Introducción a Data Mining

Análisis Cluster

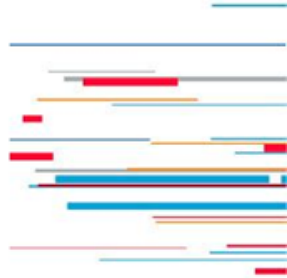
Maestría en Ciencia de Datos

Docentes: Mg. Pablo Beltramone
Mg. Leandro Kovalevski

UNIVERSIDAD AUSTRAL
CIENCIAS EMPRESARIALES

Agenda para hoy

- Concepto de clustering (o agrupamiento)
- Métodos de aglomeración
- La importancia de la elección de la medida de distancia
- Aplicaciones
- Programación en R



UNIVERSIDAD AUSTRAL
CIENCIAS EMPRESARIALES

Análisis Cluster - Introducción

Bajo el nombre de Análisis Cluster ó de conglomerados se conocen a las técnicas de agrupamiento de elementos en grupos que reflejen las relaciones y/o parecidos entre ellos.

Normalmente se agrupan individuos, pero el análisis de conglomerados puede también aplicarse para agrupar variables

Análisis Cluster - Introducción

Estos métodos se conocen también con el nombre de métodos de **clasificación no supervisada**, haciendo referencia a que no se conoce de antemano el grupo al que pertenecen los individuos (ó variables).

Algunos métodos de clasificación supervisada son: Análisis Discriminante, Árboles de clasificación, Bagging, Random Forest, K-vecinos, etc.

Análisis Cluster - Objetivo

Objetivo: Agrupar elementos en grupos homogéneos en función de las similitudes (o *similaridades*) entre ellos.

¿Existen realmente sub-grupos de individuos (o variables)?

¿Cuántos grupos forman? ¿Qué grado de parecido exigir para que dos individuos se consideren del mismo grupo?

¿Cómo medir el parecido (“similaridad”) ó el descuerdo (“distancia”) entre los individuos?

Métodos de aglomeración

En análisis cluster, los métodos de aglomeración pueden ser jerárquicos y no jerárquicos.

Los **métodos jerárquicos** tienen en cuenta el orden de parecido entre las observaciones a clasificar, el agrupamiento se realiza de forma jerárquica según la similitud de las observaciones.

Los métodos jerárquicos pueden ser:

- Divisivos ó Aglomerativos
- Monotéticos ó Politéticos

Métodos de aglomeración

Los **métodos no jerárquicos**, no tienen en cuenta el orden de parecidos para agrupar las observaciones. Son siempre politéticos y divisivos. Es necesario determinar a priori cuantos grupos se pretender formar.

Se parte de un único grupo, se lo divide en la cantidad de grupos deseados y se van pasando individuos de un grupo a otro tratando de aumentar la homogeneidad dentro de los grupos.

Métodos de aglomeración

Métodos jerárquicos → matriz de cuadrada de distancias de orden n x n para agrupar individuos (ó de orden p x p para agrupar variables)

Métodos no jerárquicos → matriz de individuos por variables ó de individuos por coordenadas en ejes factoriales

Métodos jerárquicos

Métodos de aglomeración jerárquicos

Matriz de datos de *n* individuos por *p* variables

$$\mathbf{X}_{n*p} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

$$\mathbf{S}_{n*n} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{pmatrix}$$

Matriz de similaridades (ó distancias)

Métodos de aglomeración jerárquicos

El procedimiento de agrupamiento es progresivo, respetando el grado de parecidos.

1. Cada individuo es un grupo.
2. Se forma un nuevo grupo reuniendo a los dos individuos con mayor s_{ij} (ó menor distancia).
3. Se recalculan los parecidos s_{ij} entre los grupos formados y el resto de los individuos, ó entre dos de los grupos formados.
4. Se juntan nuevamente los individuos y/o grupos más parecidos.
5. Se continúa hasta haber reunido a todos los individuos en un solo grupo.

*Los distintos pasos del procedimiento se resumen en un diagrama denominado **dendograma**, y se decide cuántos grupos considerar.*

Métodos de aglomeración jerárquicos

Ejemplo: Se quiere agrupar 7 individuos según el grado de parecidos.

Las similaridades entre los individuos son las siguientes:

	I1	I2	I3	I4	I5	I6	I7
I1	1	.	.	.	.	.	.
I2	0,5	1	.	.	.	.	.
I3	0,2	0,33	1	.	.	.	.
I4	0,15	0,8	0,12	1	.	.	.
I5	0,32	0,42	0,54	0,81	1	.	.
I6	0,7	0,74	0,63	0,26	0,3	1	.
I7	0,91	0,96	0,8	0,77	0,51	0,6	1

Métodos de aglomeración jerárquicos

Se recalculan las similaridades entre los individuos y el nuevo grupo formado:

	I1	I3	I4	I5	I6	(I2, I7)
I1	1	.	.	.	.	.
I3	0,2	1	.	.	.	.
I4	0,15	0,12	1	.	.	.
I5	0,32	0,54	0,81	1	.	.
I6	0,7	0,63	0,26	0,3	1	.
(I2, I7)	0,705	0,565	0,785	0,465	0,67	1

¿Cómo se obtienen esos valores?

