第八章 聚类分析

聚类分析的基本思想是：

将样本(或变量)按相似程度的大小聚在一起，逐一归类.

相似程度的度量

1) 样本之间定义**距离**;

2) 变量之间定义**相似系数**;

距离或相似系数代表样本或变量之间的相似程度.

8.1 个体聚类和变量聚类

假设有 n 个个体, 每个个体有 p 个观测变量(指标).

个体聚类就是根据个体与个体间观测值的差别和相似之处, 将这 n 个个体分成若干个类别.

变量聚类就是根据变量与个变量之间的差别和相似之处, 将这 p 个变量分成若干个类别.

例子:

个体聚类：欧洲各国语言的语系分类.
变量聚类：人体部位尺寸的分类.

8.2 距离、相似系数和匹配系数

8.2.1 个体聚类

个体之间的接近程度通常用Minkowski距离来度量.

设 $x=(x_1,\cdots,x_p)'$, $y=(y_1,\cdots,y_p)'$, 则 x 与 y 的Minkowski距离为

$$d(x,y)=\left(\sum_{i=1}^{p}|x_i-y_i|^m\right)^{1/m},\quad \text{其中 } m>0.$$

有

$$d(x,y)=\begin{cases}\sum_{i=1}^{p}|x_i-y_i|, & m=1, \quad \text{绝对距离;}\\ \left(\sum_{i=1}^{p}(x_i-y_i)^2\right)^{1/2}, & m=2, \quad \text{欧氏距离;}\\ \max_{1\le i\le p}|x_i-y_i|, & m=\infty, \quad \text{Chebyshev距离.}\end{cases}$$

注1. 具体问题可用合适的距离, 特别是离散型数据.

8.2.2 变量聚类

变量之间的接近程度通常用**Pearson矩相关系数**来度量.

设变量 w 和 u 在 n 个个体的观测值分别为 $(w_1, \cdots, w_n)$ 和 $(u_1, \cdots, u_n)$, 则变量 w 和 u 之间的距离(相似系数)为

$$d(w,u) = r_{w,u} = \frac{\sum_{i=1}^{n}(w_i - \bar{w})(u_i - \bar{u})}{\sqrt{\sum_{i=1}^{n}(w_i - \bar{w})^2}\sqrt{\sum_{i=1}^{n}(u_i - \bar{u})^2}},$$

其中, $\bar{w} = n^{-1}\sum_{i=1}^{n} w_i$, $\bar{u} = n^{-1}\sum_{i=1}^{n} u_i$.

注2. $r_{w,u} = 1$ 时, $w \neq u$.

变量之间的接近程度的另一个常用度量是夹角余弦, 定义如下:

设两个变量分别为 $x = (x_1, \cdots, x_n)'$ 和 $y = (y_1, \cdots, y_n)'$,

则变量 x 和 y 之间的夹角余弦为

$$d(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}.$$

此度量常用于处理图形的相似性.

8.2.3 匹配系数

假设有两个观测向量 $x = (x_1, \cdots, x_k)'$, $y = (y_1, \cdots, y_k)'$,
则 x 与 y 的匹配系数为

$$d(x, y) = \frac{\sum_{i=1}^{k} I(x_i = y_i)}{k}.$$

匹配系数常用于属性数据的相似性度量.

8.3 聚类方法

聚类方法大致分为两种类型: **谱系聚类**和**迭代分块聚类**.

谱系聚类又可分为以下两种类型,

凝聚: 类别由多到少. 一开始把每个个体(变量)自成一类, 然后将最接近的个体 (最相似的变量) 凝聚为一小类, 再将最接近(相似)的类凝聚在一起, 依次类推直到所有个体(变量)凝聚为一个大类时止.

分解: 类别由少到多. 一开始把所有个体(变量)看成一大类, 然后将它分解成两个子类, 使这两个子类最为疏远, 再将子类分为两个最为疏远的子类, 依此类推直到每个个体(变量)都自成一类时止.

关键: 类与类之间的度量.

