

2024-2025学年 第1学期(秋)



数据挖掘

第2章 认识数据

2025 年 9 月



南京大學
NANJING UNIVERSITY

目录

01

数据对象和属性

02

数据统计与可视化

03

数据相似性和相异性度量

数据对象

- **数据集**由数据对象组成
- 一个**数据对象**代表一个实体
- 例子
 - 销售数据库：客户，商店物品，销售额
 - 医疗数据库：患者，治疗信息
 - 大学数据库：学生，教授，课程信息
- 称为样本，示例，实例，数据点，对象，元组 (tuple)

Objects

Attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

数据属性

- 数据对象所描述的属性
 - 数据库中的行 - > 数据对象
 - 列 - > “属性”

Objects

Attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

属性的类型

- 常见的四类属性:

- 标称 (Nominal)

- Examples: ID numbers, zip codes

- 序数 (Ordinal)

- Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}

属性的类型

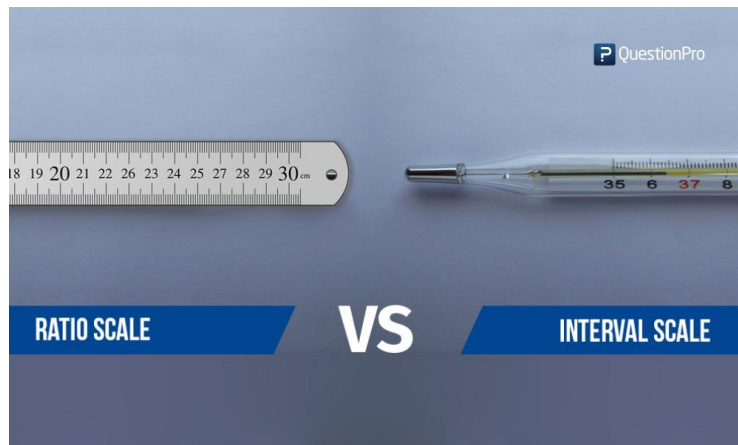
- 常见的四类属性：

- 区间 (Interval)

- Examples: calendar dates, temperatures in Celsius or Fahrenheit.

- 比率 (Ratio)

- Examples: temperature in Kelvin, length, time, counts



属性的类型

- **标称：类别，状态**

- Hair_color={黑色, 棕色, 金色, 白色}
- 婚姻状况, 职业, 身份证号码, 邮政编码

- **二进制**

- 只有2个状态（0和1）的属性
- 对称二进制两种
 - 例如，性别
- 非对称的二进制
 - 例如，医疗测试（正面与负面）

Objects

Attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

属性的类型

- 序数

- 价值观有一个有意义的顺序（排名），但不知道连续值之间的大小。
- 大小={小, 中, 大}, 等级, 成绩排名

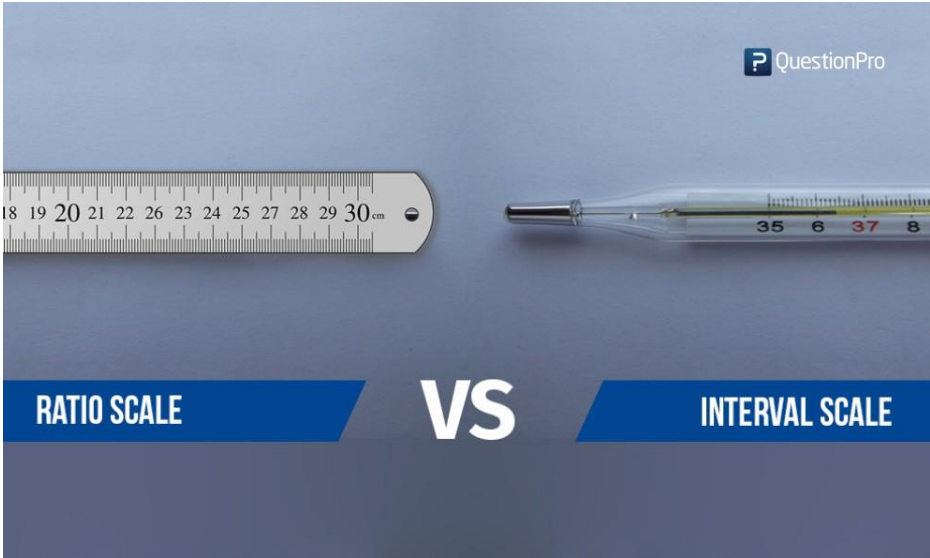
Objects

Attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

属性的类型

- 区间标度属性
 - 以单位长度顺序性度量
 - 值有序，比如温度、日历等
 - 不存在0点，倍数没有意义，比如我们平常通常不说2000年是1000年的2倍



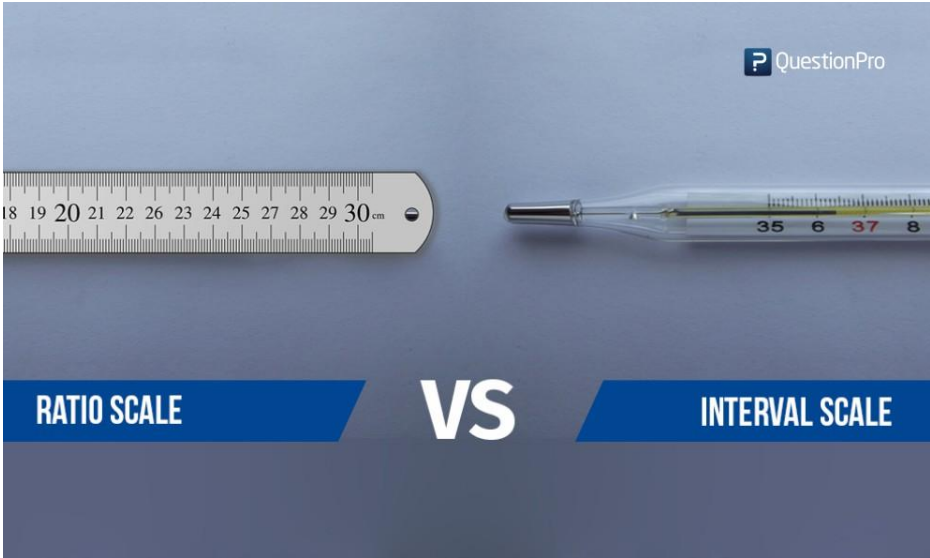
Objects

Attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

属性的类型

- 比率标度属性
 - 具有固定零点的数值属性，有序且可以计算倍数
 - 长度、重量等



Objects

Attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

属性的类型

属性类型	能否分类	能否排序	能否体现 相等间隔	有无绝对 零点
标称	✓	✗	✗	✗
序数	✓	✓	✗	✗
区间	✓	✓	✓	✗
比率	✓	✓	✓	✓

属性的类型

- 常见的四类属性：

- 标称 (Nominal)

- ✓ 标称类型的数据分类模型有哪些？

- 序数 (Ordinal)

- ✓ 序数类型的数据求均值是否合理？能否用K-means算法？

- 区间 (Interval)

- ✓ 区间类型的特征是否可以采用乘法原则？

- 比率 (Ratio)

- ✓ 哪些数据挖掘算法适用于比率标定属性？



南京大學
NANJING UNIVERSITY

目录

01

数据对象和属性

02

数据统计与可视化

03

数据相似性和相异性度量

数据统计

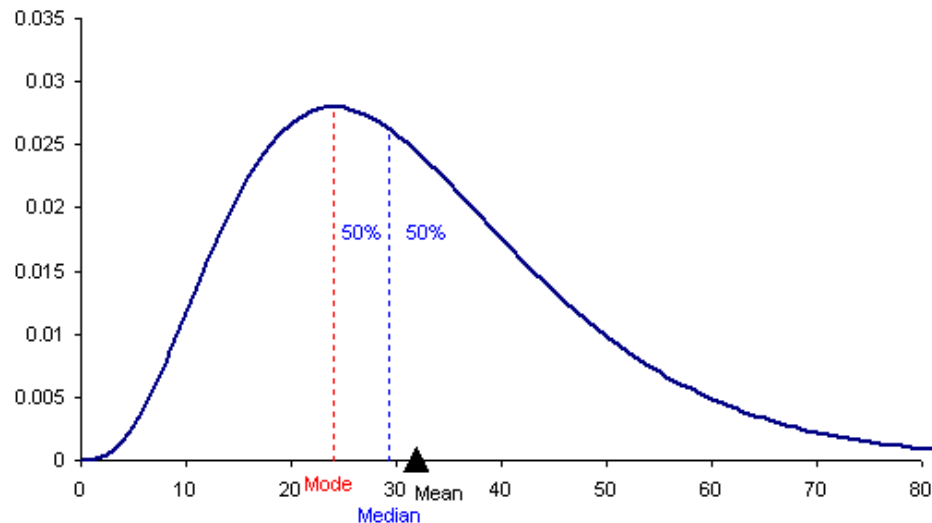
- **动机：为了更好地理解数据：**
集中趋势，分布
- **数据的统计特性**
 - **最大值，最小值，均值，中位数，方差等。**

