大数据导论
Introduction to Big Data

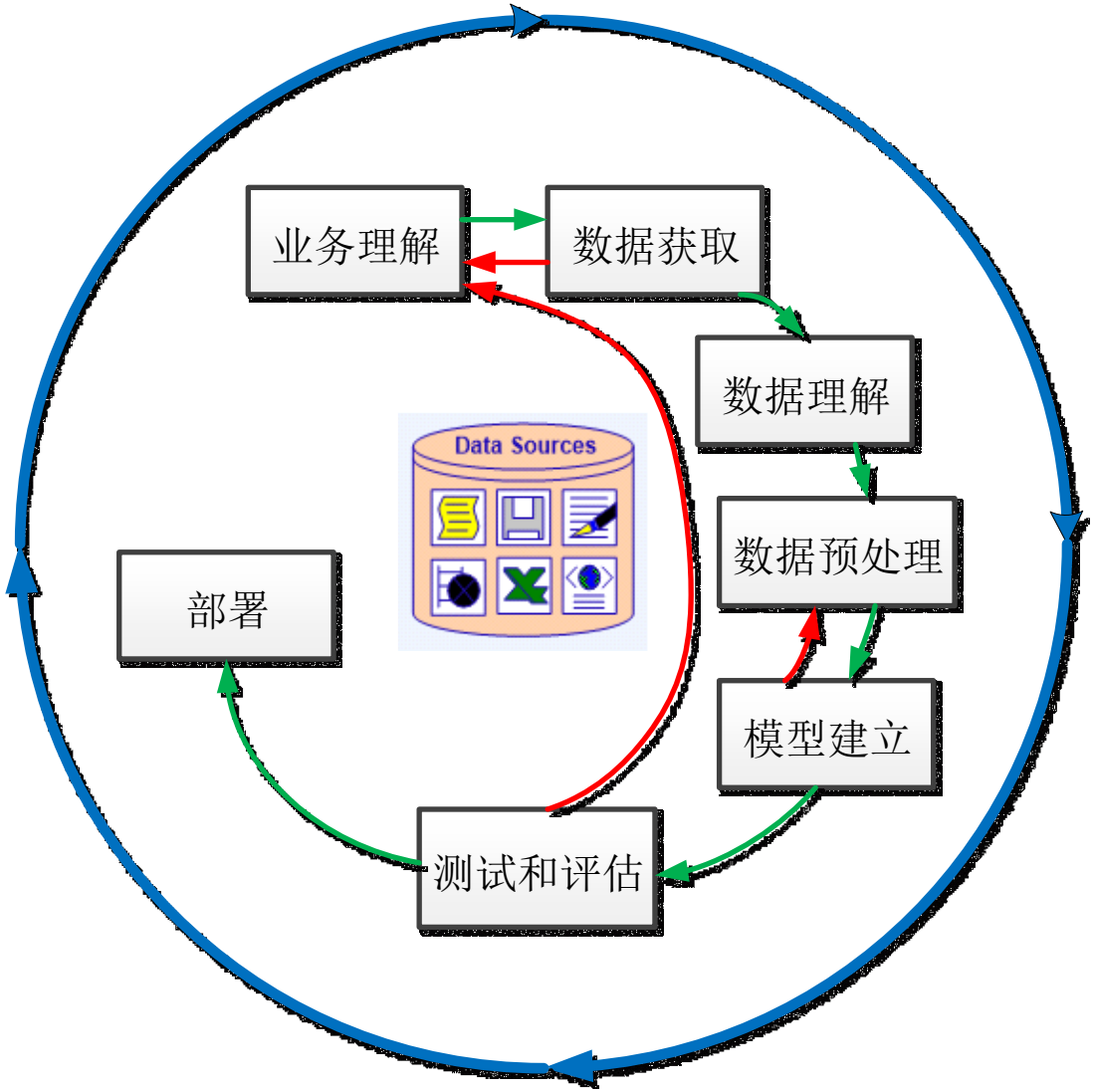
第 4.2 讲: 理解数据

叶允明

计算机科学与技术学院

哈尔滨工业大学（深圳）

大数据挖掘过程



目录

- 数据理解的主要任务
- 基于统计描述的数据理解方法
- 数据可视化的传统方法
- 高维数据的低维嵌入可视化方法

数据理解的主要任务

数据理解

- 在统计学中称为exploratory data analysis（EDA）
  - 借助人类对数据的观察和“模式”识别能力
  - 对数据进行初步研究，以便更好地理解它的特性
  - 有助于选择合适的数据预处理和数据分析、挖掘技术

任务一：每个属性取值分布统计

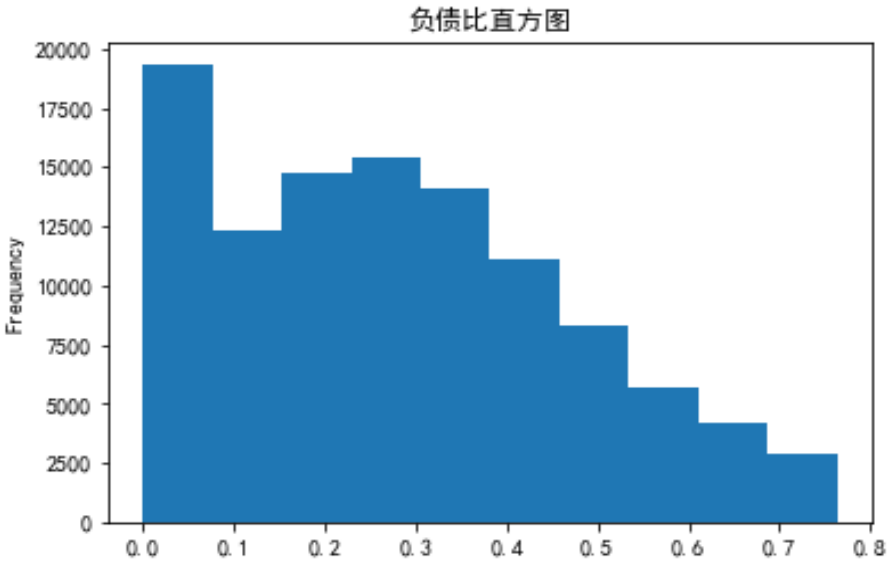
- 对给定数据集的各个属性的取值分布情况进行统计概括
- 例：结构化的信贷数据集

数值型

类别型

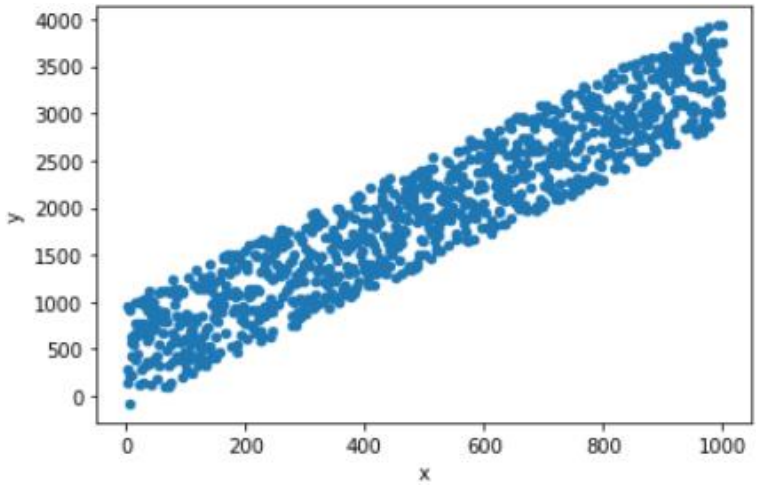
信贷数据集

id	年龄	逾期30-59天次数	负债比	收入	未清贷款	过去90天逾期次数	家属数量	年龄域
1	45	2	0.802982	9120	13	0	2	中年
2	40	0	0.121876	2600	4	0	1	青年
3	38	1	0.085113	3042	2	1	0	青年
4	30	0	0.03605	3300	5	0	0	青年
5	49	1	0.024926	63588	7	0	0	中年
6	74	0	0.375607	3500	3	0	1	老年
7	57	0	5710	1522	8	0	0	中年
8	39	0	0.20994	3500	8	0	0	青年
9	27	0	46	2525	2	0	0	青年
10	57	0	0.606291	23684	9	0	2	中年
11	30	0	0.309476	2500	5	0	0	青年
12	51	0	0.531529	6501	7	0	2	中年
13	46	0	0.298354	12454	13	0	2	中年
14	40	3	0.382965	13700	9	3	2	青年
15	76	0	477	0	6	0	0	老年
16	64	0	0.209892	11362	7	0	2	中年
17	78	0	2058	0	10	0	0	老年
18	53	0	0.188274	8800	7	0	0	中年
19	43	0	0.527888	3280	7	0	2	青年
20	25	0	0.065868	333	2	0	0	青年

