聚类分析

Cluster Analysis

§1　什么是聚类分析

- 聚类分析是研究分类问题的一种多元统计方法。所谓类，就是指相似元素的集合
- 聚类分析的研究目的

  把相似的东西归成类，根据相似的程度将研究目标进行分类。

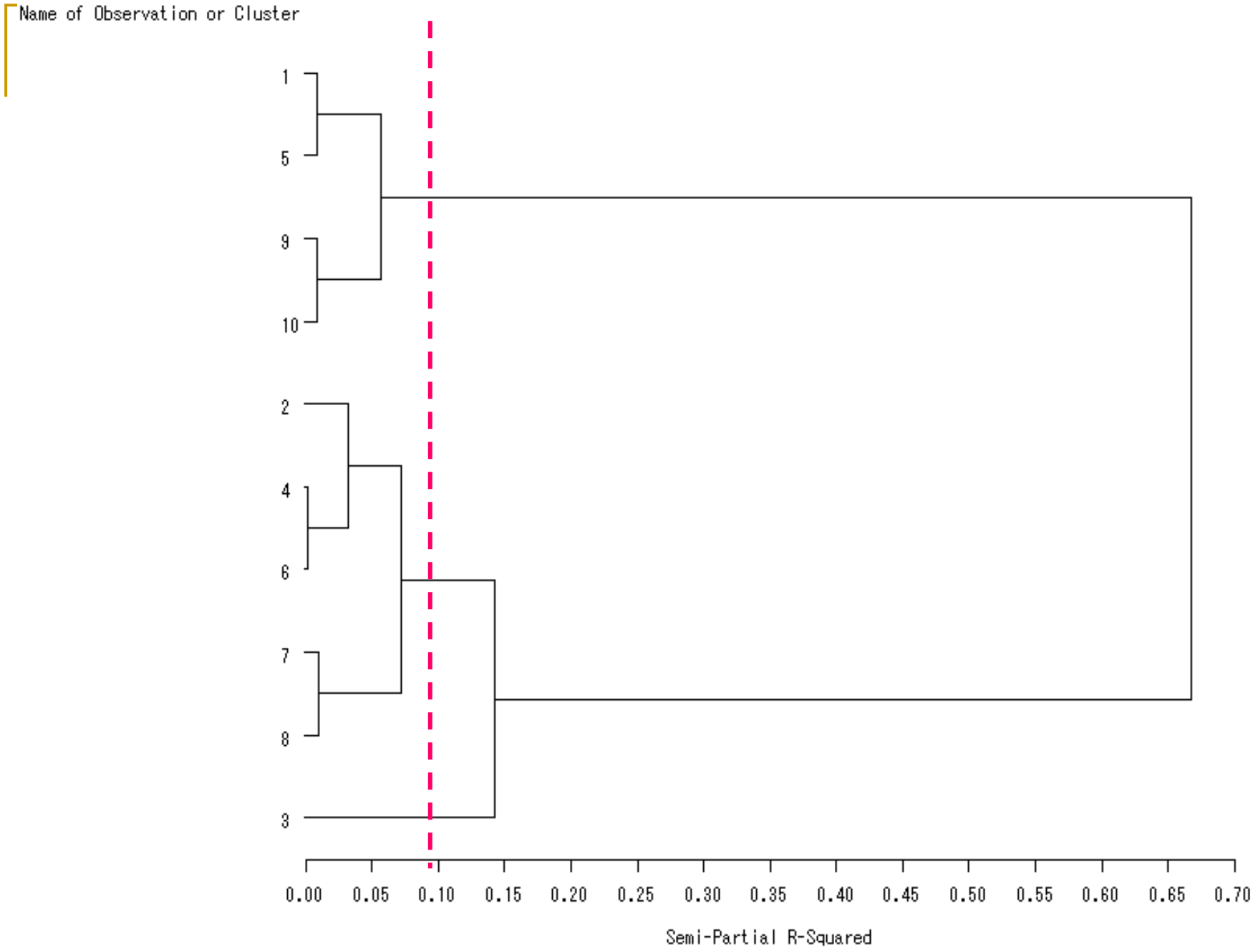
- 聚类分析的研究对象
  - R型分析----对变量进行分类
  - Q型分析----对样品进行分类

- 聚类分析研究的主要内容
  - 如何度量事物之间的**相似性**？
  - 怎样构造聚类的具体方法以达到分类的目的?

例 对10位应聘者做智能检验。3项指标X、Y和Z分别表示数学推理能力、空间想象能力和语言理解能力。其得分如下，选择合适的统计方法对应聘者进行分类。

应聘者	1	2	3	4	5	6	7	8	9	10
X	28	18	11	21	26	20	16	14	24	22
Y	29	23	22	23	29	23	22	23	29	27
Z	28	18	16	22	26	22	22	24	24	24

我们的问题是如何来选择样品间相似性的测度指标，如何将相似的类连接起来？


Name of Observation or Cluster
1
5
9
10
2
4
6
7
8
3
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
Semi-Partial R-Squared

§2　距离和相似系数

一、相似性的测度

- **距离**：测度样品之间的亲疏程度。将每一个样品看作p 维空间的一个点，并用某种度量测量点与点之间的距离，距离较近的归为一类，距离较远的点应属于不同的类。
- **相似系数**：测度变量之间的亲疏程度

2、常用的距离

(1) 明氏（Minkowski）距离

设原始数据为

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

令 d_{ij} 表示样品 x_i 与 x_j 的距离

$$d_{ij} = \left(\sum_{l=1}^{p} | x_{il} - x_{jl} |^k\right)^{\frac{1}{k}}$$

2、常用的距离

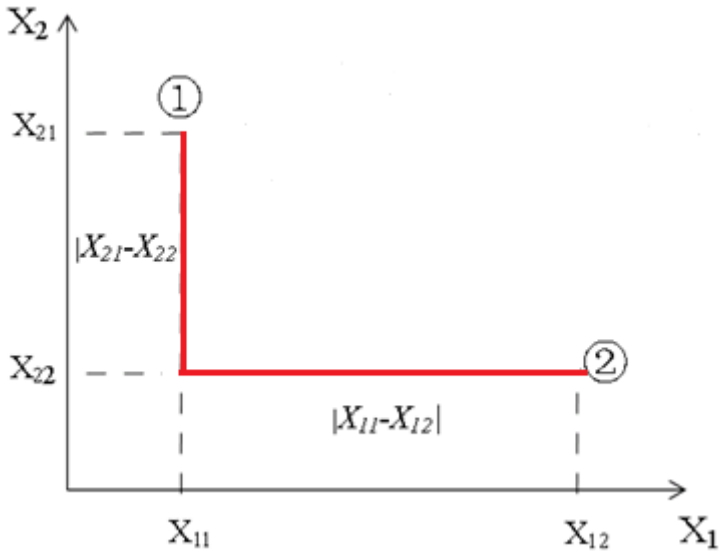
(1) 明氏（Minkowski）距离

$$d_{ij} = \left(\sum_{l=1}^{p} | x_{il} - x_{jl} |^{k}\right)^{\frac{1}{k}}$$

特别地，当k＝1时，即为**绝对值距离**

$$d_{ij} = \sum_{l=1}^{p} \left| x_{il} - x_{jl} \right|$$

CityBlock Distance
曼哈顿距离

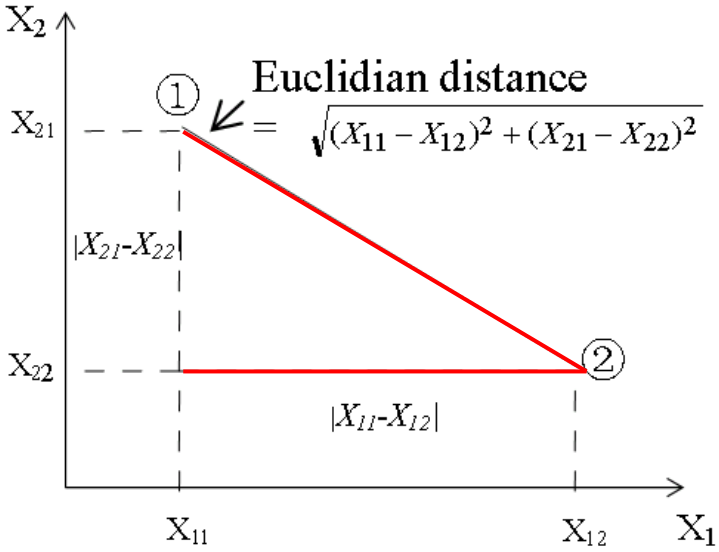


明氏距离

$$d_{ij} = \left(\sum_{l=1}^{p} | x_{il} - x_{jl} |^{k}\right)^{\frac{1}{k}}$$

当k＝2时，即为**欧氏(Euclidian)距离**

$$d_{ij} = \sqrt{\sum_{l=1}^{p} (x_{il} - x_{jl})^{2}}$$



欧氏距离有明确的空间距离概念

当k＝∞时，即为**切比雪夫距离**

$$d_{ij} = \max_{1 \le l \le p} \left| x_{il} - x_{jl} \right|$$

	x_1	x_2	x_3
1	20	7	25.2
2	18	10	36.3
3	10	5	28.9
4	4	5	11.5
5	4	3	17

计 算 d_{24}

欧氏距离

$$d_{24} = \sqrt{\sum_{l=1}^{3} (x_{2l} - x_{4l})^2}$$

$$= \sqrt{(18-4)^2 + (10-5)^2 + (36.3-11.5)^2}$$

切比雪夫距离

$$d_{24} = \max_{1 \le l \le 3} |x_{2l} - x_{4l}| = |36.3 - 11.5| = 24.8$$

明氏距离的两个缺点：

①明氏距离的数值与指标的**量纲**有关

②没有考虑各个变量之间相关性的影响

	年龄	月收入	家庭人口数	月消费支出
甲	30	20000	1	8000
乙	40	22000	3	9000

$$d = \sqrt{(30-40)^2 + (20000-22000)^2 + (1-3)^2 + (8000-8500)^2}$$

(2) 标准化的欧氏距离

设原始数据为

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

$$d_{ij} = \sqrt{\left(\frac{x_{i1}-x_{j1}}{\sqrt{s_{11}}}\right)^2 + \left(\frac{x_{i2}-x_{j2}}{\sqrt{s_{22}}}\right)^2 + \ldots + \left(\frac{x_{ip}-x_{jp}}{\sqrt{s_{pp}}}\right)^2}$$

