遼寧石油化工大學
LIAONING SHIHUA UNIVERSITY

数学模型与数学建模之

聚类分析之
系统聚类法及其SPSS实现

于晶贤

E-mail: yujingxian@126.com

主要内容：

1. 样品与样品间的距离
2. 指标和指标间的“距离”
3. 类与类间的距离
4. 常用系统聚类法
5. 例子

聚类分析（物以类聚，人以群分）

引例1 下表是30个学生的六门课的成绩。根据这30个人的成绩，对这30个学生进行分类。

序号	数学	物理	化学	语文	历史	英语
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57
……	……	……	……	……	……	……
28	77	90	85	68	73	76
29	91	82	84	54	62	60
30	78	84	100	51	60	60

引例2 下表是30个学生的六门课的成绩。根据这30个人的成绩，将六门课程分为两类。

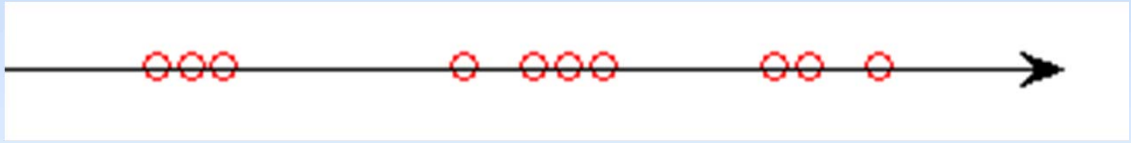
序号	数学	物理	化学	语文	历史	英语
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57
……	……	……	……	……	……	……
28	77	90	85	68	73	76
29	91	82	84	54	62	60
30	78	84	100	51	60	60

引例3 下表是中国大陆地区31个省级行政区域的月人均消费数据（单位：元），请根据消费水平对这31个省级行政区域进行分类。

城市	人均粮食支出	人均副食支出	人均烟、酒、饮料支出	人均衣着支出	人均日用杂品支出	人均水电燃料支出	人均其他非商品支出
北京	21.3	124.89	35.43	93.01	20.58	43.97	433.73
天津	21.5	122.39	29.08	55.04	11.3	54.88	288.13
河北	18.25	90.21	24.45	62.48	7.45	47.5	178.84
……	……	……	……	……	……	……	……
青 海	20.33	75.64	20.88	53.81	10.06	32.82	171.32
宁 夏	19.75	70.24	18.67	61.75	10.08	40.26	165.22
新 疆	21.03	78.55	14.35	64.98	9.83	33.87	161.67

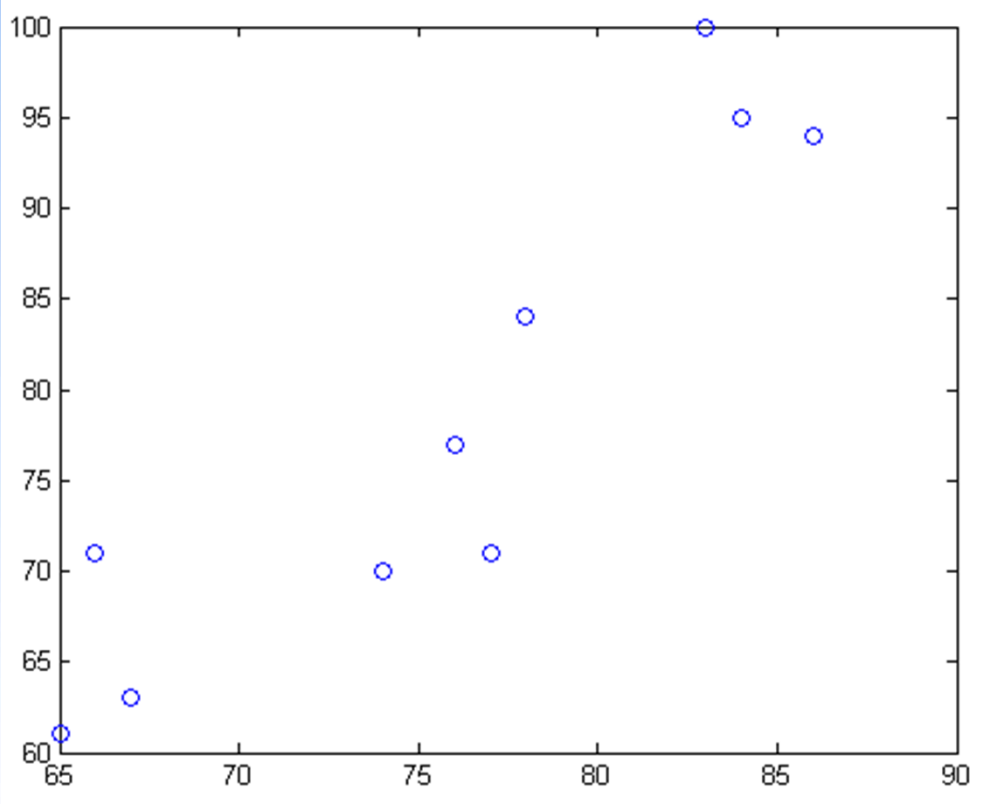
如何分类

引例1中，如果只考虑数学成绩（取前十个学生的成绩来分析），可以将十个人的分数画在x轴上，然后把接近的点放入一类。



ID	数学
1	65
2	76
3	67
4	84
5	74
6	78
7	66
8	77
9	83
10	86

引例1中，如果考虑数学成绩和物理成绩（取前十个学生的成绩来分析），可以将十个人的两个分数看作是 *xoy* 平面上的点，然后把接近的点分为一类。



ID	数学	物理
1	65	61
2	76	77
3	67	63
4	84	95
5	74	70
6	78	84
7	66	71
8	77	71
9	83	100
10	86	94

