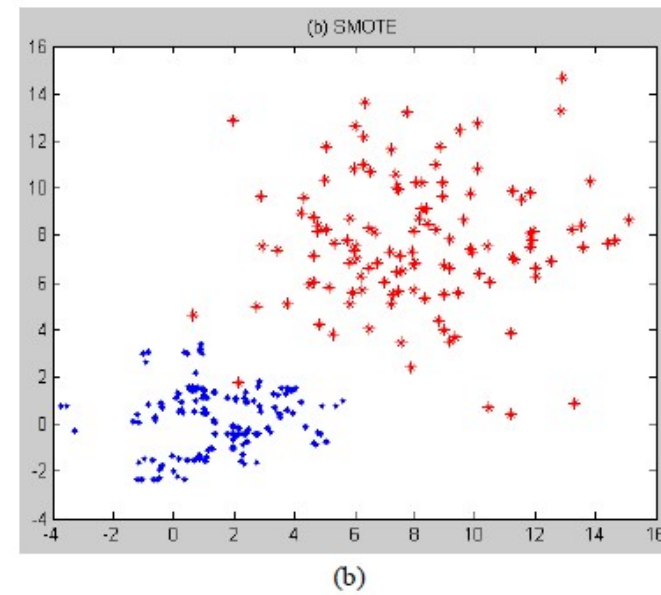
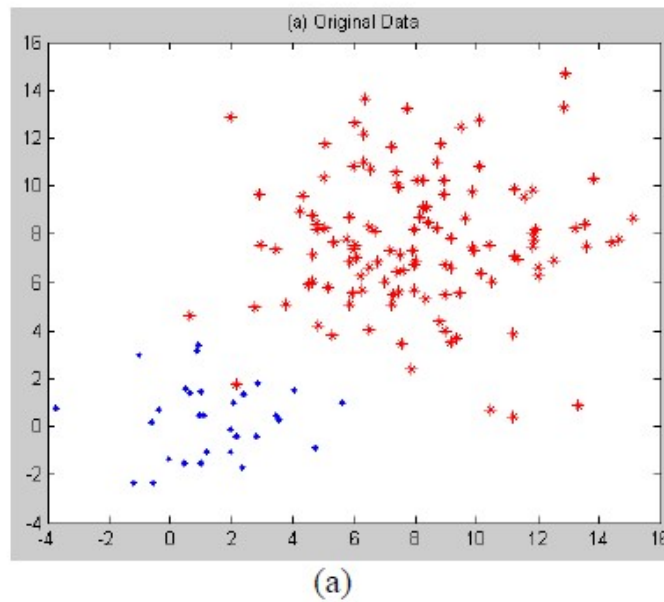
### 边缘采样



只采样分类边缘的数据用于计算。

## 1.2 抽样

### 基于插值的过采样



在少数类中选取两个临近的实例，在其之间的区域内随机生成新样例。



## 第二节

## 选择属性

- 维归约（降维）
- 特征子集选择





## 2.1 维规约（降维）

■**定义：**通过创建新属性，将一些旧属性合并在一起降低数据集的维度

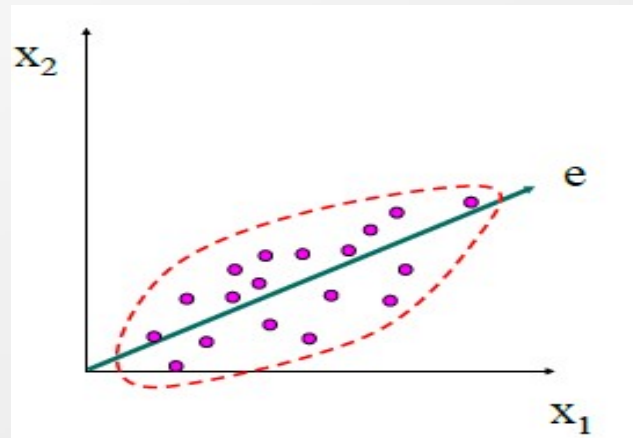
■**目的：**

- 避免维度灾难
- 降低算法复杂度
- 使数据更容易可视化
- 有助于去掉无关属性或降低维度

## 2.1 维规约（降维）

### ■常用技术：

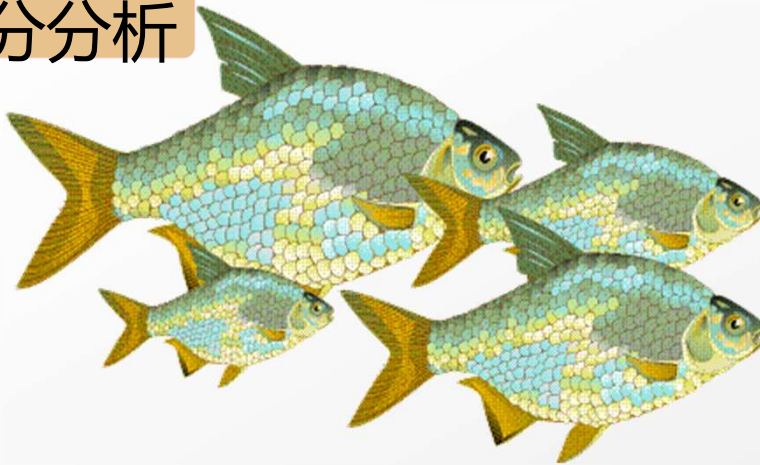
- 主成分分析（Principal Component Analysis, PCA）
  - 将数据映射到数据损失最小的方向
  - 找到协方差矩阵的特征值，最大特征值对应的方向为主成分方向



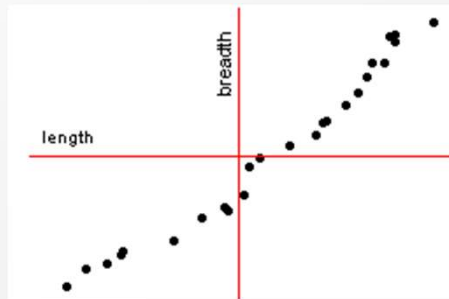
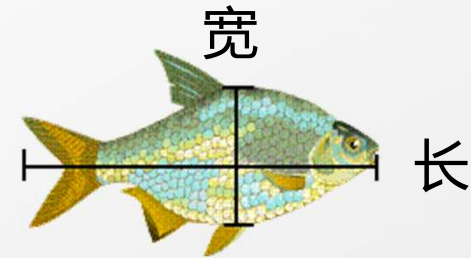


