MAT-269: Análisis de Conglomerados

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Objetivo:

El análisis de conglomerados intenta descubrir grupos (o cluster) de observaciones que son homogéneas dentro de cada grupo.

Problema:

Dividir el análisis en dos pasos fundamentales.

- Elección de la medida de proximidad (similaridad).
- Selección del algoritmo de construcción de grupos.

Nos concentraremos en tres tipos de procedimiento de agrupamiento:

- Métodos jerarquicos aglomerativos.
- Métodos tipo K-means.
- Métodos de clasificación ML.

Estas técnicas operan sobre una matriz $\boldsymbol{D} = (d_{ij}) \in \mathbb{R}^{n \times n}$ de distancias[1] entre los puntos de $\boldsymbol{X} \in \mathbb{R}^{n \times p}$,

$$\boldsymbol{D} = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & d_{nn} \end{pmatrix}.$$

Por ejemplo, podríamos usar la distancia Euclidiana,

$$d_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = \Big\{ \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \Big\}^{1/2}, \qquad i, j = 1, \dots, n.$$

Note que, si d_{ij} es una distancia, entonces $d'_{ij} = \max_{ij}\{d_{ij}\} - d_{ij}$ es una medida de proximidad.

[1] $\boldsymbol{D}$ es construída usando medidas de similaridad o de disimilaridad

Tipo de distancias:

- Norma Euclidiana con un métrica $\boldsymbol{A} > \mathbf{0}$,

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_A = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \boldsymbol{A}(\boldsymbol{x}_i - \boldsymbol{x}_j)},$$

  es usual tomar $\boldsymbol{A} = \boldsymbol{S}^{-1}$ o bien $\boldsymbol{A} = \mathrm{diag}(s_{11}^{-1}, \ldots, s_{pp}^{-1})$.

- Métrica de Minkowski

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left\{ \sum_{k=1}^{p} |x_{ik} - x_{jk}|^m \right\}^{1/m}.$$

- Métrica Canberra

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{(x_{ik} + x_{jk})}.$$

- Coeficiente de Czekanowski

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 - 2\frac{\sum_{k=1}^{p} \min(x_{ik}, x_{jk})}{\sum_{k=1}^{p} (x_{ik} + x_{jk})}.$$

Example:

Para los datos de Iris, tenemos:

$$\boldsymbol{D} = \begin{pmatrix} 0.00 & & & & & & & & & \\ 0.54 & 0.00 & & & & & & & & \\ 0.51 & 0.30 & 0.00 & & & & & & & \\ 0.65 & 0.33 & 0.24 & 0.00 & & & & & & \\ 0.14 & 0.61 & 0.51 & 0.65 & 0.00 & & & & & \\ 0.62 & 1.09 & 1.09 & 1.17 & 0.62 & 0.00 & & & & \\ 0.52 & 0.51 & 0.26 & 0.33 & 0.46 & 0.99 & 0.00 & & & \\ 0.17 & 0.42 & 0.41 & 0.50 & 0.22 & 0.70 & 0.42 & 0.00 & & \\ 0.92 & 0.51 & 0.44 & 0.30 & 0.92 & 1.46 & 0.55 & 0.79 & 0.00 & \\ \vdots & \vdots & & & & & & & & \ddots \end{pmatrix}$$

En este caso, tenemos que $\boldsymbol{D}$ es una matriz simétrica 150×150.

