

数据挖掘

Data Mining



第二章 认识数据

主讲人：丁兆云

数据挖掘

Data Mining



01

数据类型

02

数据统计汇总

03

数据相似性和
相异性度量



1 数据类型

3

- 记录数据
 - 关系记录
 - 数据矩阵
 - 文档数据
 - 交易数据
- 图形和网络
 - 万维网
 - 社会或信息网络
 - 分子结构
 - 有序
 - 时间数据：时间序列
 - 顺序数据：交易序列
 - 基因序列数据
- 视频数据的图像序列
 - 空间，图像和多媒体
 - 空间数据：地图

	Document 1	Document 2	Document 3
Team	3	0	0
Coach	0	7	1
Play	5	0	0
Ball	0	2	0
Score	2	1	1
Game	6	0	2
Win	0	0	2
Lost	2	3	0
Timeout	0	0	3
Season	2	0	0

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk



- **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
- 称为样品，示例，实例，数据点，对象，元组 (tuple)
- 数据对象所描述的属性
 - 数据库中的行 -> 数据对象
 - 列 -> “属性”

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

Objects



- 属性（变量、特性、字段、特征或维）：一个数据字段，代表一个数据对象的特征或功能。

例如 • 乘客_ID • 是否存活 • 客舱等级

- 类型：

• 标称 • 二进制 • 序数 • 区间标度 • 比率标度

	A	B	C	D	E	F	G	H	I	J	K	L
1	Passenge	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, female		38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen	female	26	0	0	STON/O2	7.925		S
5	4	1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr.	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, M	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, M	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, female		27	0	2	347742	11.1333		S
11	10	1	2	Nasser, M	female	14	1	0	237736	30.0708		C
12	11	1	3	Sandstror	female	4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, M	female	58	0	0	113783	26.55	C103	S

数据对象的别名

- ☒ A 样品
- ☒ B 实例
- ☐ C 维度
- ☒ D 元组
- ☒ E 对象

提交

属性的别名

- ☐ A 元组
- ☒ B 维度
- ☒ C 特征
- ☒ D 字段
- ☐ E 数据点

提交

所谓高维数据，指的是

- ☐ A 数据对象很多
- ☒ B 数据属性很多

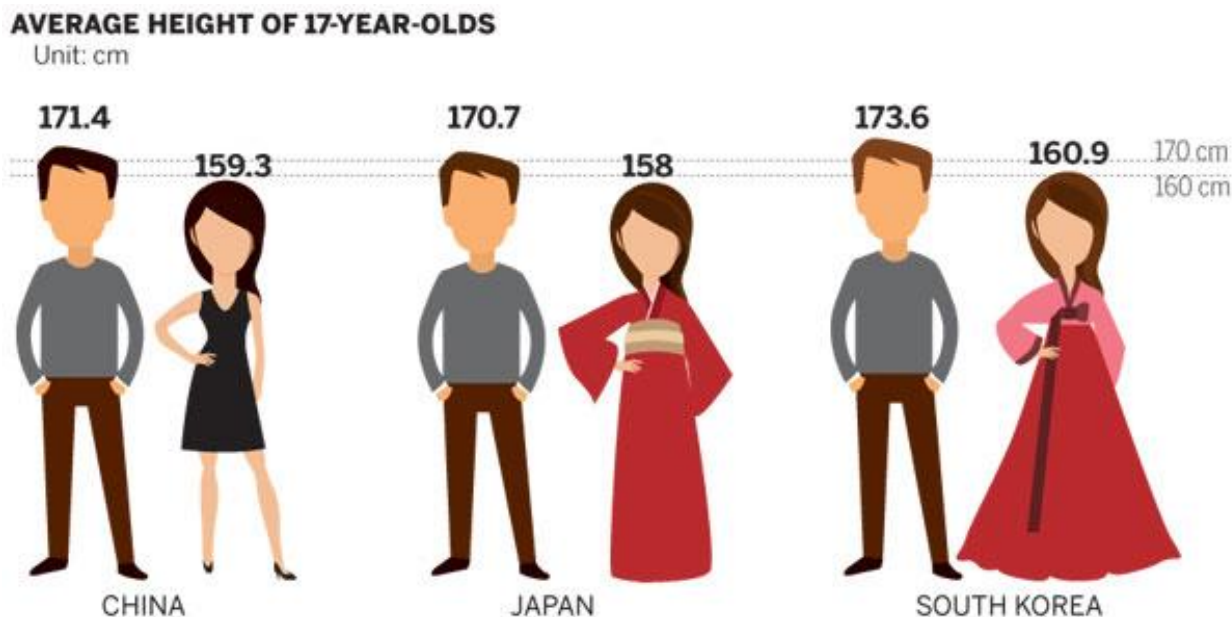
提交

所谓特征选择，是指

- ☒ A 从数据中，选择有代表性的属性
- ☐ B 从数据中，选择有代表性的数据对象

提交

- An **attribute** (属性) is a property or characteristic of an object
 - Example: 人的身高, 体重, 眼睛的颜色等
- **Attribute values** (属性值: 定义属性的特定的特征或参数) are numbers or symbols assigned to an attribute
 - Example: 175cm, 70kg, 黑色



创建时间 = 1 月 2 日

- ☐ A 创建时间表示属性, 1 月 2 日表示属性
- ☐ B 创建时间表示属性值, 1 月 2 日表示属性值
- ☒ C 创建时间表示属性, 1 月 2 日表示属性值
- ☐ D 创建时间表示属性值, 1 月 2 日表示属性

提交



- 常见的四类属性：

- **标称 (Nominal)**

- Examples: ID numbers, zip codes

- **序数 (Ordinal)**

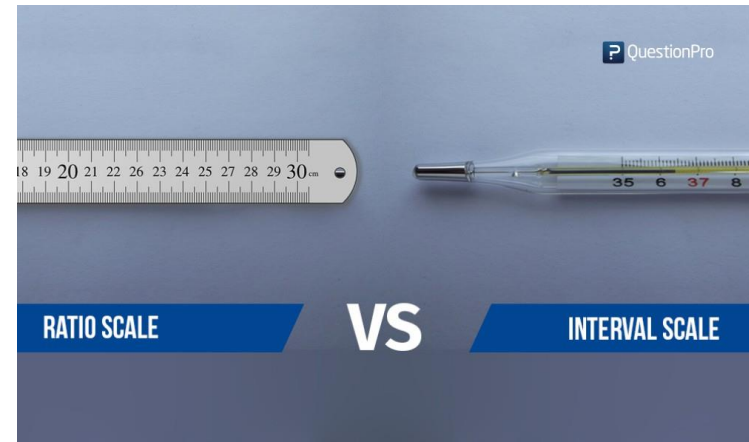
- Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}

- **区间 (Interval)**

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

- **比率 (Ratio)**

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





- 标称：类别，状态
 - Hair_color={黑色, 棕色, 金色, 红色, 红褐色, 灰色, 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
 - 对称二进制两种结果重要
 - 例如, 性别
 - 不对称的二进制结果同样重要。
 - 例如, 医疗测试 (正面与负面)
- 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, 军队排名



1.4 属性的类型

14

- 标称：类别，状态
 - Hair_color={黑色，棕色，金色，红色，红褐色，灰色，白色}
 - 婚姻状况，职业，身份证号码，邮政编码

	A	B	C	D	E	F	G	H	I	J	K	L
1	Passenger	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, female		38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, female		26	0	0	STON/O2	7.925		S
5	4	1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr.	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, M	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy, male		54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, M	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, female		27	0	2	347742	11.1333		S
11	10	1	2	Nasser, M	female	14	1	0	237736	30.0708		C
12	11	1	3	Sandstrom, female		4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, M	female	58	0	0	113783	26.55	C103	S

下列是标称类型的属性是

- ☐ A Survived: 0表示遇难, 1表示幸存
- ☐ B Pclass: 1代表Upper, 2代表Middle, 3代表Lower
- ☐ C Sex: 标识乘客性别
- ☐ D SibSp: 兄弟姐妹及配偶的个数
- ☒ E Embarked: 乘客登船口岸, 可列举

提交



● 二进制

- 只有2个状态（0和1）的属性
- 对称二进制两种结果重要
 - 例如，性别
- 不对称的二进制结果同样重要。
 - 例如，新型冠状病毒肺炎测试（阳性与阴性）

	A	B	C	D	E	F	G	H	I	J	K	L
1	Passenge	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, female		38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, female		26	0	0	STON/O2	7.925		S
5	4	1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr.	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, M	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy, male		54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, M	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, female		27	0	2	347742	11.1333		S
11	10	1	2	Nasser, M	female	14	1	0	237736	30.0708		C
12	11	1	3	Sandstror, female		4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, M	female	58	0	0	113783	26.55	C103	S

下列是对称二进制类型的属性是

- ☐ A Survived: 0表示遇难, 1表示幸存
- ☐ B Pclass: 1代表Upper, 2代表Middle, 3代表Lower
- ☒ C Sex: 标识乘客性别
- ☐ D SibSp: 兄弟姐妹及配偶的个数
- ☐ E Embarked: 乘客登船口岸, 可列举

提交

下列是非对称二进制类型的属性是

- ☒ A Survived: 0表示遇难, 1表示幸存
- ☐ B Pclass: 1代表Upper, 2代表Middle, 3代表Lower
- ☐ C Sex: 标识乘客性别
- ☐ D SibSp: 兄弟姐妹及配偶的个数
- ☐ E Embarked: 乘客登船口岸, 可列举

提交



● 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, 军队排名

	A	B	C	D	E	F	G	H	I	J	K	L
1	Passenge	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, female		38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen	female	26	0	0	STON/O2	7.925		S
5	4	1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr.	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, M	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, M	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, female		27	0	2	347742	11.1333		S
11	10	1	2	Nasser, M	female	14	1	0	237736	30.0708		C
12	11	1	3	Sandstror	female	4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, M	female	58	0	0	113783	26.55	C103	S

下列是序数类型的属性是

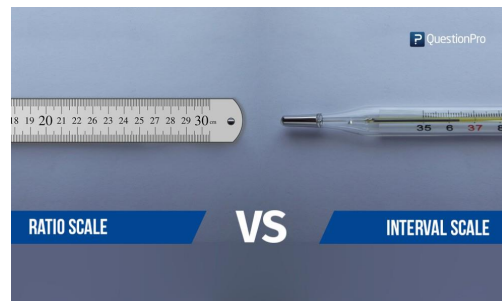
- ☐ A Survived: 0表示遇难, 1表示幸存
- ☒ B Pclass: 1代表Upper, 2代表Middle, 3代表Lower
- ☐ C Sex: 标识乘客性别
- ☐ D SibSp: 兄弟姐妹及配偶的个数
- ☐ E Embarked: 乘客登船口岸, 可列举

