目　　录

第一章 探索性数据分析简介

1. 引言

- 数据就是承载了信息的东西。
- 例如，数字、文本、图形、音频、视频、网页等。
- 对数据进行观察、研究，寻找其蕴含的规律，这就是数据分析。
- 数据分析的目的就是在本来彼此错综复杂的或者大量看似不相关的数据之间找到内在的、本质的、起作用的规律或特性。

2. 探索性数据分析的定义

- 探索性数据分析(Explorative Data Analysis, EDA)就是在较少预设或没有预设的前提下对数据进行分析。

- 探索性数据分析是对已有数据在尽量少的先验假设下通过作图、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。

3. 常用的数据变换

➢ 数据变换、预处理的目的

- 对于数据问题而言，在进行数据分析前，一般都需要对数据进行预处理，主要有 2 个目的：

（1）**无量纲化：**不同属性的数据往往具有不同的量纲，即使对同一属性，采用不同的计量单位，其数值也不同。因此，数据分析前，需对数据进行无量纲化

（2）**归一化：**不同属性的数据，其取值在数值大小有很大的差异，因此需要对数据进行预处理，将不同属性的数据变换到可比较的数值大小，可通过归一化处理，变换到[0, 1]

3. 常用的数据变换

● **线性变换**

对于原始数据矩阵为$A=(a_{ij})_{m\times n}$，其每一列代表不同的数据属性，

$i=1,\cdots,m$，$j=1,\cdots,n$。设$a_j^{\max}$是矩阵第j列中的最大值，则

$$b_{ij}=a_{ij}/a_j^{\max}$$

● **标准 0 - 1 变换**

与线性变换类似，也可以对于原始数据矩阵为$A=(a_{ij})_{m\times n}$，进行

标准 0 - 1 变换，$a_j^{\max}$和$a_j^{\min}$分别是矩阵第j列中的最大值和最

小值，则

$$b_{ij}=\frac{a_{ij}-a_j^{\min}}{a_j^{\max}-a_j^{\min}},$$

3. 常用的数据变换

- **规范化处理**

无论成本型属性还是效益型属性，向量规范化均用下式进行变换

$$b_{ij} = a_{ij} \Big/ \sqrt{\sum_{i=1}^{m} a_{ij}^2}\ ,\quad i=1,\cdots,m\ ,\quad j=1,\cdots,n.$$

- **标准化处理**

在实际问题中，每个变量都具有同等的表现力，可对数据进行标准化处理，即

$$b_{ij} = \frac{a_{ij} - \overline{a}_j}{s_j},\quad i=1,2,\cdots,m,\quad j=1,2,\cdots,n$$

第二章 统计分析

1. 一维数据的统计分析

2. 多维数据的统计分析

1. 一维数据的统计分析

- **(1) 表示位置水平的数字特征——样本均值：**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **(1) 表示位置水平的数字特征——中位数：**

将 $x_1, x_2, \cdots, x_n$ **从小到大排序为** $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$

样本中位数的位置： $\frac{n+1}{2}$

- **(1) 表示位置水平的数字特征——样本百分位数：**

将 $x_1, x_2, \cdots, x_n$ **从小到大排序为** $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$

样本百分位数的位置： $q = \frac{p}{100}(n+1)$ $\quad\quad p = 1, 2\ldots..99$

其中： p **为分位**

1. 一维数据的统计分析

- **(1) 表示位置水平的数字特征——上、下四分位数：**

75分位数与25分位数分别称为上、下四分位数，并记为

- **(2) 表示波动（分散）的统计特征——样本方差：**

样本方差：

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- **(2) 表示波动（分散）的统计特征——样本极差：**

将 $x_1, x_2, \cdots, x_n$ **从小到大排序为** $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$

样本极差：

$$R = x_{(n)} - x_{(1)}$$

1. 一维数据的统计分析

- (2) 表示波动（分散）的统计特征——**样本四分位极差**：

将 $x_1, x_2, \cdots, x_n$ 从小到大排序为 $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$

样本四分位极差：

$$R_1 = Q_3 - Q_1$$

- (2) 表示波动（分散）的统计特征——**下、上截断点**：

下、上截断点：

$$Q_1 - 1.5R_1$$

$$Q_3 + 1.5R_1$$

- (2) 表示波动（分散）的统计特征——**变异系数**：

变异系数：CV=100*标准差/样本均值(%)

1. 一维数据的统计分析

- (3)表示形状的统计特征——样本偏度：

样本偏度：

$$g_1 = \frac{n}{(n-1)(n-2)} \frac{1}{s^3} \sum_{i=1}^{n} (x_i - \overline{x})^3$$

偏度大于0，右偏（正偏）；

偏度小于0，左偏（负偏）；

偏度等于0，数据分布左右对称。

- (3)表示形状的统计特征——样本峰度：

样本峰度：

$$g_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{1}{s^4} \sum_{i=1}^{n} (x_i - \overline{x})^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

峰度大于0，有较多远离均值的极端数值；

峰度小于0，均值两侧的极端数值较少；

峰度等于0，数据分布为正态分布。

2. 多维数据的统计分析

➢ (1) 多维数据的均值向量：

总体X的观测值为 $X_i = (x_{i1}, x_{i2}, \cdots, x_{im})^T, i = 1, 2, \cdots, n$

样本观测矩阵

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} = \begin{pmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{pmatrix}$$

2. 多维数据的统计分析

➢ (1) 多维数据的均值向量:

总体X的观测值为 $X_i = (x_{i1}, x_{i2}, \cdots, x_{im})^T, i = 1, 2, \cdots, n$

样本均值向量

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \left(\frac{1}{n}\sum_{i=1}^{n} x_{i1}, \frac{1}{n}\sum_{i=1}^{n} x_{i2}, \cdots, \frac{1}{n}\sum_{i=1}^{n} x_{im}\right)^T \hat{=} (\overline{x}_1, \overline{x}_2, \cdots, \overline{x}_m)^T$$

➢ (2) 多维数据的协方差阵:

总体X的观测值为 $X_i = (x_{i1}, x_{i2}, \cdots, x_{im})^T, i = 1, 2, \cdots, n$

样本协方差阵

$$\hat{\Sigma} = \begin{pmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1m} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22} & \cdots & \hat{\sigma}_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\sigma}_{m1} & \hat{\sigma}_{m1} & \cdots & \hat{\sigma}_{mm} \end{pmatrix} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^T$$

2. 多维数据的统计分析

➢ **(2) 多维数据的协方差阵：**

$$\hat{\sigma}_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j), i, j = 1, 2, \cdots, m$$

➢ **(3) 多维数据的相关阵：**

总体X的观测值为 $X_i = (x_{i1}, x_{i2}, \cdots, x_{im})^T, i = 1, 2, \cdots, n$

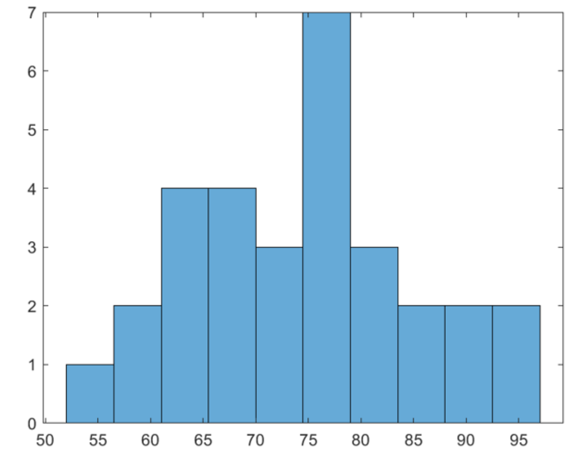
样本相关阵

$$\hat{R} = \begin{pmatrix} 1 & \hat{\rho}_{12} & \cdots & \hat{\rho}_{1m} \\ \hat{\rho}_{21} & 1 & \cdots & \hat{\rho}_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\rho}_{m1} & \hat{\rho}_{m1} & \cdots & 1 \end{pmatrix} \qquad \hat{\rho}_{ij} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii} \times \hat{\sigma}_{jj}}}, i, j = 1, 2, \cdots, m$$

