数据预处理

电子信息与通信学院 冯 镔
fengbin@hust.edu.cn

- 什么是数据
- 数据预处理的重要性
- 数据清洗
- 数据集成与转换
- 数据消减
- 相似度和相异度

- 什么是数据
- 数据预处理的重要性
- 数据清洗
- 数据集成与转换
- 数据消减
- 相似度和相异度

什么是数据

- 数据是构成数据集的基本成分
- 数据用一组刻画对象基本特征的属性描述
- 属性描述数据的性质或特性
  - 眼睛的颜色，温度等
- 属性的集合描述了一个数据对象
  - 数据对象又称作记录、实体、观测等

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
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

属性

- 属性的值是赋予给属性的数或符号
- 属性和属性的值有什么区别？
  - 同样一个属性可以映射到不同的属性值上
    - 高度可以用米和厘米来衡量
  - 不同的属性可以映射到同样的一组值上
    - ID号和年龄都可以用整数表示
    - 但是属性值的性质可以不同
      - ID的范围没有限制，而年龄有最大和最小值范围

属性

- 测量属性的方式并不以一定和属性的性质相吻合

属性的类型

- 属性有许多不同的类型
- 标称
  - 提供不同的名字以区分对象
  - ID号，眼睛的颜色，邮政编码
- 序数
  - 给数据对象排序
  - 排名，成绩
- 区间
  - 关心值之间的差
  - 温度，日期
- 比率
  - 关心值之间的差和比例
  - 绝对温度，长度

属性的类型

- 属性的类型由其具有的数值的属性决定的
  - 相异性： = ≠
  - 序数： < >
  - 加法： + -
  - 乘法： * /

- 标称：相异性
- 序数：相异性 + 序
- 区间：相异性 + 序 + 加法
- 比率：所有四种属性

属性类型	描述	例子	操作
标称	标称属性的值仅仅只是不同的名字，即标称值只提供足够的信息以区分对象 (=, ≠)	邮政编码，员工ID，眼球颜色，性别	众数，熵，列联相关，χ^2 test
序数	序数属性的值提供足够的信息确定对象的序 (<, >)	矿石硬度（好，较好，最好），成绩，街道号码	中值，百分位，秩相关，游程检验，符号检验
区间	对于区间属性，值之间的差是有意义的，即存在测量单位(+, -)	日历日期，摄氏或华氏温度	均值，标准差，皮尔逊相关，t和F检验
比率	对于比率变量，差和比率都是有意义的 (*, /)	绝对温度，货币量，计数，年龄，质量，长度，电流	几何平均，调和平均，百分比变差

属性类型	变换	注释
标称	任何一对一变换	所有员工的ID号都重新赋值，不会导致任何不同
序数	值的保序变换, i.e., *new_value = f(old_value)* f是一个单调函数.	表示概念好，较好，最好的属性可以完全等价的用值 {1, 2, 3} or { 0.5, 1, 10} 来表示.
区间	*new_value =a * old_value + b* a和b是常数	华氏和摄氏温度标度零度的位置和1度的大小不同
比率	*new_value = a * old_value*	长度可以用米和英尺度量

离散和连续属性

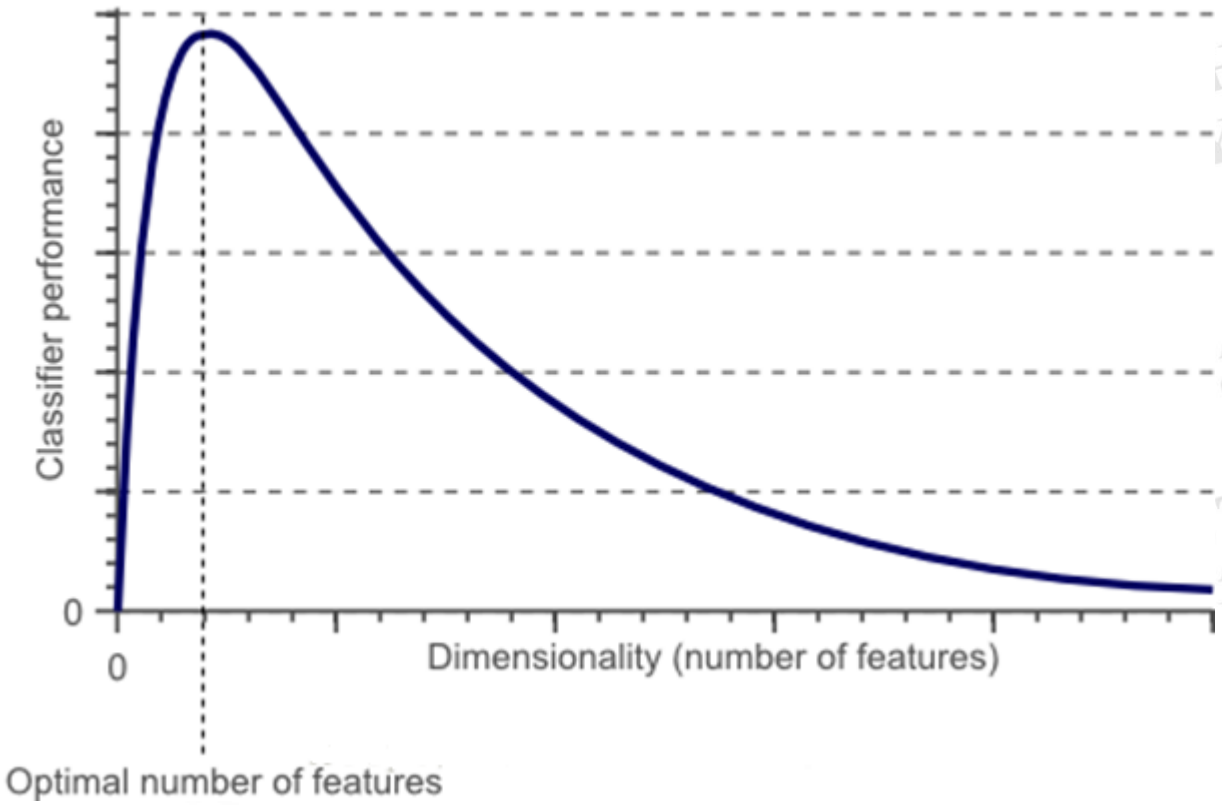
- 离散属性
  - 具有有限个，或无限个但是可数的值
  - 如：邮政编码或ID号
  - 通常用整数变量表示
  - 二元属性是离散属性的一种特殊情况
- 连续属性
  - 取实数值的属性，常用浮点数表示
  - 如：温度，高度和重量
  - 实际中，实数值只能用有限的精度测量和表示

离散和连续属性

- 下列属性分别属于什么类别
- （1）医院中的病人数
  - 离散、定量、比率
- （2）透光能力：不透明、半透明、透明
  - 离散、定性、序数
- （3）书籍的ISBN号
  - 离散、定性、标称

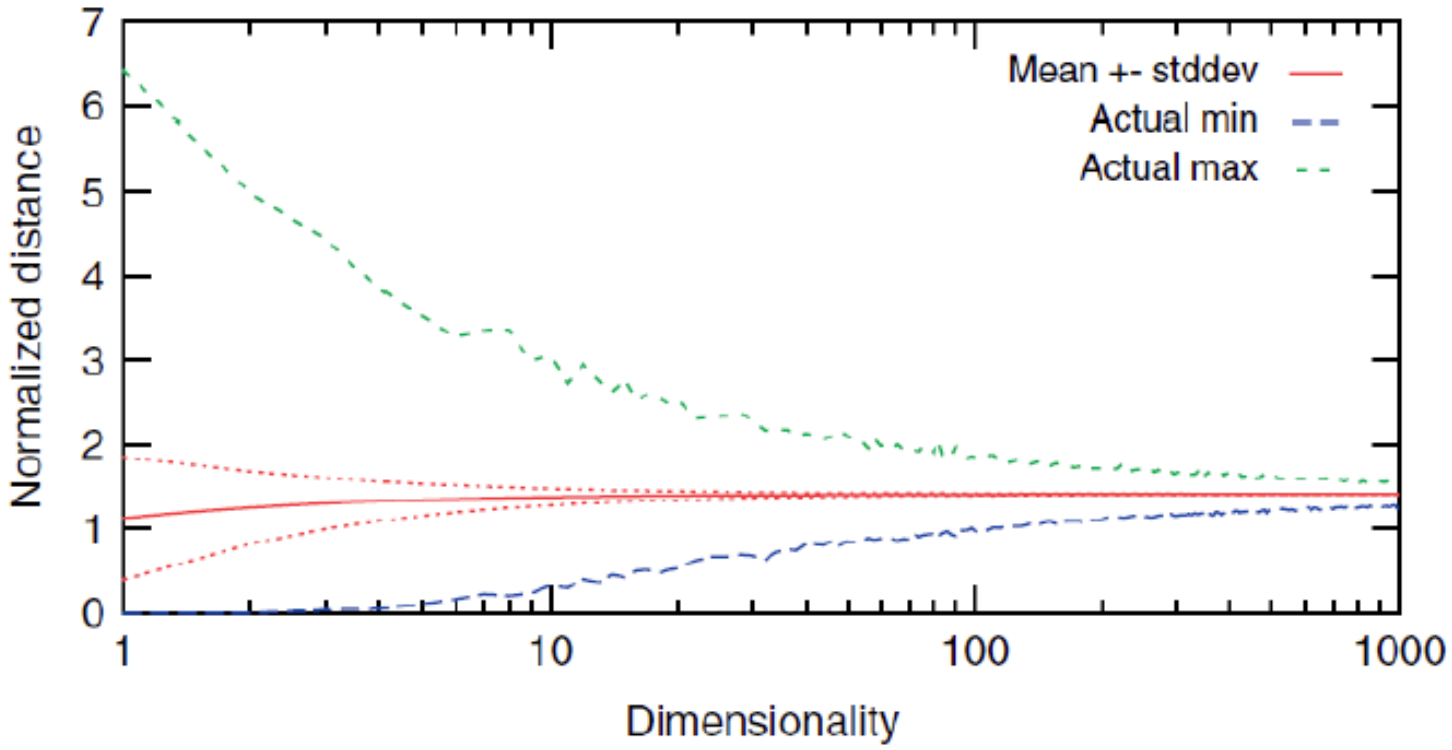
数据集的类型

- 数据集的一般特性
- 维度
  - 维灾难
  - 维归约



数据集的类型

- 距离集中效应
  - 从高斯分布[0,1]中采样10^5个样本，计算两两之间的距离，并归一化



数据集的类型

- 稀疏性
  - 非对称性数据集，大部分属性值都为0
  - 节省大量的计算时间和存储空间
- 分辨率
  - 不同分辨率尺度下的模式不同
  - 地球表面在数米的分辨率下很不平坦，但在数十公里下却相对平坦

数据集的类型

- 记录数据
  - 每个记录数据包含固定的数据字段（属性）集
  - 数据矩阵、文档数据、事务数据
- 图形数据
  - 利用图形捕获数据对象之间的联系
  - **Web**、分子结构
- 有序数据
  - 属性具有空间或时间自相关性
  - 空间数据、时间数据、序列数据

记录数据

- 一系列记录的集合，每个记录包含固定的属性集
- 几种变形

Tid	Refund	Marital Status	Taxable Income	Cheat
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

记录数据

- 数据矩阵

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

- 文档数据

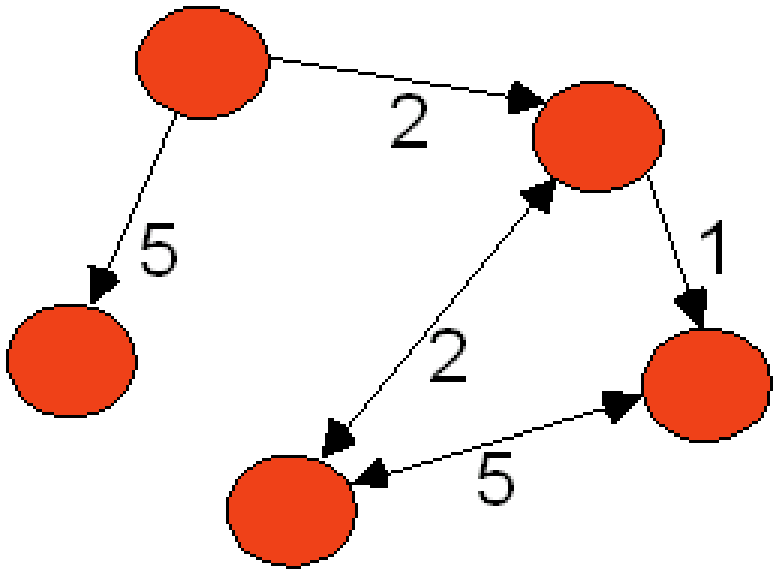
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

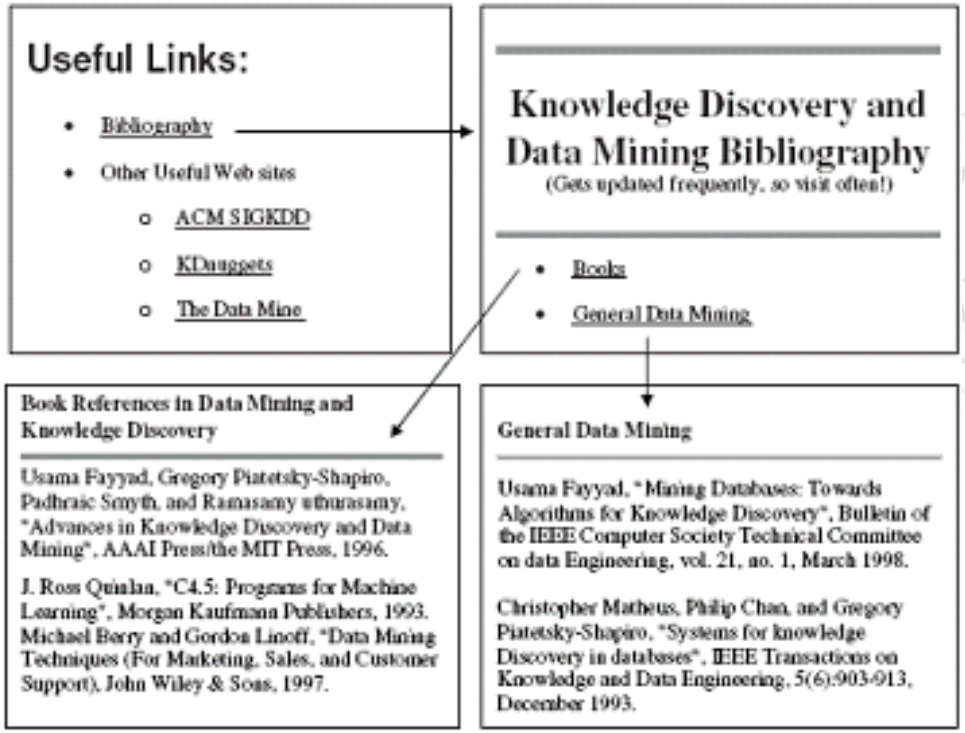
- 事务数据

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

图形数据

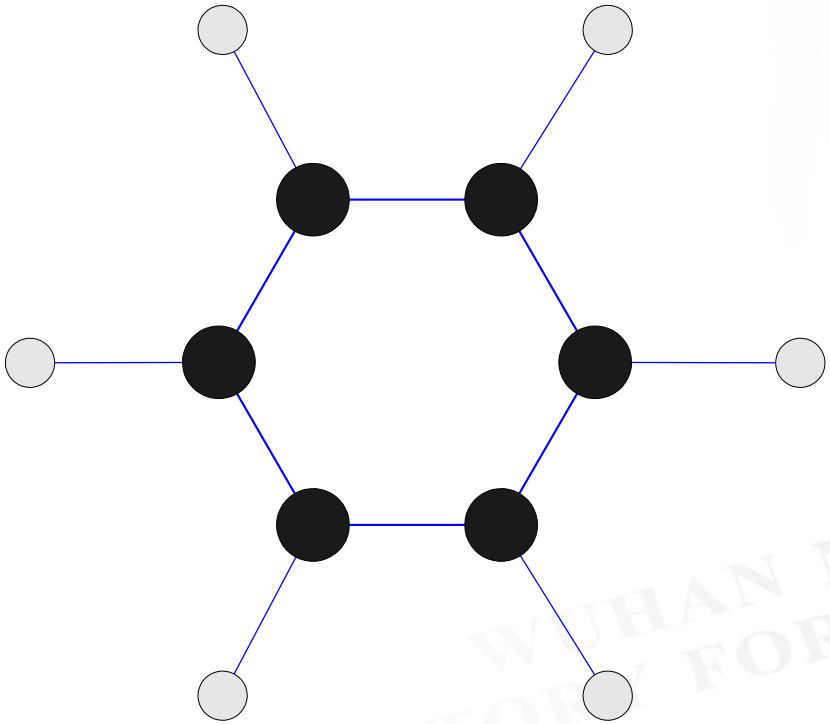
- 用图形表示对象之间的联系
  - 数据对象映射到图的结点
  - 对象之间的联系用对象之间的链和方向、权值等信息表示





