智能制造过程大数据技术

Big Data Technology in Intelligent Manufacturing Process

第三讲：数据预处理

Lecture 3: Data preprocessing

丁敏 dingmin@cug.edu.cn

中国地质大学(武汉) 自动化学院

School of Automation, China University of Geosciences

为什么要进行数据预处理?

- 初始数据集的准备和变换是数据挖掘过程中重要的步骤
- 包含大量不完整、含噪声和不一致的数据是大数据应用中典型特点
- 数据的预处理能有效提高数据质量，节约大量的时间和空间
- 大部分数据挖掘算法对输入数据的格式、质量以及规模有一定的要求

目录

- 工业数据质量
- 数据清洗
- 数据集成
- 数据变换
- 数据归约

IEEE INDUSTRIAL ELECTRONICS MAGAZINE ■ DECEMBER 2019 PP.38

目录

- 工业数据质量
- 数据清洗
- 数据集成
- 数据变换
- 数据归约

IEEE INDUSTRIAL ELECTRONICS MAGAZINE ■ DECEMBER 2019 PP.38

➢ 数据质量

◘ 数据质量决定是否能满足应用需求，涉及许多因素

- 准确性：具有正确的属性值
- 完整性：存储在数据库中的所有数据值均正确的状态
- 一致性：关联数据之间的逻辑关系正确和完整
- 时效性：数据是否及时更新
- 可信性：用户的信赖程度
- 可解释性：是否被容易理解

高质量的数据质量有利于保证高效的数据挖掘

工业数据存在的问题

- 不完整：缺失属性、缺失数值

属性 / 时间	实际功率	实际风速	
0：15	16. 69767	5. 99	65. 03
0：30	24. 82667	7. 07067	79. 61
0：45		7. 16167	87. 08

- 不正确：错误值或者异常值

属性 / 时间	实际功率	实际风速	预测风速
0：15	16. 69767	5. 99	2. 94
0：30	24. 82667	7. 07067	4
0：45	-1000	7. 16167	4. 99

- 不一致：相同属性对应的值错误，逻辑错误

属性 / 时间	实际功率	实际风速	预测风速	工作模态
0：15	16. 69767	5. 99	2. 94	1
0：30	24. 82667	7. 07067	4	3
0：45	24. 82667	7. 16167	4. 99	A

1 工业数据质量问题

- 正确的数据:
  - ✓ John Doe | john.doe@mail.com | 123 Main Street
- “脏数据”举例:

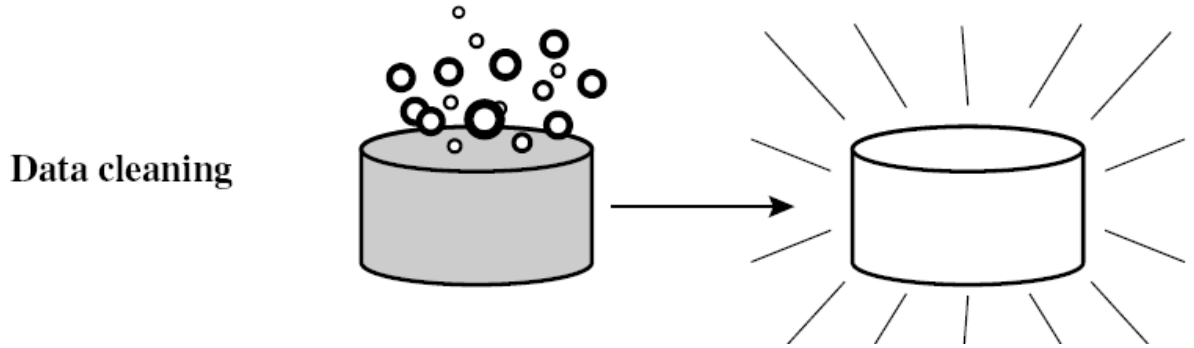
- Duplicate:
  - John Doe | john.doe@mail.com | 123 Main Street
  - John Doe | john.doe@mail.com | 123 Main St.
- Inaccurate:
  - John Doe | john.doe@mail.com | 132 Main Street
- Dated:
  - John Doe | john.doe@old_email.com | 123 Main Street

1 工业数据质量问题

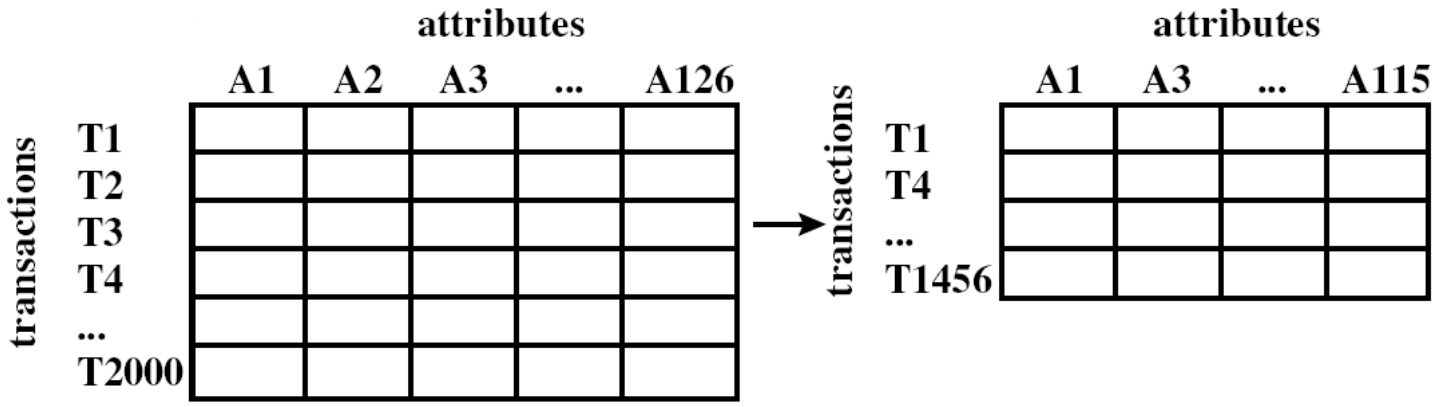
- 预处理为什么是重要的?
  - "No quality data, no quality mining results!"
    - ✓ 数据中存在的不一致以及噪声，对很多数据挖掘算法影响较大，甚至“挖掘”出错误的知识
    - ✓ 很多挖掘算法对于数据的分布等条件有限制，需要预先处理
    - ✓ 数据维数过高会引起“维数灾难”或者“过拟合”，需要进行降维等预处理
  - "It is often postulated that 50-70 percent of the time and effort in a data mining project is used in the Data Preparation Phase" -------CRISP-DM

1 工业数据质量问题

- 数据预处理
  - 主要有以下几个任务：
    - 数据清洗
    - 数据集成
    - 数据变换
    - 数据归约



-2,32,100,59,48→-0.02,0.32,1.00,0.59,0.48



目录

- 工业数据质量问题
- 数据清洗
- 数据集成
- 数据变换
- 数据归约

IEEE INDUSTRIAL ELECTRONICS MAGAZINE ■ DECEMBER 2019 PP.38

➢ 数据清洗

- 数据清洗主要用于填充缺失的值、光滑噪声并识别异常值、纠正数据中的不一致。
  - 格式内容
  - 缺失值
  - 噪声数据
  - 异常值
  - 逻辑错误

格式内容清洗

- 系统日志中，人工收集或用户填写时可能存在格式内容问题
- 种类繁多的大数据存在异构数据情况
- 内容中有不该存在的字符：数据中存在字母、字符等

方法	简介
时间、日期、数值、全半角等显示格式不一致	通常由输入端造成，处理为一致的格式
内容中有不该存在的字符	例如数据中有不应存在的汉字，字母
内容与该字段应有内容不符	例如将温度数据和压强数据互换问题

2 数据清洗

格式错误清洗实例

属性 / 时间	实际功率	实际风速	预测风速	风向	温度	压强	湿度
0：15	16.69767	5.99	2.94	65.03	3.78	92281.6	82.81
0：30	24.82667	7.07067	4	79.61	3.66C°	92269.85	83.44
0：45	NAN	7.16167	4.99	87.08	3.61	92263.59	83.74
1：00	22.32567	6.84833	5.13	92.13	3.63	92260.14	83.9
1：15	20.24833	6.566	5.01	93.25	3.58C°	92262.38	84.3
1：30	18.85633	6.26267	5.25	90.16	3.5	92262.66	84.66
1：45	23.14433	6.83867	5.51	89.01	3.47	92263.88	84.79
2：00	-1000	0	5.64	89.74	3.39	92263.3	85.15
2：15	16.13933	0.95967	5.65	91.5	3.28	92261.36	85.6

温度
3.78
3.66
3.61
3.63
3.58
3.5
3.47
3.39
3.28

缺失值

- 缺失值是指粗糙数据中由于缺少信息而造成的数据的聚类、分组、删失或截断。它指的是现有数据集中某个或某些属性的值是不完全的。
  - ✓ 工业数据中样本的个别属性会出现缺失的情况，如：NaN、None或空缺值等。

TIME 时间	AVG（CBV） 冷风流量	AVG（CBP） 冷风压力
2015-06-01 00:06	5631.67	440.34
2015-06-01 00:07	5621.07	439.89
2015-06-01 00:08		438.63
2015-06-01 00:09	5663.64	433.20
2015-06-01 00:10	5667.24	NaN
2015-06-01 00:11	5658.63	428.45

时间	产量	运行模态
12/26	7.6	1
12/27	9.3	2
12/28	8.1	1
12/29	8.6	1
12/30		1
12/31	10.4	2

缺失值处理方法

- ✓ 忽略样本：通常使用在缺少类标号样本时
- ✓ 人工填写缺失值：费事，并且当数据集很大、缺失很多值时该方法可能行不通
- ✓ 使用一个全局常量填充缺失值：将缺失的属性值用同一个常量替换
- ✓ 使用属性的中心度量填充缺失值：正常数据用均值填充，倾斜数据则用中位数填充
- ✓ 使用与给定样本同一类的属性均值或中位数：如某一时刻的产量缺失，用同一运行模态下的其他时刻的均值或者中位数来补充
- ✓ 使用最可靠的值填充缺失值：可以用回归、贝叶斯形式化方法的基于推理的工具或决策树归纳确定

➢ 缺失值清洗方法案例

◘ 例1：使用属性的中心度量填充缺失值

- 将变量的属性分为数值型和非数值型来分别进行处理
- 如果缺失值是数值型的，就根据该变量在其他所有对象的取值的平均值来填充该缺失的变量值
- 如果缺失值是非数值型的，则使用众数来补齐该缺失的变量值

属性 / 时间	实际功率	实际风速	预测风速
0：15	16. 69767	5. 99	2. 94
0：30	24. 82667	7. 07067	4
0：45	NAN	7. 16167	4. 99
1：00	22. 32567	6. 84833	5. 13

求三个时刻实际功率的均值

实际功率
16. 69767
24. 82667
20. 61637
22. 32567

这方法是建立在完全随机缺失的假设之上，会造成变量的方差和标准差变小，用与元组属于同一个“类别”的元组的均值填充

➢ 缺失值清洗

◘ 例2：使用与给定样本同一类的属性均值或中位数

- 某一时刻的产量缺失，则用同一运行模态下的其他时刻的均值或者中位数来补充

属性 / 时间	产量	运行模态
12/26	7.6	1
12/27	9.3	2
12/28	8.1	1
12/29	8.6	1
12/30		1
12/31	10.1	2