第三章 数据可视化

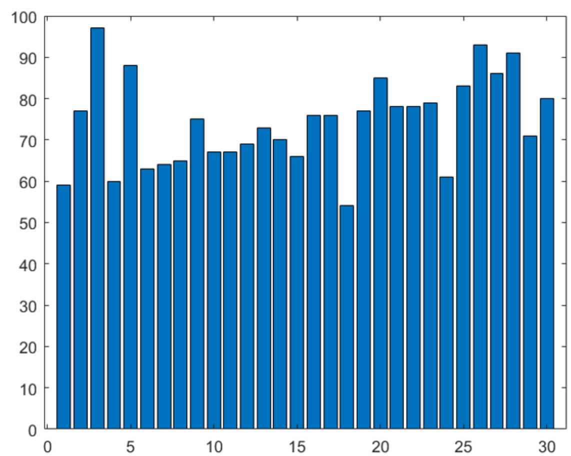
1. 直方图
2. 条形图/柱状图
3. 散点图
4. 饼状图
5. 箱型图

1. 直方图



直方图

2. 条形图/柱状图



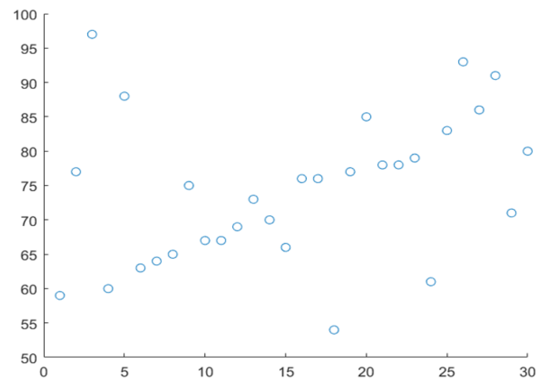
条形图

例2-1.随机抽取30名大学生，得到某课程的考试分数数据如下：

59,77,97,60,88,63,64,65,75,67,67,69,73,70,66,76,76,54,77,85,78,78,79,61,83,93,86,91,71,80.

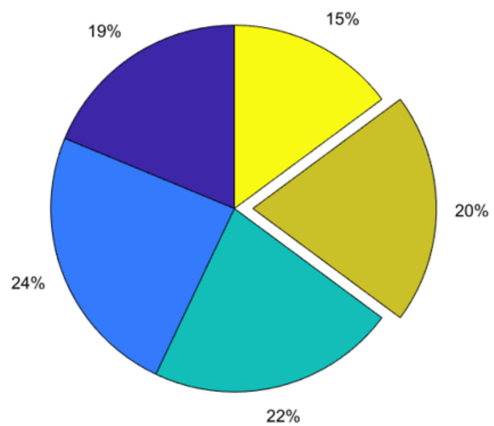
给出30名学生成绩对比的条形图

3. 散点图



散点图

4. 饼状图



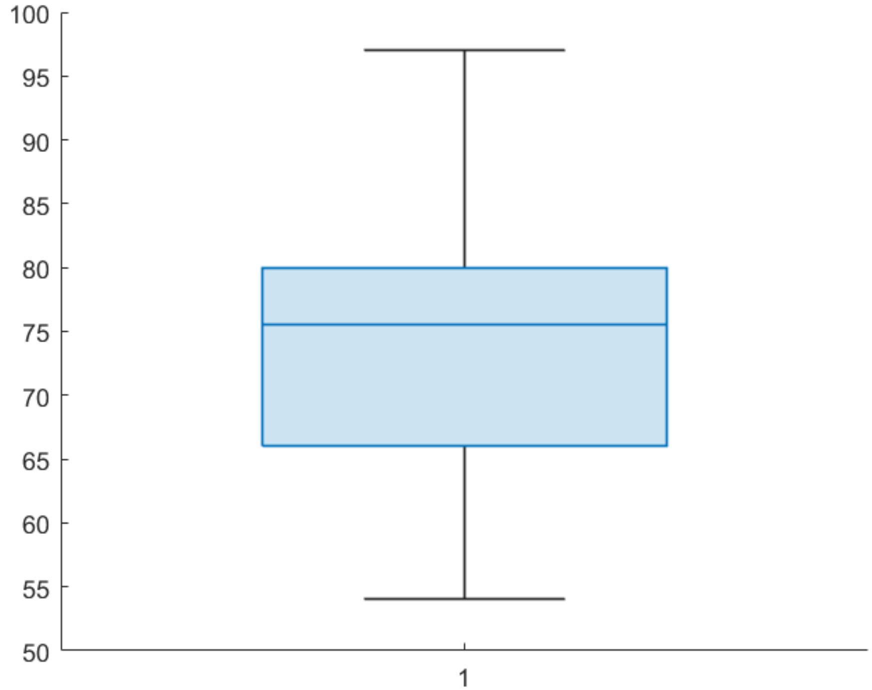
饼形图

例2-1.随机抽取30名大学生，得到某课程的考试分数数据如下：

59,77,97,60,88,63,64,65,75,67,67,69,73,70,66,76,76,54,77,85,78,78,79,61,83,93,86,91,71,80.

给出30名学生成绩对比的条形图

5. 箱型图



箱型图

例2-1.随机抽取30名大学生，得到某课程的考试分数数据如下：

59,77,97,60,88,63,64,65,75,67,67,69,73,70,66,76,76,54,77,85,78,78,79,61,83,93,86,91,71,80.

给出30名学生成绩对比的条形图

第四章 方差分析

4.1 单因素方差分析

4.2 多因素方差分析

Variance analysis

- 在科学试验和工农业生产中，常常需要分析哪几种因素对产品的产量或质量有显著性影响，并希望知道这些因素处于什么条件时生产状态最佳。

4.1 单因素方差分析

问 题 的 提 出

表 4-1

A_1	A_2	A_3	A_4
A_2	A_3	A_4	A_1
A_3	A_4	A_1	A_2
A_4	A_1	A_2	A_3

表 4-2

肥料种类 A_i	收 获 量 X_{ij}				平均产量 $\overline{X}_i$
A_1	98	96	91	66	87.75
A_2	60	69	50	35	53.50
A_3	79	64	81	70	73.50
A_4	90	70	79	88	81.75

4.1 单因素方差分析

建 立 模 型

由前所述得方差分析的一般数学模型：

$$\begin{cases} y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1,2,\cdots,k, \quad j = 1,2,\cdots,r \\ \sum_{i=1}^{k} \alpha_i = 0 \\ \varepsilon_{ij} \text{独立同分布} N(0,\sigma^2) \end{cases} \qquad (4\text{–}3)$$

μ, α_i $(i = 1,2,3,\cdots,k)$ 是未知参数.要解决的问题是：

（1）估计未知参数 μ， α_i （$i = 1,2,3,\cdots$， k）；

（2）考察 k 个因子水平对试验结果的影响有无显著差异。即检验 $H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$。