图形数据

- 具有图形对象的数据



有序数据

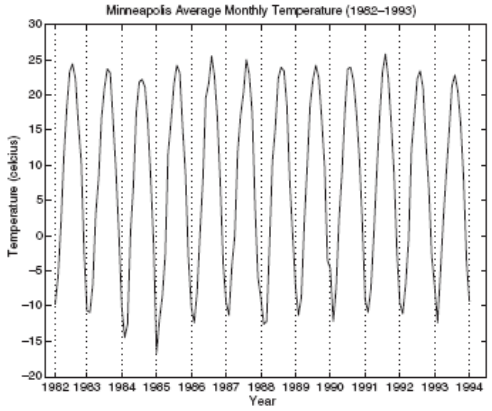
Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

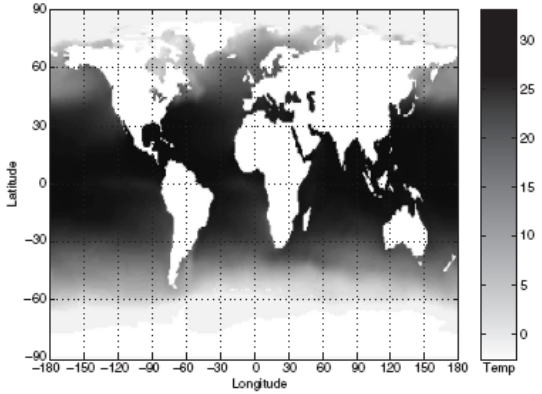
(a) Sequential transaction data.

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

(b) Genomic sequence data.



(c) Temperature time series.

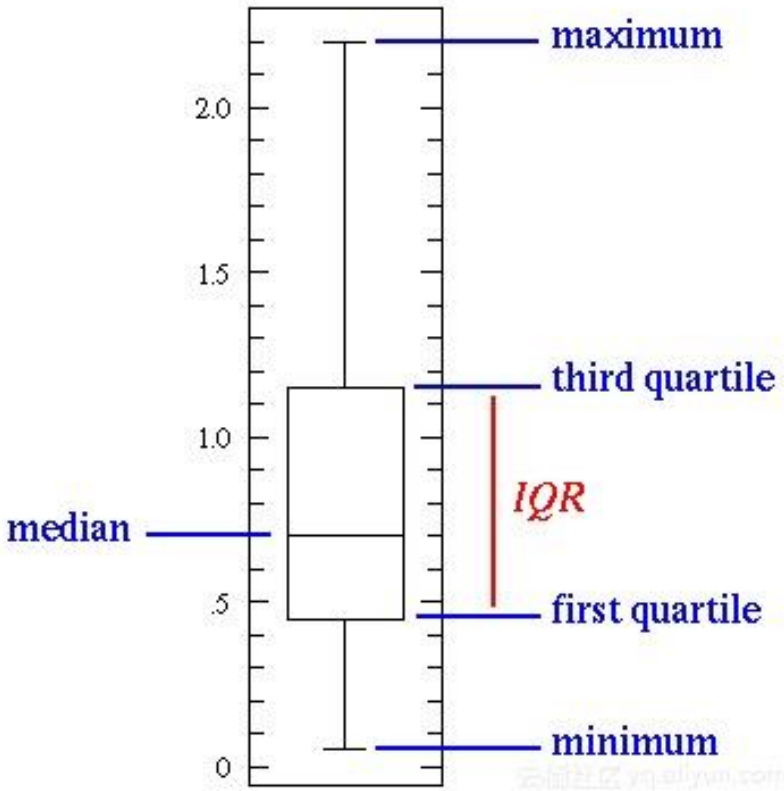


(d) Spatial temperature data.

Figure 2.4. Different variations of ordered data.

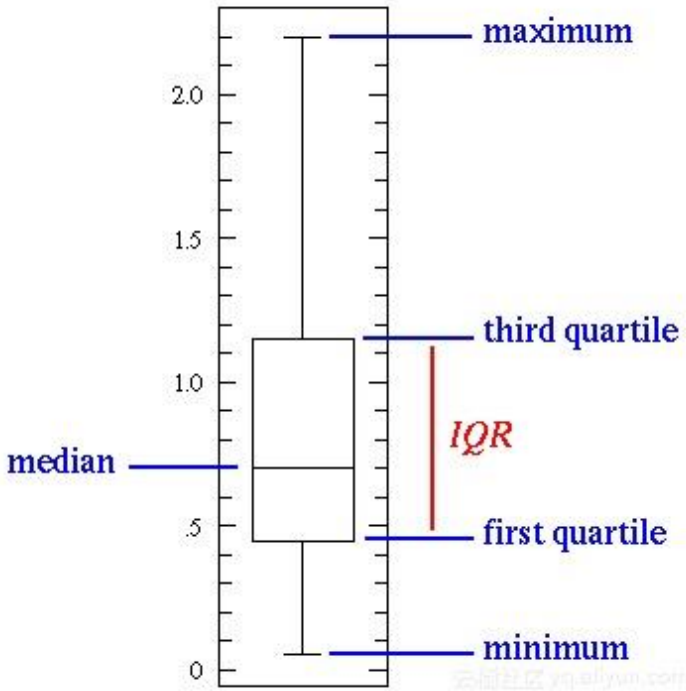
数据的统计量

- 箱型图
- 中位数median
  - 对噪声更有鲁棒性
- 二十五百分位数
- 七十五百分位数
- **IQR**
- **interquartile range,** 四分位差，内距



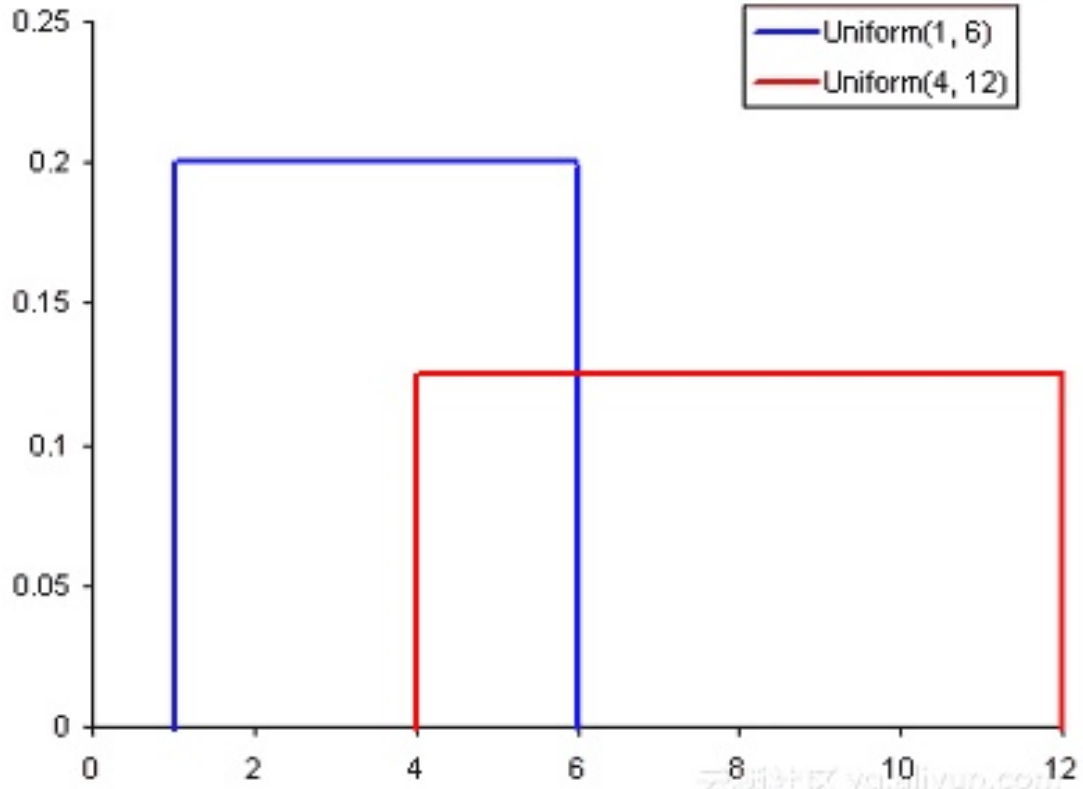
数据的统计量

- 箱形图很小，意味着很多数据点是相似的
- 箱形图较大，意味着大部分的数据点之间的差异很大
- 如果中位数接近了底部，大部分的数据具有较小的值
- 如果中位数比较接近顶部，大多数的数据具有更大的值
- 如果框上下两端的线段很长，表示数据具有很高的标准偏差和方差



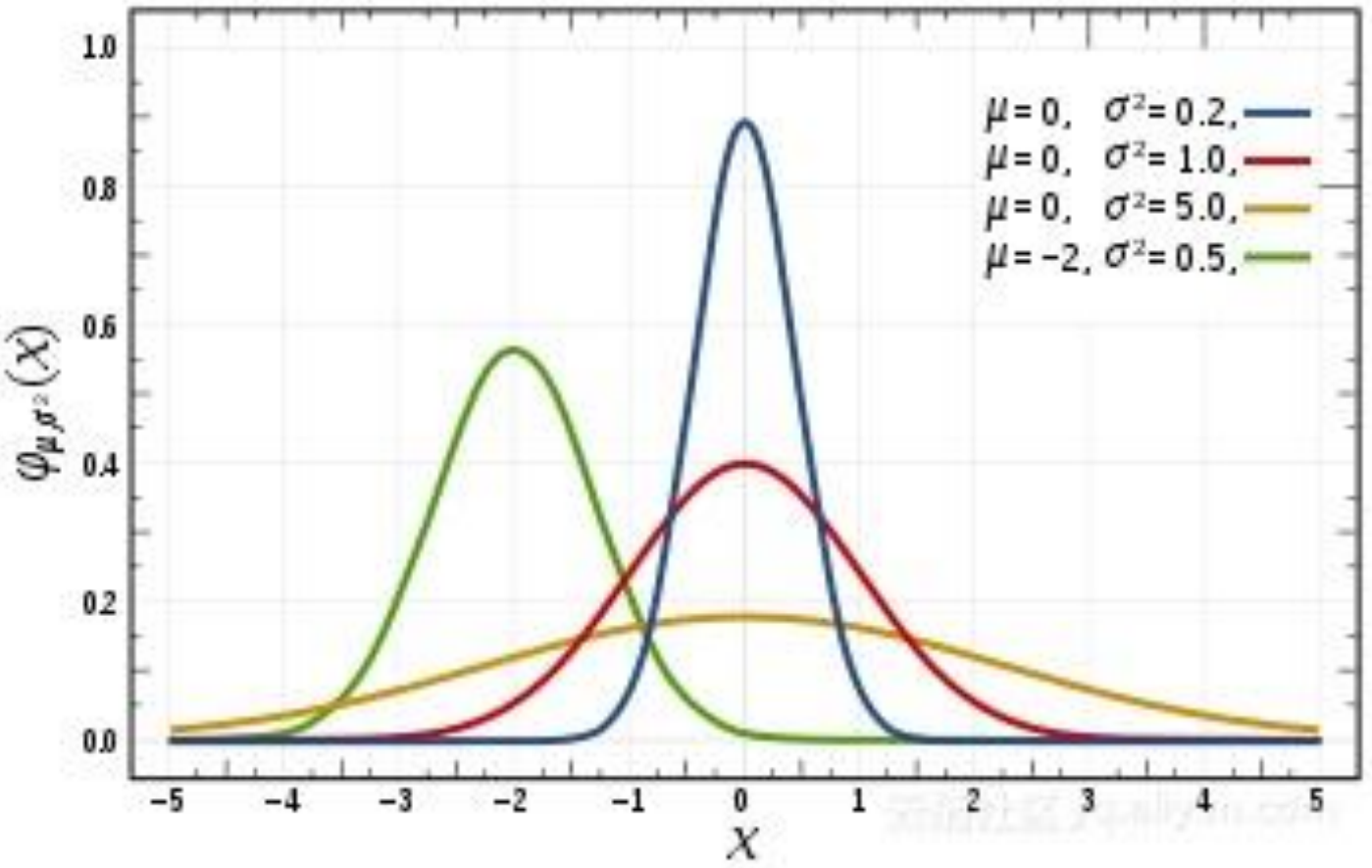
数据的统计量

- 常见的概率分布
- 均匀分布



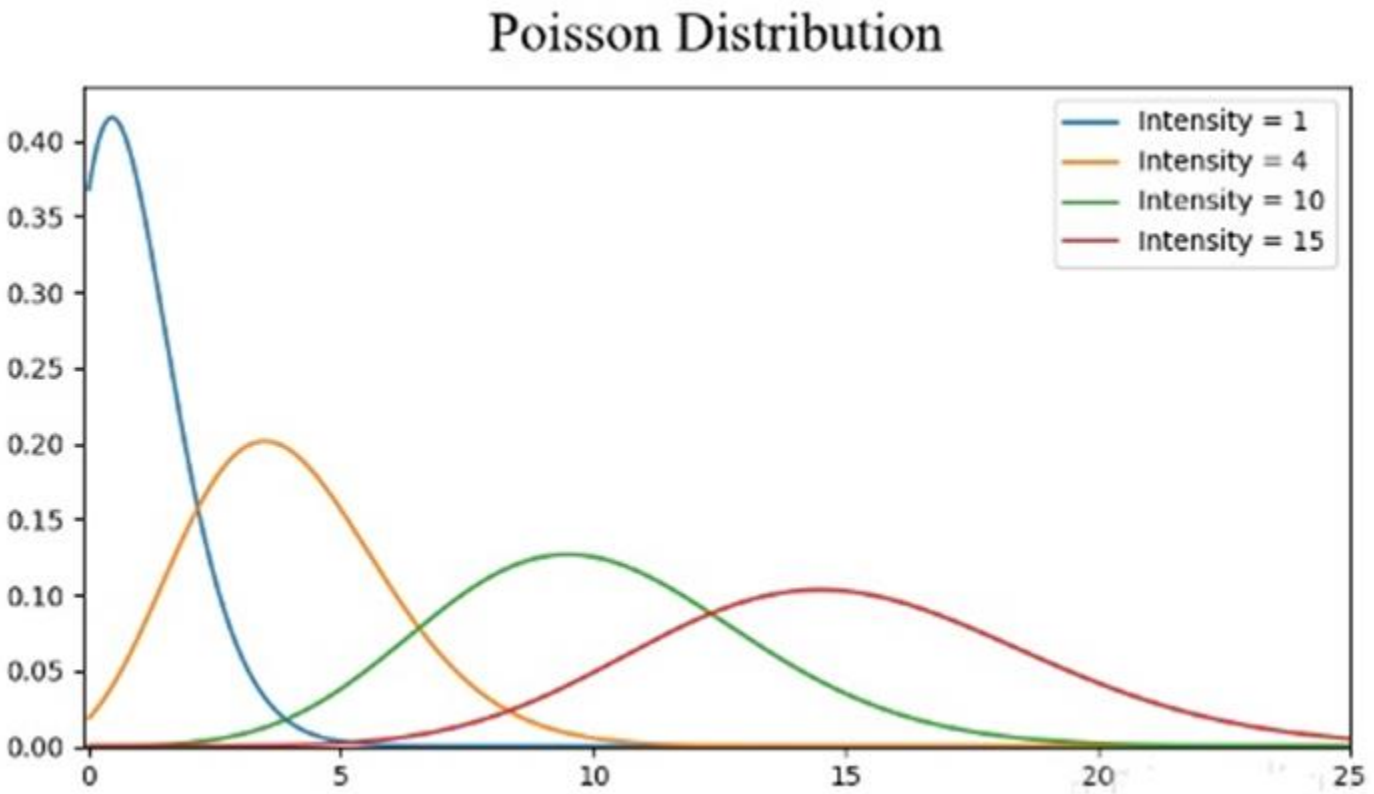
数据的统计量

- 高斯分布


1.0
0.8
0.6
0.4
0.2
0.0
$\varphi_{\mu,\sigma^2}(x)$
−5
−4
−3
−2
−1
0
1
2
3
4
5
x
μ = 0, σ² = 0.2,
μ = 0, σ² = 1.0,
μ = 0, σ² = 5.0,
μ = −2, σ² = 0.5,