提交



● 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



● 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

	A	B	C	D	E	F	G	H	I	J	K	L
1	Passenge	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, female		38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen	female	26	0	0	STON/O2	7.925		S
5	4	1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr.	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, M	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, M	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, female		27	0	2	347742	11.1333		S
11	10	1	2	Nasser, M	female	14	1	0	237736	30.0708		C
12	11	1	3	Sandstror	female	4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, M	female	58	0	0	113783	26.55	C103	S

下列是比例标度类型的属性是

- ☐ A Survived: 0表示遇难, 1表示幸存
- ☐ B Pclass: 1代表Upper, 2代表Middle, 3代表Lower
- ☐ C Sex: 标识乘客性别
- ☒ D SibSp: 兄弟姐妹及配偶的个数
- ☐ E Embarked: 乘客登船口岸, 可列举

提交



表 不同的属性类型

属性类型		描 述	例 子	操 作
分类的 (定性的)	标称	标称属性的值仅仅只是不同的名字，即标称值只提供足够的信息以区分对象 (=, ≠)	邮政编码、雇员ID号、眼球颜色、性别	众数、熵、列联相关、 χ^2 检验
	序数	序数属性的值提供足够的信息确定对象的序 (<, >)	矿石硬度、{好, 较好, 最好}、成绩、街道号码	中值、百分位、秩相关、游程检验、符号检验
数值的 (定量的)	区间	对于区间属性，值之间的差是有意义的，即存在测量单位 (+, -)	日历日期、摄氏或华氏温度	均值、标准差、皮尔逊相关、 t 和 F 检验
	比率	对于比率变量，差和比率都是有意义的 (+, -, *, /)	绝对温度、货币量、计数、年龄、质量、长度、电流	几何平均、调和平均、百分比变差



表 不同的属性类型

属性类型		描 述	例 子	操 作
分类的 (定性的)	标称	标称属性的值仅仅只是不同的名字，即标称值只提供足够的信息以区分对象 (=, ≠)	邮政编码、雇员ID号、眼球颜色、性别	众数、熵、列联相关、 χ^2 检验
	序数	序数属性的值提供足够的信息确定对象的序 (<, >)	矿石硬度、{好, 较好, 最好}、成绩、街道号码	中值、百分位、秩相关、游程检验、符号检验
数值的 (定量的)	区间	对于区间属性，值之间的差是有意义的，即存在测量单位 (+, -)	日历日期、摄氏或华氏温度	均值、标准差、皮尔逊相关、 t 和 F 检验
	比率	对于比率变量，差和比率都是有意义的 (+, -, *, /)	绝对温度、货币量、计数、年龄、质量、长度、电流	几何平均、调和平均、百分比变差



表 不同的属性类型

属性类型		描 述	例 子	操 作
分类的 (定性的)	标称	标称属性的值仅仅只是不同的名字，即标称值只提供足够的信息以区分对象 (=, ≠)	邮政编码、雇员ID号、眼球颜色、性别	众数、熵、列联相关、 χ^2 检验
	序数	序数属性的值提供足够的信息确定对象的序 (<, >)	矿石硬度、{好, 较好, 最好}、成绩、街道号码	中值、百分位、秩相关、游程检验、符号检验
数值的 (定量的)	区间	对于区间属性，值之间的差是有意义的，即存在测量单位 (+, -)	日历日期、摄氏或华氏温度	均值、标准差、皮尔逊相关、 t 和 F 检验
	比率	对于比率变量，差和比率都是有意义的 (+, -, *, /)	绝对温度、货币量、计数、年龄、质量、长度、电流	几何平均、调和平均、百分比变差



1.5 属性的类型小结

表 不同的属性类型

属性类型		描 述	例 子	操 作
分类的 (定性的)	标称	标称属性的值仅仅只是不同的名字，即标称值只提供足够的信息以区分对象 (=, ≠)	邮政编码、雇员ID号、眼球颜色、性别	众数、熵、列联相关、 χ^2 检验
	序数	序数属性的值提供足够的信息确定对象的序 (<, >)	矿石硬度、{好, 较好, 最好}、成绩、街道号码	中值、百分位、秩相关、游程检验、符号检验
数值的 (定量的)	区间	对于区间属性，值之间的差是有意义的，即存在测量单位 (+, -)	日历日期、摄氏或华氏温度	均值、标准差、皮尔逊相关、 t 和 F 检验
	比率	对于比率变量，差和比率都是有意义的 (+, -, *, /)	绝对温度、货币量、计数、年龄、质量、长度、电流	几何平均、调和平均、百分比变差

标称类型数据的可以实现数学计算

A	众数
B	中位数
C	均值
D	方差
E	相等=
F	加法+
G	除法/

提交

序数类型数据的可以实现数学计算

A	众数
B	中位数
C	均值
D	方差
E	相等=
F	加法+
G	除法/

提交

区间标度类型数据的可以实现数学计算

A	众数
B	中位数
C	均值
D	方差
E	相等=
F	加法+
G	除法/

提交

比例标度类型数据的可以实现数学计算

- | | |
|---|-----|
| A | 众数 |
| B | 中位数 |
| C | 均值 |
| D | 方差 |
| E | 相等= |
| F | 加法+ |
| G | 除法/ |

提交



- 离散属性(Discrete Attribute)
 - 有限或无限可数 (countable infinite)个值
 - 例: 邮政编码, 计数, 文档集的词
 - 常表示为整数变量.
 - 注意: 二元属性(binary attributes)是离散属性的特例
- 连续属性(Continuous Attribute)
 - 属性值为实数
 - 例: 温度, 高度, 重量.
 - 实践中, 实数只能用有限位数字的数度量和表示.
 - 连续属性一般用浮点变量表示.

身高和体重分别是什么类型

- ☐ A 身高离散、体重离散
- ☒ B 身高连续、体重连续
- ☐ C 身高连续、体重离散
- ☐ D 身高离散、体重连续

身高	体重
167	56.7
178	97.5
159	43.8
.....

提交

数据挖掘

Data Mining



01

数据类型

02

数据统计汇总

03

数据相似性和
相异性度量



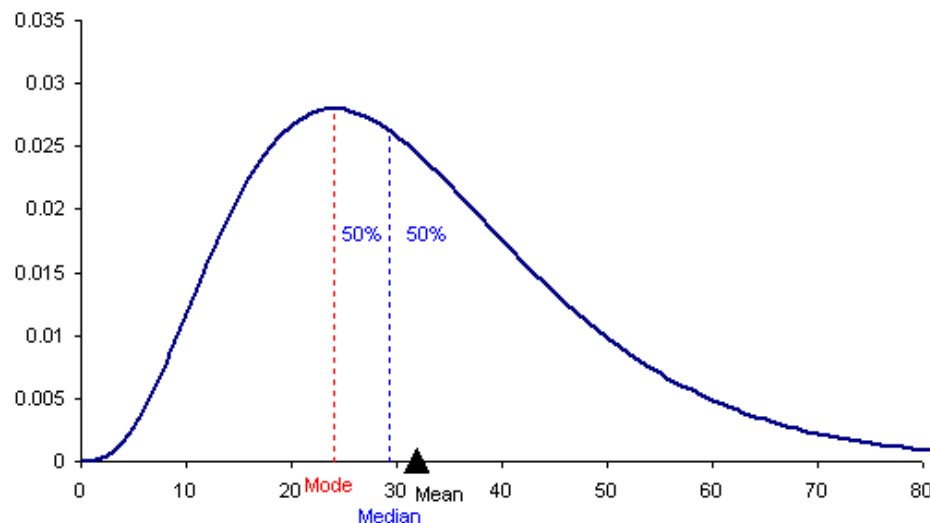
- 动机
 - 为了更好地理解数据：集中趋势，分布
- 数据的统计特性
 - 最大值，最小值，中位数，位数，离群值，方差等。

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%



- 平均值一组数据的均衡点。
- 但是，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。



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

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

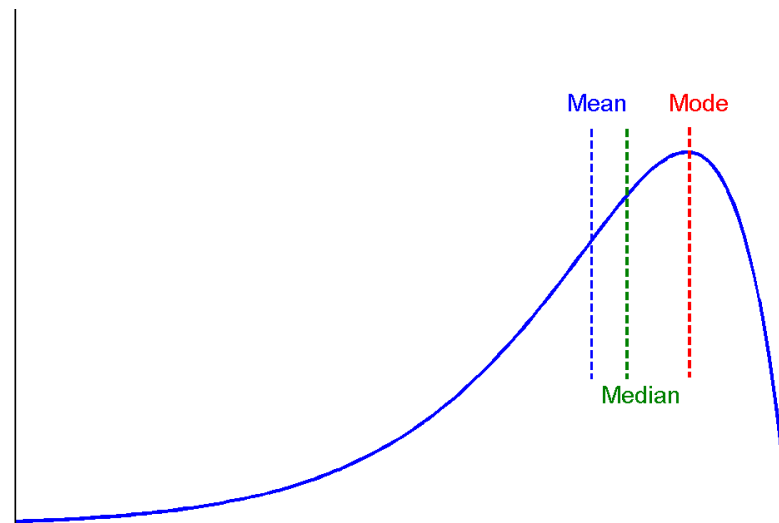
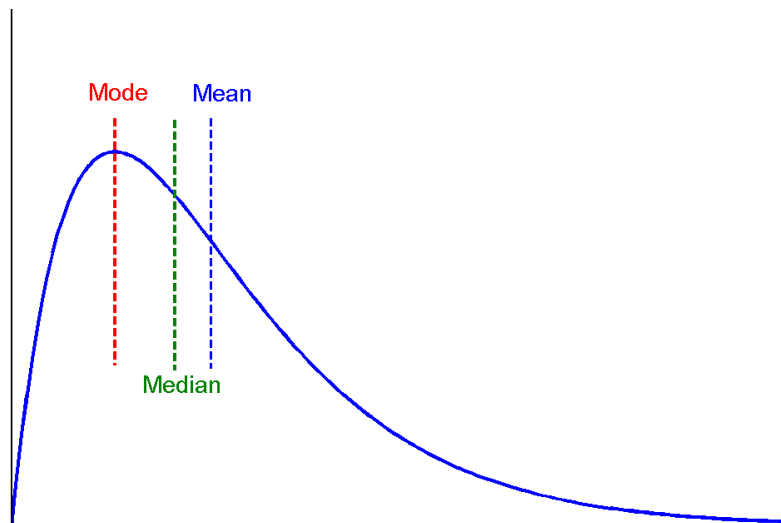
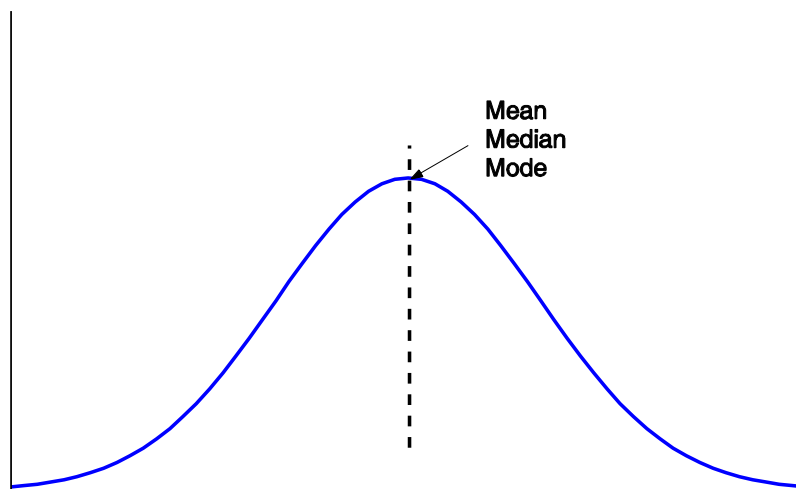
经验公式 $mean - mode = 3 \times (mean - median)$



