Introduction/Data Preprocessing

✏️ by kingno

Introduction

1、什么是数据挖掘

许多人把数据挖掘视为**数据中的知识发现（KDD）**，其过程如下：

1. 数据清理：消除噪声和删除不一致数据
2. 数据集成：多种数据源组合在一起，结果存在数据仓库中
3. 数据选择：从数据库中提取与分析任务相关的数据
4. 数据变换：通过汇总或聚集操作，把数据变换和统一成适合挖掘的形式
5. 数据挖掘：基本步骤，使用智能方法提取数据模式
6. 模式评估：根据某种度量，识别真正有趣的模式
7. 知识表示：知识可视化等

1~4为数据预处理

2、可以挖掘什么类型的数据

- 数据库数据。利用SQL语句，提取有用的信息
- 数据仓库。数据仓库是一个从**多个数据源**收集的信息存储库，放在**一致的模式**下，并且通常驻留在**单个站点**上。数据仓库通过数据清理、数据变换、数据集成、数据装入和定期数据刷新来构造。

  数据仓库围绕**主题**（如顾客、商品、供应商和活动）组织。数据存储从**历史的角度**（如过去6~12个月）提供信息，并且通常是**汇总的**。

通常，数据仓库用称作数据立方体的多维数据结构建模

数据仓库非常适合联机分析处理。**OLAP 操作**（上钻、下卷）使得用户在不同的汇总级别观察数据

数据仓库的具体信息有相关章节

- 事务数据。从事务数据库中获取信息
- 其他类型的数据

3、可以挖掘什么类型的模式

这些任务可以分类两类：描述性和预测性

- 特征化与区分。数据特征化，总结数据的一般特性
- 频繁模式、关联和相关性挖掘：比如，挖掘出尿布和啤酒的关联
- 分类与回归
- 聚类分析
- 离群点分析

Know Your Data

1、数据的基本统计描述

1.1 中心趋势度量：均值、中位数和众数

数据集的“中心”在哪里?

1. 均值:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

   加权平均:

$$\overline{x} = \frac{w_1 x_1 + w_2 x_2 + \ldots + w_N x_N}{w_1 + w_2 + \ldots + w_N}$$

然而，均值对极端值很敏感。解决方法：截尾平均，即丢掉高低极端值再取平均

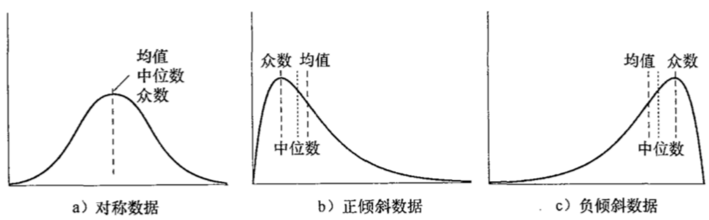
2. 中位数是有序数据的中间值，适用于倾斜数据。假如是偶数个值，取最中间两个数的平均值

当观测数据很多时，我们可以计算中位数的近似值：假定数据根据它们的 x_i 值划分成区间，并且已知每个区间的频率。则中位数为

$$\text{median} = L_1 + \left(\frac{N/2 + (\sum freq)_l}{freq_{median}}\right) width$$

- L_1 ：中位数区间的下界
- N ：整个数据集值的个数
- $(\sum freq)_l$ ：低于中位数区间的所有区间的频率和
- $freq_{median}$ ：中位数区间的频率
- $width$ ：中位数区间的宽度

3. 众数，即集合中出现最频繁的值。有单峰、双峰、三峰

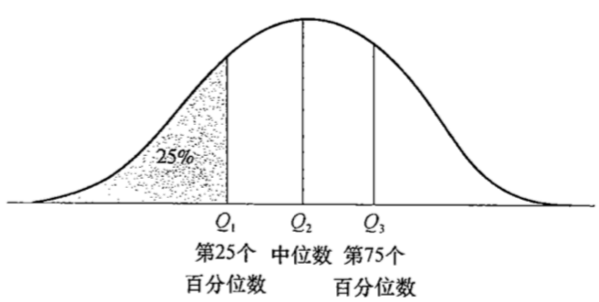


有一个经验公式：$mean - mode = 3 \times (mean - median)$

中列数：$(\min + \max)/2$

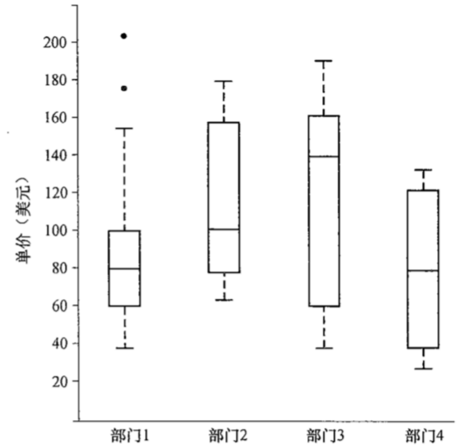
1.2 度量数据散步：极差、四分位数、方差、标准差和四分位数极差

1. 极差：$\max() - \min()$
2. 4-分位数是三个数据点，把数据分布划分为 4 个相等的部分. Q_1, Q_2, Q_3 分别对应 $25\%, 50\%, 75\%$ 的位置



3. 四分位数极差 $IQR = Q_3 - Q_1$

盒图：



- 盒子的端点在四分位数上，使得盒的长度是 IQR
- 中位数 用盒内的线标记
- 盒外的两条线（胡须）延伸到 最小 和 最大 观测值
- 胡须最长长度为 $1.5IQR$ ，剩余的点为 离群点