当前薪金

性别	薪资分组	均值	N	极小值	极大值	合计 N 的 %
女	低收入	17850.00	32	15750	19950	6.8%
	中收入	26046.07	173	20100	38850	36.5%
	高收入	49611.36	11	40800	58125	2.3%
	总计	26031.92	216	15750	58125	45.6%
男	低收入	19650.00	1	19650	19650	.2%
	中收入	29719.94	164	21300	39900	34.6%
	高收入	62346.88	93	40050	135000	19.6%
	总计	41441.78	258	19650	135000	54.4%
总计	低收入	17904.55	33	15750	19950	7.0%
	中收入	27833.95	337	20100	39900	71.1%
	高收入	60999.86	104	40050	135000	21.9%
	总计	34419.57	474	15750	135000	100.0%

中性化趋势度量：均值、中位数和众数

- **平均值**一组数据的均衡点。
 - 均值对离群值很敏感。
- 因此，**中位数和截断均值**也很常用。
- **众数**指一组数据中出现次数最多的数据值。



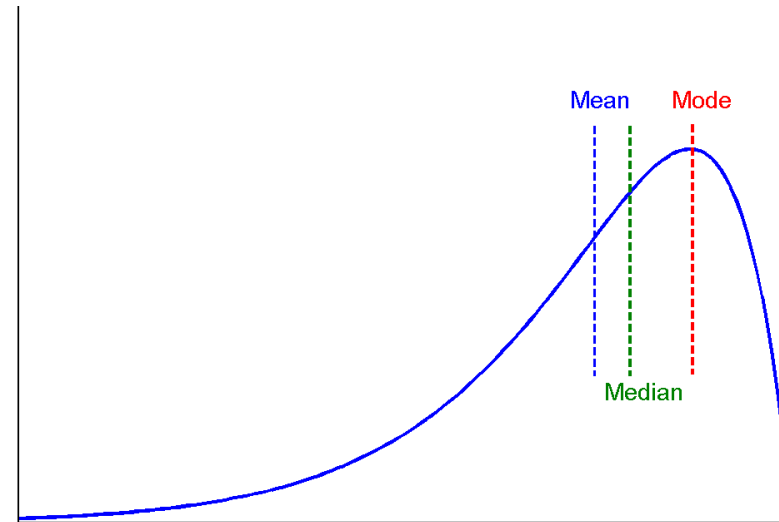
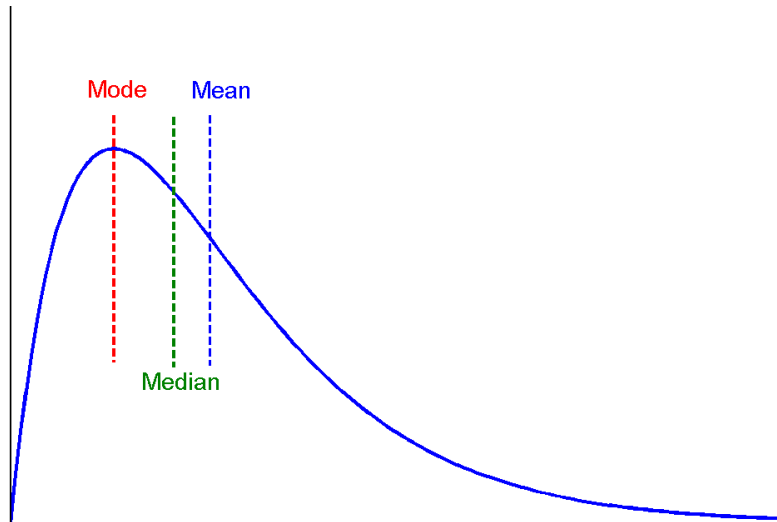
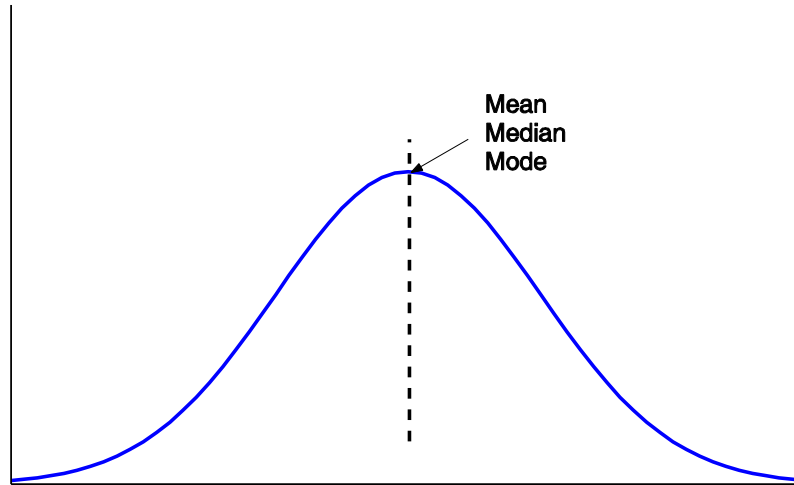
$$\text{mean}(x) = \bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r + 1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

经验公式 $mean - mode = 3 \times (mean - median)$

中性化趋势度量：均值、中位数和众数

- 中位数，均值和对称模式，正面和负面的偏斜数据



分布趋势度量

- 方差和标准差
- 分位数
 - 分位数: Q1 (第25百分位) , Q3 (第75百分位)
 - 分位数极差: $IQR = Q3 - Q1$
 - 离群点: 通常情况下, 一个值高于 $Q3 + 1.5 \times IQR$ /低于 $Q1 - 1.5 \times IQR$

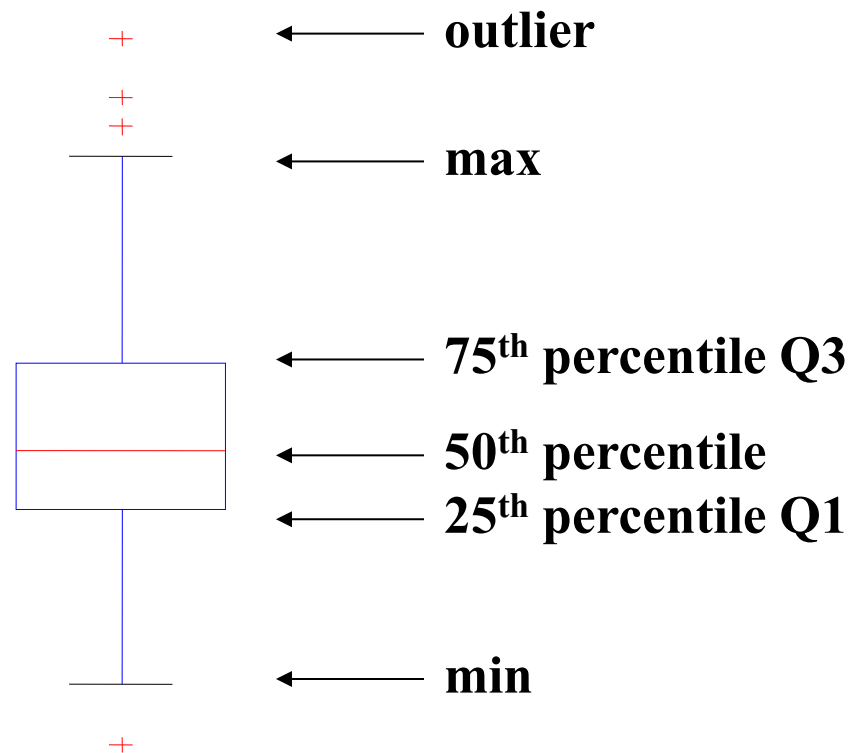
分布趋势度量

● 箱线图 (boxplot)

