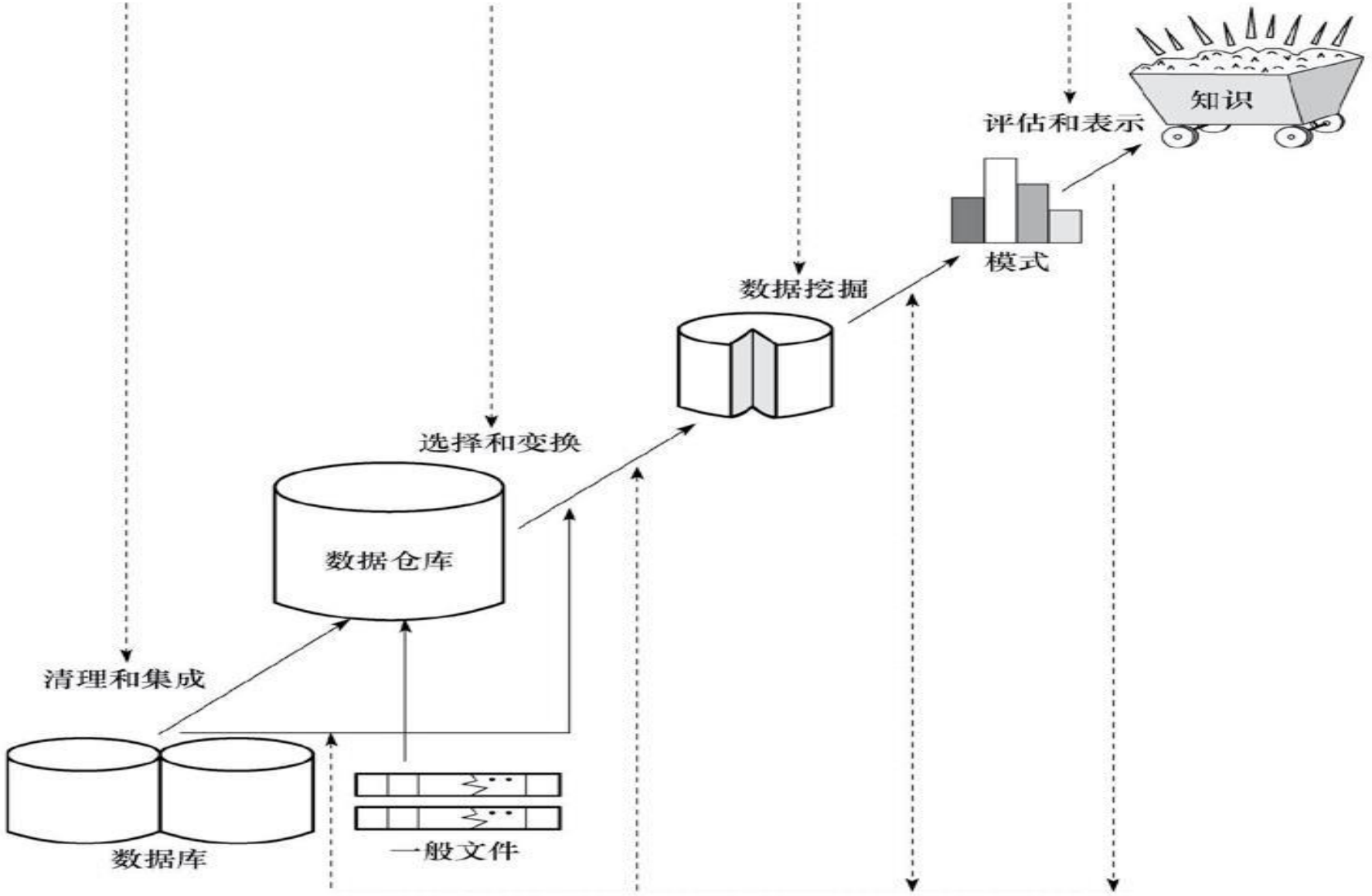
数据挖掘
第五讲• 数据预处理

数据挖掘步骤



一些常见的问题

❖ 工资怎么是-10？

准确性问题

- 该销售商品是否做了降价销售广告？

  没有相关记录

完整性问题

- 同一种商品，在不同表中类别编码不一样 ？

不一致性

数据质量：为什么要预处理（1/2）