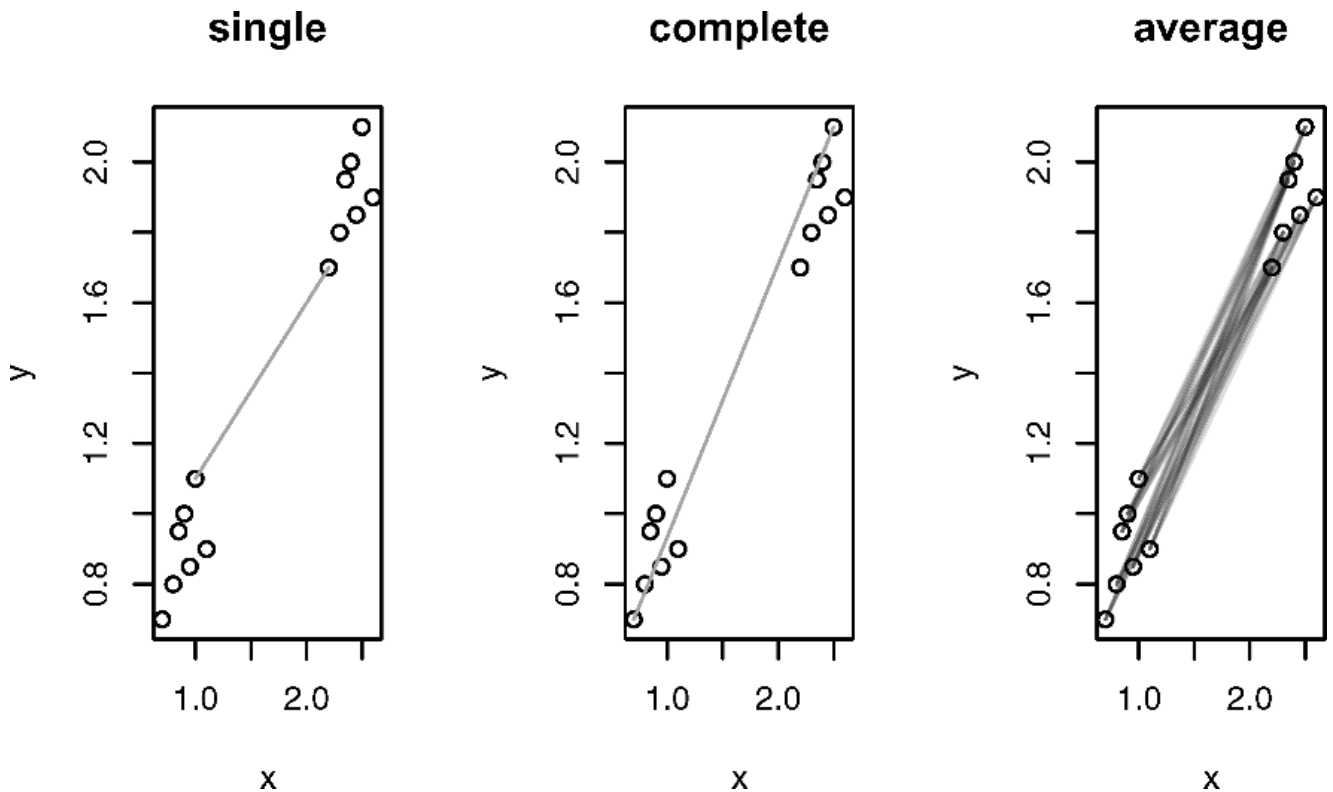
Suponga dos objetos o grupos P y Q, y sea

$$n_P = \sum_{i=1}^{n} I(\boldsymbol{x}_i \in P),$$

el número de objetos en P, y análogamente para n_Q. Considere los siguientes procedimientos para agrupar las observaciones:

- single linkage:
$$d(P, Q) = \min_{i \in P, i \in Q} \{d_{ij}\},$$
- complete linkage:
$$d(P, Q) = \max_{i \in P, i \in Q} \{d_{ij}\},$$
- average linkage:
$$d(P, Q) = \frac{1}{n_P n_Q} \sum_{i \in P} \sum_{i \in Q} d_{ij}.$$

Análisis de Conglomerados



Suponga dos objetos o grupos P y Q que están unidos, y deseamos calcular la distancia entre este nuevo grupo $P+Q$ con un grupo R, digamos:

$$d(R, P+Q) = \delta_1 d(R,P) + \delta_2 d(R,Q) + \delta_3 d(P,Q) + \delta_4 |d(R,P) - d(R,Q)|,$$

donde diferentes elecciones de las ponderaciones δ_i's da origen a distintos tipos de algoritmos aglomerativos.