8.3.1 谱系聚类法

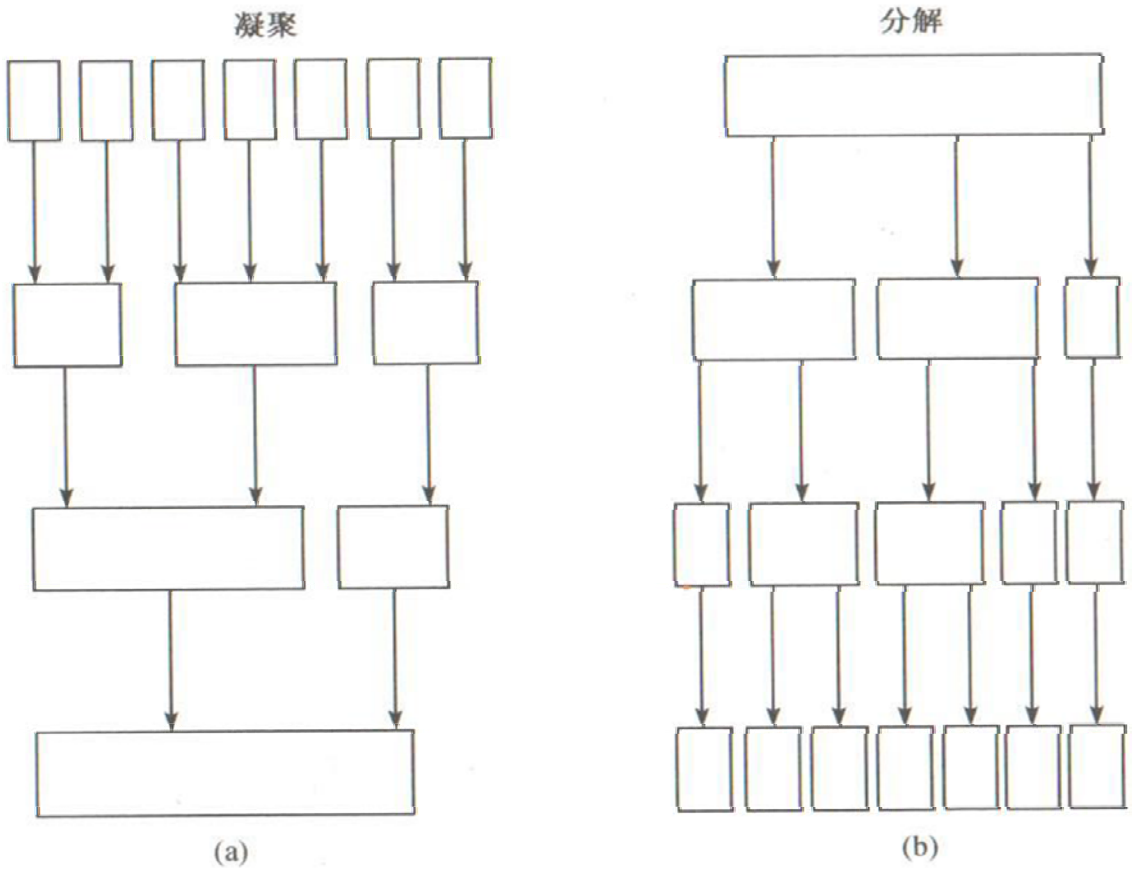


图 9.2.1　谱系图

8.3.2 类与类之间的距离

设有两个由个体或变量组成的类

$$\pi_1 = \{x_i : i \in G_1\}, \quad \pi_2 = \{x_j : j \in G_2\}.$$

类 π_1 与 π_2 之间的距离 $d(\pi_1, \pi_2)$ 常用以下定义方法:

(1) 最小距离法: $d(\pi_1, \pi_2) = \min\limits_{\{i \in G_1,\, j \in G_2\}} d(x_i, x_j).$

(2) 最大距离法: $d(\pi_1, \pi_2) = \max\limits_{\{i \in G_1,\, j \in G_2\}} d(x_i, x_j).$

(3) 类平均法:

$$d(\pi_1, \pi_2) = \frac{\sum\limits_{i \in G_1} \sum\limits_{j \in G_2} d(x_i, x_j)}{n_1 n_2},$$