数据的统计量

- 泊松分布


Poisson Distribution
Intensity = 1
Intensity = 4
Intensity = 10
Intensity = 15
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
5
10
15
20
25

- 什么是数据
- 数据预处理的重要性
- 数据清洗
- 数据集成与转换
- 数据消减
- 相似度和相异度

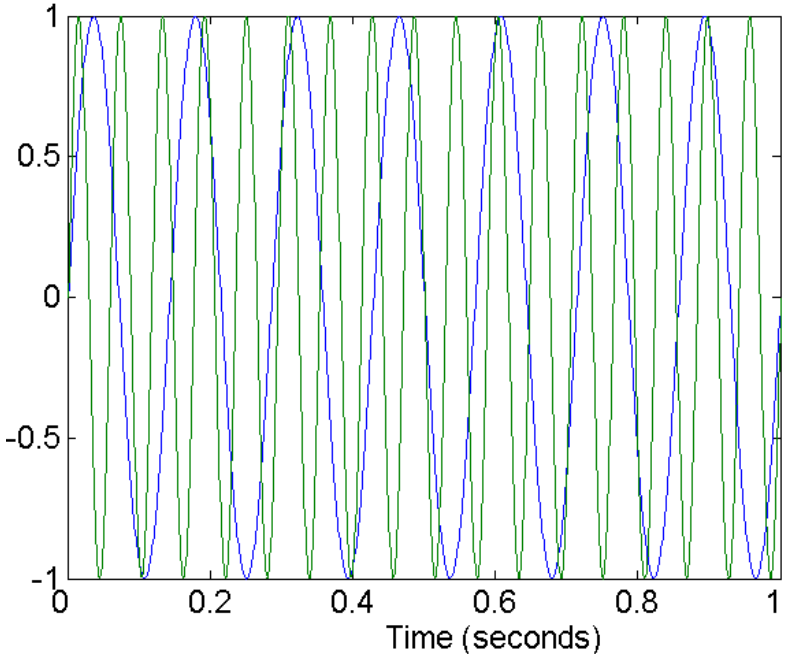
数据预处理的重要性

- 在对包含噪声、不完整、甚至是不一致数据进行数据挖掘时，需要进行数据预处理
  - 注重理解和提高数据质量将改进分析结果质量
- 高质量的决策来自高质量的数据
  - 最终达到提高所获模式知识质量的目的

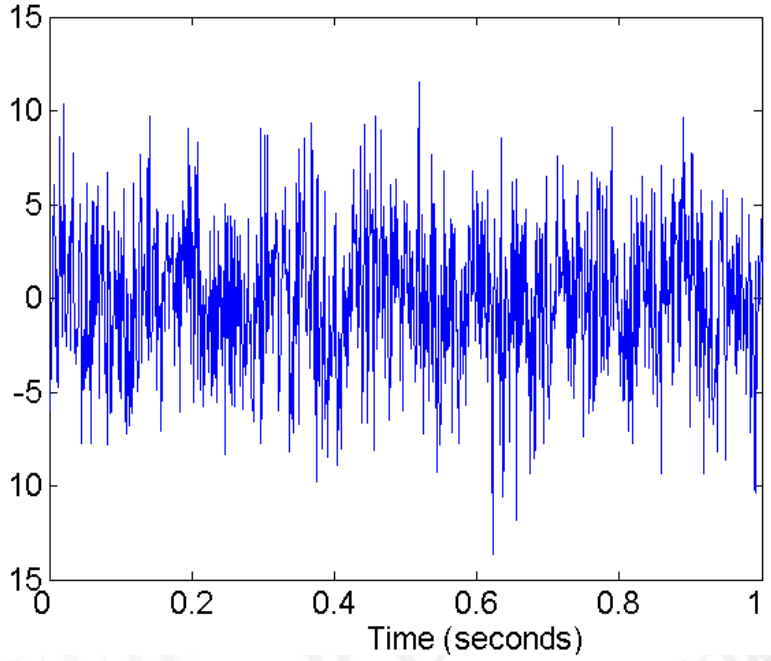
数据预处理的重要性

- 噪声数据是指数据中存在着错误或异常（偏离期望值）的数据
  - 可能是数据采集、记录、传输过程中的问题
- 不完整数据是指感兴趣的属性没有值
- 不一致数据是指数据内涵出现不一致情况
  - 作为属性的“部门编码”出现不同值

数据预处理的重要性



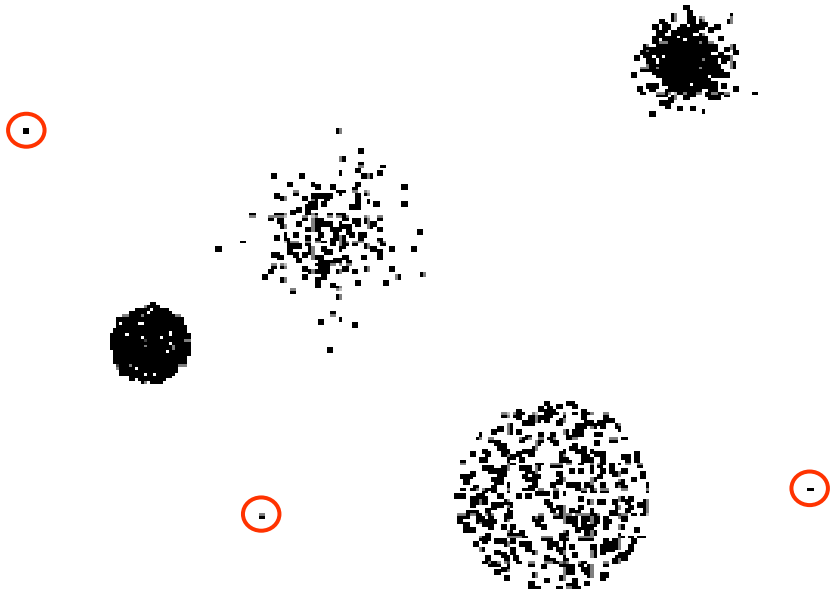
正弦波形



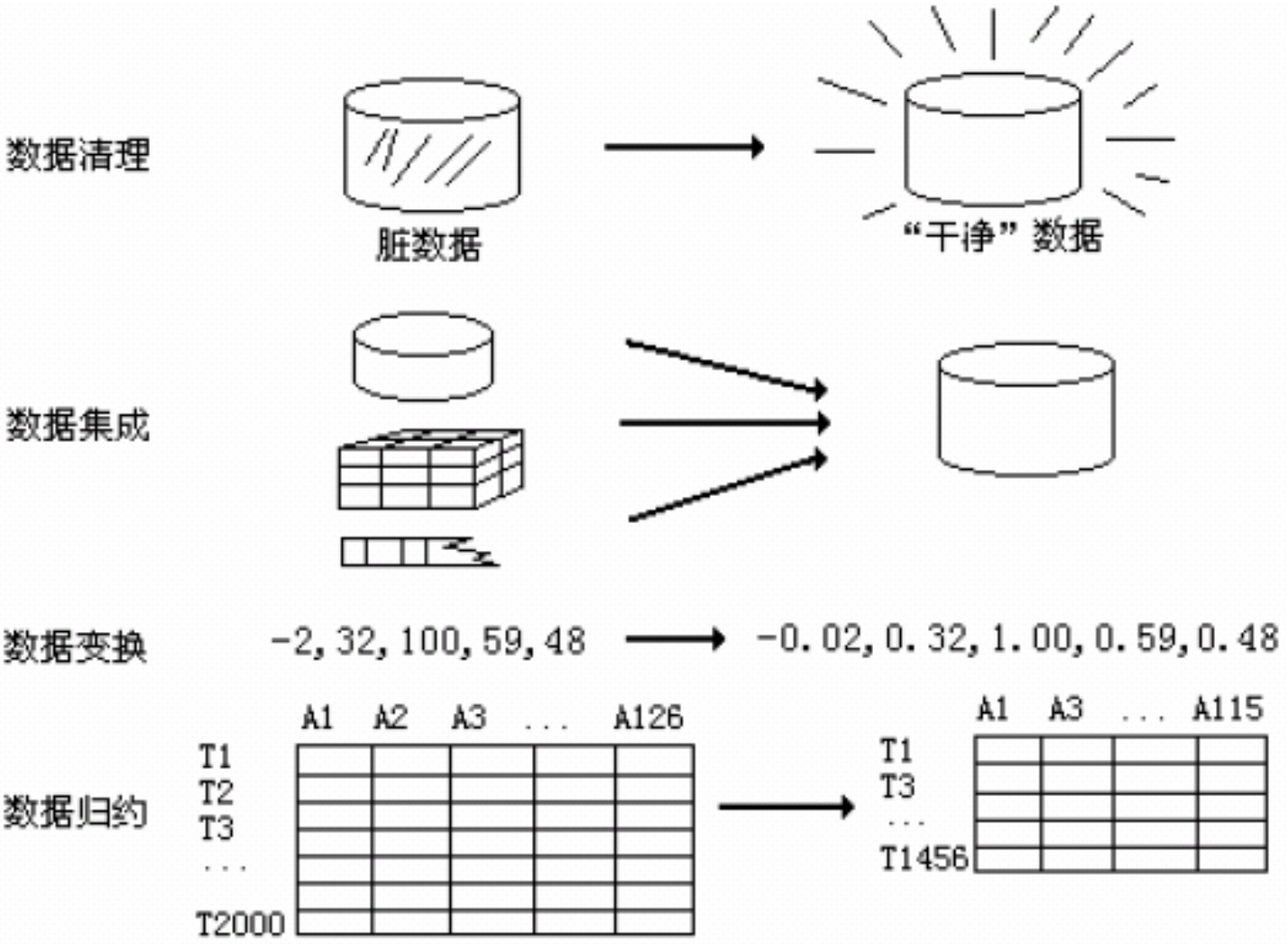
正弦波形 + 噪声

数据预处理的重要性

- 离群点：在某种意义上具有不同于数据集中其他大部分数据对象的特征的数据对象，或是相对于该属性的典型值



数据预处理的重要性


数据清理
脏数据
“干净”数据
数据集成
数据变换
-2, 32, 100, 59, 48
-0.02, 0.32, 1.00, 0.59, 0.48
数据归约
A1 A2 A3 ... A126
T1
T2
T3
...
T2000
A1 A3 ... A115
T1
T3
...
T1456

数据预处理的重要性

- 数据清洗是指消除数据中所存在的噪声以及纠正其不一致的错误
- 数据集成则是将来自多个数据源的数据合并到一起构成一个完整的数据集
- 数据转换是指将一种格式的数据转换为另一种格式的数据
- 数据消减是指通过删除冗余特征消除多余数据

- 什么是数据
- 数据预处理的重要性
- 数据清洗
- 数据集成与转换
- 数据消减
- 相似度和相异度

数据清洗

- 处理例程通常包括：
- 填补遗漏的数据值
- 平滑有噪声数据
- 识别或除去异常值(outlier)
- 解决不一致性问题

数据清洗

- 1. 遗漏数据处理
- 忽略该条记录
  - 最简单，不是很有效，尤其当遗漏比例较大
- 手工填补遗漏值
  - 可行性较差
- 利用缺省值填补遗漏值
  - 可能会误导挖掘进程，使结果产生较大误差

数据清洗

- 1. 遗漏数据处理
- 利用均值填补遗漏值
  - 利用顾客的平均收入填补收入属性中所有的遗漏值
- 利用同类别均值填补遗漏值
  - 适合分类挖掘中使用
  - 同一信用风险类别下收入属性的平均值
- 利用最可能的值填补遗漏值
  - 利用回归分析、贝叶斯公式计算

数据清洗

- **2. 噪声数据处理**

- **邻点平滑**
  - **利用周围点的数据进行平滑，bin的宽度越宽，平滑效果越明显**

排序后价格: 4, 8, 15, 21, 21, 24, 25, 28, 34

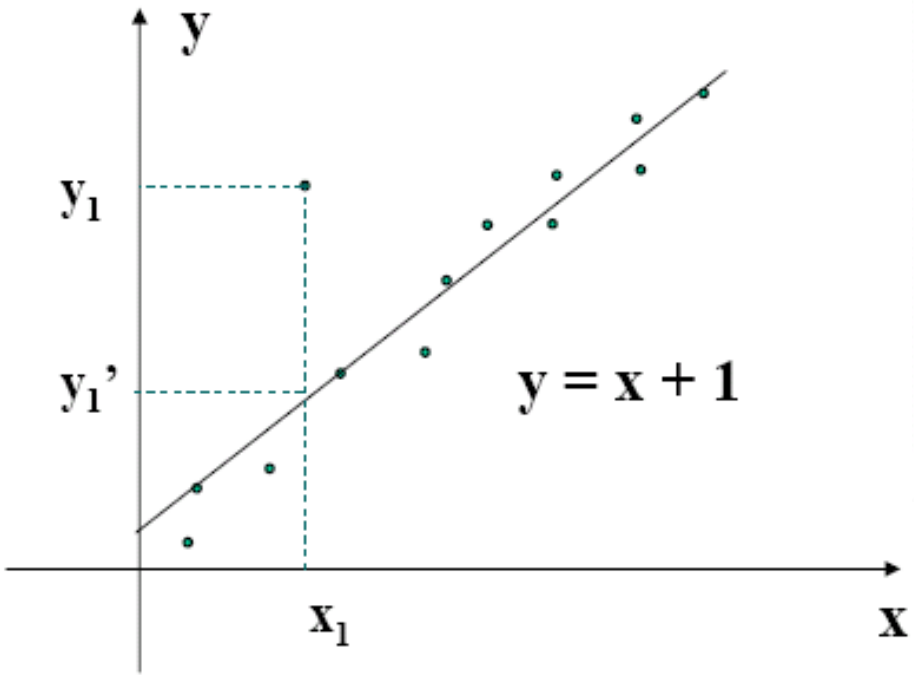
划分为等高度 bins:
- Bin 1: 4, 8, 15
- Bin 2: 21, 21, 24
- Bin 3: 25, 28,34