Sea

$$n_P = \sum_{i=1}^{n} I(\boldsymbol{x}_i \in P),$$

el número de objetos en P, y análogamente para n_Q y n_R. Por ejemplo,

Linkage	δ_1	δ_2	δ_3	δ_4
single	$1/2$	$1/2$	0	-1/2
complete	$1/2$	$1/2$	0	$1/2$
average	$1/2$	$1/2$	0	0
median	$1/2$	$1/2$	-1/4	0
centroid	$\frac{n_P}{n_P+n_Q}$	$\frac{n_Q}{n_P+n_Q}$	$-\frac{n_P n_Q}{(n_P+n_Q)^2}$	0

Algoritmo 1: Método Jerárquico Aglomerativo.

Entrada: Matriz de datos $\boldsymbol{X} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top)^\top$.

```
1 begin
2     Construir la partición más fina.
3     Calcular la matriz de distancias D.
4     do
5         Hallar dos grupos con la distancia más cercana.
6         Agrupar dos grupos en un único grupo.
7         Calcular la distancia entre los nuevos grupos y obtener una matriz
            reducida D.
8     until todos los grupos están aglomerados en X
9 end
```

Ejemplo:

Considere

$$\boldsymbol{x}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \boldsymbol{x}_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \boldsymbol{x}_3 = \begin{pmatrix} 5 \\ 5 \end{pmatrix}$$

El algoritmo inicia con $K = 3$ grupos, $P = \{\boldsymbol{x}_1\}$, $Q = \{\boldsymbol{x}_2\}$, $R = \{\boldsymbol{x}_3\}$. La matriz de distancias $\boldsymbol{D}$ es dada por

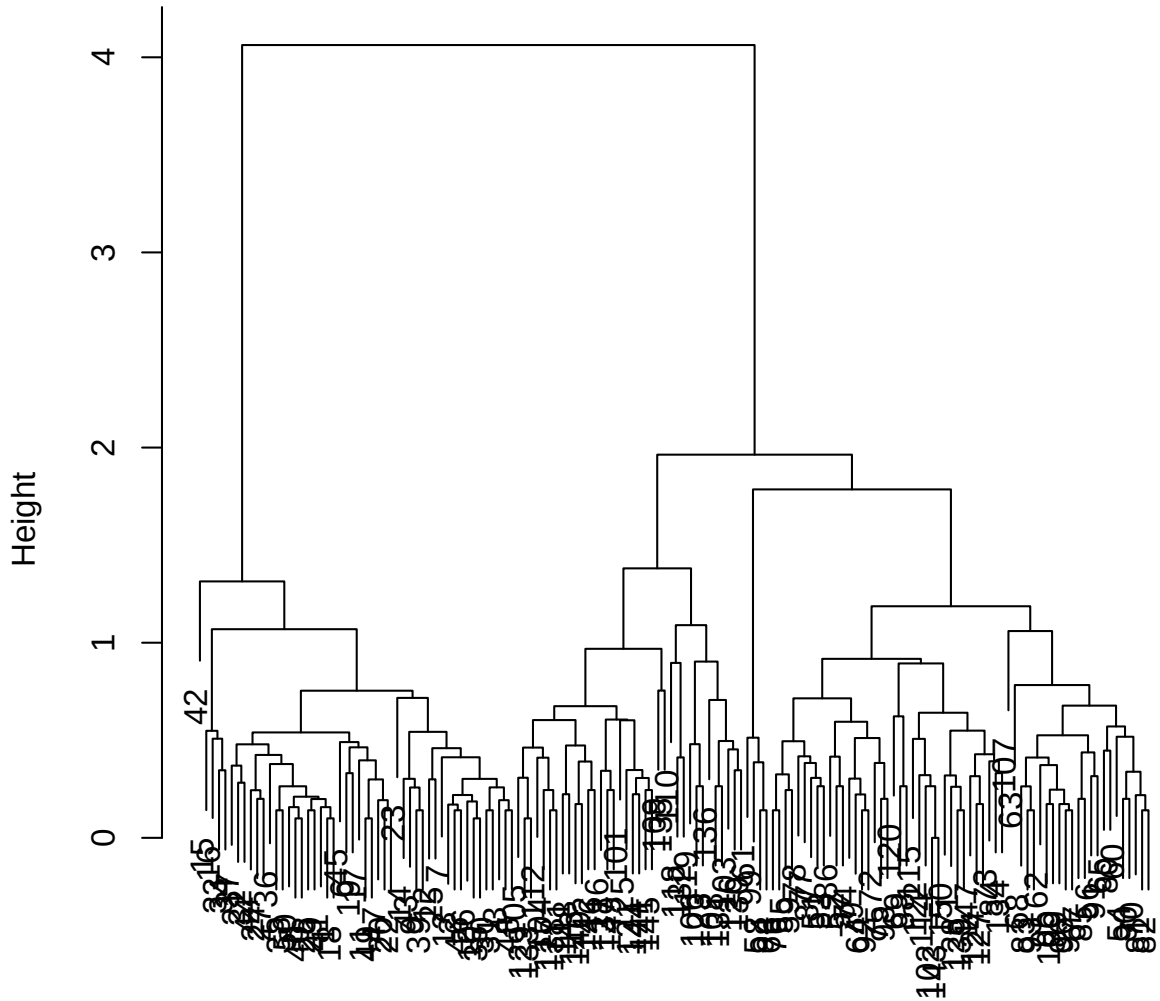
$$\boldsymbol{D} = \begin{pmatrix} 0 & 1 & 50 \\ 1 & 0 & 41 \\ 50 & 41 & 0 \end{pmatrix}.$$