参 数 估 计

记 $\bar{y}=\frac{1}{kr}\sum_{i=1}^{k}\sum_{j=1}^{r}y_{ij}$， $\bar{y}_i=\frac{1}{r}\sum_{j=1}^{r}y_{ij}$，

因此μ，α_i的矩估计为

$$\hat{\mu}=\bar{y}，\quad \hat{\alpha}_i=\bar{y}_i-\bar{y}，\quad i=1,2,3,\cdots,\ k \tag{4-4}$$

统 计 检 验

$$S_{总}^2 = \sum_{i=1}^{k}\sum_{j=1}^{r} y_{ij}^2 - \frac{T_{..}^2}{kr} \qquad T_{..} = \sum_{i=1}^{k}\sum_{j=1}^{r} y_{ij}$$

$$S_{总}^2 = S_{误}^2 + S_{组间}^2 \qquad S_{组间}^2 = \sum_{i=1}^{k}\frac{T_{i.}^2}{r} - \frac{T_{..}^2}{rk} \quad (T_{i.} = \sum_{j=1}^{r} y_{ij})$$

$$S_{误}^2 = S_{总}^2 - S_{组间}^2$$

记$n = kr$，可证当H_0成立时

$$F = \frac{S_{组间}^2/(k-1)}{S_{误}^2/(n-k)} \sim F(k-1,\ n-k) \qquad (4\text{-}8)$$

因此可利用F检验来检验$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$，

4.1 单因素方差分析

统 计 检 验

对给定水平α，由$P\{F > \lambda\} = \alpha$查$F(k-1，n-k)$表得$\lambda$。若$F > \lambda$，则拒绝$H_0$，即检验效果显著；否则接受$H_0$，即检验效果不显著。通常将有关结果列为一张表，称为方差分析表，如表 4-4

表 4-4 方差分析表

方差来源	平方和	自由度	平均平方和	F 值
因素 A(组间)	$S_{组间}^2$	$k-1$	$\dfrac{S_{组间}^2}{k-1}$	$F = \dfrac{S_{组间}^2/(k-1)}{S_{误}^2/(n-k)}$
误差（组内）	$S_{误}^2$	$n-k$	$\dfrac{S_{误}^2}{n-k}$	
总 和	$S_{总}^2$	$n-1$		

4.2多因素方差分析

问 题 的 提 出

例 4-2 考虑合成纤维弹性，影响因素为收缩率 A 和拉伸倍数 B，A、B 各有四个水平，每个水平分别作了两次试验，相应的试验结果见表 4-8

表 4-8

试验结果 因子A / 因子B		A_1 0		A_2 4		A_3 8		A_4 12	
B_1	460	71	73	73	75	76	73	75	73
B_2	520	72	73	76	74	79	77	73	72
B_3	580	75	73	78	77	74	75	70	71
B_4	640	77	75	74	74	74	73	69	69

4.2 多因素方差分析

建 立 模 型

两个因素方差分析的一般数学模型：

$$
\begin{cases}
y_{ijl} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijl}, \\
\sum_{i=1}^{k} \alpha_i = 0, \sum_{j=1}^{\mathrm{r}} \beta_{\mathrm{j}} = 0, \sum_{i=1}^{k} \gamma_{ij} = \sum_{j=1}^{r} \gamma_{ij} = 0, \\
\varepsilon_{ijl} i \cdot i \cdot d N(0,\sigma^2), \quad i = 1,\cdots,k; \quad j = 1,\cdots,r; \quad l = 1,\cdots n,
\end{cases}
$$

需要解决如下问题：

（1）估计未知参数μ，α_i，β_j，γ_{ij}

（$i = 1,\cdots,k$；$j = 1,\cdots,r$；$l = 1,\cdots,n$）；

（2）考察因子A和因子B的水平变化对试验结果的影响有无显著差异，以及因子A和因子B有无交互作用，归结为下述三个假设检验：

4.2 多因素方差分析

参 数 估 计

记

$$\bar{y} = \frac{1}{nkr}\sum_{i=1}^{k}\sum_{j=1}^{r}\sum_{l=1}^{n} y_{ijl} \qquad \bar{y}_{i\cdot} = \frac{1}{nr}\sum_{j=1}^{r}\sum_{l=1}^{n} y_{ijl}$$

$$\bar{y}_{\cdot j} = \frac{1}{nk}\sum_{i=1}^{k}\sum_{l=1}^{n} y_{ijl} \qquad \bar{y}_{ij} = \frac{1}{n}\sum_{l=1}^{n} y_{ijl}$$

完全类似于单因素方差分析，得未知参数μ，α_i，β_j，γ_{ij}的矩估计为

$$\hat{\mu} = \bar{y}，\ \hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}，\ \hat{\beta}_j = \bar{y}_{\cdot j} - \bar{y} \qquad (4\text{-}13)$$

$$\hat{\gamma}_{ij} = \bar{y}_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}，\ i = 1,2,\cdots,k;\ j = 1,2,\cdots,r$$

易证它们分别是μ，α_i，β_j，γ_{ij}的无偏估计。

4.2 多因素方差分析

统 计 检 验

H_{01}：$\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$；

H_{10}：$\beta_1 = \beta_2 = \cdots = \beta_r = 0$；

H_{11}：γ_{ij}=0，$i = 1, \cdots, k$；$j = 1, \cdots, r$。

$$S_{总}^2 = \sum_{i=1}^{k}\sum_{j=1}^{r}\sum_{l=1}^{n} (y_{ijl} - \overline{y})^2 \triangleq S_{误}^2 + S_A^2 + S_B^2 + S_{AB}^2$$

当H_{01}：$\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$成立时

$$F_1 = \frac{S_A^2/(k-1)}{S_{误}^2/kr(n-1)} \sim F(k-1,\ kr(n-1)) \quad （4\text{–}19）$$

统 计 检 验

当H_{10}： $\beta_1 = \beta_2 = \cdots = \beta_r = 0$成立时

$$F_2 = \frac{S_B^2/(r-1)}{S_{误}^2/kr(n-1)} \sim F(r-1,\ kr(n-1)) \qquad (4\text{–}20)$$

当H_{11}: $\gamma_{ij}=0$， $i=1,\cdots,k$； $j=1,\cdots,r$成立时

$$F_3 = \frac{S_{AB}^2/(k-1)(r-1)}{S_{误}^2/kr(n-1)} \sim F((k-1)(r-1),\ kr(n-1)) \qquad (4\text{–}21)$$

4.2 多因素方差分析

统 计 检 验

$$S_{总}^2=\sum_{i=1}^{k}\sum_{j=1}^{r}\sum_{l=1}^{n}(y_{ijl}-\bar{y})^2=\sum_{i=1}^{k}\sum_{j=1}^{r}\sum_{l=1}^{n}y_{ijl}^2-\frac{T_{\cdots}^2}{krn}$$

$$S_A^2=\sum_{i=1}^{k}\sum_{j=1}^{r}\sum_{l=1}^{n}(\bar{y}_{i\cdot}-\bar{y})^2=rn\sum_{i=1}^{k}(\bar{y}_{i\cdot}-\bar{y})^2=\frac{1}{rn}\sum_{i=1}^{k}T_{i\cdot\cdot}^2-\frac{T_{\cdots}^2}{krn}$$

$$S_B^2=\sum_{i=1}^{k}\sum_{j=1}^{r}\sum_{l=1}^{n}(\bar{y}_{\cdot j}-\bar{y})^2=kn\sum_{j=1}^{r}(\bar{y}_{\cdot j}-\bar{y})^2=\frac{1}{kn}\sum_{j=1}^{r}T_{\cdot j\cdot}^2-\frac{T_{\cdots}^2}{krn}$$