## 2.1 维规约 (降维)

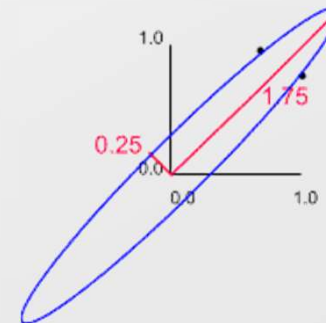
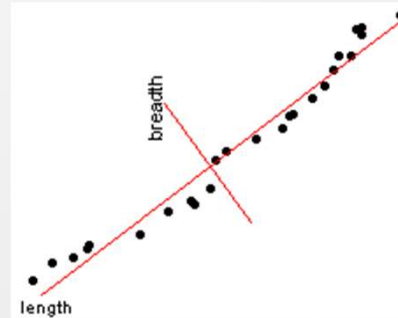
### 主成分分析



同一种类的鱼群



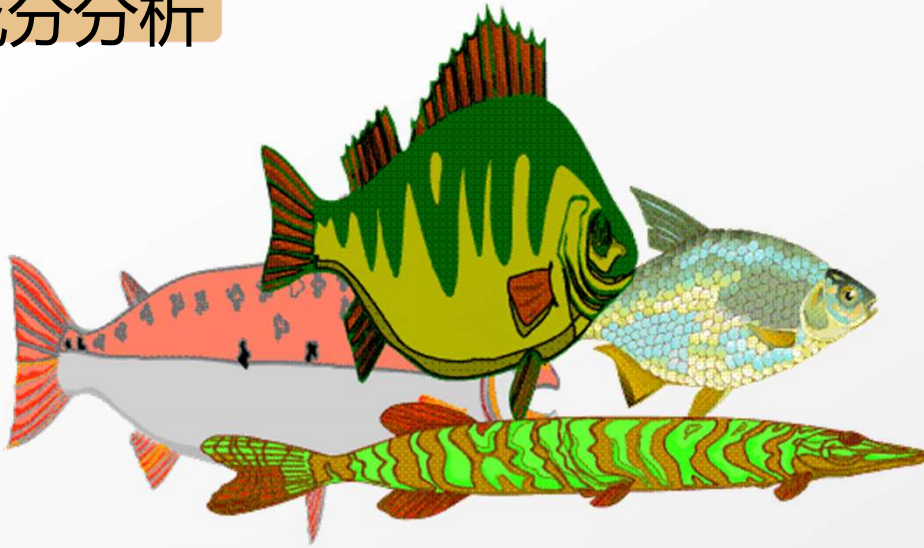
数据分布



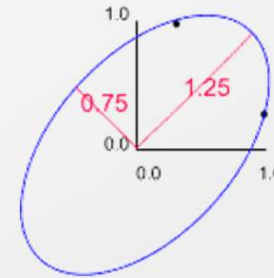
PCA投影

## 2.1 维规约（降维）

### 主成分分析



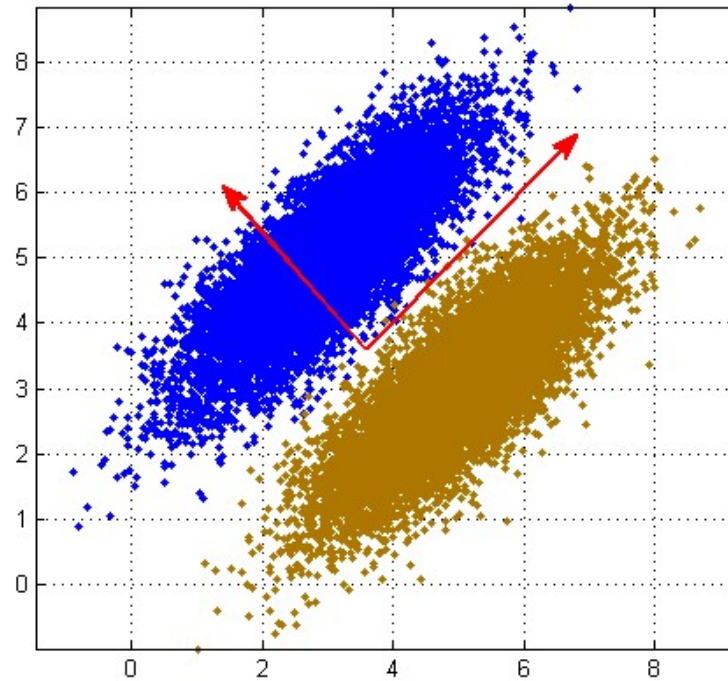
不同种类的鱼群



1. “长”和“宽”相关性减小
2. 方差差异也减小
3. PCA投影信息损失变大