- min, Q1, median, Q3, max;
单独添加胡须表示离群点

Max: 不高于 $Q3 + 1.5 * IQR$

Min: 不低于 $Q1 - 1.5 * IQR$



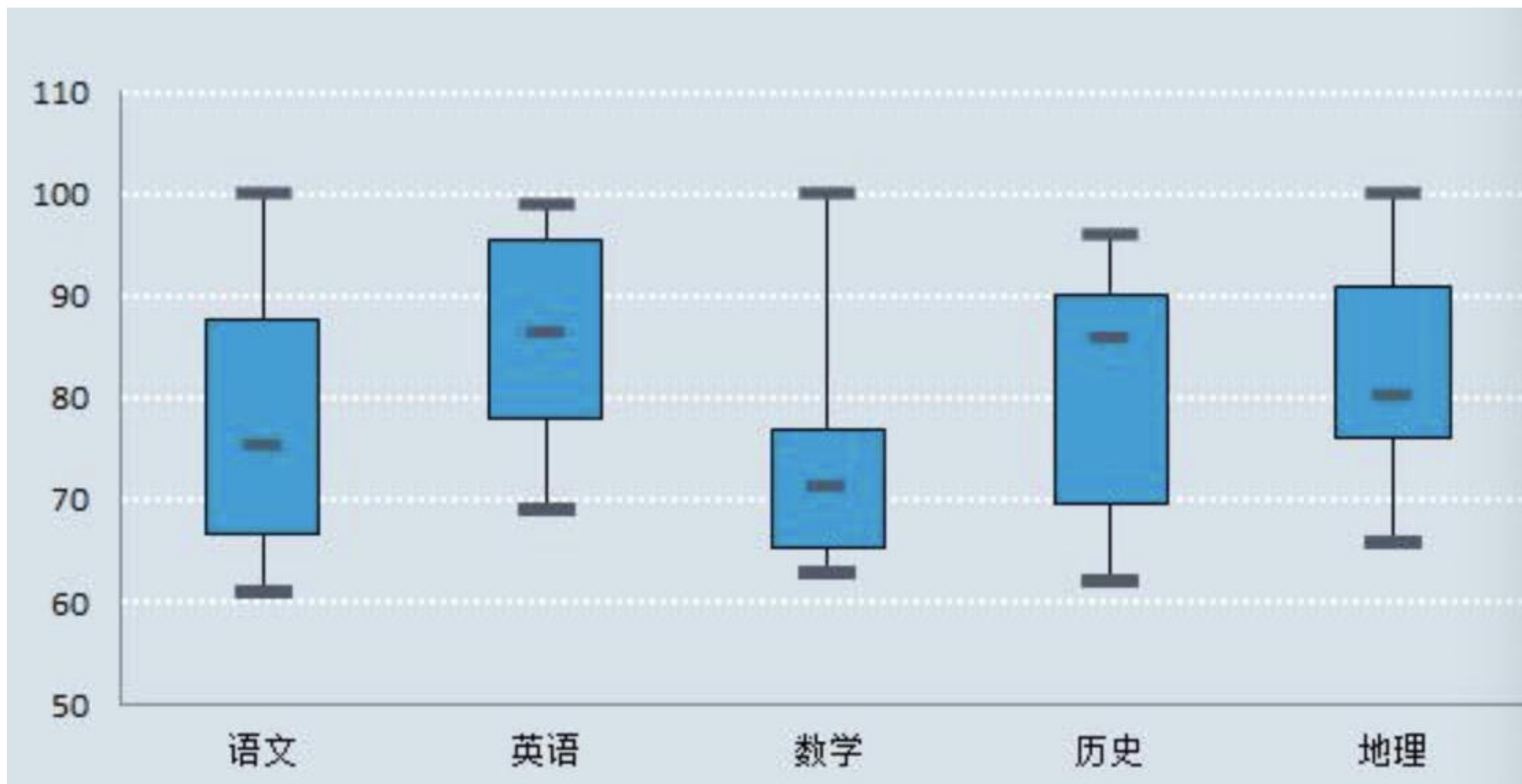
分布趋势度量—例子

- 现在一家电商公司要卖两个同类型的商品，它们的一周销量（单位：个）如下：
 - 商品A：10, 10, 10, 11, 12, 12, 12
 - 商品B：3, 5, 6, 11, 16, 17, 19
- 它们的平均数一样，中位数也一样，可它们的真实情况呢？

$$s^2 = \frac{(x_1 - \bar{X})^2 + (x_2 - \bar{X})^2 + (x_3 - \bar{X})^2 + \dots + (x_n - \bar{X})^2}{n}$$

- 上述公式是总体数据集的方差计算，当数据集为部分抽样样本时，n应该改为n-1。数据集足够大时，两者的误差也可以忽略不计。

数据可视化 —— 箱线图分析



箱线图能够分析多个属性数据的离散度差异性

数据可视化 —— 箱线图分析

- IRIS: 鸢尾花 (sepal:萼片,petal:花瓣)

- ▣ sepal length in cm

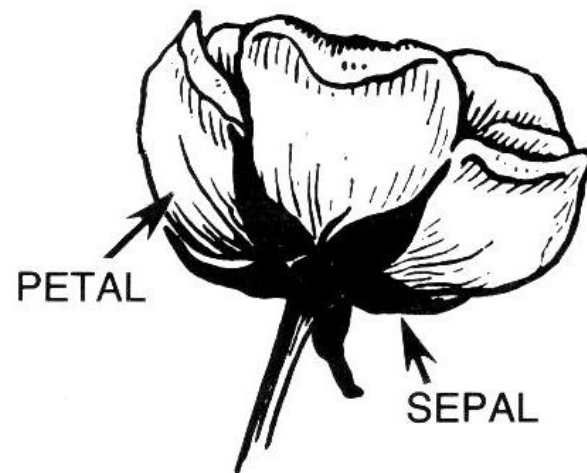
- ▣ sepal width in cm

- ▣ petal length in cm

- ▣ petal width in cm

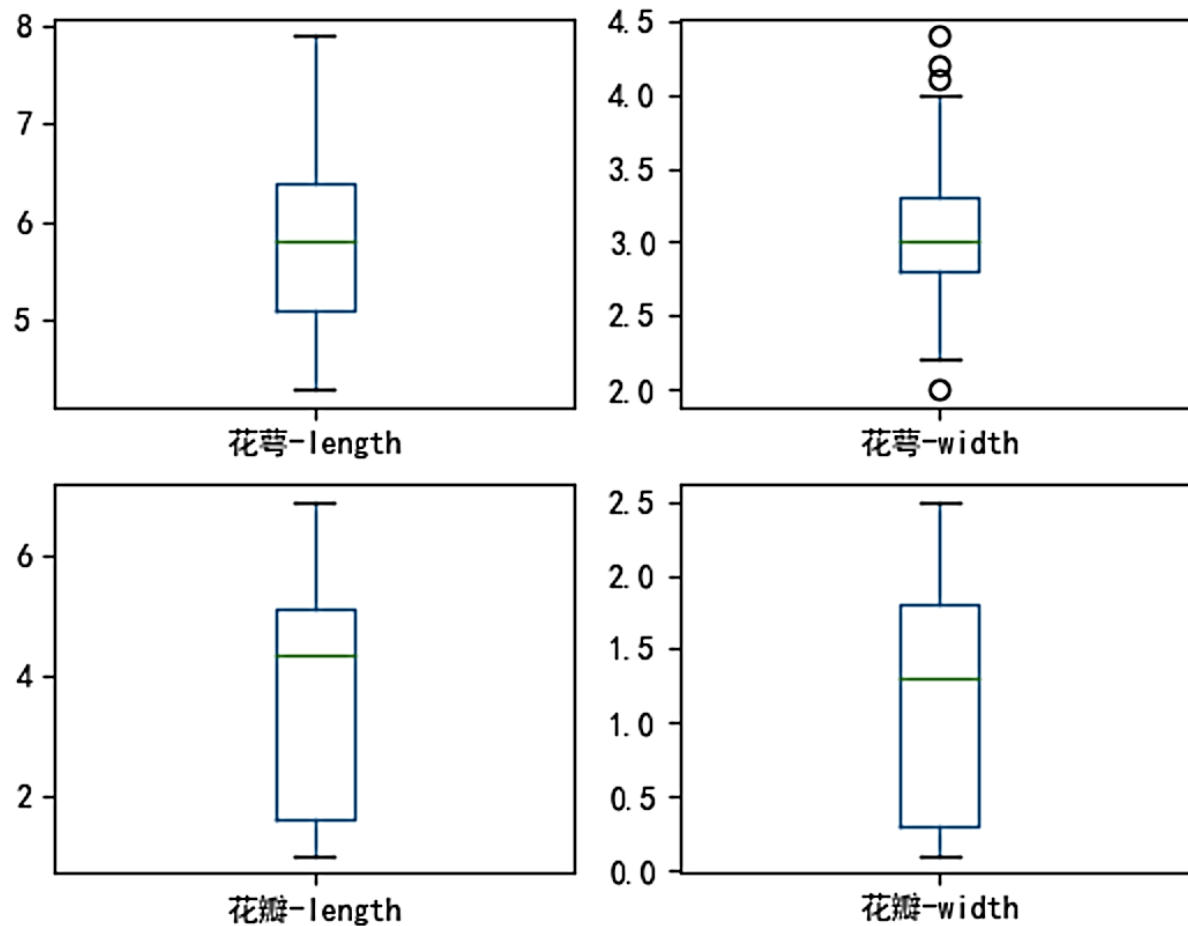
- ▣ class:

- Iris Setosa
- Iris Versicolour
- Iris Virginica



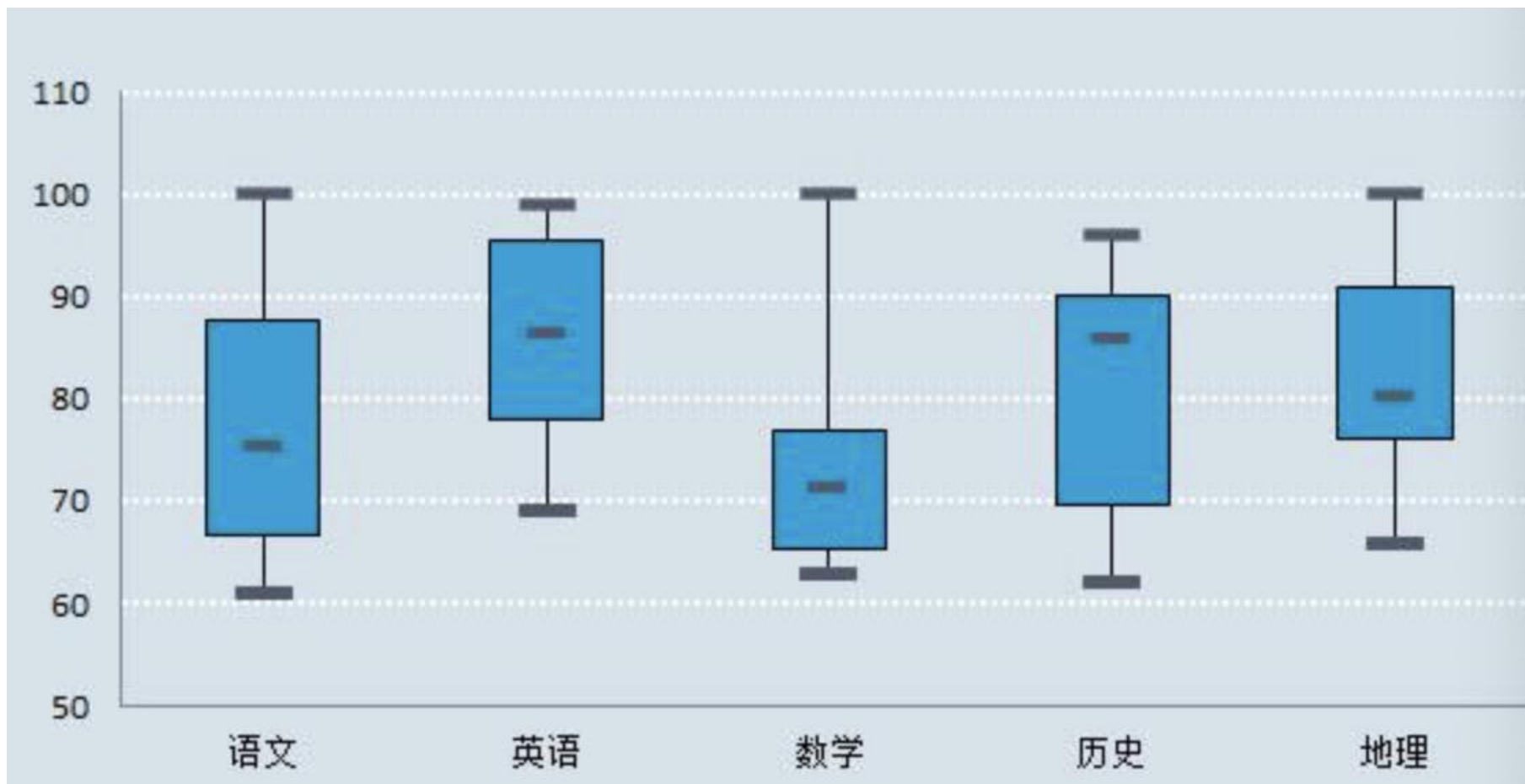
```
6.7,3.0,5.2,2.3,Iris-virginica
6.3,2.5,5.0,1.9,Iris-virginica
6.5,3.0,5.2,2.0,Iris-virginica
6.2,3.4,5.4,2.3,Iris-virginica
5.9,3.0,5.1,1.8,Iris-virginica
5.1,3.8,1.6,0.2,Iris-setosa
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
```

数据可视化 —— 箱线图分析



参考: https://blog.csdn.net/H_lukong/article/details/90139700

数据可视化 —— 箱线图分析

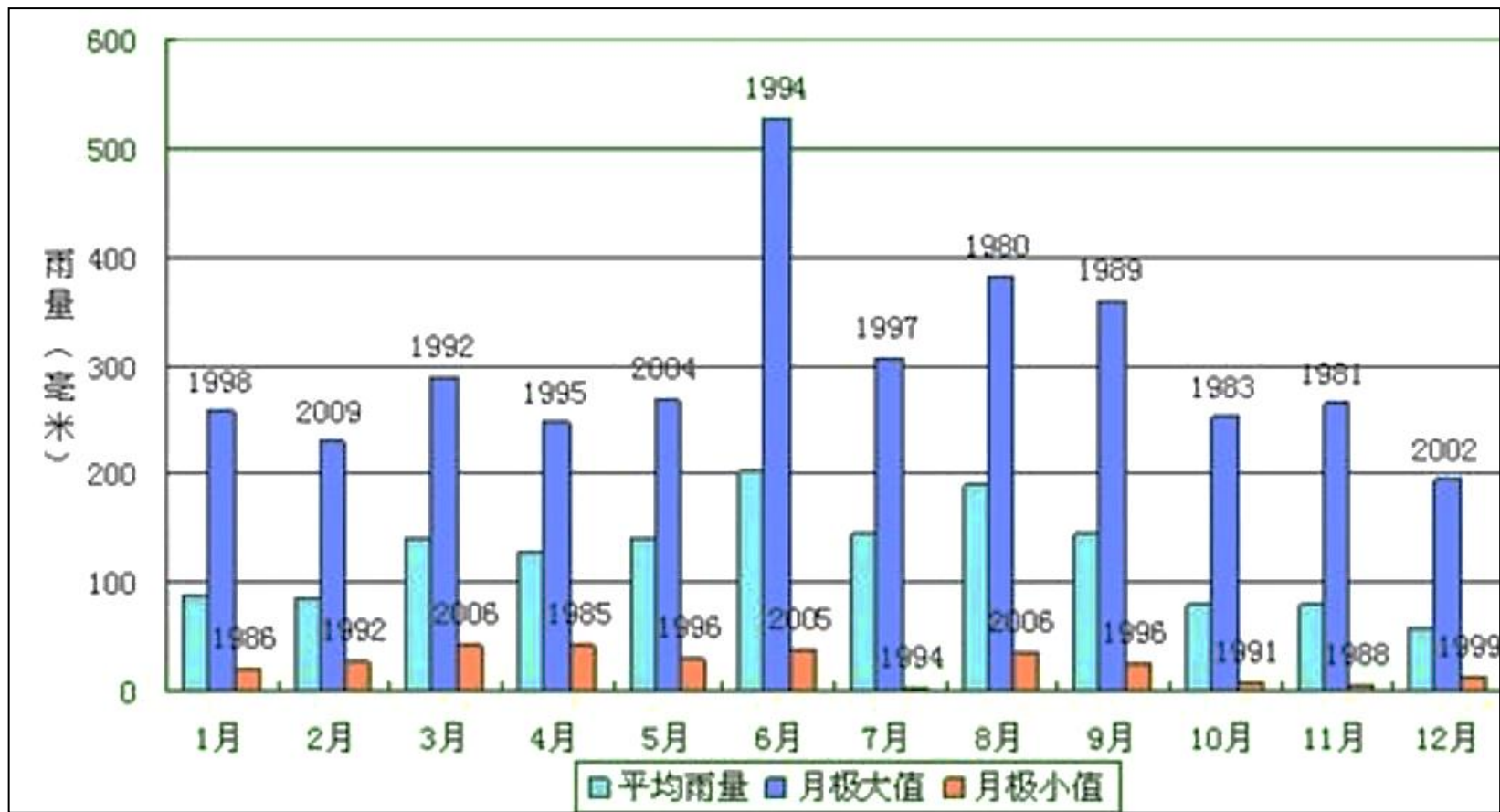


- 盒状图能够分析**多个属性数据**的**离散度差异性**
- 如果希望分析**单个属性**在**各个区间**的**变化分布**怎么办?
 - 例如：如果希望分析**语文成绩**在每个**分数段**的**变化分布**

数据可视化 —— 直方图分析

- 直方图

- 用来分析单个属性在各个区间变化分布



数据可视化 —— 直方图分析

- IRIS (sepal:萼片,petal:花瓣)