其中, $n_1 = \#\{x_i : i \in G_1\}, \quad n_2 = \#\{x_j : j \in G_2\}.$

(4) 重心法: $d(\pi_1, \pi_2) = d(\bar{x}_1, \bar{x}_2)$,

$$\text{其中}, \bar{x}_1 = \frac{\sum_{i \in G_1} x_i}{n_1}, \ \bar{x}_2 = \frac{\sum_{j \in G_2} x_j}{n_2}.$$

(5) 离差平方和法: $d(\pi_1, \pi_2) = \dfrac{n_1 n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2)'(\bar{x}_1 - \bar{x}_2)$.

注3: 离差平方和即组间平方和, 事实上有

$$\begin{aligned} d(\pi_1, \pi_2) &= \frac{n_1 n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2)'(\bar{x}_1 - \bar{x}_2) \\ &= n_1(\bar{x}_1 - \bar{x})'(\bar{x}_1 - \bar{x}) + n_2(\bar{x}_2 - \bar{x})'(\bar{x}_2 - \bar{x}), \\ \text{其中 } \bar{x} &= \frac{1}{n_1 + n_2}\left(\sum_{i \in G_1} x_i + \sum_{j \in G_2} x_j\right). \end{aligned}$$

注4: 各种距离的递推公式

假设最初有类G_p, G_q, G_k, 它们之间的距离分别为D_{pq}, D_{pk}和D_{qk}.
通过凝聚, 类G_p和类G_q能够合并为一个新的类G_r.
记$n_p = \#G_p, n_q = \#G_q, n_k = \#G_k, n_r = \#G_r = n_p + n_q$.
则类G_k与新类G_r的距离为

$$D_{kr}^2 = \alpha_p D_{kp}^2 + \alpha_q D_{kq}^2 + \beta D_{pq}^2 + \gamma |D_{kp}^2 - D_{kq}^2|,$$

其中,

方　法	α_p	α_q	β	γ
最小距离法	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
最大距离法	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
重心法	$\frac{n_p}{n_r}$	$\frac{n_q}{n_r}$	$-\frac{n_p n_q}{n_r^2}$	0
类平均法	$\frac{n_p}{n_r}$	$\frac{n_q}{n_r}$	0	0
离差平方和法	$\frac{n_k+n_p}{n_k+n_r}$	$\frac{n_k+n_q}{n_k+n_r}$	$-\frac{n_k}{n_k+n_r}$	0

谱系聚类法基于类与类的距离来对类进行凝聚或分解.

谱系聚类法的缺陷:

1) 凝聚: 一旦两个个体(变量)被分入同一个类中, 则它们以后不再会分入两个不同的类中.

2) 分解: 一旦两个个体(变量)被分入两个不同的类中, 则它们以后不再会并入同一个类中.

3) 个体被错误分类无法纠正.

4) 计算量大.

例1: 有六个样本, 每个样本只测一个指标.

样本	G_1	G_2	G_3	G_4	G_5	G_6
观测	1	2	5	7	9	10

采用**凝聚法**和**最小距离法**给它们聚类.

(1) 样本间采用**绝对距离**, 结果如下;

表1

类	G_1	G_2	G_3	G_4	G_5
G_2	1				
G_3	4	3			
G_4	6	5	2		
G_5	8	7	4	2	
G_6	9	8	5	3	1

(2) 表1中最小值是1, 故将 G_1 和 G_2 合并成 G_7, 将 G_5 和 G_6 合并成 G_8.

(3) 计算 G_7 和 G_8 之间以及与其它类的距离, 结果如下;

表2

类	G_7	G_3	G_4
G_3	3		
G_4	5	2	
G_8	7	4	2

(4) 表2中最小值是2, 故将 G_3, G_4 和 G_8 合并成 G_9.

(5) 计算 G_7 和 G_9 的距离, 结果如下;

表3

类	G_7
G_9	3

(6) 将 G_7 和 G_9 合并成 G_{10}, 凝聚结束.

再采用**分解法**和**最大距离法**给它们聚类.