2.1 中性化趋势度量：均值、中位数和众数

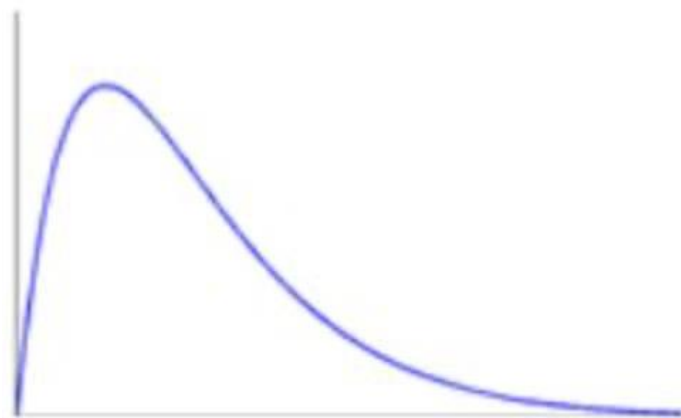
36

- 中位数，均值和对称模式，正面和负面的偏斜数据



中位数、平均数、众数三者的关系

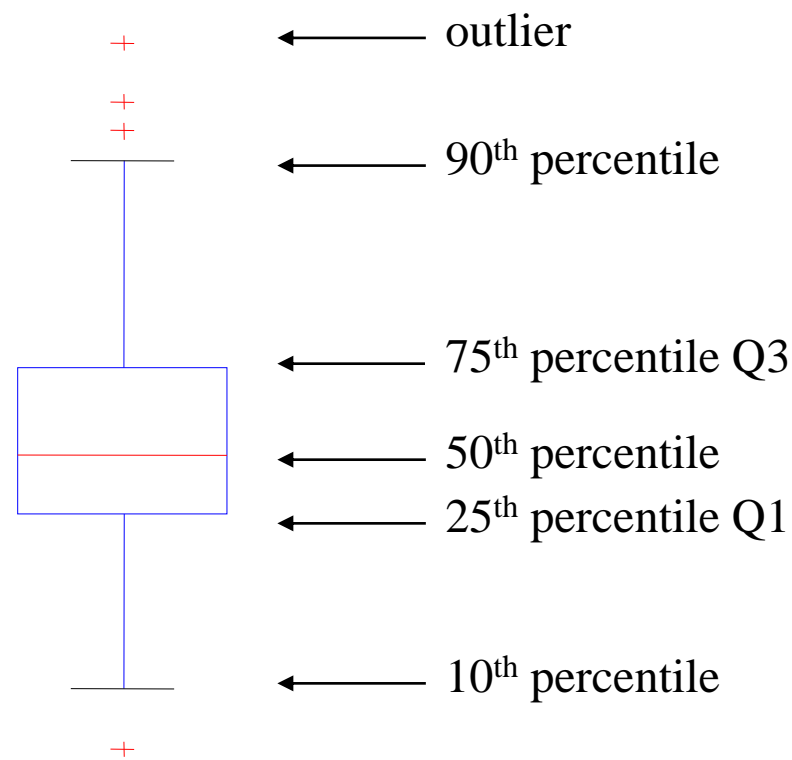
- ☐ A 中位数=平均数=众数
- ☐ B 中位数>平均数>众数
- ☒ C 平均数>中位数>众数
- ☐ D 中位数<平均数<众数



提交



- 四分位数
 - 四分位数: Q1 (第25百分位), Q3 (第75百分位)
 - 四分位数极差: $IQR = Q3 - Q1$
- 五点概况:
 - min, Q1, median, Q3, max
- 盒状图 (boxplot) : min, Q1, median, Q3, max; 单独添加胡须表示离群点
- 离群点: 通常情况下, 一个值高于/低于 $1.5 \times IQR$



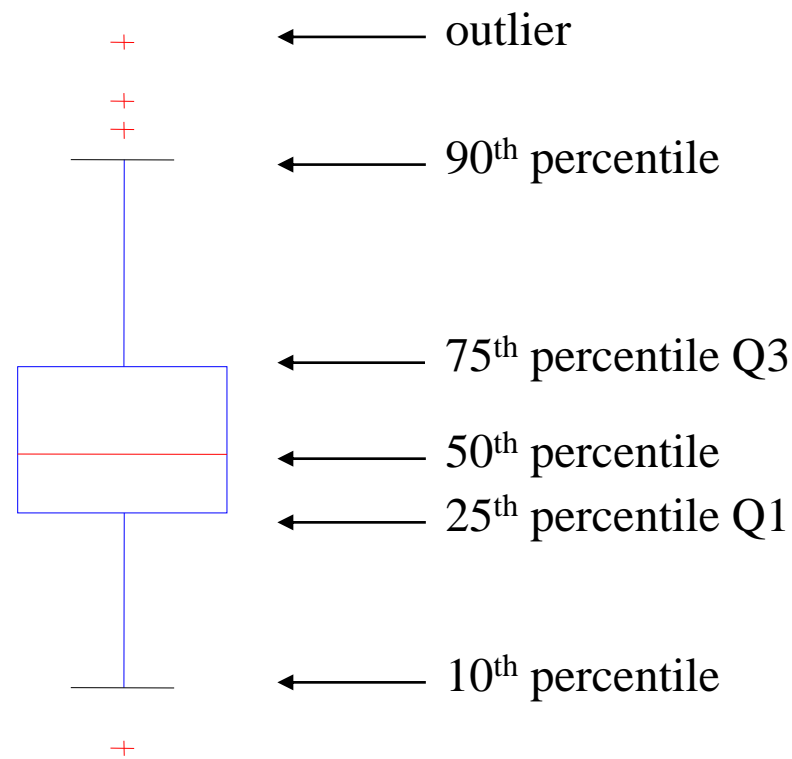
$$IQR = Q3 - Q1$$

$$\max = Q3 + 1.5 \times IQR$$

$$\min = Q1 - 1.5 \times IQR$$



- **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： $IQR = Q3 - Q1$
- **五点概况：**
 - min, Q1, median, Q3, max
- **盒状图（boxplot）：** min, Q1, median, Q3, max；单独添加胡须表示离群点
- **离群点：** 通常情况下，一个值高于/低于 $1.5 \times IQR$



$$IQR = Q3 - Q1$$

$$\max = Q3 + 1.5 * IQR$$

$$\min = Q1 - 1.5 * IQR$$



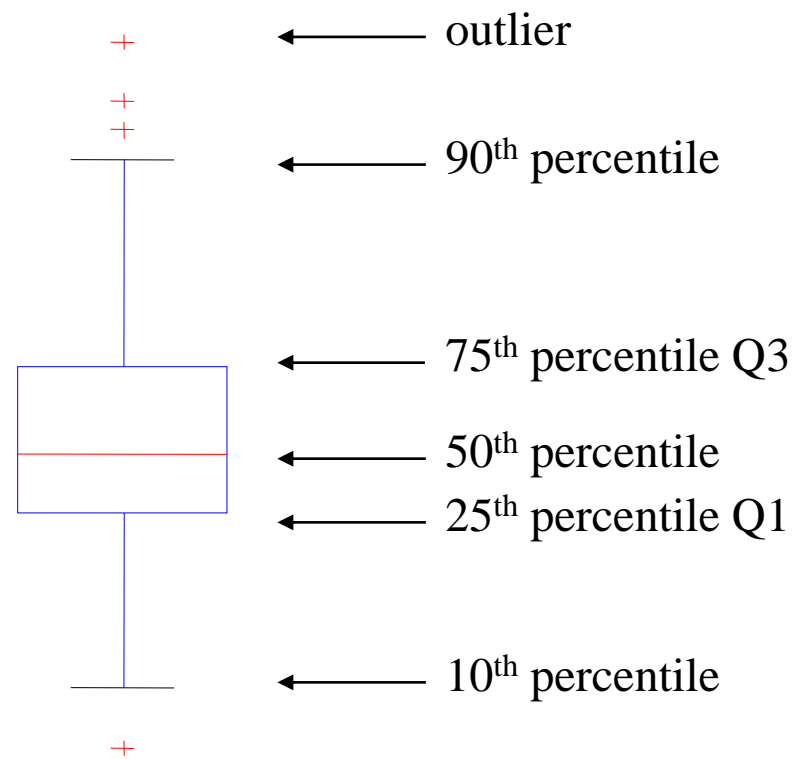
- 现在一家电商公司要卖两个同类型的商品，它们的一周销量（单位：个）如下：
 - 商品A：10, 10, 10, 11, 12, 12, 12
 - 商品B：3, 5, 6, 11, 16, 17, 19
- 它们的平均数一样，中位数也一样，可它们的真实情况呢？