$$= \sqrt{\frac{1}{s_{11}}\left(x_{i1}-x_{j1}\right)^2 + \frac{1}{s_{22}}\left(x_{i2}-x_{j2}\right)^2 + \ldots + \frac{1}{s_{pp}}\left(x_{ip}-x_{jp}\right)^2}$$

$$= \sqrt{\sum_{l=1}^{p} \frac{\left(x_{il}-x_{jl}\right)^2}{s_{ll}}}$$

$$\begin{bmatrix} \frac{x_{11}-\bar{x}_1}{\sqrt{S_{11}}} & \frac{x_{12}-\bar{x}_2}{\sqrt{S_{22}}} & \cdots & \frac{x_{1p}-\bar{x}_p}{\sqrt{S_{pp}}} \\ \frac{x_{21}-\bar{x}_1}{\sqrt{S_{11}}} & \frac{x_{22}-\bar{x}_2}{\sqrt{S_{22}}} & \cdots & \frac{x_{2p}-\bar{x}_p}{\sqrt{S_{pp}}} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \frac{x_{n1}-\bar{x}_1}{\sqrt{S_{11}}} & \frac{x_{n2}-\bar{x}_2}{\sqrt{S_{22}}} & \cdots & \frac{x_{np}-\bar{x}_p}{\sqrt{S_{pp}}} \end{bmatrix}$$

(3)马氏距离

由印度著名统计学家马哈拉诺比斯(Mahalanobis)所定义的一种距离，其计算公式为：

$$d_{ij} = \left[\left(x_{i1} - x_{j1}, x_{i2} - x_{j2}, \cdots, x_{ip} - x_{jp} \right) S^{-1} \begin{pmatrix} x_{i1} - x_{j1} \\ x_{i2} - x_{j2} \\ \vdots \\ x_{ip} - x_{jp} \end{pmatrix} \right]^{\frac{1}{2}}$$

$$= \left[\left(x_i - x_j \right)' S^{-1} \left(x_i - x_j \right) \right]^{\frac{1}{2}}$$

- 马氏距离又称为广义欧氏距离。
- 马氏距离考虑了观测变量之间的相关性。如果假定各变量之间相互独立，即观测变量的协方差矩阵是对角矩阵，此时马氏距离就是标准化的欧氏距离。
- 马氏距离不受指标量纲及**指标间相关性**的影响

（4）兰氏 Canberra距离

$$d_{ij} = \frac{1}{p}\sum_{l=1}^{p}\frac{\left|x_{il} - x_{jl}\right|}{x_{il} + x_{jl}}$$

- 不受量纲影响
- 对大的奇异值不敏感，特别适合于高度偏倚的数据
- 没有考虑指标之间的相关性

二、变量间相似系数的算法

（1）相关系数

变量 x_j 和 x_k 的相关系数：

$$r_{jk}=\frac{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)\left(x_{ik}-\bar{x}_k\right)}{\left[\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)^2\sum_{i=1}^{n}\left(x_{ik}-\bar{x}_k\right)^2\right]^{\frac{1}{2}}}=\frac{\sigma_{xy}}{\sqrt{s_{xx}s_{yy}}}$$

（2）夹角余弦

$$c_{jk}=\frac{\sum_{i=1}^{n}x_{ij}x_{ik}}{\left(\sum_{i=1}^{n}x_{ij}^2\sum_{i=1}^{n}x_{ik}^2\right)^{\frac{1}{2}}}$$

各种聚类方法

- 系统聚类法 hierarchical clustering method

  简单，直观。

- 快速聚类法（动态聚类法）

  快速，动态。

§3 系统聚类法

系统聚类法的基本思想

先将n个样品各自看成一类，然后规定样品之间的“距离”和类与类之间的距离。选择**距离最近**的两类合并成一个新类，计算新类和其它类（各当前类）的距离，再将距离最近的两类合并。这样，每次合并减少一类，**直至所有的样品都归成一类为止**。

系统聚类法的基本步骤：

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. **计算新类与各当前类的距离**。
5. 重复步骤3、4，合并距离最近的两类为新类，直到所有的类并为一类为止。
6. 画聚类谱系图。
7. 决定类的个数和类。

系统聚类方法：

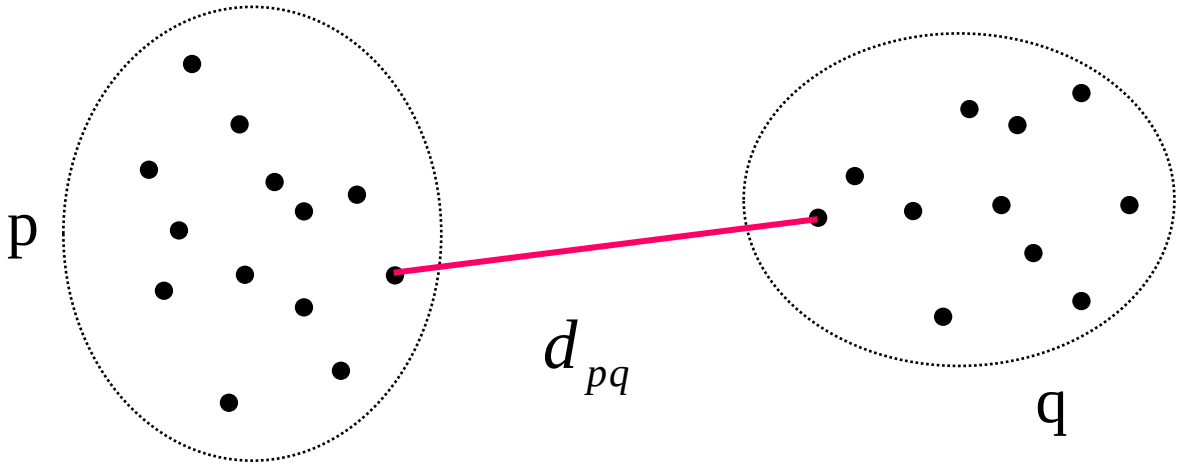
1. 最短距离法 (Single linkage)
2. 最长距离法 (Complete method)
3. 中间距离法 (Median method)
4. 重心法 (Centroid method)
5. 类平均法 (Average linkage)
6. 离差平方和法（Ward method ）