分类准则

距离近的样品聚为一类

数据的一般的格式

	指标 1 X_1	指标 2 X_2	…	指标 j X_j	…	指标 p X_p
样品 1	x_{11}	x_{12}	…	x_{1j}	…	x_{1p}
样品 2	x_{21}	x_{22}	…	x_{2j}	…	x_{2p}
⋮	⋮	⋮	⋱	⋮	…	⋮
样品 i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{ip}
⋮	⋮	⋮	…	⋮	⋱	⋮
样品 n	x_{n1}	x_{n2}	…	x_{nj}	…	x_{np}

样品与样品之间的常用距离（样品*i*与样品*j*）

绝对值距离：$d(\vec{x}_i,\vec{x}_j)=\sum_{k=1}^{p}|x_{ik}-x_{jk}|$

欧氏距离：$d(\vec{x}_i,\vec{x}_j)=\sqrt{\sum_{k=1}^{p}(x_{ik}-x_{jk})^2}$

Minkowski距离：$d(\vec{x}_i,\vec{x}_j)=\left[\sum_{k=1}^{p}(x_{ik}-x_{jk})^q\right]^{\frac{1}{q}}$

Chebyshev距离：$d(\vec{x}_i,\vec{x}_j)=\max_{1\le k\le p}|x_{ik}-x_{jk}|$

马氏距离：$d(\vec{x}_i,\vec{x}_j)=(\vec{x}_i-\vec{x}_j)'\Sigma^{-1}(\vec{x}_i-\vec{x}_j)$

其中：$\vec{x}_i=(x_{i1},x_{i2},\cdots,x_{ip})'$　$\vec{x}_j=(x_{j1},x_{j2},\cdots,x_{jp})'$

Σ为样本的协方差矩阵

ID	数学	物理
1	65	61
2	76	77
3	67	63
4	84	95
5	74	70
6	78	84
7	66	71
8	77	71
9	83	100
10	86	94

绝对值距离：

$$d(\vec{x}_1,\vec{x}_2)=\sum_{k=1}^{p}|x_{1k}-x_{2k}|=27$$

欧氏距离：

$$d(\vec{x}_1,\vec{x}_2)=\sqrt{\sum_{k=1}^{p}(x_{1k}-x_{2k})^2}=19.416$$

Minkowski距离：

$$d(\vec{x}_1,\vec{x}_2)=\left[\sum_{k=1}^{p}(x_{1k}-x_{2k})^3\right]^{\frac{1}{3}}=17.573$$

Chebyshev距离：

$$d(\vec{x}_1,\vec{x}_2)=\max_{1\le k\le p}|x_{1k}-x_{2k}|=16$$

马氏距离：

$$d(\vec{x}_1,\vec{x}_2)=(\vec{x}_1-\vec{x}_2)'\Sigma^{-1}(\vec{x}_1-\vec{x}_2)=2.2305$$

指标与指标之间的常用“距离”（指标*i*与指标*j*）

相关系数：
$$\rho(X_i, X_j) = \frac{\sum_{k=1}^{p}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{p}(x_{ki} - \overline{x}_i)^2 \sum_{k=1}^{p}(x_{kj} - \overline{x}_j)^2}}$$

夹角余弦：
$$r(X_i, X_j) = \frac{\sum_{k=1}^{p} x_{ki} x_{kj}}{\sqrt{\sum_{k=1}^{p} x_{ki}^2 \sum_{k=1}^{p} x_{kj}^2}}$$

序号	数学	物理	化学	语文	历史	英语
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57

相关系数：$$\rho(\text{数学},\text{语文})=\frac{\sum_{k=1}^{p}\left(x_{ki}-\overline{x}_i\right)\left(x_{kj}-\overline{x}_j\right)}{\sqrt{\sum_{k=1}^{p}\left(x_{ki}-\overline{x}_i\right)^2\sum_{k=1}^{p}\left(x_{kj}-\overline{x}_j\right)^2}}=-0.663$$

夹角余弦：$$r(\text{数学},\text{语文})=\frac{\sum_{k=1}^{p}x_{ki}x_{kj}}{\sqrt{\sum_{k=1}^{p}x_{ki}^2\sum_{k=1}^{p}x_{kj}^2}}=0.983$$

类与类之间的常用距离

1.由一个样品组成的类是最基本的类；如果每一类都由一个样品组成，那么样品间的距离就是类间距离。

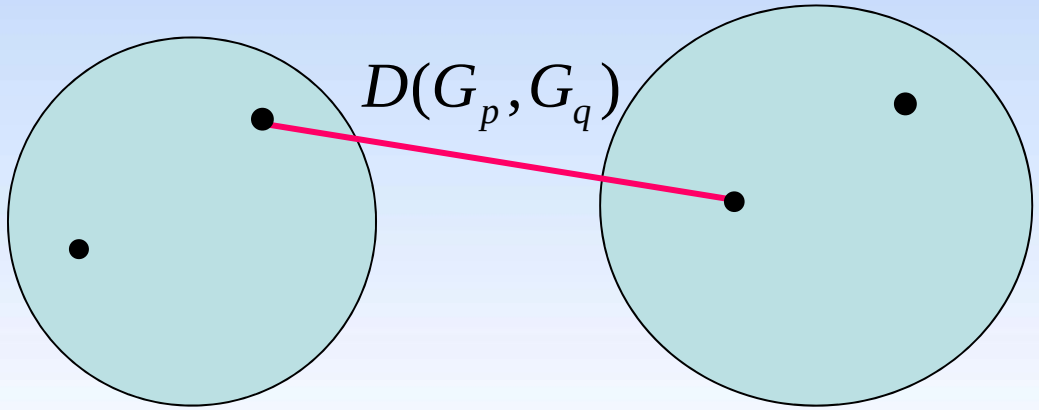
2.如果某一类包含不止一个样品，那么就要确定类间距离，类间距离是基于样品间距离定义的，大致有如下几种定义方式：