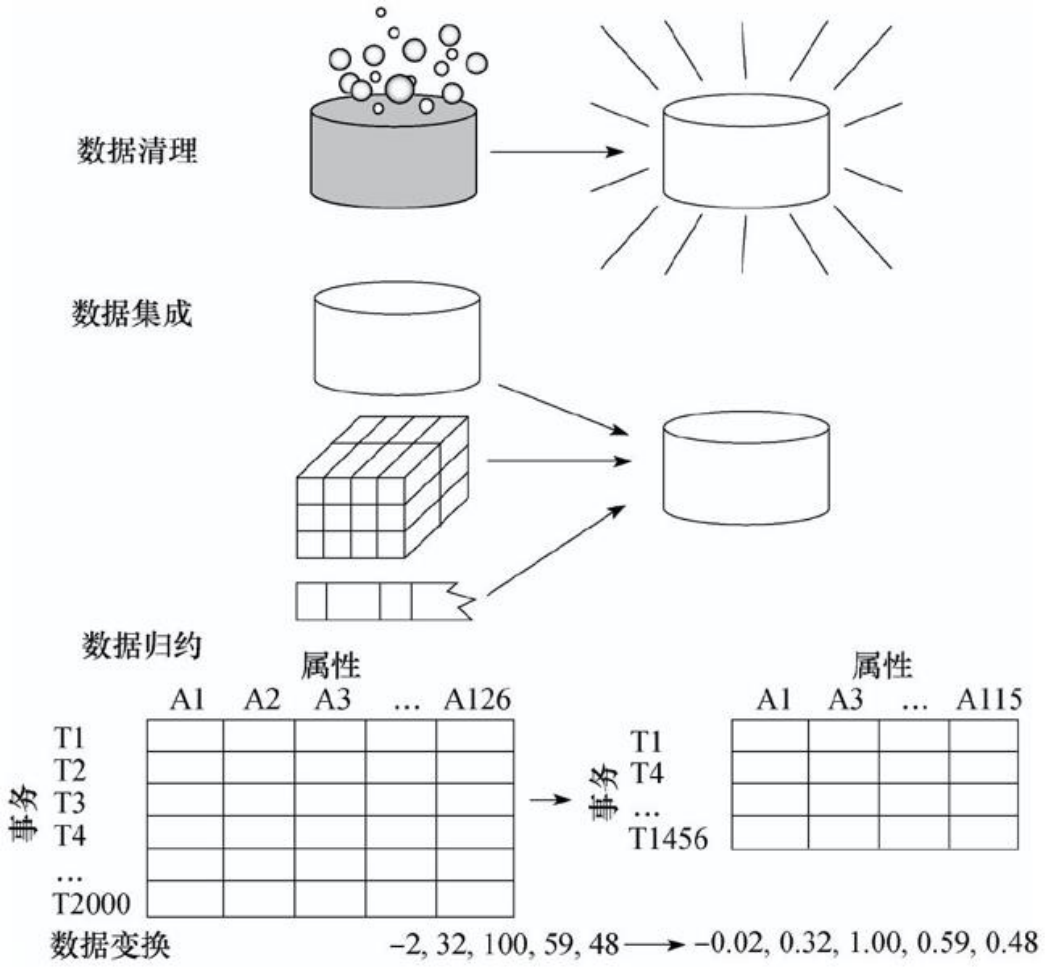
- 不正确或含噪声：包含错误或存在偏离期望的值
  - 采集: 设备故障，人或计算机错误，故意掩盖
  - 传输: 缓冲区大小限制出现偏差
- 不完整：缺少属性值或者缺少某些感兴趣的属性,或仅包含聚集数据
  - 理解错误，人、硬件或软件的问题
- 不一致：同一属性采用的编码或命名不同导致存在差异
  - 不同数据源命名约定不同，输入字段格式不一致