记 $G_0 = \{G_1, \cdots, G_6\}$, 即, G_0 是全集.

表4

	G_1	G_2	G_3	G_4	G_5	G_6
$G_0 \backslash G_i$	1	1	2	2	1	1

选出 G_3 作为新类的第一候选. 再计算

表5

	$\{G_3, G_1\}$	$\{G_3, G_2\}$	$\{G_3, G_4\}$	$\{G_3, G_5\}$	$\{G_3, G_6\}$
$G_0 \backslash \{G_3, G_i\}$	1	1	2	1	1

$\{G_3, G_4\}$ 是新类的候选. 再计算

表6

	$\{G_3, G_4, G_1\}$	$\{G_3, G_4, G_2\}$	$\{G_3, G_4, G_5\}$	$\{G_3, G_4, G_6\}$
$G_0 \backslash \{G_3, G_4, G_i\}$	1	1	1	1

由于此时的最大距离都是1, 第一次分解结束, 新分出的两类分别是

$$\{G_1, G_2, G_5, G_6\} \text{ 和 } \{G_3, G_4\}.$$

再对新的两类作类似分解. 记 $G_{01} = \{G_1, G_2, G_5, G_6\}$.

表7

	G_1	G_2	G_5	G_6
$G_{01} \backslash G_i$	1	1	1	1

选出 G_1 作为新类的候选. 计算

表8

	$\{G_1, G_2\}$	$\{G_1, G_5\}$	$\{G_1, G_6\}$
$G_{01}\backslash\{G_1, G_i\}$	7	1	1

$\{G_1, G_2\}$ 是新类的候选. 再计算

表9

	$\{G_1, G_2, G_5\}$	$\{G_1, G_2, G_6\}$
$G_{01}\backslash\{G_1, G_2, G_i\}$	1	1

第二次分解结束. 新分出四类, 分别是

$$\{G_1, G_2\},\ \{G_5, G_6\},\ \{G_3\},\ \{G_4\}.$$

最后完全分解为每个个体为一类, 分解结束, 即

$$\{G_1\},\ \{G_2\},\ \{G_3\},\ \{G_4\},\ \{G_5\},\ \{G_6\}.$$

8.3.3 迭代分块聚类

迭代分块聚类法是一种动态聚类法, 其搜索迭代步骤大致如下:

(1) 初始分类 将 n 个个体(变量)初始分为 k 类, 其中, 类的个数 k 可以事先给定, 也可以在聚类过程中逐步确定.

(2) 修改分类 对每个个体(变量)逐一进行搜索, 若将某个个体(变量)分入另一个类后对分类有所改进, 则将其移入改进最多的那个类, 否则其不移动, 仍在原来的类中.

(3) 重复迭代 在对每个个体(变量)逐一都进行搜索之后, 重复第**(2)**步, 直到任何一个个体(变量)都不需要移动为止, 从而得到最终分类.

迭代分块聚类 – K均值法

1) 初始分类 将 n 个个体初始分成 k 类, k 事先给定.

2) 修改分类 计算初始 k 类的重心. 然后对每个个体逐一计算它到初始 k 类的距离(通常用该个体到类的重心的欧氏距离). 若该个体到其原来的类的距离最近, 则它保持类不变, 否则它移入离其距离最近的类. 重新计算由此变动的两个类的重心.

由于初始分类数 k 事先给定, 且迭代过程中不断计算类的重心, 故称该聚类方法为k均值法(k-means).

3) 重复迭代 在对所有个体都逐一进行验证, 是否需要修改分类之后, 重复步骤**2)**, 直到没有个体需要移动为止, 从而得到最终分类.

动态K均值法

事先给定3个数: 类别数 k, 阀值 c_1 和 $c_2, c_2 > c_1 > 0$.

1) 选取聚点

取前 k 个个体作为初始聚点, 计算这 k 个聚点两两之间的距离, 若最小的距离比 c_1 小, 则将最小距离的这两个聚点合并在一起, 并用它们的重心作为新的聚点. 重复上述过程, 直到所有的聚点两两之间的距离都不比 c_1 小时为止.