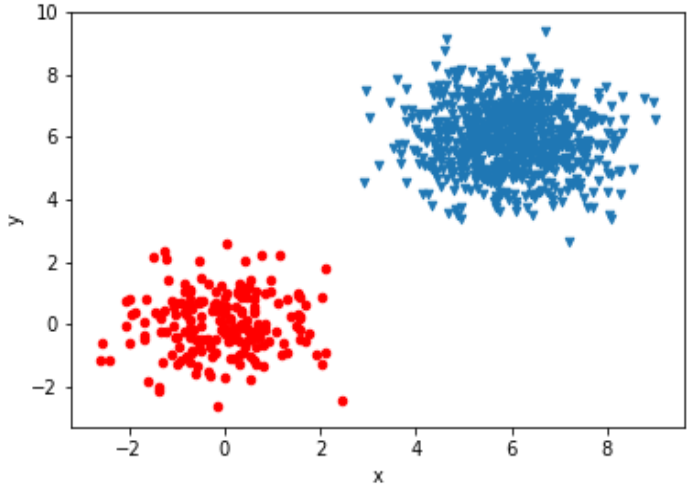


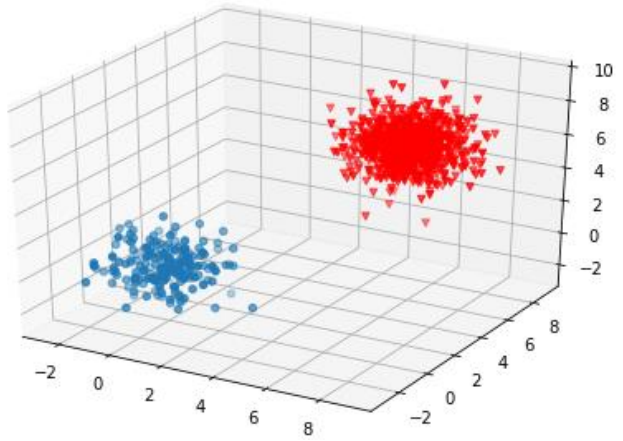
任务二：多个属性取值分布统计

- 对给定数据集的多个属性的取值分布情况进行统计概括
- 例如，利用散点图查看属性之间具有的相关性，数据的分布情况



属性间具有相关性



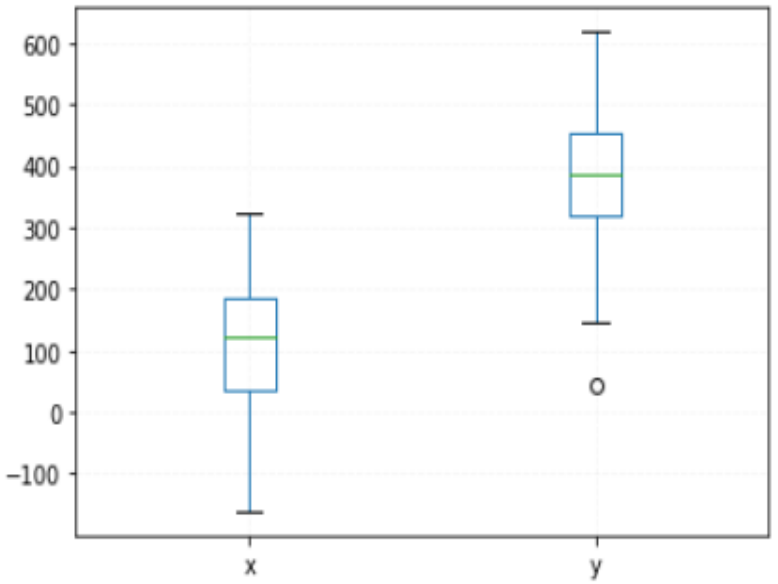


数据的分布情况

任务三：数据的总体质量评估

- 数据存在误差属性值缺失噪声和不一致性等潜在的数据质量问题
- 例：原始的信贷数据集

信贷数据集								
id	年龄	逾期30-59天次数	负债比	收入	未清贷款	过去90天逾期次数	家属数量	年龄域
1	45	2	0.802982	9120	13	0	2	中年
2	40	0	0.121876	2600	4	0	1	青年
3	38	1	0.085113	3042	2	1	0	青年
4	30	0	0.03605	3300	5	0	0	青年
5	49	1	0.024926	63588	7	0	0	中年
6	74	0	0.375607	3500	3	0	1	老年
7	57	0	5710	1522	8	0	0	中年
8	39	0	0.20994	3500	8	0	0	青年
9	27	0	46	2525	2	0	0	青年
10	57	0	0.606291	23684	9	0	2	中年
11	30	0	0.309476	2500	5	0	0	青年
12	51	0	0.531529	6501	7	0	2	中年
13	46	0	0.298354	12454	13	0	2	中年
14	40	3	0.382965	13700	9	3	2	青年
15	76	0	477	NA	6	0	0	老年
16	64	0	0.209892	11362	7	0	2	中年
17	78	0	2058	NA	10	0	0	老年
18	53	0	0.188274	8800	7	0	0	中年
19	43	0	0.527888	3280	7	0	2	青年
20	25	0	0.065868	333	2	0	0	青年



理解数据的主要方法

- 汇总统计方法（Summary statistics），或称为概况性统计描述
- 传统可视化方法（Visualization），
  - 包括OLAP（Online Analytical Processing）
- 基于数据挖掘结果的可视化方法
  - 通过聚类、离群点检测的结果分析数据的特点
  - 高维数据的低维嵌入（low-dimensional embedding）可视化

基于统计描述的数据理解方法

集中趋势（central tendency）度量

- 数值型数据
  1. 均值
  2. 分位数
  3. 几何均值
- 类别型数据
  1. 众数

数值型数据的集中趋势度量方法

- 均值

1. 简单平均数

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_m}{m} = \frac{\sum x_i}{m}$$

2. 截尾均值

$$\bar{x}_\alpha = \frac{x_{[m\alpha+1]} + x_{[m\alpha+2]} + \cdots + x_{[m-m\alpha]}}{m - 2m\alpha}$$

	裁判1	裁判2	裁判3	裁判4	裁判5	总得分
简单平均数	83	95	96	98	100	94.4
截尾均值	83	95	96	98	100	96.3

数值型数据的集中趋势度量方法

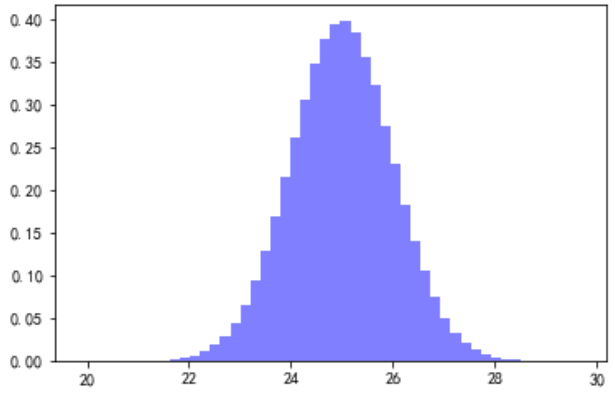
- 中位数 $x_{\left[\frac{m+1}{2}\right]}$
- 四分位数 $x_{\left[\frac{m+1}{4}\right]}$ $x_{\left[\frac{3(m+1)}{4}\right]}$
- 几何平均数 $G = \sqrt[m]{x_1 \times x_2 \times \cdots \times x_m} = \sqrt[m]{\prod x_i}$

例：收益率计算. 假定某基金产品近三年的年收益率（按复利计算）：第一年6%（比率1.06），第二年5%，第三年9%。则该基金产品的平均年收益率：

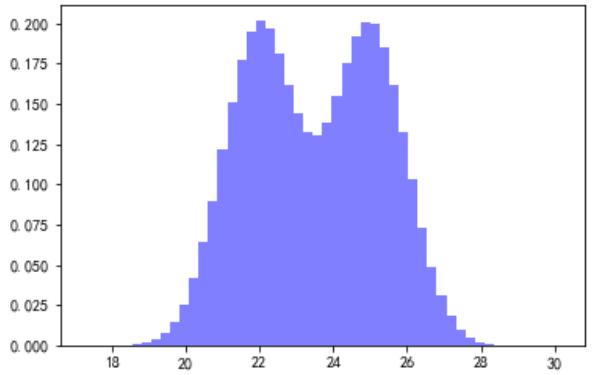
$$\sqrt[3]{1.06 \times 1.05 \times 1.09} - 1 = 6.653\%。$$

