大数据分析

数据预处理


QUESTIONNAIRE
Very often
Sometimes
Always


cash register
cashier
optical scanner
electronic payment terminal
grocery bags
bagger

- 数据质量分析
- 数据特征分析
- 数据预处理

数据质量分析

- 数据质量分析是数据预处理的前提，是数据挖掘分析结论有效性和准确性的基础，其主要任务是检查原始数据中是否存在脏数据
- 脏数据一般是指不符合要求，以及不能直接进行相应分析的数据，在常见的数据挖掘工作中，脏数据包括：
  - 缺失值
  - 异常值
  - 不一致的值
  - 重复数据及含有特殊符号（如#、￥、*）的数据

缺失值

- 数据的缺失主要包括记录的缺失和记录中某个字段信息的缺失，两者都会造成分析结果不准确，缺失值产生的原因：
  - 有些信息暂时无法获取，或者获取信息的代价太大。
  - 有些信息是被遗漏的。
  - 属性值不存在。

异常值

- 异常值是指样本中的个别值，其数值明显偏离其余的观测值。
- 异常值分析是检验数据是否有录入错误，是否含有不合常理的数据。
- 异常值也称为离群点，异常值的分析也称为离群点的分析。
- 异常值分析方法主要有：简单统计量分析、3σ原则、箱型图分析、基于聚类的离群点检测。

不一致的值

- 数据不一致性是指数据的矛盾性、不相容性。直接对不一致的数据进行挖掘，可能会产生与实际相违背的挖掘结果。
- 在数据挖掘过程中，不一致数据的产生主要发生在数据集成的过程中，可能是由于被挖掘数据是来自于不同的数据源、重复存放的数据未能进行一致性地更新造成的

数据特征分析

- 一般可通过绘制图表、计算某些特征量等手段进行数据的特征分析。
- 特征分析的方法有：
  - 分布分析
  - 对比分析
  - 统计量分析
  - 周期性分析
  - 贡献度分析
  - 相关性分析

分布分析

- 分布分析能揭示数据的分布特征和分布类型。
- 对于定量数据，欲了解其分布形式是对称的、还是非对称的，通常可做出频率分布直方图进行直观地分析；
- 对于定性分类数据，可用饼图和条形图直观地显示分布情况。

分布分析-定量数据

- 频率分布直方图一般按照以下步骤：
  - 求极差
  - 决定组距与组数
  - 决定分点
  - 列出频率分布表
  - 绘制频率分布直方图

- 遵循的主要原则有：
  - 各组之间必须是相互排斥的
  - 各组必须将所有的数据包含在内
  - 各组的组宽最好相等

实例

- 下表是描述菜品“捞起生鱼片”在2014年第二个季度的销售数据，绘制销售量的频率分布表、频率分布图，对该定量数据做出相应的分析。

日期	销售额	日期	销售额	日期	销售额
2014/4/1	420	2014/5/1	1770	2014/6/1	3960
2014/4/2	900	2014/5/2	135	2014/6/2	1770
2014/4/3	1290	2014/5/3	177	2014/6/3	3570
2014/4/4	420	2014/5/4	45	2014/6/4	2220
2014/4/5	1710	2014/5/5	180	2014/6/5	2700
…	…	…	…	…	…
2014/4/30	450	2014/5/30	2220	2014/6/30	2700
		2014/5/31	1800		

第一步：求极差

极差 = 最大值 - 最小值 = 3960-45=3915

第二步：分组

取组距为500

组数 = 极差/组距 = 3915/500=7.83≈8

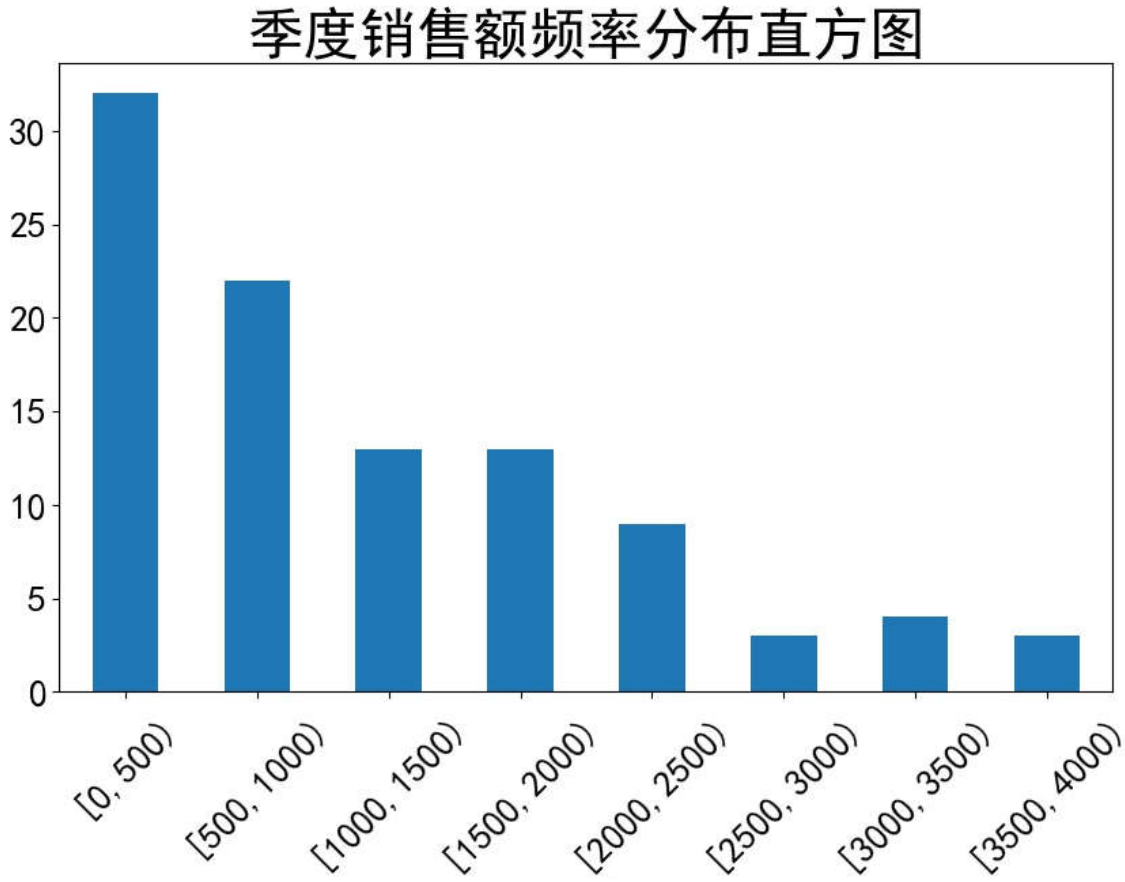
第三步：决定分点，如下表：

[0，500）	[500，1000）	[1000，1500）	[1500，2000）
[2000，2500）	[2500，3000）	[3000，3500）	[3500，4000）