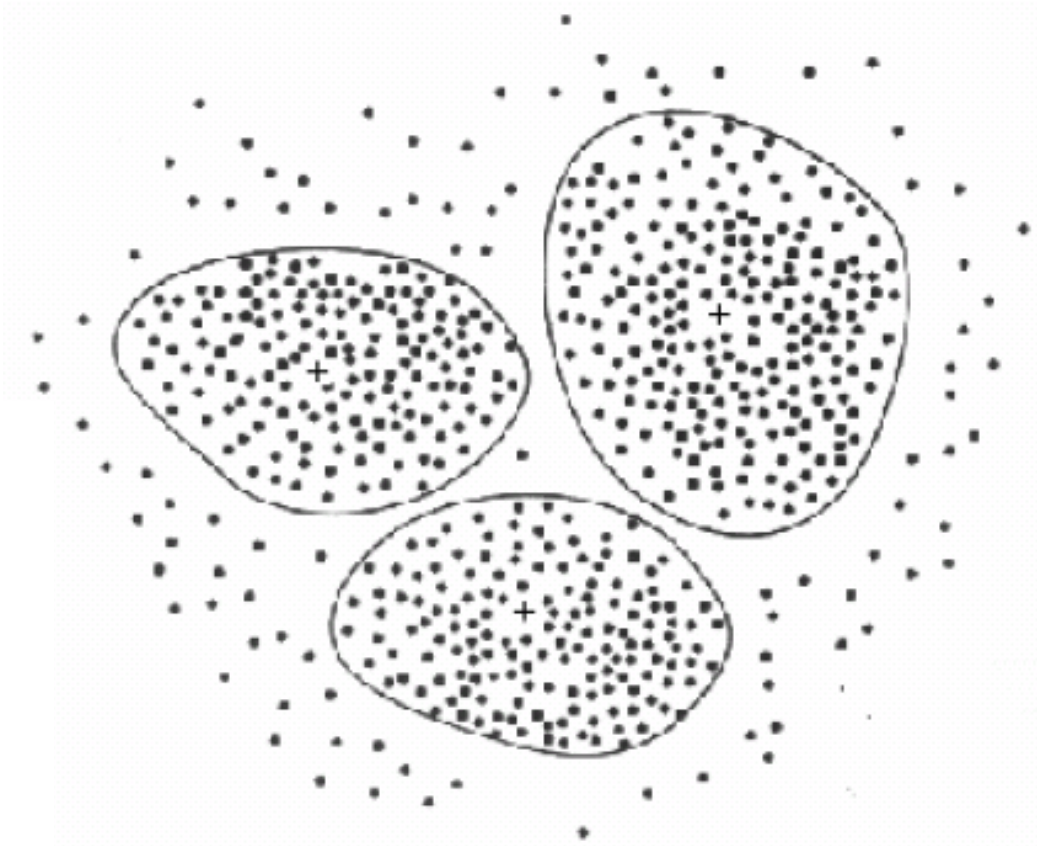
根据bin均值进行平滑:
- Bin 1: 9, 9, 9
- Bin 2: 22, 22, 22
- Bin 3: 29, 29, 29

根据bin边界进行平滑:
- Bin 1: 4, 4, 15
- Bin 2: 21, 21, 24
- Bin 3: 25, 25, 34

数据清洗

- 聚类分析
- 人机结合检查
- 回归方法





- 噪声对象可能是离群点吗？
- 噪声对象总是离群点吗？
- 离群点总是噪声对象吗？
- 噪声能将典型值变成例外值，或反之吗？

离群点检测---Local Outlier Factor

- LOF：基于密度的经典算法
  - SIGMOD 2000
- 基于统计的方法：假设数据服从特定的数据分布
  - 假设往往不成立
- 基于聚类的方法：DBSCAN…
  - 不能量化每个数据点的异常程度

离群点检测---Local Outlier Factor

- 基于距离对密度进行定义
- ***k*-distance**：在距离数据点 p 最近的几个点中，第 k 个最近的点与点 p 之间的距离
- **Rechability distance**：给定参数k时数据点 o 的***k*-distance**和数据点p与点o之间的直接距离的最大值

$$reach\_dist_k(p,o) = \max\{k\text{-}distance(o), d(p,o)\}$$

离群点检测---Local Outlier Factor

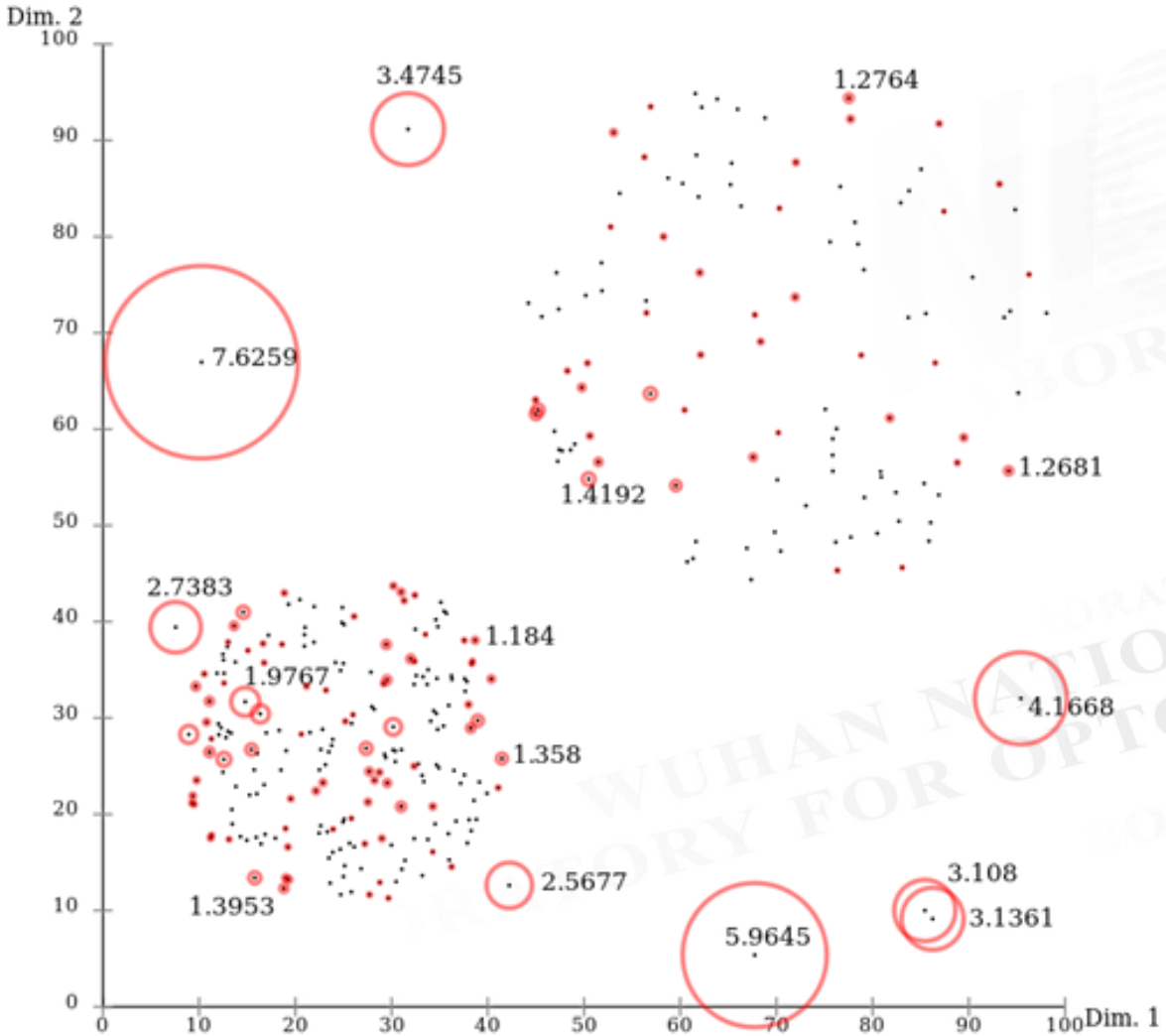
- **Local Rechability Density**：邻近的数据点平均可达距离的倒数

$$lrd_k(p) = \frac{1}{\frac{\sum_{o \in N_k(p)} reach\_dist_k(p,o)}{|N_k(p)|}}$$

- **Local Outlier Factor**：p的邻居的平均局部可达密度与数据点p局部可达密度的比值

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{lrd(o)}{lrd(p)}}{|N_k(p)|} = \frac{\sum_{o \in N_k(p)} lrd(o)}{|N_k(p)|} / lrd(p)$$

离群点检测---Local Outlier Factor



离群点检测---Local Outlier Factor

- 算法流程：
  - 1. 对于每个数据点，计算它与其它所有点的距离，并排序
  - 2. 对于每个数据点，找到它的K近邻，计算LOF得分
- 算法复杂度：$O(n^2)$
  - FastLOF(Goldstein，2012): 先将整个数据随机的分成多个子集，然后在每个子集里计算，

例子

➢1. 导入预处理所需要的库

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

➢2. 导入数据集

```
dataset = pd.read_csv('my_data.csv')
```

例子

➢3. 读入数据

```
X = dataset.iloc[:, :-1].values
```

```
y = dataset.iloc[:, 3].values
```

	Animal	Age	Worth	Friendly
0	Cat	4.0	72000.0	No
1	Dog	17.0	48000.0	Yes
2	Moose	6.0	54000.0	No
3	Dog	8.0	61000.0	No
4	Moose	4.0	NaN	Yes

例子

➢4. 导入sklearn库中的imputer类

```
from sklearn.preprocessing import Imputer

imputer = Imputer(missing_values = np.nan, strategy = 'mean', axis = 0)
```

```
imputer = imputer.fit(X[:, 1:3])
```

例子

➢5. 填充数据

```
X[:, 1:3] = imputer.transform(X[:, 1:3])
```

```
[['Cat' 4.0 72000.0]
 ['Dog' 17.0 48000.0]
 ['Moose' 6.0 54000.0]
 ['Dog' 8.0 61000.0]
 ['Moose' 4.0 63777.77777777778]
 ['Cat' 15.0 58000.0]
 ['Dog' 8.666666666666666 52000.0]
 ['Cat' 12.0 79000.0]
 ['Moose' 5.0 83000.0]
 ['Cat' 7.0 67000.0]]
```

- 什么是数据
- 数据预处理的重要性
- 数据清洗
- 数据集成与转换
- 数据消减
- 相似度和相异度

数据集成与转换

- 将来自多个数据源的数据合并到一起
- 集成常常会引起数据的不一致或冗余
  - 在一个数据库中一个人的姓取“Bill”，而在另一个数据库中则取“B”
- 有时还需要进行数据清洗以便消除可能存在的数据冗余
- 数据转换主要是对数据进行规格化操作
  - 尤其是使用基于对象距离的挖掘算法时

数据集成与转换

- 1. 模式集成问题
  - 属性的相互匹配
  - 人工干预
- 2. 冗余问题
  - 如果一个属性可以从其他属性中推演出来，那就是冗余属性
    - 月收入→平均月收入
- 3. 数据值冲突检测与消除
  - 来自不同数据源的属性值或许不同
  - 如比例尺度不同

数据集成与转换

- 数据转换就是将数据转换或归并以构成一个适合数据挖掘的描述形式，包括：
- 1. 合并处理
  - 对数据进行总结或合计操作
  - 如对每天的销售额进行合计操作获得每月的总额，多用于对数据进行多粒度的分析
- 2. 泛化处理
  - 用更抽象（更高层次）的概念来取代低层次的数据对象
- 3. 规格化

数据集成与转换

- 规格化是将一个属性取值范围投射到一个特定范围之内，以消除数值型属性因大小不一而造成挖掘结果的偏差
- 最大最小规格化方法
  - 保留了原数据中存在的关系
  - 当遇到超出目前属性取值范围的数值，将会引起系统错误

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A$$

数据集成与转换

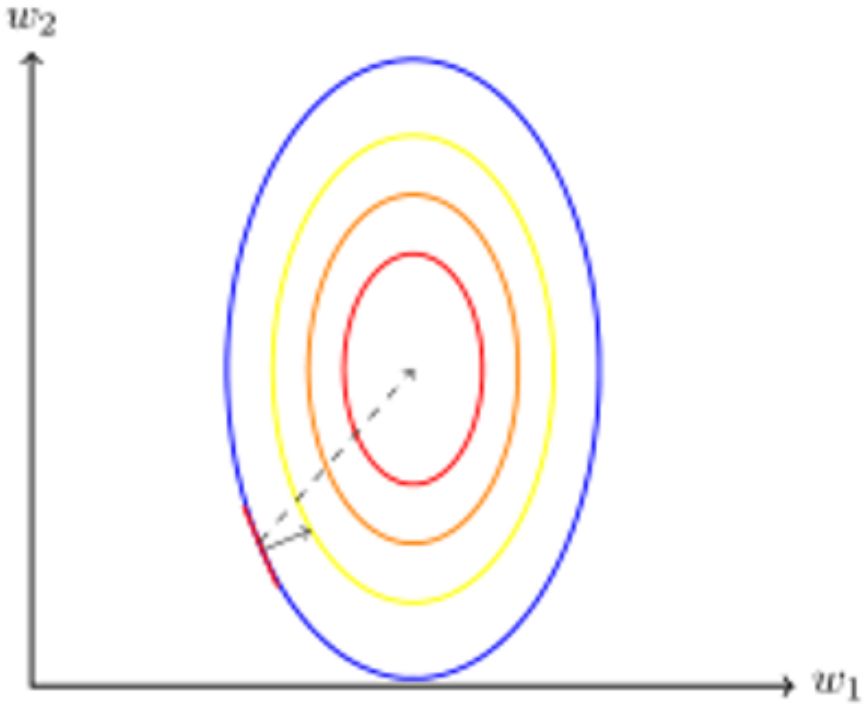
- 零均值规格化方法
  - 根据属性的均值和偏差对其进行规格化，常用于属性最大值和最小值未知，或用最大最小规格化方法时会出现异常数据的情况

$$v' = \frac{v - \overline{A}}{\sigma_A}$$

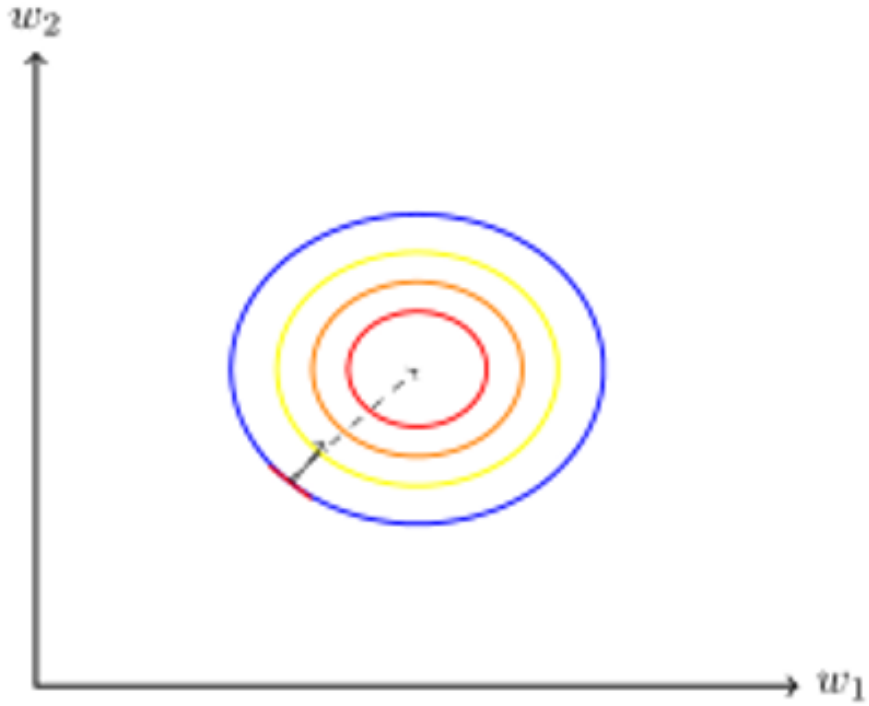
- 十基数变换规格化方法
  - 通过移动属性小数点位置达到规格化目的
  - 所移动的小数位取决于属性绝对值的最大值

$$v' = \frac{v}{10^j}$$

数据集成与转换



未归一化数据的梯度



归一化数据的梯度

- 什么是数据
- 数据预处理的重要性
- 数据清洗
- 数据集成与转换
- 数据消减
- 相似度和相异度

数据消减

- 对大规模数据库内容进行复杂的数据分析常常需要消耗大量的时间，使得这样的分析显得不现实和不可行
- 数据消减（**data reduction**）的目的是在（基本）不影响最终挖掘结果的条件下，缩小所挖掘数据的规模

数据消减

- 数据聚合（data aggregation）
  - 构造数据立方
- 维数消减（dimension reduction）
  - 维归约
- 数据块消减（numerosity reduction）
  - 采用更简单的数据表达形式取代原有的数据
  - 参数模型，非参数模型等