类别型数据的集中趋势度量方法

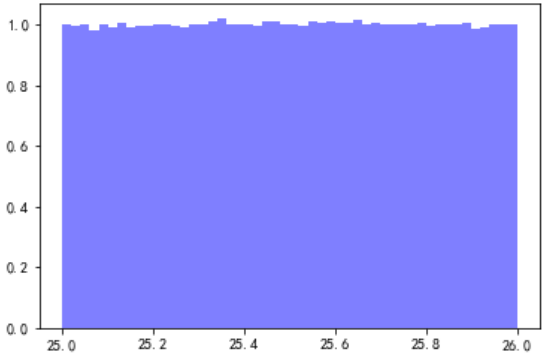
- 众数(mode)：一组数据中出现频率最高的属性值



单个众数

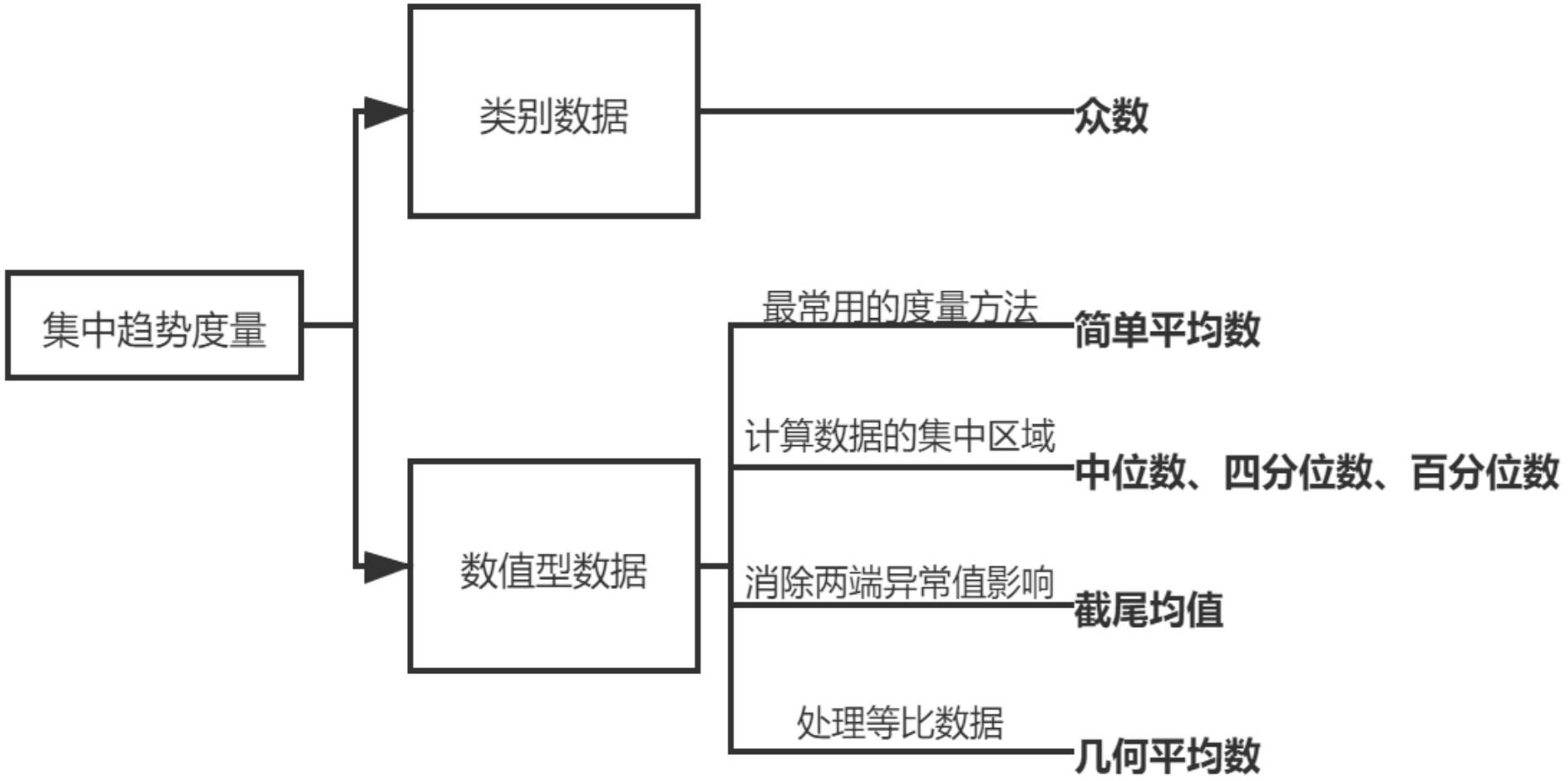


多个众数



没有众数

集中趋势度量方法比较


集中趋势度量
类别数据
众数
数值型数据
最常用的度量方法
简单平均数
计算数据的集中区域
中位数、四分位数、百分位数
消除两端异常值影响
截尾均值
处理等比数据
几何平均数

散布（spread）程度度量

- 数值型数据
  1. 极差
  2. 四分位距
  3. 方差、标准差
- 类别数据
  1. 异众比率

数值型数据的散布程度度量方法

- 极差

$$R = X_{max} - X_{min}$$

- 四分位距

$$\boldsymbol{x}_{\left[\frac{3(\boldsymbol{m}+1)}{4}\right]} - \boldsymbol{x}_{\left[\frac{\boldsymbol{m}+1}{4}\right]}$$

- 方差

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_m - \bar{x})^2}{m-1} = \frac{\sum(x_i - \bar{x})^2}{m-1}$$