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

- 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。



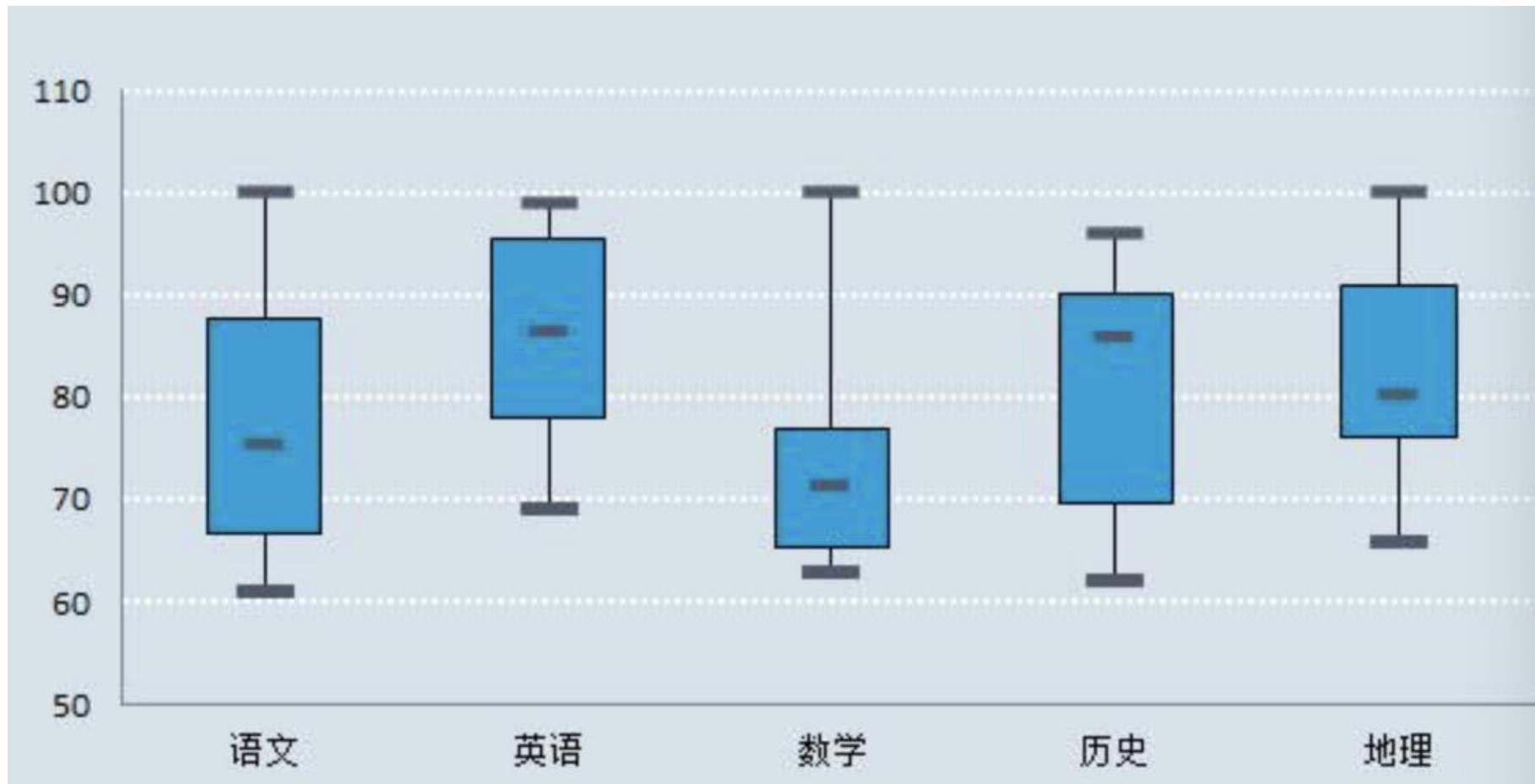
- 方差和标准差
- 分位数
 - 分位数：Q1（第25百分位），Q3（第75百分位）
 - 分位数极差：IQR = Q3 - Q1
- 五点概况：
 - min, Q1, median, Q3, max
- 盒状图（boxplot）：min, Q1, median, Q3, max；单独添加胡须表示离群点
- 离群点：通常情况下，一个值高于/低于 $1.5 \times \text{IQR}$