上述 6 种方法归类的**基本步骤一致**，只是**类与类之间的距离**有不同的定义。

一、最短距离法

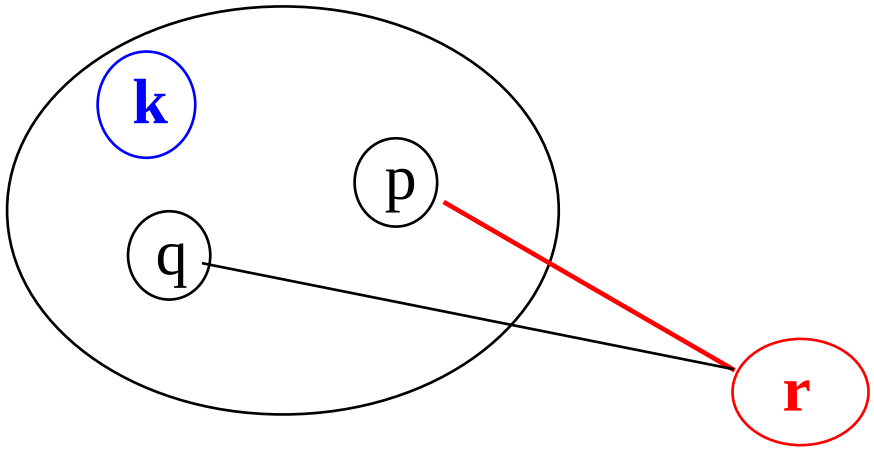
定义类p与q之间的距离为两类最近样品的距离，即

$$d_{pq} = \min_{i \in p, j \in q} \{d_{ij}\}$$



设类p与 q合并成一个新类，记为k，则k与任一类r 的距离是

$$d_{kr} = \min\{d_{pr} \quad , \quad d_{qr}\}$$



例　最短距离法

设抽取5个样品，每个样品观察2个指标，

x_1：您每月大约喝多少瓶啤酒，

x_2：您对“饮酒是人生的快乐”这句话的看法如何？观察数据如下，对这5个样品分类。

	x_1	x_2
1	20	7
2	18	10
3	10	5
4	4	5
5	4	3

1.计算5个样品两两之间的距离 d_{ij}（采用欧氏距离），

记为距离矩阵 $D=(d_{ij})_{n\times n}$

	②	③	④	⑤
①	3.6	10.2	16.12	16.49
②		9.43	14.87	15.65
③			6	6.32
④				**2**

2. 合并距离最小的两类为新类，按顺序定为第 6 类。

$d_{45}=2$为最小，　⑥$=\{4,5\}$

3、计算新类⑥与各当前类的距离，

$$d_{61}=\min\{d_{41},d_{51}\}=\min\{16.12,16.49\}=16.12$$

$$d_{62}=\min\{d_{42},d_{52}\}=\min\{14.87,15.65\}=14.87$$

$$d_{63}=\min\{d_{43},d_{53}\}=6$$

得距离矩阵如下：

	②	③	⑥
①	**3.6**	10.2	16.12
②		9.43	14.87
③			6

4、重复步骤2、3，合并距离最近的两类为新类，直到所有的类并为一类为止。

$d_{12}=3.6$ 为最小，⑦= $\{1,2\}$

$d_{73}=\min\{d_{13},d_{23}\}=9.43$　　$d_{76}=\min\{d_{16},d_{26}\}=14.87$

	⑥	⑦
③	**6**	9 .43
⑥		14.87

5、$d_{36}=6$ 为最小，⑧= $\{3,6\}$

$d_{87}=\min\{d_{37},d_{67}\}=9.43$

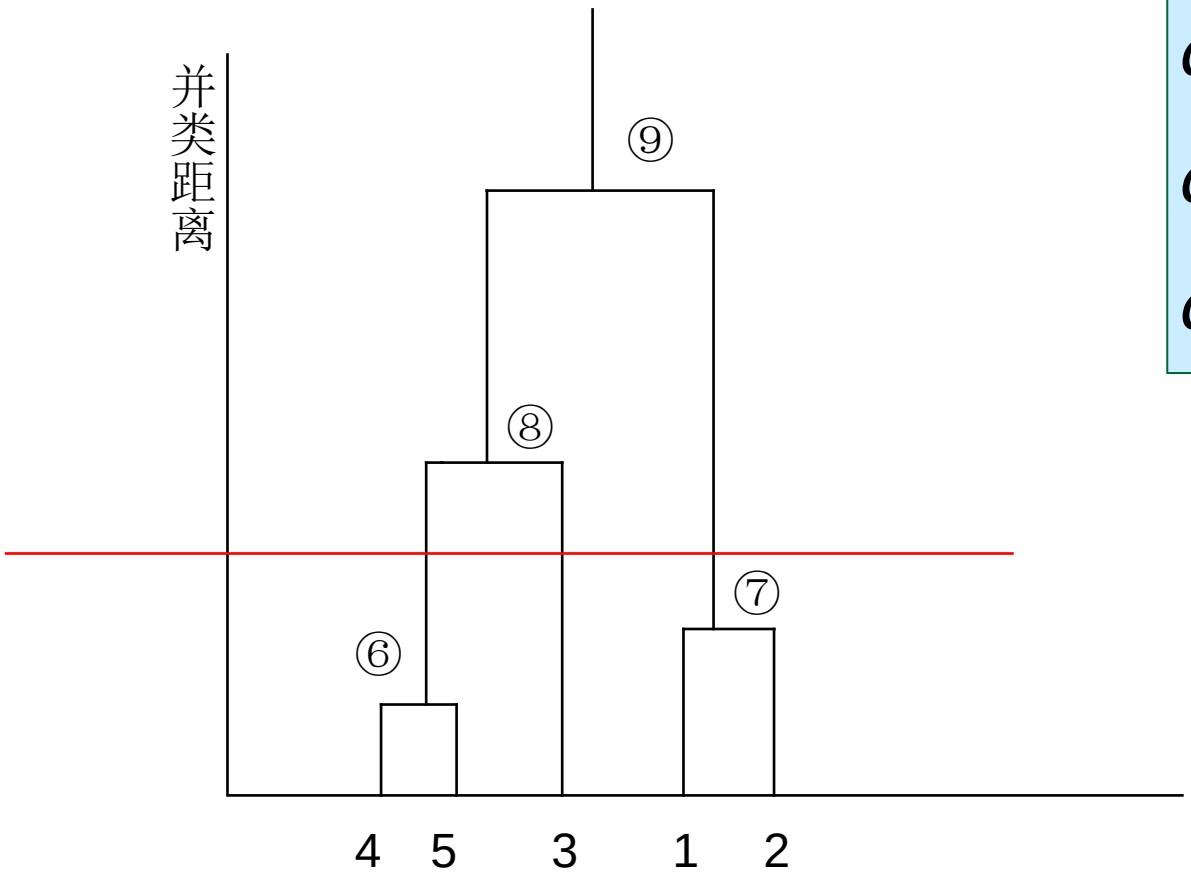
6、按聚类的过程画聚类谱系图

$$d_{4,5}=2$$

$$d_{1,2}=3.6$$

$$d_{3,6}=6$$

$$d_{7,8}=9.43$$



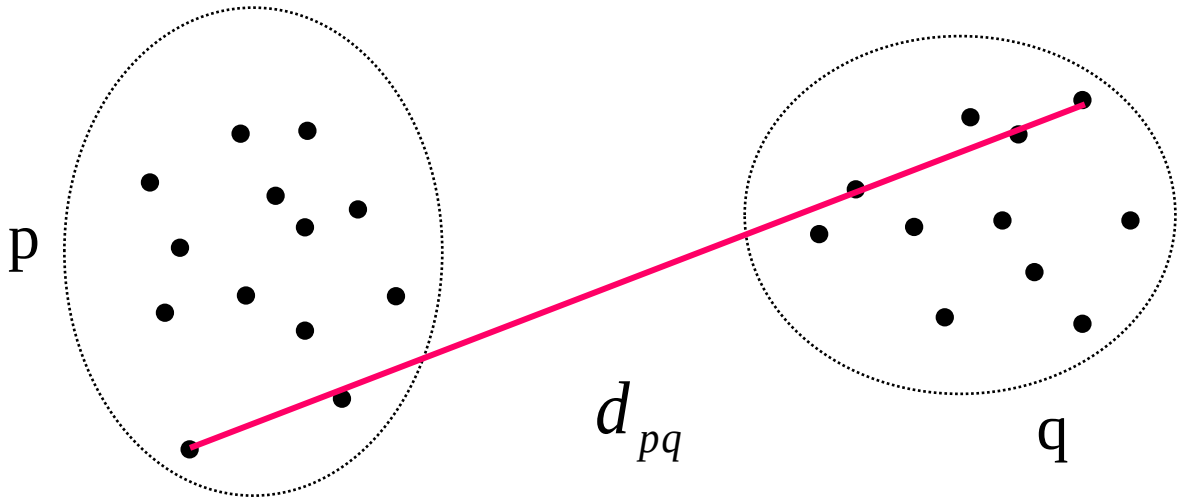
7、决定类的个数与类。

观察此图，我们可以把5个样品分为3类，$\{1,2\}$、$\{3\}$、$\{4,5\}$。

二、最长距离法

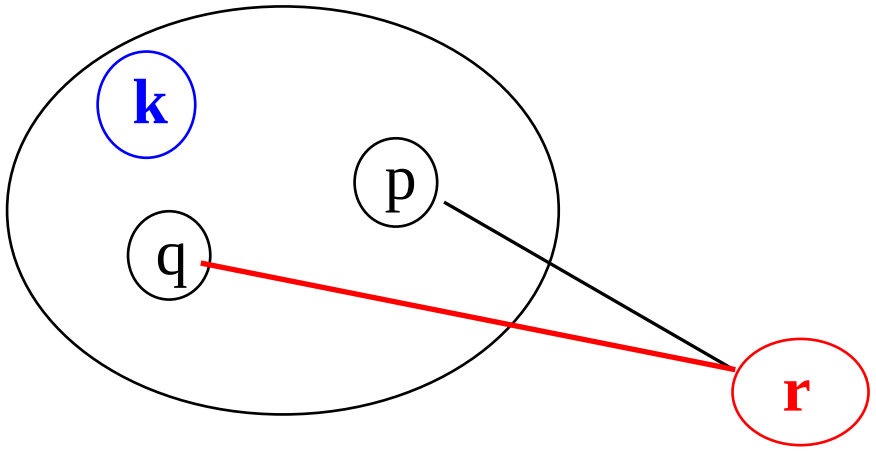
定义类p与q之间的距离为两类最远样品的距离，即

$$d_{pq} = \max_{i \in p, j \in q} \{d_{ij}\}$$



设类p与 q合并成一个新类，记为k，则k与任一类r 的距离是

$$d_{kr} = \max\{d_{pr} \quad , \quad d_{qr}\}$$



例 最长距离法

1.计算5个样品两两之间的距离 d_{ij}（采用欧氏距离），

记为距离矩阵 $D=(d_{ij})_{n\times n}$

	②	③	④	⑤
①	3.6	10.2	16.12	16.49
②		9.43	14.87	15.65
③			6	6.32
④				**2**

2. 合并距离最小的两类为新类，按顺序定为第 6 类。

$d_{45}=2$为最小，　⑥$=\{4,5\}$

3、计算新类⑥与各当前类的距离，

$$d_{61}=\max\{d_{41},d_{51}\}=\max\{16.12,16.49\}=16.49$$

$$d_{62}=\max\{d_{42},d_{52}\}=\max\{14.87,15.65\}=15.65$$

$$d_{63}=\max\{d_{43},d_{53}\}=6.32$$

得距离矩阵如下：

	②	③	⑥
①	**3.6**	10.2	16.49
②		9.43	15.65
③			6.32