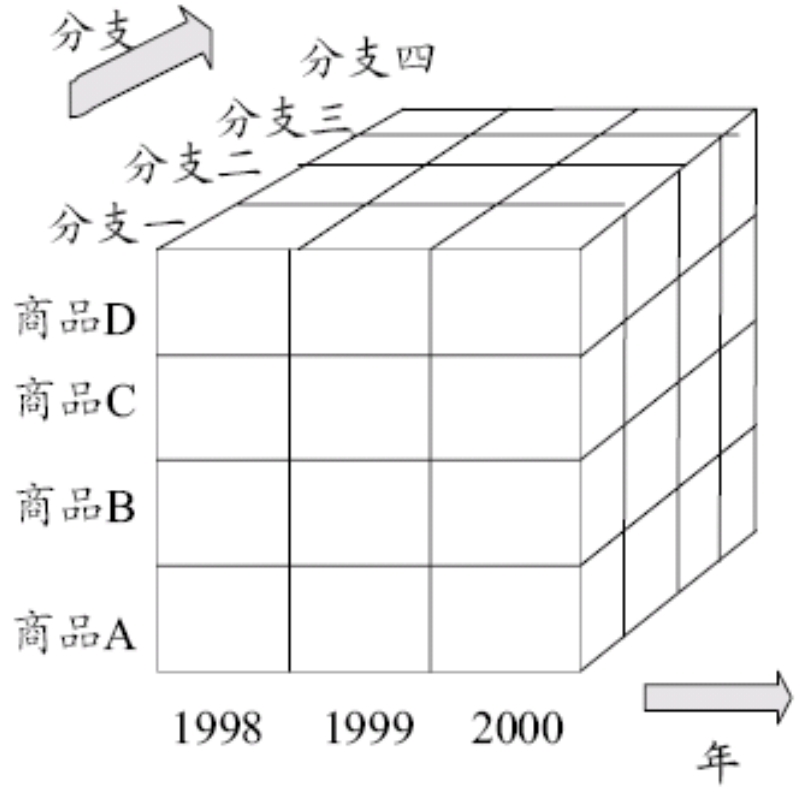
数据聚合

- 从三个维度对公司原始销售数据进行合计所获得的数据立方

年=1999
年=1998
年=1997

季度	销售额
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

年	销售额
1997	$1,568,000
1998	$2,356,000
1999	$3,594,000



维数消减

- 许多属性是与挖掘任务无关或冗余的
  - 挖掘顾客是否会在商场购买mp3播放器的分类规则时，顾客的电话号码可能与挖掘任务无关
- 维数消减通过消除多余和无关的属性而有效减小数据集的规模
- 属性子集选择
  - 寻找出最小的属性子集，同时确保新数据子集的概率分布尽可能接近原来数据集的概率分布

维数消减

- 全搜索难以实现
  - 包含d个属性的集合共有2^d个不同子集
- 1. 逐步添加
- 2. 逐步消减
- 3. 消减与添加结合
- 4. 决策树归纳

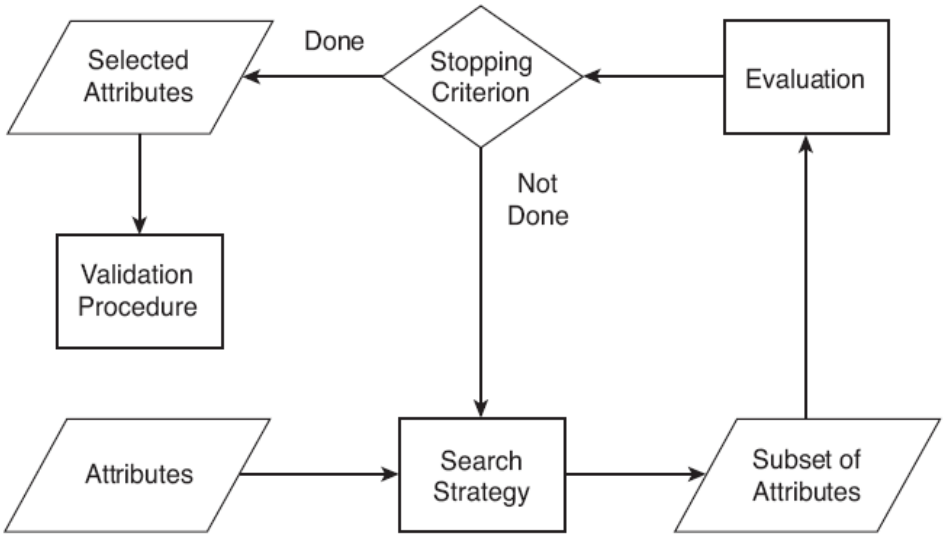


Figure 2.11. Flowchart of a feature subset selection process.

维数消减

向前选择

初始属性集：
{A1,A2,A3,A4,A5,A6}

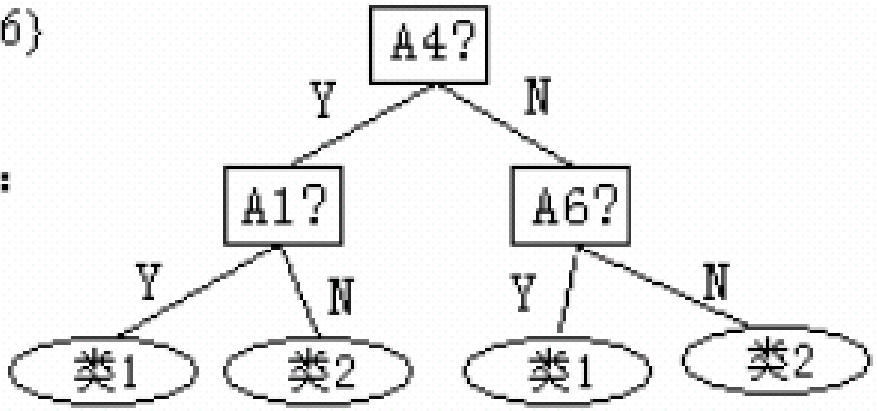
初始化归约集：
{}
→ {A1}
→ {A1,A4}
→归约后的属性集：
{A1,A4,A6}

向后删除

初始属性集：
{A1,A2,A3,A4,A5,A6}
→ {A1,A3,A4,A5,A6}
→ {A1,A4,A5,A6}
→归约后的属性集：
{A1,A4,A6}

判定树归纳

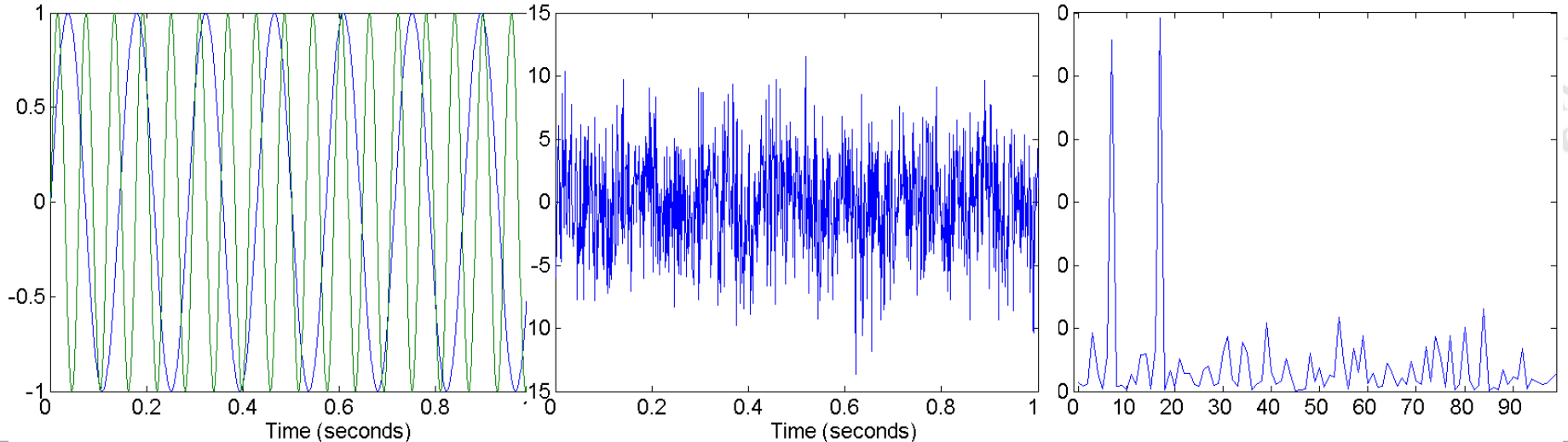
初始属性集：
{A1,A2,A3,A4,A5,A6}



→归约后的属性集：
{A1,A4,A6}

维数消减

- 特征创建：由原来的属性集创建新的属性集，更有效的捕获数据集中的重要信息
- 1. 特征提取
  - 需要特定领域知识
- 2. 数据映射
- 3. 特征构造



维数消减

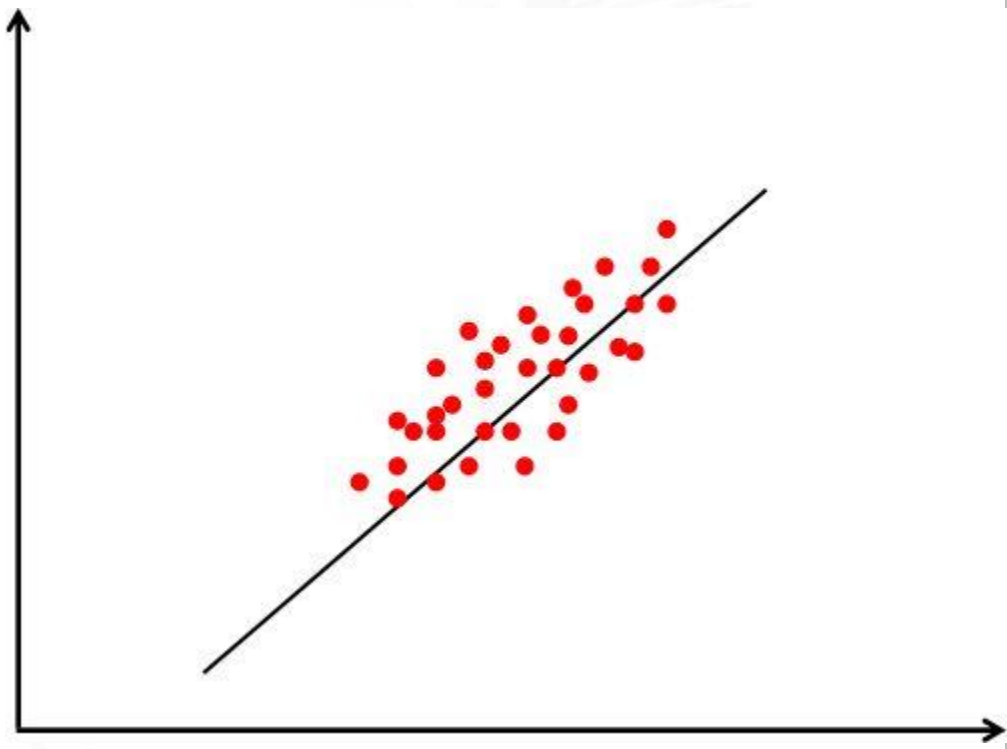
- 线性代数技术PCA
  - Principal Components Analysis

- 用于连续属性的线性代数技术，找出新的正交属性（主成分），是原属性的线性组合，并捕获数据的最大变差

- 广泛应用在人脸识别，图像压缩，模式识别等领域

维数消减

- 向重构误差最小（方差最大）的方向做线性投影
- 一种无监督的学习方法，不能直接用于分类和回归



PCA-背景知识

- 假设有一个分布的采样

$$X = [\,1\ 2\ 4\ 6\ 12\ 15\ 25\ 45\ 68\ 67\ 65\ 98\,]$$

- 均值

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- 标准差（SD）

$$s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{(n-1)}}$$

  - 为什么除以n-1，而不是n？

PCA-背景知识

- 自由度：可以不受约束，自由变化的变量的个数
  - 因为需要计算均值，n个样本的自由度为n-1

- 假设总体X的方差存在，且$Var(X)=\sigma^2$

- 考察$S^2=\sum_{i=1}^{N}(X_i-\overline{X})^2$ 和σ^2之间的关系

PCA-背景知识

$$S^2 = \sum_{i=1}^{N}(X_i - \overline{X})^2 = \sum_{i=1}^{N}[(X_i - \mu) - (\overline{X} - \mu)]^2$$

$$= \sum_{i=1}^{N}[(X_i - \mu)^2 - 2(X_i - \mu)(\overline{X} - \mu) + (\overline{X} - \mu)^2]$$

$$= \sum_{i=1}^{N}(X_i - \mu)^2 - 2(\overline{X} - \mu)\sum_{i=1}^{N}(X_i - \mu) + N(\overline{X} - \mu)^2$$

$$= \sum_{i=1}^{N}(X_i - \mu)^2 - 2N(\overline{X} - \mu)^2 + N(\overline{X} - \mu)^2$$

$$= \sum_{i=1}^{N}(X_i - \mu)^2 - N(\overline{X} - \mu)^2$$

PCA-背景知识

$$E(S^2) = \sum_{i=1}^{N} E[(X_i - \mu)^2] - NE[(\bar{X} - \mu)^2]$$

$$= N\sigma^2 - NE[(\frac{1}{N}\sum_{i=1}^{N} X_i - \mu)^2]$$

$$= N\sigma^2 - NE[\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2]$$

$$= N\sigma^2 - \sigma^2 = (N-1)\sigma^2$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)}}$$

PCA-背景知识

- 方差(Var)

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)}$$

- 协方差(Cov)
  - 在两维数据之间计算，分析不同维数据之间的联系

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

PCA-背景知识

	Hours(H)	Mark(M)	Hi - H^	Mi - M^	(Hi - H^) * (Mi - M^)
Data	9	39	-4.92	-23.42	115.2264
	15	56	1.08	-6.42	-6.9336
	25	93	11.08	30.58	338.8264
	14	61	0.08	-1.42	-0.1136
	10	50	-3.92	-12.42	48.6864
	18	75	4.08	12.58	51.3264
	0	32	-13.92	-30.42	423.4464
	16	85	2.08	22.58	46.9664
	5	42	-8.92	-20.42	182.1464
	19	70	5.08	7.58	38.5064
	16	66	2.08	3.58	7.4464
	20	80	6.08	17.58	106.8864
Totals	167	749			1352.4168
Averages	13.92	62.42			112.7014

PCA-背景知识

- 协方差矩阵

$$C^{n\times n} = (c_{i,j},\ c_{i,j} = cov(Dim_i, Dim_j))$$

- 假设有一个三维的数据集合，则协方差矩阵为

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$

PCA-背景知识

- 特征矢量

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 11 \\ 5 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4 \times \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

- 性质
  - 特征矢量是针对方阵的
  - 并不是所有的方阵都有特征矢量
  - 特征矢量的个数和方阵的维数相同

PCA-背景知识

- 性质...
  - 如果对特征矢量进行缩放，和矩阵相乘后，会得到和缩放前同样的倍数关系

$$2 \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 6 \\ 4 \end{pmatrix} \quad \begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 6 \\ 4 \end{pmatrix} = \begin{pmatrix} 24 \\ 16 \end{pmatrix} = 4 \times \begin{pmatrix} 6 \\ 4 \end{pmatrix}$$