因此, 此时聚点的个数可能小于 k.

2) 初始分类

对余下的 $n-k$ 个个体逐一进行计算. 对输入的一个个体, 分别计算它到所有聚点的距离. 若该个体到所有聚点的距离都大于 c_2, 则它作为一个新的聚点, 这时所有聚点两两之间的距离都不比 c_1 小; 否则将它归入离它最近的那一类, 并重新计算接受该个体的那个类的重心以代替该类原来的聚点. 然后重复步骤**1)**, 再次验证所有聚点两两之间的距离是否都不比 c_1 小, 如果比 c_1 小就将其合并, 直到所有聚点两两之间的距离都不比 c_1 小时止.

该步完成后, 聚点的个数可能小于 k, 也可能大于 k.

3) 重复迭代

在对所有个体都逐一进行验证, 是否需要修改分类之后, 重复步骤**2)**, 直到没有个体需要移动为止, 从而得到最终分类.

这时, 最终个体的类别数不一定是 k.

注5. 选取最初的k个个体时, 可以随机地从所有个体中抽取k个个体作为初始聚点, 然后进行迭代聚类, 看最终分类是否一致, 以检验聚类的稳定性.

注6. 选取初始聚点的密度法. 给定两个数$d_2 > d_1 > 0$. 以每个个体为球心, 以d_1为半径划球, 落在这个球内的个体数(不包括球心)就称为这个点的密度. 首先选取最大密度的个体作为第一聚点. 再选取次大密度的个体, 如果它和第一聚点的距离小于d_2则取消该个体, 如果不小于d_2, 则把该个体作为第二聚点. 依次按个体的密度有大到小进行选择, 将和每个已选聚点的距离不小于d_2的个体作为一个新的聚点, 直至选完所有个体. 将最终选出的聚点作为初始聚点. 常取$d_2 = 2d_1$.

注7. 在初始分类步骤时, 剩余个体进入的次序也有可能影响最终分类.

8.3.4 数据变换

谱系聚类法和迭代分块聚类法的最终分类都与数据的量纲有关,
通常采用尺度变换或标准化变换等数据变换方法来消除量纲的影响.

变换公式

$$\begin{cases} \text{尺度变换:} & x \mapsto \frac{x}{c}; \\ \text{标准化变换:} & x \mapsto \frac{x-\mu}{c}. \end{cases}$$

变换参数表

变换名称	位置参数 μ	尺度参数 c
总和(SUM)	0	$\sum_{i=1}^{n} x_i$
欧氏总和	0	$\sqrt{\sum_{i=1}^{n} x_i^2}$
标准化变换	$\bar{x}$	$\sqrt{\sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{n-1}}$
绝对极大值	0	$\max\{\lvert x_1\rvert, \cdots, \lvert x_n\rvert\}$
极差	$x_{(1)}$	$x_{(n)} - x_{(1)}$
绝对中位数	$m = median\{x_1, \cdots, x_n\}$	$median\{\lvert x_1 - m\rvert, \cdots, \lvert x_n - m\rvert\}$

8.4 有序样本的聚类

前述聚类方法处理的样本是相互独立的, 在聚类时也是平等的.

但在一些实际问题中, 样本是有序的, 在对其进行聚类时, 样本的顺序是不能被打乱的.

例如:

1) 要通过岩心对地层进行分类, 而岩心所在的位置 (即样本次序)在分类时是必须保持有序的.
2) 研究气候演变的历史时, 样本是按由古至今的年代排列的, 在进行演变史归类时样本的顺序不能乱.

问题的描述: 假设有 n 个有序样本 $x_1, \cdots, x_n$. 如果将它们分成 k 类, 每一类的样本数为 n_i, $1 \le i \le k$. 则每一类的样本必须是

$$\{x_{\sum_{j=1}^{i} n_{j-1}+1}, x_{\sum_{j=1}^{i} n_{j-1}+2}, \cdots, x_{\sum_{j=1}^{i+1} n_{j-1}}\}, \quad 1 \le i \le k,$$

其中 $n_0 = 0$.