求三个同一模态产量的均值

产量
7.6
9.3
8.1
8.6
8.1
10.1

多选题 3分

设置

一般常见的缺失值处理方法有

A 替换法

B 最近邻插补

C 回归法

D 插值法

提交

➢ 噪声数据

- 最简单的平滑方法：分箱法
  - ✓ 存储的值被分布到一些“桶”或箱中
  - ✓ 分箱方法参考近邻的值，因此它属于局部平滑
- 分箱方法步骤

① 首先排序数据，并将它们分到等频（等深）的箱中

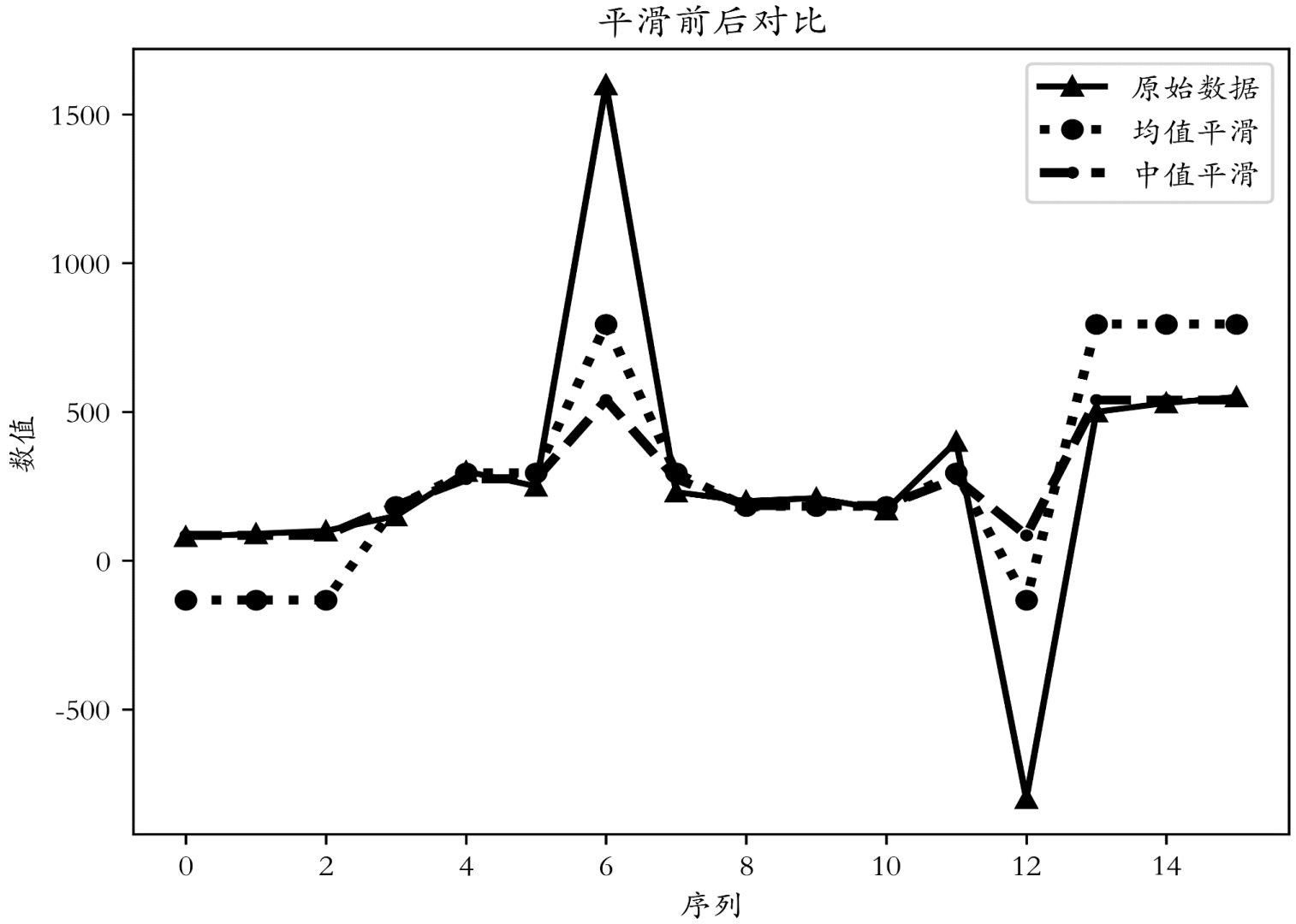
② 平滑各个分箱中的数据：

- ✓ 箱均值平滑：箱中每一个值用箱的平均数替换
- ✓ 箱中位数平滑：箱中每一个值用箱的中位数替换
- ✓ 箱边界平滑：箱中每一个值用离它最近的箱边界值替换

分箱法进行数据平滑实例

原始数据	80，90，100，150，300，250，1600，230，200，210，170，400，-800，500，530，550			
排序后	-800，80，90，100，150，170，200，210，230，250，300，400，500，530，550，1600			
等频分箱	-800，80，90，100	150，170，200，210	230，250，300，400	500，530，550，1600
箱均值平滑	-132.5	182.5	295	795
平滑后	-132.5，-132.5，-132.5，182.5，295.0，295.0，795.0，295.0，182.5，182.5，182.5，295.0，-132.5，795.0，795.0，795.0			
箱中位数平滑	85	185	275	540
平滑后	85.0，85.0，85.0，185.0，275.0，275.0，540.0，275.0，185.0，185.0，185.0，275.0，85.0，540.0，540.0，540.0			

- 分箱法进行数据平滑实例

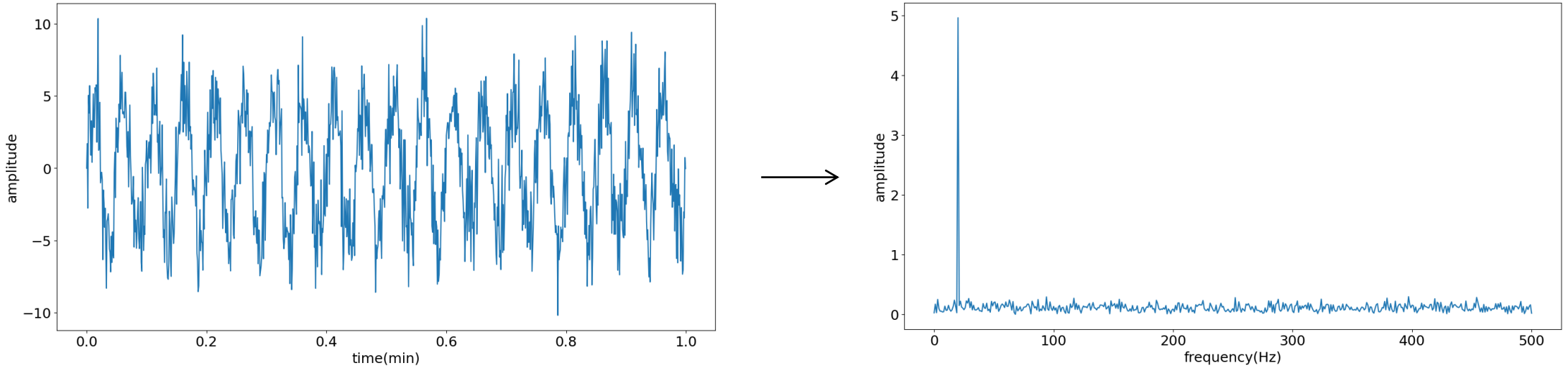


➢ 噪声数据

◘ 傅立叶变换（时间序列）

傅立叶变换（Fourier Transform，简称FT）常用于数字信号处理，它的目的是将时间域上的信号转变为频率域上的信号： $f(t) = a_0 + \sum_{i=1}^{N}(a_i cosw_i t + b_i sinw_i t)$

经傅立叶变换后得到信号的频谱图，可以观察各频率对应的分量



➢ 噪声数据

◘ 傅立叶变换去噪

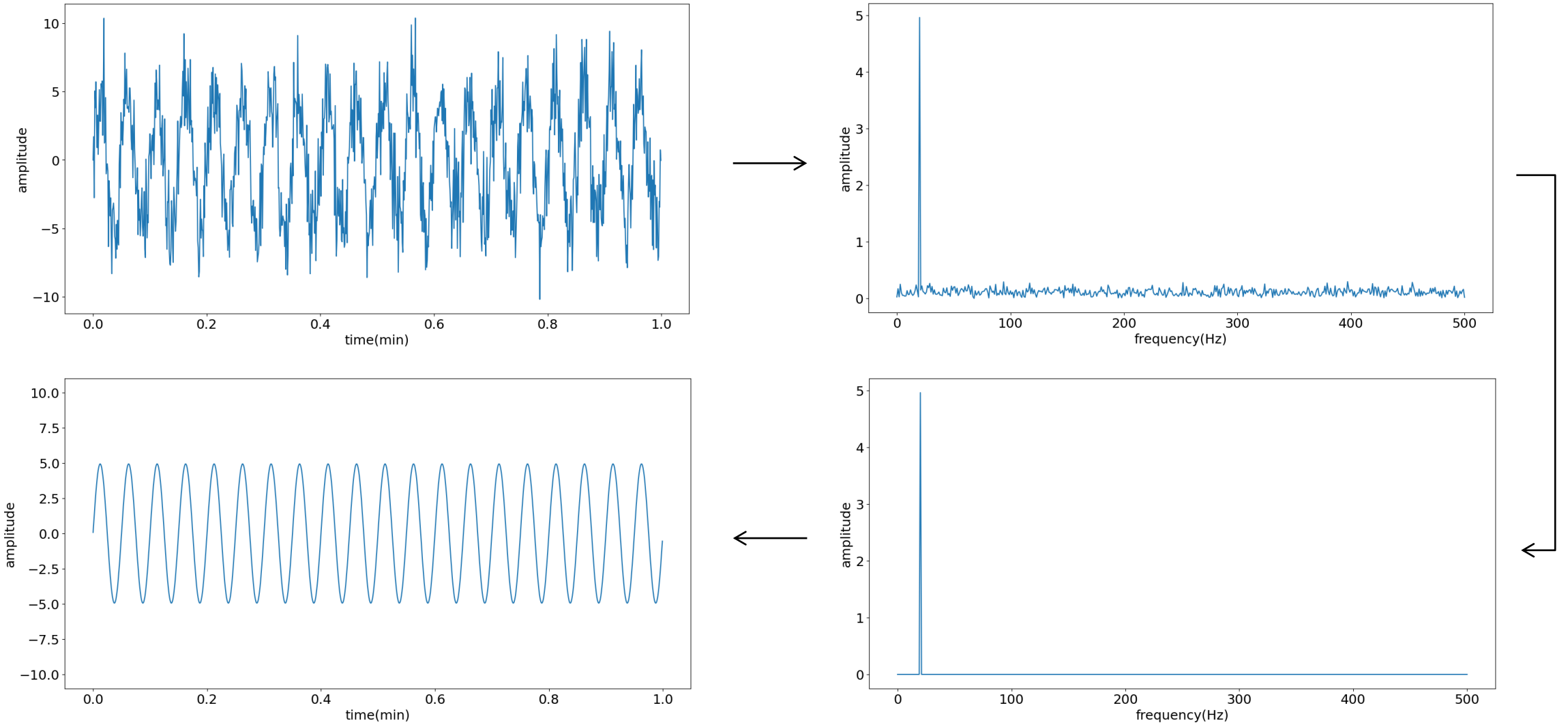
傅立叶变换可用于信号的频域分析与去噪，基本过程如下：

- 对时域信号x_n进行傅立叶变换，得到各频率下的信号幅值
- 设定阈值，使用低通滤波器或带通滤波器滤除噪声频率
- 使用傅立叶反变换重构时域信号

◘ 傅立叶变换的局限性

- 不能刻画时间域上信号的局部特征
- 对含突变的信号的处理效果不好

- 傅立叶变换去噪实例


amplitude
time(min)
frequency(Hz)
amplitude
frequency(Hz)
amplitude
time(min)