$$\text{IQR} = Q3 - Q1$$

$$\text{max} = Q3 + 1.5 * \text{IQR}$$

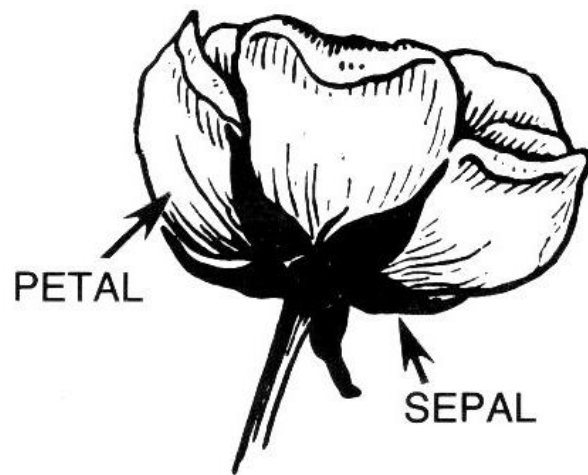
$$\text{min} = Q1 - 1.5 * \text{IQR}$$



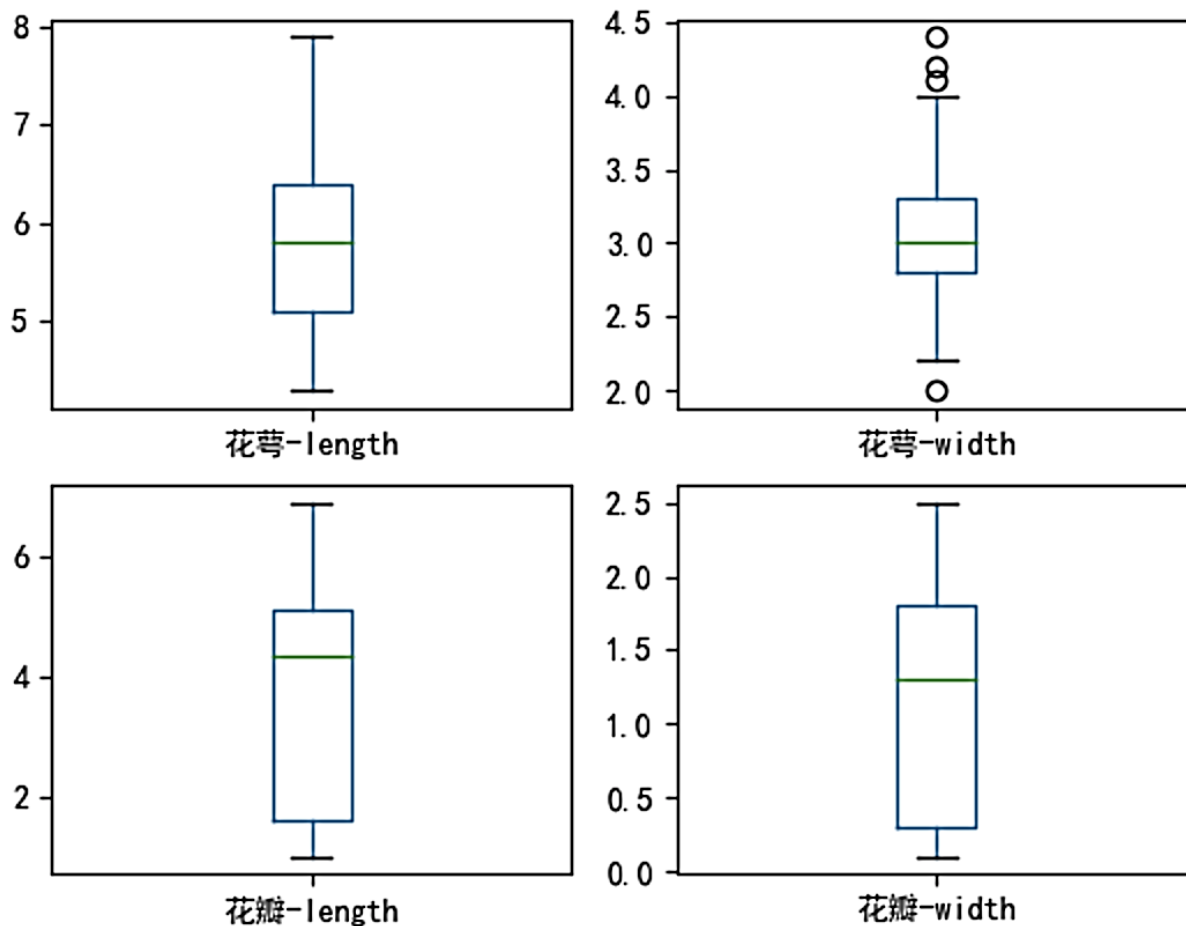
盒状图能够分析**多个属性数据**的**离散度差异性**

- 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

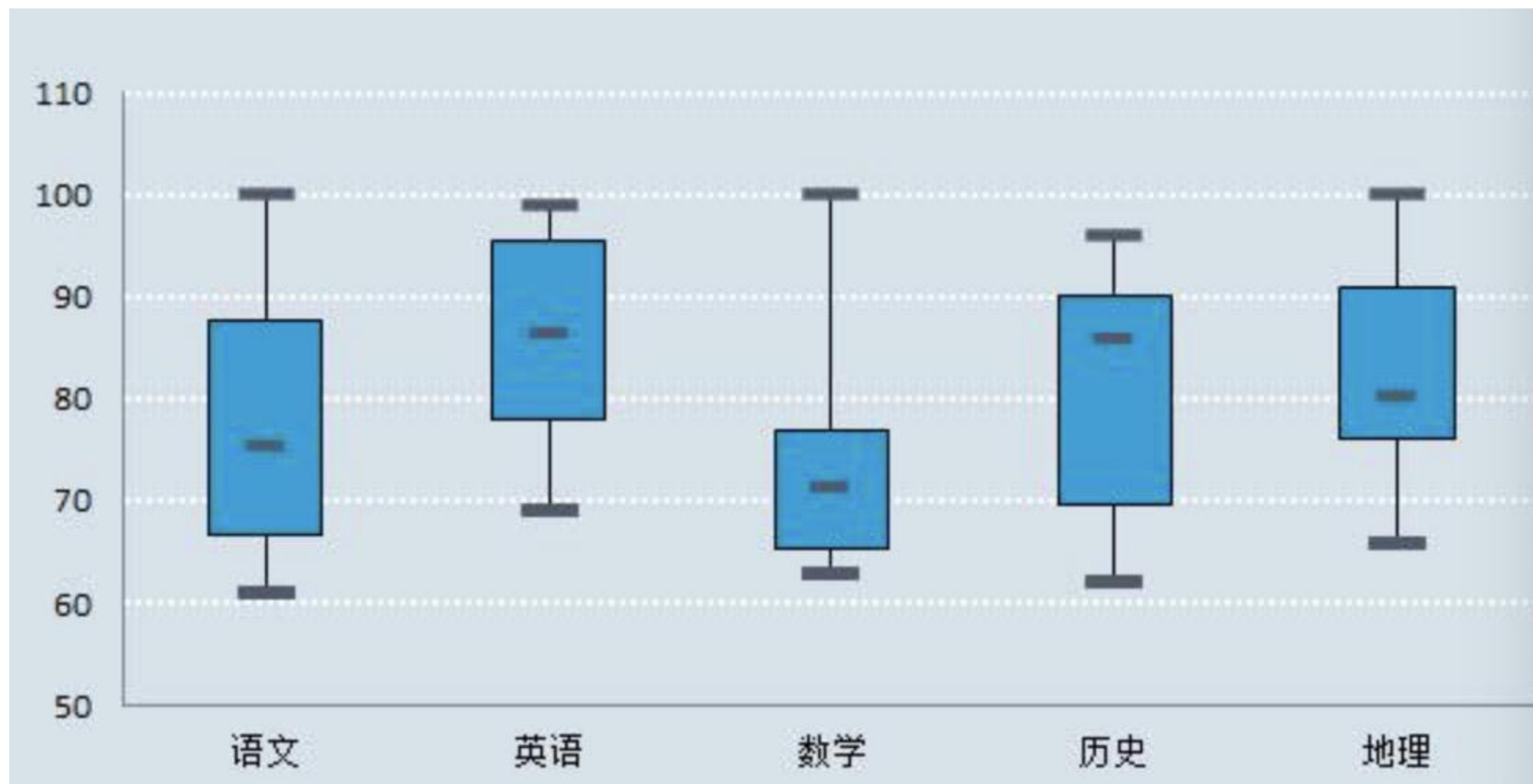


```
6.7,3.0,5.2,2.3,Iris-virginica
6.3,2.5,5.0,1.9,Iris-virginica
6.5,3.0,5.2,2.0,Iris-virginica
6.2,3.4,5.4,2.3,Iris-virginica
5.9,3.0,5.1,1.8,Iris-virginica
5.1,3.8,1.6,0.2,Iris-setosa
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor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



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

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



盒状图能够分析**多个属性数据**的**离散度差异性**

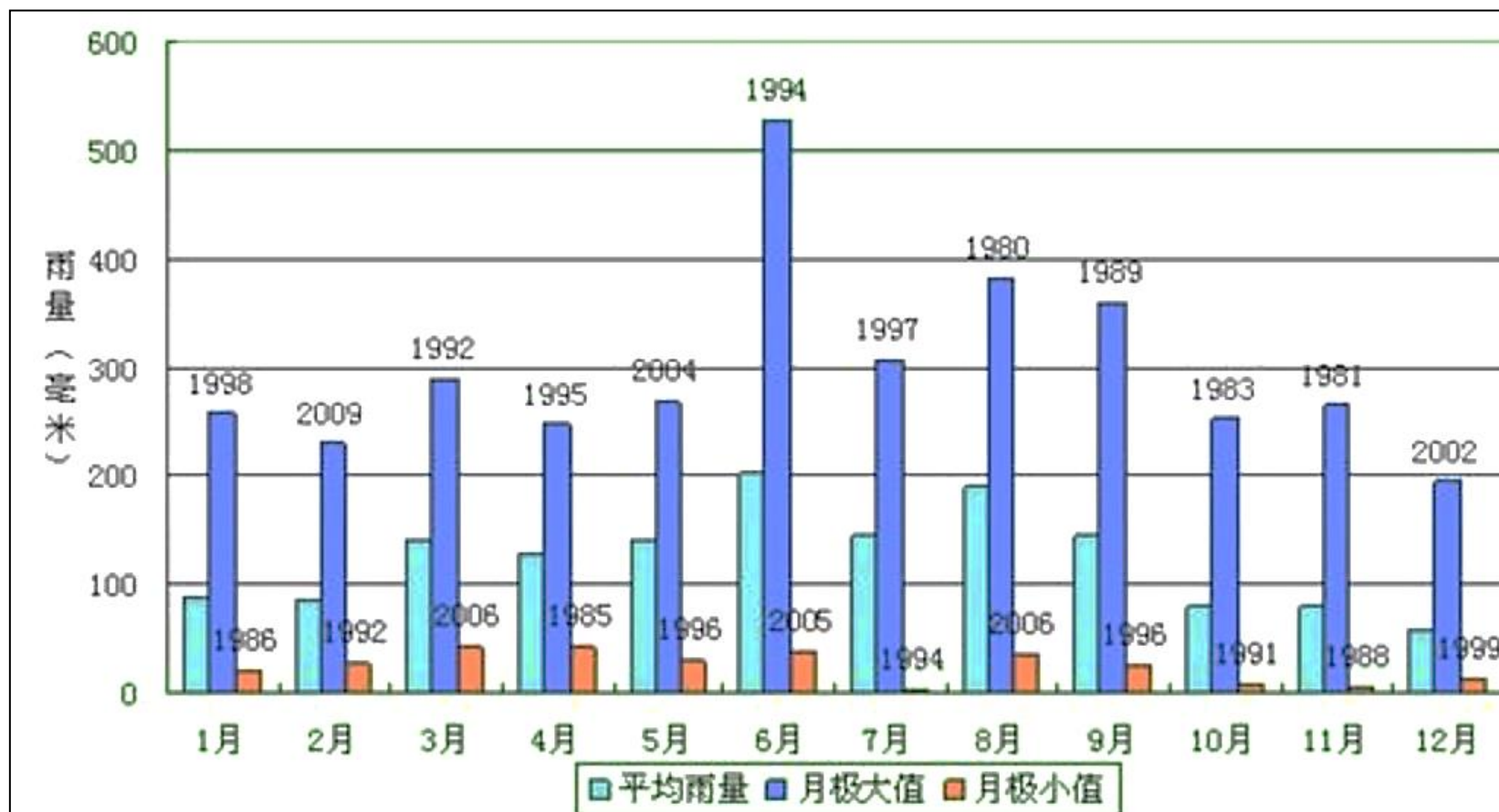
如果希望分析**单个属性**在**各个区间**的**变化分布**怎么办？

例如：如果希望分析**语文成绩**在每个**分数段**的**变化分布**



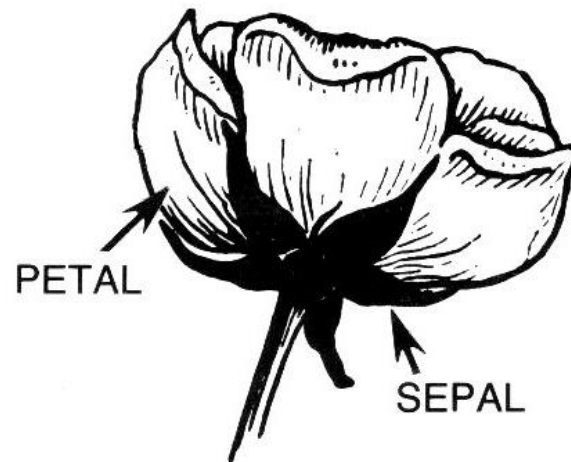
● 直方图

- 用来分析**单个属性**在**各个区间变化分布**

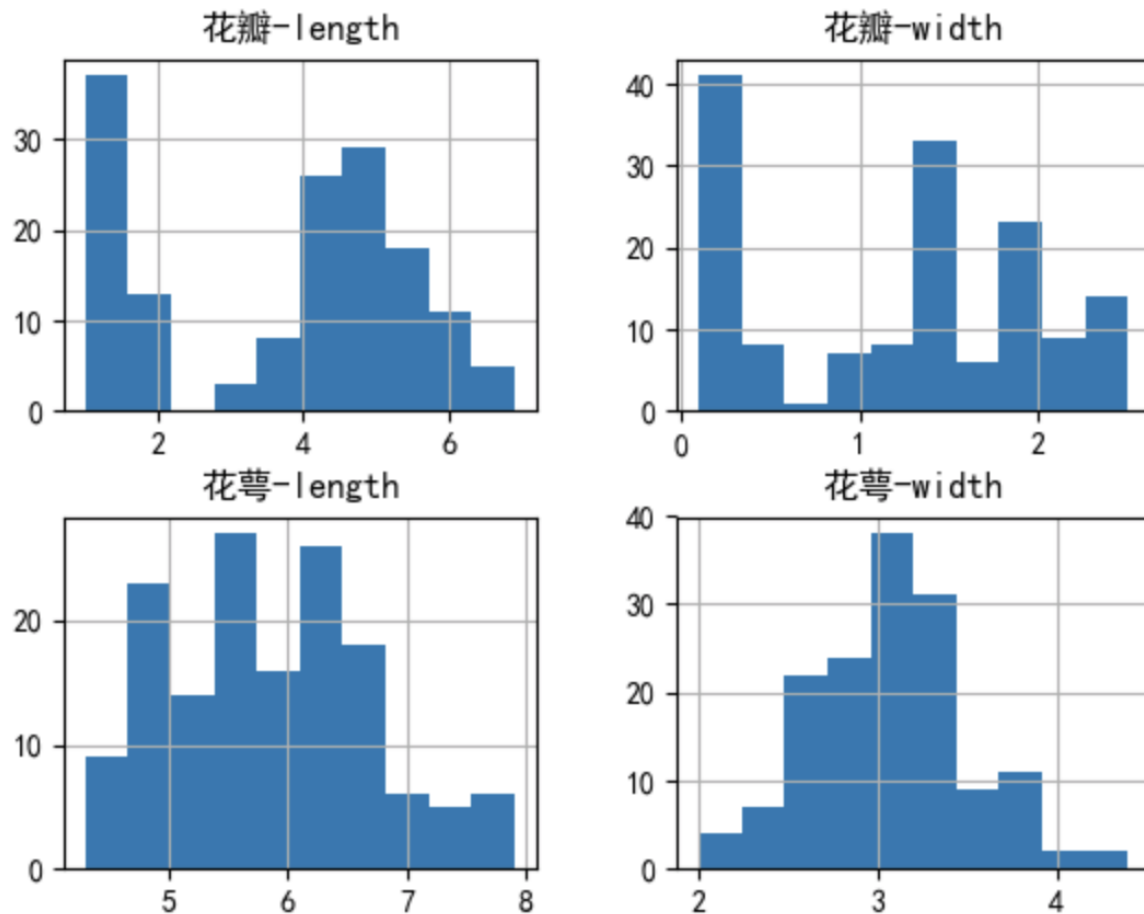


- IRIS (sepal:萼片,petal:花瓣)

- 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
```