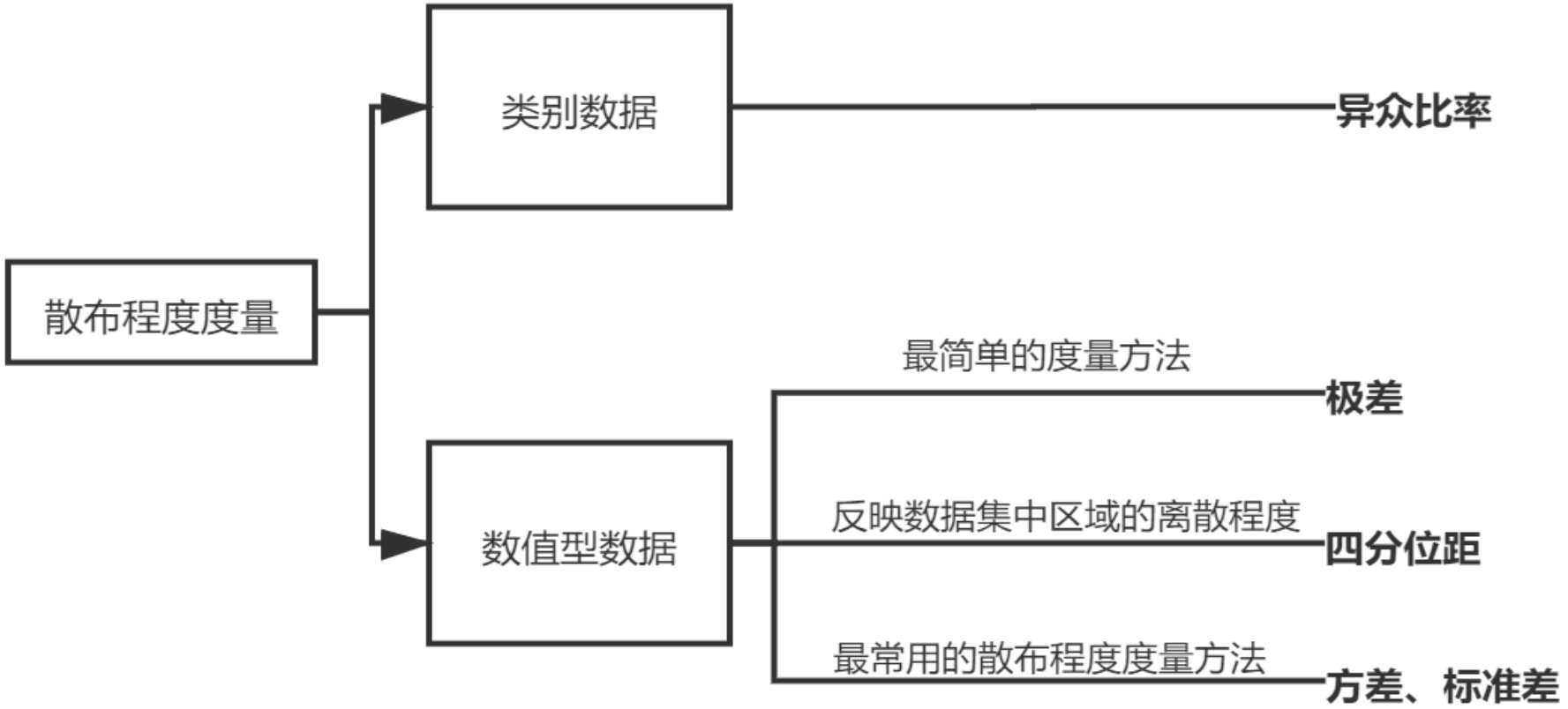
类别型数据的散布程度度量方法

- 异众比率

$$\rho = \frac{m - f}{m}$$

  - 其中，ρ表示异众比率，m表示数据集的对象总数，f表示数据集中取值为众数的数据对象总数。
  - 显然，异众比率越大，则数据越分散，众数对数据集的代表性越差

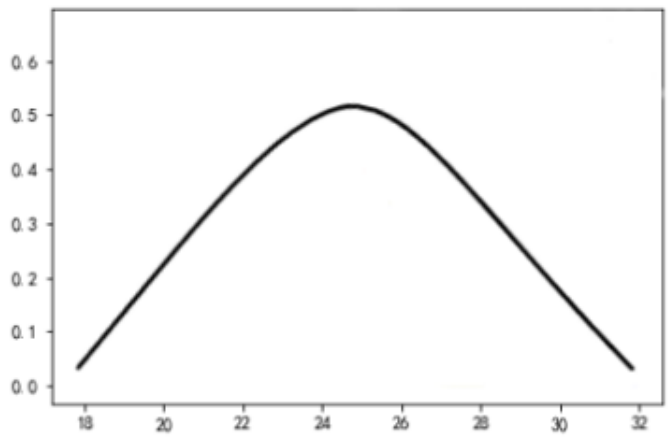
散布程度度量方法比较


散布程度度量
类别数据
异众比率
数值型数据
最简单的度量方法
极差
反映数据集中区域的离散程度
四分位距
最常用的散布程度度量方法
方差、标准差

偏态度量

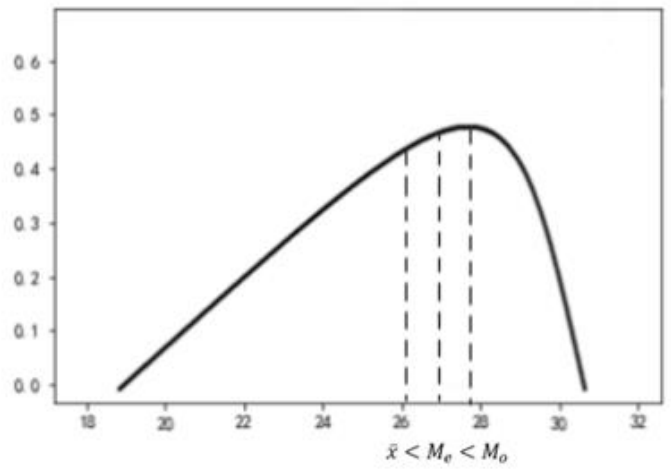
- 偏态（skewness）是描述数值型数据分布的对称性的度量指标
- 数据分布的对称性可以分为对称、左偏或右偏。

$$S_k = \frac{m\sum(x_i - \bar{x})^3}{(m-1)(m-2)s^3}$$



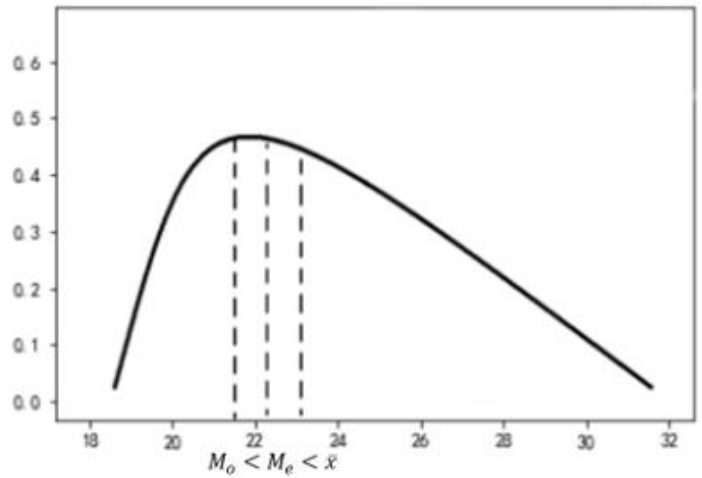
对称分布

Sk小于0



左偏分布

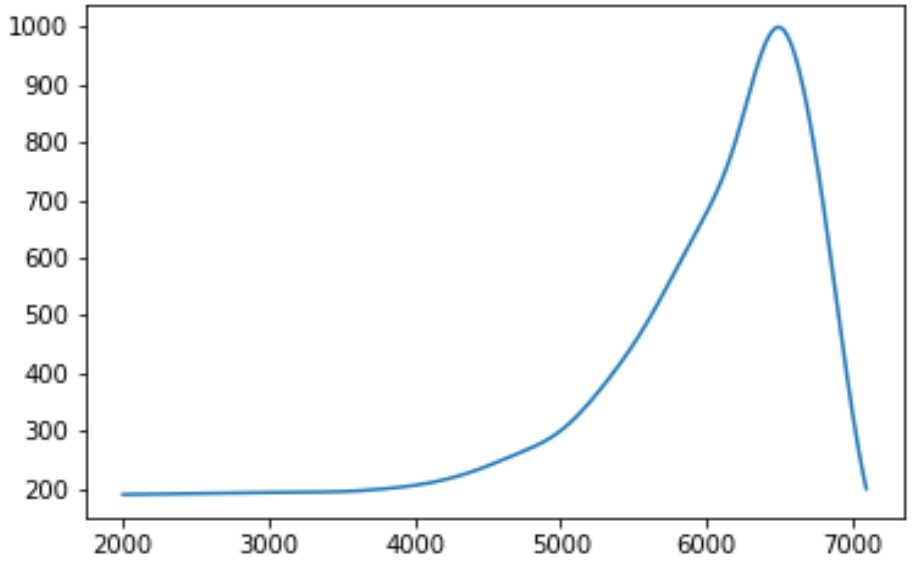
Sk大于0



右偏分布

偏态度量

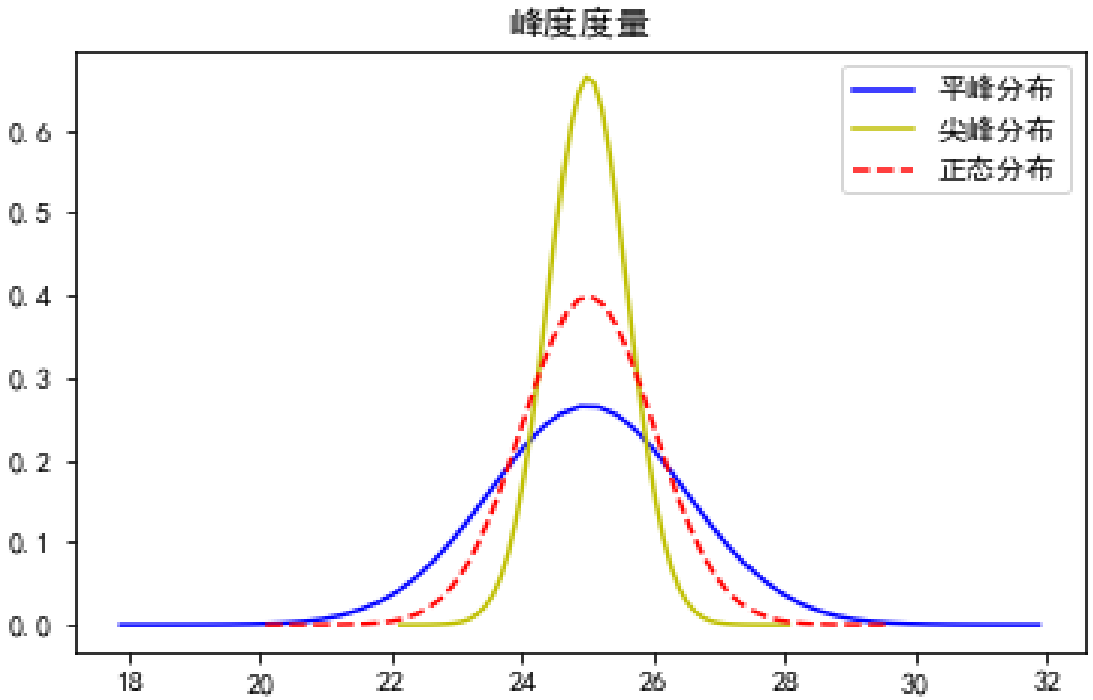
- 在数据分布的统计描述中，偏态度量是集中趋势度量和散布度量的有益补充
- 例：某地收入统计