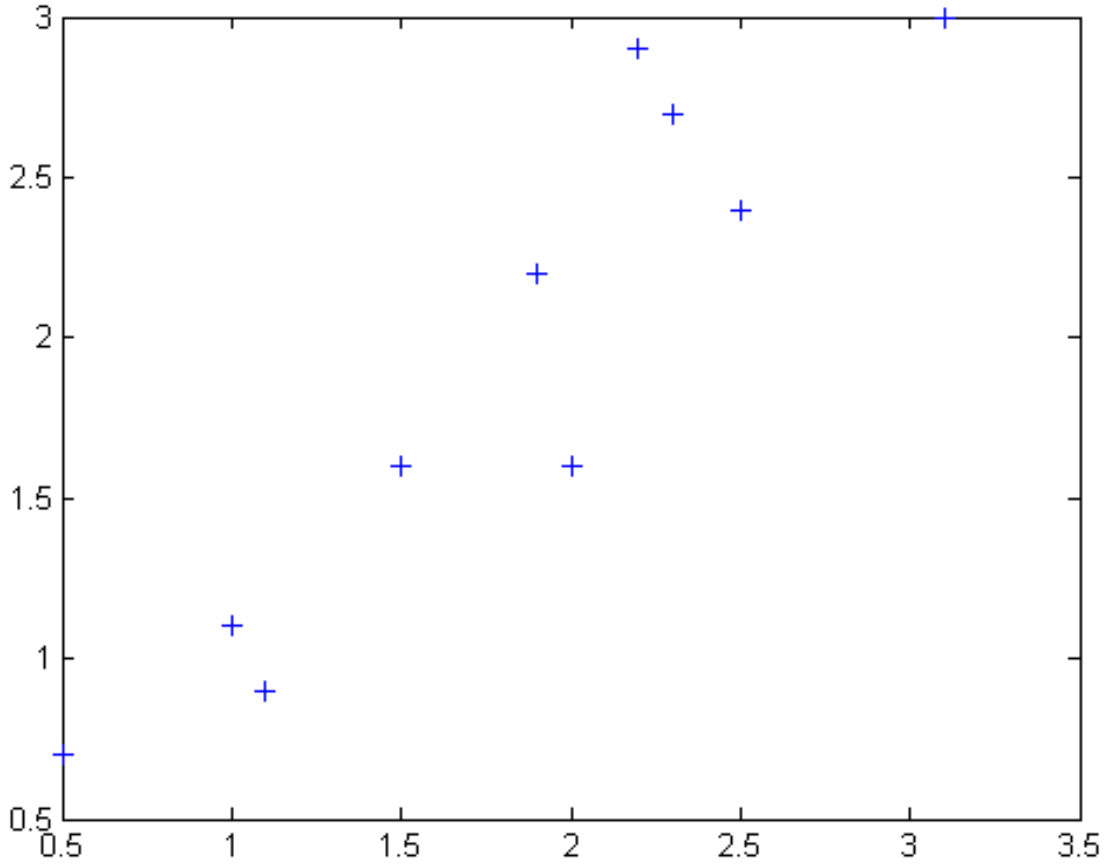
  - 所有特征矢量是正交的
- 为了研究方便，通常将特征矢量都归一化

$$\begin{pmatrix} 3 \\ 2 \end{pmatrix} \div \sqrt{13} = \begin{pmatrix} 3/\sqrt{13} \\ 2/\sqrt{13} \end{pmatrix}$$

PCA

- 步骤一：获取数据

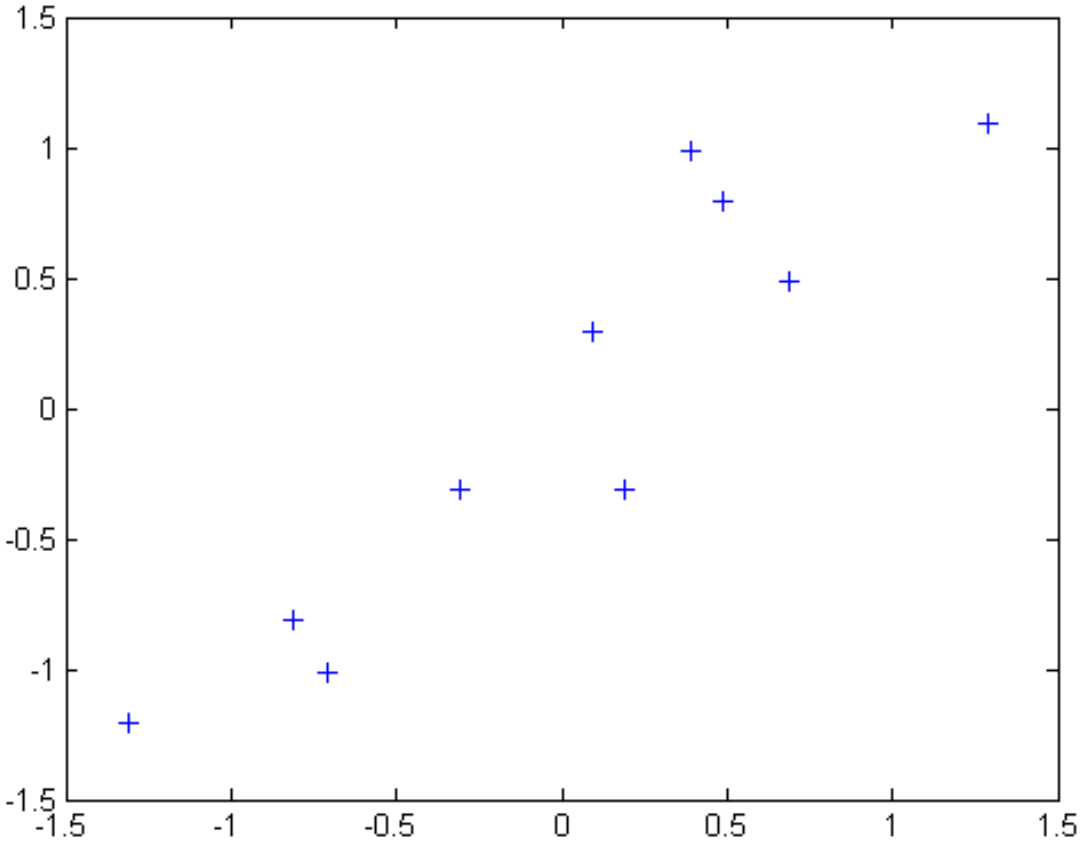
x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9



PCA

- 步骤二：减去平均值

x	y
0.69	0.49
-1.31	-1.21
0.39	0.99
0.09	0.29
1.29	1.09
0.49	0.79
0.19	-0.31
-0.81	-0.81
-0.31	-0.31
-0.71	-1.01



PCA

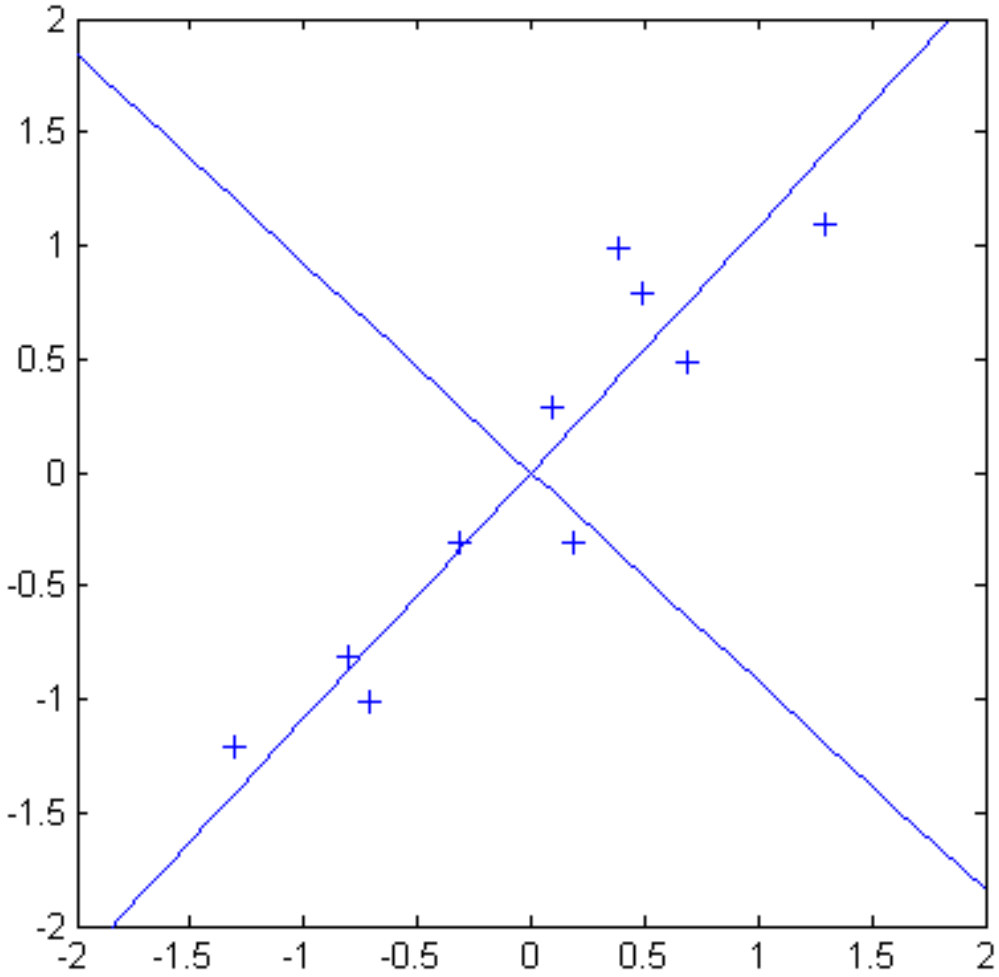
- 步骤三：计算协方差矩阵

$$\text{cov} = \begin{pmatrix} 0.617, & 0.615 \\ 0.615, & 0.717 \end{pmatrix}$$

- 步骤四：计算协方差矩阵的特征矢量和特征值

$$\text{eigenvalues} = \begin{pmatrix} 0.049 \\ 1.284 \end{pmatrix} \qquad \text{eigenvectors} = \begin{pmatrix} -0.735, & -0.678 \\ 0.678, & -0.735 \end{pmatrix}$$

PCA



PCA

- 步骤五：选择分量组成特征矢量
  - 按照特征值大小排序
  - 保留重要的分量
  - 组成特征矢量

$$\text{FeatureVector} = (eig_1\ eig_2\ eig_3\ \ldots.\ eig_n)$$

$$\begin{pmatrix} -0.678, & -0.735 \\ -0.735, & 0.678 \end{pmatrix} \qquad \begin{pmatrix} -0.678 \\ -0.735 \end{pmatrix}$$

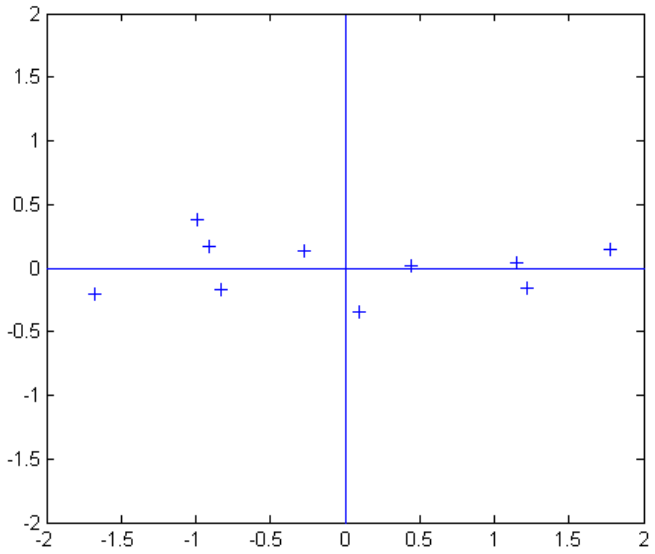
PCA

- 步骤六：推导新的数据集

FinalData = RowFeatureVector x RowDataAdjust

- 保留两个特征矢量

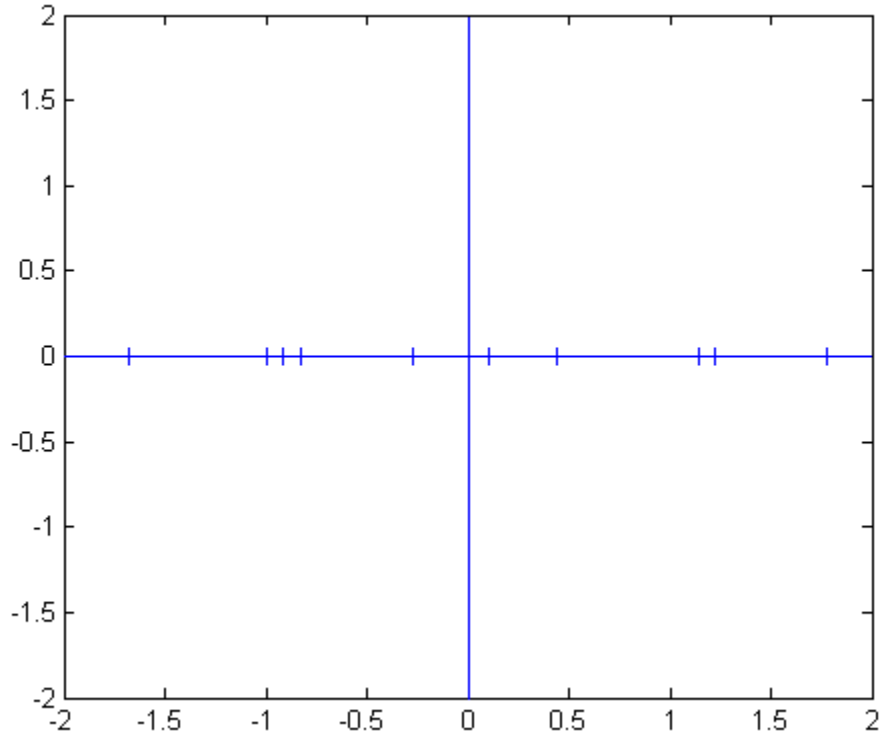
-0.83	1.78	-0.99	-0.27	-1.68	-0.91	0.1	1.14	0.44	1.22
-0.18	0.14	0.38	0.13	-0.21	0.18	-0.35	0.05	0.02	-0.16



PCA

- 保留最重要矢量

-0.83	1.78	-0.99	-0.27	-1.68	-0.91	0.1	1.14	0.44	1.22



PCA

- 恢复原始数据

$$FinalData = RowFeatureVector \times RowDataAdjust$$

- ==》

$$RowDataAdjust = RowFeatureVector^{-1} \times FinalData$$

- ==》

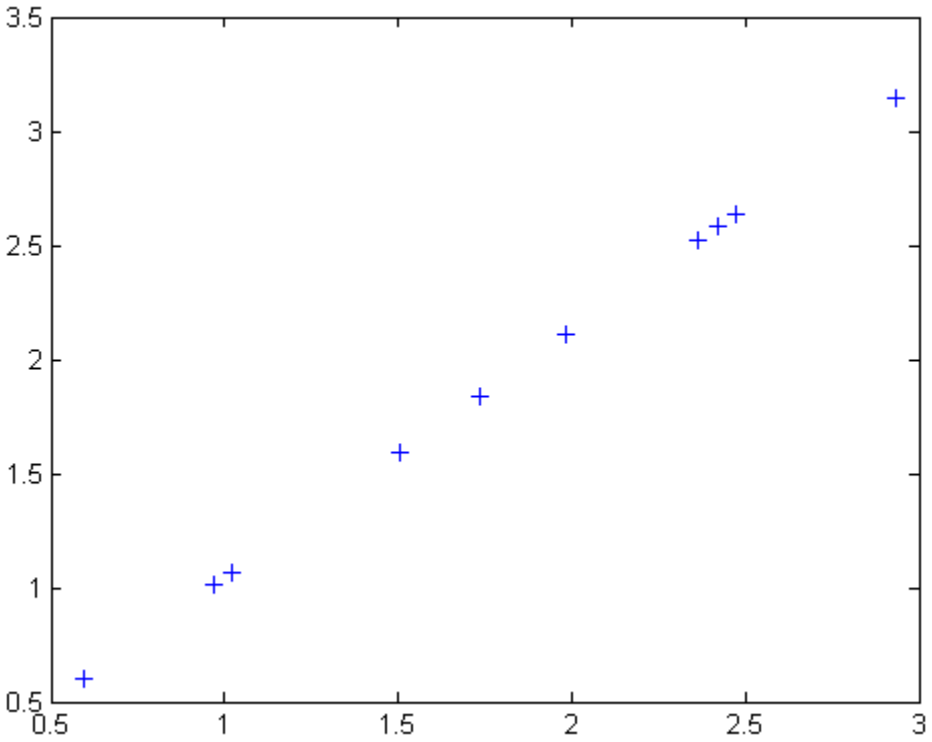
$$RowDataAdjust = RowFeatureVector^{T} \times FinalData$$