8.4.1 有序样本可能的分类数目

定理1. 对于有序样本, n 个样本分成 k 类的所有可能分法数为

$$R(n,k) = C_{n-1}^{k-1}.$$

证明: $n-1$ 个“间隔”里放 $k-1$ 个“隔板”.

8.4.2 最优分割法（Fisher算法）

设 m 维的有序样本为 $x_1, x_2, \cdots, x_n$, 最优分割法的步骤如下:

第一步: 定义类的直径

设某一类 G_{ij} 的样本是 $\{x_i, x_{i+1}, \cdots, x_j\}$, $j > i$,

记它们的均值为

$$\bar{x}_{ij} = \frac{1}{j-i+1} \sum_{k=i}^{j} x_k.$$

记 G_{ij} 的直径为

$$D(i,j) = \sum_{k=i}^{j} (x_k - \bar{x}_{ij})'(x_k - \bar{x}_{ij}). \qquad (1)$$

当$m=1$时, 有时也用如下的直径

$$D(i,j)=\sum_{k=i}^{j}|x_k-\tilde{x}_{ij}|,$$

其中$\tilde{x}_{ij}$是$\{x_i,x_{i+1},\cdots,x_j\}$的中位数.

第二步: 定义目标函数

将n个有序样本分成k类. 设某一种分法为:

$$P(n,k):\ \{x_{i_1},x_{i_1+1},\cdots,x_{i_2-1}\},\{x_{i_2},x_{i_2+1},\cdots,x_{i_3-1}\},$$

$$\cdots,\{x_{i_k},x_{i_k+1},\cdots,x_{i_{k+1}}\},$$

其中分点是$1=i_1<i_2<\cdots<i_{k+1}=n$.

定义该分类的目标函数为

$$obj[P(n,k)]=\sum_{j=1}^{k}D(i_j,i_{j+1}-1).$$

第三步：精确最优解的算法（递推算法）

$$obj[P^*(n,2)] = \min_{2\le j\le n}\{D(1,j-1)+D(j,n)\}, \qquad (2)$$

其中 $P^*(n,2)$ 是 $k=2$ 时的最优分类.

$$obj[P^*(n,k)] = \min_{k\le j\le n}\{obj[P^*(j-1,k-1)]+D(j,n)\}. \qquad (3)$$

如果样本要分成k类,

首先找j_k, 使得(3)达到极小, 即

$$obj[P^*(n,k)] = obj[P^*(j_k, k-1)] + D(j_k, n), \qquad (4)$$

因而得$G_k = \{x_{j_k}, x_{j_k+1}, \cdots, x_n\}$.