峰度度量

- 数据分布的峰度是描述数值型数据分布的几何形态陡峭程度的统计量
  - 以正态分布为基准定义的

$$Kurt = \frac{m(m+1)\sum(x_i - \bar{x})^4 - 3[\sum(x_i - \bar{x})^2]^2(m-1)}{(m-1)(m-2)(m-3)s^4}$$



数据可视化的传统方法

数据可视化

- 可视化是将数据转换为可视的形式，以图形或表格的形式显示信息，以便能够借此分析数据的特征，以及数据项或属性之间的关系。
- 数据可视化是理解数据的重要工具
  - 人类具有分析大量视觉信息的能力
  - 可以检测普遍模式和趋势
  - 可以检测异常值和异常模式

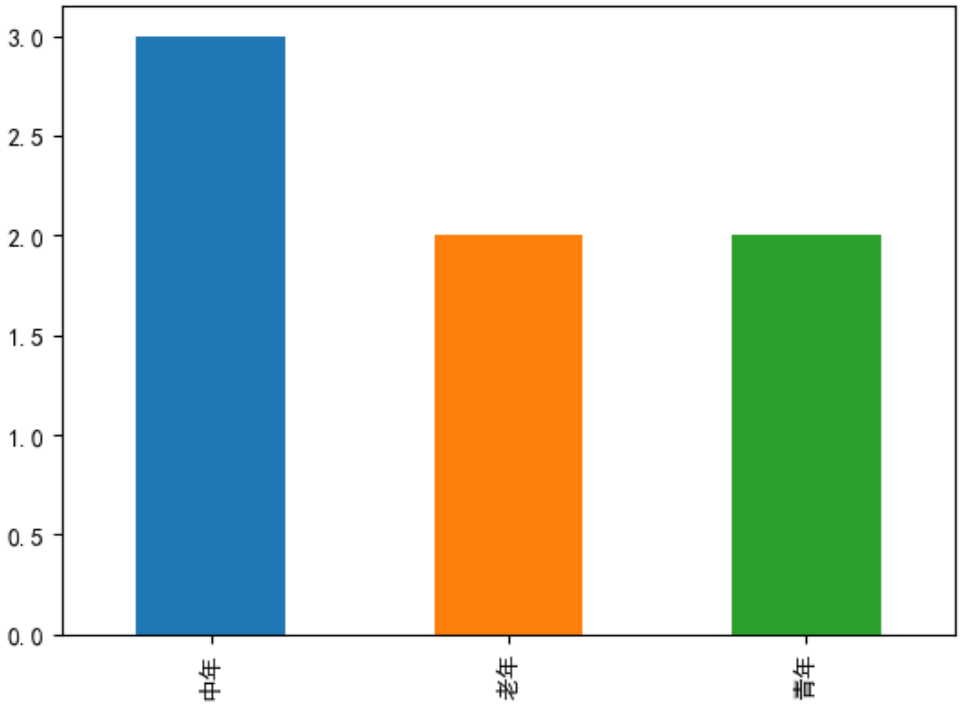
鸢尾花数据集

- 使用鸢尾花（Iris）数据集说明数据探索技术
  - 该数据集可从UCI的机器学习库中得到 http://www.ics.uci.edu/~mlearn/MLRepository.html
  - 来自统计学家Douglas Fisher
  - 三种花（类）：
    - Setosa (Se)
    - Virginica (Vi)
    - Versicolour (Ve)
  - 四个（非类）属性:
    - 萼片宽度和长度
    - 花瓣宽度和长度

鸢尾花Virginica的图片. Robert H. Mohlenbrock. USDA NRCS. 1995. 东北湿地植物志：野外办公室植物物种指南。东北国家技术中心，切斯特，宾夕法尼亚州