记号：G_p和G_q是两个类，$D(G_p,G_q)$是这两个类的距离。

$\vec{x}_i \in G_p \quad \vec{x}_j \in G_q \quad d(\vec{x}_i,\vec{x}_j)$是这两个样品的距离。

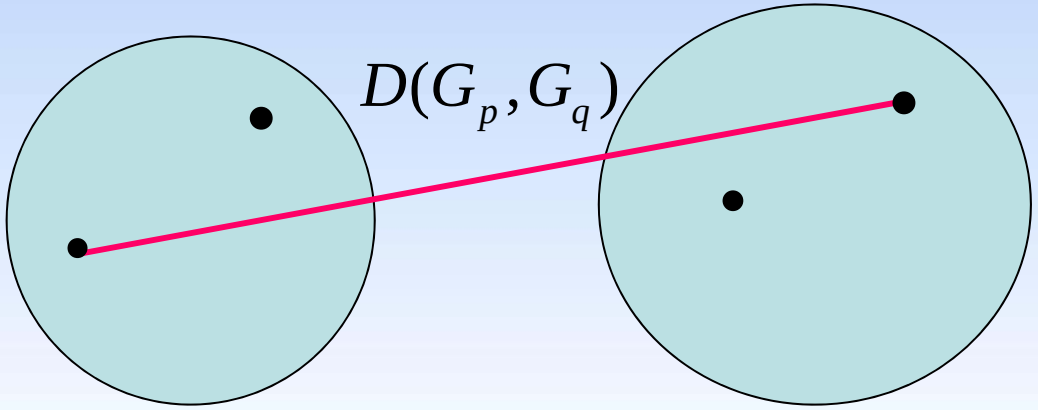
最短距离法:
(Nearest Neighbor)

$$D(G_p, G_q) = \min d(\vec{x}_i, \vec{x}_j)$$



最长距离法:
(Furthest Neighbor)

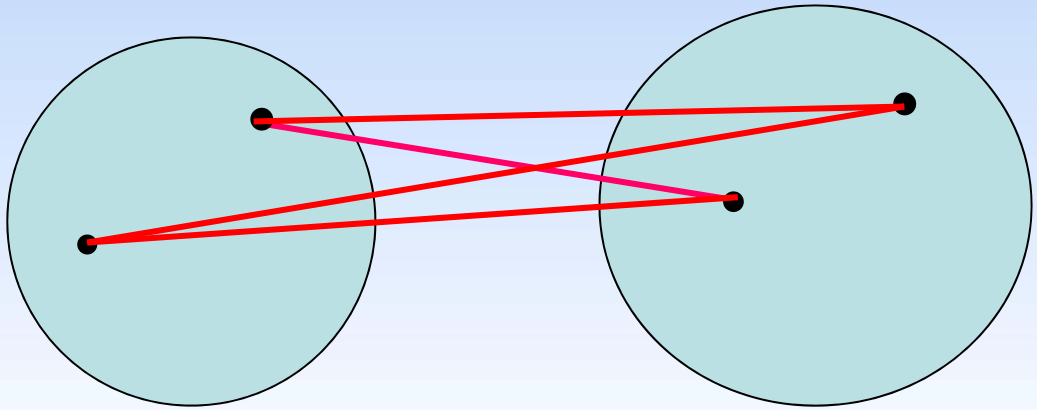
$$D(G_p, G_q) = \max d(\vec{x}_i, \vec{x}_j)$$



组间平均连接法:

(Between-group Linkage)

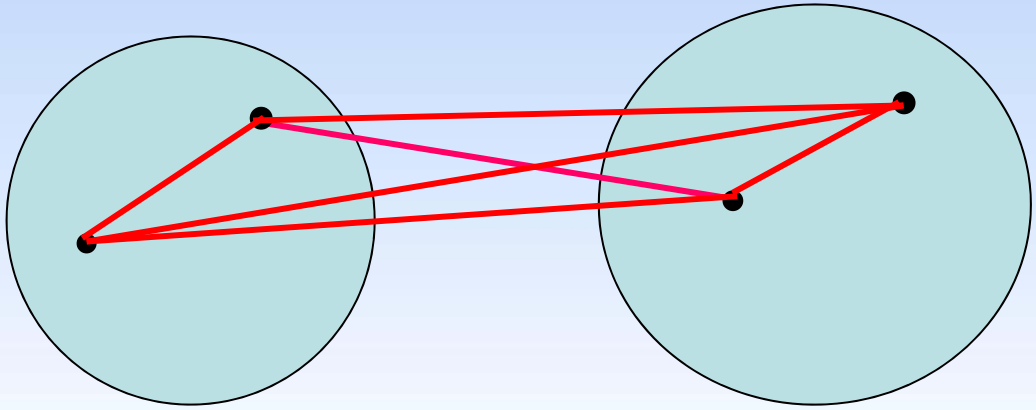
$$D(G_p, G_q) = \frac{d_1 + d_2 + d_3 + d_4}{4}$$