第四步：绘制频率分布表

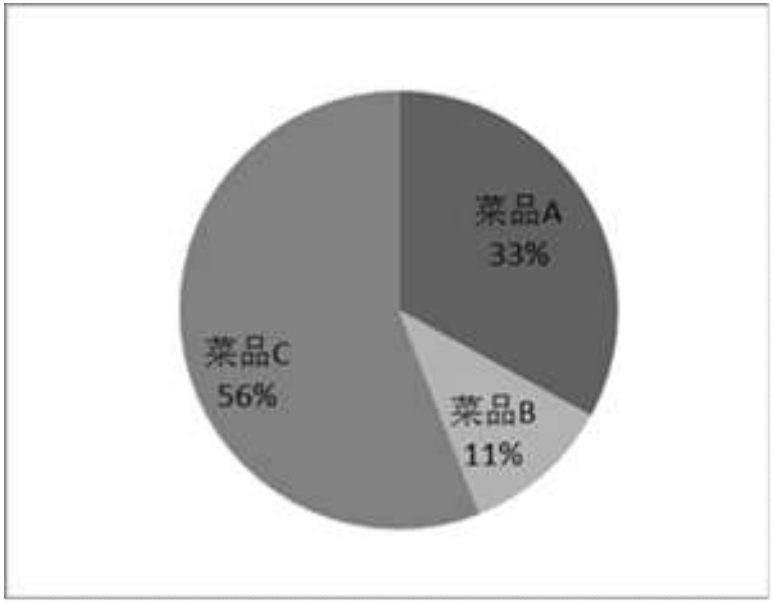
组段	组中值 x	频数	频率 f	累计频率
[0，500）	250	15	16.48%	16.48%
[500，1000）	750	24	26.37%	42.85%
[1000，1500）	1250	17	18.68%	61.54%
[1500，2000）	1750	15	16.48%	78.02%
[2000，2500）	2250	9	9.89%	87.91%
[2500，3000）	2750	3	3.30%	92.31%
[3000，3500）	3250	4	4.40%	95.60%
[3500，4000）	3750	3	3.30%	98.90%
[4000，4500）	4250	1	1.10%	100.00%

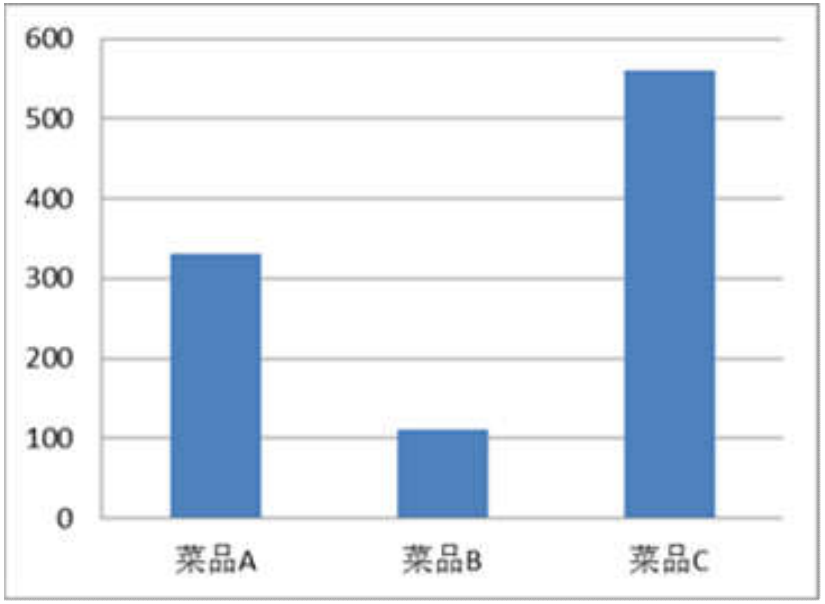
第五步：绘制频率分布直方图，以2014年第二季度捞起生鱼片每天的销售额为横轴，以各组段的频率密度（频率与组距之比）为纵轴



分布分析-定性数据

- 对于定性变量，常常根据变量的分类类型来分组，可以采用饼图和条形图来描述定性变量的分布。





对比分析

- 对比分析是指把两个相互联系的指标数据进行比较，从数量上展示和说明研究对象规模的大小、水平的高低、速度的快慢，以及各种关系是否协调。
- 对比分析主要有以下两种形式：
  - 绝对数比较
  - 相对数比较