$$S_{AB}^2=\sum_{i=1}^{k}\sum_{j=1}^{r}\sum_{l=1}^{n}(\bar{y}_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y})^2=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{r}T_{ij\cdot}^2-\frac{T_{\cdots}^2}{krn}-S_A^2-S_B^2$$

$$S_{误}^2=\sum_{i=1}^{k}\sum_{j=1}^{r}\sum_{l=1}^{n}(y_{ijl}-\bar{y}_{ij})^2=S_{总}^2-S_A^2-S_B^2-S_{AB}^2$$

$$T_{\cdots}=\sum_{i=1}^{k}\sum_{j=1}^{r}\sum_{l=1}^{n}y_{ijl}\;,\qquad T_{ij\cdot}=\sum_{l=1}^{n}y_{ijl}\qquad T_{i\cdot\cdot}=\sum_{j=1}^{r}\sum_{l=1}^{n}y_{ijl}$$

$$T_{\cdot j\cdot}=\sum_{i=1}^{k}\sum_{l=1}^{n}y_{ijl}\qquad i=1,\cdots,k;\;\; j=1,\cdots,r$$

4.2 多因素方差分析

统 计 检 验

表 4-7 方差分析表

方差来源	平方和	自由度	平均平方和	F 值
因素 A	S_A^2	$k-1$	$S_A^2/(k-1)$	$F_1=\dfrac{S_A^2/(k-1)}{S_{误}^2/kr(n-1)}$
因素 B	S_B^2	$r-1$	$S_B^2/(r-1)$	$F_2=\dfrac{S_B^2/(r-1)}{S_{误}^2/kr(n-1)}$
$A\times B$	S_{AB}^2	$(k-1)\times(r-1)$	$\dfrac{S_{AB}^2}{(k-1)(r-1)}$	$F_3=\dfrac{S_{AB}^2/(k-1)(r-1)}{S_{误}^2/kr(n-1)}$
误差	$S_{误}^2$	$kr(n-1)$	$S_{误}^2/kr(n-1)$	
总和	$S_{总}^2$	$krn-1$		

第五章 典型相关分析

对于两个变量，用它们的相关系数来衡量它们之间的线性相关关系。当考虑一个变量与一组变量的线性相关关系时，用它们的多重相关系数来衡量。但是，在许多实际问题中，常常会碰到两组变量之间的线性相关性问题研究。

5.1 典型相关分析简介

典型相关分析是分析两组变量之间相关性的一种统计分析方法。典型相关分析的基本思想和主成分分析的基本思想相似，它将两组变量之间的多重线性相关性研究，转化为少数几对综合变量之间的简单线性相关性的研究，并且这几对变量所包含的相关性信息，几乎覆盖了原变量组的全部相关信息。

5.2 典型相关分析的基本思想

典型相关分析方法的基本原理是：考虑研究的两组变量为x组和y组，x组有p个变量 $(x_1, x_2, \cdots, x_p)$**，y组有q个变量** $(y_1, y_2, \cdots, y_q)$ **，分别对这两组变量进行线性组合，再计算该对加权组合变量的简单相关系数，然后以这个简单相关系数当做这两组变量之间相关性的衡量指标，即**

5.2 典型相关分析的基本思想

$$s = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p$$

$$t = \beta_1 y_1 + \beta_2 y_2 + \cdots + \beta_q y_q$$

这样的 s 和 t 称为**典型变量**，典型变量 s 和 t 之间的相关系数称为**典型相关系数**。

注：1. 典型变量成对出现，权值系数改变，有很多。

2. 典型相关系数最大的那对典型变量称为第1（对）典型变量；典型相关系数第2大的那对典型变量称为第2（对）典型变量；以此类推。

5.3 典型相关分析的具体计算

令

$$x_1^* = \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1p}x_p$$

$$= (\alpha_{11}, \alpha_{12}, \cdots, \alpha_{1p})(x_1, x_2, \cdots, x_p)^T \hat{=} \alpha_1^T x$$

$$y_1^* = \beta_{11}y_1 + \beta_{12}y_2 + \cdots + \beta_{1q}y_q$$

$$= (\beta_{11}, \beta_{12}, \cdots, \beta_{1q})(y_1, y_2, \cdots, y_q)^T \hat{=} \beta_1^T y$$

典型相关系数为

$$\rho(x_1^*, y_1^*) = \frac{Cov(x_1^*, y_1^*)}{\sqrt{Var(x_1^*) \times Var(y_1^*)}}$$

$$= \frac{Cov(\alpha_1^T x, \beta_1^T y)}{\sqrt{Var(\alpha_1^T x) \times Var(\beta_1^T y)}} = \frac{\alpha_1^T \Sigma_{12} \beta_1}{\sqrt{\alpha_1^T \Sigma_{11} \alpha_1 \times \beta_1^T \Sigma_{22} \beta_1}}$$

5.3 典型相关分析的具体计算

max $\rho(x_1^*, y_1^*) = \alpha_1^T \Sigma_{12} \beta_1$ ⇨ α_1, β_1

s.t. $\alpha_1^T \Sigma_{11} \alpha_1 = 1 \quad \beta_1^T \Sigma_{22} \beta_1 = 1$

构造拉格朗日函数 $Q = \alpha_1^T \Sigma_{12} \beta_1 - \frac{1}{2}\lambda_1(\alpha_1^T \Sigma_{11} \alpha_1 - 1) - \frac{1}{2}\lambda_2(\beta_1^T \Sigma_{22} \beta_1 - 1)$

$$\begin{cases} \frac{\partial Q}{\partial \alpha_1} = \Sigma_{12}\beta_1 - \lambda_1 \Sigma_{11}\alpha_1 = 0 \\ \frac{\partial Q}{\partial \beta_1} = \Sigma_{21}\alpha_1 - \lambda_2 \Sigma_{22}\beta_1 = 0 \\ \frac{\partial Q}{\partial \lambda_1} = -\frac{1}{2}(\alpha_1^T \Sigma_{11}\alpha_1 - 1) = 0 \\ \frac{\partial Q}{\partial \lambda_2} = -\frac{1}{2}(\beta_1^T \Sigma_{22}\beta_1 - 1) = 0 \end{cases}$$

前两个方程分别乘以 α_1^T, β_1^T ，有

$$\begin{cases} \alpha_1^T \Sigma_{12}\beta_1 = \lambda_1 \alpha_1^T \Sigma_{11}\alpha_1 = \lambda_1 \\ \beta_1^T \Sigma_{21}\alpha_1 = \lambda_2 \beta_1^T \Sigma_{22}\beta_1 = \lambda_2 \end{cases}$$

故 $\lambda_1 = \lambda_2 = \lambda \quad \rho(x_1^*, y_1^*) = \alpha_1^T \Sigma_{12}\beta_1 = \lambda$

5.3 典型相关分析的具体计算

从而前两个方程变为

$$\Sigma_{12}\beta_1 = \lambda\Sigma_{11}\alpha_1$$

$$\Sigma_{21}\alpha_1 = \lambda\Sigma_{22}\beta_1$$

故 $\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\beta_1 = \lambda^2\beta_1$

$$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\alpha_1 = \lambda^2\alpha_1$$

令

$$A = \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \qquad B = \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

则

$$B\beta_1 = \lambda^2\beta_1 \qquad A\alpha_1 = \lambda^2\alpha_1$$

5 典型相关分析

典型相关分析的步骤： $(x_1, x_2, \cdots, x_p)$　$(y_1, y_2, \cdots, y_q)$

Step1： 计算x组的协方差阵为Σ_{11}，y组的协方差阵为Σ_{22}，x与y的协方差阵为Σ_{12}

Step2： 计算　$A = \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$　$B = \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$