组内平均连接法

(Within-group Linkage)

$$D(G_p, G_q) = \frac{d_1 + d_2 + d_3 + d_4 + d_5 + d_6}{6}$$



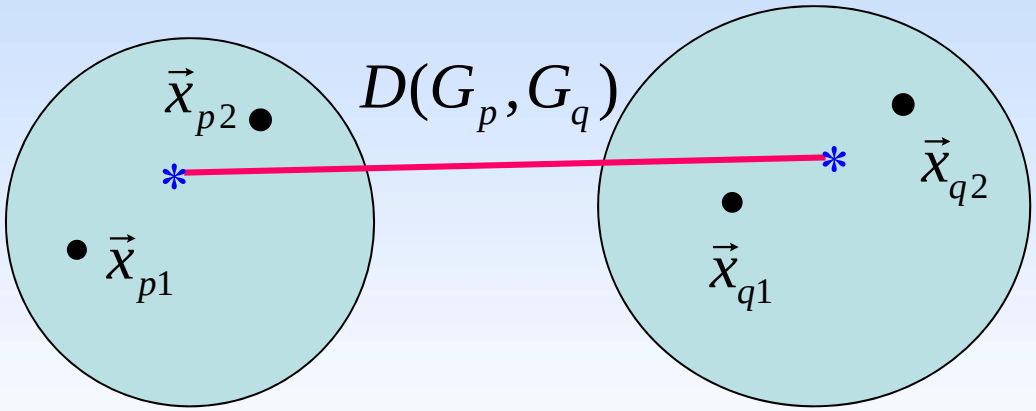
重心法:

(Centroid clustering)

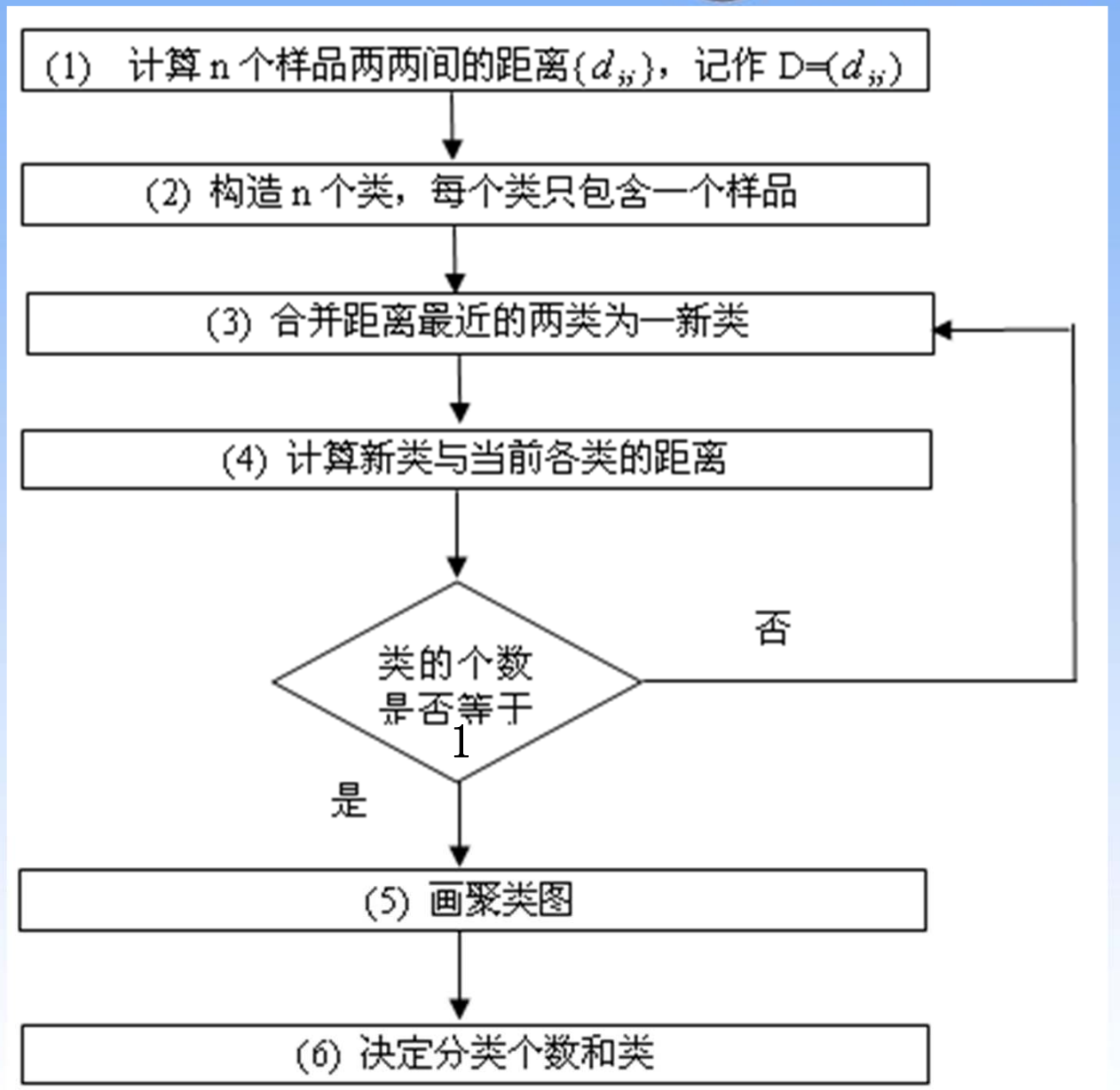
$$D(G_p, G_q) = d(\overline{x}_p, \overline{x}_q)$$

两个类的重心如下:

$$\overline{x}_p = \frac{\vec{x}_{p1} + \vec{x}_{p2}}{2} \qquad \overline{x}_q = \frac{\vec{x}_{q1} + \vec{x}_{q2}}{2}$$



系统聚类法过程


(1) 计算 n 个样品两两间的距离{d_{ij}}，记作 D=(d_{ij})
(2) 构造 n 个类，每个类只包含一个样品
(3) 合并距离最近的两类为一新类
(4) 计算新类与当前各类的距离
类的个数是否等于1
否
是
(5) 画聚类图
(6) 决定分类个数和类

最短距离系统聚类法

根据五个学生的六门课的成绩，对这五个学生进行分类

ID	数学	物理	化学	语文	历史	英语
学生1	65	61	72	84	81	79
学生2	77	77	76	64	70	55
学生3	67	63	49	65	67	57
学生4	80	69	75	74	74	63
学生5	74	70	80	84	81	74

1. 写出样品间的距离矩阵(以欧氏距离为例)

$$D_0=\begin{pmatrix} 0 & & & & \\ 38.9 & 0 & & & \\ 39.7 & 32.2 & 0 & & \\ 26.5 & 15.9 & 32.4 & 0 & \\ 15.8 & 30.9 & 43.6 & 18.2 & 0 \end{pmatrix}\begin{matrix} G_1 \\ G_2 \\ G_3 \\ G_4 \\ G_5 \end{matrix}$$

2. 将每一个样品看做是一个类，即 G_1,G_2,G_3,G_4,G_5

观察 $D(G_1,G_5)=15.8$最小，故将G_1与G_5聚为一类，记为G_6.