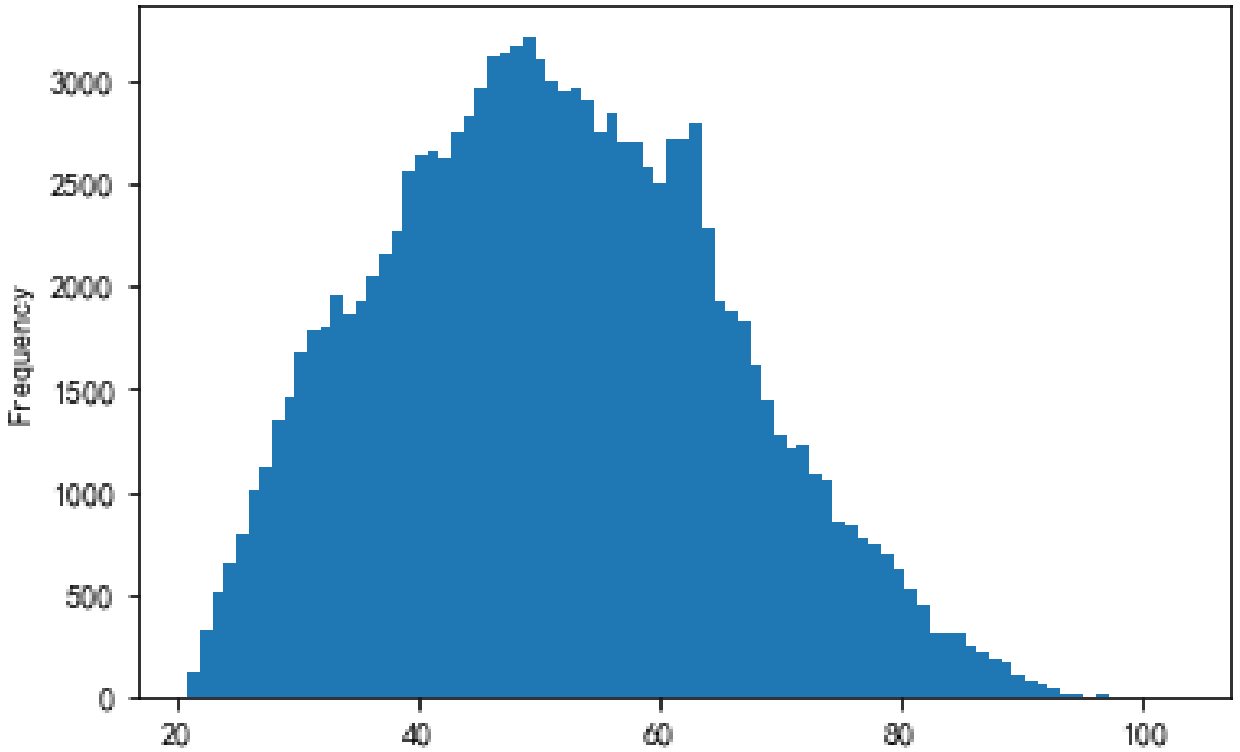
柱状图

- 用于类别型数据：图中每一个类别的计数统计以“条柱”的形式展示，其中数据的类别的频次比较以“条柱”的高度作为评判标准。



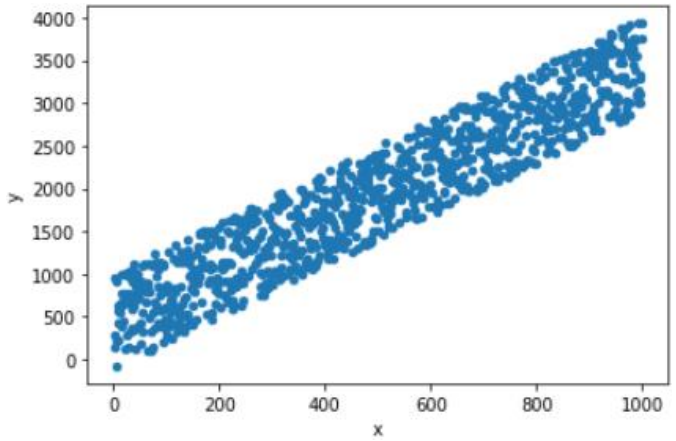
直方图

- 直方图用于数值型数据：数值属性值频次的统计可视化
  - 其横坐标为数值的取值跨度，纵坐标为数值的频率或频次

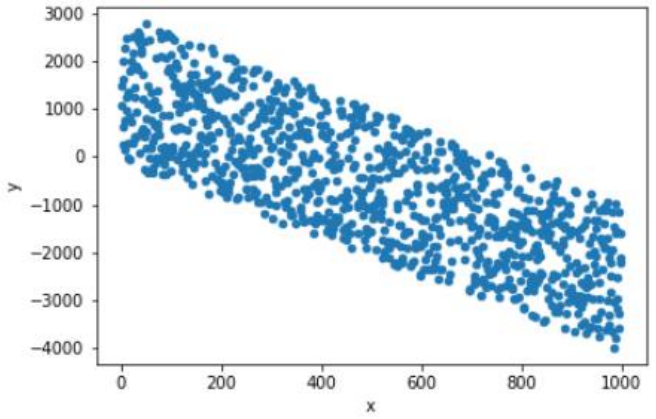


散点图

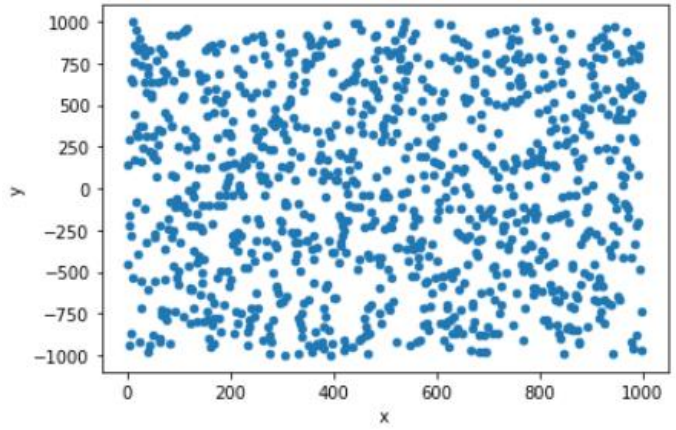
- 散点图主要应用于数值型数据，其作用主要是观察两个（或三个）数值型属性的数据之间的关联性