## 2.1 维规约 (降维)

### 主成分分析

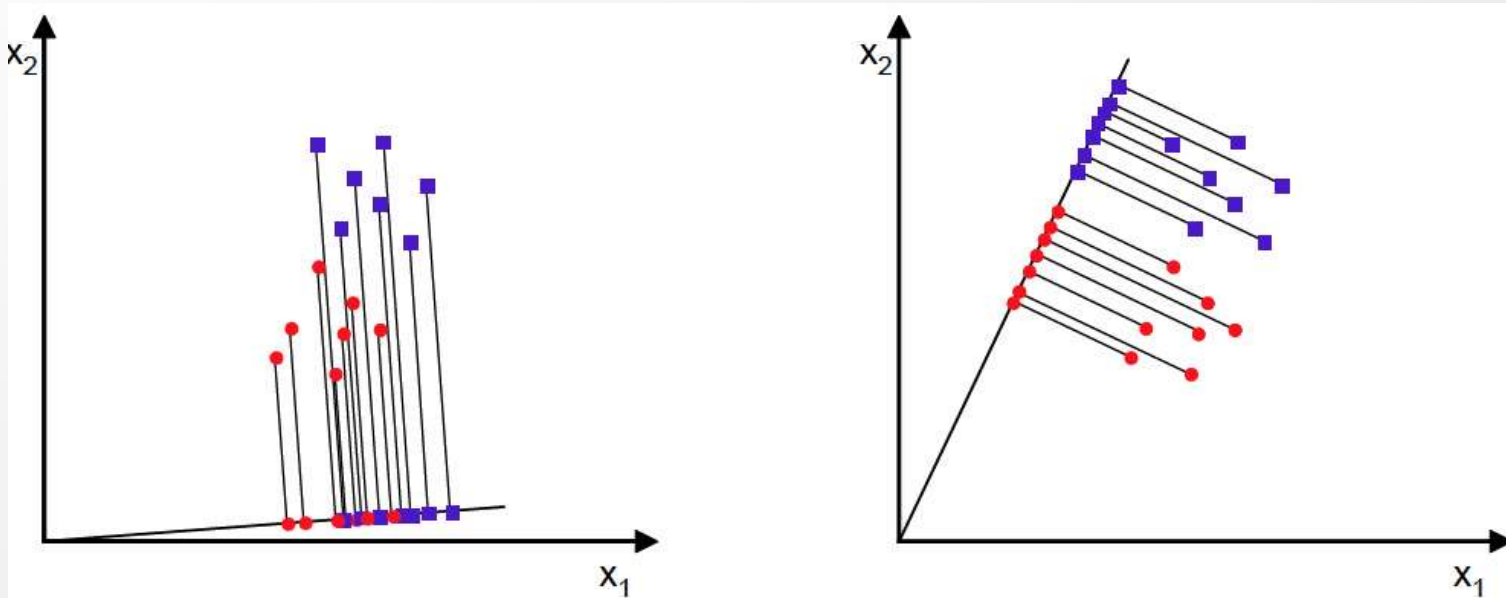


加入类别信息.....

## 2.1 维规约 (降维)

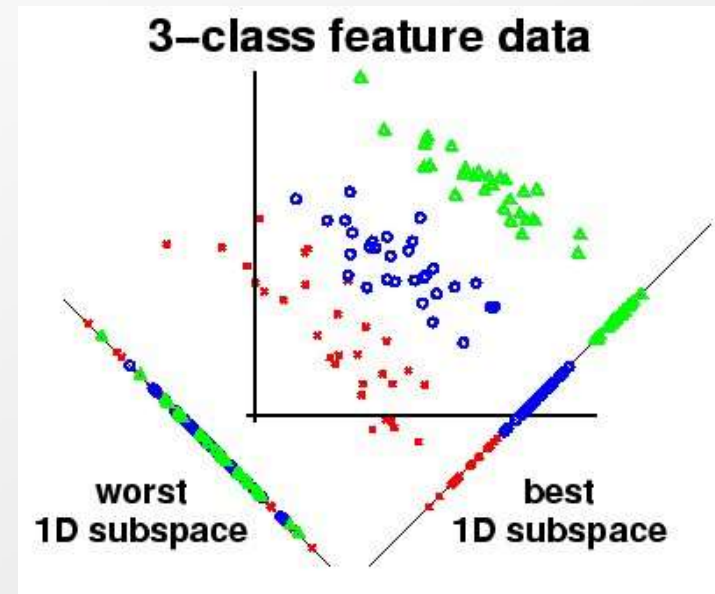
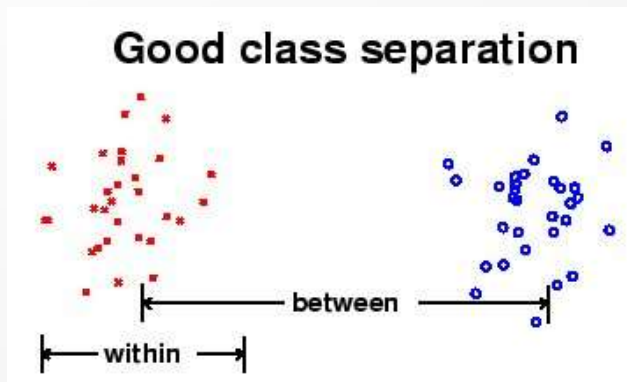
### 线性判别分析

- 有监督的降维方法
- 找到一个投影方向
- 不同类别的实例在该方向上的投影能最大程度分开



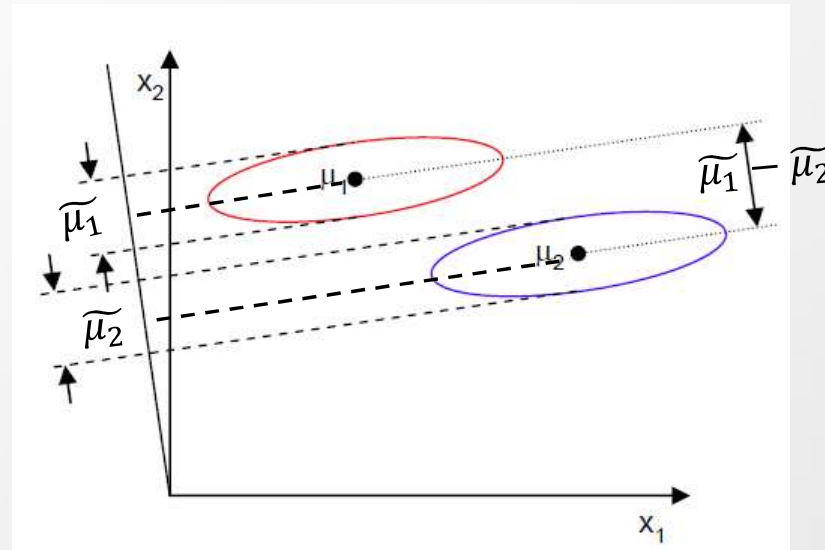
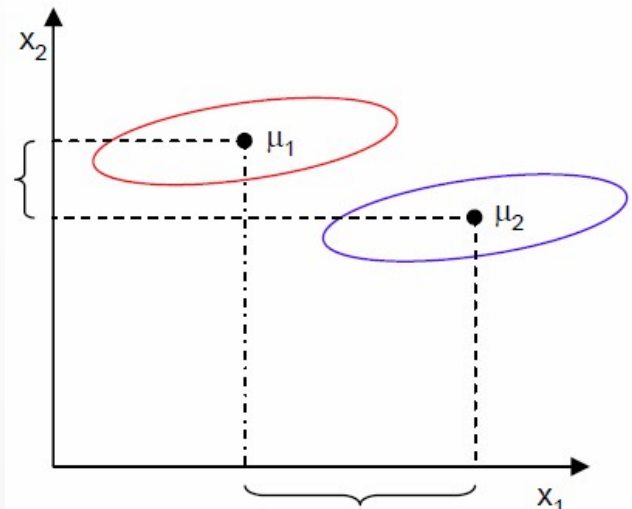
## 2.1 维规约 (降维)

如何度量不同类别实例的区分程度?