Step3： 确定非零特征根的数量　$\min(p,q)$

Step4： 由A或B计算，q个非零特征值排序为$\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_q^2$
$\lambda_1, \lambda_2, \cdots, \lambda_q$为典型相关系数

Step5： 由$A\alpha_i = \lambda_i^2\alpha_i$和$B\beta_i = \lambda_i^2\beta_i$，计算特征向量$\alpha_i$和$\beta_i$，为典型变量的线性组合系数

5 典型相关分析

Matlb中进行典型相关分析：

Matlab在其统计与机器学习工具箱中，提供了进行典型相关分析的函数：

[A,B,r] = canoncorr(X,Y)

其中，X, Y分别为x组和y组的数据观测矩阵；

A, B矩阵中每一列，分别为x组和y组的线性组合系数

r 向量给出，每对典型变量的典型相关系数

第六章 判别分析

6.1 距离判别

6.2 Fisher判别

6.3 Bayes判别

1. Mahalanobis距离的概念

定义 设x,y是从均值为μ，协方差为Σ的总体A中抽取的样本，则总体A内两点x与y的 Mahalanobis 距离（简称马氏距离）定义为

$$d(x,y)=\sqrt{(x-y)^T\Sigma^{-1}(x-y)},$$

定义样本x与总体A的 Mahalanobis 距离为

$$d(x,A)=\sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}.$$

2.距离判别的判别准则和判别函数

在这里讨论两个总体的距离判别，分协方差相同和协方差不同两种进行讨论。

设总体A和B的均值向量分别为μ_1和μ_2，协方差阵分别为Σ_1和Σ_2，今给一个样本x，要判断x来自哪一个总体。

$$x \in \begin{cases} A, d(x,A) \le d(x,B), \\ B, d(x,A) > d(x,B). \end{cases}$$

2.距离判别的判别准则和判别函数

(1) 首先考虑协方差相同，即

$$\mu_1 \neq \mu_2, \quad \Sigma_1 = \Sigma_2 = \Sigma.$$

$$d^2(x,B) - d^2(x,A) = 2(x-\bar{\mu})^T \Sigma^{-1}(\mu_1 - \mu_2) \qquad \bar{\mu} = \frac{\mu_1 + \mu_2}{2}$$

令

$$w(x) = (x - \bar{\mu})^T \Sigma^{-1}(\mu_1 - \mu_2),$$

称$w(x)$为两总体距离的判别函数，因此判别准则变为

$$x \in \begin{cases} A, & w(x) \geq 0, \\ B, & w(x) < 0. \end{cases}$$

2.距离判别的判别准则和判别函数

(2) 再考虑协方差不同的情况，即

$$\mu_1 \neq \mu_2, \quad \Sigma_1 \neq \Sigma_2,$$

对于样本x，在方差不同的情况下，判别函数为

$$w(x) = (x-\mu_2)^T \Sigma_2^{-1}(x-\mu_2) - (x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1) .$$

6.2 Fisher判别

1. Fisher判别基本原理

Fisher 的判别思想是变换多元观测 x 到一元观测 y，使得由总体 X_1, X_2 产生的 y 尽可能的分离开来。

对于 p 维的 x，需判定 x 属于 X_1, X_2 哪个 p 维总体？

做线性组合：$y = a^T x$。

X_1, X_2 的均值向量分别为 μ_1, μ_2（均为 p 维），且有公共的协方差矩阵 Σ（$\Sigma > 0$），则：

$$\mu_{y_1} = E(y \mid y = a^T x, x \in X_1) = a^T \mu_1,$$

$$\mu_{y_2} = E(y \mid y = a^T x, x \in X_2) = a^T \mu_2,$$

其方差为

$$\sigma_y^2 = \mathrm{Var}(y) = a^T \Sigma a,$$

6.2 Fisher判别

1. Fisher判别基本原理

求a，使得max $\dfrac{(\mu_{y_1}-\mu_{y_2})^2}{\sigma_y^2}=\dfrac{[a^T(\mu_1-\mu_2)]^2}{a^T\Sigma a}=\boxed{\dfrac{(a^T\delta)^2}{a^T\Sigma a}}$

达到最大，其中$\delta=\mu_1-\mu_2$

定理 6.1 x为p维随机变量，设$y=a^Tx$，当选取$a=c\Sigma^{-1}\delta$，$c\neq0$为常数时，上式达到最大。

特别当$c=1$时，线性函数

$$y=a^Tx=(\mu_1-\mu_2)^T\Sigma^{-1}x$$

称为 Fisher 线性判别函数。令

$$K=\frac{1}{2}(\mu_{y_1}+\mu_{y_2})=\frac{1}{2}(a^T\mu_1+a^T\mu_2)=\frac{1}{2}(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1+\mu_2).$$

1. Fisher判别基本原理

定义判别函数

$$W(x)=(\mu_1-\mu_2)^T\Sigma^{-1}x-K=\left[x-\frac{1}{2}(\mu_1+\mu_2)\right]^T\Sigma^{-1}(\mu_1-\mu_2)$$

则判别规则可改写成

$$\begin{cases} x\in X_1,\text{当}x\text{使得}W(x)\geq 0, \\ x\in X_2,\text{当}x\text{使得}W(x)<0. \end{cases}$$

1.误判概率与误判损失

某样本x实际是来自X_1，但被判为X_2的概率为

$$P(2|1)=P(x\in R_2|X_1),$$

来自X_2，但被判为X_1的概率为

$$P(1|2)=P(x\in R_1|X_2).$$

类似地，来自X_1被判为X_1的概率，来自X_2被判为X_2的概率分别为

$$P(1|1)=P(x\in R_1|X_1),$$

$$P(2|2)=P(x\in R_2|X_2).$$

6.3 Bayes判别

p_1, p_2 为总体 X_1 和 X_2 的先验概率，且 $p_1 + p_2 = 1$

$$P(\text{正确地判为}X_1) = P(\text{来自}X_1,\text{被判为}X_1) = P(1|1) \times p_1$$

$$P(\text{误判到}X_1) = P(\text{来自}X_2,\text{被判为}X_1) = P(1|2) \cdot p_2$$

类似地有

$$P(\text{正确地判为}X_2) = P(2|2) \cdot p_2\,,$$

$$P(\text{误判到}X_2) = P(2|1) \cdot p_1.$$

$L(1|2)$: 来自 X_2 误判为 X_1 引起的损失

$L(2|1)$: 来自 X_1 误判为 X_2 引起的损失

$L(1|1) = L(2|2) = 0$。

6.3 Bayes判别

平均误判损失（Expected Cost of Misclassification，简记为 ECM）如下

$$\mathrm{ECM}(R_1, R_2) = L(2|1)P(2|1)p_1 + L(1|2)P(1|2)p_2$$

一个合理的判别规则应使 ECM 达到极小

2.两总体的Bayes判别

由上面叙述，要选择样本空间Ω的一个划分R_1和$R_2=\Omega-R_1$使得平均损失ECM达到极小。

定理 6.3 极小化平均误判损失ECM的R_1和R_2为

$$R_1=\left\{x:\frac{f_1(x)}{f_2(x)}\geq\frac{L(1|2)}{L(2|1)}\cdot\frac{p_2}{p_1}\right\},$$

$$R_2=\left\{x:\frac{f_1(x)}{f_2(x)}<\frac{L(1|2)}{L(2|1)}\cdot\frac{p_2}{p_1}\right\},$$