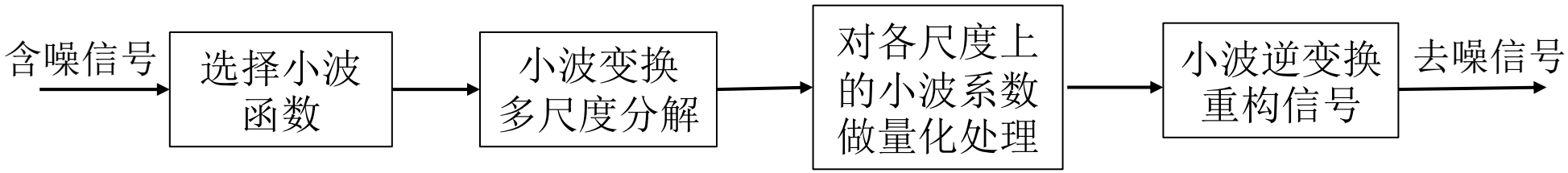
➢ 噪声数据

□ 小波变换

小波变换（wavelet transform，WT）是一种新的变换分析方法，能够在时间（空间）频率的局部化分析，通过伸缩平移运算对信号逐步进行多尺度细化。

□ 小波变换优点：

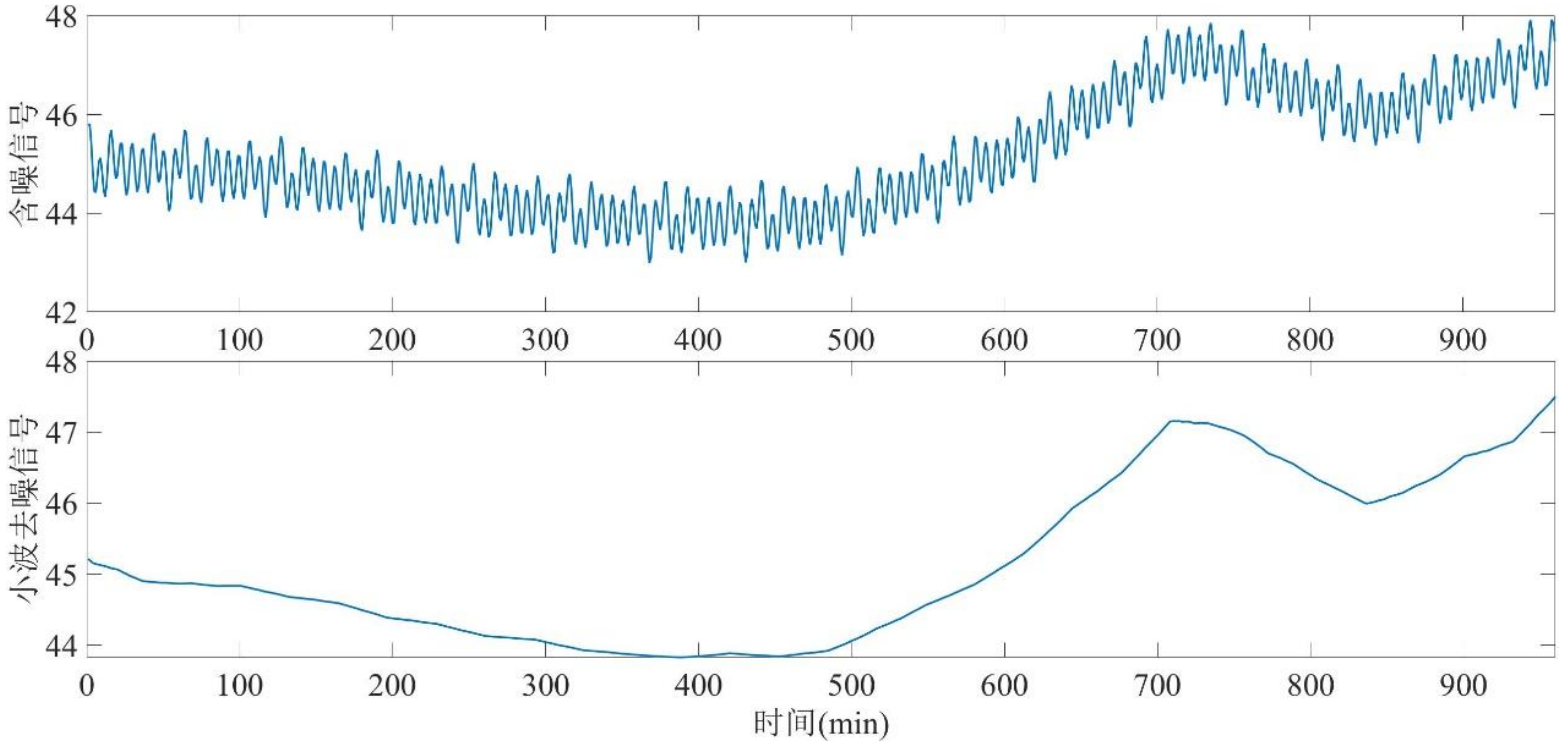
- 小波变换是时间(空间)频率的局部化分析，通过伸缩平移运算对信号(函数)逐步进行多尺度细化，可聚焦到信号的任意细节
- 小波变换去噪可以很好的保护有用的信号尖峰和突变信号



➢ 噪声数据

◘ 小波变换去噪实例

✓ 利用小波变换去噪方法，对某钢铁厂的煤气利用率数据中部分含噪信号去噪

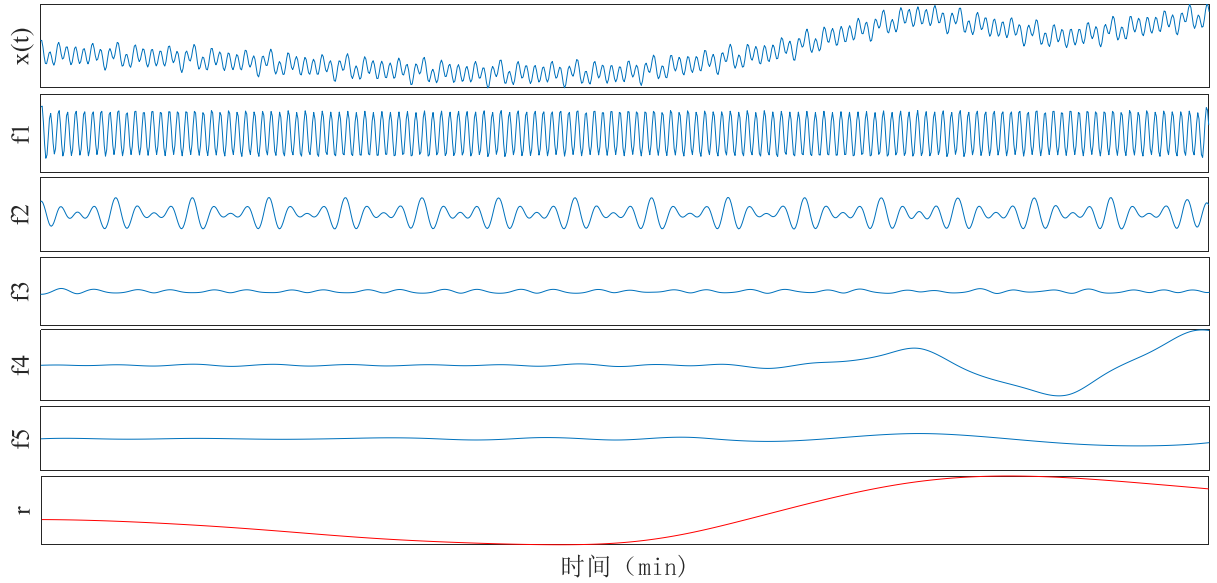


➢ 噪声数据

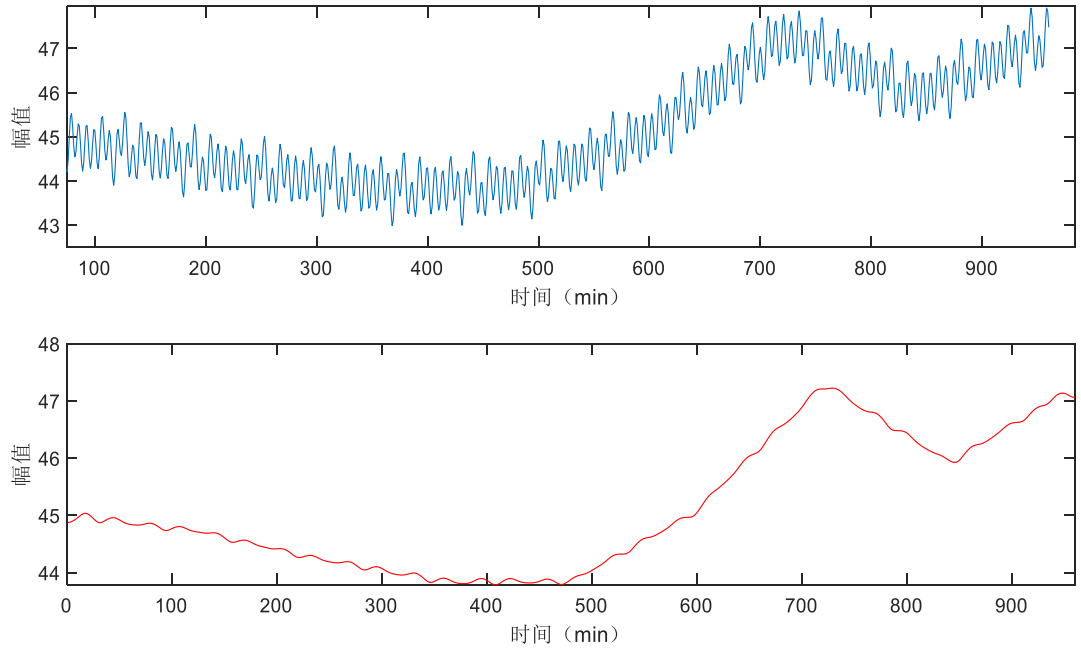
- 经验模态分解(Empirical Mode Decomposition，EMD)：依据数据自身的时间尺度特征来进行信号分解，无须预先设定任何基函数
- 特点：
  - ✓ 可以应用于任何类型信号的分解
  - ✓ 处理非平稳及非线性数据上具有非常明显的优势
  - ✓ 把复杂信号分解成有限个本征模函数IMF
  - ✓ 分解出来的各IMF分量包含原信号不同时间尺度的局部特征信号

- 经验模态分解实例
  - ✓ 将含噪信号经过上述步骤的EMD分解，将原始信号分解成$f_1 - f_5$的5个本征模函数与分解筛出的信号残余分量r的线性叠加，最后将5个本征模函数进行信号重构