## 2.1 维规约 (降维)

### 线性判别分析



Fisher 准则:

$$J = \frac{|\widetilde{\mu}_1 - \widetilde{\mu}_2|^2}{S_1^2 + S_2^2}$$



最大化类之间的距离



最小化同类实例的散度，即每个实例到均值点的距离之和



## 第二节

## 选择属性

- 维归约
- 特征子集选择





## 2.2 特征子集选择

- **冗余特征**

- 重复了包含在一个或多个其他属性中的许多或所有信息
- 例如，一种产品的购买价格和所支付的销售税额包含许多相同的信息

- **不相关特征**

- 包含对于当前数据挖掘任务几乎完全没用的信息
- 例如，学生的ID号码对于预测学生的总平均成绩是不相关的





## 2.2 特征子集选择

### ■理想方法

1. 找到所有特征子集
2. 分别作为数据挖掘算法的输入
3. 选择算法性能最好的特征子集

假设数据集中每条数据记录有 $n$ 个属性，则有多少特征子集？



## 2.2 特征子集选择

- **嵌入方法**

- 由算法本身决定使用哪些属性和忽略哪些属性

- **过滤方法**

- 使用某种独立于数据挖掘任务的方法，在数据挖掘算法运行前进行特征选择

- **包装方法**

- 将目标数据挖掘算法作为黑盒，使用类似理想算法，但通常不枚举所有可能的子集



## 2.2 特征子集选择

■ **嵌入方法**：基于线性模型的特征选择

$$x = \underline{w_0} + \underline{w_1}a_1 + \underline{w_2}a_2 + \cdots + \underline{w_k}a_k$$

递归特征消除 (recursive feature elimination)

1. 建一个模型
2. 根据系数进行特征排序
3. 移除最低位的特征
4. 重复以上操作，直至剩余特征数目达到阈值

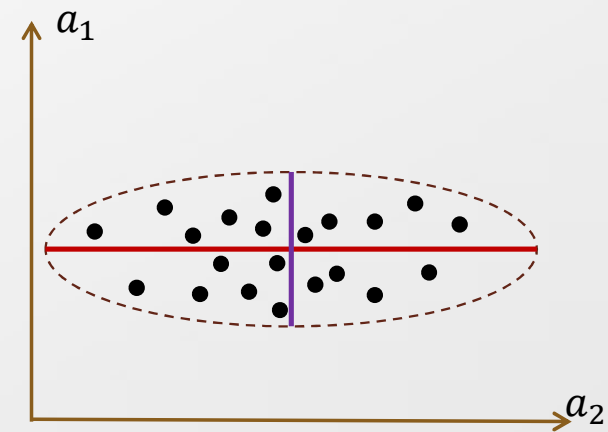
## 2.2 特征子集选择

### ■过滤方法：基于统计值的方法

➤特征在训练数据中所有取值的方差 $\sigma$

□ $\sigma$ 小，意味着特征在所有实例的取值差别不大，特征的区分能力不强

□ $\sigma$ 小于给定阈值，该特征被过滤





## 2.2 特征子集选择

### ■ 过滤方法局限性

➤ 无法查探出冗余特征

□ 判断：

若A特征和B特征的取值之间存在 $a=2b$ 的关系，A特征与分类密切相关，那么B特征也与分类密切相关



## 第三节 改变属性

- 特征创建
- 离散化和二元化
- 变量变换





## 3.1 特征创建

■**定义：**由原来的属性创建新的属性集，更有效地捕获数据集中的重要的信息

- 特征提取
- 映射数据到新的空间
- 特征构造



## 3.1 特征创建

### ■特征提取

- 由原来数据创建新的特征集
- 例如，人脸识别的图像数据，从图像像素集合构造与人脸高度相关的某些类型的边和区域等



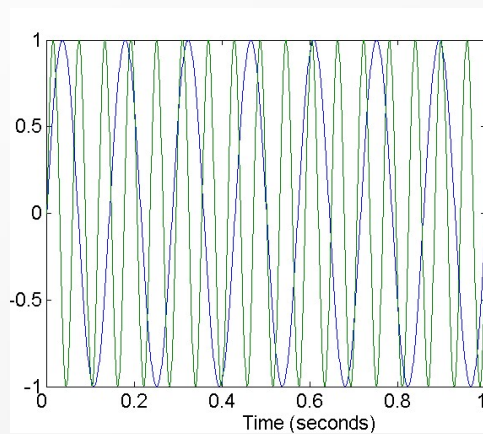


## 3.1 特征创建

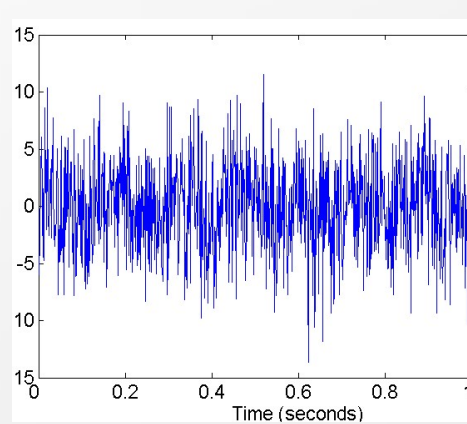
- 映射数据到新的空间
  - 使用一种完全不同的视角挖掘数据
  - 例如，傅里叶变换、小波变换

