聚类分析

TUTU

什么是聚类分析

♣ 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

♣ 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

距离和相似系数

♣ 明式距离：$d_{ij} = \left[\sum_{l=1}^{p} |x_{il} - x_{jl}|^k\right]^{\frac{1}{k}}$

- 绝对值距离：$k = 1$ 时，$d_{ij} = \sum_{l=1}^{p} |x_{il} - x_{jl}|$
- 欧氏距离：$k = 2$ 时，$d_{ij} = \left[\sum_{l=1}^{p} (x_{il} - x_{jl})^2\right]^{\frac{1}{2}}$
- 标准化欧氏距离：$d_{ij} = \left[\sum_{l=1}^{p} \frac{(x_{il} - x_{jl})^2}{s_{ll}}\right]^{\frac{1}{2}}$
- 切比雪夫距离：$k = \infty$ 时，$d_{ij} = \max_{1 \le l \le q} |x_{il} - x_{jl}|$
- 缺点：
  - 明氏距离的数值与指标的量纲有关
  - 没有考虑各个变量之间相关性的影响

距离和相似系数

♣ 马氏距离：$\boldsymbol{S}$ 是样品观测数据矩阵的协方差矩阵，

$\boldsymbol{x}_i = (x_{1i}, x_{2i}, \cdots, x_{ni})^{\mathrm{T}}$，

$$d_{ij} = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)}$$

马氏距离不受指标量纲及指标间相关性的影响

距离和相似系数

♣ 夹角余弦：变量 $\boldsymbol{x}_{(i)}, \boldsymbol{x}_{(j)}$ 的夹角余弦定义为

$$c_{ij} = \frac{\sum_{i=1}^{n} x_{ij} x_{ik}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2 \sum_{i=1}^{n} x_{ik}^2}}$$

♣ 相似系数：$\bar{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ki}, \bar{x}_j = \frac{1}{n}\sum_{k=1}^{n} x_{kj} (i, j = 1, 2, \cdots, p)$,

$$r_{ij} = \frac{\sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n} (x_{ki} - \bar{x}_i)^2 \sum_{k=1}^{n} (x_{kj} - \bar{x}_j)^2}} = \frac{\sigma_{xy}}{\sqrt{s_{xx} s_{yy}}}$$

系统聚类法

♣ 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。

♣ 优点：

简单，直观

系统聚类法

♣ 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

系统聚类法

♣ 最短距离法：类 G_p 和 G_q 之间的距离为

$$D_{pq} = \begin{cases} \min\limits_{i\in G_p, j\in G_q} d_{ij}, & p \neq q \\ 0, & p = q \end{cases}$$

若聚类出现新类 $G_k = G_p \cup G_q$，则 $D_{kr} = \min\limits_{i\in G_k, j\in G_r} d_{ij} =$

$$\min\left\{\min_{i\in G_p, j\in G_r} d_{ij}, \min_{i\in G_q, j\in G_r} d_{ij}\right\} = \min\{D_{pr}, D_{qr}\}$$

♣ 最长距离法：类 G_p 和 G_q 之间的距离为

$$D_{pq} = \begin{cases} \max\limits_{i\in G_p, j\in G_q} d_{ij}, & p \neq q \\ 0, & p = q \end{cases}$$

若聚类出现新类 $G_r = G_p \cup G_q$，则 $D_{kr} = \max\limits_{i\in G_k, j\in G_r} d_{ij} =$

$$\max\left\{\max_{i\in G_p, j\in G_r} d_{ij}, \max_{i\in G_q, j\in G_r} d_{ij}\right\} = \max\{D_{pr}, D_{qr}\}$$

系统聚类法

♣ 中间距离法：类 G_p 和 G_q 并为新类 G_r 后，递推公式为

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

♣ 重心法：类 G_p 和 G_q 之间的距离为 $D_{pq} = D_{\bar{x}_p\bar{x}_q}$，递推公式为

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

♣ 类平均法：类 G_p 和 G_q 之间的距离为 $D_{pq}^2 = \frac{1}{n_p n_q}\sum_i\sum_j D_{ij}^2$，递推公式为 $D_{kr}^2 = \frac{n_p}{n_k}D_{pr}^2 + \frac{n_q}{n_k}D_{qr}^2$

♣ 离差平方和法 (Ward 法)：类 G_p 和 G_q 之间的距离为 $D_{pq}^2 = \Delta S_{pq} = \frac{n_p n_q}{n_p + n_q}D_{\bar{x}_p\bar{x}_q}^2$，递推公式为

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

系统聚类法的具体性质

♣ 单调性：

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

♣ 空间的浓缩或扩张：

- $D(A) \geq D(B)$：A 的每个元素都不小于 B
- 若有$D(AK) \geq D(BK)$对所有 K，则称 A 比 B 使空间扩张或 B 比 A 使空间浓缩

系统聚类法的具体性质

♣ 确定类的个数：

- 观察
- $R^2 = 1 - \frac{P_G}{T}$，T 是数据的总离差平方和，P_G 是组内离差平方和，R^2 比较大合适
- 伪 F 统计量：$F = \frac{(T - P_G)/(G-1)}{P_G/(n-G)}$，取伪 F 统计量较大而类数较小的聚类水平
- 伪 t^2 统计量：$t^2 = \frac{B_{KL}}{(W_K + W_L)/(N_K + N_L - 2)}$，其中 W_L 和 W_K 分别是的类内离差平方和，W_M 是将 K 和 L 合并为第 M 类的离差平方和，$B_{KL} = W_M - W_K - W_L$ 为合并导致的类内离差平方和的增量，伪 t^2 统计量比较小合适

系统聚类法 SAS 代码

♣ SAS 代码:

```
/*method有single-最短距离法，complete-最长距离法，median-中
    间距离法，centroid-重心法，average-类平均法，ward-离差平
    方和法(Ward法)*/
proc cluster data=yourdata method=ward outtree=outtree
    standard;
id region;
run;

/*画树形图*/
proc tree data=outtree horizontal out=result n=5;
run;
```

快速聚类法

♣ 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

♣ 优点和缺点：

- 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

♣ 基本步骤：

1. 选择凝聚点
2. 初始分类
3. 修改分类

快速聚类法

♣ 选择凝聚点和确定初始分类：

- 人为选择
- 重心法
- 密度法：半径 d 内的样本数，从大到小选
- 最大最小法：先选择所有样品中最远的两个样品为凝聚点 $\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}$，选择第三个凝聚点 $\boldsymbol{x}_{i3}$，使$\boldsymbol{x}_{i3}$ 与前面两个凝聚点的距离最小者等于所有其余样品与 $\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}$ 的较小距离中的最大的.

  假设已选择了 l 个，第 $l+1$ 个凝聚点满足：$(t = 1, 2, \cdots, l)$

$$\min\{d(\boldsymbol{x}_{i_{(l+1)}}, \boldsymbol{x}_{i_t})\} = \max\{\min[d(\boldsymbol{x}_j, \boldsymbol{x}_{i_t}), j \neq i_t]\}$$

快速聚类法 SAS 代码

♣ SAS 代码:

```
/*先标准化*/
proc standard data=yourdata m=0 std=1 out=stout;
run;

/*选择凝聚点种子，分几类就几个种子*/
data seed;
set stout;
if _n_=10 then output;
if _n_=20 then output;
if _n_=30 then output;
run;

/*进行快速聚类*/
proc fastclus maxclusters=3 data=stout seed=seed mean=stat
    out=output;
run;
```