对比分析-相对数比较

- 由于研究目的和对比基础不同，相对数可以分为以下几种：

1)结构相对数

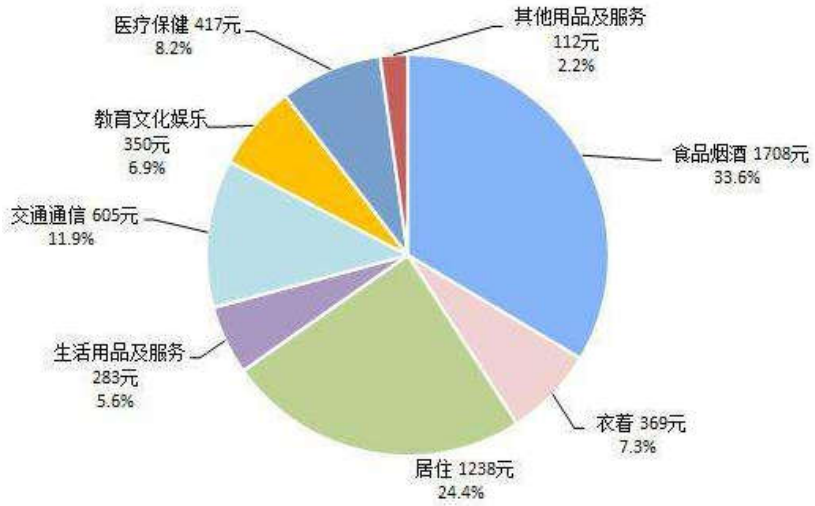
2)比例相对数

3)比较相对数

4)强度相对数

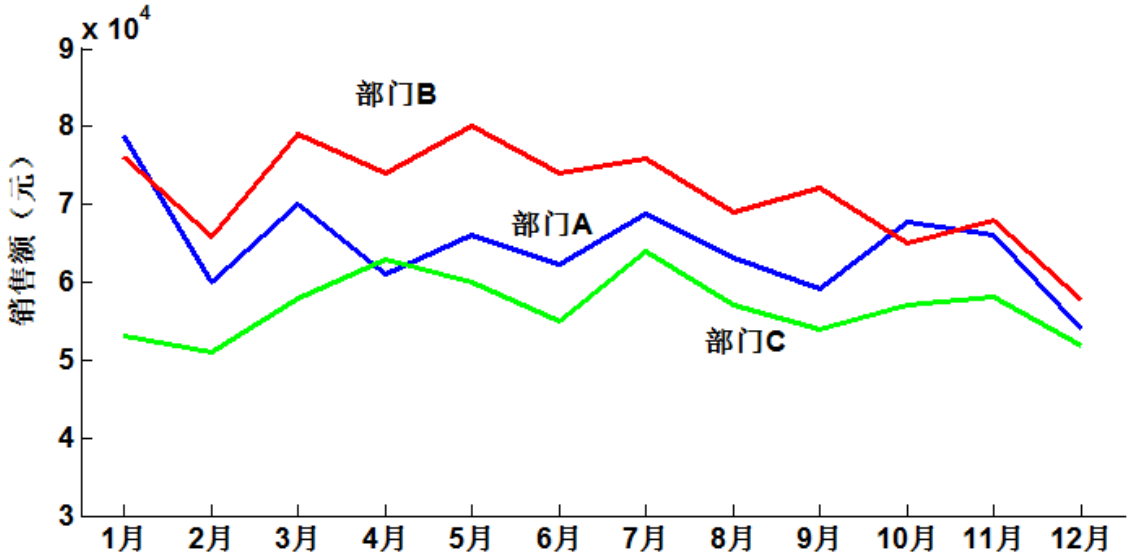
5)计划完成程度相对数

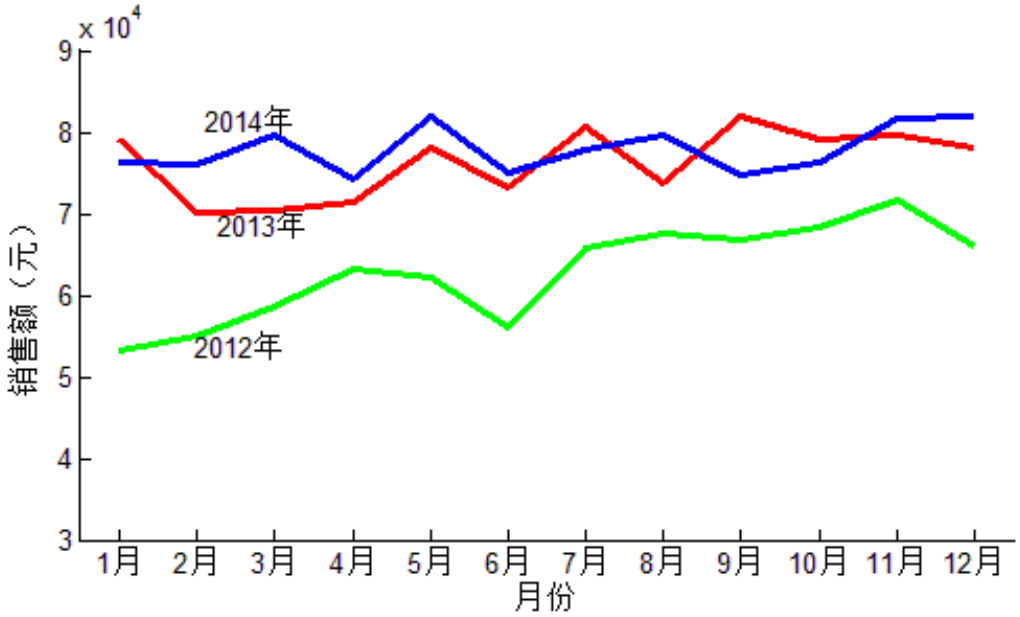
6)动态相对数



对比分析——具体事例

- 各菜品的销售数据，从时间的维度上分析，可以看到甜品部A、海鲜部B、素菜部C三个部门之间的销售金额随时间的变化趋势，了解在此期间哪个部门的销售金额较高，趋势比较平稳
- 也可以从单一部门做分析，了解各月份的销售对比情况





统计量分析

- 用统计指标对定量数据进行统计描述，常从**集中趋势**和**离中趋势**两个方面进行分析。
- 描述集中趋势的度量指标主要有：均值、中位数
- 描述离中趋势的度量指标主要有：极差、标准差、变异系数、四分位间距

均值

- 设 x_1，x_2，…，x_n 是某个数值属性X的n个观测值，

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 \ldots + x_n}{n}$$

每个x_i可以与一个权值w_i相关联。权值反映它们所依附的对应值的意义、重要性或出现的频率。

$$\bar{X} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{w_1 x_1 + w_2 x_2 \ldots + w_n x_n}{w_1 + w_2 \ldots + w_n}$$

加权平均

中位数

- 有序数值的中间值，即把数据较高的一半与较低的一半分开的值。
- 假设给定某属性X的n个值，按递增排序。
  - n为奇数，中位数是该有序集中的中间值；
  - n为偶数，中位数取最中间两个值的平均值。

统计量	意义	不足
平均值	平均数是反映数据集中趋势最常用的统计量，它能充分利用数据所提供的信息	受极端值的影响较大
中位数	中位数是一个位置代表值，表明一组数据中有一半的数据大于（或小于）中位数，计算简便，不受极端值的影响	不能充分利用所有数据信息

极差

- 设 x_1，x_2，…，x_n 是某个数值属性X的n个观测值，极差是最大值与最小值之差。

方差

- 设 x_1，x_2，…，x_n 是某个数值属性X的n个观测值，

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

其中，σ_X为标准差

低方差意味着数据集中分布在均值附近，高方差表示数据散布在一个大的值域中

统计量	意义	不足
极差	反映一组数据的波动范围，计算简单	不能充分利用所有数据信息
方差	反映一组数据的波动大小，方差越大，数据的波动就越大，方差越小，数据的波动越小	计算烦琐，单位与原数据单位不一致

变异系数

- 变异系数度量标准差相对于均值的离中趋势

$$CV = \frac{\sigma}{\bar{x}} \times 100\%$$

其中，σ为标准差，$\bar{x}$ 为均值

四分位间距

设有容量为n的样本观察值 $x_1, x_2, \cdots, x_n$，
按从小到大的顺序排列成 $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$.
p分位数$(0 < p < 1)$记为x_p，它具有以下的性质：

⑴ 至少有 np 个观察值小于或等于 x_p；

⑵ 至少有 $n(1-p)$ 个观察值大于或等于 x_p.