6.3 Bayes判别

两总体 Bayes 判别的步骤：

（1）新样本点 $x_0 = [x_{01}, x_{02}, \cdots, x_{0p}]^T$ 的密度函数比

$f_1(x_0) / f_2(x_0)$；

（2）损失比 $L(1|2) / L(2|1)$；

（3）先验概率比 p_2 / p_1。

（4）$$\begin{cases} x \in X_1, & 当x使得 \dfrac{f_1(x)}{f_2(x)} \geq \dfrac{L(1|2)}{L(2|1)} \cdot \dfrac{p_2}{p_1}, \\ x \in X_2, & 当x使得 \dfrac{f_1(x)}{f_2(x)} < \dfrac{L(1|2)}{L(2|1)} \cdot \dfrac{p_2}{p_1}. \end{cases}$$

6.3 Bayes判别

下面列举三种特殊情况：

（1） 当 $p_2 / p_1 = 1$

$$\begin{cases} x \in X_1, & 当x使得 \dfrac{f_1(x)}{f_2(x)} \geq \dfrac{L(1|2)}{L(2|1)}, \\ x \in X_2, & 当x使得 \dfrac{f_1(x)}{f_2(x)} < \dfrac{L(1|2)}{L(2|1)}. \end{cases}$$

（2） 当 $L(1|2) / L(2|1) = 1$ 时

$$\begin{cases} x \in X_1, & 当x使得 \dfrac{f_1(x)}{f_2(x)} \geq \dfrac{p_2}{p_1}, \\ x \in X_2, & 当x使得 \dfrac{f_1(x)}{f_2(x)} < \dfrac{p_2}{p_1}. \end{cases}$$

（3） $p_1 / p_2 = L(1|2) / L(2|1) = 1$ 时

$$\begin{cases} x \in X_1, & 当x使得 \dfrac{f_1(x)}{f_2(x)} \geq 1, \\ x \in X_2, & 当x使得 \dfrac{f_1(x)}{f_2(x)} < 1. \end{cases}$$

第七章 聚类分析

7.1 聚类标准

7.2 系统聚类法

7.3 K均值聚类法

7.4 谱聚类法

7.5 基于密度的聚类法

7.1 聚类标准

聚类分析又称群分析，它是研究分类问题的一种多元统计分析。所谓类通俗地说，就是指相似元素的集合。要将相似元素聚为一类，通常选取元素的许多共同指标，然后通过分析元素的指标值来分辨元素间的差距，从而达到分类的目的。聚类分析可以分为 Q 型聚类（样本聚类）、R 型聚类（指标聚类）。

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

1. 样本的相似性度量

（1）绝对值距离

$$d_1(x,y)=\sum_{k=1}^{p}\left|x_k-y_k\right|,$$

（2）欧氏距离

$$d_2(x,y)=\left[\sum_{k=1}^{p}\left|x_k-y_k\right|^2\right]^{\frac{1}{2}},$$

（3）Chebyshev 距离

$$d_\infty(x,y)=\max_{1\le k\le p}\left|x_k-y_k\right|.$$

（4）马氏（Mahalanobis）距离

$$d(x,y)=\sqrt{(x-y)^T\Sigma^{-1}(x-y)}$$

2. 指标(变量)的相似性度量

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

(1) 相关系数

记变量x_j的取值

$$[x_{1j},x_{2j},\cdots,x_{mj}]^T\in R^m (j=1,2,\cdots,p)。$$

$$r_{jk}=\frac{\sum_{i=1}^{m}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)}{\left[\sum_{i=1}^{m}(x_{ij}-\bar{x}_j)^2\sum_{i=1}^{m}(x_{ik}-\bar{x}_k)^2\right]^{\frac{1}{2}}},\quad j,k=1,2,\cdots,p$$

其中，$\bar{x}_j=\frac{1}{m}\sum_{i=1}^{m}x_{ij}$。对指标(变量)进行聚类分析时，利用相关系数矩阵$R=(r_{ij})_{p\times p}$是最多的。

(2) 夹角余弦

也可以直接利用两变量x_j与x_k的夹角余弦r_{jk}来定义它们的相似性度量，有

$$r_{jk}=\frac{\sum_{i=1}^{m}x_{ij}x_{ik}}{\left(\sum_{i=1}^{m}x_{ij}^2\sum_{i=1}^{m}x_{ik}^2\right)^{\frac{1}{2}}},\quad j,k=1,2,\cdots,p.$$

7.1 聚类标准

3. 类与类之间的相似性度量——样本类G1和G2

（1）最短距离法（nearest neighbor or single linkage method）

$$D(G_1,G_2)=\min_{\substack{x_i\in G_1\\ y_j\in G_2}}\{d(x_i,y_j)\},$$

它的直观意义为两个类中最近两点间的距离。

（2）最长距离法（farthest neighbor or complete linkage method）

$$D(G_1,G_2)=\max_{\substack{x_i\in G_1\\ y_j\in G_2}}\{d(x_i,y_j)\},$$

它的直观意义为两个类中最远两点间的距离。

（3）重心法（centroid method）

$$D(G_1,G_2)=d(\bar{x},\bar{y}),$$

其中$\bar{x},\bar{y}$分别为G_1,G_2的重心。

（4）类平均法（group average method）

$$D(G_1,G_2)=\frac{1}{n_1 n_2}\sum_{x_i\in G_1}\sum_{x_j\in G_2} d(x_i,x_j),$$

它等于G_1,G_2中两两样本点距离的平均，式中n_1,n_2分别为G_1,G_2中的样本点个数。

（5）离差平方和法（sum of squares method）

若记

$$D_1=\sum_{x_i\in G_1}(x_i-\bar{x}_1)^T(x_i-\bar{x}_1),$$

$$D_2=\sum_{x_j\in G_2}(x_j-\bar{x}_2)^T(x_j-\bar{x}_2),$$

$$D_{12}=\sum_{x_k\in G_1\cup G_2}(x_k-\bar{x})^T(x_k-\bar{x}),$$

其中

$$\bar{x}_1=\frac{1}{n_1}\sum_{x_i\in G_1}x_i,\quad \bar{x}_2=\frac{1}{n_2}\sum_{x_j\in G_2}x_j,\quad \bar{x}=\frac{1}{n_1+n_2}\sum_{x_k\in G_1\cup G_2}x_k,$$

则定义

$$D(G_1,G_2)=D_{12}-D_1-D_2.$$

7.1 聚类标准

3. 类与类之间的相似性度量——指标类G1和G2

（1）指标类与类间的最长距离法

在最长距离法中，定义两类变量的距离为

$$R(G_1,G_2)=\max_{\substack{x_j\in G_1\\ x_k\in G_2}}\{d_{jk}\},$$

其中$d_{jk}=1-|r_{jk}|$或$d_{jk}^2=1-r_{jk}^2$，这时，$R(G_1,G_2)$与两类中相似性最小的两变量间的相似性度量值有关。

（2）指标类与类间的最短距离法

在最短距离法中，定义两类变量的距离为

$$R(G_1,G_2)=\min_{\substack{x_j\in G_1\\ x_k\in G_2}}\{d_{jk}\},$$

其中$d_{jk}=1-|r_{jk}|$或$d_{jk}^2=1-r_{jk}^2$，这时，$R(G_1,G_2)$与两类中相似性最大的两个变量间的相似性度量值有关。

7.2 系统聚类法

系统聚类法是最常用的一种聚类方法，其基本思想是将样品各看成一类，然后定义类与类之间的距离，将距离最短的两类合并为一个新类，再计算新类与其它类之间的距离，将距离最短的两类合并为一个新类，如此下去，直到合并为一个大类为止。