计算新类与其余各类之间的距离，得到新的距离矩阵 D_1

$$D(G_6,G_2)=\min\{D(G_1,G_2),D(G_5,G_2)\}=\min\{38.9,30.9\}=30.9$$

$$D(G_6,G_3)=\min\{D(G_1,G_3),D(G_5,G_3)\}=\min\{39.7,43.6\}=39.7$$

$$D(G_6,G_4)=\min\{D(G_1,G_4),D(G_5,G_4)\}=\min\{26.5,18.2\}=18.2$$

$$D_1=\begin{pmatrix} 0 & & & \\ 30.9 & 0 & & \\ 39.7 & 32.2 & 0 & \\ 18.2 & 15.9 & 32.4 & 0 \end{pmatrix}\begin{matrix} G_6 \\ G_2 \\ G_3 \\ G_4 \end{matrix}$$

3. 观察 $D(G_2,G_4)=15.9$最小，故将G_2与G_4聚为一类，记为G_7.

计算新类与其余各类之间的距离，得到新的距离矩阵D_2

$$D(G_7,G_6)=\min\{D(G_2,G_6),D(G_4,G_6)\}=\min\{30.9,18.2\}=18.2$$

$$D(G_7,G_3)=\min\{D(G_2,G_3),D(G_4,G_3)\}=\min\{32.2,32.4\}=32.2$$

$$D_2=\begin{pmatrix}0 & & \\ 18.2 & 0 & \\ 32.2 & 39.7 & 0\end{pmatrix}\begin{matrix}G_7\\ G_6\\ G_3\end{matrix}$$

4. 观察 $D(G_6,G_7)=18.2$最小，故将G_6与G_7聚为一类，记为G_8.

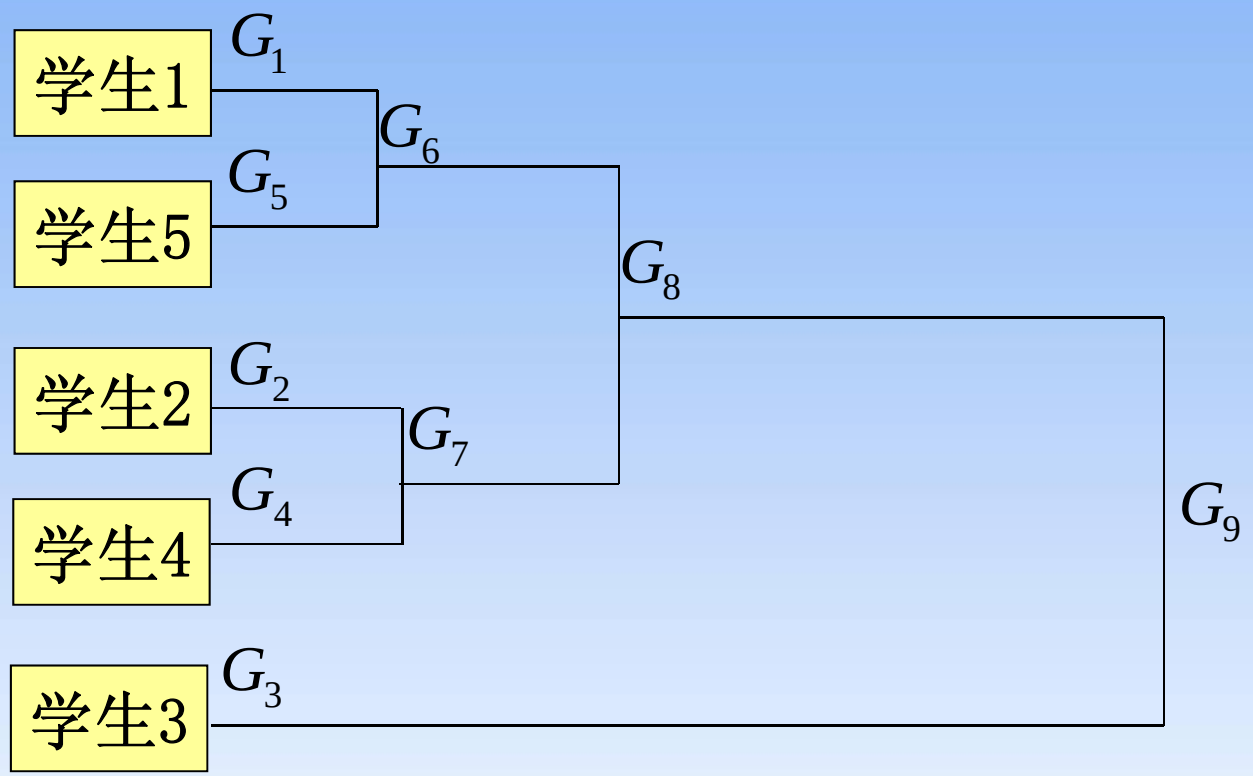
计算新类与其余各类之间的距离，得到新的距离矩阵 D_3

$$D(G_8,G_3)=\min\{D(G_6,G_3),D(G_7,G_3)\}=\min\{39.7,32.2\}=32.2$$

$$D_3=\begin{pmatrix}0 & \\ 32.2 & 0\end{pmatrix}\begin{matrix}G_8\\ G_3\end{matrix}$$

5. 最后将G_8与G_3聚为一类，记为G_9.