4. 方差

$$\text{样本方差:}\ s^2 = \frac{1}{N-1}(x_i - \overline{x})^2$$

$$\text{总体方差:}\ \sigma^2 = \frac{1}{N}(x_i - \overline{x})^2$$

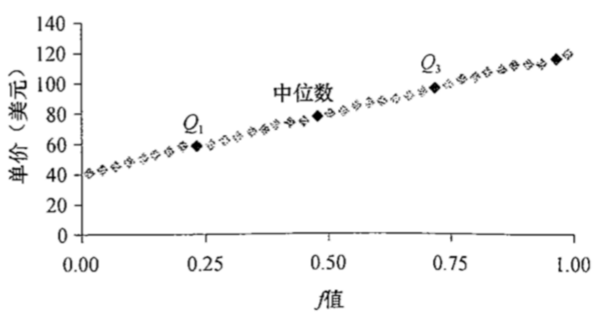
   标准差是方差的平方根

5. 正态分布的三 σ 准则：$(u-3\sigma, u+3\sigma)$ 区间中包含了 99.7% 的值

1.3 数据的基本统计描述的图形显示

分位数图、分位数-分位数图、直方图、散点图

1. 分位数图，可用于观察单变量数据的分布。将数据递增排序，每一个观测值 x_i 和一个百分数 f_i 配对，表示大约 $f_i \times 100\%$ 的数据小于 x_i



2. 分位数-分位数图（Q-Q图），使得用户可以观察一个分布到另一个分布是否有漂移

假设我们有两个观测集，取自两个不同的部门，它们是 $x_1, x_2, \ldots x_N$ 和 $y_1, \ldots, y_M$，我们简单地对着 x_i 画 y_i，其中 x_i 和 y_i d欧式对应数据集的第 $(i-0.5)/N$ 个分位数

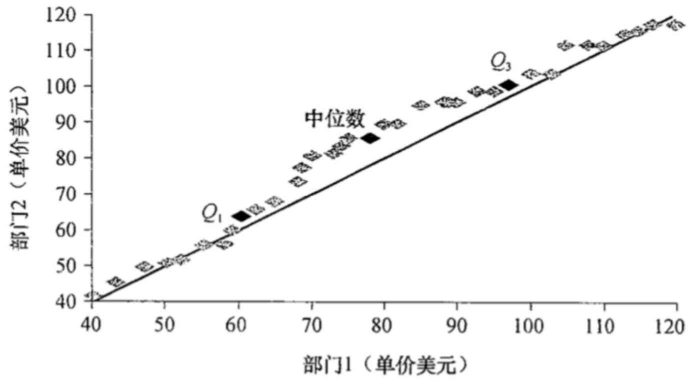
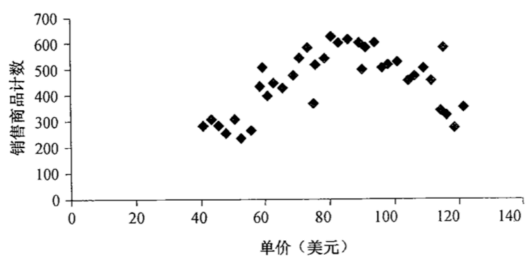


图 2.5 两个不同部门的单价数据的分位数 – 分位数图

3. 直方图

4. 散点图，可用于猜测相关性



2、度量数据的相似性和相异性

2.1 数据矩阵与相异性矩阵

- 数据矩阵，也叫对象-属性结构，存放 n 个对象，每个对象 p 个属性

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- 相异性矩阵，也叫对象-对象局结构，存放 n 个对象两两之间的邻近度。$d(i,j)$ 表示两个对象之间相异性的度量

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

2.2 标称属性的邻近性度量

在如下公式中，$sim(i,j) = 1 - d(i,j)$

1. 匹配率：m 是匹配的数目，p 是对象的属性总数

$$sim(i,j) = \frac{m}{p}$$

2.3 二元属性的邻近性度量

表 2.3　二元属性的列联表

		对象 j		
		1	0	sum
对象 i	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

- 对称的二元相异性

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

- 非对称的二元相异性

$$d(i,j) = \frac{r+s}{q+r+s}$$

- Jaccard 系数

$$sim(i,j) = 1 - d(i,j) = \frac{q}{q+r+s}$$

2.4 闵可夫斯基距离

$$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \ldots + |x_{ip} - x_{jp}|^h}$$

- $i = (x_{i1}, \ldots, x_{ip})$ 和 $j = (x_{j1}, \ldots, x_{jp})$ 是两个有 p 个属性的对象
- $h = 1$：曼哈顿距离
- $h = 2$：欧几里得距离
- $h \to \infty$：上确界距离，$d(i,j) = \max|x_{if} - x_{jf}|$

2.5 序数属性的邻近性度量

略

2.6 混合类型的邻近性度量

略

2.7 余弦相似度

$$sim(x,y) = \frac{x \cdot y}{||x|| \times ||y||}$$

- x, y： 向量
- $||x|| = \sqrt{x_1^2 + \ldots + x_p^2}$

Data Preprocessing

主要任务：数据清理、数据集成、数据归约、数据变换

1、噪声数据

分箱：把数据有序地分配到一些桶中，在桶中进行局部光滑

按price（美元）排序后的数据：4, 8, 15, 21, 21, 24, 25, 28, 34

划分为（等频的）箱：

箱1：4, 8, 15
箱2：21. 21, 24
箱3：25, 28, 34

用箱均值光滑：

箱1：9, 9, 9
箱2：22, 22, 22
箱3：29, 29, 29

用箱边界光滑：

箱1：4, 4, 15
箱2：21, 21, 24
箱3：25, 25, 34

图 3.2 数据光滑的分箱方法

> 老师突然在群里说不能带自己整理的笔记。懒得继续整理了