- 加上均值

$$RowOriginalData = (RowFeatureVector^{T} \times FinalData) + OriginalMean$$

PCA

- 重建数据

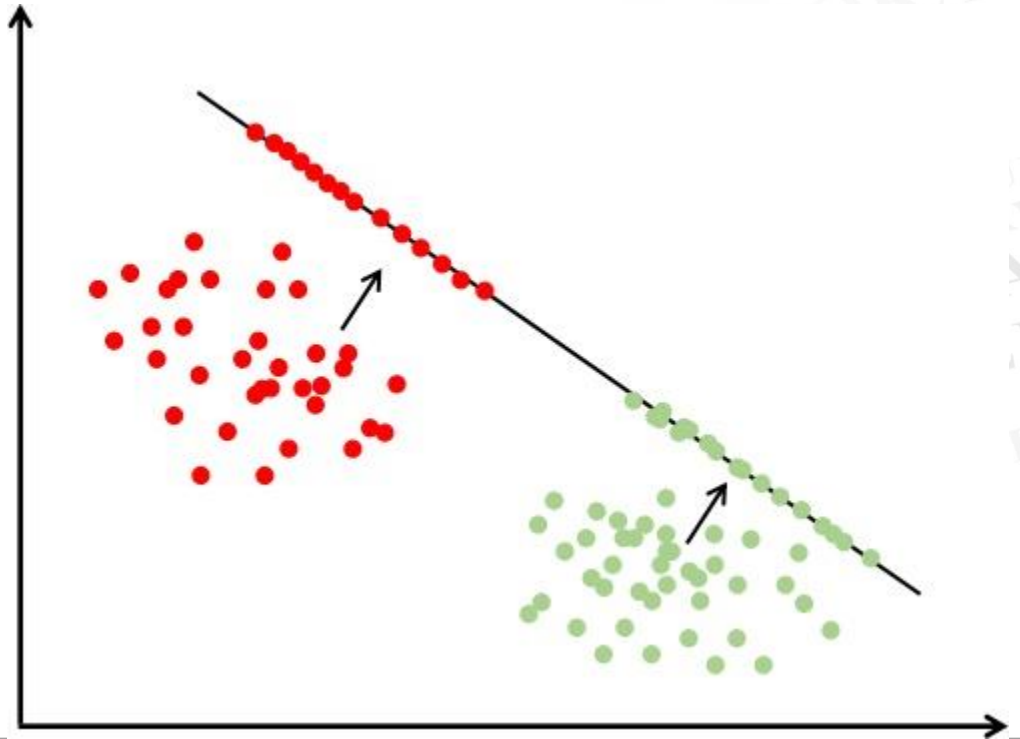


奇异值分解

- **SVD (Singular Value Decomposition)**
- 特征值分解只能针对方阵处理，**SVD**可以处理任意矩阵的通用情况
- $A=U\Sigma V^T$
- 如果**A**是**N*M**的矩阵，则**U**是**N*N**的方阵，**Σ**是**N*M**的矩阵，V^T是**N*N**的矩阵
- 左奇异向量，奇异值，右奇异向量

LDA

- 线性判别分析
  - **Linear Discriminant Analysis**
- 向最大化类间差异、最小化类内差异的方向线性投影



LDA

- 优化目标

$$L(\mathrm{W}) = \frac{\left|\mathrm{W}^{\mathrm{T}} \mathrm{S}_B \mathrm{W}\right|}{\left|\mathrm{W}^{\mathrm{T}} \mathrm{S}_W \mathrm{W}\right|}$$

- **LDA是一种有监督的机器学习算法**
- 不能直接用于分类和回归问题，需要借助其他算法对降维后的向量进行分类

数据块消减

- 数据块消减方法主要包含参数和非参数两种基本方法，来取代原有的数据
- 参数方法利用一个模型来表示原有的数据
  - 线性回归：**Y=a+bX**
  - 模型的准确程度决定了消减后数据的准确程度，不够灵活
- 非参数方法则是存储利用直方图、聚类或取样而获得的消减后数据
  - 与数据本身的分布无关

直方图

- 直方图就是根据属性的数据分布将其分成若干不相交的区间，每个区间的高度与其出现的频率成正比
- 价格清单

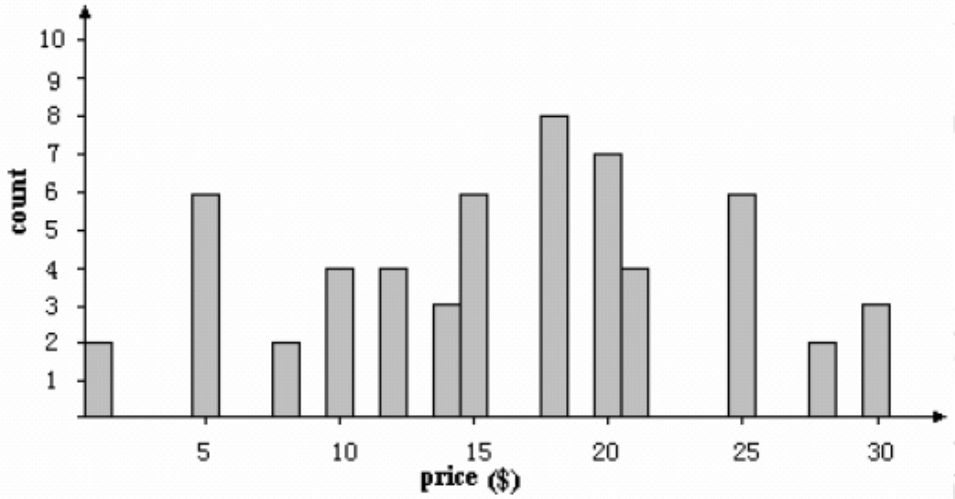
1(2), 5(5), 8(2)

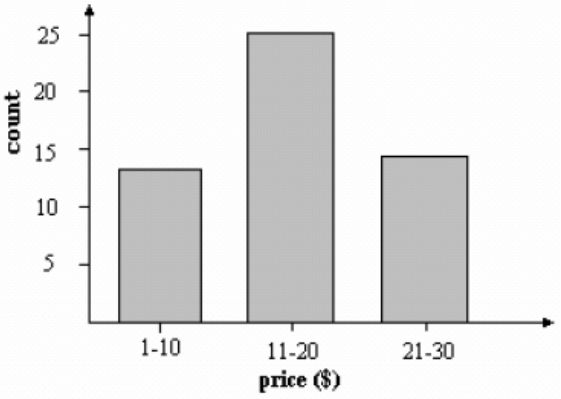
10(4), 12, 14(3)

15(5), 18(8), 20(7)

21(4), 25(5), 28

30(3)





直方图

- 构造直方图的方法有以下几种：
- 等宽法
- 等高法
- **V-Optimal法**
  - 在指定bin个数的情况下，所有可能的直方图中，变化最小的直方图
  - **Divide {1,5,8,10} into 3 buckets**
  - **Optimal:{1},{5},{8,10}**

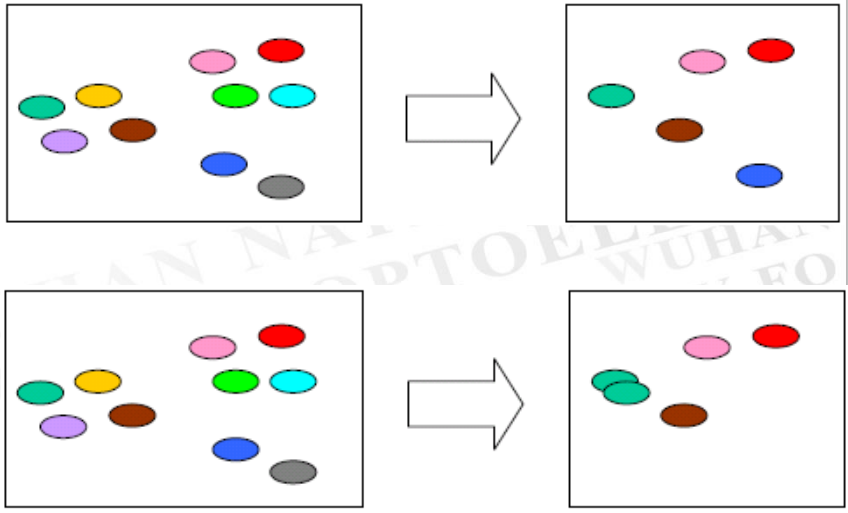
$$\text{variance} = \frac{1}{n-1}\sum_i \left(x_i - \bar{x}\right)^2$$

聚类

- 聚类是将原数据集合分成由类似的数据组成的多个类的过程
- 同类中的数据彼此相似，而不同类中的数据彼此不相似
- 相似通常利用空间中的距离来衡量
  - **Diameter:**聚类中任意两个对象间的最大距离
  - **Centroid distance:**类中每个对象到质心的平均距离
- 聚类的有效性依赖于实际数据的内在规律

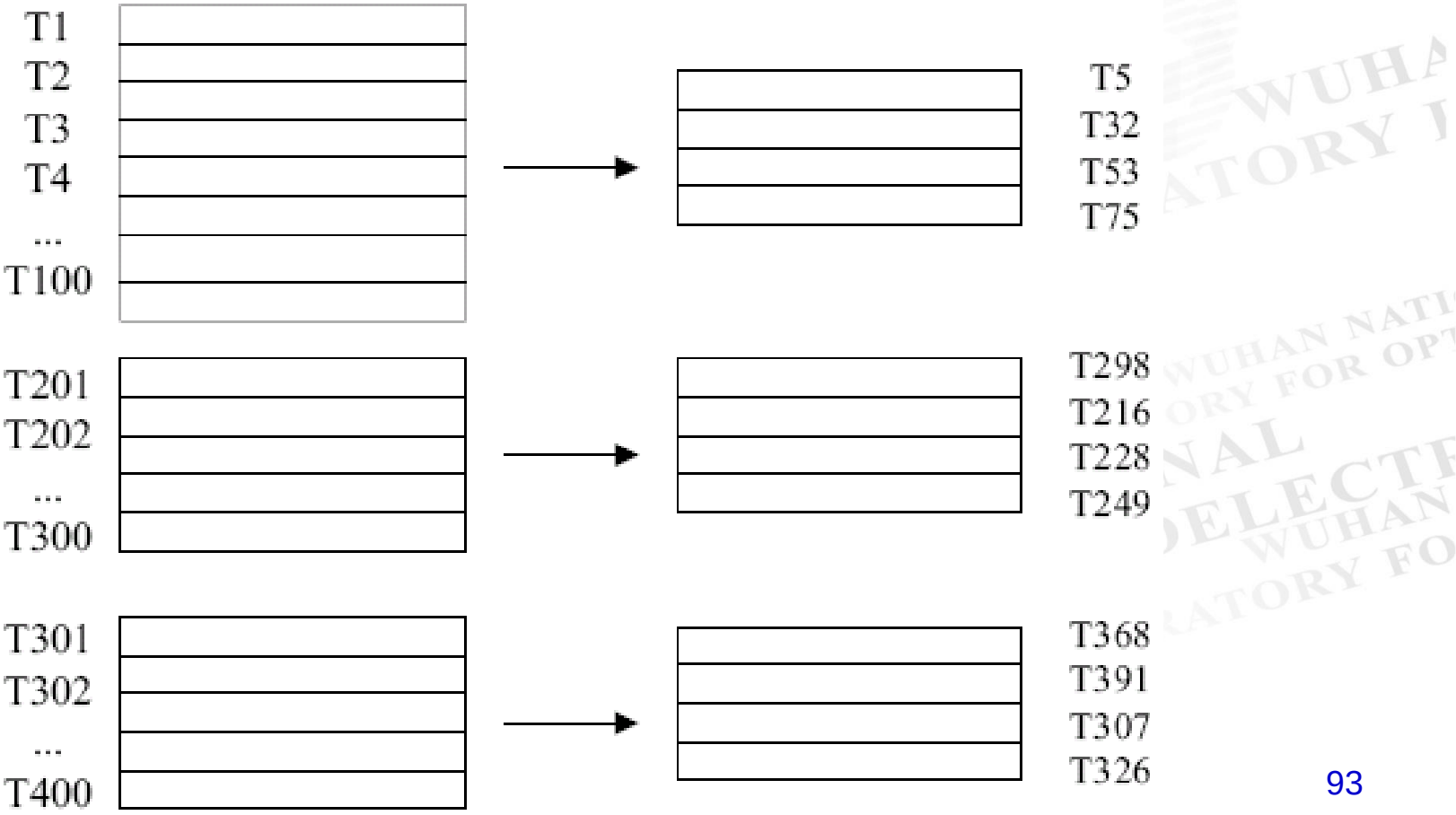
采样

- 采样是利用一小部分数据（子集）来代表一个大数据集
- 假设一个大数据集为D,其中包括N个数据，几种主要采样方法有：
- 1. 无替换简单随机采样
  - SRSWOR
- 2. 有替换简单随机采样
  - SRSWR

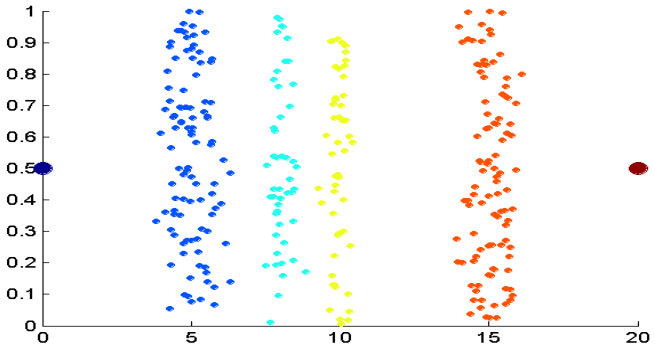


采样

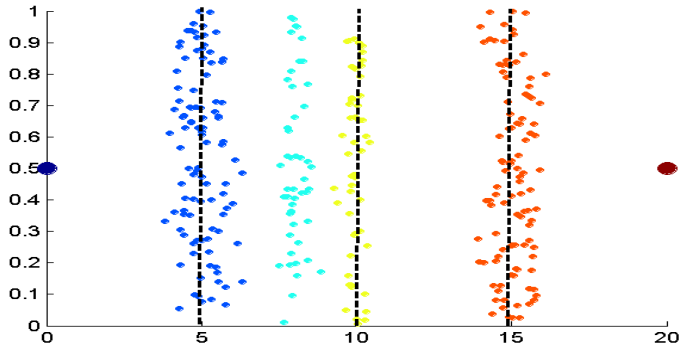
➢3. 分层采样方法


T1
T2
T3
T4
...
T100
T5
T32
T53
T75
T201
T202
...
T300
T298
T216
T228
T249
T301
T302
...
T400
T368
T391
T307
T326

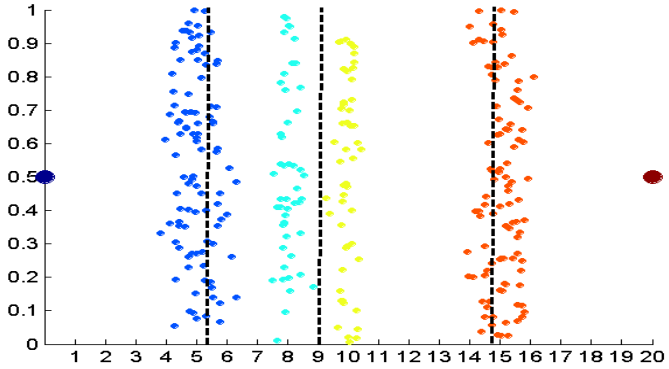
离散化和二值化



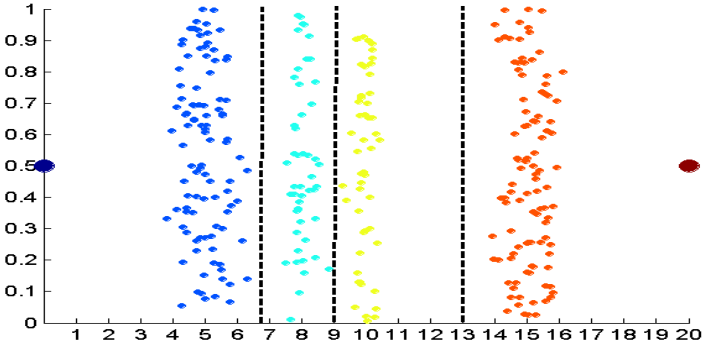
Data



Equal interval width



Equal frequency



K-means

- 什么是数据
- 数据预处理的重要性
- 数据清洗
- 数据集成与转换
- 数据消减
- 相似度和相异度

相似度和相异度

相似度和相异度

- 相似度
  - 两个数据对象相似程度的数值度量
  - 对象越相似，值越大
  - 取值范围通常在[0,1]之间
- 相异度
  - 两个对象差异程度的数值度量
  - 对象越类似，相异度就越低
  - 最小相异度通常为0，上限可变
- 邻近度可用相似度或相异度表示