$$x_p = \begin{cases} x_{(\lfloor np \rfloor + 1)}, & \text{当}np\text{不是整数}, \\ \dfrac{1}{2}[x_{(np)} + x_{(np+1)}], & \text{当}np\text{是整数}. \end{cases}$$

p=1/4:第一四分位数Q_1
p=3/4:第三四分位数Q_3

$$Q_3 - Q_1 = IQR$$

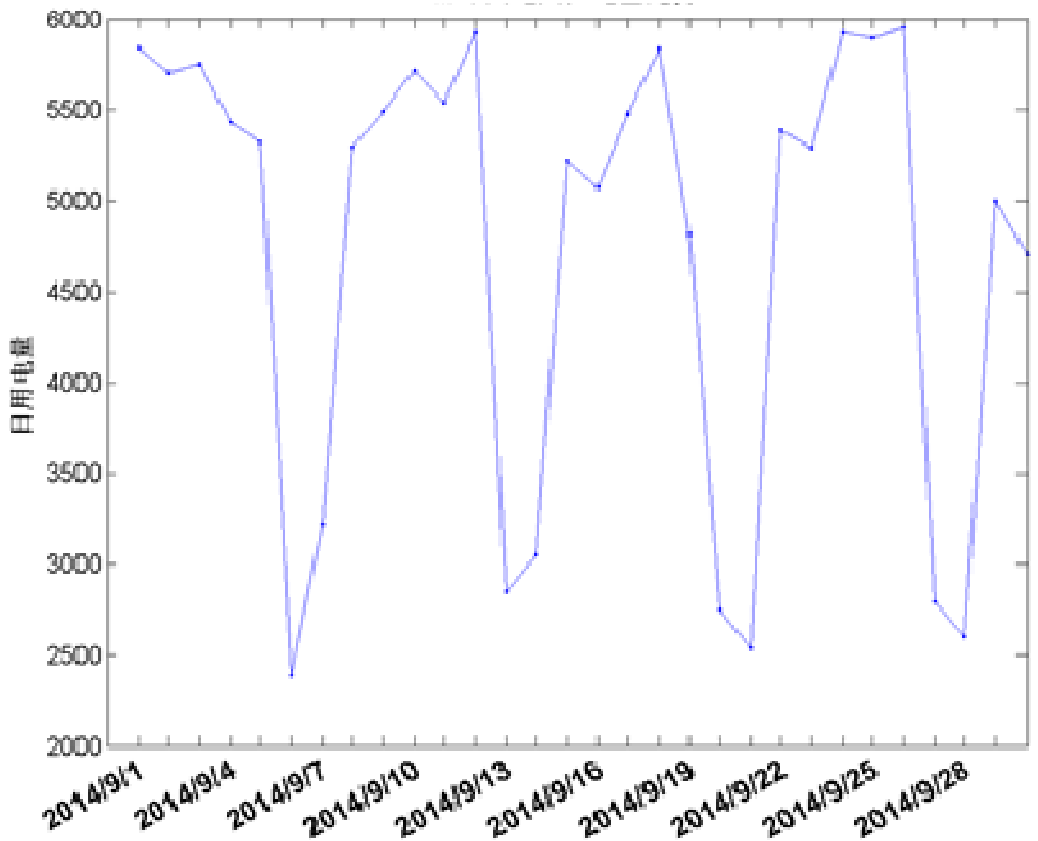
称为四分位数间距.

周期性分析

- 周期性分析是探索某个变量是否随着时间变化而呈现出某种周期变化趋势。
- 时间尺度相对较长的有年度周期性趋势、季节性周期趋势；
- 相对较短的一般有月度周期性趋势、周度周期性趋势，甚至更短的天、小时周期性趋势。

周期性分析-实例

- 某用电单位在2014年9月份日用电量的时序图：

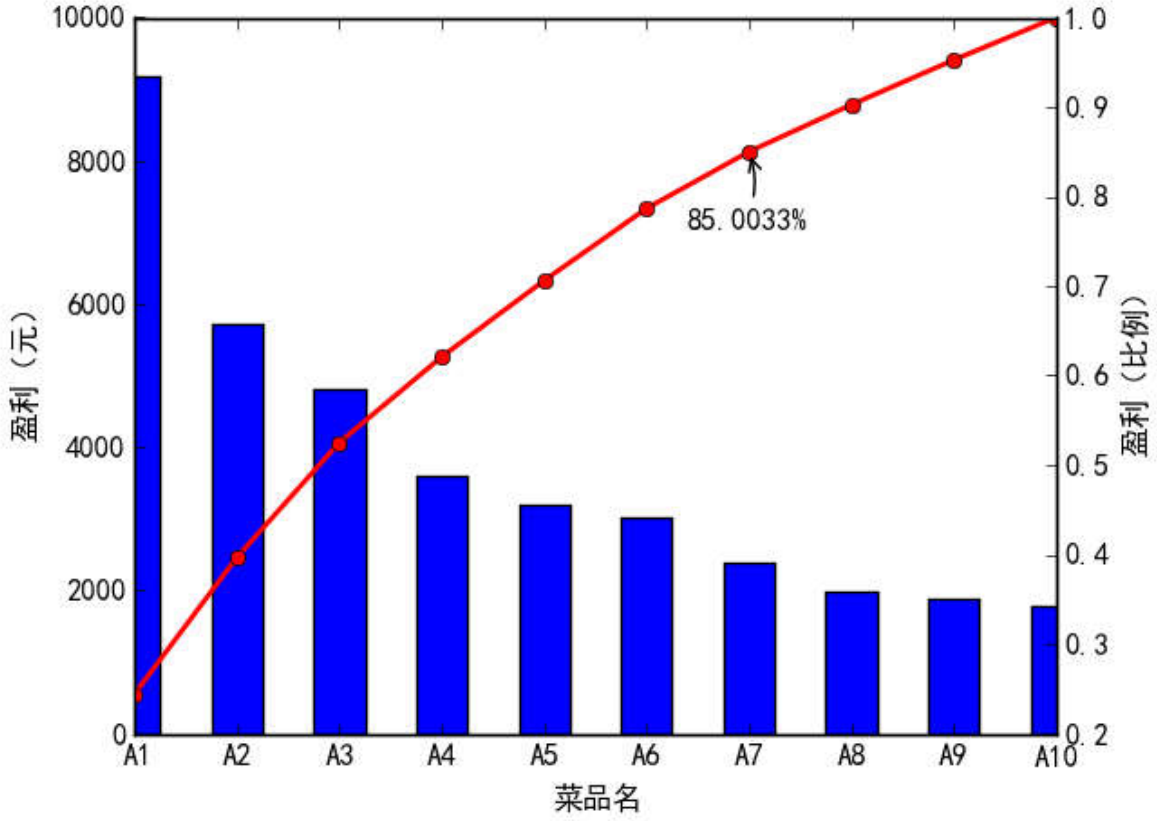


贡献度分析

- 贡献度分析又称帕累托分析，它的原理是帕累托法则，又称20/80定律。
- 比如对一个公司来讲，80%的利润常常来自于20%最畅销的产品；而其他80%的产品只产生了20%的利润。