La menor distancia en $\boldsymbol{D}$ se encuentra entre los grupos P y Q. De esta forma estos grupos se deben combinar en $P + Q = \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$. Usando single linkage, obtenemos

$$\begin{aligned} d(R, P+Q) &= \tfrac{1}{2} d(R,P) + \tfrac{1}{2} d(R,Q) - \tfrac{1}{2} |d(R,P) - d(R,Q)| \\ &= \tfrac{1}{2} d_{13} + \tfrac{1}{2} d_{23} - \tfrac{1}{2} |d_{13} - d_{23}| = \tfrac{50}{2} + \tfrac{41}{2} - \tfrac{|50-41|}{2} = 41. \end{aligned}$$

y la matriz de distancias *reducida* adopta la forma $\boldsymbol{D}_* = \begin{pmatrix} 0 & 41 \\ 41 & 0 \end{pmatrix}$. Detenemos el algoritmo uniendo los grupos R y $P + Q$ para formar el cluster $\boldsymbol{X}$, la matriz de datos original.

Análisis de Conglomerados



K-means busca particionar los n individuos en K grupos, digamos $G_1, G_2, \ldots, G_K$. El tipo más común de algoritmo halla una partición que minimice la suma de cuadrados dentro-de-grupo,

$$WGSS = \sum_{j=1}^{q} \sum_{r=1}^{K} \sum_{i \in G_r} (x_{ij} - \overline{x}_j^{(r)})^2,$$

donde $\overline{x}_j^{(r)} = \frac{1}{n_i} \sum_{i \in G_r} x_{ij}$.

n	k	Num. de particiones posibles
15	3	2 375 101
20	4	45 232 115 901
25	8	690 223 721 118 368 580
100	5	10^{68}

Algoritmo 2: Método K-medias.

Entrada: Matriz de datos $\boldsymbol{X} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top)^\top$.

```
1 begin
2     Hallar una partición inicial de los individios en los K grupos.
3     do
4         Proceder a través de la lista de elementos y asignar una observación al
            grupo cuyo centroide (media) sea más cercano.
5         Recalcular centroides.
6     until no se pueda hacer más asignaciones.
7 end
```

Observación:

El método de K-medias sufre principalmente de dos problemas:

- No es invariante a transformaciones de escala.
- Impone una estructura "esférica" a los datos.