EMD分解



信号重构



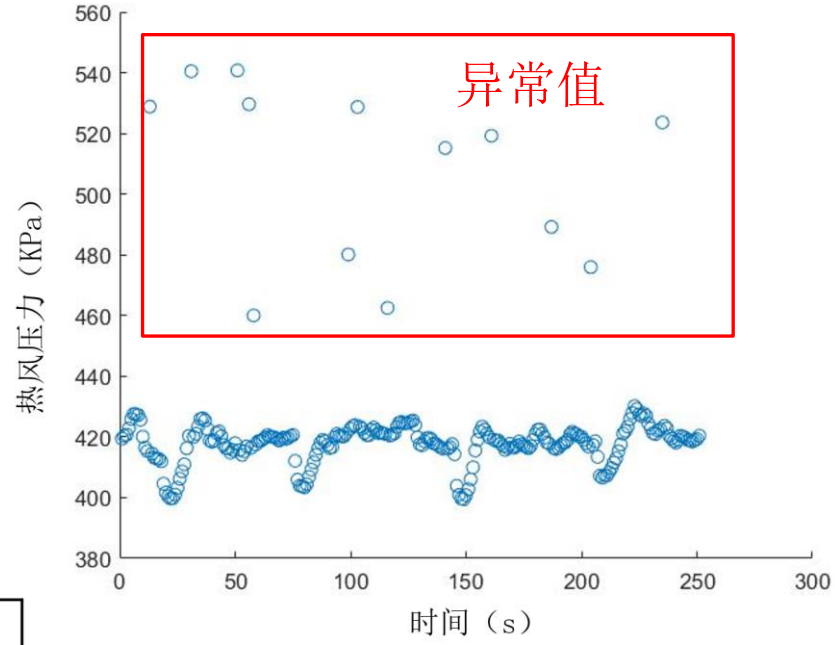
2 数据清洗

➢ 异常值清洗

- 异常值：检测数据有输入错误或者不合常理的数据。

  - ✓ 简单统计分析：对数据进行简单的描述性统计分析。

  例如，结合机理知识判断变量取值是否正常。



TIME 时间	AVG（CBV） 冷风流量	AVG（CBP） 冷风压力
2015-06-01 00:00	5665.07	432.41
2015-06-01 00:01	5670.73	433.32
2015-06-01 00:02	-5000	433.64
2015-06-01 00:03	5636.75	435.62
2015-06-01 00:04	5629.01	438.64
2015-06-01 00:05	5631.91	440.38

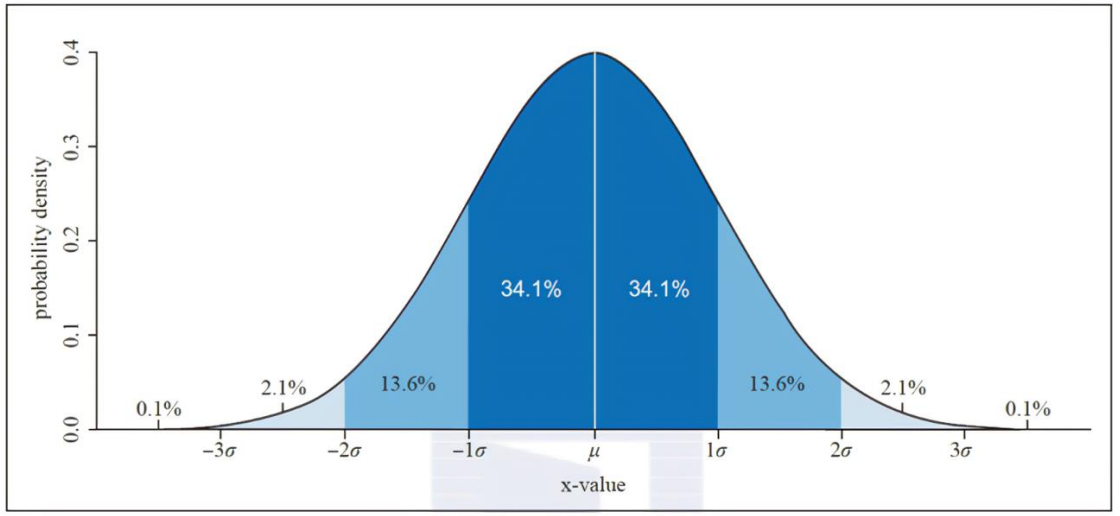
删除该行 →

TIME 时间	AVG（CBV） 冷风流量
2015-06-01 00:00	5665.07
2015-06-01 00:01	5670.73
2015-06-01 00:03	5636.75
2015-06-01 00:04	5629.01
2015-06-01 00:05	5631.91

➢ 异常值清洗

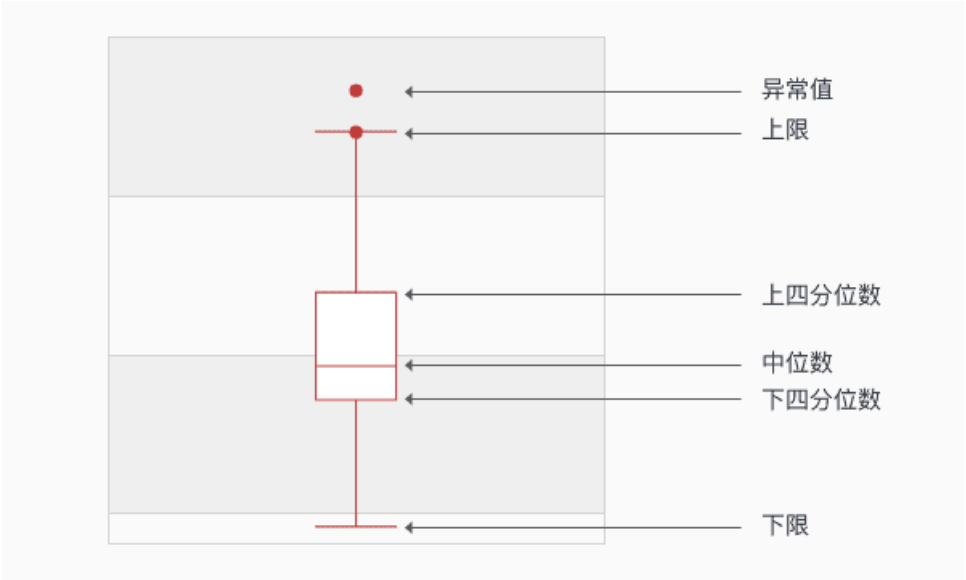
✓ 3σ原则

针对服从正态分布的数据，基于3σ原则，异常值为一组测定值中与平均值的偏差超过3倍标准差的值。



✓ 箱线图分析

箱线图识别异常值标准：小于$Q_1 - 1.5 * IQR$或大于$Q_3 + 1.5 * IQR$的值

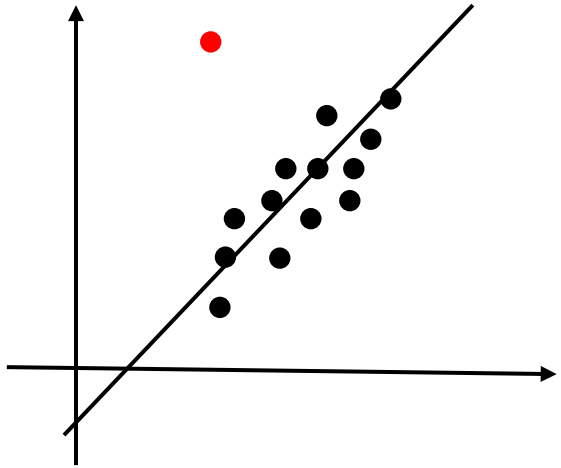


➢ 异常值清洗

✓ 基于模型检测

构造概率分布模型并计算对象符合模型的概率，低概率的点为异常点。

- 模型是簇的集合，异常值是不显著属于任何簇的对象；
- 模型是回归的，异常值是相对远离预测值的对象。



线性回归检测异常点

✓ 基于距离

在对象之间定义邻近性度量，异常对象是那些远离其他对象的对象。

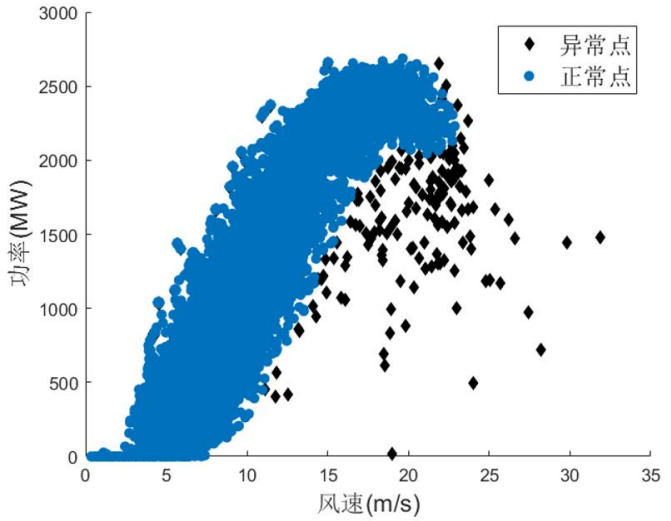
➢ 异常值清洗

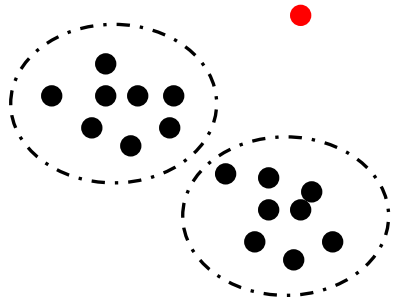
✓ 基于密度

当对象的局部密度显著低于它的大部分近邻时将其分类为异常值，适合非均匀分布的数据。

✓ 基于聚类

一个对象如果不强属于任何簇，则判定该对象是一个基于聚类的异常值。





K-means聚类检测异常点

➢ 逻辑错误清洗

- 根据简单的逻辑分析发现数据问题

去重

需要注意：1. 人为判断的重复不一定是重复
2. 重复可能是大数据学习的需要
例如深度学习对抗设计

去除不合理值

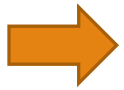
主要为人工参与产生的错误：例如某人年龄为20岁，婚龄为15年。需要根据具体大数据问题进行细节判断，例如当风速数据为5m/s时，将风电功率数据限制为10-30W之间

2 数据清洗

- 逻辑错清洗实例：高功率对应低风速值

属性 / 时间	实际功率	实际风速	预测风速	风向	温度	压强	湿度
0：15	16.69767	5.99	2.94	65.03	3.78	92281.6	82.81
0：30	24.82667	7.07067	4	79.61	3.66℃	92269.85	83.44
0：45	NAN	7.16167	4.99	87.08	3.61	92263.59	83.74
1：00	22.32567	6.84833	5.13	92.13	3.63	92260.14	83.9
1：15	20.24833	6.566	5.01	93.25	3.58℃	92262.38	84.3
1：30	18.85633	6.26267	5.25	90.16	3.5	92262.66	84.66
1：45	23.14433	6.83867	5.51	89.01	3.47	92263.88	84.79
2：00	-1000	0	5.64	89.74	3.39	92263.3	85.15
2：15	16.13933	0.95967	5.65	91.5	3.28	92261.36	85.6

删除该行



属性 / 时间	实际功率	实际风速
0：15	16.69767	5.99
0：30	24.82667	7.07067
0：45	NAN	7.16167
1：00	22.32567	6.84833
1：15	20.24833	6.566
1：30	18.85633	6.26267
1：45	23.14433	6.83867
2：00	-1000	0

单选题 3分

设置

异常值处理3σ准则中，σ的含义为

A 正态分布的方差

B 正态分布的标准差

C 原始数据的方差

D 原始数据的标准差

提交

目录

- 工业数据质量问题
- 数据清洗
- **数据集成**
- 数据变换
- 数据归约

IEEE INDUSTRIAL ELECTRONICS MAGAZINE ■ DECEMBER 2019 PP.38

➢ 数据集成

- 数据集成：多个数据源中的数据合并，存放在一个一致的数据存储中。
  - ✓ 减少结果数据集的冗余和不一致，提高数据挖掘的准确度和速度
    - 实体识别问题
    - 冗余
    - 样本重复
    - 数据冲突