贡献度分析-实例

- 就餐饮企业来讲，可以重点改善盈利最高的80%的菜品，或者重点发展综合影响最高的80%的部门。

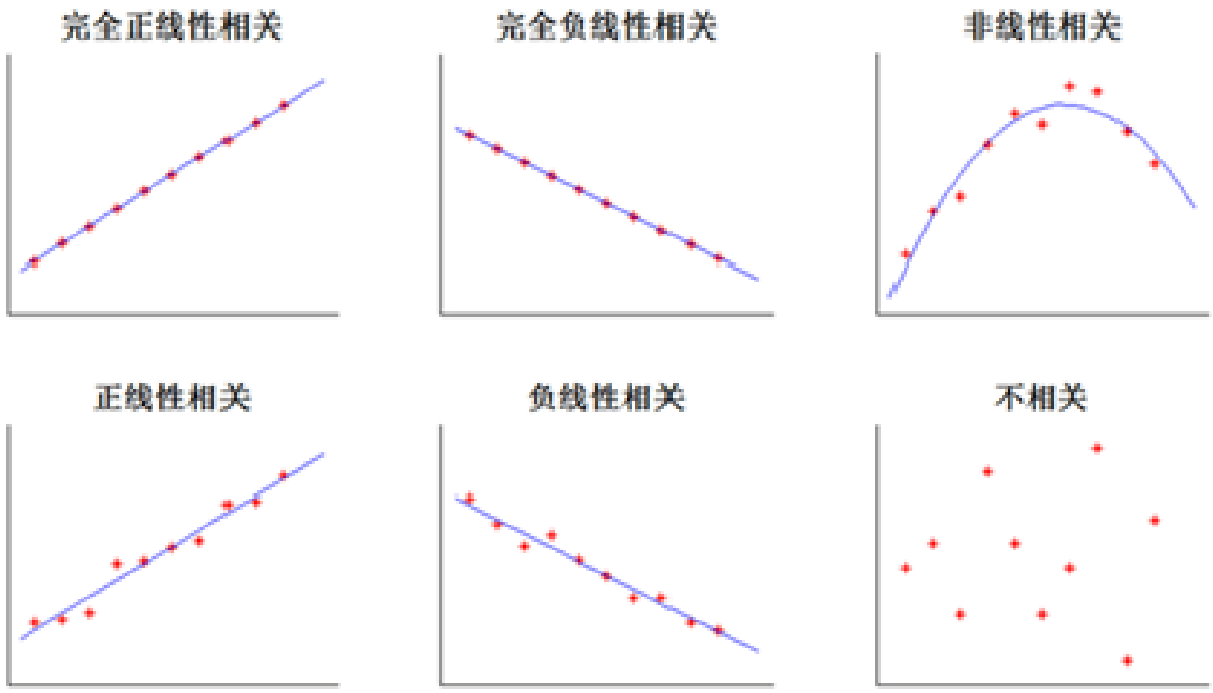


相关性分析

- 分析连续变量之间线性的相关程度的强弱，并用适当的统计指标表示出来的过程称为相关分析。
- 相关性分析方法主要有：
  - 直接绘制散点图
  - 绘制散点图矩阵
  - 计算相关系数

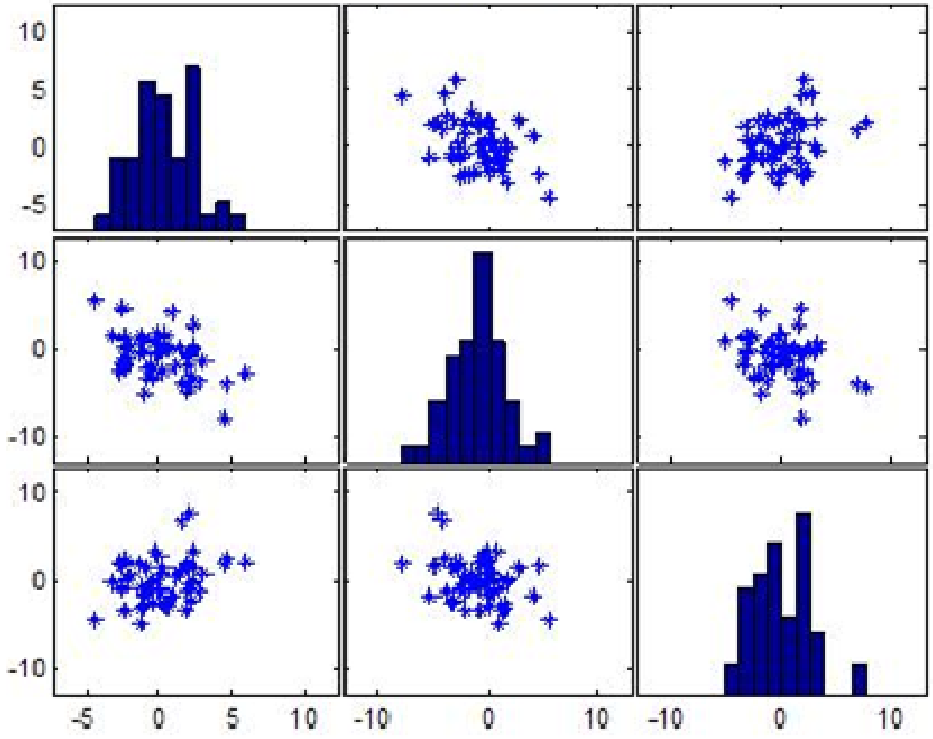
相关性分析-直接绘制散点图

- 判断两个变量是否具有线性相关关系的最直观的方法是直接绘制散点图



相关性分析-绘制散点图矩阵

- 需要同时考察多个变量间的相关关系时，若一一绘制它们间的简单散点图，十分麻烦。
- 此时可利用散点图矩阵来同时绘制各自变量间的散点图，这样可以快速发现多个变量间的主要相关性。



相关性分析——计算相关系数

- 为了更加准确的描述变量之间的线性相关程度，可以通过计算相关系数来进行相关分析。
- Pearson相关系数，一般用于分析两个连续性变量之间的关系

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

- $|r|\le 0.3$ 极弱线性相关或不存在线性相关
- $0.3<|r|\le 0.5$ 低度线性相关
- $0.5<|r|\le 0.8$ 显著线性相关
- $|r|>0.8$ 高度线性相关

数据预处理

- 数据预处理的目的：
  - 提高数据的质量；
  - 让数据更好地适应特定的挖掘技术和工具。
- 数据预处理的主要任务包括数据取样，数据清洗，数据集成，数据变换和数据规约。

数据取样

- 抽取数据的标准，一是相关性，二是可靠性，三是有效性。
- 抽样：主要依赖随机化技术，从数据中随机选出一部分样本。
- 过滤：依据限制条件，仅选择符合要求的数据参与下一步骤的计算。