## 3.1 特征创建

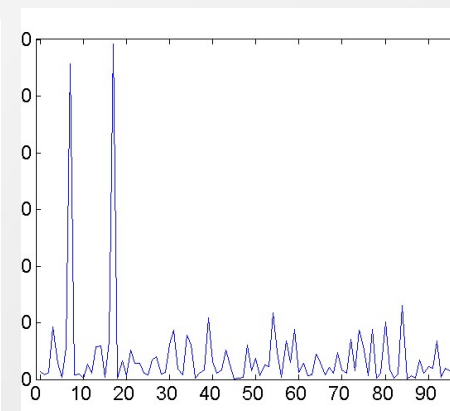
- 傅里叶变换（音频处理）
- 小波变换（图像处理）



(a) 两个时间序列



(b) 噪声时间序列



(c) 功率频谱

傅里叶变换应用：识别时间序列数据中的基本频率



## 第三节 改变属性

- 特征创建
- 离散化和二元化
- 变量变换





## 3.2 离散化和二元化

### ■ 离散化

- 将连续属性变换成分类属性

### ■ 二元化

- 将连续或离散属性变换成一个或多个二元属性



## 3.2 离散化和二值化

### ■二值化

➤关联规则挖掘：关心属性的出现

分类值	整数值	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
Awful	0	1	0	0	0	0
Poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
Good	3	0	0	0	1	0
Great	4	0	0	0	0	1

分类值	整数值	$x_1$	$x_2$	$x_3$
Awful	0	0	0	0
Poor	1	0	0	1
OK	2	0	1	0
Good	3	0	1	1
Great	4	1	0	0



## 3.2 离散化和二元化

### ■二元化

➤多分类问题转换成二分类问题

➤原因：

□有些算法只能处理二分类问题

□针对多分类问题的算法速度很慢或难以实现

➤如何转换：

□将数据集分解为多个二分类问题

□在每个子集上运行学习算法

□输出各个分类器结果的组合

## 3.2 离散化和二值化

### ■ 一对多 (one vs. rest)

➤ 将一个多分类数据集分成针对每个类别的二分类数据集

A1	A2	A3	Class
$a_{11}$	$a_{12}$	$a_{13}$	YES
$a_{21}$	$a_{22}$	$a_{23}$	NO
$a_{31}$	$a_{32}$	$a_{33}$	NO

A1	A2	A3	Class
$a_{11}$	$a_{12}$	$a_{13}$	$C_1$
$a_{21}$	$a_{22}$	$a_{23}$	$C_2$
$a_{31}$	$a_{32}$	$a_{33}$	$C_3$

A1	A2	A3	Class
$a_{11}$	$a_{12}$	$a_{13}$	NO
$a_{21}$	$a_{22}$	$a_{23}$	YES
$a_{31}$	$a_{32}$	$a_{33}$	NO

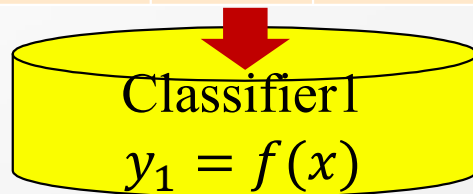
A1	A2	A3	Class
$a_{11}$	$a_{12}$	$a_{13}$	NO
$a_{21}$	$a_{22}$	$a_{23}$	NO
$a_{31}$	$a_{32}$	$a_{33}$	YES

## 3.2 离散化和二元化

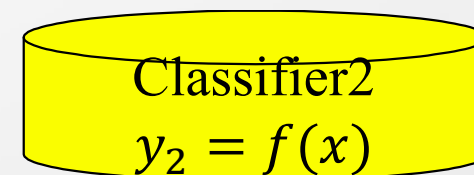
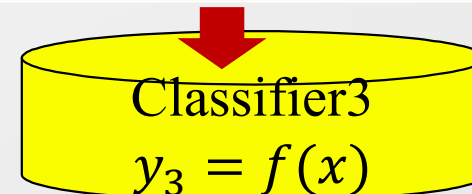
### ■ 一对多 (one vs. rest)

➤ 为二分类数据集构建分类器

A1	A2	A3	Class
$a_{11}$	$a_{12}$	$a_{13}$	YES
$a_{21}$	$a_{22}$	$a_{23}$	NO
$a_{31}$	$a_{32}$	$a_{33}$	NO



A1	A2	A3	Class
$a_{11}$	$a_{12}$	$a_{13}$	NO
$a_{21}$	$a_{22}$	$a_{23}$	NO
$a_{31}$	$a_{32}$	$a_{33}$	YES



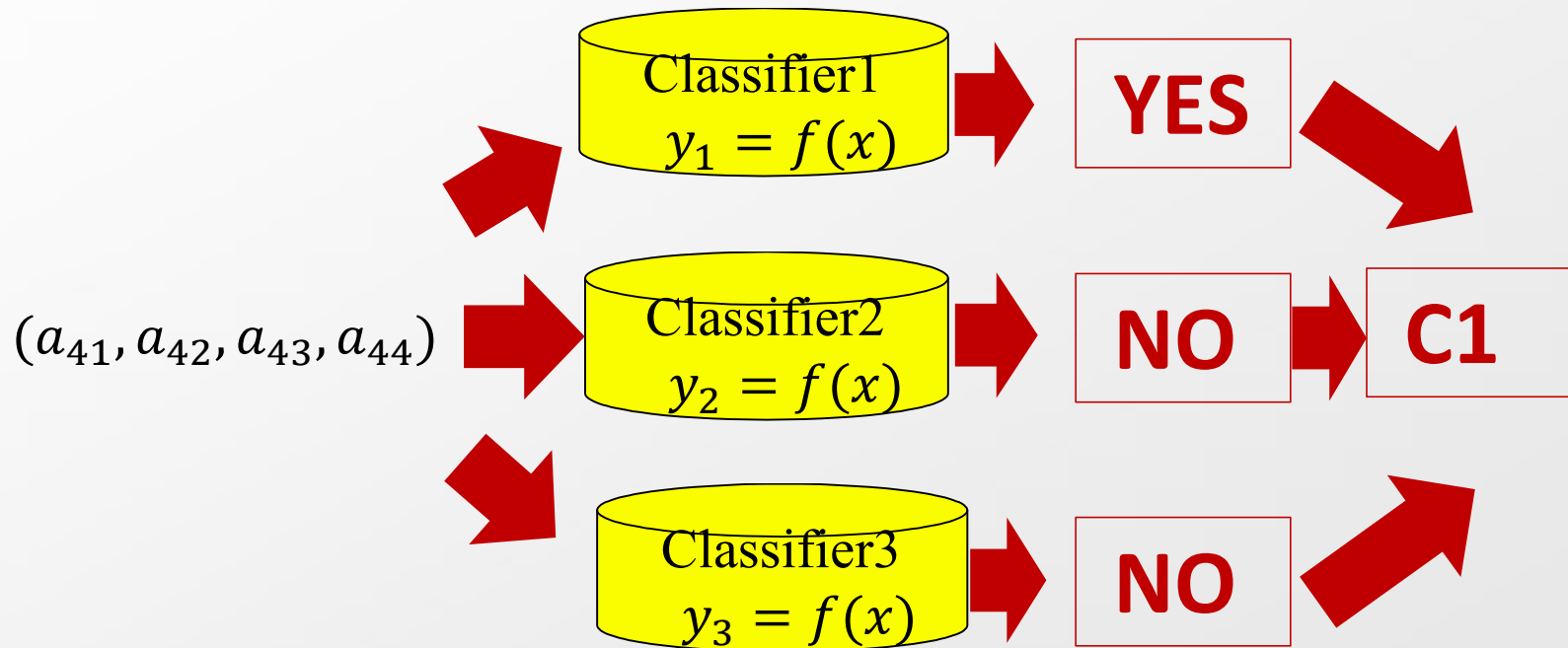
A1	A2	A3	Class
$a_{11}$	$a_{12}$	$a_{13}$	NO
$a_{21}$	$a_{22}$	$a_{23}$	YES
$a_{31}$	$a_{32}$	$a_{33}$	NO