系统聚类法的一般步骤

（1）将每个样品独自聚成一类，构造n个类。

（2）根据所确定的样品距离公式，计算n个样品（或变量）两两间的距离，构造距离矩阵，记为$D_{(0)}$。

（3）把距离最近的两类归为一新类，其它样品仍各自聚为一类，共聚成$n-1$类。

（4）计算新类与当前各类的距离，将距离最近的两个类进一步聚成一类，共聚成$n-2$类。以上步骤一直进行下去，最后将所有的样品聚成一类。

（5）画聚类谱系图。

（6）决定类的个数及各类包含的样品数，并对类做出解释。

7.3 K均值聚类法

1. K均值聚类基本思想

算法的思想是假定样本集中的全体样本可分为C类，并选定C个初始聚类中心，然后，根据最小距离原则将每个样本分配到某一类中，之后不断迭代计算各类的聚类中心，并依据新的聚类中心调整聚类情况，直到迭代收敛或聚类中心不再改变。

2. K均值聚类算法步骤

K均值聚类算法描述如下：

（1）初始化。设总样本集$G=\{\omega_j, j=1,2,\cdots,n\}$是$n$个样品组成的集合，聚类数为$C$（$2\le C\le n$），将样本集$G$任意划分为$C$类，记为$G_1,G_2,\cdots,G_C$，计算对应的$C$个初始聚类中心，记为$m_1,m_2,\cdots,m_C$，并计算$J_e$。

（2）$G_i=\Phi$（$i=1,2,\cdots,C$），按最小距离原则将样品ω_j（$j=1,2,\cdots,n$）进行聚类，即若$d(\omega_j,G_k)=\min\limits_{1\le i\le C} d(\omega_j,m_i)$，则$\omega_j\in G_k$，$G_k=G_k\cup\{\omega_j\}$，$j=1,2,\cdots,n$。 重新计算聚类中心

$$m_i=\frac{1}{n_i}\sum_{\omega_j\in G_i}\omega_j,\quad i=1,2,\cdots,C,$$

式中，n_i为当前G_i类中的样本数目。并重新计算J_e。

$$\boxed{J_e=\sum_{i=1}^{C}\sum_{\omega\in G_i}\left\|\omega-m_i\right\|^2,}$$

（3）若连续两次迭代的J_e不变，则算法终止，否则算法转（2）。

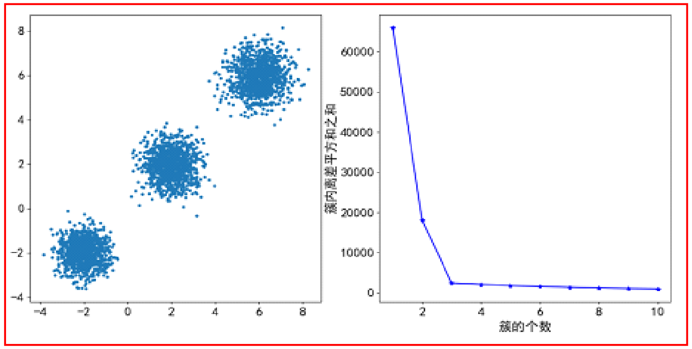
实际计算时，可以不计算J_e，只要聚类中心不发生变化，算法即可终止。

7.3 K均值聚类法

3. 如何确定K均值聚类的聚类数k值

1.拐点法

簇内离差平方和拐点法的思想很简单，就是在不同的k值下计算簇内离差平方和，然后通过可视化的方法找到“拐点”所对应的k值。重点关注的是斜率的变化，当斜率由大突然变小时，并且之后的斜率变化缓慢，则认为突然变换的点就是寻找的目标点，因为继续随着簇数k的增加，聚类效果不再有大的变化。

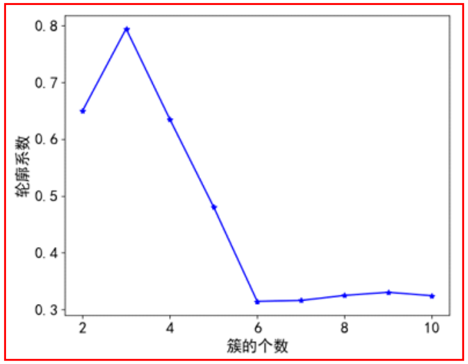


2.轮廓系数法

定义样本点i的轮廓系数

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)},$$

k个簇的总轮廓系数定义为所有样本点轮廓系数的平均值



7.4 谱聚类法

谱聚类是从图论中演化出来的算法，后来在聚类中得到了广泛的应用。它的主要思想是把所有的数据看做空间中的点，这些点之间可以用边连接起来。

距离较远的两个点之间的边权重值较低，而距离较近的两个点之间的边权重值较高，通过对所有数据点组成的图进行切图，让切图后不同的子图间边权重和尽可能的低，而子图内的边权重和尽可能的高，从而达到聚类的目的。

7.4 谱聚类法

上述方法是Ng，Jordan和Weiss（NJW）提出的，NJW算法具体步骤如下：

Step1. 生成关联矩阵A (相似矩阵)； $A_{ij} = \exp[-\frac{\| X_i - X_j \|^2}{2\sigma^2}], i \neq j$

Step2. 利用A中的元素构造矩阵D； $D_{ii} = \sum_{j=1}^{n} A_{ij}$

Step3. 利用公式计算L； $L = D^{-1/2} A D^{-1/2}$

Step4. 计算L的特征向量和特征值； $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ $u_1, u_2, \cdots, u_k$

Step5. 把L的前k个特征值所对应的特征向量存入矩阵U $U = [u_1, u_2, \cdots, u_k]$

7.4 谱聚类法

Step6. 把U的行向量归一化，使其具有单位长度，归一化的矩阵命名为Y;

$$y_{ij}=\frac{u_{ij}}{\sqrt{\sum_{j}u_{ij}^{2}}}$$

Step7. 对Y的行向量进行K均值聚类;

Step8. 如果Y的第i行被分配到第j个类，则把Xi分配到第j个类。

SpectralClustering(n_clusters, n_init, gamma)

7.5 基于密度的聚类法

ε-邻域：给定对象O，半径ε内的区域称为该对象O的ε -邻域。

核心对象：如果给定对象O的ε- 邻域内的样本点数大于或等于MinPts，则称该对象O为核心对象。

MinPts 为人为预先指定的最小点数，阈值参数。

直接密度可达：给定一个对象集合D，如果p在q的ε -邻域内，且q是一个核心对象，则称对象p从对象q出发是直接密度可达的。

密度可达：对于样本集合D，如果存在一个对象链p1,p2,…,pn，使得p1=q，pn=p，并且pi属于D（i=1,2,…,n），p(i+1)是pi关于 ε 和 MinPts 直接密度可达的，则称对象p从对象q出发是密度可达的。

7.5 基于密度的聚类法

密度相连：如果存在对象q属于D，使对象p1和p2都是从q关于 ε 和MinPts密度可达的，那么对象p1、p2是关于ε 和MinPts 密度相连的。

DBSCAN聚类：由**密度可达**关系导出的**最大密度相连**的**样本集合**，即为最终聚类簇（一簇 即为 一类）。

7.5 基于密度的聚类法

那么怎么才能找到这样的簇样本集合呢？DBSCAN使用的方法流程很简单：

(1) 任意选择一个没有类别的核心对象作为种子，

(2) 然后找到所有这个核心对象能够密度可达的样本集合，即为一个聚类簇。

(3) 接着继续选择另一个没有类别的核心对象去寻找密度可达的样本集合，这样就得到另一个聚类簇。

(4) 一直运行到所有核心对象都有类别为止。

DBSCAN(eps=1.5,min_samples=4)