然后找j_{k-1}, 使得

$$obj[P^*(j_k - 1, k-1)] = obj[P^*(j_{k-1} - 1, k-2)] + D(j_{k-1}, j_k - 1),$$

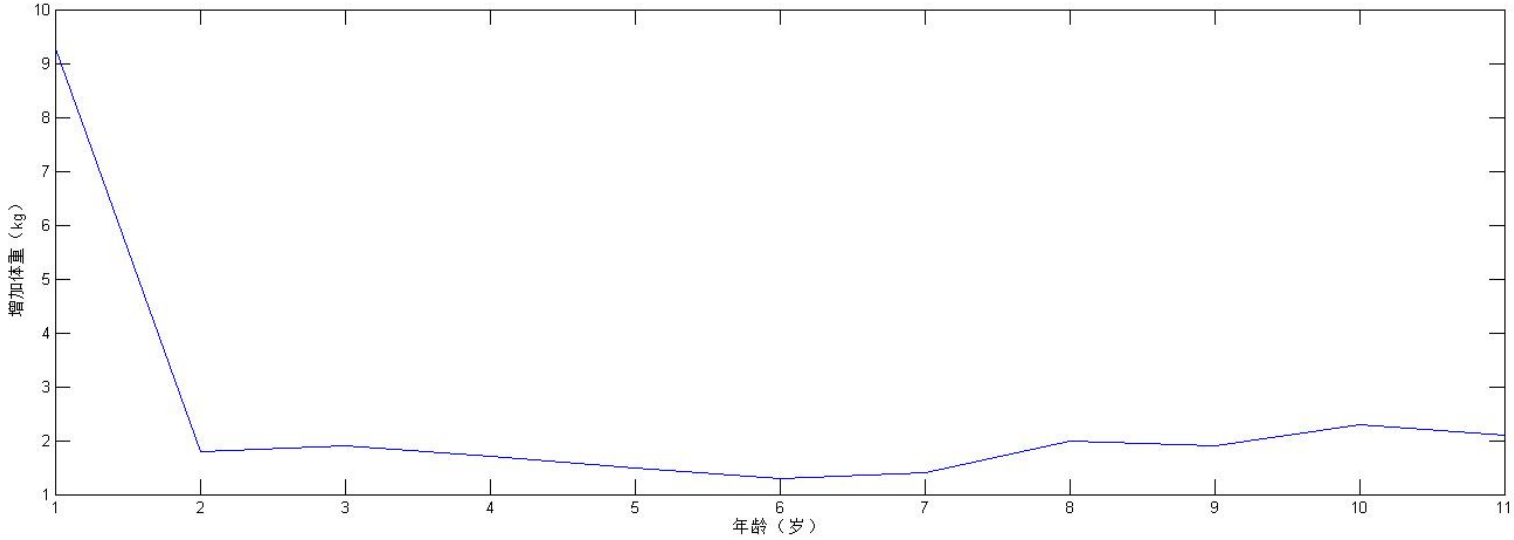
再得到$G_{k-1} = \{x_{j_{k-1}}, \cdots, x_{j_k-1}\}$;

以此类推, 得到类$G_1, G_2, \cdots, G_k$.

例3. 为了解儿童的生长发育规律, 统计了男孩从出生到十一岁每年平均增长的重量如下:

年龄	1	2	3	4	5	6	7	8	9	10	11
增加重量(kg)	9.3	1.8	1.9	1.7	1.5	1.3	1.4	2.0	1.9	2.3	2.1

这是一个有序样本的聚类问题. 用最优分割法来确定类.



(i) 计算直径 $\{D(i,j)\}$, 采用公式(1)的定义, 结果见表10.

表10. 直径 $D(i,j)$

j \ i	1	2	3	4	5	6	7	8	9	10
2	28.125									
3	37.007	0.005								
4	42.208	0.020	0.020							
5	45.992	0.088	0.080	0.020						
6	49.128	0.232	0.200	0.080	0.020					
7	51.100	0.280	0.232	0.088	0.020	0.005				
8	51.529	0.417	0.393	0.308	0.290	0.287	0.180			
9	51.980	0.469	0.454	0.393	0.388	0.370	0.207	0.005		
10	52.029	0.802	0.800	0.774	0.773	0.708	0.420	0.087	0.080	
11	52.182	0.909	0.909	0.895	0.889	0.793	0.452	0.088	0.080	0.020

(ii) 计算最小目标函数, $\{obj[P^*(i,j)],\ 3 \le i \le 11,\ 2 \le j \le 10\}$, 结果见表11.

表11. 最小目标函数 $obj[P^*(i,j)]$

i \ j	2	3	4	5	6	7	8	9	10
3	0.005 (2)								
4	0.020 (2)	0.005 (4)							
5	0.088 (2)	0.020 (5)	0.005 (5)						
6	0.232 (2)	0.040 (5)	0.020 (6)	0.005 (6)					
7	0.280 (2)	0.040 (5)	0.025 (6)	0.010 (6)	0.005 (8)				
8	0.417 (2)	0.280 (8)	0.040 (8)	0.025 (8)	0.010 (8)	0.005 (8)			
9	0.469 (2)	0.285 (8)	0.045 (8)	0.030 (8)	0.015 (8)	0.010 (8)	0.005 (8)		
10	0.802 (2)	0.367 (8)	0.127 (8)	0.045 (10)	0.030 (10)	0.015 (10)	0.010 (10)	0.005 (10)	
11	0.909 (2)	0.368 (8)	0.128 (8)	0.065 (10)	0.045 (11)	0.030 (11)	0.015 (11)	0.010 (11)	0.005 (11)

注: 括号里的数字表示使得 $obj[P^*(i,j)]$ 达到最小的指标, 即(3)式的解.