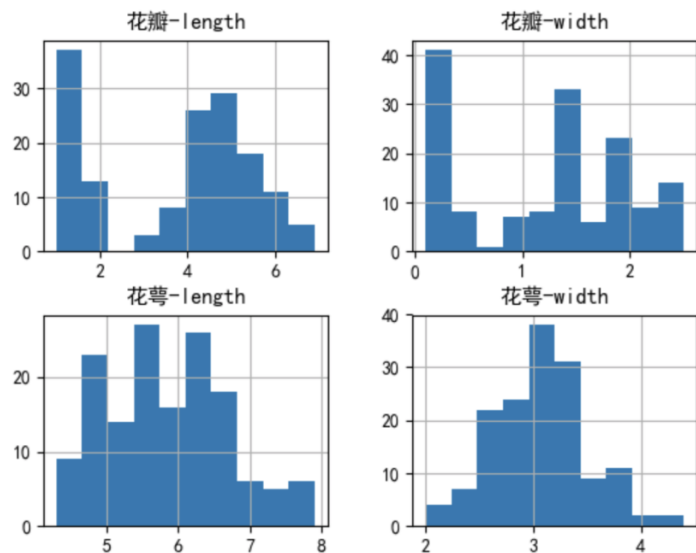
参考: https://blog.csdn.net/H_lukong/article/details/90139700