第八章 降维分析

8.1 主成分分析

8.2 因子分析

8.1 主成分分析PCA

主成分分析PCA基本思想

主成分分析通过将原来指标重新组合成一组新的相互无关的几个综合指标，来消除原有指标间的相关性，由几个相互无关的综合指标尽可能多地反映原来指标的信息，从而实现降维的一种方法。所构造的几个综合指标就称为 主成分。

8.1 主成分分析PCA

主成分分析的步骤

设有m个指标变量$x_1, x_2, \cdots, x_m$，它在第i次观测中的取值为

$$a_{i1}, a_{i2}, \cdots, a_{im}\ (i = 1, 2, \cdots, n),$$

将它们写成矩阵形式

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix},$$

矩阵A称为观测阵。

主成分分析的步骤

对于观测数据矩阵$A=(a_{ij})_{n\times m}$。按如下步骤进行 PCA 分析

（1）对原来的m个指标进行标准化，得到标准化的指标变量

$$y_j=\frac{x_j-\mu_j}{s_j},\quad j=1,2,\cdots,m,$$

其中，$\mu_j=\frac{1}{n}\sum_{i=1}^{n}a_{ij}$，$s_j=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_{ij}-\mu_j)^2}$。对应地，得到标准化的数据矩阵$B=(b_{ij})_{n\times m}$，其中$b_{ij}=\frac{a_{ij}-\mu_j}{s_j}$，$i=1,2,\cdots,n$，$j=1,2,\cdots,m$。

主成分分析的步骤

（2）根据标准化的数据矩阵B求出协方差阵 $\sum$ 或 相关系数矩阵 $R=(r_{ij})_{m\times m}$，

（3）计算协方差阵 $\sum$ 或 相关系数矩阵R的特征值$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$，及对应的标准正交化特征向量$u_1, u_2, \cdots, u_m$，其中$u_j=[u_{1j}, u_{2j}, \cdots, u_{mj}]^T$，由特征向量组成$m$个新的指标变量

$$\begin{cases} y_1 = u_{11}x_1 + u_{21}x_2 + \cdots + u_{m1}x_m, \\ y_2 = u_{12}x_1 + u_{22}x_2 + \cdots + u_{m2}x_m, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_m = u_{1m}x_1 + u_{2m}x_2 + \cdots + u_{mm}x_m. \end{cases}$$

式中y_1是第 1 主成分，y_2是第 2 主成分，…，y_m是第m主成分。

主成分分析的步骤

（4）计算主成分贡献率及累计贡献率，主成分F_j的贡献率为

$$w_j = \frac{\lambda_j}{\sum_{k=1}^{m} \lambda_k}, \quad j = 1, 2, \cdots, m,$$

前i个主成分的累计贡献率为

$$\sum_{k=1}^{i} \lambda_k \Big/ \sum_{k=1}^{m} \lambda_k.$$

一般取累计贡献率达 85%以上的特征值$\lambda_1, \lambda_2, \cdots, \lambda_k$所对应的第 1、第 2、…、第$k\,(k \le p)$主成分。

(5)最后利用得到的主成分$y_1, y_2, \cdots, y_k$分析问题，或者继续进行评价、回归、聚类等其他建模。

2. 因子分析模型

因子分析有确定的模型，观察数据在模型中被分解为公共因子、特殊因子和误差三部分。因子分析中的因子是一个比较抽象的概念。

设p个变量$X_i(i=1,2,\cdots,p)$，如果表示为

$$X_i=\mu_i+\alpha_{i1}F_1+\cdots+\alpha_{im}F_m+\varepsilon_i,\quad (m\le p)$$

或

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}=\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}+\begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2m} \\ \vdots & \vdots & & \vdots \\ \alpha_{p1} & \alpha_{p2} & \cdots & \alpha_{pm} \end{bmatrix}\begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix}+\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix},$$

或

$$X-\mu=\Lambda F+\varepsilon,$$

2. 因子分析模型

称$F_1,F_2,\cdots,F_m$为公共因子，是不可观测的变量，Λ矩阵称为因子载荷矩阵。ε_i是特殊因子，是不能被前m个公共因子包含的部分。并且满足

$$E(F)=0,\ E(\varepsilon)=0,\ \mathrm{Cov}(F)=I_m,$$

$$D(\varepsilon)=\mathrm{Cov}(\varepsilon)=\mathrm{diag}(\sigma_1^2,\sigma_2^2,\cdots,\sigma_p^2),\ \mathrm{Cov}(F,\varepsilon)=0.$$

称上述因子模型为正交因子模型

因子载荷矩阵的估计方法—主成分分析

设$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$为样本相关系数矩阵$R$的特征值$\eta_1,\eta_2,\cdots,\eta_p$为相应的标准正交化特征向量。设$m < p$，则因子载荷矩阵$\Lambda$为

$$\Lambda = [\sqrt{\lambda_1}\eta_1, \sqrt{\lambda_2}\eta_2, \cdots, \sqrt{\lambda_m}\eta_m],$$

特殊因子的方差用$R - \Lambda\Lambda^T$的对角元来估计，即

$$\sigma_i^2 = 1 - \sum_{j=1}^{m} \alpha_{ij}^2$$

$\boxed{R = \Lambda\Lambda^T}$

先对变量进行标准化变换

$\boxed{\sum = \Lambda\Lambda^T + \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2)}$ → $\boxed{R = \Lambda\Lambda^T + \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2)}$

8.2 因子分析

因子载荷矩阵的估计方法—因子分析

主因子方法是对主成分方法的修正，假定我们首先对变量进行标准化变换。则 $\boxed{R = \Lambda\Lambda^T + \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2)}$

$$R = \Lambda\Lambda^T + D,$$

其中 $D = \mathrm{diag}\{\sigma_1^2, \cdots, \sigma_p^2\}$。记

$$R^* = \Lambda\Lambda^T = R - D,$$

称 R^* 为约相关系数矩阵，R^* 对角线上的元素是 h_i^2。

在实际应用中，特殊因子的方差一般都是未知的可以通过一组样本来估计。估计的方法有如下几种

因子载荷矩阵的估计方法—因子分析法

（1）取$\hat{h}_i^2=1$，在这种情况下主因子解与主成分解等价。

（2）取$\hat{h}_i^2=\max\limits_{j\neq i}|r_{ij}|$，这意味着取$X_i$与其余的$X_j$的简单相关系数的绝对值最大者。

记

$$R^*=R-D=\begin{bmatrix}\hat{h}_1^2 & r_{12} & \cdots & r_{1p}\\ r_{21} & \hat{h}_2^2 & \cdots & r_{2p}\\ \vdots & \vdots & & \vdots\\ r_{p1} & r_{p2} & \cdots & \hat{h}_p^2\end{bmatrix},$$

因子载荷矩阵的估计方法—因子分析法

直接求R^*的前p个特征值$\lambda_1^* \geq \lambda_2^* \geq \cdots \geq \lambda_p^*$，和对应的正交特征向量$u_1^*, u_2^*, \cdots, u_p^*$，得到如下的因子载荷矩阵

$$\Lambda = \left[\sqrt{\lambda_1^*} u_1^* \quad \sqrt{\lambda_2^*} u_2^* \quad \cdots \quad \sqrt{\lambda_p^*} u_p^* \right].$$