➢ 数据集成

- 实体识别问题：将来自多个信息源的现实世界的等价实体进行匹配。
  - ✓ 例：高炉炼铁过程中冷风流量在一个数据库存储中为CBV，而在另一个数据库存储为CBV_ID。
- 冗余：如果一个属性能由另一个或几个属性“导出”，则这个属性可能是冗余的。
  - 标称数据：卡方检验检测
  - 数值属性：相关系数、协方差评估

➢ 数据集成

- 样本重复：一组实体数据中存在两个或多个相同的样本
  - ✓ 例：数据库对同一时刻的数据进行了重复记录
- 数据冲突：数据集成时，由于不同数据源的表示方式、度量方法或编码存在区别，数据值可能存在冲突。
  - ✓ 例：重量属性可能在一个系统中以国际单位存放，而在另一个系统中以英制单位存放。

目录

- 工业数据质量问题
- 数据清洗
- 数据集成
- **数据变换**
- 数据归约

IEEE INDUSTRIAL ELECTRONICS MAGAZINE ■ DECEMBER 2019 PP.38

4 数据变换

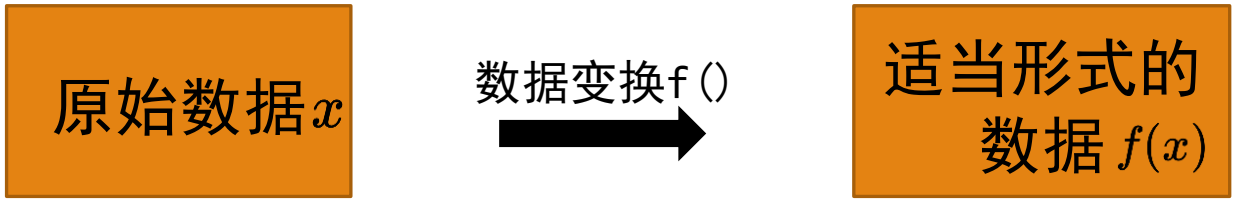
➢ 数据变换

- 数据变换主要是对数据进行规范化处理，将数据转换为适当的形式
- 数据变换策略
  - 光滑：去掉数据中的噪音。如分箱、小波变换和经验模态分解等
  - 属性（特征）构造：由给定的属性构造新的属性并添加到属性集
  - 聚集：对数据进行汇总和聚集
  - 规范化：把属性数据按照比例缩放，使之落入一个特定的小区间，如-1.0到1.0或0.0-1.0
  - 离散化：数值属性（如年龄）的原始值用区间标签（如0-10,11-20等）替换

➢ 简单的函数变换

- 简单函数变换是对原始数据进行某些数学函数变换，常用的变换包含平方、开方、取对数、差分运算等：

$$x' = x^2$$
$$x' = \sqrt{x}$$
$$x' = \log x$$



- 简单的函数变换常用来将原始数据变换成易于处理的适当形式数据
  - 时间序列分析：简单的对数变换、差分运算
  - 取值范围较为宽泛的分布：对数变换

➢ 规范化：消除指标间量纲和取值范围差异

- 最小-最大（Min-Max）规范化：to [new_min_A, new_max_A]

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

特点：新数据的加入使得min，max会发生变化

- 标准分数/Z分数（Z-score）规范化(μ_A: 均值, σ_A: 标准差):

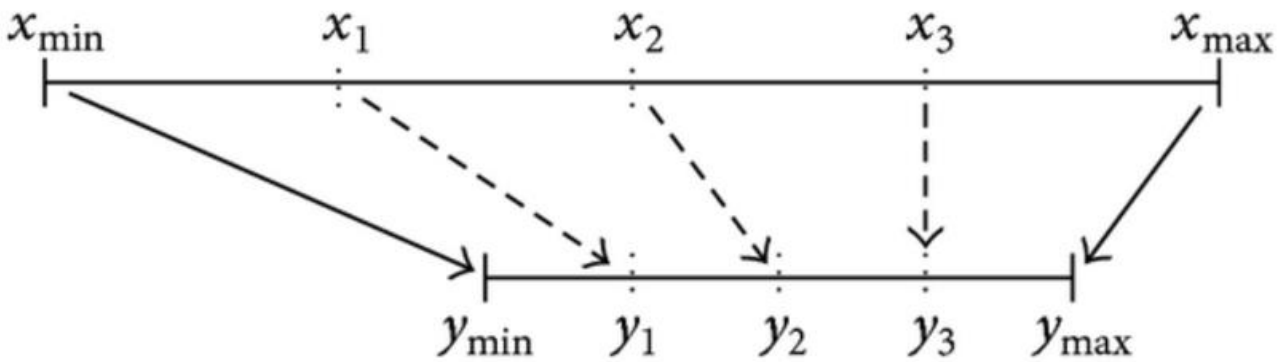
$$v' = \frac{v - \mu_A}{\sigma_A}$$

特点：适用于数据最大值、最小值未知，或最小-最大规范化结果不合理时

- 小数定标规范化 $v' = \frac{v}{10^j}$ j是使Max($|v'|$) < 1的最小整数
- 零均值变换

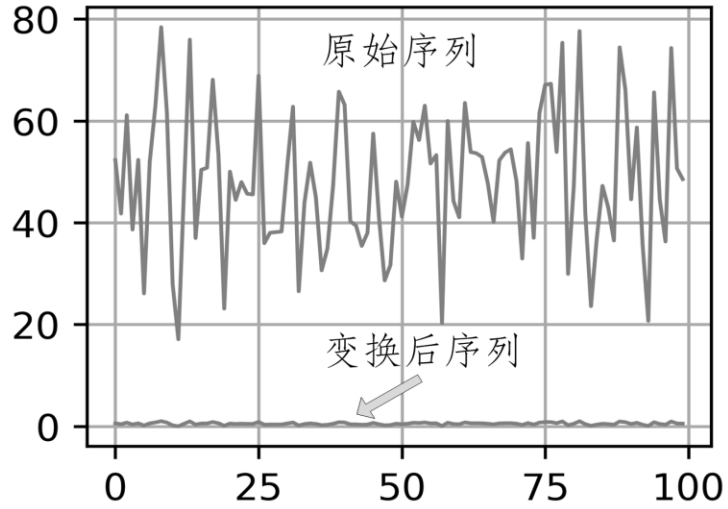
➢ 最小-最大规范化 (Min-Max Normalization)

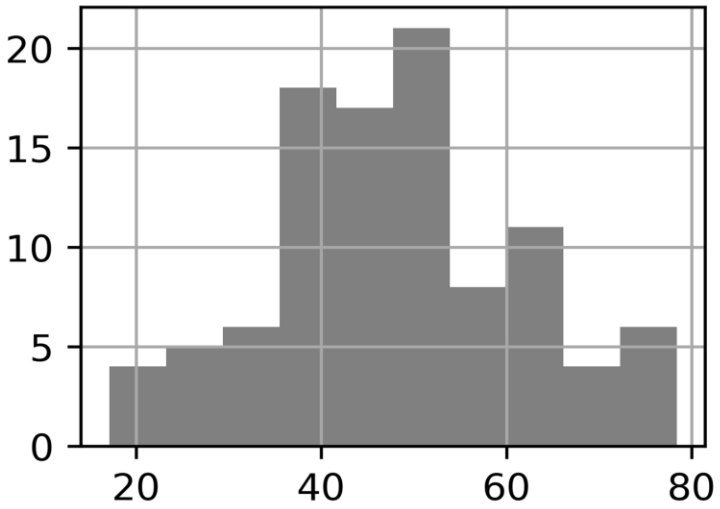
$$v' = \frac{v - v_{min}}{v_{max} - v_{min}}(v'_{max} - v'_{min}) + v'_{min}$$

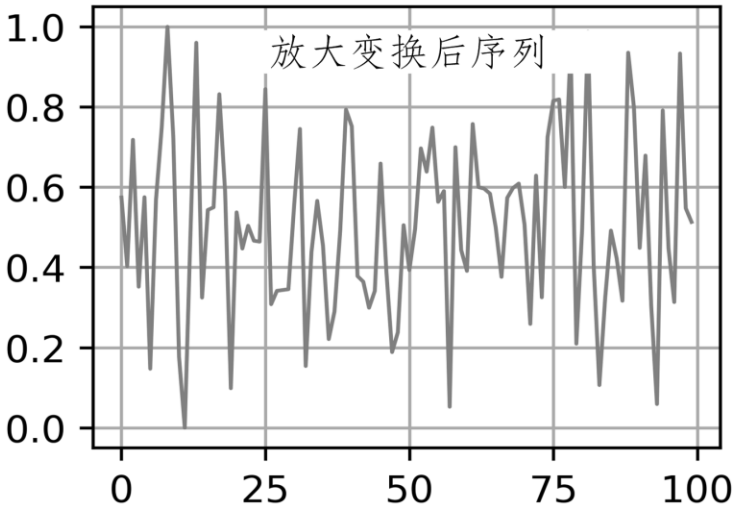


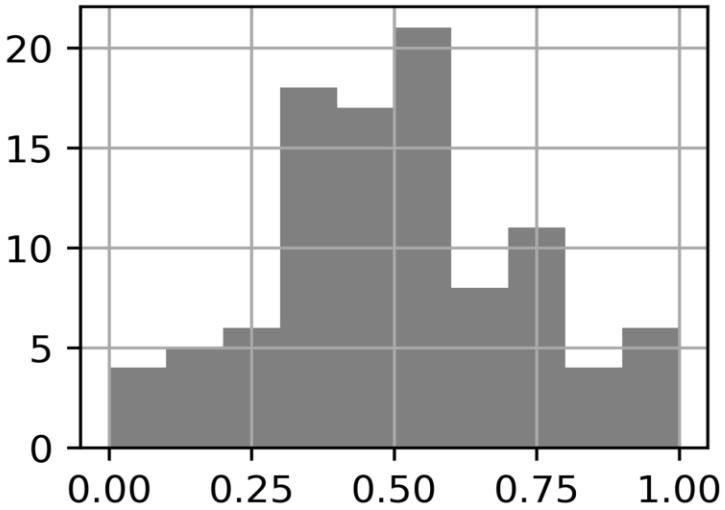
- 当多个属性的数值分布区间相差较大时，使用最小-最大规范化，可以让这些属性值变换到同一个区间，这对于属性间的比较以及计算对象之间的距离很重要

- 最小-最大规范化 (Min-Max Normalization)









➢ 最小-最大规范化 (Min-Max Normalization)

例： 假设属性热风温度的最小值与最大值分别为1114℃和1326℃，现在的风温值为1200℃，我们想把热风温度映射到区间[0,1]，根据最小-最大规范化风温值1200℃将变换为：

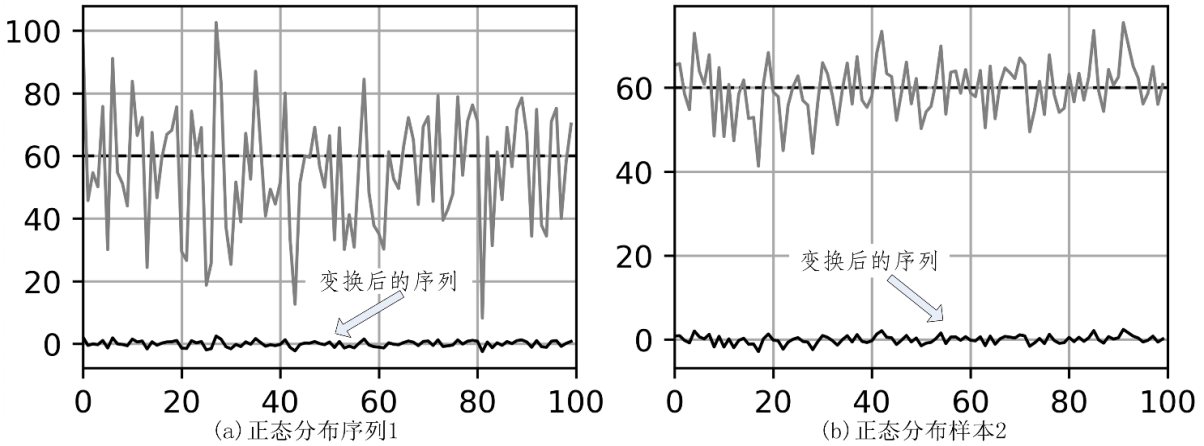
$$\frac{1200-1114}{1326-1114}(1-0)+0=0.406$$