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



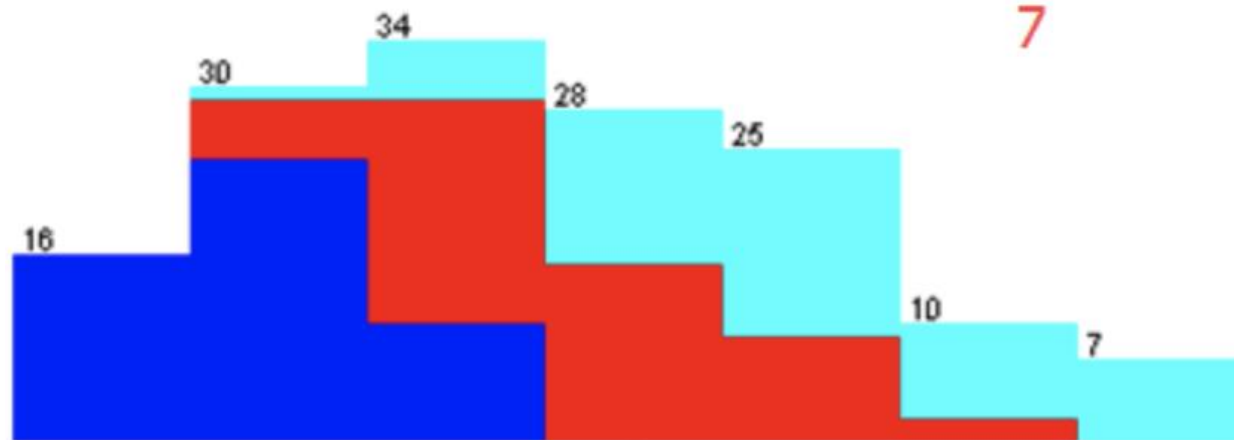
2.3 数据可视化 —— 直方图分析案例

50



题目：分别在一个图中用3种颜色表示单个属性数据（每个属性画一个图，4个直方图）在不同类别下的直方图分布（用python实现画图）

Class: class (Nom) Visualize All

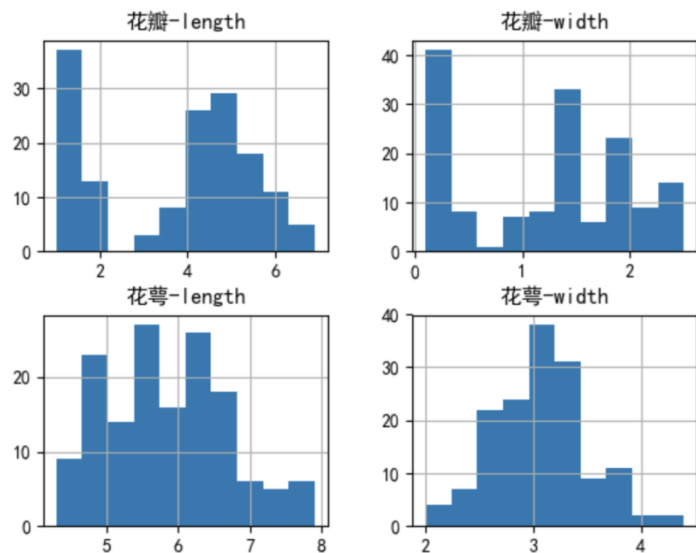


这种直方图有什么用呢？



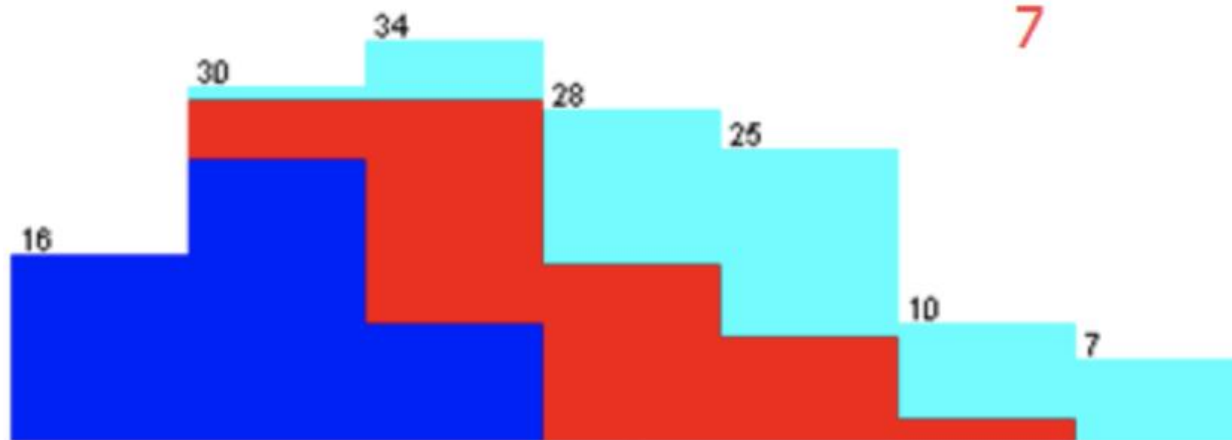
2.3 数据可视化 —— 直方图分析案例

51



题目：分别在一个图中用3种颜色表示单个属性数据（每个属性画一个图，4个直方图）在不同类别下的直方图分布（用python实现画图）

Class: class (Nom) Visualize All



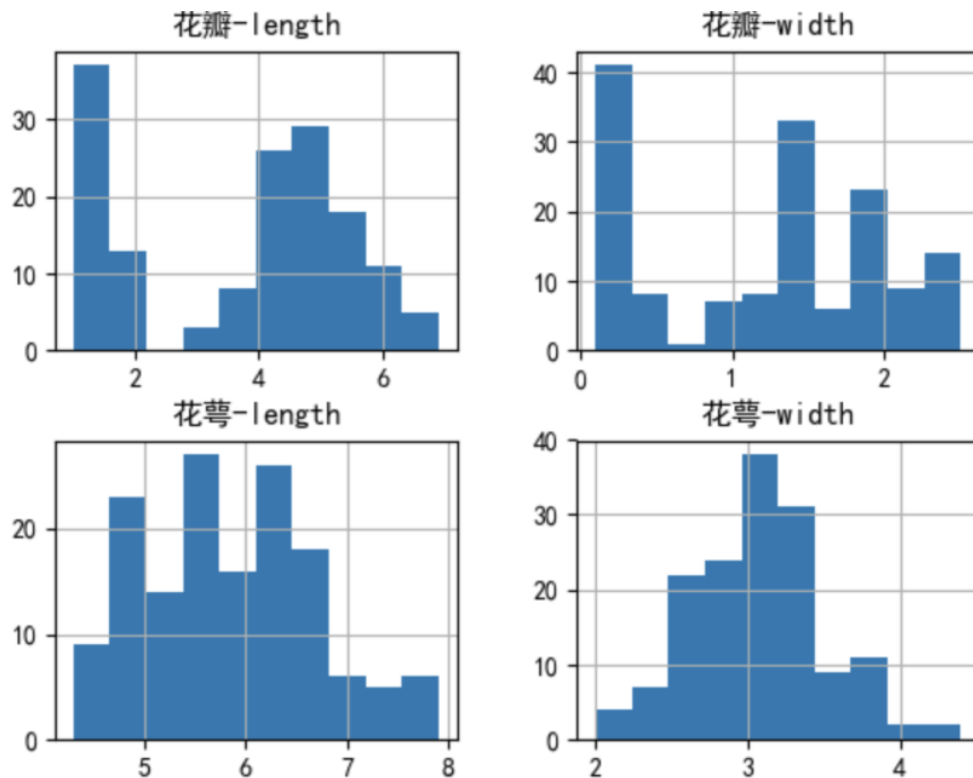
这种直方图有什么用呢？

分类算法特征分析：
我们可以看到花瓣长度在3个类别下的分布具有差异



● 直方图

- 用来分析**单个属性**在**各个区间变化分布**



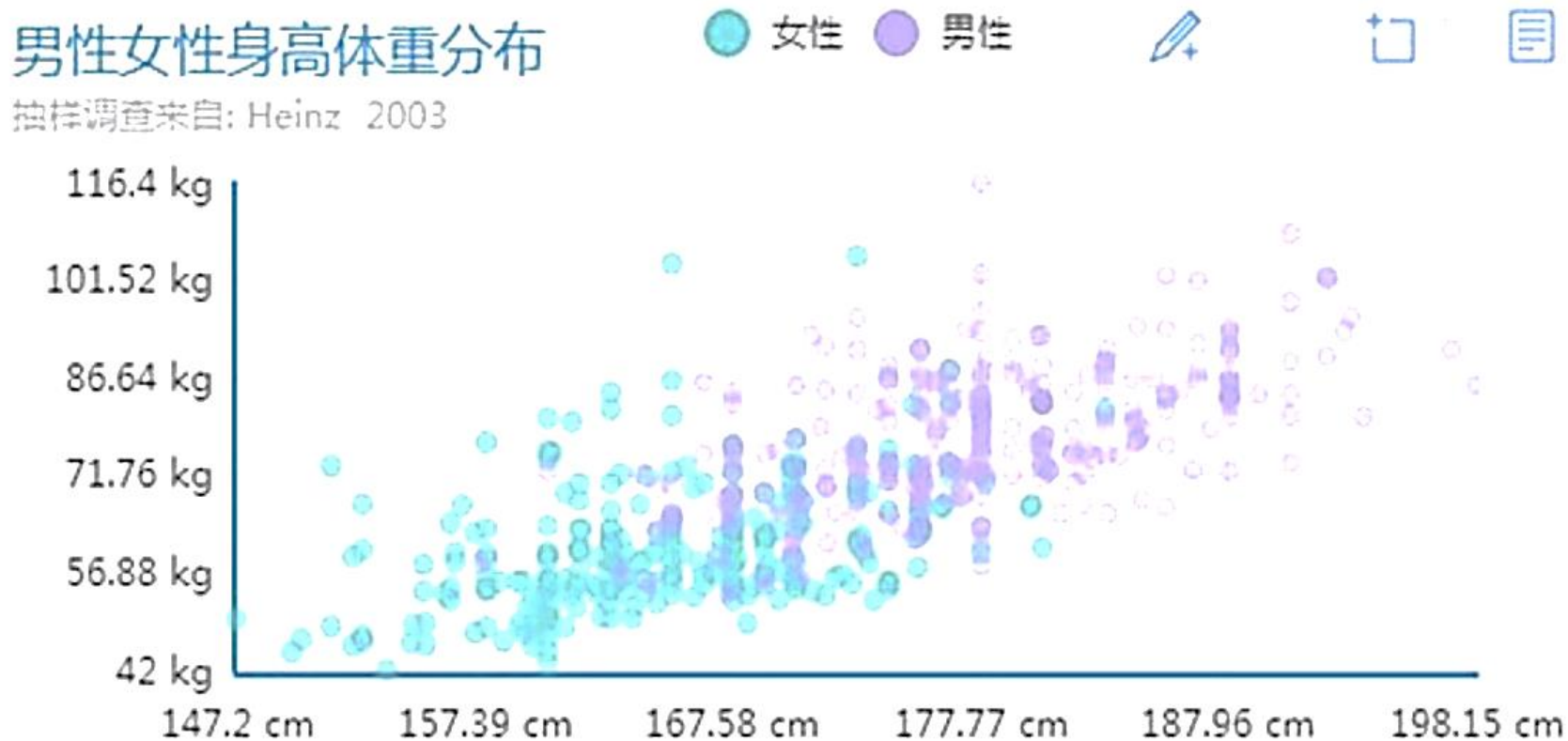
如果希望分析2个属性数据的**关联关系**，怎么办？

例如：如果希望分析**花瓣长度**和**花萼长度**的**关联关系**



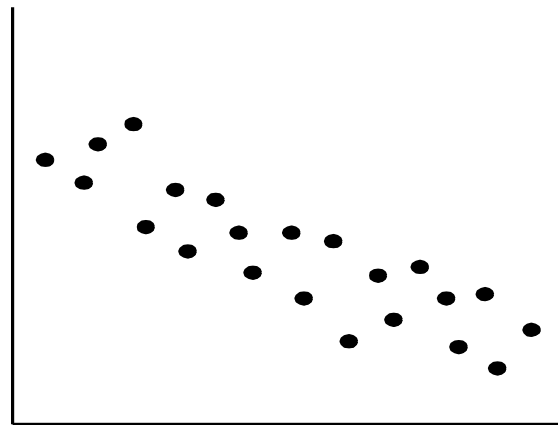
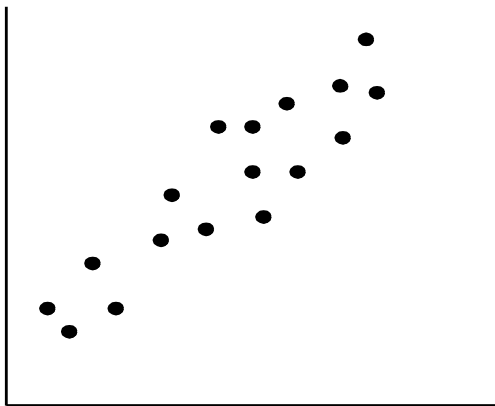
- 散点图