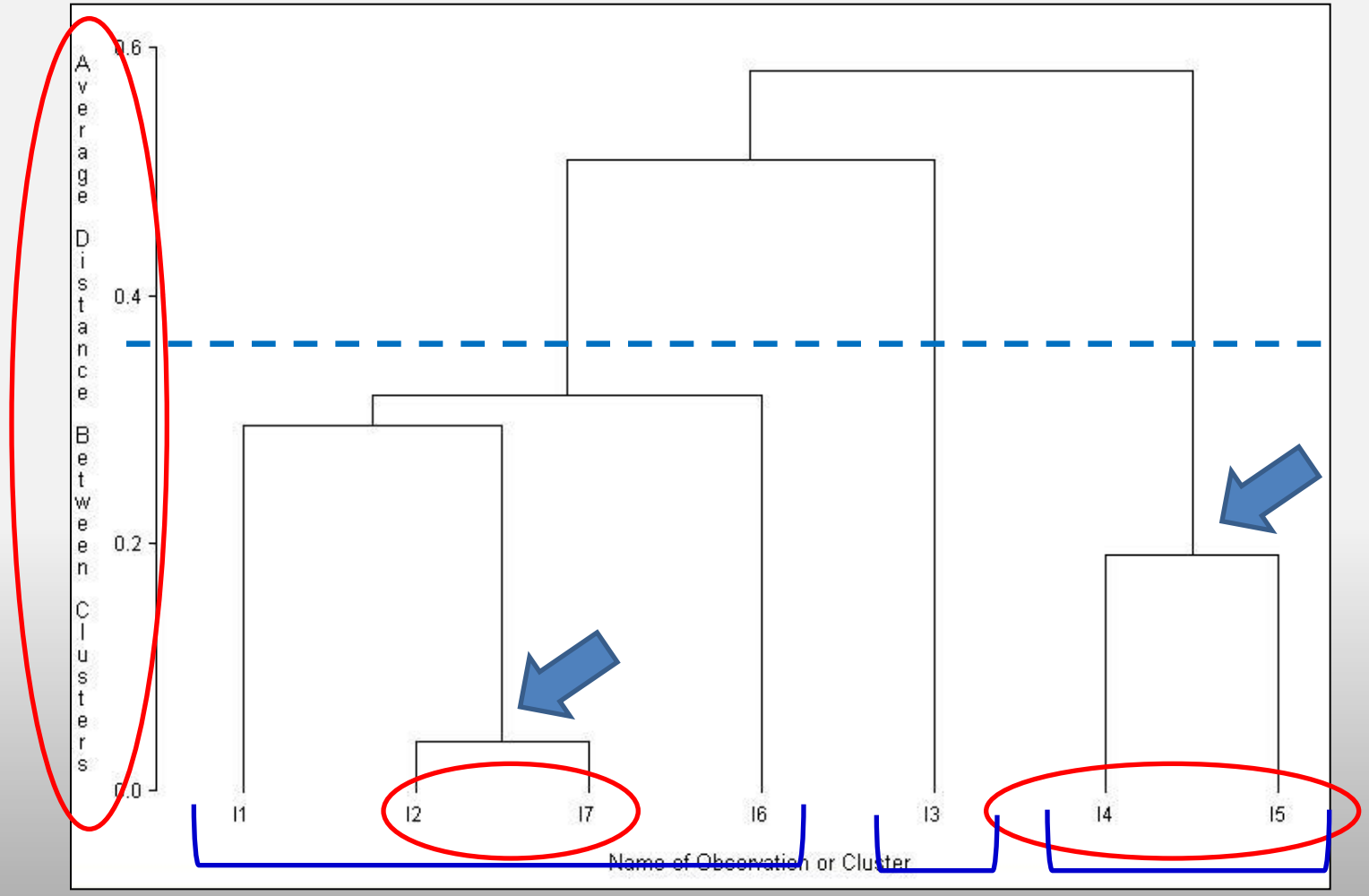
Métodos de aglomeración jerárquicos

Resumiendo el procedimiento:

Paso	Número de grupos	Individuos/grupos unidos		Individuos en el grupo formado	Distancia	Empates
1	7	-	-	-	-	
2	6	I2	I7	2	0,04	
3	5	I4	I5	2	0,19	
4	4	I1	CL6	3	0,3	
5	3	CL4	I6	4	0,32	
6	2	CL3	I3	5	0,51	
7	1	CL2	CL5	7	0,58	