聚类的谱系图



最长距离系统聚类法

1. 写出样品间的距离矩阵(以欧氏距离为例)

$$D_0=\begin{pmatrix} 0 & & & & \\ 38.9 & 0 & & & \\ 39.7 & 32.2 & 0 & & \\ 26.5 & 15.9 & 32.4 & 0 & \\ 15.8 & 30.9 & 43.6 & 18.2 & 0 \end{pmatrix}\begin{matrix} G_1 \\ G_2 \\ G_3 \\ G_4 \\ G_5 \end{matrix}$$

2. 将每一个样品看做是一个类，即 G_1,G_2,G_3,G_4,G_5

观察 $D(G_1,G_5)=15.8$最小，故将G_1与G_5聚为一类，记为G_6.

计算新类与其余各类之间的距离，得到新的距离矩阵 D_1

$$D(G_6,G_2)=\max\{D(G_1,G_2),D(G_5,G_2)\}=\max\{38.9,30.9\}=38.9$$

$$D(G_6,G_3)=\max\{D(G_1,G_3),D(G_5,G_3)\}=\max\{39.7,43.6\}=43.6$$

$$D(G_6,G_4)=\max\{D(G_1,G_4),D(G_5,G_4)\}=\max\{26.5,18.2\}=26.5$$

$$D_1=\begin{pmatrix} 0 & & & \\ 38.9 & 0 & & \\ 43.6 & 32.2 & 0 & \\ 26.5 & 15.9 & 32.4 & 0 \end{pmatrix}\begin{matrix} G_6 \\ G_2 \\ G_3 \\ G_4 \end{matrix}$$

3. 观察 $D(G_2,G_4)=15.9$最小，故将G_2与G_4聚为一类，记为G_7.

计算新类与其余各类之间的距离，得到新的距离矩阵 D_2

$$D(G_7,G_6)=\max\{D(G_2,G_6),D(G_4,G_6)\}=\max\{38.9,26.5\}=38.9$$

$$D(G_7,G_3)=\max\{D(G_2,G_3),D(G_4,G_3)\}=\max\{32.2,32.4\}=32.4$$

$$D_2=\begin{pmatrix}0&&\\38.9&0&\\32.4&43.6&0\end{pmatrix}\begin{matrix}G_7\\G_6\\G_3\end{matrix}$$

4. 观察 $D(G_3,G_7)=32.4$最小，故将G_3与G_7聚为一类，记为G_8.

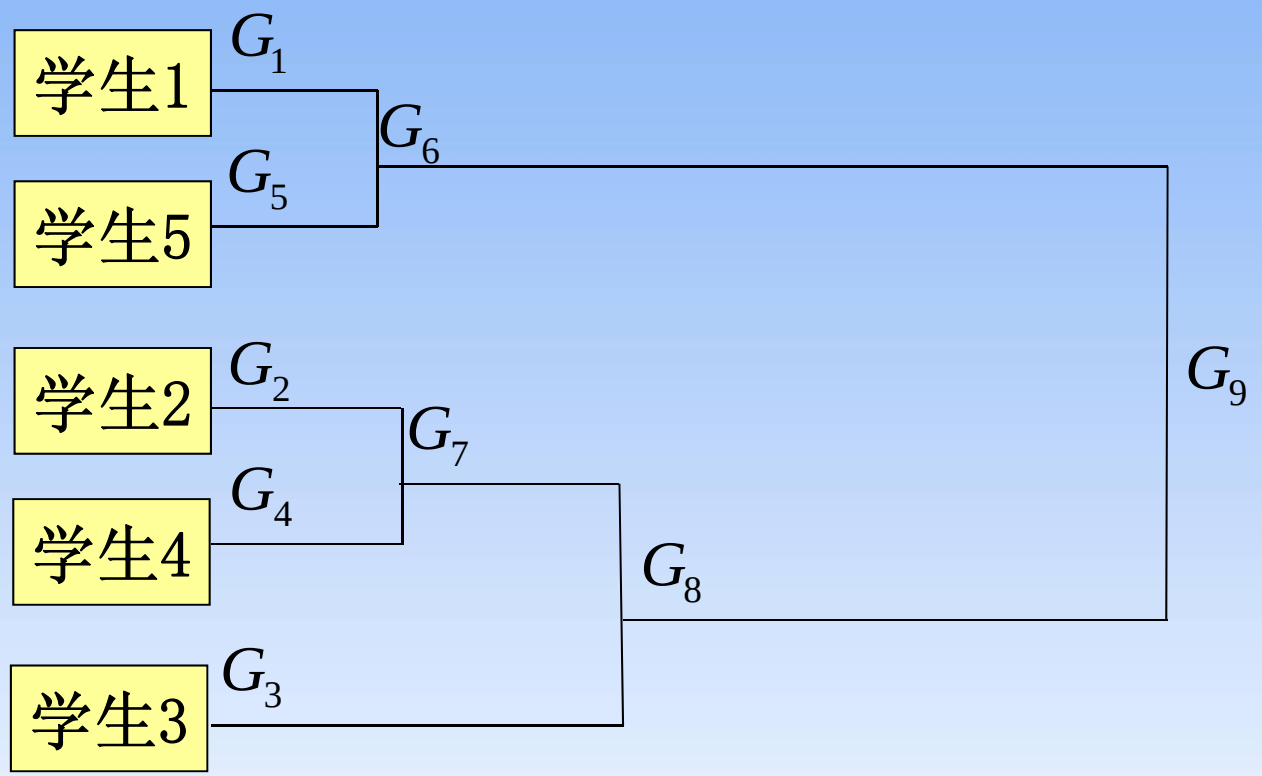
计算新类与其余各类之间的距离，得到新的距离矩阵 D_3

$$D(G_8,G_6)=\max\{D(G_3,G_6),D(G_7,G_6)\}=\max\{43.6,38.9\}=43.6$$

$$D_3=\begin{pmatrix}0&\\43.6&0\end{pmatrix}\begin{matrix}G_8\\G_6\end{matrix}$$

5. 最后将G_8与G_6聚为一类，记为G_9.