- ▣ sepal length in cm

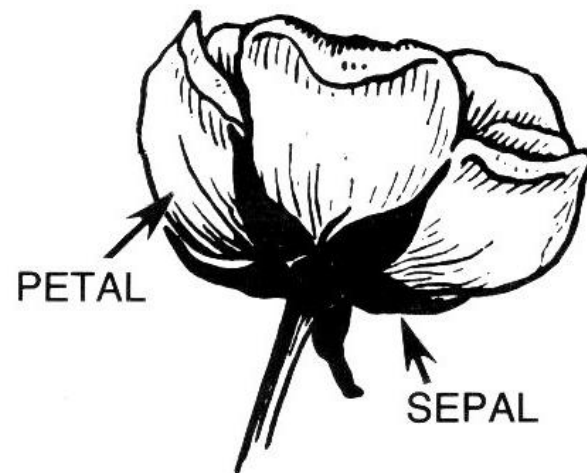
- ▣ sepal width in cm

- ▣ petal length in cm

- ▣ petal width in cm

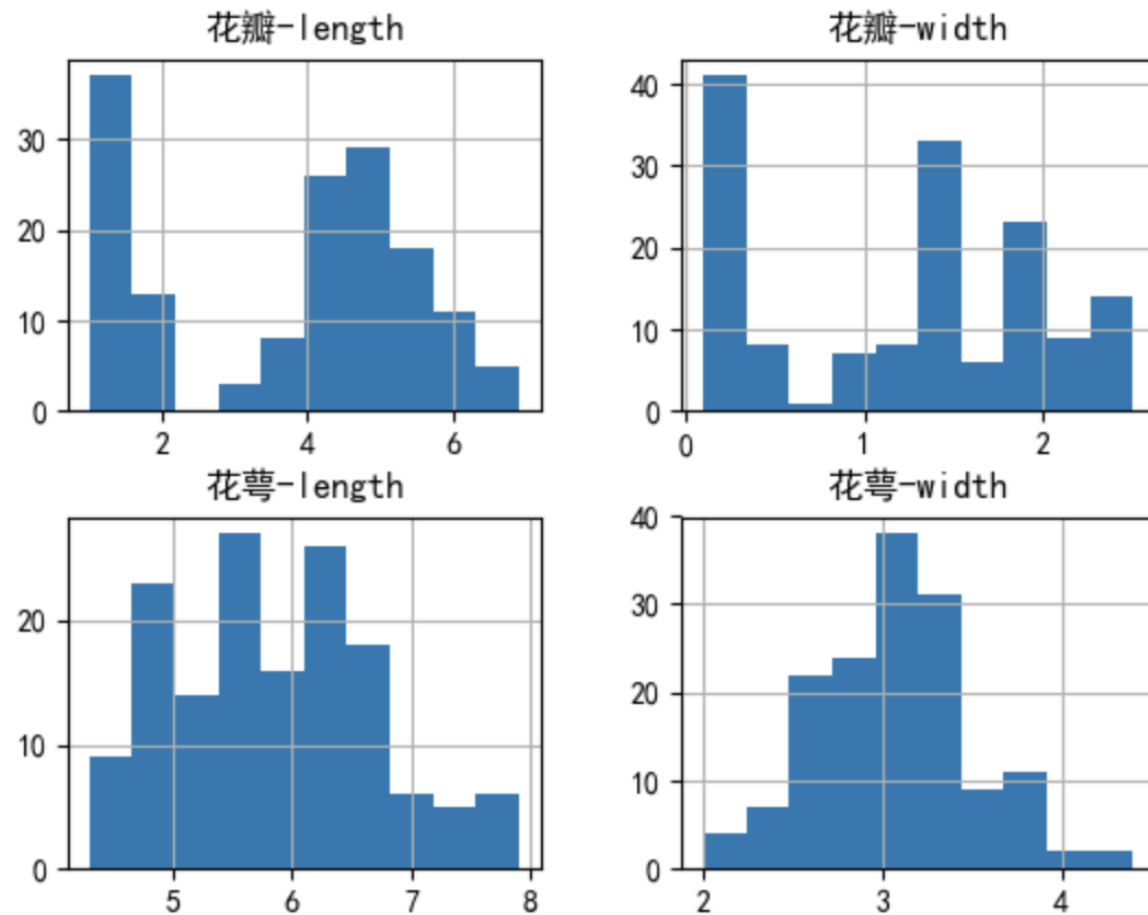
- ▣ class:

- Iris Setosa
- Iris Versicolour
- Iris Virginica



```
6.7,3.0,5.2,2.3,Iris-virginica
6.3,2.5,5.0,1.9,Iris-virginica
6.5,3.0,5.2,2.0,Iris-virginica
6.2,3.4,5.4,2.3,Iris-virginica
5.9,3.0,5.1,1.8,Iris-virginica
5.1,3.8,1.6,0.2,Iris-setosa
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
```