Z分数变换（Z –score）规范化

$$v' = \frac{v - \bar{v}}{\sigma_v}$$

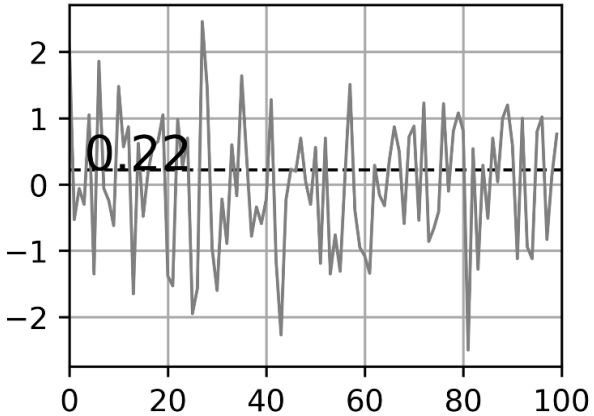
- 以标准差为单位度量，原始数离开其平均数之上或之下多少个标准差
- 变换后，平均数近似为0，标准差近似为1
- 对满足不同正态分布的多个属性进行*Z-score*变换，可以将这些正态分布都化成标准正态分布，充分利用标准正态分布的性质，对不同属性的数据进行分析和相互比较

➢ Z分数变换（Z –score）规范化

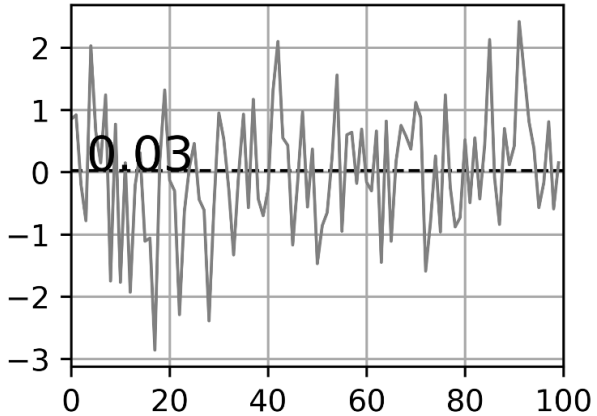


(a) 正态分布序列1

(b) 正态分布样本2



(c) 样本1变换为标准分之后



(d) 样本2变换为标准分之后

$$p(X \leqslant 0.22) = 58.71\% \quad p(X \leqslant 0.03) = 51.20\%$$

➢ Z分数变换（Z –score）规范化

例：假设属性热风温度的均值和标准差分别为1189℃和374℃。使用Z分数规范化，值1200℃被转换为：

$$\frac{1200-1189}{374}=0.06$$

小数定标规范化

- 小数定标规范化通过移动属性v的小数点位置进行规范化，小数点的移动位数依赖于v的最大绝对值。v的值被规范化为v'：

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

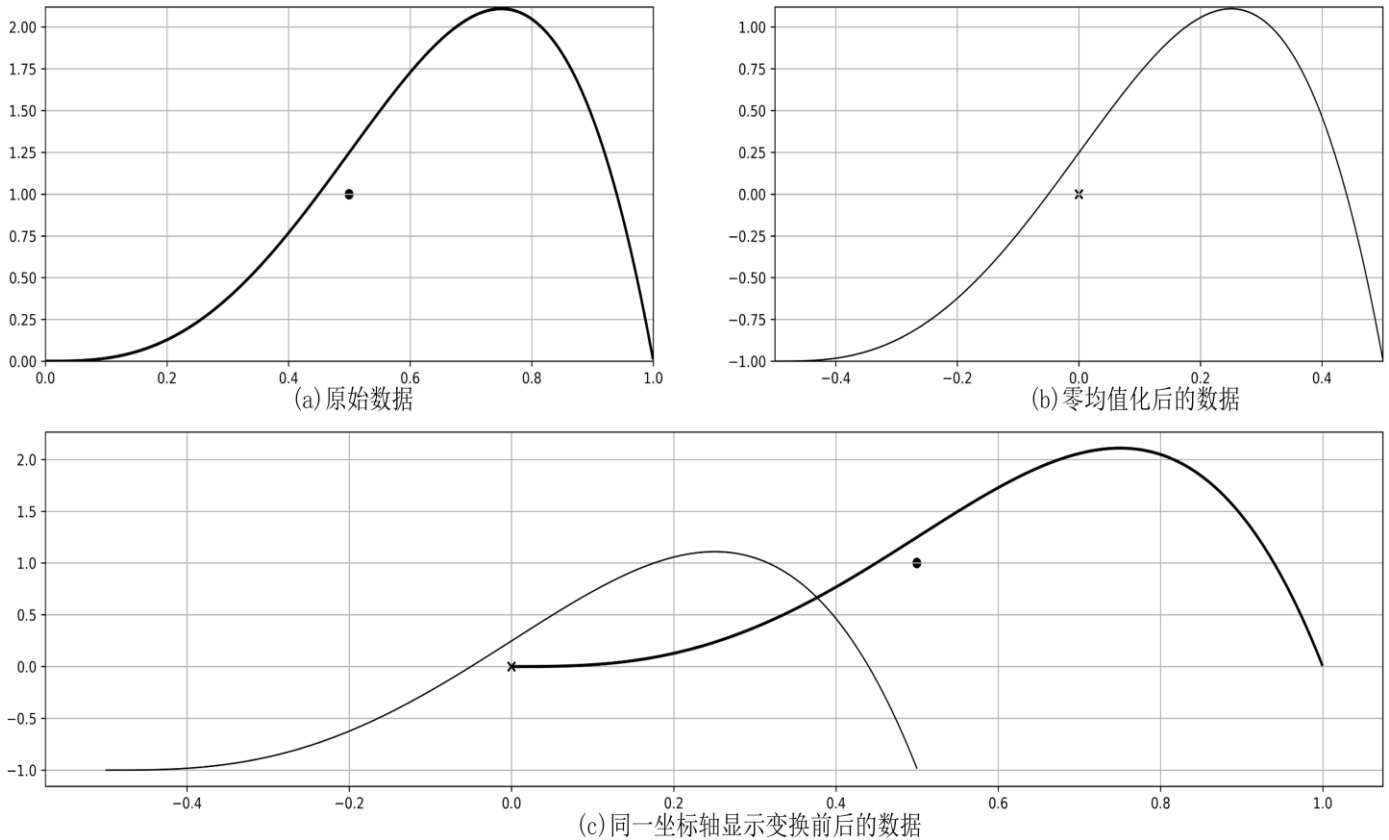
j是使得$(\max|v'|) < 1$的最小整数

小数定标规范化

- 假设v的取值是由-986到917，v的最大绝对值为986。使用小数定标规范化时，我们用每个值除以1000（即，j=3）。这样，-986被规范化为-0.986，而917被规范化为0.917。

零均值化变换

- 每一个属性的数据都减去这个属性的均值，各属性的数据和与均值都为零
- 多个属性经过零均值化变换后，都以零为均值分布，各属性的方差不发生变化，各属性间的协方差也不发生变化
- 零均值化变换在很多场合得到应用，例如对信号数据零均值化，可以消除直流分量的干扰



多选题 4分

设置

一般常见的数据归一化方法有

A Z分数变换

B 最小-最大规范化

C 零均值化

D 小数定标规范化

提交

➢ 数据离散化

- 离散化：将一个连续属性的范围划分为区间
  - 使用间隔标签替换实际的数据值
  - 通过离散化减少数据量
  - 为进一步的分析做准备，例如，分类
- 常见方法
  - 分箱离散化：基于指定的箱个数的自上向下的分裂方式，一种无监督的离散化技术
  - 直方图分析离散化：使用等频直方图划分数据，使得每个分区包括相同个数的数据样本
  - 聚类、决策树离散化

目录

- 工业数据质量问题
- 数据清洗
- 数据集成
- 数据变换
- 数据归约

IEEE INDUSTRIAL ELECTRONICS MAGAZINE ■ DECEMBER 2019 PP.38

5 数据归约

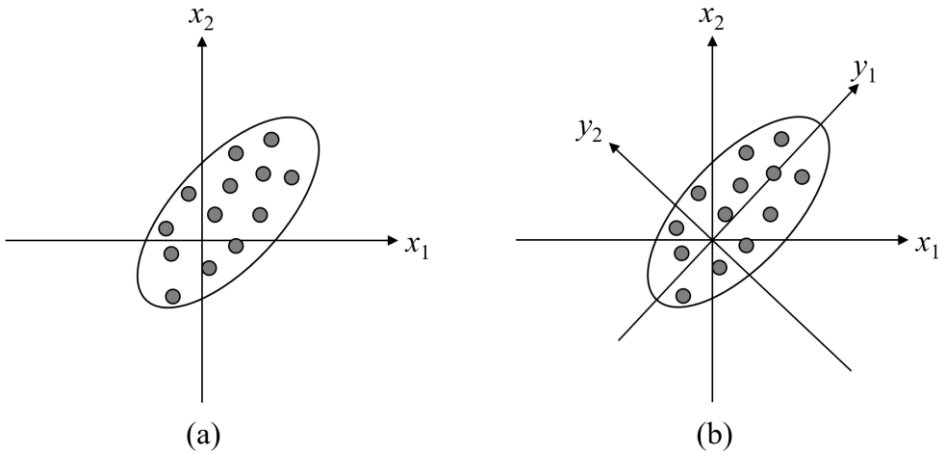
➢ 数据归约概述

- 数据归约是指在尽可能保持数据原貌的前提下，最大限度地精简数据量
- 数据归约技术可以用来得到数据集的归约表示，它小得多，但仍可以保持原始数据的完整性
- 数据归约策略主要分为：
  - 维归约（Dimensionality Reduction）：从原有的维度中删除不重要或不相关属性
  - 数量归约（Numerosity Reduction）：用替代的、较小的数据表示形式替换原数据
  - 数据压缩（Data Compression）：维归约和数量规约也可以视为某种形式的数据压缩

➢ 维归约

◘ 主成分分析（Principal Component Analysis, PCA）

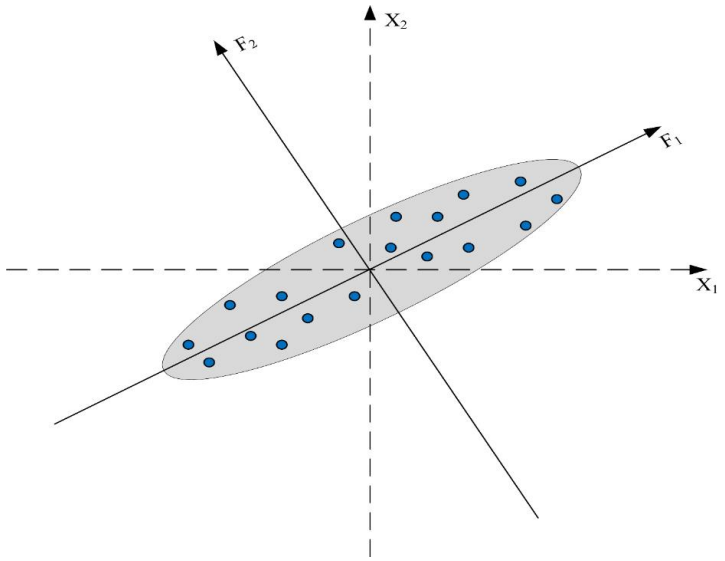
- 由于多个变量之间往往存在着一定程度的相关性
- 是一种通过线性变换，将原始数据的多个变量组合成相互正交的少数几个能充分反映总体信息的指标，以便于进一步分析
- 作用：简化原始数据、降维，在保证精度前提下增加计算效率
- 方法：求协方差矩阵的特征向量，用这些特征向量定义新空间