## 3.2 离散化和二值化

### ■ 一对多 (one vs. rest)

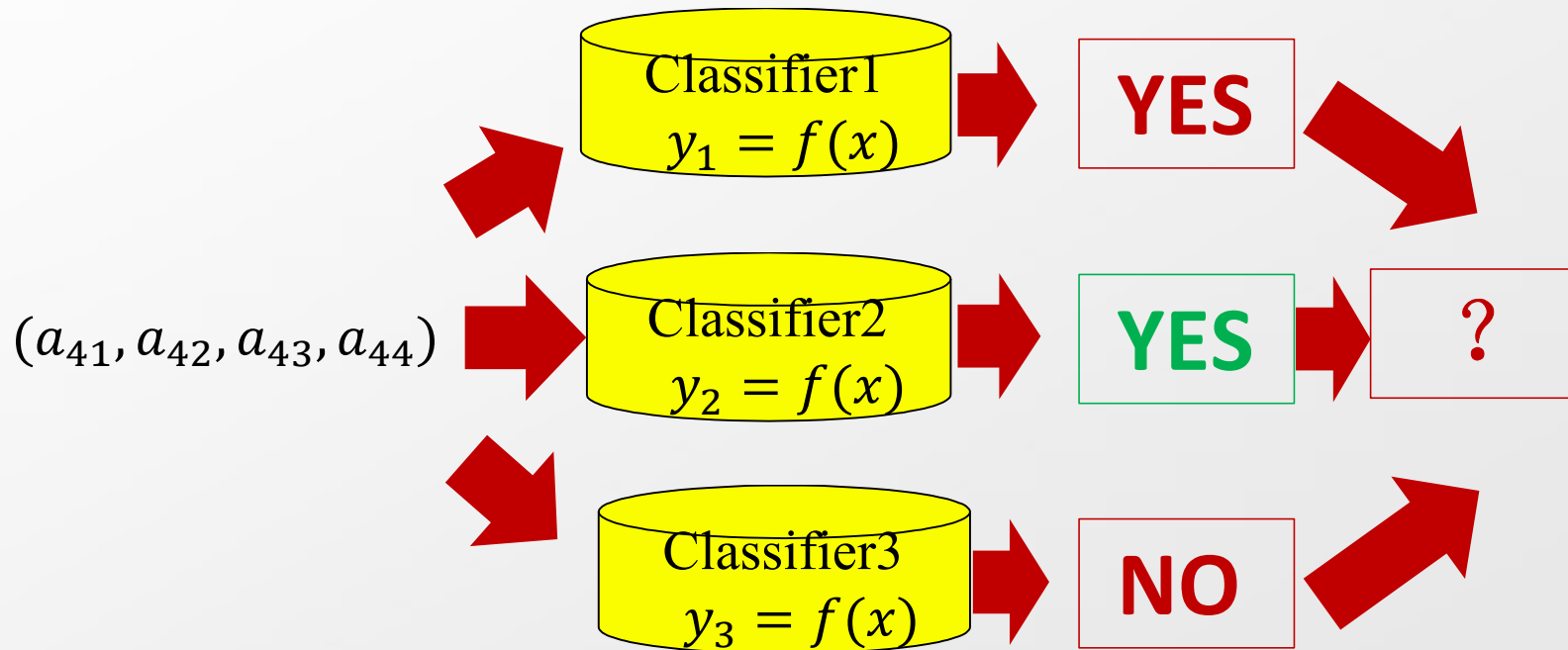
- 用所有分类器对新数据分类，选择YES分数最高的类标



## 3.2 离散化和二元化

### ■ 一对多 (one vs. rest)

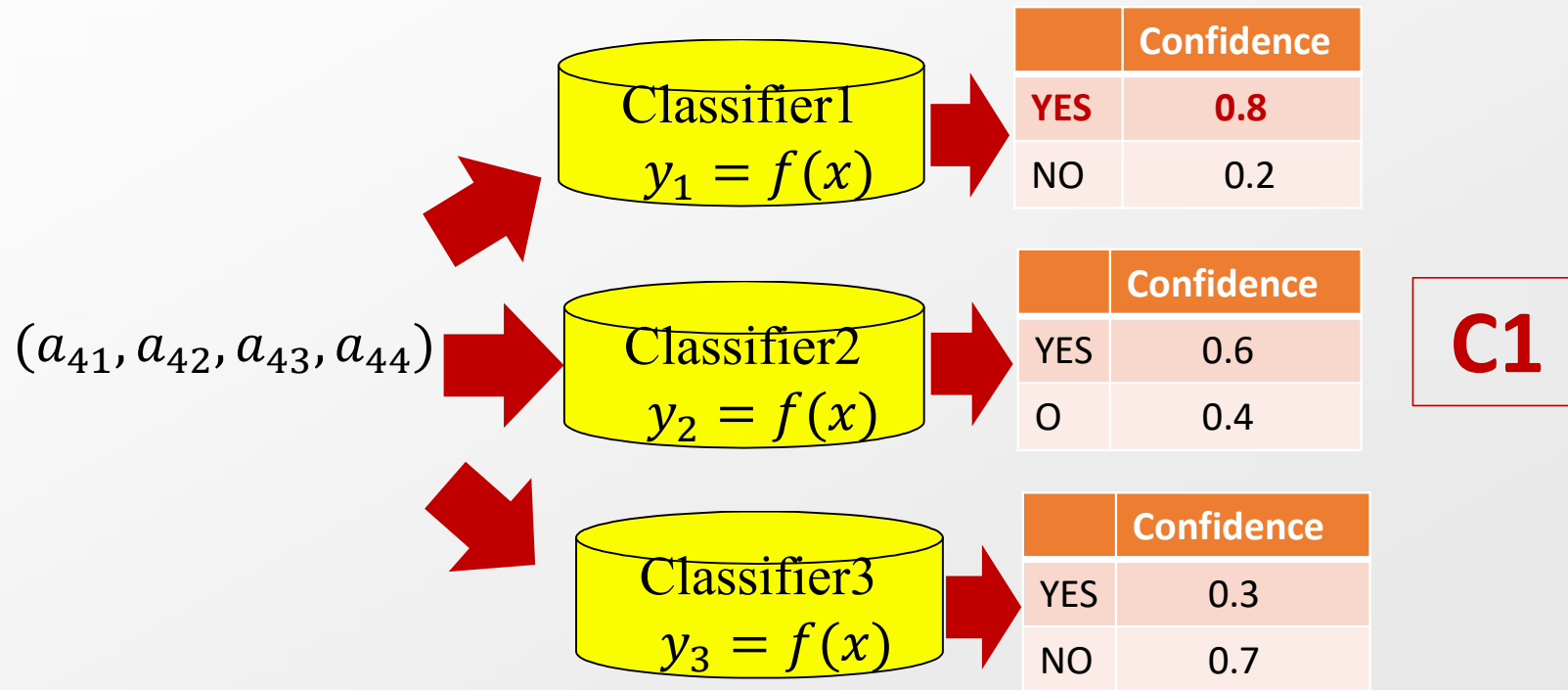
- 用所有分类器对新数据分类，选择YES分数最高的类标



## 3.2 离散化和二元化

### ■ 一对多 (one vs. rest)

➤ 用所有分类器对新数据分类，选择YES分数最高的类标




## 3.2 离散化和二元化


### ■ 成对分类 (pairwise classification)

➤ 将一个多分类数据集分成两两分类的二分类数据集


A1	A2	A3	Class
$a_{11}$	$a_{12}$	$a_{13}$	$C_1$
$a_{21}$	$a_{22}$	$a_{23}$	$C_2$
$a_{31}$	$a_{32}$	$a_{33}$	$C_3$