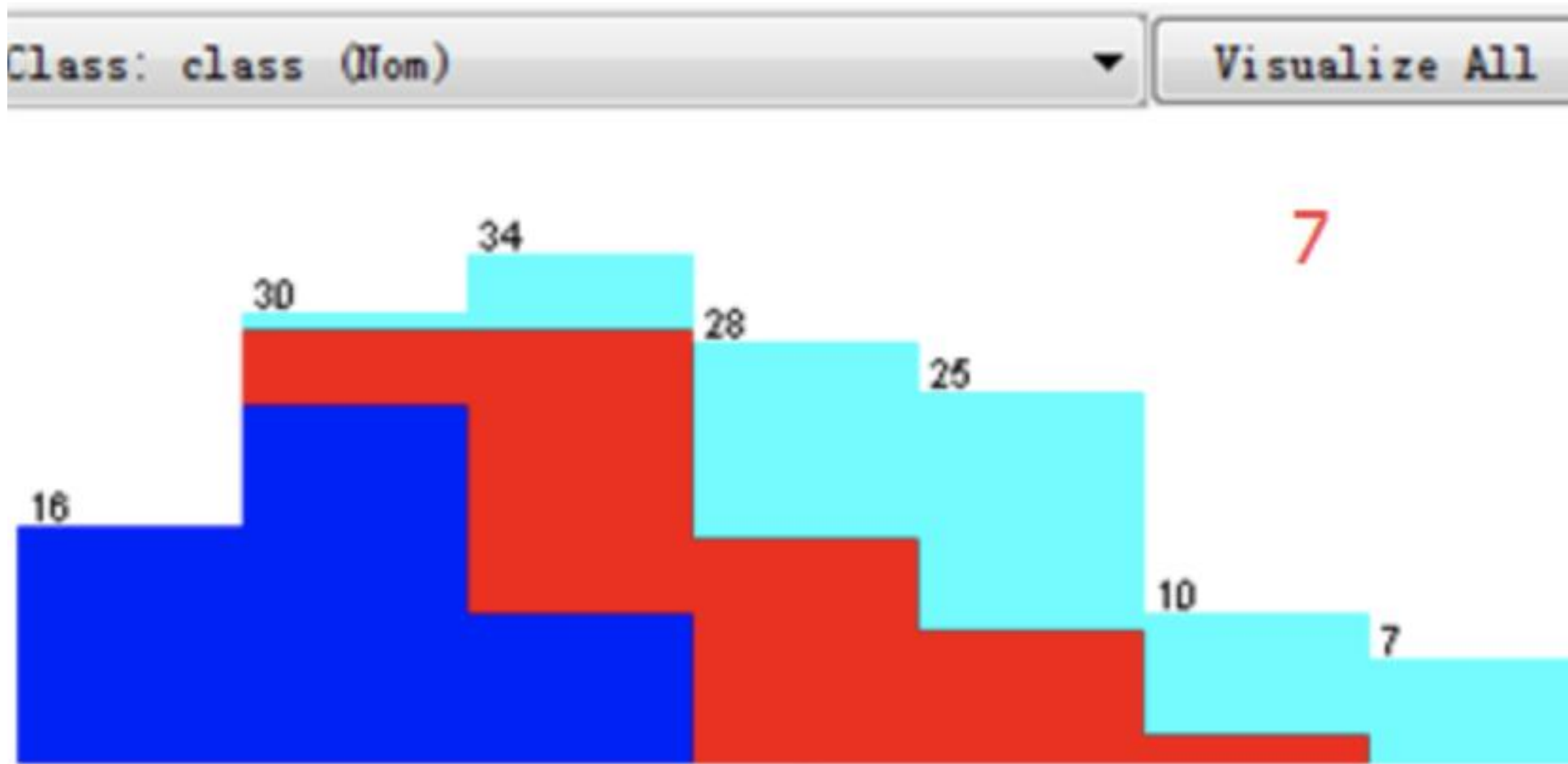
数据可视化 —— 直方图分析案例



参考: https://blog.csdn.net/H_lukong/article/details/90139700

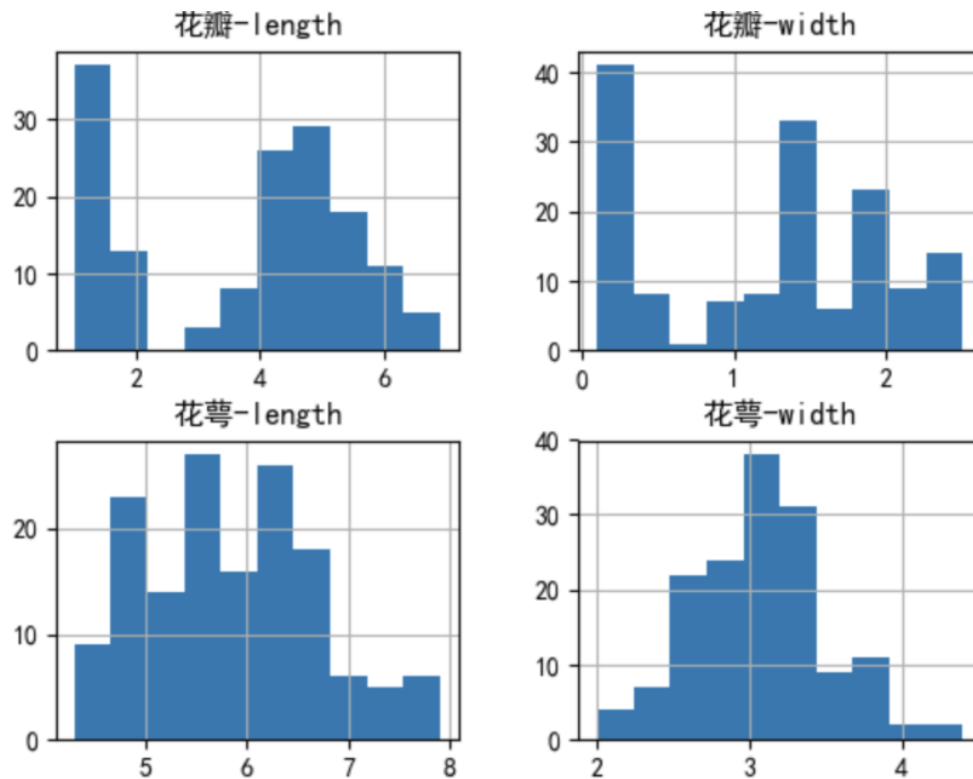
数据可视化 —— 直方图分析案例

分类算法特征分析： 我们可以看到花萼宽度在3个类别下的分布具有差异



数据可视化 —— 直方图分析案例

- **直方图**：用来分析**单个属性**在**各个区间**变化分布

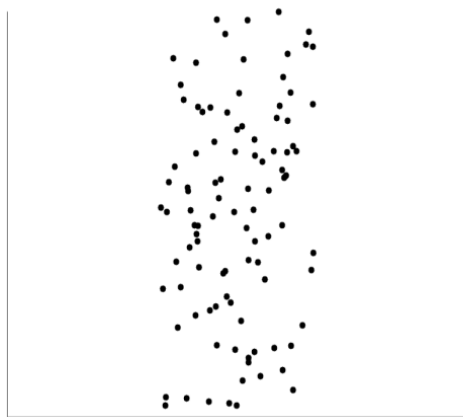
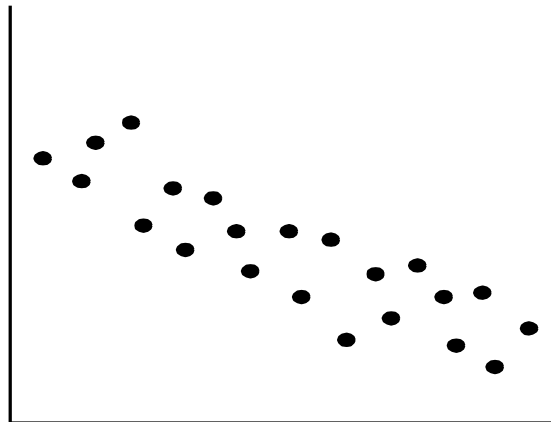
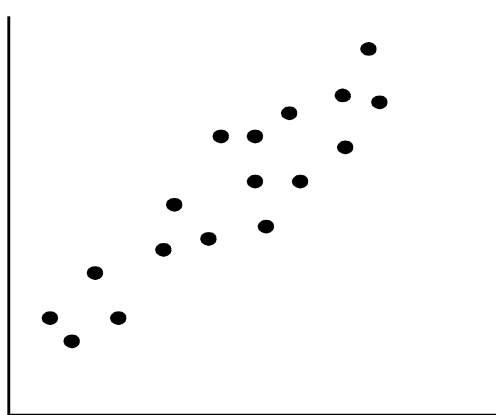


如果希望分析2个属性数据的**关联关系**，怎么办？
例如：如果希望分析**花瓣长度和花萼长度**的关联关系

数据可视化 —— 散点图分析案例

- 散点图

- 用来显示两组数据的相关性分布



数据可视化 —— 散点图分析

- IRIS (sepal:萼片,petal:花瓣)

- ▣ sepal length in cm

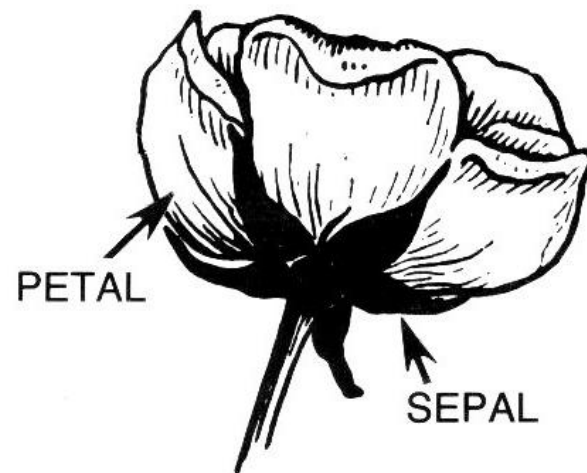
- ▣ sepal width in cm

- ▣ petal length in cm

- ▣ petal width in cm

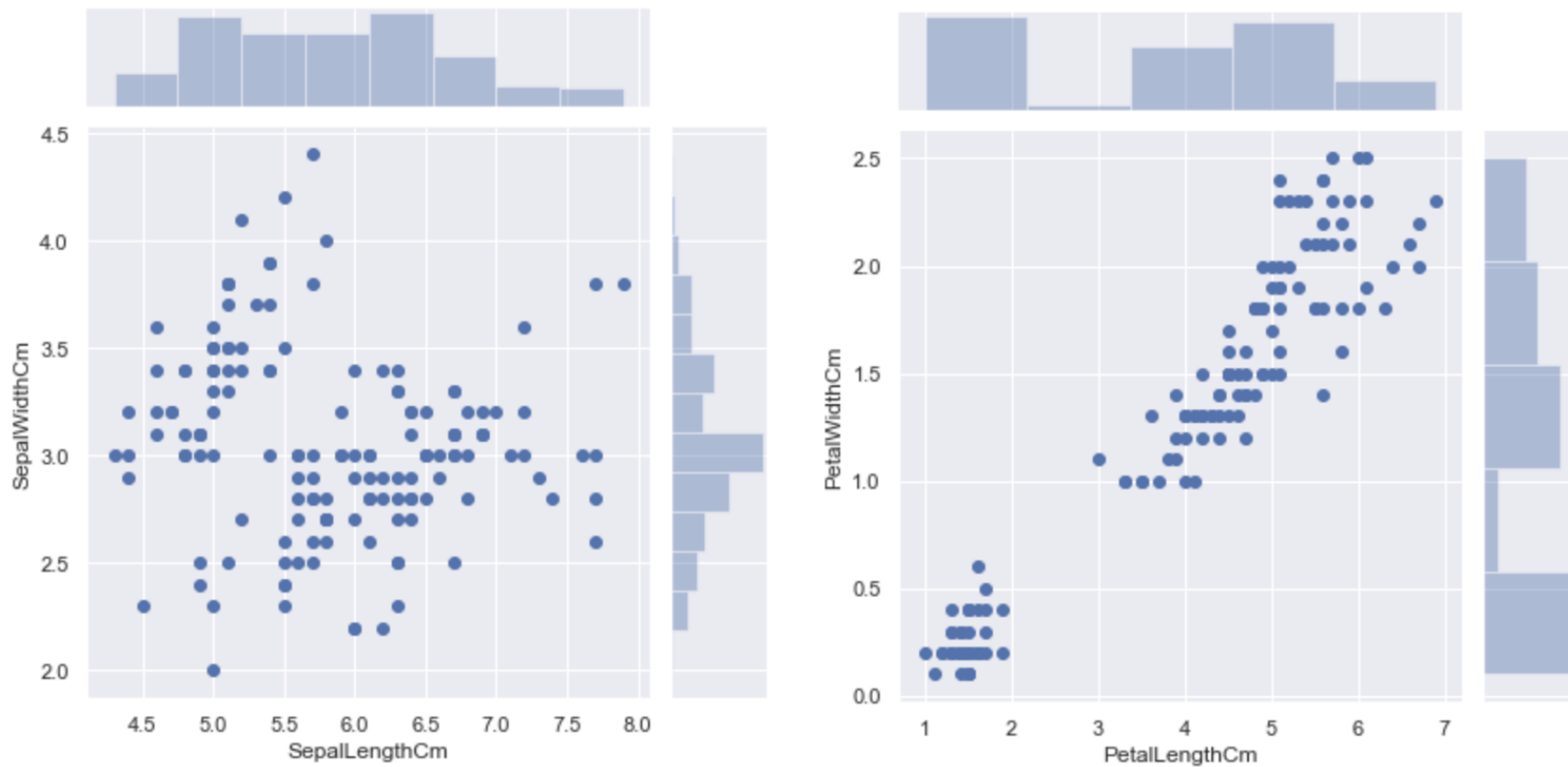
- ▣ class:

- Iris Setosa
- Iris Versicolour
- Iris Virginica



```
6.7,3.0,5.2,2.3,Iris-virginica
6.3,2.5,5.0,1.9,Iris-virginica
6.5,3.0,5.2,2.0,Iris-virginica
6.2,3.4,5.4,2.3,Iris-virginica
5.9,3.0,5.1,1.8,Iris-virginica
5.1,3.8,1.6,0.2,Iris-setosa
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
```

数据可视化 —— 散点图分析案例



参考: <https://www.cnblogs.com/star-zhao/p/9847082.html>

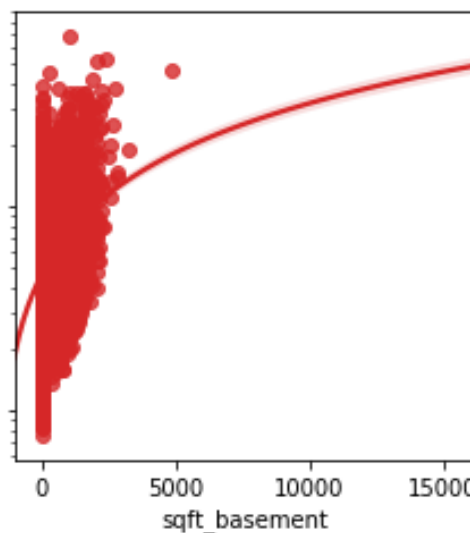
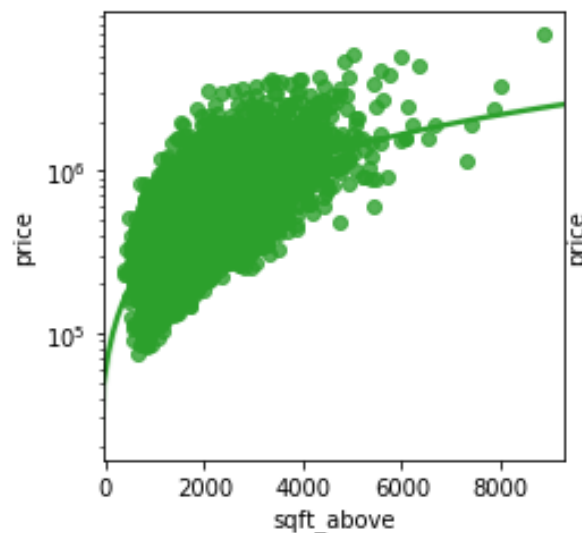
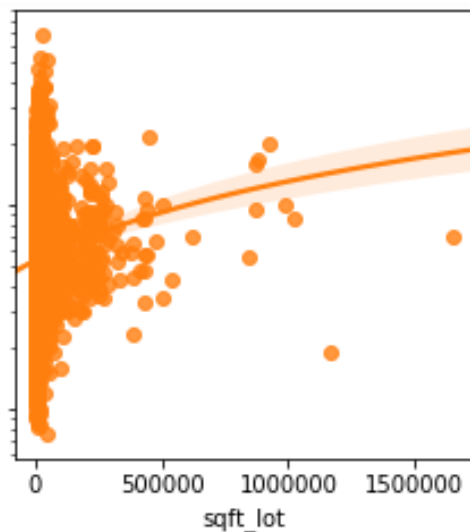
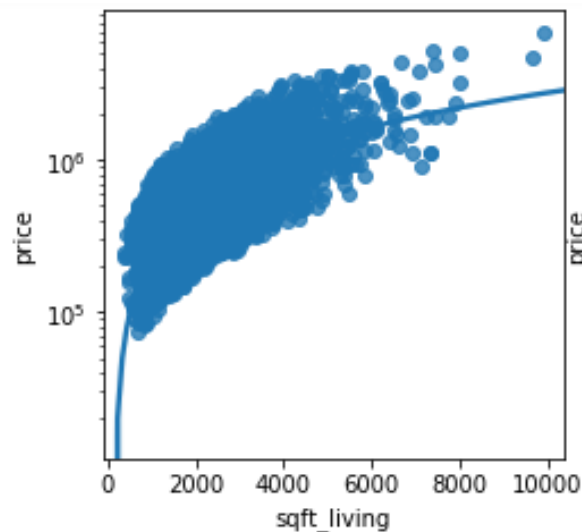
散点图分析在数值预测中的应用-房价预测

房屋销售价格以及房屋的基本信息建立模型，来预测在此期间其他房屋的销售价格

销售日期	销售价格	卧室数	浴室数	房屋面积	停车面积	楼层数	房屋评分	建筑面积	地下室面积	建筑年份	修复年份	纬度	经度
20150302	545000	3	2.25	1670	6240	1	8	1240	430	1974	0	47.6413	-122.113
20150211	785000	4	2.5	3300	10514	2	10	3300	0	1984	0	47.6323	-122.036
20150107	765000	3	3.25	3190	5283	2	9	3190	0	2007	0	47.5534	-122.002
20141103	720000	5	2.5	2900	9525	2	9	2900	0	1989	0	47.5442	-122.138
20140603	449500	5	2.75	2040	7488	1	7	1200	840	1969	0	47.7289	-122.172
20150506	248500	2	1	780	10064	1	7	780	0	1958	0	47.4913	-122.318
20150305	675000	4	2.5	1770	9858	1	8	1770	0	1971	0	47.7382	-122.287
20140701	730000	2	2.25	2130	4920	1.5	7	1530	600	1941	0	47.573	-122.409
20140807	311000	2	1	860	3300	1	6	860	0	1903	0	47.5496	-122.279
20141204	660000	2	1	960	6263	1	6	960	0	1942	0	47.6646	-122.202
20150227	435000	2	1	990	5643	1	7	870	120	1947	0	47.6802	-122.298
20140904	350000	3	1	1240	10800	1	7	1240	0	1959	0	47.5233	-122.185
20140902	385000	3	2.25	1630	1598	3	8	1630	0	2008	0	47.6904	-122.347
20150413	235000	2	1	930	10505	1	6	930	0	1930	0	47.4337	-122.329
20140930	350000	3	1	1300	10236	1	6	1300	0	1971	0	47.5028	-121.77
20150507	1350000	4	1.75	2000	3728	1.5	9	1820	180	1926	0	47.643	-122.299
20140530	459900	3	1.75	2580	11000	1	7	1290	1290	1951	0	47.5646	-122.181
20140723	430000	6	3	2630	8800	1	7	1610	1020	1959	0	47.7166	-122.293
20141003	718000	5	2.75	2930	7663	2	9	2930	0	2013	0	47.5308	-122.184

基本idea：哪些属性跟房价相关

散点图分析在数值预测中的应用-房价预测



- 房屋面积
- 停车面积
- 建筑面积
- 地下室面积

越是强相关，说明该属性
对预测房价更有作用



南京大學
NANJING UNIVERSITY

目录

01

数据对象和属性

02

数据统计与可视化

03

数据相似性和相异性度量

度量数据的相似性和相异性

- 数据矩阵

- N个数据, p个维度

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

- 相异矩阵

- N个数据点, 记录两点之间的距离
- 三角矩阵

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

度量数据的相似性和相异性

- **相似度Similarity**
 - 度量两个数据对象有多相似
 - 值越大就表示数据对象越相似
 - 通常取值范围为 $[0,1]$
- **相异度Dissimilarity** (e.g., distance)
 - 度量两个数据对象的差别程度
 - 值越小就表示数据越相似
 - 最小相异度通常为0
- **邻近性Proximity**
 - 指相似度或者相异度