聚类的谱系图



其它系统聚类法

组间平均连接系统聚类法

组内平均连接系统聚类法

重心系统聚类法

注：这些方法的差别就是在计算新类与其余各类间的距离，如需学习详细内容，可参考多元统计分析相关书籍。

参考教材：《多元统计分析》，何晓群，中国人民大学出版社，2008.
《多元统计分析》，于秀林，中国统计出版社，2006.

聚类分析需要注意的问题

1. 对于一个实际问题要根据分类的目的来选取指标，指标选取的不同分类结果一般也不同。

2. 样品间距离定义方式的不同，聚类结果一般也不同。

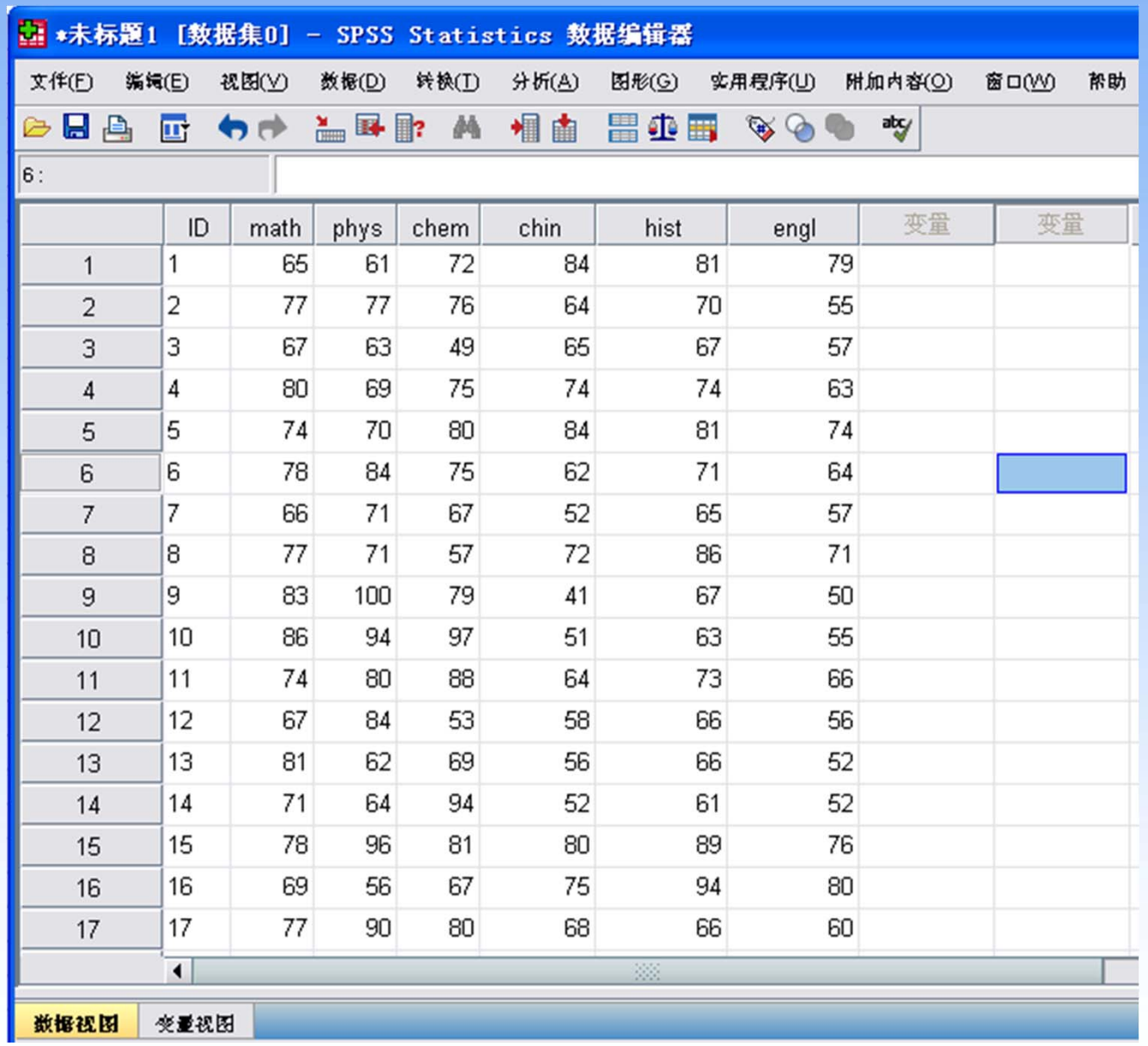
3. 聚类方法的不同，聚类结果一般也不同（尤其是样品特别多的时候）。最好能通过各种方法找出其中的共性。