数据质量：为什么要预处理（2/2）

- 时效性：未能及时更新数据导致数据不准确
  - 高端销售代理月销售红利分布更新不及时问题
- 可信性：多少数据是用户信赖的
  - 数据库错误，对用户产生影响，即使修改，用户不再相信
- 可解释性：数据是否容易理解
  - 使用了很多会计编码，销售部门不理解

高质量的数据是高质量决策的基础

数据预处理


数据清理
数据集成
数据归约
属性
A1 A2 A3 … A126
事务
T1
T2
T3
T4
…
T2000
属性
A1 A3 … A115
事务
T1
T4
…
T1456
数据变换
-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Outline

数据清理

- 数据清理
  - 通过填写空缺的值，平滑噪声并识别离群点、纠正数据中的不一致。
- 基本方法
  - 缺失值处理
  - 噪声数据处理

数据清理（1）：缺失值

- 许多元组的一些属性没有记录值，怎样才能为该属性填上丢失的值?
  - 忽略元组;
  - 人工填写空缺值;
  - 使用一个全局常量填充空缺值，比如用一个常数（unknown）来替换所有空缺的值;
  - 使用属性的平均值填充空缺值;
  - 使用与给定元组属同类的所有样本的平均值;
  - 使用最可能的值填充空缺值，可以使用回归，或决策树确定推理获得。

数据清理（2）：噪声数据

- 噪声（noise）
  - 被测量的变量的随机误差。

- 如何才能“光滑”数据，去掉噪声？
  - 分箱 binding
  - 回归 Regression
  - 聚类 Clustering

分箱（binding）

- 分箱是根据考察数据的近邻值来平滑有序数据的值。
- 步骤如下：
  - 数据排序
  - 划分到等频箱中（即每箱包含相同的数据个数）
  - 用箱中的均值、中位数或者边界光滑数据
    - 按箱平均值平滑(smoothing by bin means)，即箱中每一个值用箱的均值替换。例如每箱3个进行分箱，然后用这三个值的平均值代替箱中的值
    - 按箱中值平滑(smoothing by bin medians)
    - 按箱边界平滑(smoothing by bin boundaries)：箱中每一个值替换为距离最近的边界值。

分箱例子

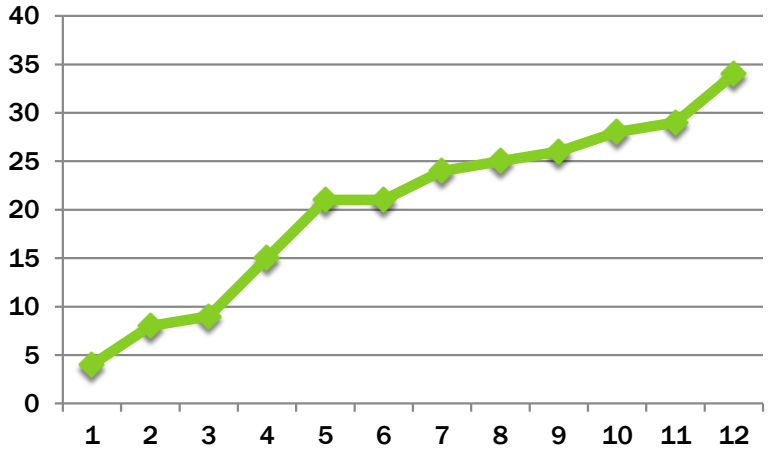
- 排序后的价格
  - 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