A1	A2	A3	Class
$a_{11}$	$a_{12}$	$a_{13}$	$C_1$
$a_{21}$	$a_{22}$	$a_{23}$	$C_2$



A1	A2	A3	Class
$a_{11}$	$a_{12}$	$a_{13}$	$C_1$
$a_{31}$	$a_{32}$	$a_{33}$	$C_3$

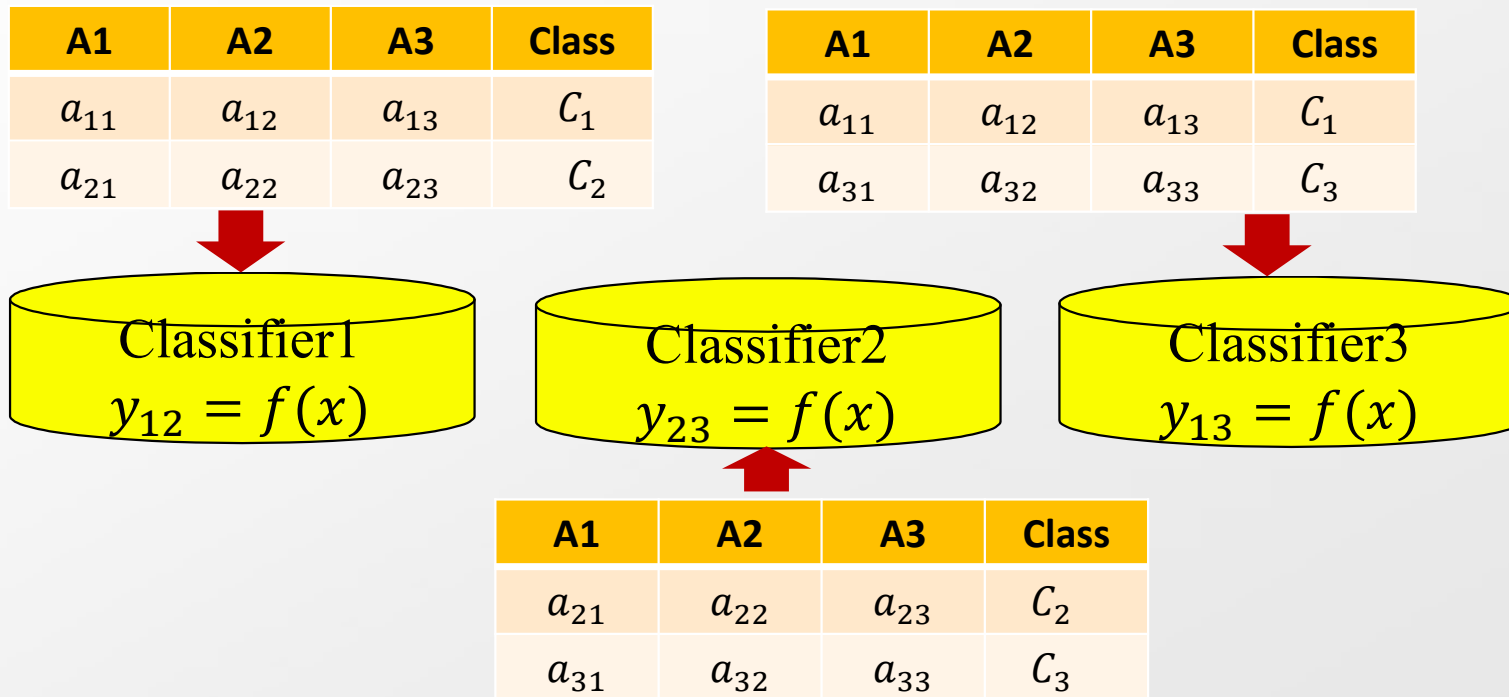


A1	A2	A3	Class
$a_{21}$	$a_{22}$	$a_{23}$	$C_2$
$a_{31}$	$a_{32}$	$a_{33}$	$C_3$

## 3.2 离散化和二元化

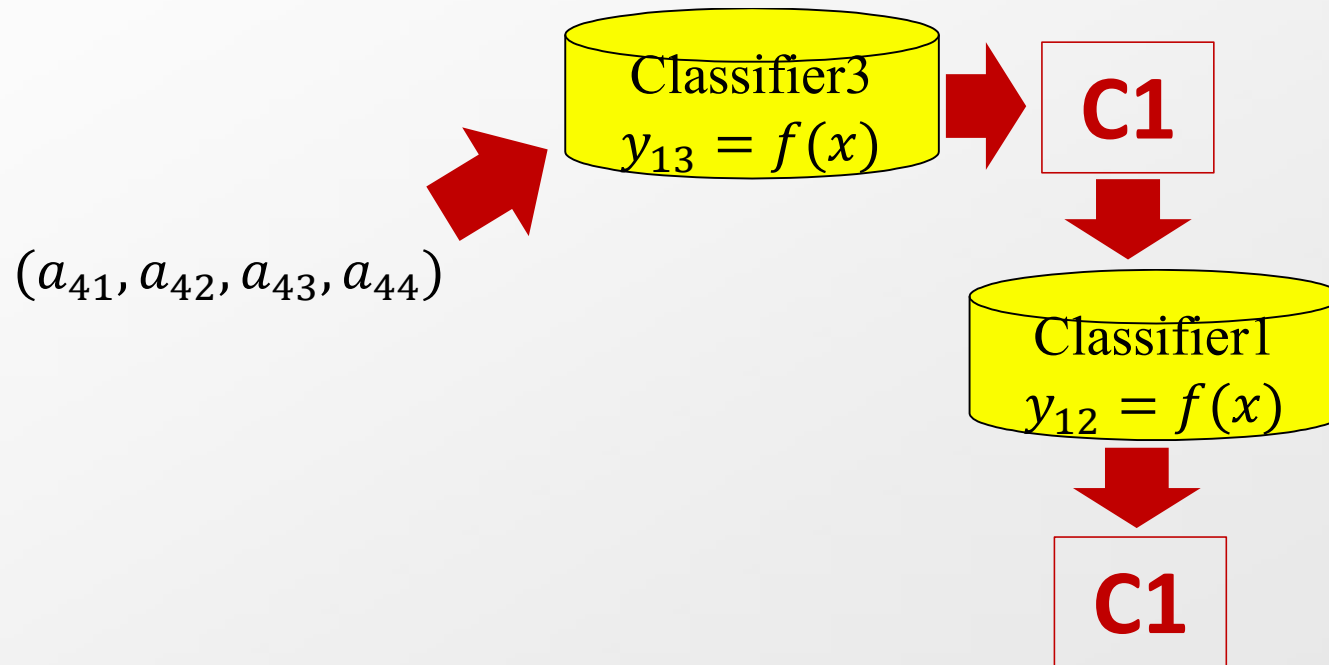
### ■ 成对分类 (pairwise classification)

➤ 为二分类数据集构建分类器



## 3.2 离散化和二元化

- 成对分类 ( pairwise classification )
  - 用训练好的分类器进行两两比较, 选择分值较高的类标





## 3.2 离散化和二元化

### ■ 一对多 vs. 成对分类

- 一对多的方法对置信度的计算结果更为敏感，需要仔细调整参数
- 训练复杂度：假设训练集包含 $n$ 种类别 $m$ 个实例，各种类别均匀分布
  - 一对多：训练 $n$ 个分类器，每个分类器 $m$ 个训练实例
  - 成对分类：训练 $n(n-1)/2$ 个分类器，每个分类器 $2m/n$ 个训练实例
  - 当分类器的计算复杂度越高，成对分类优势越明显