4、重复步骤2、3，合并距离最近的两类为新类，直到所有的类并为一类为止。

$d_{12}=3.6$ 为最小，⑦= $\{1,2\}$

$d_{73}=\max\{d_{13},d_{23}\}=10.2$　　$d_{76}=\max\{d_{16},d_{26}\}=16.49$

	⑥	⑦
③	**6.32**	10 .2
⑥		16.49

5、$d_{36}=6.32$ 为最小，⑧= $\{3,6\}$

$d_{87}=\max\{d_{37},d_{67}\}=16.49$

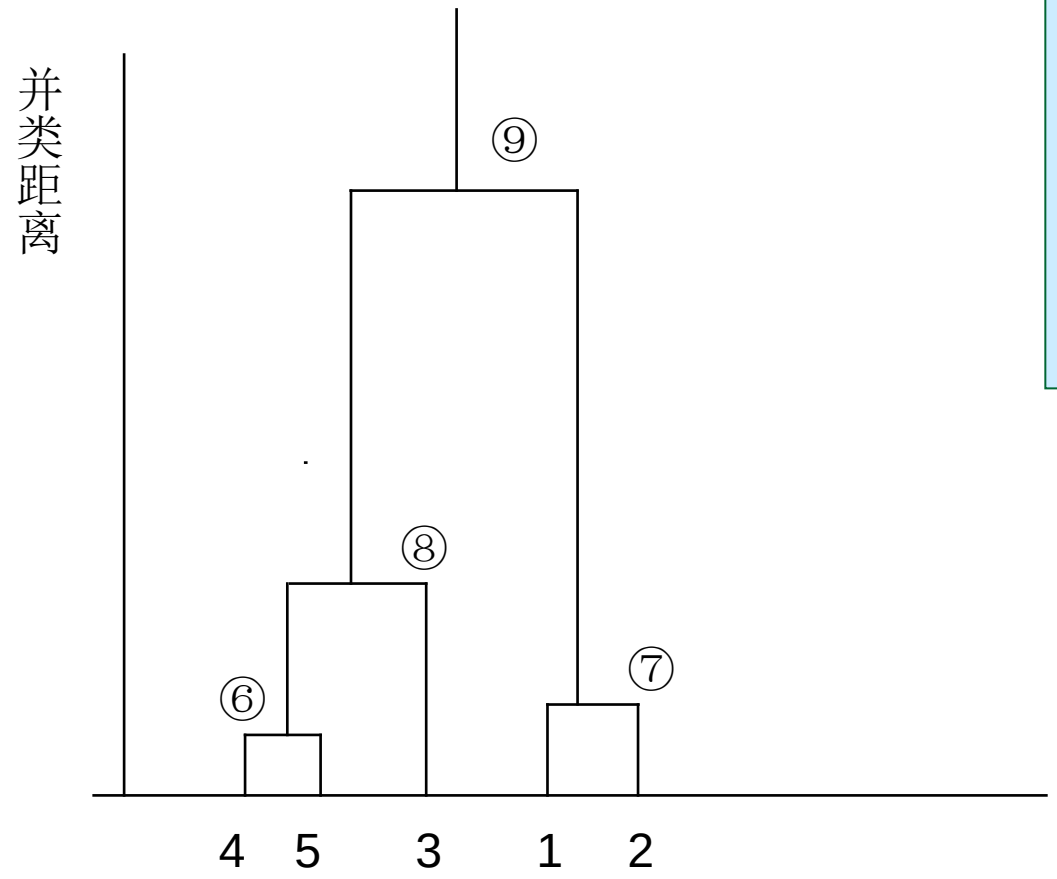
6、按聚类的过程画聚类谱系图

$$d_{4,5}=2$$

$$d_{1,2}=3.6$$

$$d_{3,6}=6.32$$

$$d_{7,8}=16.49$$

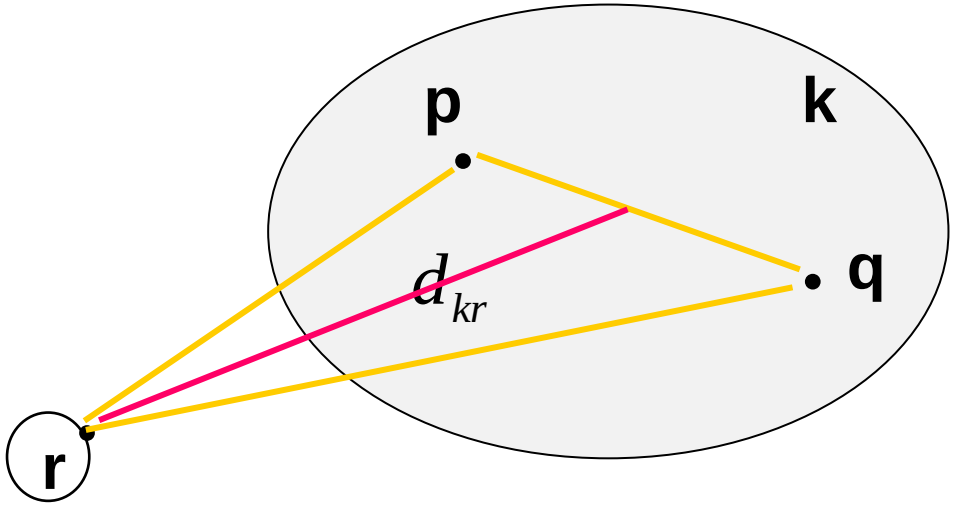


7、决定类的个数与类。

观察此图，我们可以把5个样品分为3类，$\{1,2\}$、$\{3\}$、$\{4,5\}$。

三、中间距离法

定义类与类之间的距离既不采用两类之间最近的距离，也不采用两类之间最远的距离，而是采用介于两者之间的距离，故称为中间距离法。



$$d_{kr}^2 = \frac{1}{2}d_{pr}^2 + \frac{1}{2}d_{qr}^2 - \frac{1}{4}d_{pq}^2$$

例　中间距离法

1.计算5个样品两两之间的距离 d_{ij}（采用欧氏距离），

记为距离矩阵 $D=(d_{ij})_{n\times n}$

d_{ij}^2	②	③	④	⑤
①	13	104	260	272
②		89	221	245
③			36	40
④				**4**

2. 合并距离最小的两类为新类，按顺序定为第 6 类。

$d_{45}^2=4$为最小，　⑥$=\{4,5\}$

3、计算新类⑥与各当前类的距离，

$$d_{61}^2=\frac{1}{2}d_{41}^2+\frac{1}{2}d_{51}^2-\frac{1}{4}d_{45}^2=\frac{1}{2}\times 260+\frac{1}{2}\times 272-\frac{1}{4}\times 4=265$$

$$d_{62}^2=\frac{1}{2}d_{42}^2+\frac{1}{2}d_{52}^2-\frac{1}{4}d_{45}^2=\frac{1}{2}\times 221+\frac{1}{2}\times 245-\frac{1}{4}\times 4=232$$

$$d_{63}^2=\frac{1}{2}d_{43}^2+\frac{1}{2}d_{53}^2-\frac{1}{4}d_{45}^2=\frac{1}{2}\times 36+\frac{1}{2}\times 40-\frac{1}{4}\times 4=37$$

得距离矩阵如下：

d_{ij}^2	②	③	⑥
①	**13**	104	265
②		89	232
③			37

4、重复步骤2、3，合并距离最近的两类为新类，直到所有的类并为一类为止。

$d_{12}^2 = 13$ 为最小，⑦= $\{1,2\}$

$$d_{73}^2 = \frac{1}{2}d_{13}^2 + \frac{1}{2}d_{23}^2 - \frac{1}{4}d_{12}^2 = \frac{1}{2}\times 104 + \frac{1}{2}\times 89 - \frac{1}{4}\times 13 = 93.25$$

$$d_{76}^2 = \frac{1}{2}d_{16}^2 + \frac{1}{2}d_{26}^2 - \frac{1}{4}d_{12}^2 = \frac{1}{2}\times 265 + \frac{1}{2}\times 232 - \frac{1}{4}\times 13 = 245.25$$

d_{ij}^2	⑥	⑦
③	**37**	93.25
⑥		245.25

5、$d_{36}^2 = 37$ 为最小，⑧= $\{3,6\}$

$$d_{87}^2 = \frac{1}{2}d_{37}^2 + \frac{1}{2}d_{67}^2 - \frac{1}{4}d_{36}^2 = \frac{1}{2}\times 93.25 + \frac{1}{2}\times 245.25 - \frac{1}{4}\times 37 = 160$$