数据过滤

- 在大数据处理过程中，数据过滤可以采用数据库的基本操作来实现，将过滤条件转换为**选择**操作来实现。

SQL语句：

```
SELECT *
FROM  Student
WHERE Sdept IN ('CS','MA','IS' );
```

数据抽样

- 随机抽样
- 系统抽样
- 分层抽样
- 加权抽样
- 整群抽样

随机抽样

- 随机抽样常常用于总体个数较少时，它的主要特征是从总体中逐个抽取，操作简单易行。
- 抽签法、随机数法、水库抽样。。。

抽签法

- 把总体中的N个个体的编号写在号签上，将号签放在一个容器中，搅拌均匀后，每次从中抽取一个号签，连续抽取n次，就得到一个容量为n 的样本。

随机数法

- 利用随机数表、随机数骰子或计算机产生的随机数进行抽样。

  C语言：rand函数

水库抽样

- 在有限的存储空间里解决无限数据的等概率抽样问题。
- 输入：一组数据，但大小未知。
- 输出：这组数据的k个均匀抽样。
- 对于这个问题有三点要求：

1) 仅允许扫描数据一次。

2) 空间复杂度为O(k)。

3) 扫描数据的前n个数据时(n>k)，要求保存当前已扫描数据的k个均匀抽样。

水库抽样算法

1、申请一个长度为k的数组A保存抽样。

2、保存首先接收到的k个数据。

3、当接收到第i（i>k）个新数据t时，随机生成[1,i]间的随机数j，若j≤k，则以t替换A[j]。

水库抽样算法（续）

- 前k个数据

直接保存在数组A中。前k个数被选中的概率都是1。

- 第K+1个数据

- 第k+1个数据被选中的概率是k/(k+1)
- 水库中原有数据被替换掉的概率是(k/k+1)*(1/k)=1/(k+1)
- 未被替换的概率就是1-1/(k+1)=k/(k+1)
- 所有数据出现的概率相等

水库抽样算法（续）

- 第K+2个数据
- 第k+2个数据被选中的概率是k/(k+2)
- 被k+2个数据替换的概率是：(k/k+2)*(1/k)=1/(k+2)，未被替换的概率就是1-1/(k+2)=(k+1)/(k+2)
- 水库中原有数据留在水库的概率:k/(k+1) *(k+1)/(k+2)=k/(k+2)

- 所有数据出现的概率相等

系统抽样

- 将总体分成均衡的几个部分，然后按照预先定出的规则，从每一部分抽取一个个体，得到所需要的样本。
- 步骤如下：
  - 先将总体的N个个体编号
  - 确定分段间隔k，对编号进行分段。
  - 在第一段用简单随机抽样确定第一个个体编号l（$l \leq k$）
  - 按照一定的规则抽取样本。通常是将l加上间隔k得到第2个个体编号（l+k），再加上k得到第3个个体编号（l+2k），依次进行下去，直到获取整个样本。

系统抽样（续）

- 为了解某大学一年级新生英语学习的情况，拟从503名大学一年级学生中抽取50名作为样本。
  - 将503名学生用随机方式编号为1,2,3，…，503.
  - 用抽签法或随机数表法，剔除3个个体，这样剩下500名学生，对剩下的500名学生重新编号，或采用补齐号码的方式。
  - 确定分段间隔k=500/50=10，将总体分为50个部分，每一部分包含10个个体，第1部分的个体编号为1,2,3,…,10,；第2部分为11,12,13，…,20；以此类推
  - 在第1部分用简单随机抽样确定起始的个体编号，例如5
  - 依次在第2部分，第3部分，…，第50部分，取出号码为15,25，…，495，这样得到一个容量为50的样本。

分层抽样

- 将总体分成互不交叉的层，然后按照一定的比例，从各层独立地抽取一定数量的个体，将各层取出的个体合在一起作为样本
- 适用于总体中的个体有明显差异的情况
- 每个个体被抽到的概率相等，为N/M

分层抽样（续）

- 一个公司的职工有500人，其中不到30岁的有125人，30~40岁的有280人，40岁以上的有95人。为了了解这个单位职工的血压情况，如何从中抽取一个容量为100的样本？
- 由于职工年龄与血压有关，故采用分层抽样方法进行抽取
- 因为样本容量与总体的个数的比为1:5，所以在各年龄段抽取1/5，依次为25、56、19.

加权抽样

- 通过对总体中的各个样本设置不同的数值系数，使样本呈现出希望的相对重要性程度。
- 研究一款啤酒的口味是否需要改变，那么不同参与度的购买者的观点也应该有不同的权值，如：
  - 经常购买该啤酒的客户的权值为3，
  - 偶尔购买该啤酒的客户权值为1，
  - 从不购买的客户的权值为0.1

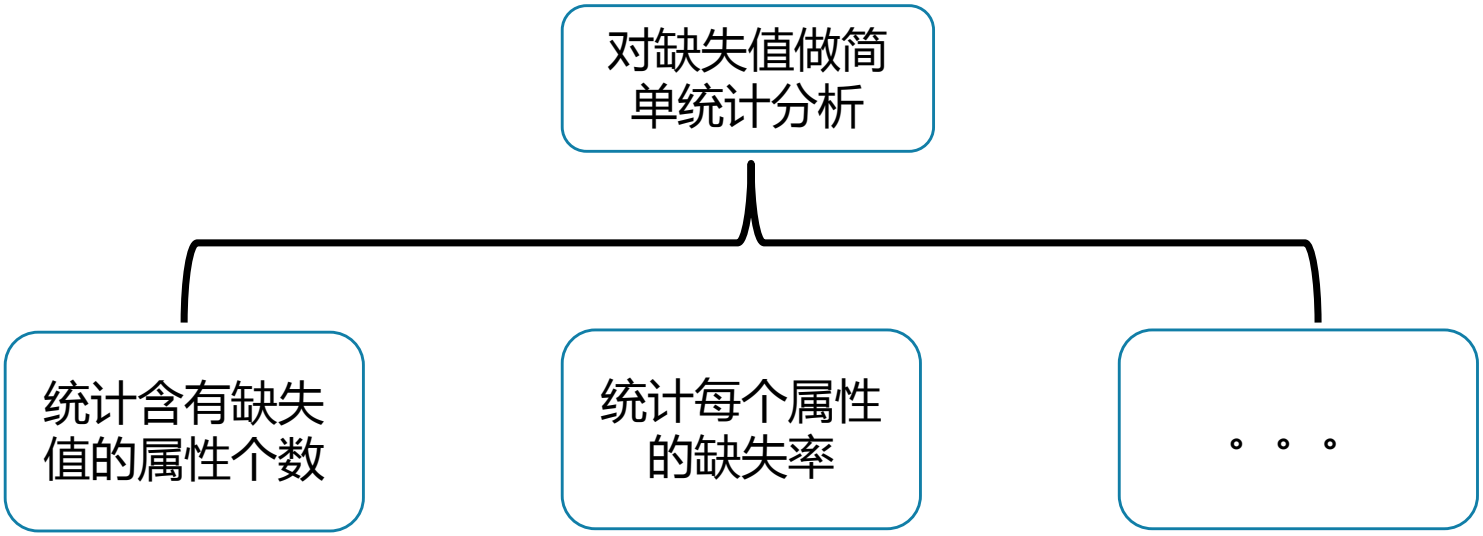
整群抽样

- 又称聚类抽样，是将总体中的个体归并成若干个互不交叉、互不重复的集合，称为群，然后以群为抽样单位抽取样本，具体步骤如下：
  - 将总体分成i个互不重复的部分，每个部分为一个群
  - 根据各群样本量，确定应该抽取的群数
  - 用简单随机抽样或系统抽样方法，从i个群中抽取确定的群数