- 用来显示两组数据的相关性分布



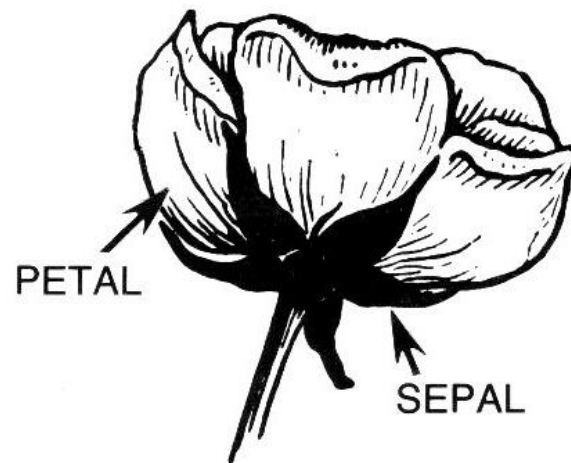


- 散点图
 - 用来显示两组数据的相关性分布



- 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

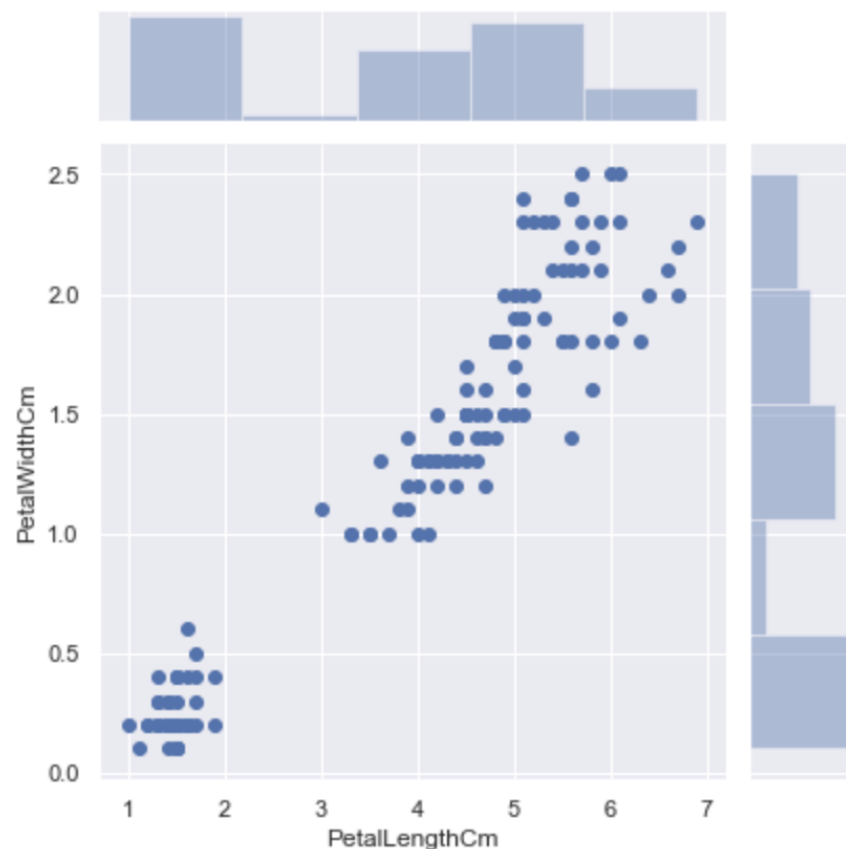
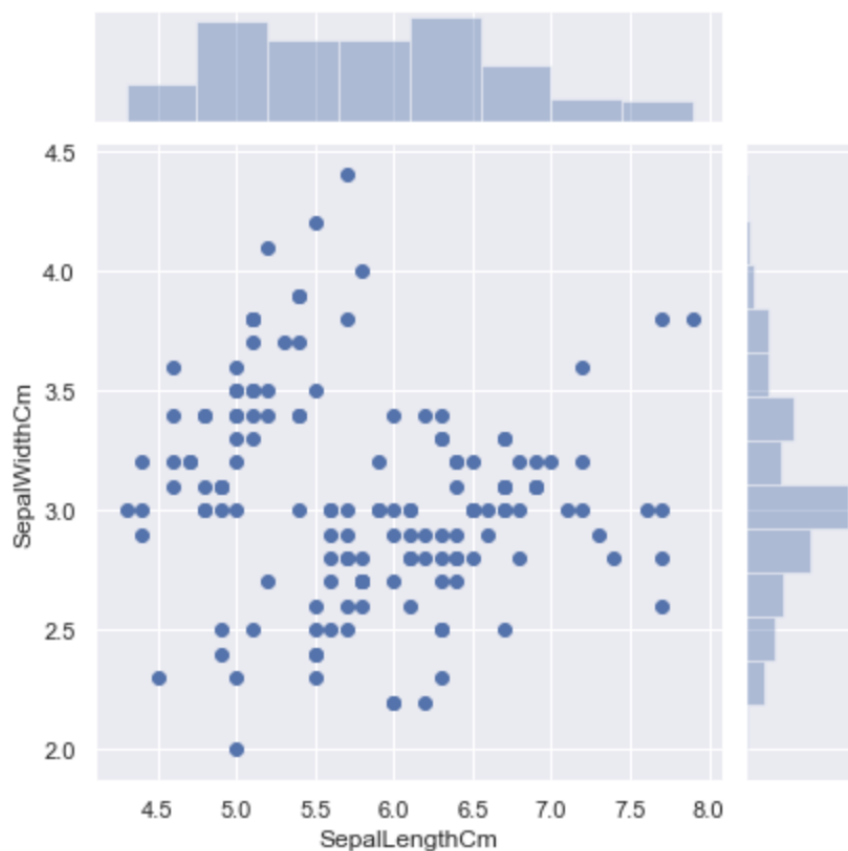


```
6.7,3.0,5.2,2.3,Iris-virginica
6.3,2.5,5.0,1.9,Iris-virginica
6.5,3.0,5.2,2.0,Iris-virginica
6.2,3.4,5.4,2.3,Iris-virginica
5.9,3.0,5.1,1.8,Iris-virginica
5.1,3.8,1.6,0.2,Iris-setosa
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor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



2.3 数据可视化 —— 散点图分析案例

56



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

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



2.3 数据可视化 —— 散点图分析

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

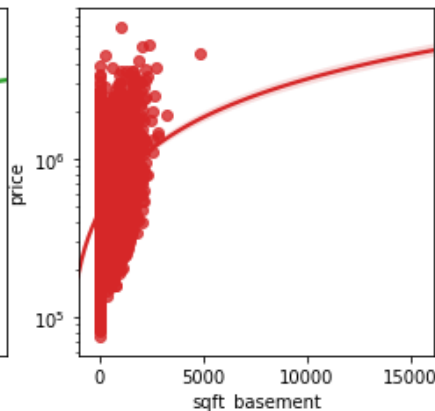
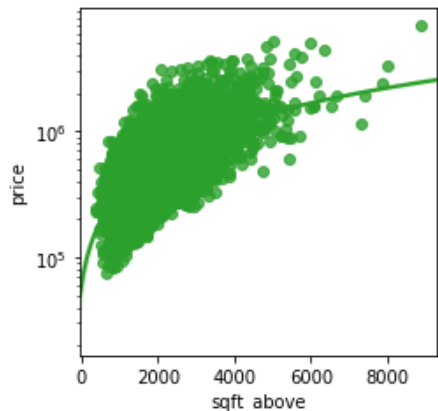
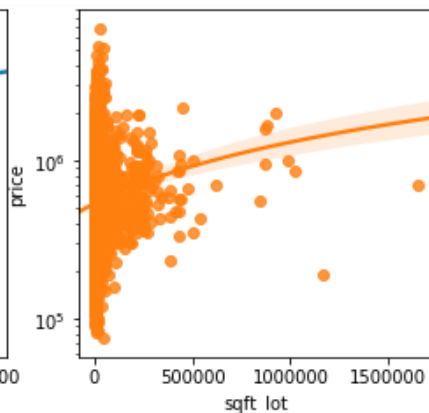
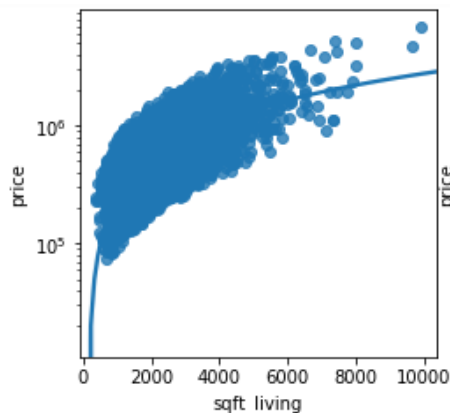
基本idea：哪些属性跟房价相关



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。



- 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**



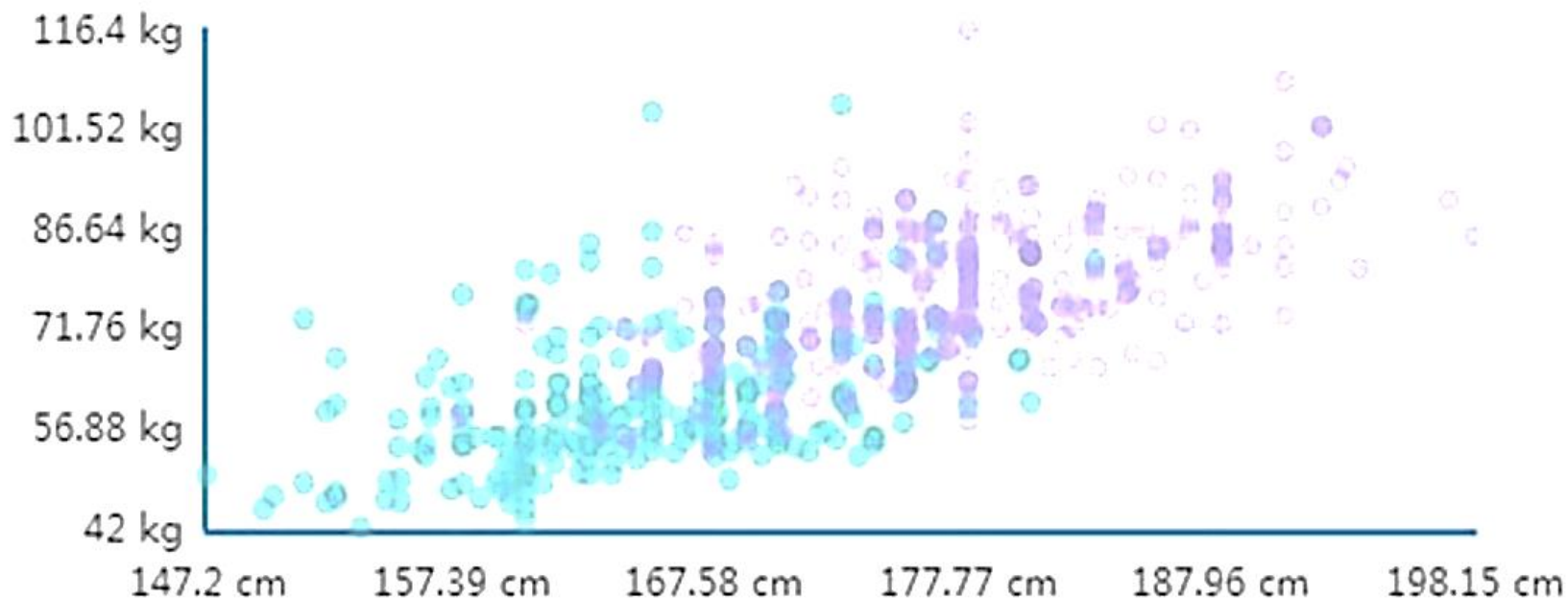
- 散点图

- 用来显示两组数据的相关性分布

男性女性身高体重分布

抽样调查来自: Heinz 2003

● 女性 ● 男性



关于盒状图、直方图、散点图，下面说法正确的是：

- ☒ A 盒状图能够分析多个属性数据的离散度差异性
- ☒ B 直方图用来分析单个属性在各个区间变化分布
- ☒ C 散点图用来显示两组数据的相关性分布
- ☐ D 直方图用于展示不同类别的数据分布

提交

数据挖掘

Data Mining



01

数据类型

02

数据统计汇总

03

数据相似性和
相异性度量



- **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



● 数据矩阵

- N个数据, p个维度

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

● 相异矩阵

- N个数据点, 记录两点之间的距离

- 三角矩阵

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



- 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

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

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

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



3.3 二值属性的邻近性度量

66

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

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



3.3 二值属性的邻近性度量

67

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

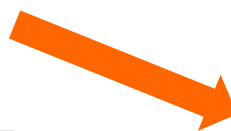
- 一个邻接表

		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


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



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



3.3 二值属性的邻近性度量

69

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)

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

● 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

$$d(jack, mary) = \text{[填空1]}$$

$$d(jack, jim) = \text{[填空2]}$$

$$d(jim, mary) = \text{[填空3]}$$

作答



- 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

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

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

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



3.4 数值属性的邻近性度量

72

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



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)
- $d(i, j) \leq d(i, k) + d(k, j)$ (三角不等性)



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

- $h = 1$: 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”.

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

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]

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

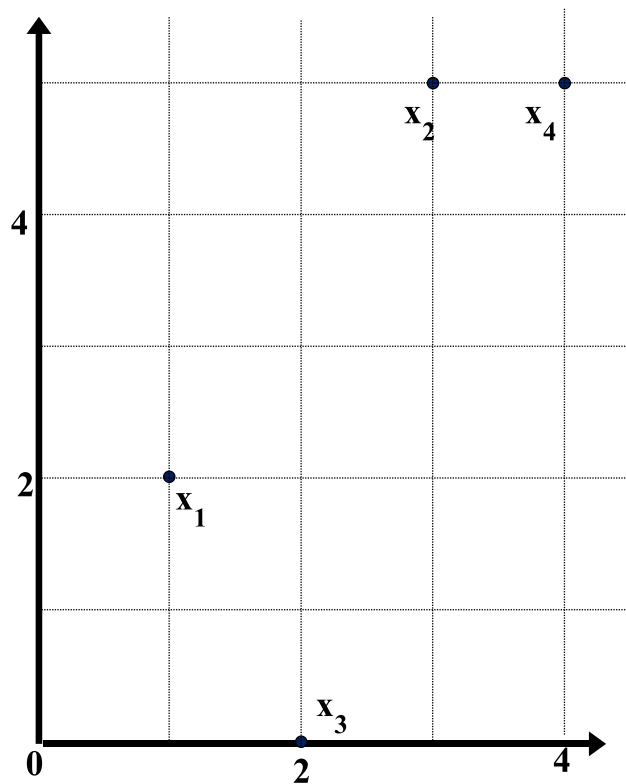
作答



3.4 数值属性的邻近性度量

76

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

- 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

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

● 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

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

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

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

- 求这两篇文档的余弦相似性： [填空1]

作答



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

● 例如：

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

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

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

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

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

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



案例1 广州白云机场流量预测

● 竞赛题目

- 机场拥有巨大的旅客吞吐量，与巨大的人员流动相对应的则是巨大的服务压力。安防、安检、突发事件应急、值机、行李追踪等机场服务都希望能够预测未来的旅客吞吐量，并据此提前调配人力物力，更好的为旅客服务。
- 本次大赛以广州白云机场真实的客流数据为基础，每天数万离港旅客在机场留下百万级的数据记录。希望参赛队伍通过数据算法来构建客流量预测模型。
- 选手需要预测未来三小时（9月25日15:00:00到18:00）的时间窗口里，机场内每个WIFI AP点每10分钟内的平均设备连接数量。



案例1 广州白云机场流量预测

● 竞赛数据

三个字段分别为：wifi_ap_tag，字符串，描述WIFI接入的AP点；passenger_count，整数，描述在某一时刻接入该WIFI AP的设备数量；time_stamp，字符串，描述该时刻（精确到秒）。下面的例子是这张表的实际数据：

wifi_ap_tag	passenger_count	time_stamp	
E1-1A-1<E1-1-01>	15	2016-09-10-18-55-04	
E1-1A-2<E1-1-02>	15	2016-09-10-18-55-04	
E1-1A-3<E1-1-03>	38	2016-09-10-18-55-04	
E1-1A-4<E1-1-04>	19	2016-09-10-18-55-04	

我们关注的机场区域，是E1,E2,E3,EC,T1,W1,W2,W3,WC。E和W分别代表机场的东侧和西侧登机区，各有三个小区；EC和WC是指连接安检区和登机区的走廊。分布见下图：WiFi_ap_tag的前四位字符，代表了该WIFI AP所在区域和楼层，例如名称的前四位“T1-1”意味着在T1航站楼的1楼。



案例2

新浪微博转发量预测

● 竞赛题目

对于一条原创博文而言，转发、评论、赞等互动行为能够体现出用户对于博文内容的兴趣程度，也是对博文进行分发控制的重要参考指标。本届赛题的任务就是根据抽样用户的原创博文在发表一天后的转发、评论、赞总数，建立博文的互动模型，并预测用户后续博文在发表一天后的互动情况。



数据挖掘

Data Mining

Any Questions?

谢谢!