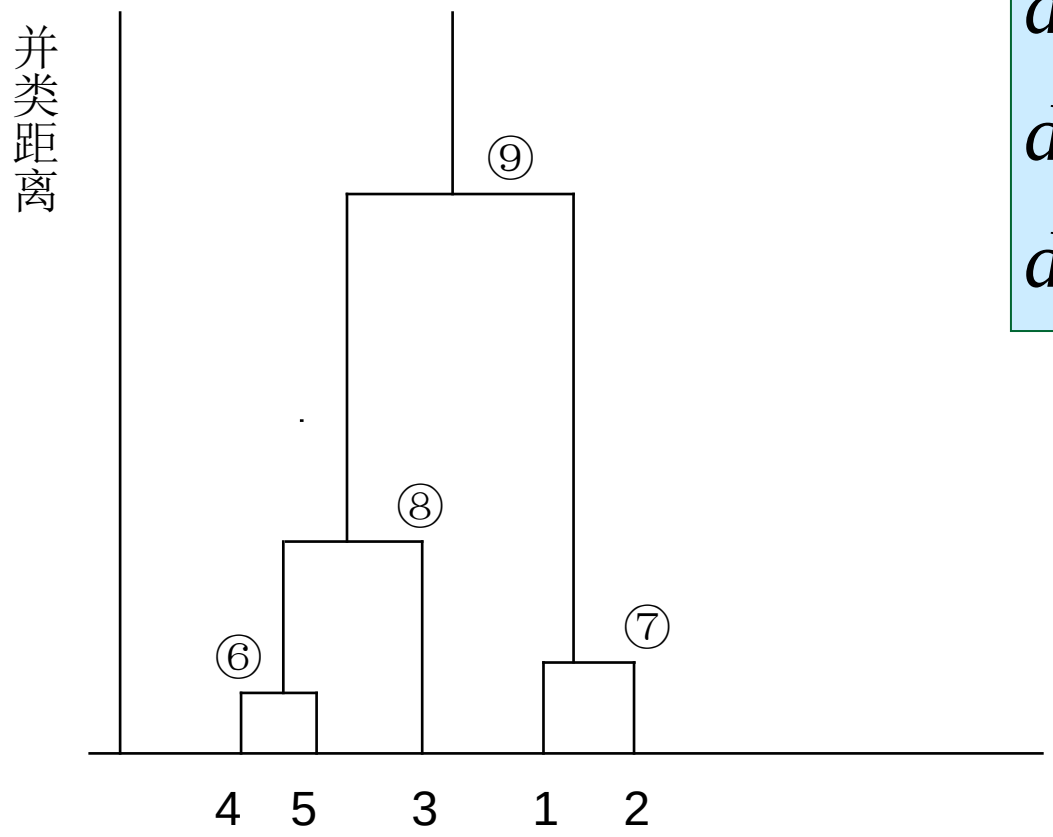
6、按聚类的过程画聚类谱系图

$$d_{4,5}=2$$

$$d_{1,2}=3.6$$

$$d_{3,6}=6.08$$

$$d_{7,8}=12.65$$



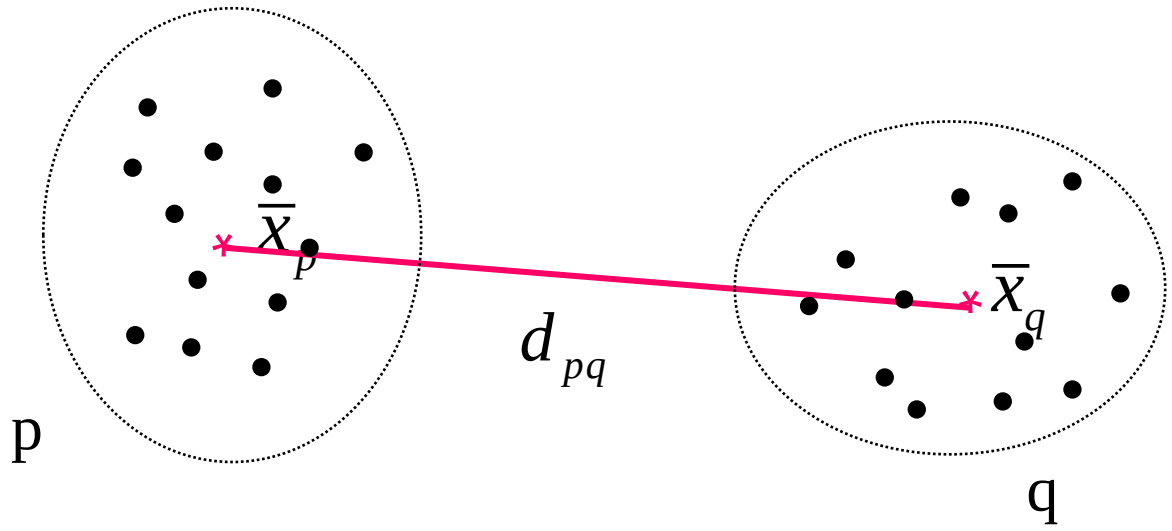
7、决定类的个数与类。

观察此图，我们可以把5个样品分为3类，$\{1,2\}$、$\{3\}$、$\{4,5\}$。

四、重心法（Centroid)

类与类之间的距离就考虑用重心之间的距离表示。设p与q的重心分别是 $\bar{x}_p$ 和 $\bar{x}_q$ ，则类p和q的距离为

$$d_{pq} = d_{\bar{x}_p \bar{x}_q}$$



设聚类到某一步，类p与 q分别有样品 n_p 、n_q 个，

将p和q合并为k，则k类的样品个数为 $n_k = n_p + n_q$

它的重心是 $\overline{x_k} = \frac{1}{n_k}\left(n_p \overline{x_p} + n_q \overline{x_q}\right)$

某一类 r 的重心是 $\overline{x_r}$ ，它与新类k的距离是

$$d_{kr}^2 = \left(\overline{x_k} - \overline{x_r}\right)'\left(\overline{x_k} - \overline{x_r}\right)$$

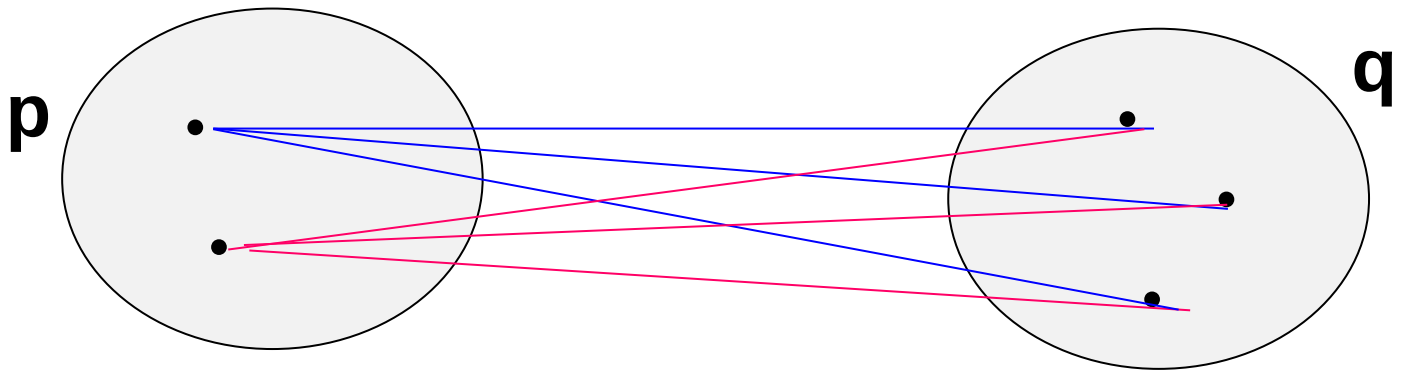
经推导可以得到如下递推公式：

$$d_{kr}^2 = \frac{n_p}{n_k} d_{pr}^2 + \frac{n_q}{n_k} d_{qr}^2 - \frac{n_p}{n_k}\frac{n_q}{n_k} d_{pq}^2$$

五、类平均法（Average)

定义两类之间的距离平方为这两类元素两两之间距离平方的平均

$$d_{pq}^2 = \frac{1}{n_p n_q} \sum_{i \in p} \sum_{j \in q} d_{ij}^2$$



$$\frac{1}{6}(d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2)$$

设聚类到某一步，类p与 q分别有样品 n_p 、n_q 个，

将p和q合并为k，则k类的样品个数为 $n_k = n_p + n_q$

k类与任一类 r 的距离为

$$d_{kr}^2 = \frac{1}{n_k n_r} \sum_{i \in r} \sum_{j \in k} d_{ij}^2$$

$$= \frac{1}{n_k n_r} \left(\sum_{i \in r} \sum_{j \in p} d_{ij}^2 + \sum_{i \in r} \sum_{j \in q} d_{ij}^2 \right)$$

$$= \frac{1}{n_k n_r} (n_p n_r d_{pr}^2 + n_q n_r d_{qr}^2) = \frac{n_p}{n_k} d_{pr}^2 + \frac{n_q}{n_k} d_{qr}^2$$