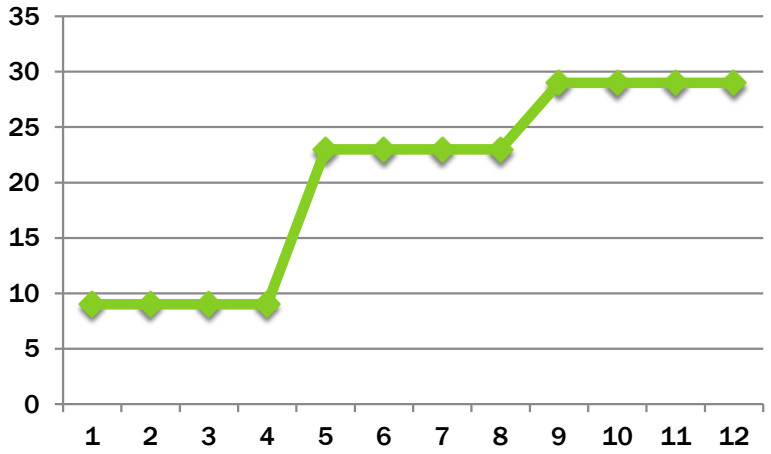
划分为等频箱:
- 箱 1: 4, 8, 9, 15
- 箱 2: 21, 21, 24, 25
- 箱 3: 26, 28, 29, 34

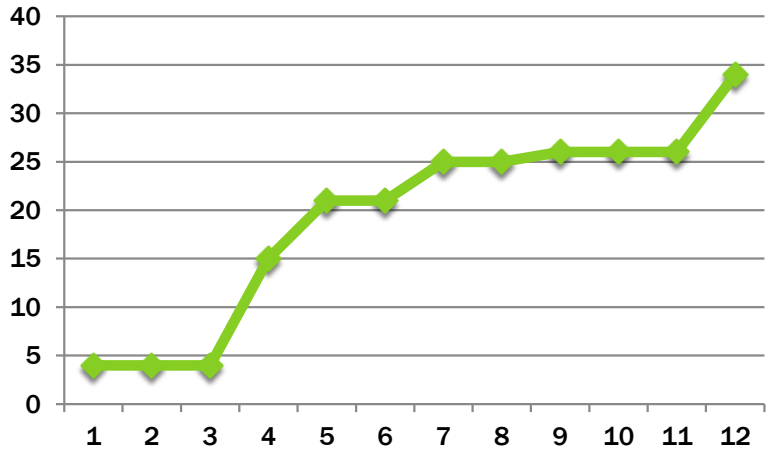
用箱均值光滑:
- 箱 1: 9, 9, 9, 9
- 箱 2: 23, 23, 23, 23
- 箱 3: 29, 29, 29, 29