表中第一列括号里的数都是2, 表明对一切形如 $\{x_1, x_2, \cdots, x_j\}$, $2 \le j \le 11$ 的类, 如欲再分成两类, 都是以 $G_1 = \{x_1\}$ 和 $G_2 = \{x_2, \cdots, x_j\}$ 为最优分法.

(iii) 进行分类

假定要分的类别的总数 k 已定. 以分三类为例, 此时 $k = 3$.

$$obj[P^*(11,3)] = 0.368, \text{ 相应的指标 } j_3 = 8.$$

分类时, 首先分出第三类 $G_3 = \{x_8, x_9, x_{10}, x_{11}\}$.

再对前7个样本 $\{x_1, \cdots, x_7\}$ 进行分解. 由表11知

$$obj[P^*(7,2)] = 0.280, \text{ 相应的指标 } j_2 = 2.$$

再分类即得 $G_2 = \{x_2, x_3, x_4, x_5, x_6, x_7\}$, $G_1 = \{x_1\}$.

最后的分类结果为:

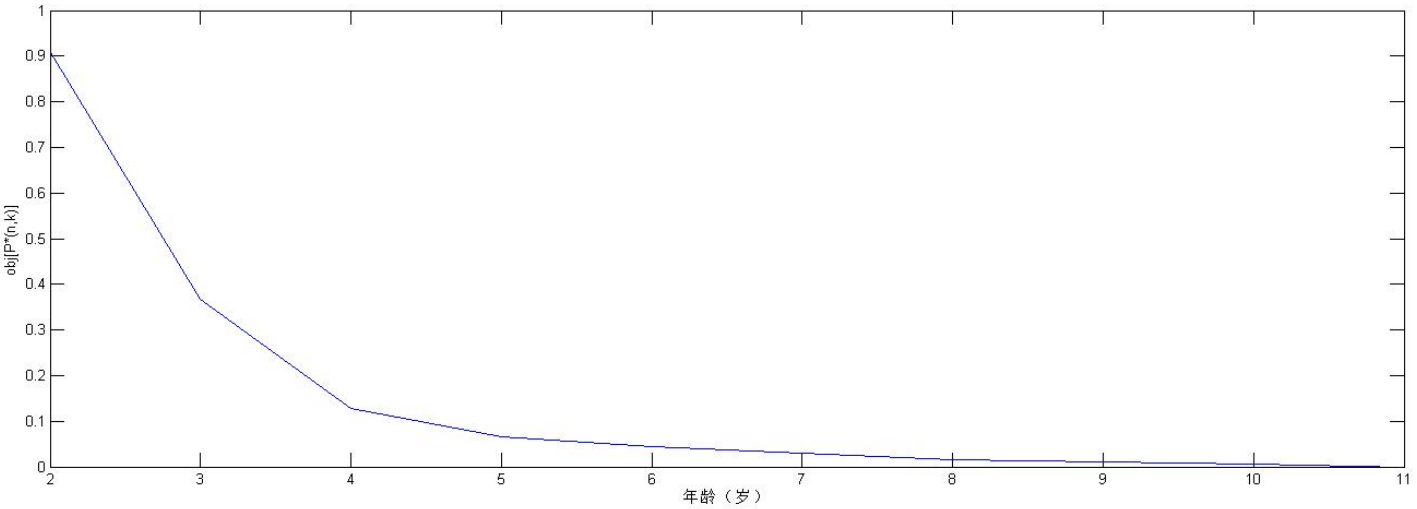
$G_1 = \{x_1\}$, $G_2 = \{x_2, x_3, x_4, x_5, x_6, x_7\}$, $G_3 = \{x_8, x_9, x_{10}, x_{11}\}$.

k 取其它值时, 分类方法类似.

(iv) 总类数的确定方法

1) 根据专业知识确定.

2) 作 $(k, obj[P^*(n,k)])$ 图, 由图确定合适的 k.



从曲线变化率看, 分成3类或4类最为合适.