Métodos de aglomeración jerárquicos

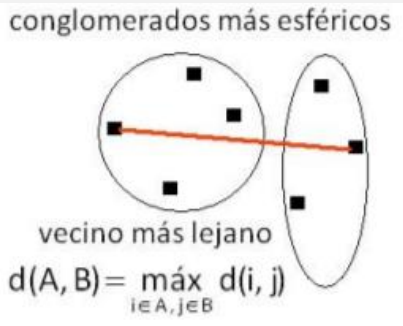
Resumiendo el procedimiento gráficamente:

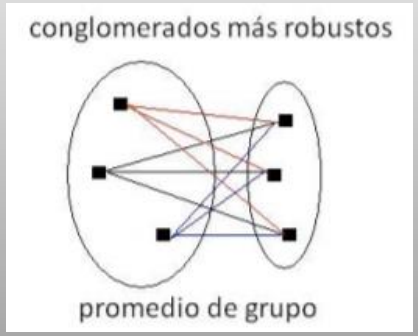


Métodos de aglomeración jerárquicos – criterios de aglomeración

Existen distintos **criterios** posibles **de aglomeración** (de cómo medir la distancia a un grupo)

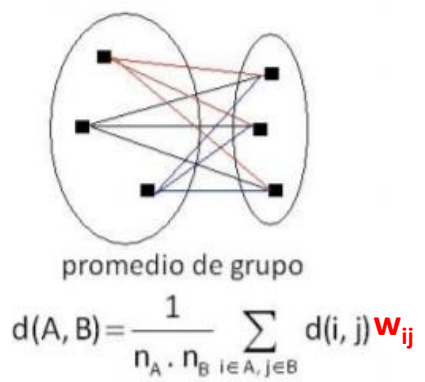
- *encadenamiento simple:* se elige como la distancia entre un individuo y un grupo al mínimo de las distancias entre el individuo y cada elemento del grupo.
- *encadenamiento completo*: se elige como la distancia entre un individuo y un grupo al máximo de las distancias entre el individuo y cada elemento del grupo.
- *promedio simple*: la distancia está dada por el promedio de las distancias entre el individuo y cada elemento del grupo.

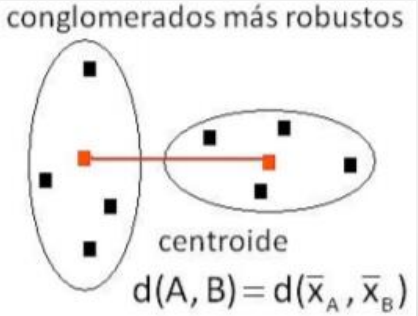


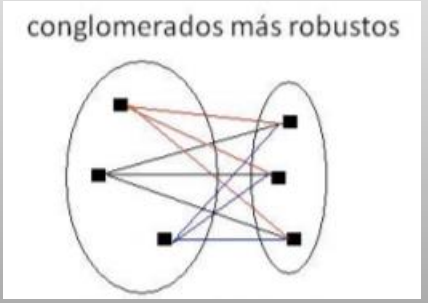


Métodos de aglomeración jerárquicos – criterios de aglomeración

- *promedio ponderado*: la distancia está dada por un promedio ponderado de las distancias entre el individuo y cada elemento del grupo.
- *distancia centroide (o prototipo)*: se elige un representante del grupo (por ejemplo, podría ser el individuo “promedio”) y la distancia al grupo está dada por la distancia a ese representante elegido.
- *distancia mediana*: la distancia está dada por la mediana de las distancias entre el individuo y cada elemento del grupo.

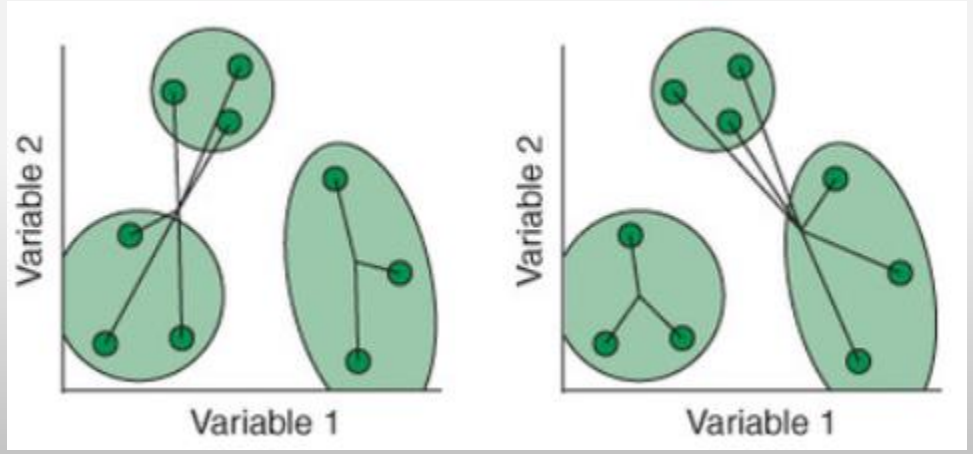






Métodos de aglomeración jerárquicos – criterios de aglomeración

- *mínima variancia de Ward*: la distancia está dada por la pérdida de homogeneidad cuando se unen dos grupos. En su cálculo intervienen el vector de medias de cada grupo y la cantidad de individuos en los grupos.