## 第三节 改变属性

- 特征创建
- 离散化和二元化
- 变量变换







## 3.3 变量变换

■ **定义**：用于变量（属性）的所有值的变换

➤ 简单函数

□ 常用变换包括： $x^k, \log x, 1/x, \sqrt{x}, e^x, \sin x, |x|$

□ 目的：

✓ 转换变量的取值范围

✓ 更好地比较两个属性值的不同

➤ 规范化或标准化



## 3.3 变量变换

### ■ 不同尺度属性值的标准化

➤ 若实例有属性 $A_1$ 和 $A_2$ ， $A_1$ 的取值范围是 $[1, 1000]$ ， $A_2$ 取值范围是 $[0.1, 1]$ ，则在计算两个实例之间的距离时，哪个属性更重要？

极小极大归一化

$$a_{ij} = \frac{v_{ij} - \min_{1 \leq k \leq N} v_{kj}}{\max_{1 \leq k \leq N} v_{kj} - \min_{1 \leq k \leq N} v_{kj}}$$

均值标准化

$$a_{ij} = \frac{v_{ij} - \bar{v}_{ij}}{\delta_{ij}}$$



# 小结

- 选择数据
  - 聚集
  - 抽样
- 选择属性
  - 维归约
  - 特征子集选择
- 改变属性
  - 特征创建
  - 离散化和二元化
  - 变量变换



张冬松  
[dszhang@nudt.edu.cn](mailto:dszhang@nudt.edu.cn)



# 谢谢! Q&A

THANKS FOR YOUR ATTENTION