➢ 主成分分析的几何解释

- 由两个维度变量构成的数据，可以看成是二维空间上的点
- 如果这些数据在二维平面上形成一个椭圆形状的点阵
- 这个椭圆有一个长轴和一个短轴。在短轴方向上，数据变化很少
- 在极端的情况，短轴如果退化成一点，那只有在长轴的方向才能够解释这些点的变化了；这样，由二维到一维的降维就自然完成了

- ✓ 坐标轴通常不和椭圆的长短轴平行。需要寻找椭圆的长短轴，并进行变换，使得新维度方向和椭圆的长短轴平行
- ✓ 如果长轴维度变量代表了数据包含的大部分信息，就用该维度代替原先的两个维度变量（舍去次要的一维），降维就完成了
- ✓ 椭圆（球）的长短轴相差得越大，降维也越有道理



➢ 主成分分析的具体步骤

设有m条n维数据。

（1）将原始数据按列组成n行m列矩阵X

（2）将X的每一行进行零均值化，即减去这一行的均值

（3）求出协方差矩阵 $C=\frac{1}{m}XX^T$

（4）求出协方差矩阵的特征值及对应的特征向量

（5）将特征向量按对应特征值大小从上到下按行排列成矩阵，取前k行组成矩阵P

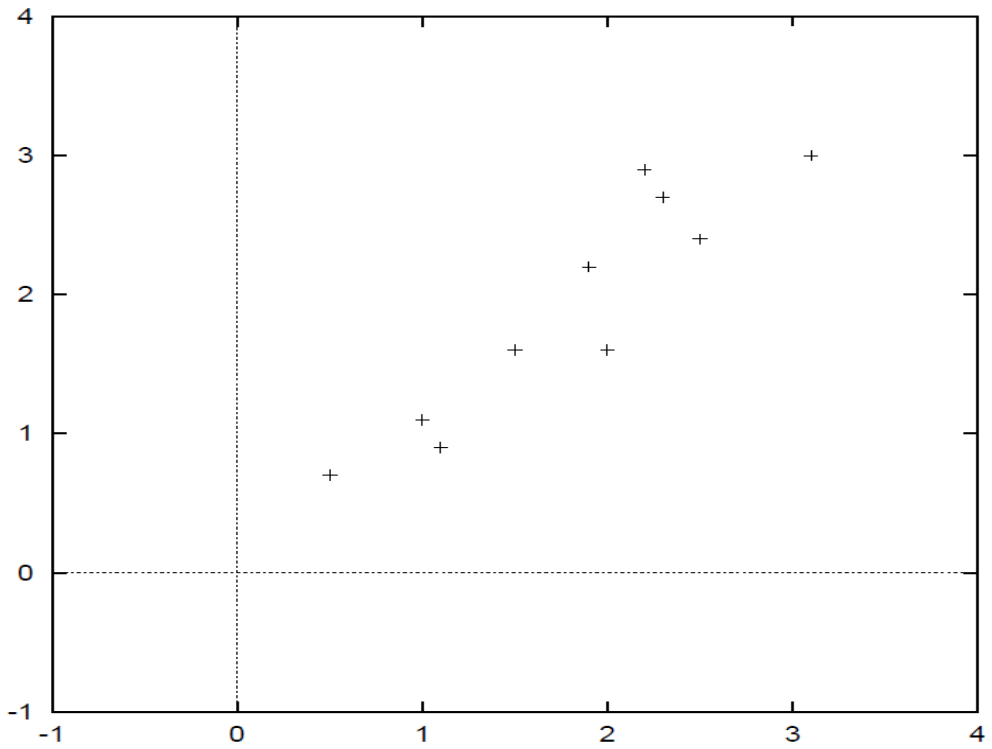
（6）$Y=PX$即为降维到k维后的数据

➢ 主成分分析实例：一个二维数据

◘ 步骤1：数据的表示：

- **每个变量有n个观察值，数据表示成一个$n \times p$矩阵；**

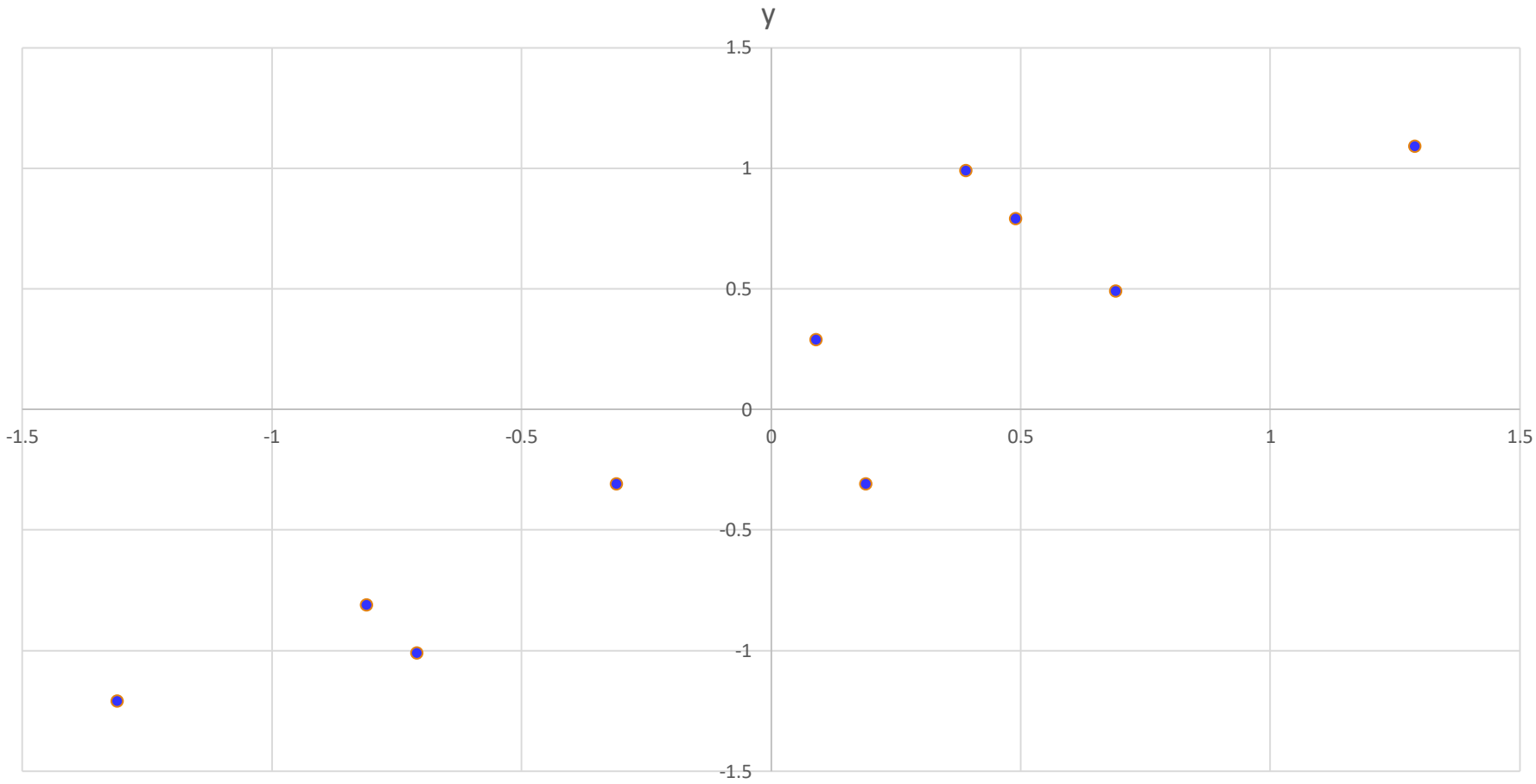
x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9



- PCA步骤2：
  - 数据变换：零均值化
  - 在数据的每一维上减掉这一维的均值，使得每一维数据零均值化

✓ 变换后的数据：

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3：
  - 计算协方差矩阵
  - 协方差矩阵包含了各个变量之间全部的关系信息

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

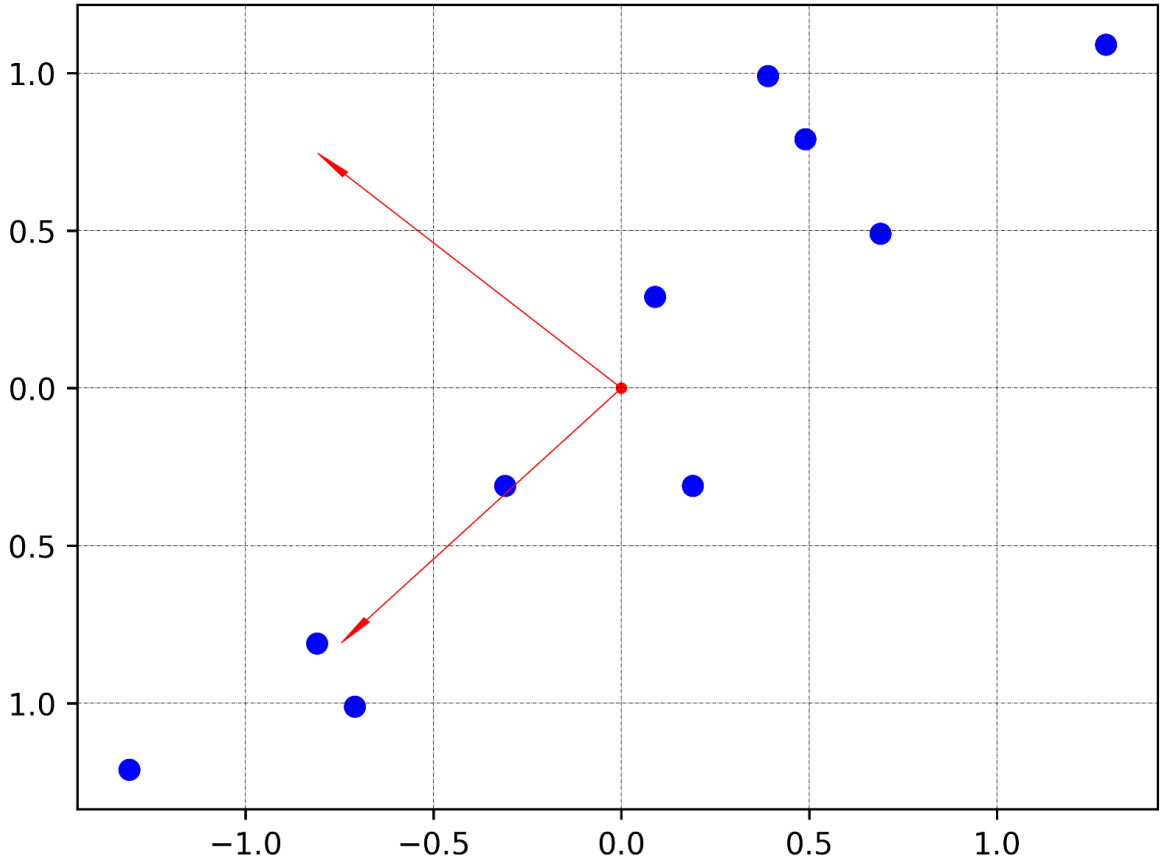
$$\mathrm{cov}=\begin{pmatrix}0.616555556 & 0.615444444\\ 0.615444444 & 0.716555556\end{pmatrix}$$