标称属性的邻近性度量

- 标称属性可以取两个或多个状态
- 方法: 简单匹配
 - m: 匹配次数, p: 属性总数

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

id	属性1	属性2	属性3	属性4
1	弹琴	跳高	唱歌	背诗
2	弹琴	跳远	跳舞	读书

$$d(1, 2) = \frac{4 - 1}{4}$$

二值属性的邻近性度量

- 二值属性的邻近性度量例子

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender 是对称属性，其余都是非对称属性，假设只计算非对称属性
- Y 和P 的值为 1, N的值为 0

$$d(jack, mary) = ?$$

$$d(jack, jim) = ?$$

$$d(jim, mary) = ?$$

二值属性的邻近性度量

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- 一个邻接表

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

- 距离度量对称的二值变量

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

二值属性的邻近性度量

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N


- 一个邻接表

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

- 距离度量对称的二值变量

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

- 距离度量非对称的二值变量


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

二值属性的邻近性度量

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- 一个邻接表

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

- 距离度量对称的二值变量

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

- 距离度量非对称的二值变量

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

通常正常用户占大多数，因此 t 远大于 q ，使得 t 作为分母，将导致值非常小，而失去比较意义

二值属性的邻近性度量

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N


- 一个邻接表

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

- 距离度量对称的二值变量

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

- 距离度量非对称的二值变量


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

- Jaccard 系数 (杰卡德系数)

$$\text{sim}_{\text{Jaccard}}(i, j) = \frac{q}{q + r + s}$$

填空题

- 二值属性的邻近性度量例子

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender 是对称属性，其余都是非对称属性，假设只计算非对称属性
- Y 和P 的值为 1, N的值为 0

$$d(jack, mary) = \text{[填空1]}$$

$$d(jack, jim) = \text{[填空2]}$$

$$d(jim, mary) = \text{[填空3]}$$

二值属性的邻近性度量

- 二值属性的邻近性度量例子

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender 是对称属性，其余都是非对称属性，假设只计算非对称属性
- Y 和P 的值为 1, N的值为 0

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

数值属性的邻近性度量

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

数值属性的邻近性度量

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

- 闵可夫斯基距离

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

- 性质

- $d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$ (正定性)
- $d(i, j) = d(j, i)$ (对称性)
- $d(i, j) \leq d(i, k) + d(k, j)$ (三角不等式)

数值属性的邻近性度量

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

- **h = 1: 曼哈顿距离**

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$$

- **h = 2: 欧氏距离**

$$d(i, j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{ip} - x_{jp}|^2)}$$

- **h $\rightarrow \infty$. “上确界距离”.**

$$d(i, j) = \lim_{h \rightarrow \infty} \left(\sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f |x_{if} - x_{jf}|$$

填空题

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

- $h = 1$: 曼哈顿距离, 求x1和x2之间的曼哈顿距离: [填空1]

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$

- $h = 2$: 欧氏距离, 求x1和x2之间的欧式距离: [填空2]

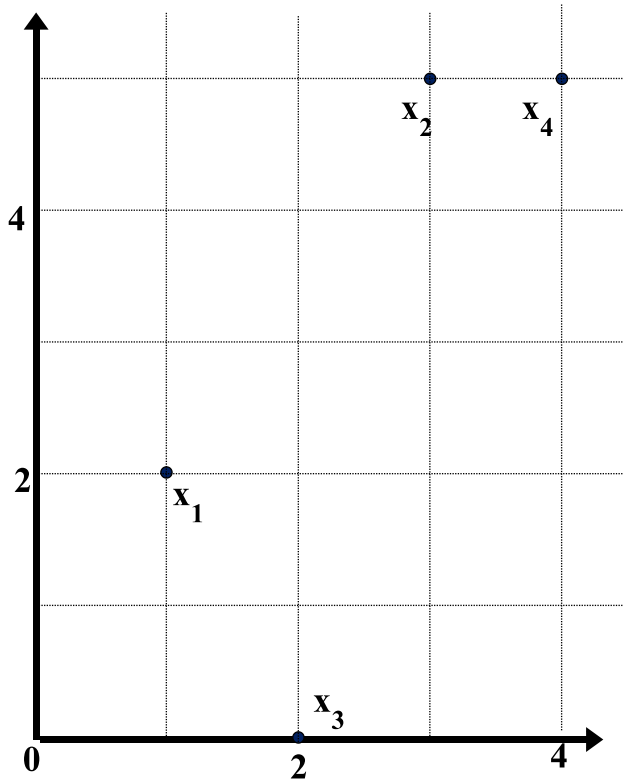
$$d(i, j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2)}$$

- $h \rightarrow \infty$. “上确界距离”, 求x1和x2之间的上确界距离: [填空3]

$$d(i, j) = \lim_{h \rightarrow \infty} \left(\sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f |x_{if} - x_{jf}|$$

数值属性的邻近性度量

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



曼哈顿距离 (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

欧氏距离 (L_2)

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

上确界距离

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

余弦相似性

- 余弦相似性

- 一个文档可以用词频向量来表示 (注意: 词的对齐)

<i>Document</i>	<i>team coach</i>		<i>hockey</i>	<i>baseball</i>	<i>soccer</i>	<i>penalty</i>	<i>score</i>	<i>win</i>	<i>loss</i>	<i>season</i>
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

Baseball wins a score in the season (0,0,1,1,0,1,1,0,1)

In the season, soccer loss a score (0,0,0,0,1,0,1,0,1,1)

- 余弦度量

- $\cos(d1, d2) = (d1 \bullet d2) / (||d1|| \cdot ||d2||)$,

填空题

- 余弦相似性

- 一个文档可以用词频向量来表示（注意：词的对齐）

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

Baseball wins a score in the season (0,0,1,1,0,1,1,0,1)

In the season, soccer loss a score (0,0,0,0,1,0,1,0,1,1)

- 余弦度量

- $\cos(d_1, d_2) = (d_1 \bullet d_2) / (||d_1|| \cdot ||d_2||)$,

$$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$$

$$d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$$

- 求这两篇文档的余弦相似性： [填空1]

余弦相似性

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / \|d_1\| \|d_2\| ,$$

- 例如:

$$d1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$$

$$d2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$$

$$d1 \bullet d2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25$$

$$\|d1\| = (5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$$

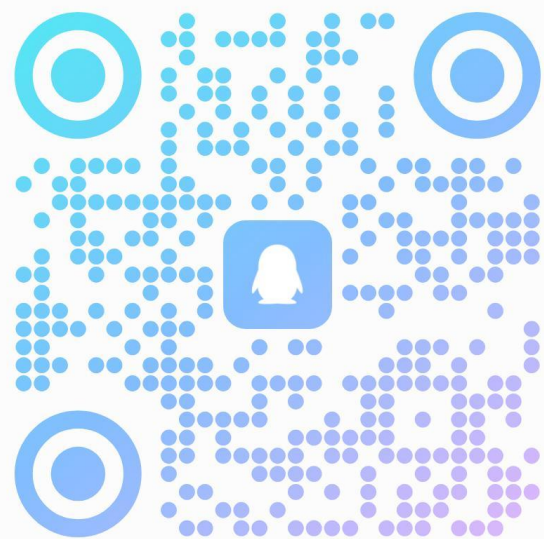
$$\|d2\| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5} = (17)^{0.5} = 4.12$$

$$\cos(d1, d2) = 0.94$$



数据挖掘 2025 课程群

群号: 732748314



扫一扫二维码，加入群聊



如果有人问起...

号码走势图										综合分析表																																						
大乐透号码分布图										--请选择期号-- 至 --请选择期号-- 查看																																						
期号	号码	中奖号码分布图																																														
		前区																				后区																										
		01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	01	02	03	04	05	06	07	08	09	10	11	12
09075	01 07 11 32 33 + 09 11	1					7				11																					32	33									9	11					
09076	03 09 11 18 24 + 04 05		3					9		11							18						24																4	5								
09077	07 20 28 32 34 + 06 08						7													20								28				32	34						6	8								
09078	11 23 26 31 34 + 05 11									11													23			26					31		34					5						11				
09079	01 02 07 13 34 + 04 10	1	2				7						13																				34					4						10				
09080	07 16 19 29 35 + 02 11						7										16		19										29					35		2										11		
09081	02 15 20 25 33 + 03 06		2													15				20					25							33						3			6							
09082	05 11 26 29 32 + 08 11				5						11															26		29		32												8			11			
09083	02 05 28 34 35 + 01 07		2		5																							28					34	35	1					7								
09084	05 09 19 26 35 + 02 10				5			9											19							26								35		2									10			
期号	号码	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	01	02	03	04	05	06	07	08	09	10	11	12

幸存者偏差