数据清洗

- 数据清洗主要是删除原始数据集中的无关数据、重复数据，平滑噪声数据，处理缺失值、异常值等。

缺失值的分析



缺失值处理

- 处理缺失值的方法可分为三类：删除记录、数据插补和不处理。其中常用的数据插补方法见下表。

插补方法	方法描述
均值/中位数/众数插补	根据属性值的类型，用该属性取值的平均数/中位数/众数进行插补
使用固定值	将缺失的属性值用一个常量替换。如广州一个工厂普通外来务工人员的"基本工资"属性的空缺值可以用 2015 年广州市普通外来务工人员工资标准 1895 元/月，该方法就是使用固定值
最近临插补	在记录中找到与缺失样本最接近的样本的该属性值插补
回归方法	对带有缺失值的变量，根据已有数据和与其有关的其他变量（因变量）的数据建立拟合模型来预测缺失的属性值
插值法	插值法是利用已知点建立合适的插值函数 $f(x)$，未知值由对应点 x_i 求出的函数值 $f(x_i)$ 近似代替

缺失值处理-拉格朗日插值法

- 对于空间上已知的n个点 可以找到一个n-1次多项式

$y = a_0 + a_1 x \quad + a_2 x^2 + \cdots + a_{n-1} x^{n-1}$，使此多项式曲线经过这n个点

- 将n个点的坐标$(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)$代入多项式，解得拉格朗日插值多项式

$$L(x) = \sum_{i=0}^{n} y_i \prod_{j=o, j \neq i}^{n} \frac{x - x_j}{x_i - x_j}$$

- 将缺失的函数值对应的点 代入插值多项式得到缺失值的近似值

已知三个点A(3,10),B(6,8),C(9,4),求D(10,?)

$$\mathrm{L}(x) = 10\frac{(x-6)(x-9)}{(3-6)(3-9)} + 8\frac{(x-3)(x-9)}{(6-3)(6-9)} + 4\frac{(x-3)(x-6)}{(9-3)(9-6)} = \frac{-x^2 + 3x + 90}{9}$$

D(10,?) —L(10)→ D(10,20/9)

缺失值处理-实例

- 餐饮系统中的销量数据可能出现缺失值，下表为某餐厅一段时间的销量表，其中的缺失值和异常值，用拉格朗日插值对缺失值进行插补。

4275.255

时间	2015/2/25	2015/2/24	2015/2/23	2015/2/22	2015/2/21	2015/2/20
销售额（元）	3442.1	3393.1	3136.6	3744.1	6607.4	4060.3
时间	2015/2/19	2015/2/18	2015/2/16	2015/2/15	2015/2/14	2015/2/13
销售额（元）	3614.7	3295.5	2332.1	2699.3	空值	3036.8

4156.86

异常值分析

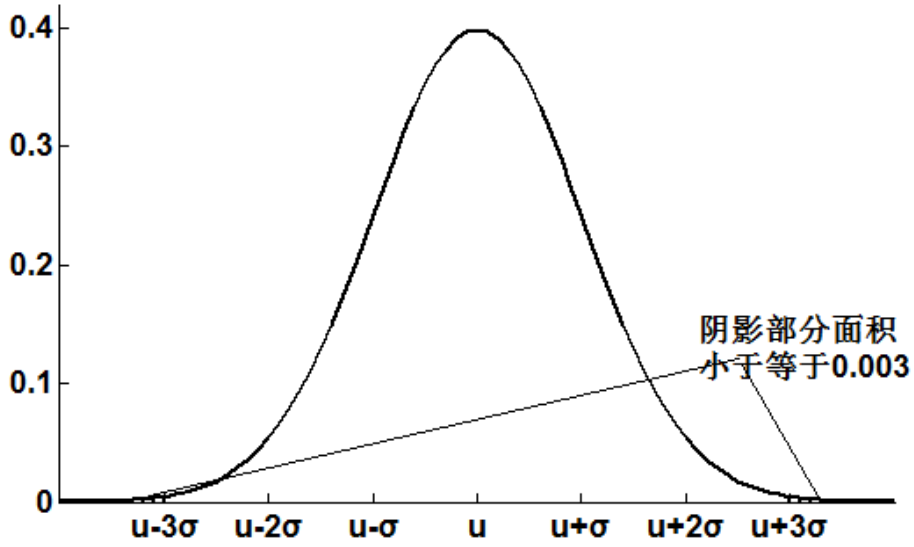
- 异常值分析方法主要有：
  - 简单统计量分析
  - 3σ原则
  - 箱型图分析
  - 基于聚类的离群点检测

简单统计量分析

- 需要的统计量主要是最大值和最小值，判断这个变量中的数据是不是超出了合理的范围
- 如身高的最大值为5米，则该变量的数据存在异常。

3σ原则

- 如果数据服从正态分布，在3σ原则下，异常值被定义为一组测定值中与平均值的偏差超过三倍标准差的值。
- 在正态分布的假设下，距离平均值 3σ 之外的值出现的概率≤0.003，属于极个别的小概率事件。



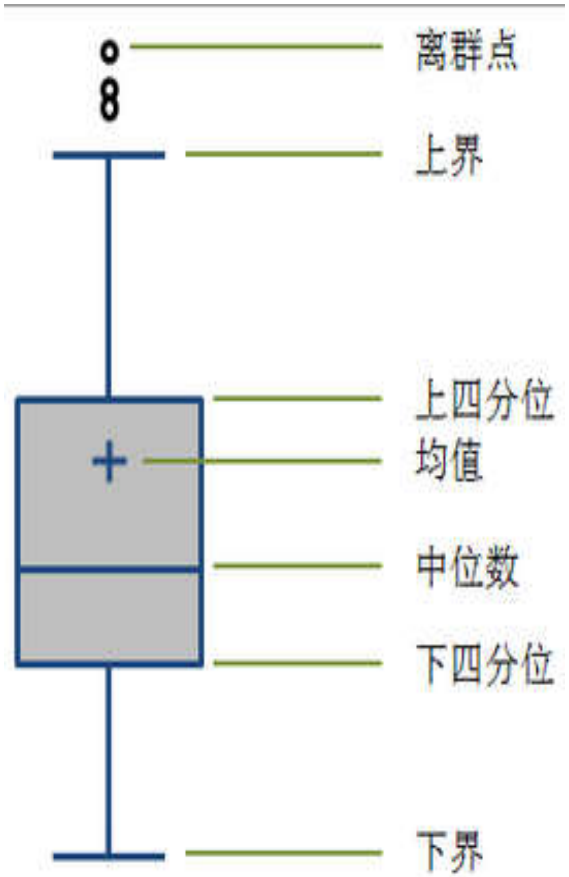
箱线图分析

第一四分位数Q_1与第三四分数Q_3之间的距离：

$Q_3 - Q_1 = IQR$ 称为四分位数间距.