- PCA步骤4：
  - 计算协方差矩阵的特征值和特征向量

$$eigenvalues = \begin{pmatrix} 0.0490833989 \\ 1.28402771 \end{pmatrix}$$

$$eigenvectors = \begin{pmatrix} -0.735178656 & -0.677873399 \\ 0.677873399 & -0.735178656 \end{pmatrix}$$

- ✓ 特征向量与数据



PCA步骤5：

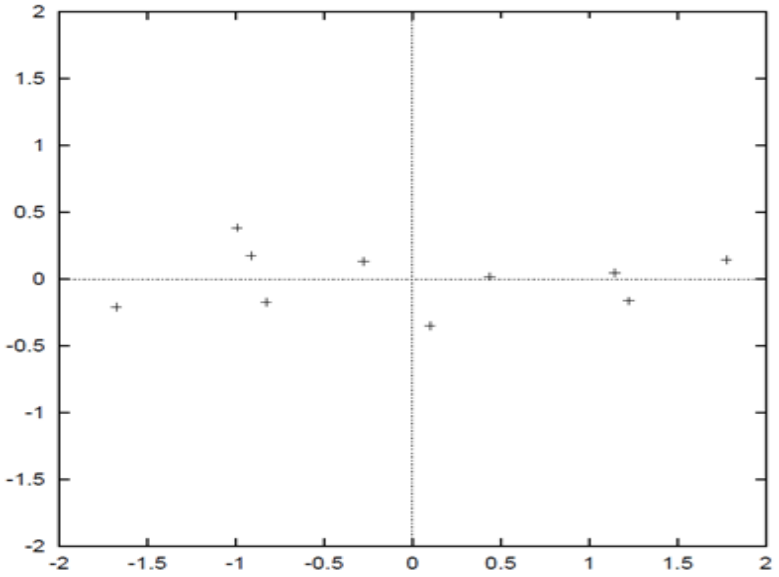
- 主成分选择
  - ✓ 从上图可以清楚地看出，较大特征值对应的特征向量的方向上，包含了数据集合中主要的信息量
  - ✓ 将特征值从大到小排列，较小的值可以忽略
- 上例中，我们选择的特征向量：

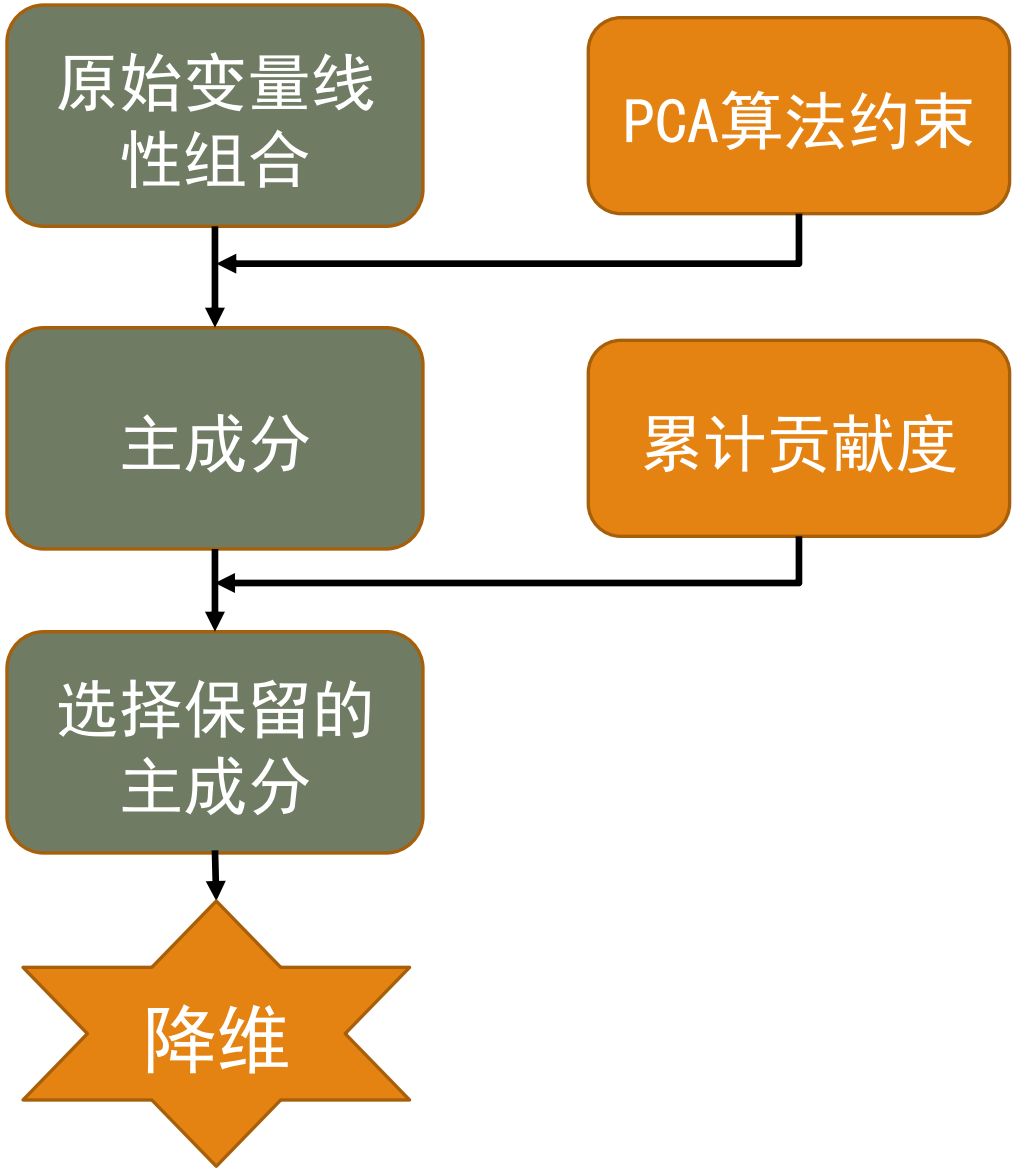
$$\begin{pmatrix} -0.677873399 \\ -0.735178656 \end{pmatrix}$$

PCA步骤6：

- 得到新的特征向量
  - ✓ $y=Ax$, 这里y为主成分向量, A为主成分变换矩阵, x为原始数据向量（零均值化后）；
  - ✓ 矩阵的乘法是将原始样本点分别往特征向量对应的轴上做投影。
  - ✓ 取前面得到的最大的几个特征值所相应的特征向量作为A的行即可。
  - ✓ 上述例子中，第一个样本的两维数据变成一个综合变量：
    -0.677873399 ×0.69-0.73518 ×0.49=-0.82797
  - ✓ 第二个综合变量类似。得到：

-0.82797018	-0.175115307
1.77758033	0.142857227
-0.992197494	0.384374989
-0.274210416	0.130417207
-0.167580142	-0.209498461
-0.912949103	0.175282444
0.0991094375	-0.349824698
1.14457216	0.0464172582
0.0428046137	0.0177646297
1.22382056	-0.162675287




原始变量线性组合
PCA算法约束
主成分
累计贡献度
选择保留的主成分
降维

维归约

主成分分析特点

- 分析基础：变量间具有相关性
- 分析目的：降维的同时尽可能少的损失原始数据信息
- 分析方法：基于约束条件对变量做线性组合进行优化求解
- 约束条件：1. 降维后主元间互不相关

  2. 主成分按照方差大小依次降序排列

  3. 每个主元方差尽可能大

  4. 线性组合的系数向量模长为1

单选题 3分

设置

主成分分析的第一步为：

A 特征根分解

B 特征向量求取

C 选取主元个数

D 数据标准化

提交

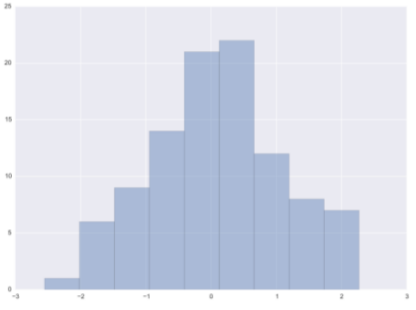
➢ 数量归约

◘ 参数方法

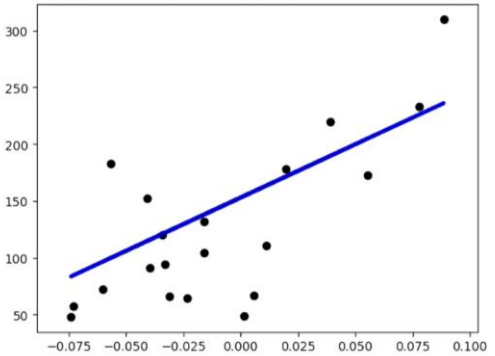
- 使用模型估计数据，只需存放模型参数，而非真实数据（离群点也可能存放）
- 主要方法：回归、对数-线性模型

◘ 非参数方法

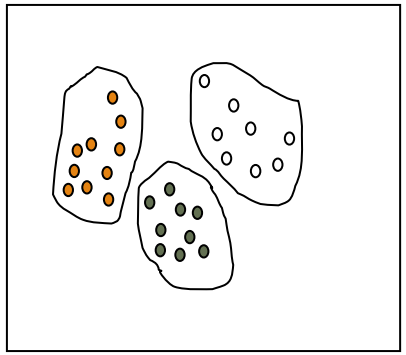
- 不用确定模型
- 主要方法：直方图、聚类、抽样等



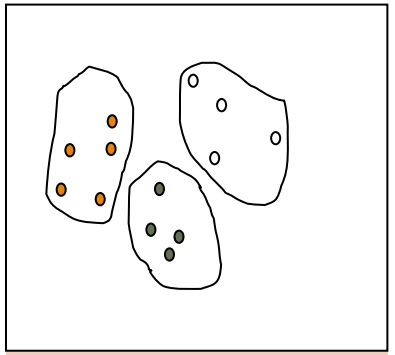
直方图



线性回归



分层抽样



对原始数据
进行聚类

6 本章小结

- 数据质量:准确性、完整性、一致性、及时性、可信度、可解释性
- 数据清理:格式内容清洗、填补缺失值，光滑噪声、识别异常值
- 数据集成:实体识别问题、删除冗余、样本重复、数据冲突
- 数据变换：规范化、数据离散化
- 数据归约：维归约、数量归约

参考文献

- Aggarwal and C. C. Data mining: the textbook. Heidelberg: Springer, 2015.
- J. Han, J. Pei, and M. Kamber. Data mining: concepts and techniques. Amsterdam: Elsevier, 2011.
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- T. Dasu, T. Johnson, S. Muthukrishnan, V. Shkapenyuk. Mining Database Structure; Or, How to Build a Data Quality Browser. SIGMOD'02
- H. V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Technical Committee on Data Engineering, 20(4), Dec. 1997
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- E. Rahm and H. H. Do. Data Cleaning: Problems and Current Approaches. IEEE Bulletin of the Technical Committee on Data Engineering. Vol.23, No.4
- V. Raman and J. Hellerstein. Potters Wheel: An Interactive Framework for Data Cleaning and Transformation, VLDB'2001
- T. Redman. Data Quality: Management and Technology. Bantam Books, 1992