六、离差平方和法（Ward法）

设变量X的n个样品观察值为：

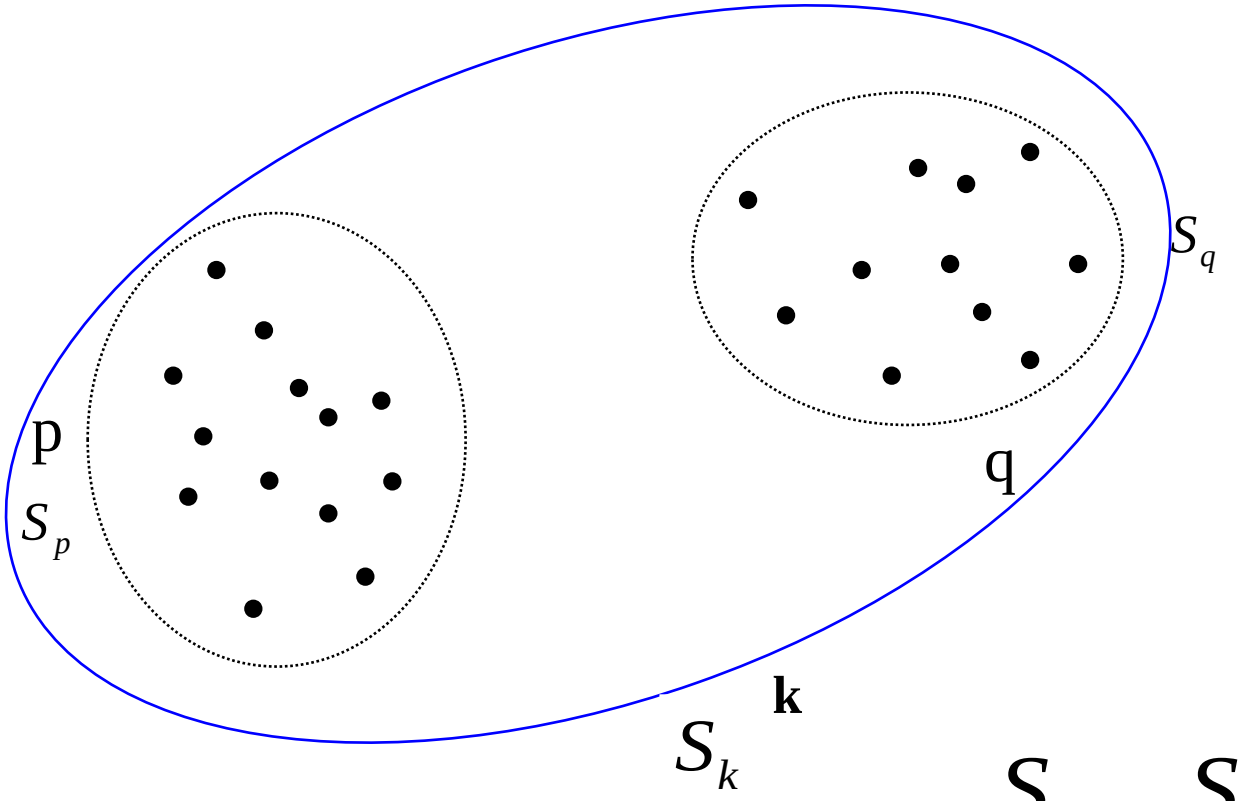
$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

n个样品的离差平方和为：

$$\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2+\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2+\cdots+\sum_{i=1}^{n}(x_{ip}-\bar{x}_p)^2$$

$$=\sum_{i=1}^{n}(X_i-\overline{X})'(X_i-\overline{X})$$

反映样品之间的差异程度



S_k $S_p + S_q$

设类p和q分别含有n_p、n_q个样品，其离差平方和分别记为

S_p 和 S_q

- 大群的离差平方和 > 原来两个群的离差平方和之和

$$S_k > S_p + S_q$$

$$S_k = S_p + S_q + \frac{n_p n_q}{n_p + n_q} d^2_{\overline{x}_p \overline{x}_q}$$

把增加的量记为 Δs_{pq}

定义类p和q之间的距离为：

$$d^2_{pq} = \Delta S_{pq} = \frac{n_p n_q}{n_p + n_q} d^2_{\overline{x}_p \overline{x}_q}$$

可以推得新类 k与任一类 r 的距离：

$$d_{kr}^2 = \frac{n_p + n_r}{n_k + n_r} d_{pr}^2 + \frac{n_q + n_r}{n_k + n_r} d_{qr}^2 - \frac{n_r}{n_k + n_r} d_{pq}^2$$

最短距离法 $$d_{pq} = \min_{i \in p, j \in q} \{d_{ij}\}$$

最长距离法 $$d_{pq} = \max_{i \in p, j \in q} \{d_{ij}\}$$

中间距离法 $$d_{kr}^2 = \frac{1}{2} d_{pr}^2 + \frac{1}{2} d_{qr}^2 - \frac{1}{4} d_{pq}^2$$

重心法 $$d_{kr}^2 = \frac{n_p}{n_k} d_{pr}^2 + \frac{n_q}{n_k} d_{qr}^2 - \frac{n_p}{n_k} \frac{n_q}{n_k} d_{pq}^2$$

类平均法 $$d_{kr}^2 = \frac{n_p}{n_k} d_{pr}^2 + \frac{n_q}{n_k} d_{qr}^2$$

离差平方和法 $$d_{kr}^2 = \frac{n_p + n_r}{n_k + n_r} d_{pr}^2 + \frac{n_q + n_r}{n_k + n_r} d_{qr}^2 - \frac{n_r}{n_k + n_r} d_{pq}^2$$

系统聚类的基本性质

1、单调性

设D_k是系统聚类法中第K次并类时的距离，如果$D_1 < D_2 < \ldots$，则称并类距离具有单调性。

可以证明除了中间距离法和重心法之外，其他的系统聚类法均满足单调性

2、空间的浓缩或扩张

两个同阶矩阵D（A）和D（B），如果D（A）的每一个元素不小于D（B）的相应元素，则记为

D（A）≧D（B）。

若有两种系统聚类法A和B，在第K步的距离矩阵记为 D（AK）和D（BK），

若有D（AK） ≧ D（BK）对所有K，则称A比B使空间扩张或B比A使空间浓缩。

最短距离法

	②	③	⑥
①	**3.6**	10.2	16.12
②		9.43	14.87
③			6

	⑥	⑦
③	**6**	9 .43
⑥		14.87

最长距离法

	②	③	⑥
①	**3.6**	10.2	16.49
②		9.43	15.65
③			6.32

	⑥	⑦
③	**6.32**	10 .2
⑥		16.49

确定类的个数

1. **给定阈值**——通过观测聚类图，给出一个合适的阈值T。

2. 统计量 $R^2 = 1 - \frac{P_G}{T}$

其中T是数据的总离差平方和，P_G是组内离差平方和。

R^2比较大，说明分G个类时类内的离差平方和比较小，也就是说分G类是合适的。

分类越多，每个类的类内的离差平方和就越小，R^2也就越大；所以我们只能取合适的G，使得 R^2足够大，而G本身很小，随着G的增加，R^2的增幅不大。比如，假定分4类时，R^2 =0. 8；下一次合并分三类时，R^2下降了许多，R^2 =0. 32，则分4 类是合适的。