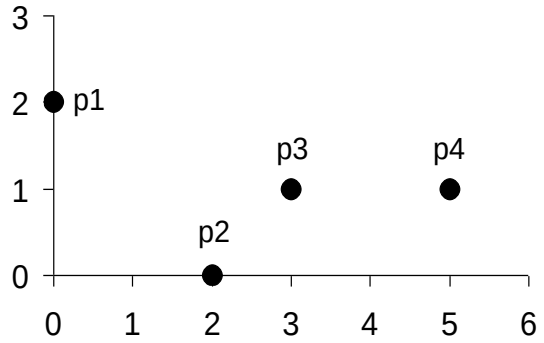
简单属性之间的相似度/相异度

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

数据对象之间的相异度

➢欧几里德距离（Euclidean distance）

$$dist = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

数据对象之间的相异度

- Minkowski距离是欧式距离的普通形式
  - 闵可夫斯基距离

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

- r=1，城市块距离，如汉明距离
- r=2，欧式距离
- r= ∞，上确界距离

Minkowski距离

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

距离的性质

- 1. 非负性： 对于所有的p和q，$d(p, q) \geq 0$ ，只有当p=q时为0；
- 2. 对称性：对于所有的p和q，$d(p, q) = d(q, p)$
- 3. 三角不等式：对于所有的p，q，r，$d(p, r) \leq d(p, q) + d(q, r)$
- 满足上述三个性质的测度称为度量
- 并不是所有的相异度都满足这三个性质

距离的性质

- 集合差：有两个集合A和B，给出如下定义，A-B是不在B中的A中元素的集合
  - 例如，A={1 2 3 4},B={2 3 4}, A-B={1}, B-A={}
  - 距离定义d(A, B)=size(A-B)不满足非负性的第二部分、对称性及三角不等式
  - 距离定义修改为d(A, B)=size(A-B)+size(B-A)则满足所有性质
- 证明！！！

相似度的性质

- 相似度通常具有如下典型性质

- 仅当p=q时，$s(p, q) = 1$
- 对于所有的p和q， $s(p, q) = s(q, p)$

二元矢量的相似性度量

- 设p和q是两个对象，都由n个二元属性组成
- 用下面四个系数来计算相似性
  - M01 = the number of attributes where p was 0 and q was 1
  - M10 = the number of attributes where p was 1 and q was 0
  - M00 = the number of attributes where p was 0 and q was 0
  - M11 = the number of attributes where p was 1 and q was 1
- 简单匹配和Jaccard系数
  - SMC = number of matches / number of attributes

  = (M11 + M00) / (M01 + M10 + M11 + M00)
  - J = number of 11 matches / number of not-both-zero attributes values

  = (M11) / (M01 + M10 + M11)

SMC vs. Jaccard

p = 1 0 0 0 0 0 0 0 0 0
q = 0 0 0 0 0 0 1 0 0 1

M_{01} = 2 (the number of attributes where p was 0 and q was 1)
M_{10} = 1 (the number of attributes where p was 1 and q was 0)
M_{00} = 7 (the number of attributes where p was 0 and q was 0)
M_{11} = 0 (the number of attributes where p was 1 and q was 1)

SMC = $(M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00})$ = (0+7) / (2+1+0+7) = 0.7

J = (M_{11}) / $(M_{01} + M_{10} + M_{11})$ = 0 / (2 + 1 + 0) = 0

余弦相似度

- 如果d_1和d_2是两个文档矢量，则它们之间的余弦相似度为
  - $\cos(d_1, d_2) = (d_1 \bullet d_2) / \|d_1\| \|d_2\|$
- 例子：

d1 = 3 2 0 5 0 0 0 2 0 0
d2 = 1 0 0 0 0 0 0 1 0 2

d1 • d2= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5
||d1|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^0.5 = 6.481
||d2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^0.5 = 2.245

cos(d1, d2) = 0.3150

相关性

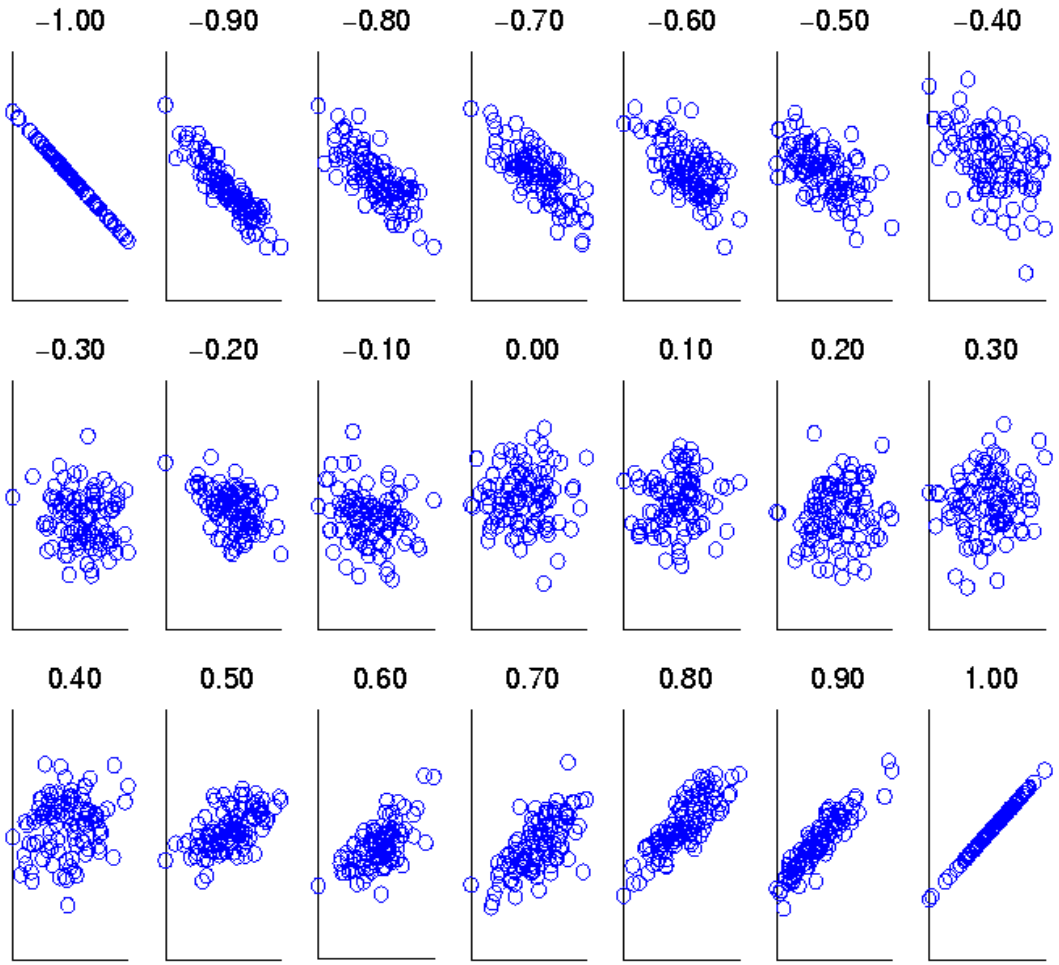
- 两个具有二元变量或连续变量的数据对象之间的相关性是对象属性之间的线性联系的度量
- 皮尔森相关：衡量两个随机变量的相关性

$$p'_k = (p_k - mean(p)) / std(p)$$

$$q'_k = (q_k - mean(q)) / std(q)$$

$$correlation(p,q) = p' \bullet q'$$

相关性可视化


−1.00
−0.90
−0.80
−0.70
−0.60
−0.50
−0.40
−0.30
−0.20
−0.10
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00

邻近度计算问题

- 1. 当属性具有不同尺度或相关时如何处理
- 2. 当对象包含不同类型的属性时如何计算对象之间的邻近度
- 3. 当属性具有不同的权重时如何处理

邻近度计算问题

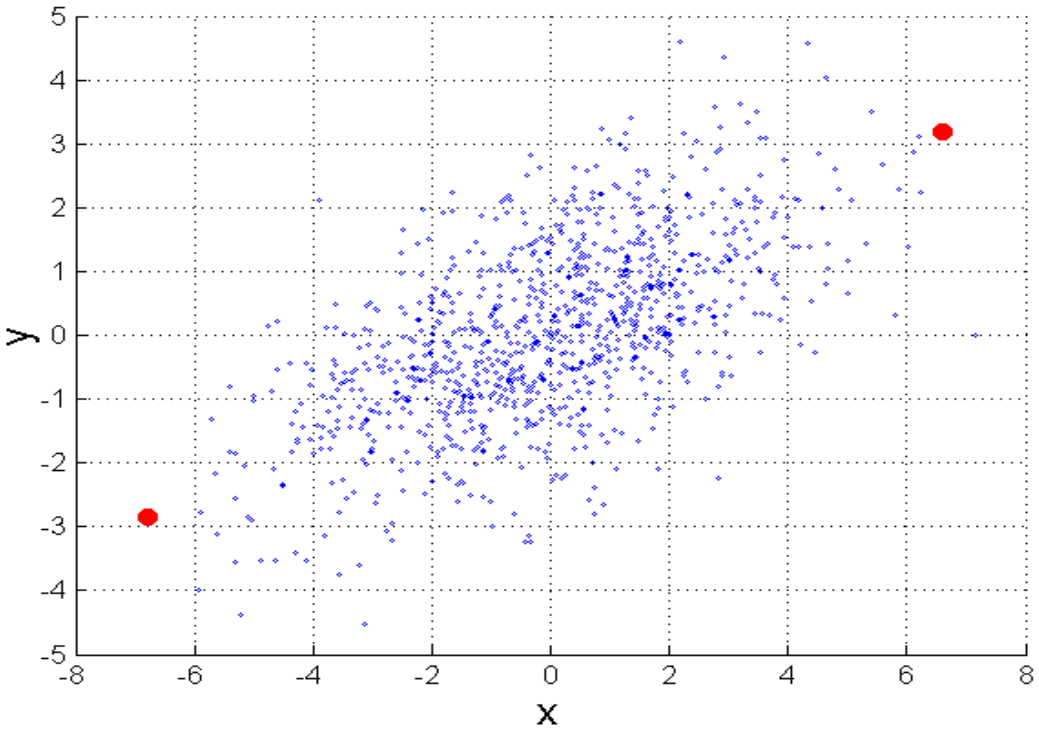
- **Mahalanobis是一种概率意义上的距离**
  - 马氏距离

$$d(x,y)=\sqrt{(x-y)^T S(x-y)}$$

- 要保证根号内的值非负，即矩阵**S**必须是半正定的
- 当矩阵**S**为单位矩阵时，**Mahalanobis**距离退化为欧氏距离
- **S**可以通过训练样本集的协方差矩阵得到，也可以通过训练样本学习得到

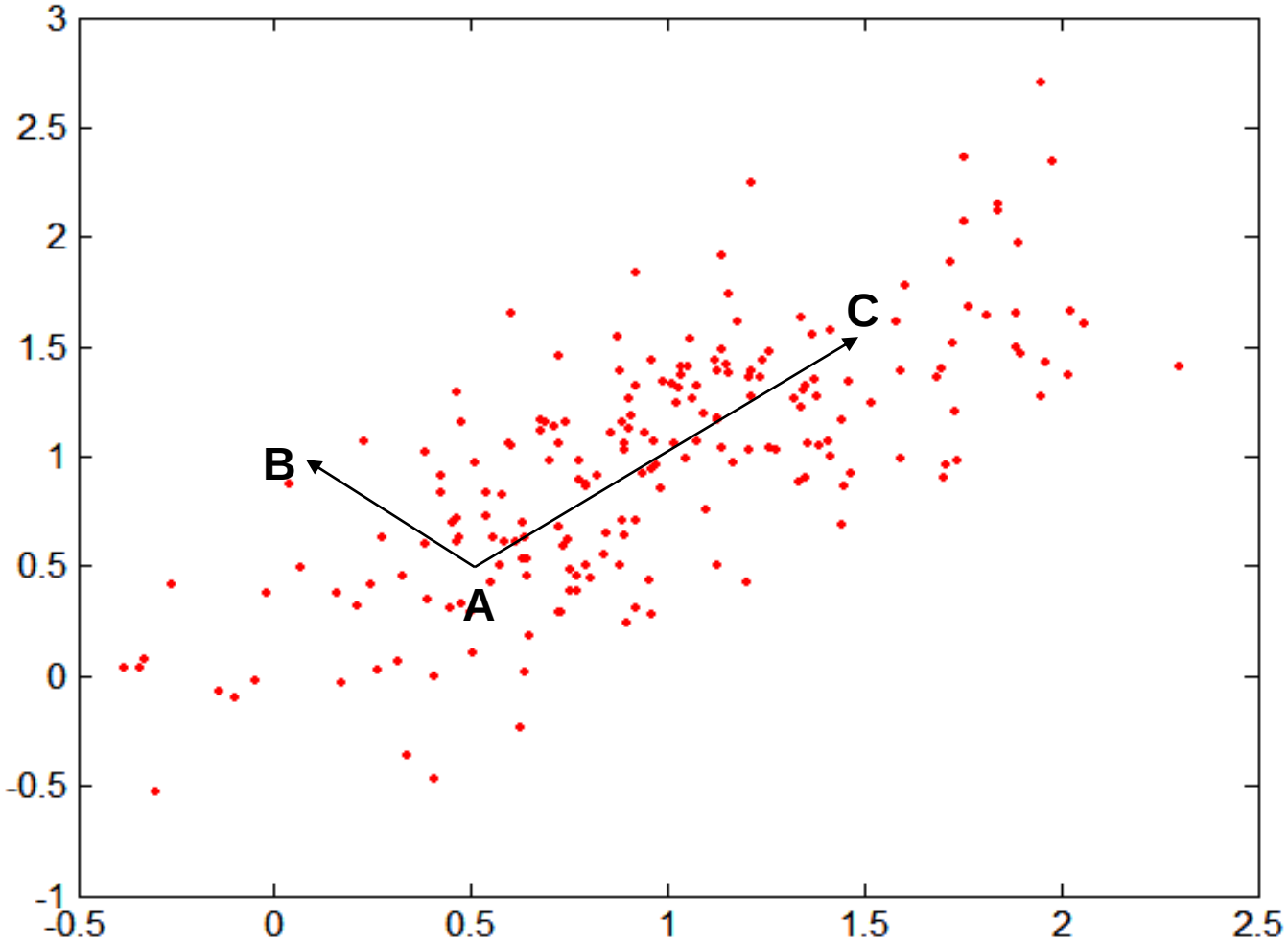
邻近度计算问题

$$\Sigma_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)$$



红点之间的欧式距离为14.7, Mahalanobis距离为6.

邻近度计算问题



Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

组合不同属性的相似度

- 有时候属性具有许多不同的类型，需要将其组合起来得到一个统一的相似度

1. For the k^{th} attribute, compute a similarity, s_k, in the range $[0, 1]$.
2. Define an indicator variable, δ_k, for the k_{th} attribute as follows:

$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is a binary asymmetric attribute and both objects have a value of 0, or if one of the objects has a missing values for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$

3. Compute the overall similarity between the two objects using the following formula:

$$similarity(p, q) = \frac{\sum_{k=1}^{n} \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$

使用权值组合相似度

- 不同属性对相似度的贡献（重要性）不同
  - 使用权重w_k，其值在[0,1]之间，之和为1

$$similarity(p,q) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$

$$distance(p,q) = \left(\sum_{k=1}^{n} w_k |p_k - q_k|^r \right)^{1/r}$$