正相关



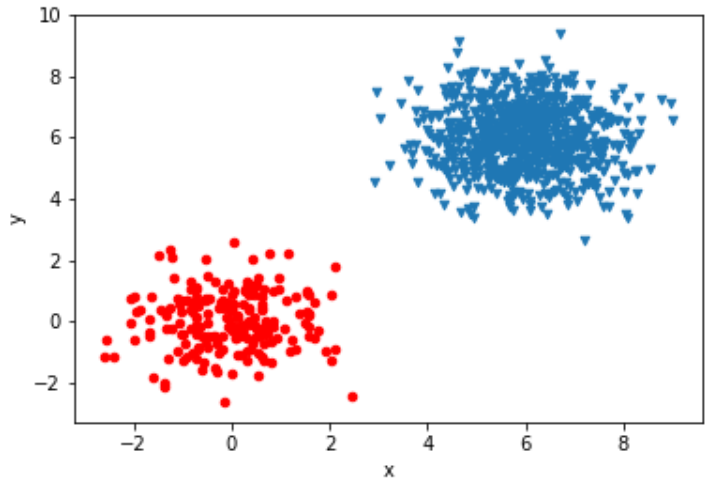
负相关



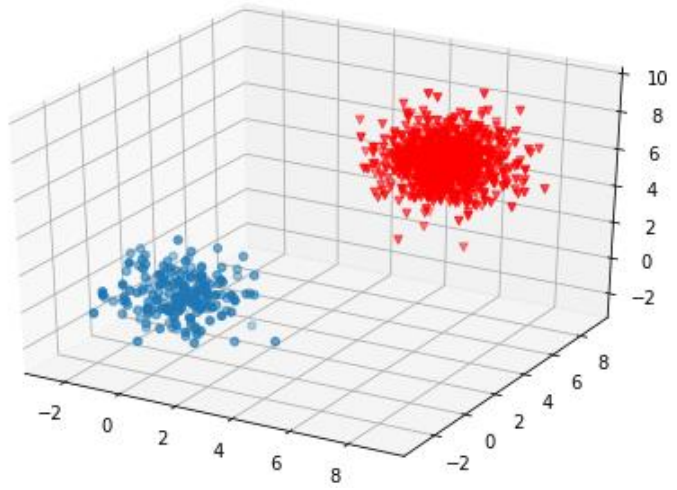
两个属性间无明显关系

散点图（二）

- 散点图可以直观的反映集中数据和离群数据的分布，还能直接反映数据集内存在的分组情况，还可反映属性组合的数据区分能力

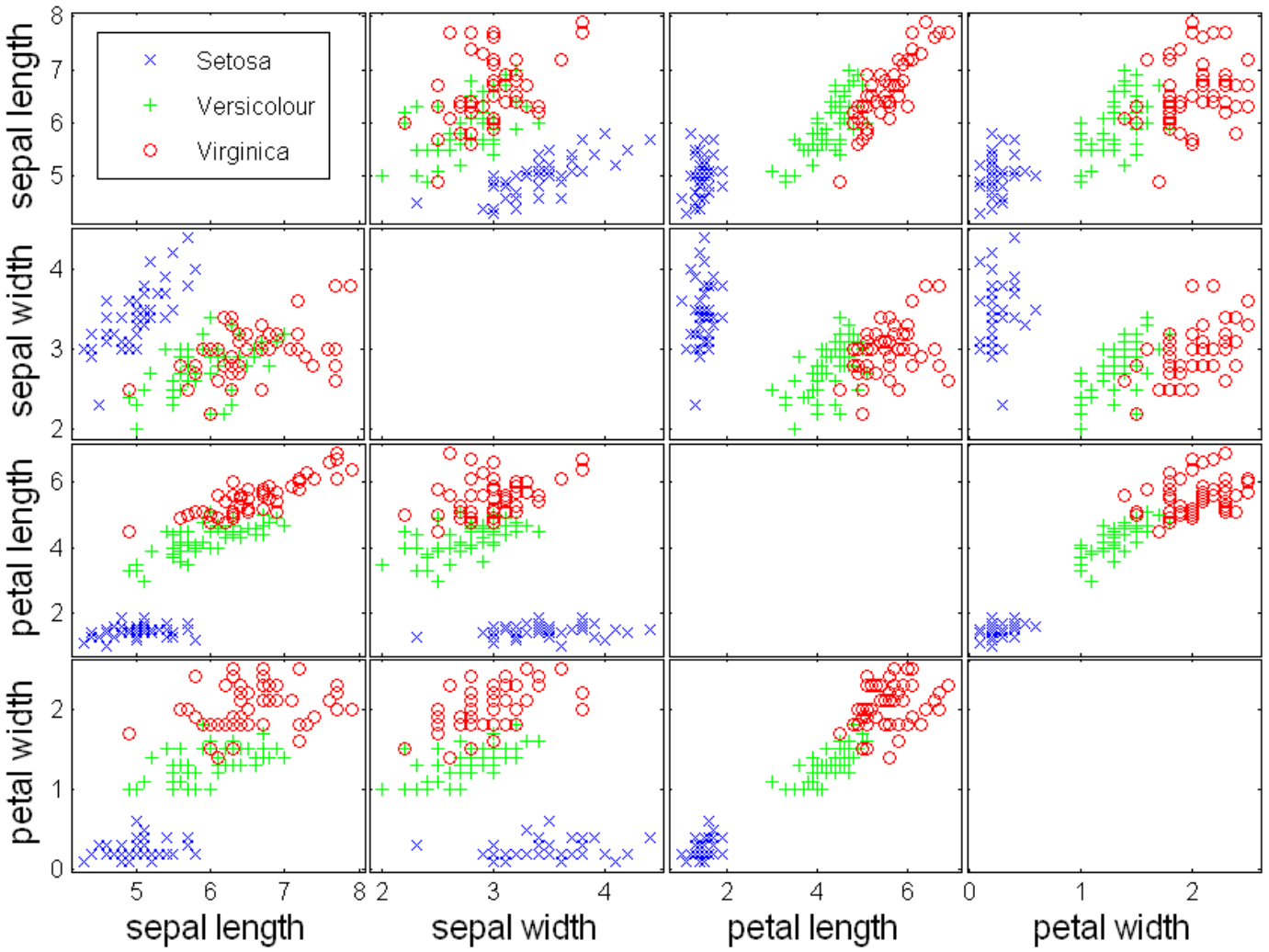


二维散点图



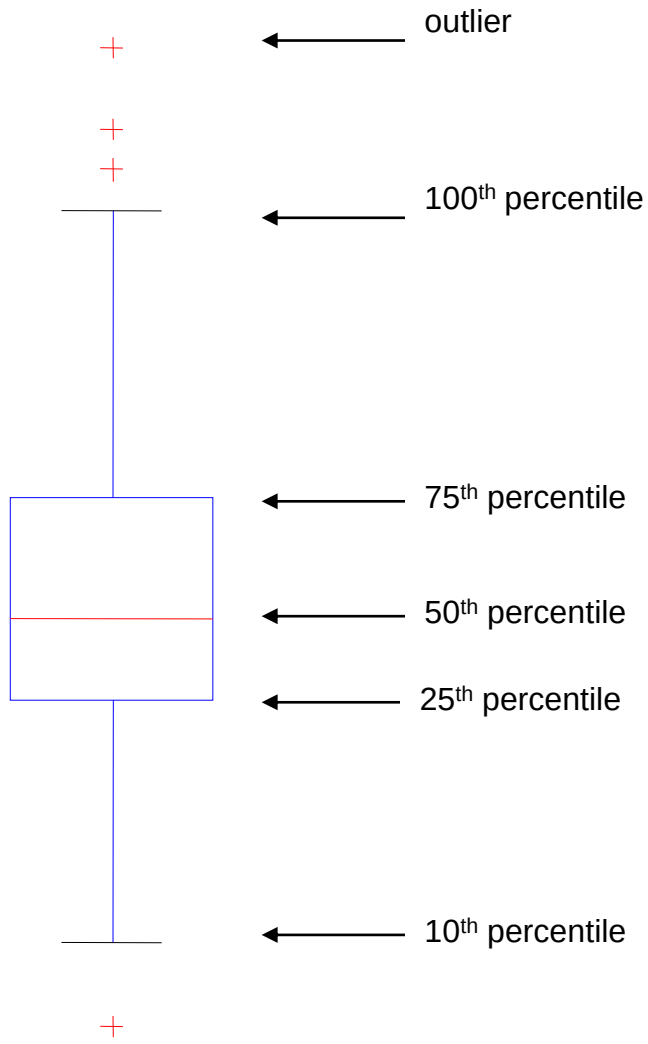
三维散点图

Iris数据集的散点图矩阵



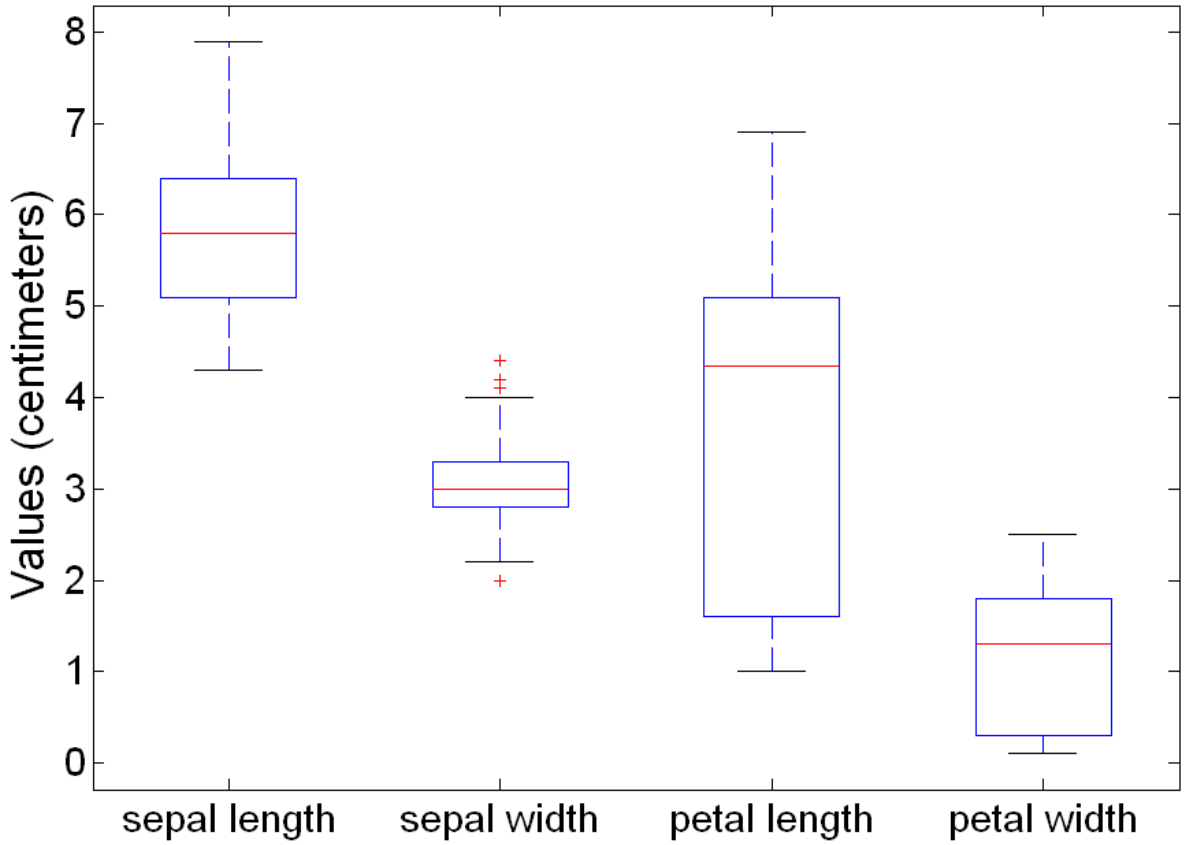
箱线图

- 数值型属性，五数归纳（Five Number Summary）
- Minimum, Q1(25%), Median, Q3(75%), Maximum
- 箱线图(Box Plots)
  - 另一种显示数据分布的方法
  - 箱线图的箱体：四分位距(IQR, Interquartile range), IQR =Q3-Q1
  - 上下边缘（whisker）：
    min(Q3+1.5IQR, Maximum)
    max(Q1-1.5IQR, Minimum)
  - 这样计算上下边缘可区分异常值