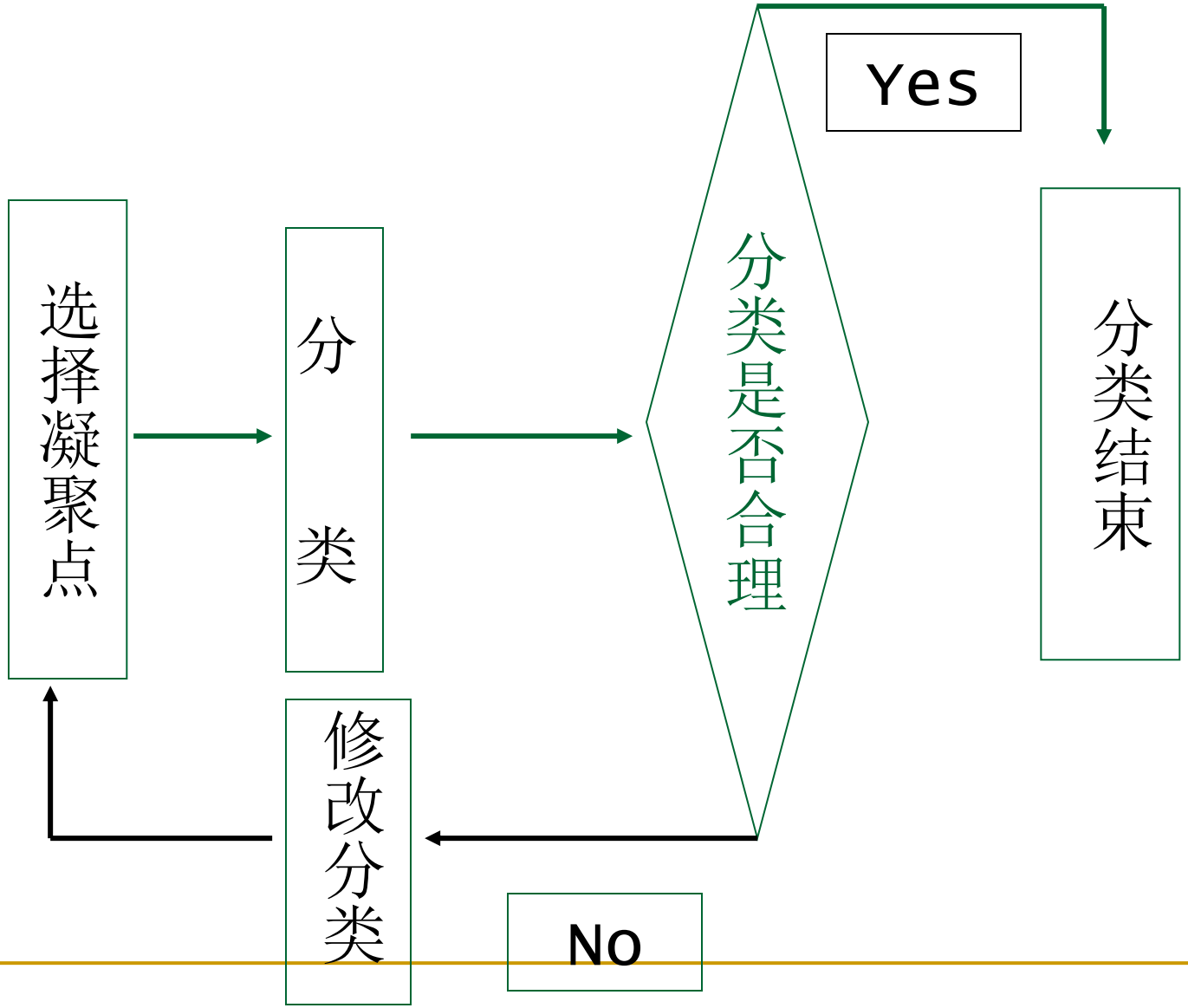
动态聚类法

系统聚类法是一种比较成功的聚类方法。然而当样本点数量十分庞大时，则是一件非常繁重的工作，且聚类的计算速度也比较慢。比如在市场抽样调查中，有4万人就其对衣着的偏好作了回答，希望能迅速将他们分为几类。这时，采用系统聚类法就很困难，而动态聚类法就会显得方便，适用。

动态聚类使用于大型数据。

动态聚类法

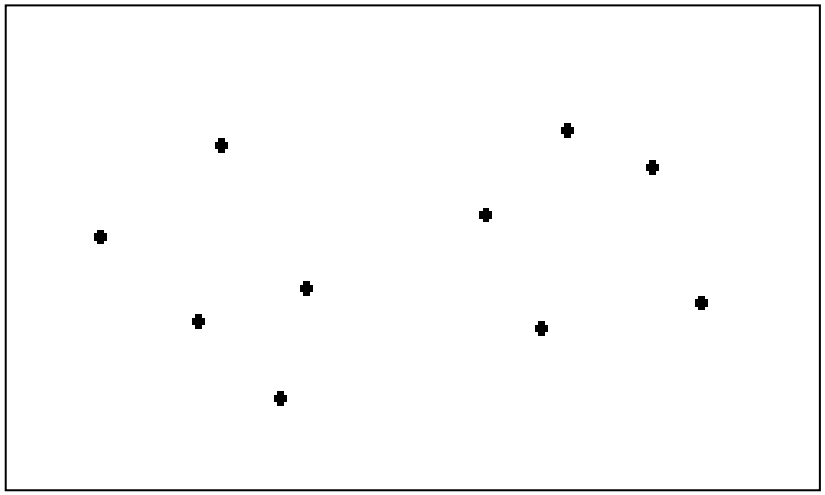
- 基本思想：选取若干个样品作为凝聚点，计算每个样品和凝聚点的距离，进行初始分类，然后根据初始分类计算其重心，再进行第二次分类，一直到所有样品不再调整为止。


选择凝聚点
分 类
分类是否合理
Yes
分类结束
修改分类
No

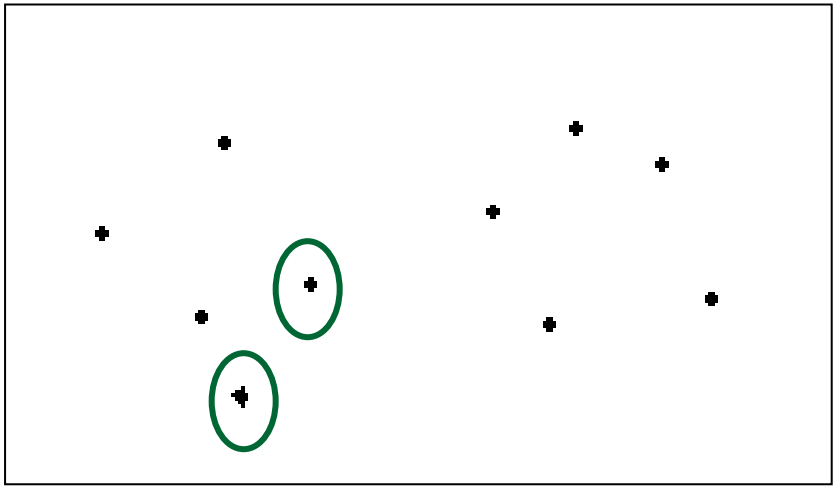
例如我们要把图中的点分成两类。快速聚类的步骤：

1、随机选取两个点 $x_1^{(1)}$ 和 $x_2^{(1)}$ 作为凝聚点。

（a）空间的群点



(b) 任取两个凝聚点



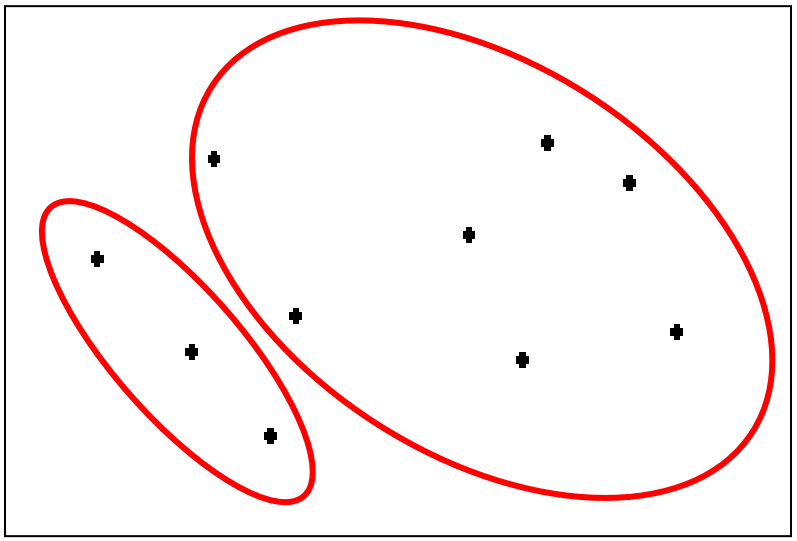
2、对于任何点 x_k ，分别计算 $d(x_k, x_1^{(1)})$ 和 $d(x_k, x_2^{(1)})$

3、若 $d(x_k, x_1^{(1)}) < d(x_k, x_2^{(1)})$ ，则将 x_k 划为第一类，否则划给第二类。于是得图（c）的两个类。

2、对于任何点 x_k，分别计算$d(x_k, x_1^{(1)})$和 $d(x_k, x_2^{(1)})$

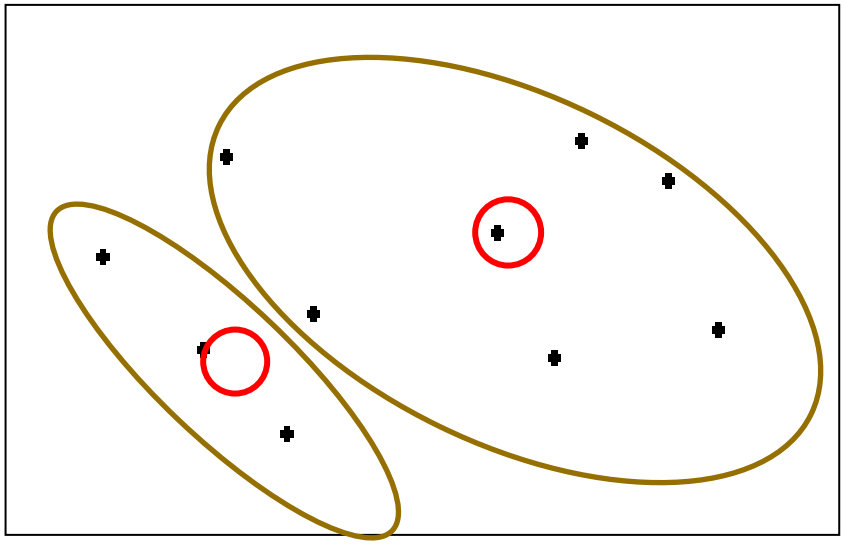
3、若 $d(x_k, x_1^{(1)}) < d(x_k, x_2^{(1)})$ ，则将 x_k 划为第一类，否则划给第二类。于是得图（c）的两个类。