用箱边界光滑:
- 箱 1: 4, 4, 4, 15
- 箱 2: 21, 21, 25, 25
- 箱 3: 26, 26, 26, 34

分箱例子

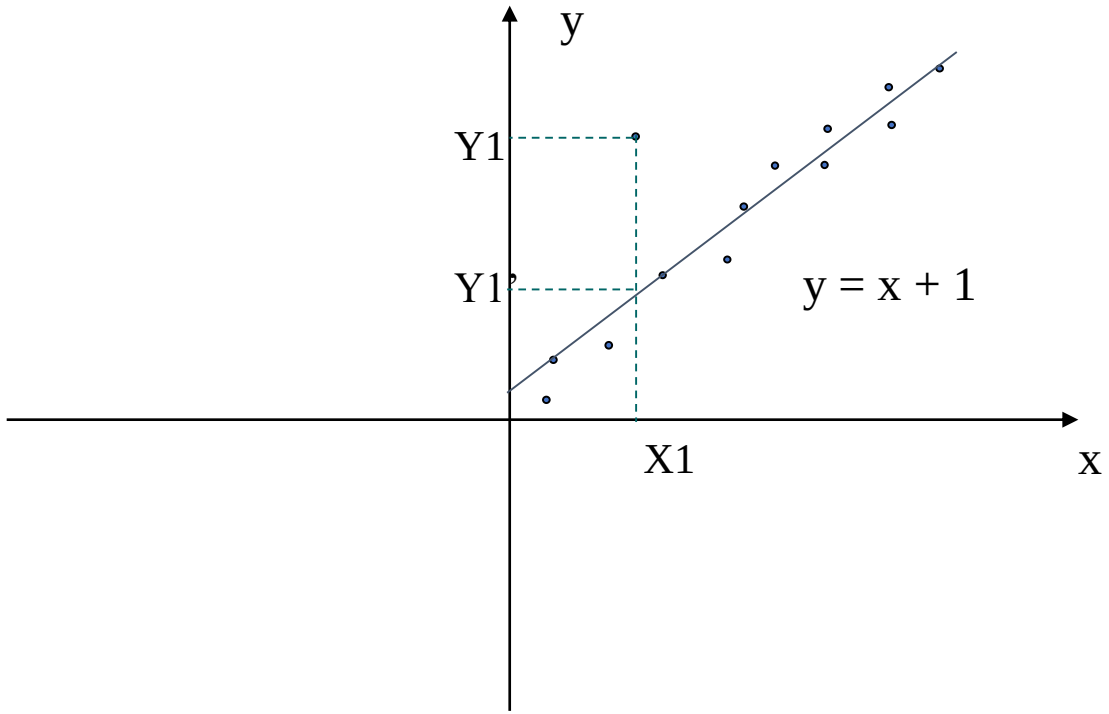

40
35
30
25
20
15
10
5
0
1 2 3 4 5 6 7 8 9 10 11 12


35
30
25
20
15
10
5
0
1 2 3 4 5 6 7 8 9 10 11 12


40
35
30
25
20
15
10
5
0
1 2 3 4 5 6 7 8 9 10 11 12

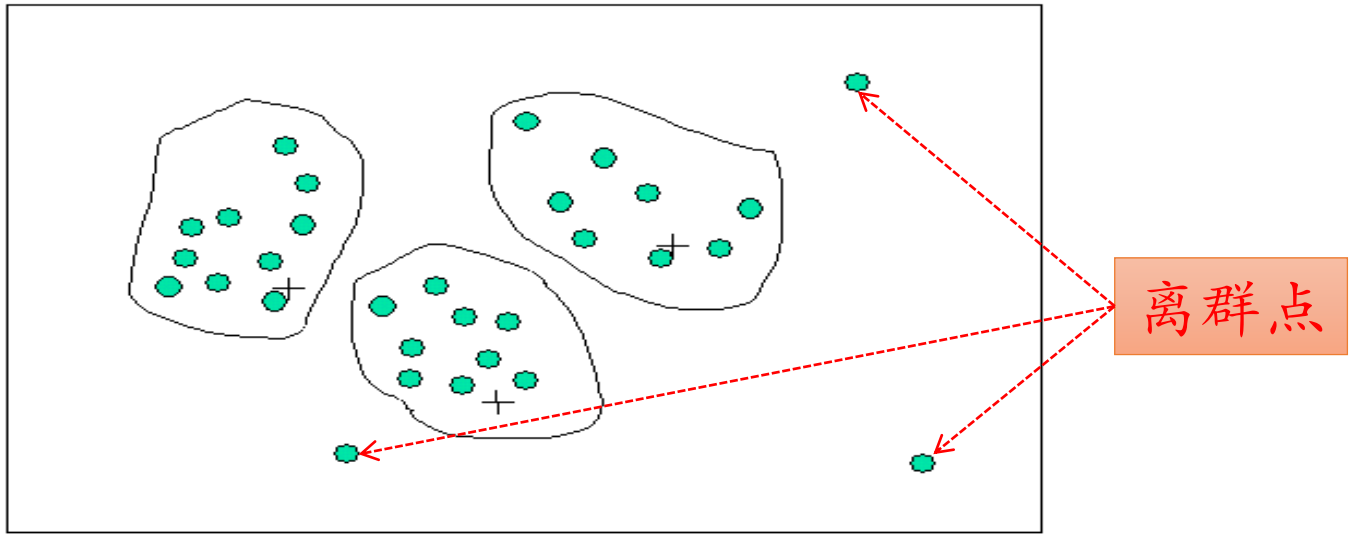
回归

- 用一个函数拟合数据来光滑数据
  - 如线性回归、多元线性回归



聚类(clustering)