El procedimiento de agrupamiento por ML es basado en asumir G subpoblaciones

$$f_j(\boldsymbol{x}; \boldsymbol{\theta}_j), \qquad \boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_G^\top)^\top.$$

Además, se introduce un vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^\top$ donde $\gamma_i = k$ si $\boldsymbol{x}_i$ pertenece a la k-ésima población.

De este modo, el problema de agrupamiento resulta de escoger $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_G^\top)^\top$ y $\boldsymbol{\gamma}$ maximizando la verosimilitud:

$$L(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} f_{\gamma_i}(\boldsymbol{x}_i; \boldsymbol{\theta}_{\gamma_i}).$$

Bajo normalidad tenemos $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, $j = 1, \ldots, G$ y los MLE de $\boldsymbol{\mu}_j$ son

$$\overline{\boldsymbol{x}} = \frac{1}{n_j} \sum_{i \in A_j} \boldsymbol{x}_i,$$

con $A_j = \{i : \gamma_i = j\}$ y n_j es el número de elementos de A_j. En este caso, la función de log-verosimilitud perfilada adopta la forma:

$$\ell(\boldsymbol{\theta}, \boldsymbol{\gamma}) = c - \frac{n}{2} \sum_{j=1}^{G} \Big\{ \operatorname{tr} \boldsymbol{S}_j \boldsymbol{\Sigma}_j^{-1} + \log |\boldsymbol{\Sigma}_j| \Big\}.$$

Una manera de caracterizar este tipo de funciones de densidad es asumiendo una mezcla discreta de densidades, como:

$$f(\boldsymbol{x}; \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{j=1}^{G} \pi_j f_j(\boldsymbol{x}; \boldsymbol{\theta}_j),$$

donde $\boldsymbol{x}$ es un vector aleatorio p-dimensional, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)^\top$, y $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_G^\top)^\top$, con π_j siendo las proporciones de la mezcla y f_j las densidades que componen la mezcla. Además,

$$\sum_{j=1}^{G} \pi_j = 1.$$

Una vez estimados los parámetros de la mezcla, las observaciones pueden ser asociadas con un particular cluster en basde de la probabilidad posterior estimada,

$$\widehat{\mathrm{P}}(\text{cluster } j | \boldsymbol{x}_i) = \frac{\widehat{\pi}_j f_j(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}_j)}{f(\boldsymbol{x}_i; \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\theta}})}, \qquad j = 1, \ldots, G. \tag{1}$$

Dada una muestra de observaciones $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ desde una mezcla discreta de densidades, tenemos la función de log-verosimilitud

$$\ell_n(\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\boldsymbol{x}_i; \boldsymbol{\pi}, \boldsymbol{\theta}).$$

En el caso de que el j-ésimo componente siga una distribución normal multivariada con media $\boldsymbol{\mu}_j$ y covarianza $\boldsymbol{\Sigma}_j$, se puede mostrar que

$$\widehat{\pi}_j = \frac{1}{n} \sum_{i=1}^{n} \widehat{\mathsf{P}}(j|\boldsymbol{x}_i),$$

$$\widehat{\boldsymbol{\mu}}_j = \frac{1}{n\widehat{\pi}_j} \sum_{i=1}^{n} \widehat{\mathsf{P}}(j|\boldsymbol{x}_i)\boldsymbol{x}_i,$$

$$\widehat{\boldsymbol{\Sigma}}_j = \frac{1}{n} \sum_{i=1}^{n} \widehat{\mathsf{P}}(j|\boldsymbol{x}_i)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_j)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_j)^\top,$$

donde $\widehat{\mathsf{P}}(j|\boldsymbol{x}_i)$ son las probabilidades estimadas definidas en (1). Este procedimiento es un caso particular del algoritmo EM para estimación ML en mezclas discretas.