数据科学基础 I (Matlab)

东北大学

数据探索性分析

东北大学

探索性数据分析

Exploratory Data Analysis（EDA）是指对已有的数据在尽量少的先验假定下进行探索，并通过作图、制表、方程拟合、计算特征量等手段探索数据的结构和规律的一种数据分析方法。

Reference：Hoaglin,D.C. 1982. Exploratory data analysis. In encyclopedia of statistical sciences, Volume 2

抗性分析

- 集中趋势（众数、中位数、四分位数、和值、均值）
- 离中趋势（极差、标准差、方差、极小、极大）
- 分布分析（偏态、峰态）
- 频度分析（分组、频次）

集中趋势与离中趋势

类别	指标	含义	Matlab函数
集中趋势	众数(Mode)	一组数据中出现最多的变量值	mode
	中位数(Median)	一组数据排序后处于中间位置的变量值	median
	四分位数(Quartile)	一组数据排序后处于25%和75%位置上的值	quartile
	和值(Sum)	一组数据相加后的值	sum
	均值(Mean)	一组数据相加后除以数据的个数的值	mean
离中趋势	极大值(Maximum)	某变量所有取值的最大值	max
	极小值(Minimum)	某变量所有取值的最小值	min
	极差(Range)	极大值与极小值之差	---
	标准差(Std Dev)	数据相对于均值的离散程度	std
	方差(Variance)	标准差的平方	var

分布分析与频度分析

类别	指标	含义	Matlab函数
分布分析	偏态(Skewness)	描述数据分布的对称性	skewness
	峰态(Kurtosis)	描述数据分布的平峰或尖峰程度	kurtosis
频度分析	组距	对离散数据进行分组时每一组的范围	---
	频数	每一组内数据的出现次数	---
频率分析	周期	时序数据重复的时间间隔	---
	频率	单位时间内时序信号重复的次数	---

频率分析（周期性分析）

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。

- 年度周期性趋势
- 季节性周期趋势
- 月度周期性趋势
- 周度周期性趋势
- 天周期性趋势
- 小时周期性趋势
- ……

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

- 规模的大小
- 水平的高低
- 速度的快慢
- ……

关键：选择合适的对比标准

对比分析

- 绝对数对比
- 相对数对比
  - 结构相对数
  - 比例相对数
  - 比较相对数
  - 强度相对数
  - 完成率相对数
  - 动态相对数

重新表达

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

简单函数变换就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

数据标准化（归一化）

- 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

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

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

其中max为样本数据的最大值， min为样本数据的最小值。

数据变换——规范化

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。

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

其中 $\bar{x}$为原始数据的均值，σ 为原始数据的标准差。

数据变换——规范化

- 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。

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

模式发现

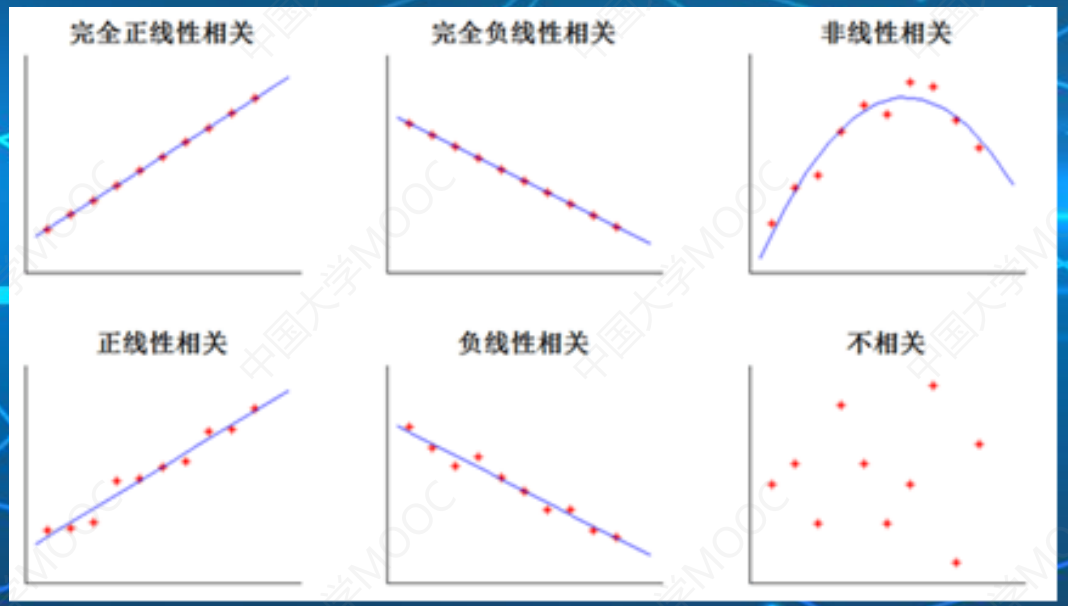
- 相关性是探索数据模式的基本方法
- 降维可以适当降低计算的复杂度
- 数据巡查是从多个角度对数据进行观察，以进一步发现潜在模式
- 一些聚类方法可以将数据以相似性进行分组

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。

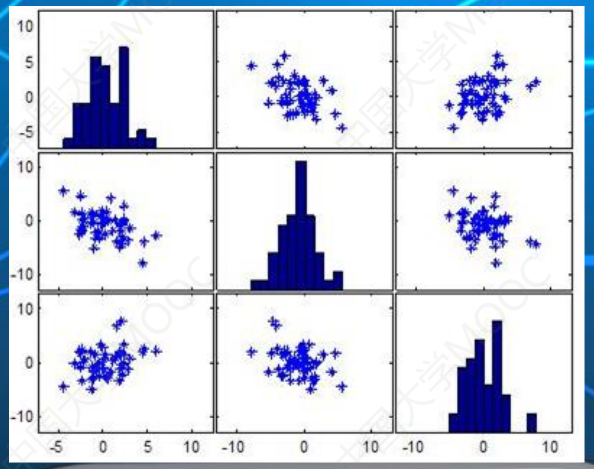
- 直接绘制散点图
- 绘制散点图矩阵
- 计算相关系数

相关性分析——直接绘制散点图



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，利用散点图矩阵来同时绘制各自变量间的散点图，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

通过计算相关系数来进行相关分析，可以更加准确的描述变量之间的线性相关程度。在二元变量的相关分析过程中比较常用的有：

- Pearson相关系数
- Spearman秩相关系数