- 通过聚类检测离群点：
  - 聚类将相似的值组织成群或类，落在群或类外的值就是孤立点，也就是噪声数据。



Outline

数据集成

- 数据集成
  - 将多个数据源中的数据结合起来存放在一个一致的数据存储（如数据仓库）中；
  - 源数据可能包括多个数据库，数据立方体或一般文件。
- 重要问题：
  - 实体识别问题（模式集成和对象匹配）
  - 冗余和相关分析
  - 元组重复
  - 数据值冲突的检测与处理

数据集成（1）：实体识别问题

- 实体识别问题
  - 来自多个信息源的现实世界的等价实体如何才能匹配?
  - 模式集成:
    - 确定来自不同数据源的属性是否表示同一个实体;
    - 例：如何确信一个数据库custom_id和另一个数据库中的cust_number指的实体是同一个实体?
  - 利用元数据-关于数据的数据（属性名称、含义、数据类型、属性取值范围），避免模式集成中的错误。

数据集成（2）：冗余和相关分析

- 冗余（redundancy）：
  - 一个属性是冗余的，如果它能由另一个或另一组数据“导出”，属性命名的不一致也可能导致数据集中的冗余。
  - 对于数值属性，可以采用相关分析检测到。
  - 对于离散属性，可以采用卡方检验发现。

相关分析检测方法

- **卡方检验**（chi-square test，χ2检验），用来验证两个总体标称属性之间是否存在关联
- 设属性 A 和B分别有c个不同的值 和r个不同的值：
  - $a_1, a_2, \ldots, a_c$ $b_1, b_2, \ldots, b_r$
  - A和B描述的数据元组用如下的（r*c）的相依表示：
  - 其中 o_{ij} /e_{ij} 分别表示元组在属性A上取a_i，在属性B上取b_j的联合事件的观测频度和期望频度。

	b_1	b_2	…	b_r
a_1				
a_2				
…			o_{ij} /e_{ij}	
a_c				

相关分析检测方法

- 利用卡方确定相关性

  计算 χ^2 值:

$$\chi^2 = \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

其中的求和运算是对于r*c个单元上计算，

$$e_{ij} = count(A=a_i) * count(B=b_j) / N$$

- χ^2检验假设属性A和B独立。对于r*c的自由表，检验自由度为（r-1）*（c-1）。在一定的置信水平(如：0.0001)，在卡方分布的表中查询拒绝该假设的卡方值。
- 若计算值大于查到的卡方值，则拒绝这个假设，说明属性是统计相关的。

例3.1 Gender和 preferred reading 相关性

	男	女	合计
小说	250（90）	200（360）	450
非小说	50（210）	1000(840)	1050
合计	300	1200	1500

$$\chi^2 = \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

判断方法:
1 卡方值结果：507.93
2　自由度1，0.001置信水平
3　拒绝假设的值为10.828

强相关

相关分析检测方法

- **数值数据的相关系数**(也称为皮尔逊积矩系数Pearson’s product moment coefficient)

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum(a_i b_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B}$$

N是元组个数，$\bar{A}$ 和$\bar{B}$ 是A，B的均值，σ_A和 σ_B分别是属性A 和 B的 标准差，$\Sigma(a_i b_i)$ 是AB 叉积的和. a_i 和 b_i 是元组 i中属性A 和B 的值。

- 如果 $r_{A,B} > 0$, **A 和 B 是正相关的**，意味着A的值随着B的值的增加而增加，该值越大，相关性越强.
- 如果 $r_{A,B} < 0$, A 和 B 是**负相关**的;
- 如果 $r_{A,B} = 0$, A 和 B 是**独立的**;