若数据小于 $Q_1 - 1.5IQR$ 或大于$Q_3 + 1.5IQR$，

则认为它是疑似异常值 .



异常值处理

- 在数据预处理时，异常值是否剔除，需视具体情况而定，因为有些异常值可能蕴含着有用的信息。异常值处理常用方法见下表：

异常值处理方法	方法描述
删除含有异常值的记录	直接将含有异常值的记录删除。
视为缺失值	将异常值视为缺失值，利用缺失值处理的方法进行处理。
平均值修正	可用前后两个观测值的平均值修正该异常值。
不处理	直接在具有异常值的数据集上进行挖掘建模。

数据集成

- 大数据分析需要的数据往往分布在不同的数据源中，数据集成就是将多个数据源合并存放在一个一致的数据存储位置（如数据仓库）中的过程。
- 在数据集成时，来自多个数据源的实体的表达形式是不一样的，不一定是匹配的，要考虑实体识别问题和属性冗余问题，从而把源数据在最底层上加以转换、提炼和集成。

数据集成-实体识别

- 实体识别的任务是检测和解决同名异义、异名同义、单位不统一的冲突。
- 同名异义：数据源A中的属性ID描述的是菜品编号，而数据源B中的属性ID是订单编号，即ID描述的是不同的实体。
- 异名同义：数据源A中的sales_dt和数据源B中的sales_date都是描述销售日期的，即A. sales_dt= B. sales_date。
- 单位不统一：描述同一个实体分别用的是国际单位和中国传统的计量单位。

数据集成-冗余属性识别

- 数据集成往往导致数据冗余，如：
  - 同一属性多次出现
  - 同一属性命名不一致导致重复
- 仔细整合不同源数据能减少甚至避免数据冗余与不一致，以提高数据挖掘的速度和质量。对于冗余属性要先分析，检测后再将其删除。

数据变换

- 数据变换主要是对数据进行规范化的操作，将数据转换成“适当的”格式，以适用于挖掘任务及算法的需要。

数据变换-简单函数变换

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

$$x' = x^2$$

$$x' = \sqrt{x}$$

$$x' = \log(x)$$

$$\nabla f(x_k) = f(x_{k+1}) - f(x_k)$$

数据变换-规范化

- 数据规范化（归一化）处理是将数据按照比例进行缩放，使之落入一个特定的区域，从而进行综合分析。
- 常用的规范化方法：最小-最大规范化、零-均值规范化、小数定标规范化

数据变换-规范化（续）

- 最小-最大规范化：也称为离差标准化，是对原始数据的线性变换，使结果值映射到[0,1]之间。

$$x^* = \frac{x - min}{max - min}$$

- 零-均值规范化:也叫标准差标准化，经过处理的数据的平均数为0，标准差为1。

$$x^* = \frac{x - \overline{x}}{\sigma}$$

- 小数定标规范化:通过移动属性值的小数位数，将属性值映射到[-1，1]之间，移动的小数位数取决于属性值绝对值的最大值。

$$x^* = \frac{x}{10^k}$$

数据变换-连续属性离散化

- 连续属性离散化就是在数据的取值范围内设定若干个离散的划分点，将取值范围划分为一些离散化的区间，最后用不同的符号或整数值代表落在每个子区间中的数据值。
- 常用的离散化方法：

• 等宽法-将属性的值域分成具有相同宽度的区间，区间的个数由数据本身的特点决定或者用户指定，类似于制作频率分布表

• 等频法-将相同数量的记录放进每个区间。

• 基于聚类分析的方法-首先将连续属性的值用聚类算法进行聚类，然后再将聚类得到的簇进行处理，合并到一个簇的连续属性值做同一标记。

数据变换——属性构造

- 进行防窃漏电诊断建模时，已有的属性包括供入电量、供出电量。为了判断是否存在有窃漏电行为的大用户，需要构造一个新的关键指标--线损率，该过程就是构造属性。
- 新构造的属性线损率计算公式如下：

线损率 =（供入电量-供出电量）/供入电量

- 线损率的范围一般在3%~15%，如果远远超过该范围，就可以认为该条线路的大用户很大可能存在窃漏电等用电异常行为。

数据归约

- 数据归约是将海量数据进行归约，归约之后的数据仍接近于保持原数据的完整性，但数据量小得多。
- 数据归约的意义：
- 降低无效、错误数据对建模的影响，提高建模的准确性
- 少量且具代表性的数据将大幅缩减数据挖掘所需的时间
- 降低储存数据的成本

数据归约-属性归约

- 属性归约常用方法有：合并属性、逐步向前选择、逐步向后删除、决策树归纳、主成分分析
- 合并属性

  初始属性集： $\{A_1, A_2, A_3, A_4, B_1, B_2, B_3, C\}$

$$\begin{matrix}\{A_1, A_2, A_3, A_4\} \to A; \\ \{B_1, B_2, B_3\} \to B.\end{matrix} \quad \Rightarrow \quad \{A, B, C\}$$

- 逐步向前选择

  初始属性集： $\{A_1, A_2, A_3, A_4, A_5, A_6\}$

$$\{\} \Rightarrow \{A_1\} \Rightarrow \{A_1, A_4\} \Rightarrow \{A_1, A_4, A_6\}$$

数据归约-属性归约

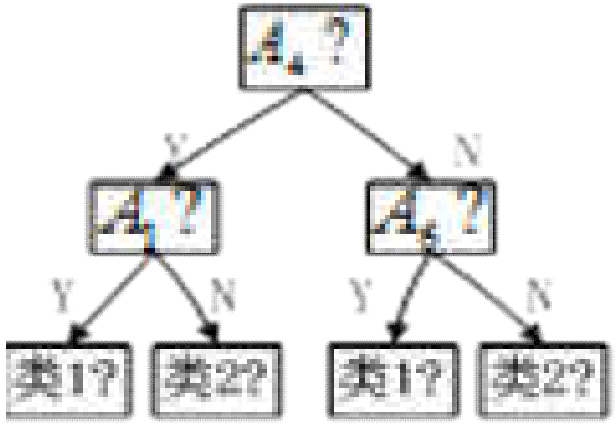
- 逐步向后删除
  初始属性集：$\{A_1, A_2, A_3, A_4, A_5, A_6\}$

  $\Rightarrow \{A_1, A_3, A_4, A_5, A_6\} \Rightarrow \{A_1, A_4, A_5, A_6\} \Rightarrow \{A_1, A_4, A_6\}$

- 决策树归约
  初始属性集：

$\{A_1, A_2, A_3, A_4, A_5, A_6\}$

$\Rightarrow \{A_1, A_4, A_6\}$



数据归约——属性归约

- 主成分分析

  用较少的变量去解释原始数据中的大部分变量，即所谓主成分，来代替原始变量进行建模。