4. 要注意指标的量纲，量纲差别太大会导致聚类结果不合理。

5. 聚类分析的结果可能不令人满意，因为我们所做的是一个数学的处理，对于结果我们要找到一个合理的解释。

系统聚类法的SPSS实现

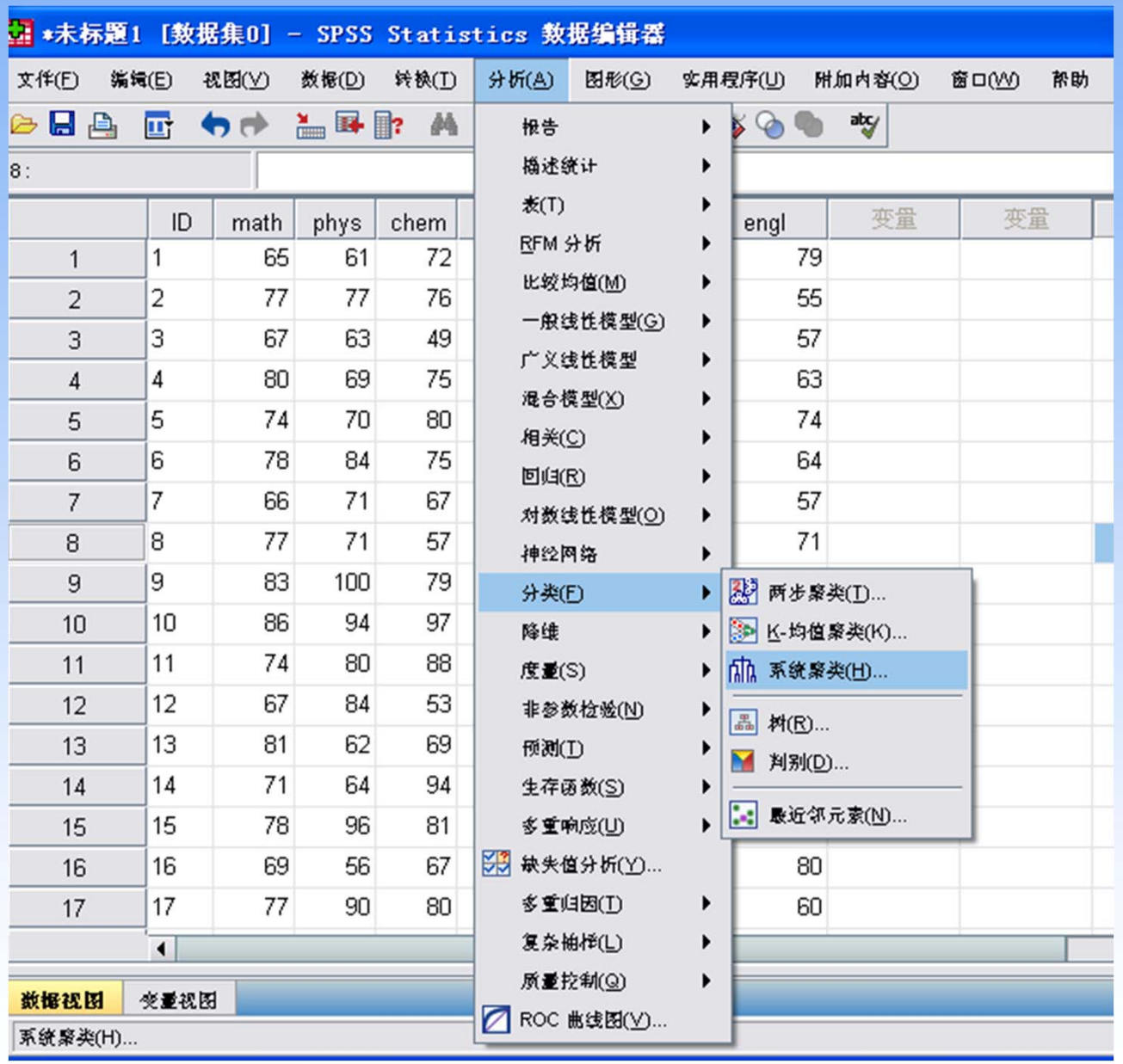
录入数据



	ID	math	phys	chem	chin	hist	engl	变量	变量
1	1	65	61	72	84	81	79		
2	2	77	77	76	64	70	55		
3	3	67	63	49	65	67	57		
4	4	80	69	75	74	74	63		
5	5	74	70	80	84	81	74		
6	6	78	84	75	62	71	64		
7	7	66	71	67	52	65	57		
8	8	77	71	57	72	86	71		
9	9	83	100	79	41	67	50		
10	10	86	94	97	51	63	55		
11	11	74	80	88	64	73	66		
12	12	67	84	53	58	66	56		
13	13	81	62	69	56	66	52		
14	14	71	64	94	52	61	52		
15	15	78	96	81	80	89	76		
16	16	69	56	67	75	94	80		
17	17	77	90	80	68	66	60		

系统聚类法的SPSS实现

选择方法



系统聚类法的SPSS实现

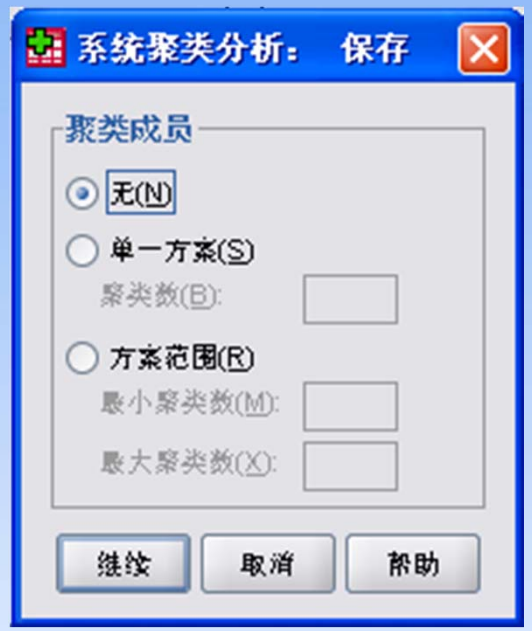
统计量选项

绘制选项

系统聚类法的SPSS实现

方法选项

保存选项



谢 谢 大 家