相关分析检测方法

- 数值数据的**协方差**（Covariance）：期望值分别为$E[X]$与$E[Y]$的两个实随机变量X与Y之间的协方差$Cov(X,Y)$定义为：

$$\mathrm{Cov}(A,B) = \mathrm{E}((A - \mathrm{E}(A))(B - \mathrm{E}(B)))$$

$$= \frac{\sum_{i=1}^{N}(a_i - \overline{A})(b_i - \overline{B})}{n}$$

$$r_{A,B} = \mathrm{cov}(A,B) / \sigma A \sigma B$$

$$\mathrm{cov}(A,B) = E(A \cdot B) - \overline{A}\,\overline{B})$$

相关分析检测方法

- **协方差：**表示两个变量的总体误差
  - 如果两个变量的变化趋势一致，即其中一个大于自身的期望值，另外一个也大于自身的期望值，那么两个变量间的**协方差就是正值**。
  - 如果两个变量的变化趋势相反，那么两个变量间的**协方差就是负值**。
  - A和B独立，则协方差为0；反之，在符合多元正态分布的情况下，**协方差0蕴含独立性**。

方差和协方差之间的关系？

例3.2 Allelectronics 和 HighTech的股票价格

时间点	Allelectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

判断其是否同时上涨

数据集成（3）：数据值冲突的检测与处理

- 数据冲突：
  - 对于现实同一个实体，来自不同数据源的属性值可能不同；
  - 可能是因为表示、尺度、编码不同导致的。
- 两个例子
  - 例1：对于现实世界的同一实体，来自不同数据源的属性值可能不同，因为表示、比例或编码不同。
  - 例2：重量属性可能在一个系统中以公制单位存放，而在另一个系统中用英制存放。

数据变换（1）：概述

- 数据变换将数据转换成适合挖掘的形式：
  - 平滑(Smoothing)：去掉数据中的噪声；主要技术有分箱、聚类和回归；
  - 属性构造(Attribute/Feather construction)：构造新的属性并添加到属性集中，以帮助挖掘；
  - 聚集(Aggregation)：对数据进行汇总和聚集；
  - 规范化(Normalization)：将属性数据按比例缩放，使之落入一个小的特定区间，如 0到1之间；
  - 数据概化(Generalization)：使用概念分层，用高层次概念值替换低层“原始”数据。

数据变换（2）：属性构造

- 属性构造
  - 由给定的属性构造和添加新的属性，以有利于挖掘。
  - 比如，根据属性heigh 和 width可以构造 area属性。

- 属性构造可以发现关于数据属性间联系的丢失信息，这对知识发现是有用的。

数据变换（3）：规范化

- 为什么规范化
  - 不同的属性取值范围不一样，为了避免数据对度量单位选择的依赖性，对数据进行处理，使数据落入较小的共同区域；
  - 规范化对于神经网络、基于距离的分类、聚类等算法十分重要。
- 规范化方法
  - 最小-最大规范化
  - z-score规范化
  - 小数定标规范化

最小-最大规范化

- 对原始数据进行线性变换
  - 假定A的属性最大值和最小值分别是max，min。
  - 设A中的任一值是V，则V的值经过规范化后则为：

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

  - 规范化后的区间是 $[new\_\min_A, new\_\max_A]$
  - 特点：保持原数据之间的关系；可检测新数据是否“越界”。

  - 例：income的最大，最小值分别为9000，2000，则将它的值映射到[0,1]时，若income的值6800规范后为：

    (6800-2000)/(9000-2000)*(1-0)+0=0.686

z-score规范化

- z-score规范化（或零-均值规范化）
  - 属性A的值基于A的平均值和标准差规范化；
  - 假设A的值V规范后为V’，由下式计算：

$$v_i' = \frac{v_i - \bar{A}}{s_A}$$

  - 特点：当属性A的最大最小值未知，或离群点对规范化影响较大时，可使用z-score规范化。

  - 例：假设属性income的平均值和方差分别为：5400，1600，则值7360的规范后的值为：

    (7360-5400)/1600=1.225

小数定标规范化

- 通过移动属性A的值的小数点位置进行规范化

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

  - j是使得 $\max(|v_i'|)<1$ 的最小整数

- 例：假设A的值取值范围为[-986,917]，求规范化后取值范围？
  - A的最大绝对值是986;
  - 因此，j=3;
  - -986被规范化为-0.986;
  - 917被规范化为0.917。

Outline