箱线图应用示例

- 箱线图可以用来比较属性的类区分能力（discrimination capability）

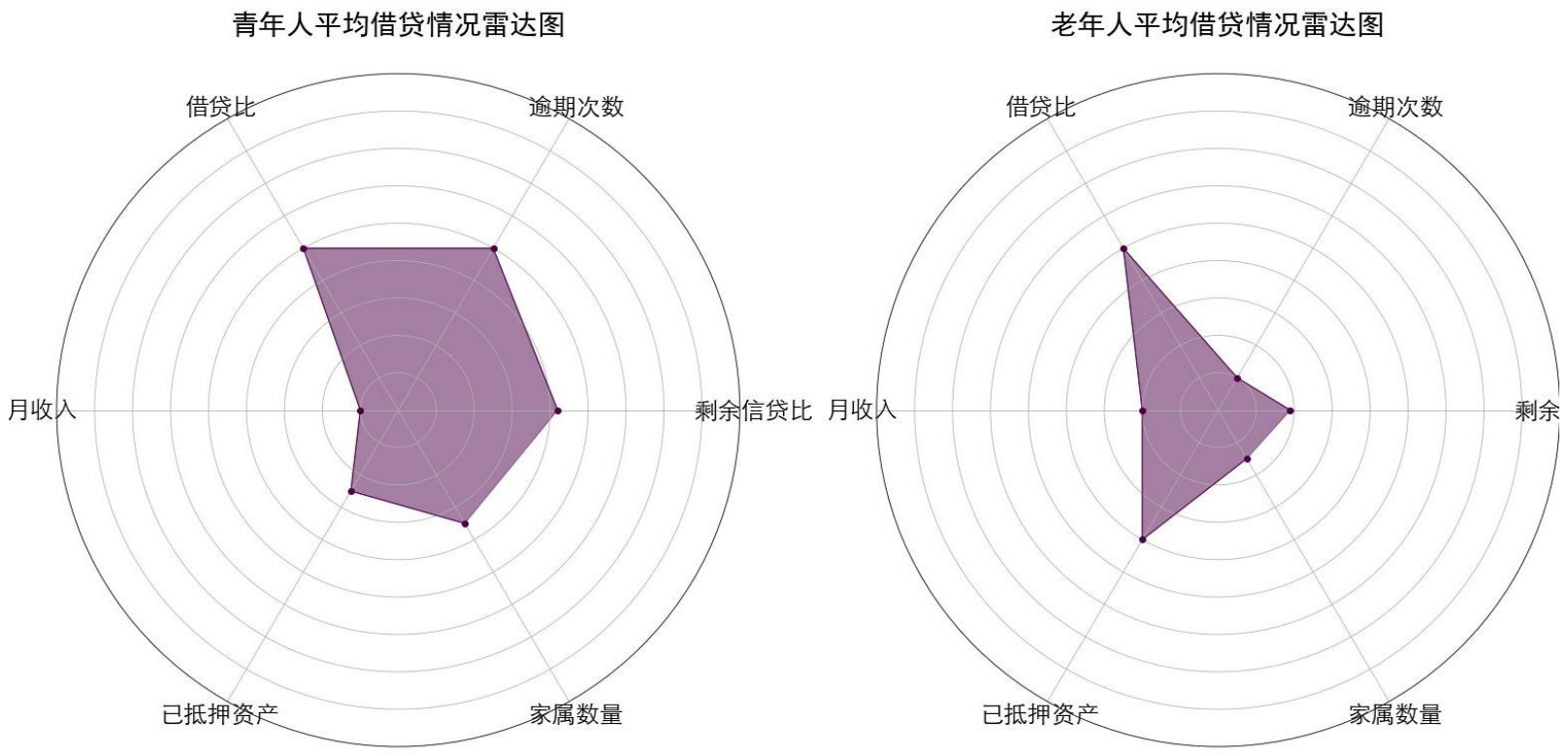


数据矩阵图

- 数据矩阵图，将数值映射到不同的颜色上，从而可以将一张数据规模不大的表进行可视化
- 例：1949-1960年各月的航班飞行次数累积表

雷达图

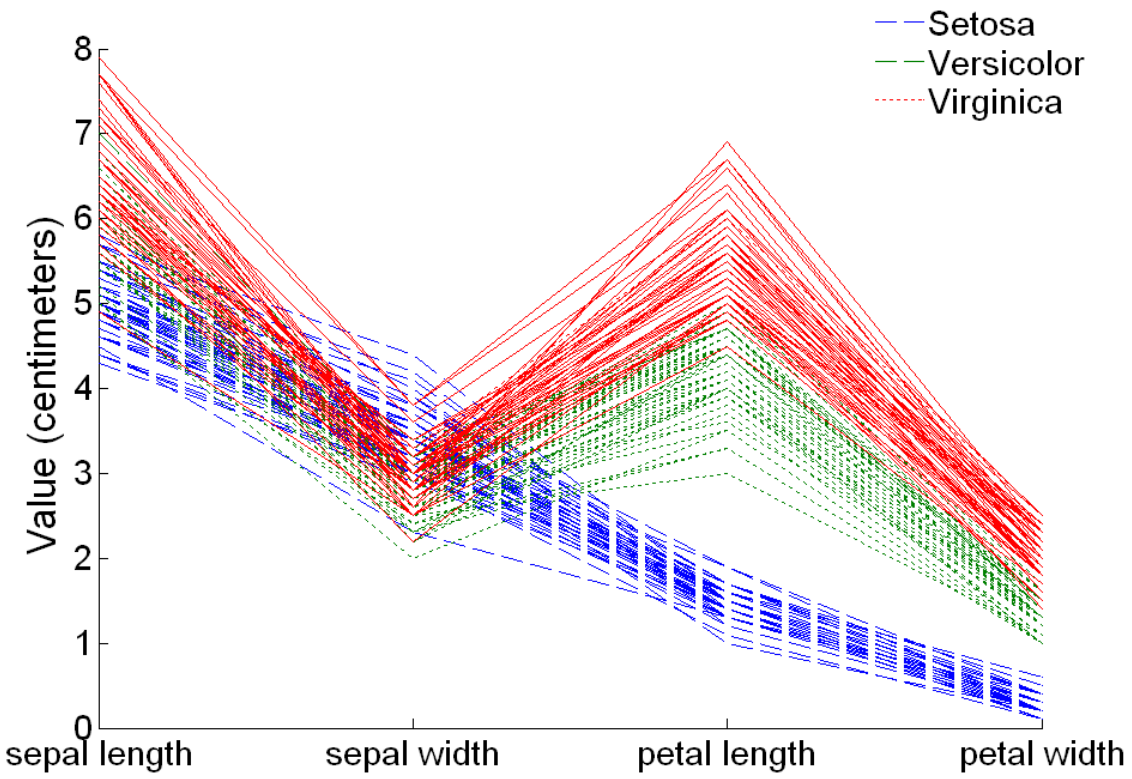
- 可对每一个（类）多维数据对象进行可视化

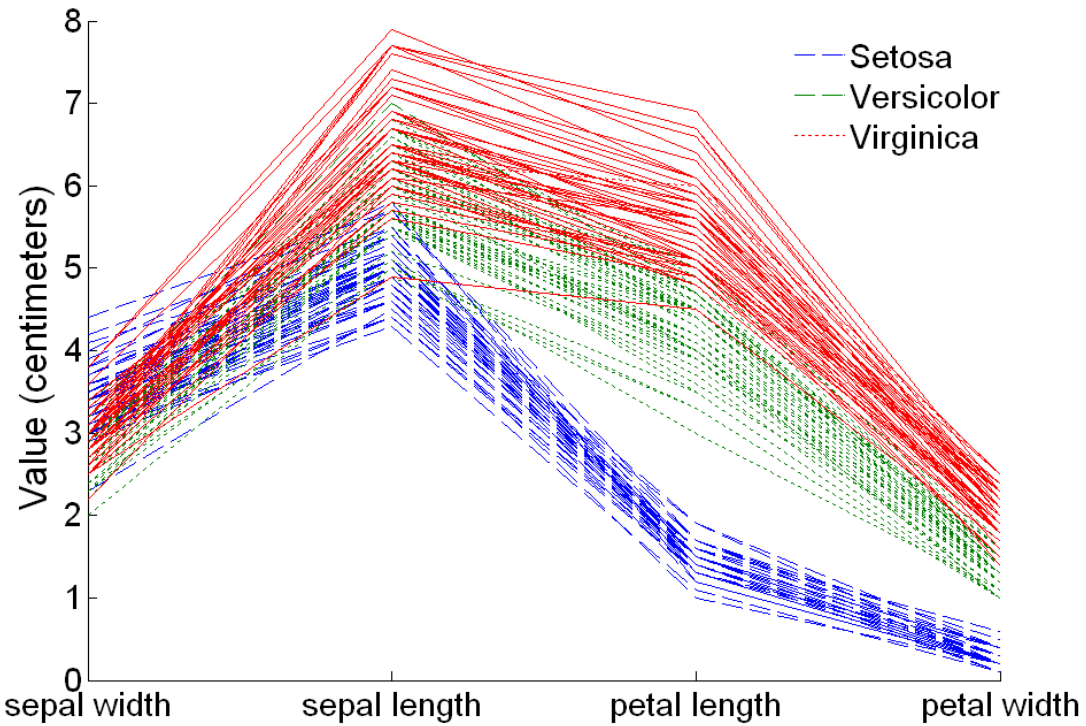


平行坐标图

- 平行坐标系(Parallel Coordinates)
  - 用于绘制高维数据的属性值
  - 使用一组平行坐标轴
  - 对象每个属性的值被绘制为对应坐标轴上的点，将这些点用线连接起来
  - 因此，每个对象表示为一条线
  - 通常，对象趋于分成少数几个组，组内的点具有类似的属性值
  - 在查看此类分组时，对属性进行排序非常重要

平行坐标图示例


Setosa
Versicolor
Virginica
Value (centimeters)
0
1
2
3
4
5
6
7
8
sepal length
sepal width
petal length
petal width


Setosa
Versicolor
Virginica
Value (centimeters)
0
1
2
3
4
5
6
7
8
sepal width
sepal length
petal length
petal width

致谢

- 一小部分图表、文字参考教材、互联网等资料，仅供公益性的学习参考，在此表示感谢！如有版权要求请联系：yym@hit.edu.cn，谢谢！