(c) 第一次分类

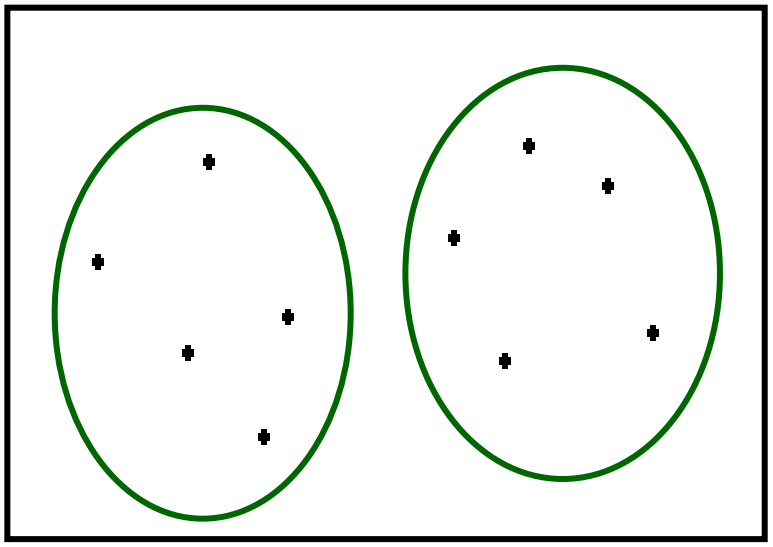


4、分别计算两个类的重心，则得 $x_1^{(2)}$ 和 $x_2^{(2)}$，以其为新的凝聚点，对空间中的点进行重新分类，得到新分类。

(d) 求各类中心



(e) 第二次分类



动态聚类法

- 优点：计算量小，方法简便，可以根据经验，先作主观分类。
- 缺点：结果受选择凝聚点好坏的影响，分类结果不稳定。

选择凝聚点和确定初始分类

凝聚点就是一批有代表性的点，是欲形成类的中心。凝聚点的 选择直接决定初始分类，对分类结果也有很大的影响，由于凝聚点 的不同选择，其最终分类结果也将出现不同。故选择时要慎重．通 常选择凝聚点的方法有：

(1) **人为选择**，当人们对所欲分类的问题有一定了解时，根据经验，预先确定分类个数和初始分类，并从每一类中选择一个有代表性的样品作为凝聚点。

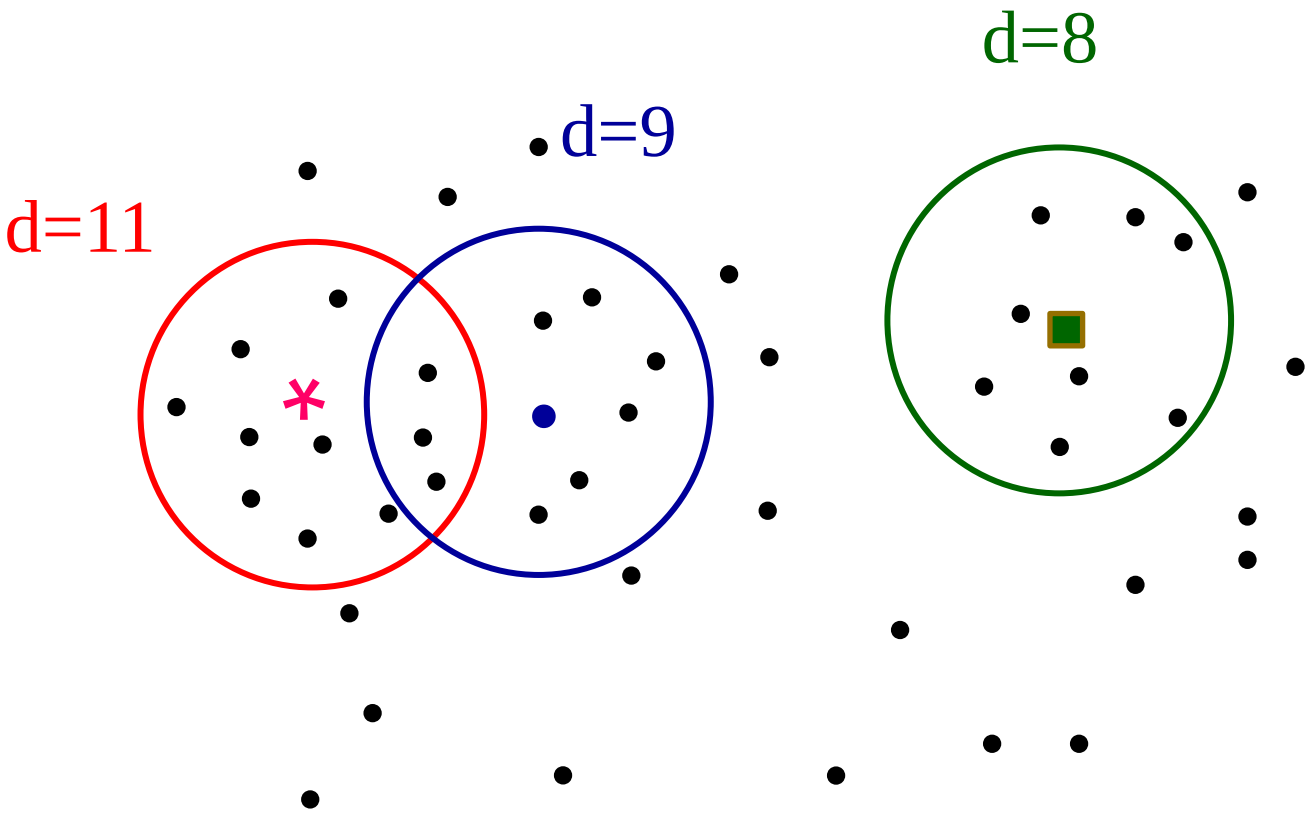
(2) **重心法**

将数据人为地分为A类，计算每一类的重心，将重心作为凝聚点。

(3) 密度法

- 以某个正数d为半径，以每个样品为球心，落在这个球内的样品数(不包括作为球心的样品)称为这个样品的密度。
- 计算所有样品点的密度，选择密度最大的样品为第一凝聚点。
- 选出密度次大的样品点，若它与第一个凝聚点的距离大于2d，则将其作为第二个凝聚点；否则舍去这点。
- 按密度由大到小依次考察，直至全部样品考察完毕为止。此方法中，d要给得合适，太大了使凝聚点个数太少，太小了使凝聚点个数太多。

(3) 密度法



(4) 人为地选择一正数d，首先以所有样品的均值作为第一凝聚点。然后依次考察每个样品，若某样品与已选定的凝聚点的距 离均大于d，该样品作为新的凝聚点，否则考察下一个样品。

动态聚类法的基本步骤：

第一，选择凝聚点；

第二，初始分类；

对于取定的凝聚点，视每个凝聚点为一类，将每个样品根据定义的距离向最近的凝聚点归类。

第三，修改分类

得到初始分类，计算各类的重心，以这些重心作为新的凝聚点，重新进行分类，重复步骤2，3，直到分类的结果与上一步的分类结果相同，表明分类已经合理为止。

例1：某商店5位售货员的销售量和教育程度如下表：

售货员	1	2	3	4	5
销售量（千件）	1	1	6	8	8
教育程度	1	2	3	2	0

对这5位售货员分类。

1.选择凝聚点

计算各样品点两两之间的距离，得到如下的距离矩阵

	②	③	④	⑤
①	1	$\sqrt{29}$	$\sqrt{50}$	$\sqrt{50}$
②		$\sqrt{26}$	$\sqrt{49}$	$\sqrt{53}$
③			$\sqrt{5}$	$\sqrt{13}$
④				$\sqrt{4}$

$d_{25}=\sqrt{53}$为最大。可选择2和5作为凝聚点。

2.初始分类

对于取定的凝聚点，视每个凝聚点为一类，将每个样品根据定义的距离，向最近的凝聚点归类。

	②G_1	⑤G_2
1	1	$\sqrt{50}$
3	$\sqrt{26}$	$\sqrt{13}$
4	$\sqrt{49}$	$\sqrt{4}$

得到初始分类为：G_1：$\{1,2\}$

G_2：$\{3,4,5\}$

3.修改分类

计算G_1和G_2的重心：G_1的重心（1,1.5），
G_2的重心（7.33,1.67）

以这两个重心点作为凝聚点，再按最小距离原则重新聚类

	G_1	G_2
1	$\sqrt{0.25}$	$\sqrt{40.52}$
2	$\sqrt{0.25}$	$\sqrt{40.18}$
3	$\sqrt{27.25}$	$\sqrt{3.54}$
4	$\sqrt{49.15}$	$\sqrt{0.56}$
5	$\sqrt{51.25}$	$\sqrt{3.24}$

得到分类结果：　$G_1:\{1,2\}$

$G_2:\{3,4,5\}$

修改前后所分的类相同，故可停止修改。

5个售货员可分为两类

$\{1,2\}$ 和 $\{3,4,5\}$ 。

动态聚类的特征：

快速、动态

结果不稳定，受凝聚点选择的影响

浙江工商大學
ZHEJIANG GONGSHANG UNIVERSITY

Thank You