¿Cómo construir las distancias o similaridades?

¿Cómo construir las distancias o *similaridades*?

Depende del tipo de información disponible y de la clase de objetos que pretendo agrupar.

Variables dicotómicas – Para agrupar individuos

Existen distintas alternativas para calcular la *similaridad* entre dos individuos a través de *p* variables dicotómicas según la importancia que se le quiera dar a las coincidencias presencia-presencia (1-1) y las coincidencias ausencia-ausencia (0-0):

Métodos de aglomeración jerárquicos – Distancias y *similaridades*

Llamamos con "*a*" a la cantidad de variables en que ambos individuos toman el valor 1, es decir, a la frecuencia de coincidencias 1-1, "*b*" a la frecuencia de pares 1-0, "*c*" a la frecuencia de pares 0-1 y con "*d*" a la frecuencia de coincidencias 0-0 a través de las p variables.

		Ind 2	
		1	0
Ind 1	1	a	b
	0	c	d

a + b + c + d = p

Coeficientes de similaridad para agrupar objetos (p variables dicotómicas)

Coeficiente	Característica
(a+d)/p	Igual "peso" a las coincidencias 1-1 y 0-0
2(a+d)/[2(a+d)+b+c]	Doble "peso" a las coincidencias 1-1 y 0-0 que a las no coincidencias
(a+d)/[a+d+2(b+c)]	Doble "peso" a las no coincidencias que a las coincidencias
a/p	No intervienen las coincidencias 0-0 en el numerador
a/(a+b+c)	No se consideran las coincidencias 0-0
2a/(2a+b+c)	No se consideran las coincidencias 0-0 y se le da "peso" doble a las coincidencias 1-1
a/[a+2(b+c)]	No se consideran las coincidencias 0-0 y se le da "peso" doble a las no coincidencias
a/(b+c)	Razón de coincidencias 1-1 y no coincidencias

Métodos de aglomeración jerárquicos – Distancias y *similaridades*

Variables cuantitativas (intervalo/razón) – Para agrupar individuos

Entre las alternativas más comunes encontramos las distancias euclídea, de Minkosky y de Mahalanobis.

La distancia entre el individuo *i* y el individuo *j* con coordenadas $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$ y $\mathbf{x}_j = (x_{j1}, x_{j2}, ..., x_{jp})$ respectivamente está dada por:

Euclídea: d(i,j) $= \sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\ldots+(x_{ip}-x_{jp})^2} = \sqrt{\sum_{k=1}^{p}(x_{ik}-x_{jk})^2}$

Minkosky: d(i,j) $= \sqrt[H]{\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|^H}$

Mahalanobis: d(i,j) $= \sqrt{(\boldsymbol{x}_i-\boldsymbol{x}_j)'*S^{-1}*(\boldsymbol{x}_i-\boldsymbol{x}_j)}$

Métodos de aglomeración jerárquicos – Distancias y *similaridades*

Cuando los datos originales no son conmensurables conviene estandarizar, tipificar o normalizarlos en un paso previo al cálculo de distancias. Algunas opciones pueden ser:

$$x_{ik}^{*} = \frac{x_{ik} - \overline{x}_k}{S_k} \qquad x_{ik}^{*} = \frac{x_{ik}}{\max_i x_{ik} - \min_i x_{ik}} \qquad x_{ik}^{*} = \frac{x_{ik}}{\max_i x_{ik}} \qquad x_{ik}^{*} = \frac{x_{ik}}{\sqrt{\sum_{i=1}^{n} x_{ik}^2}}$$

Mezcla de variables categóricas y cuantitativas

En este caso no existe una única metodología.

Algunas alternativas pueden ser:

- Usar coordenadas principales para resumir la información de las variables cualitativas y tratar todo como variables cuantitativas.
- Transformar las variables numéricas para que sus distancias varíen entre 0 y 1.
- Usar distancia a partir de la *similaridad* de Gower.

Mezcla de variables categóricas y cuantitativas

Similaridad de Gower

La similaridad entre el individuo *i* y el individuo *i'*

está dada por la fórmula:

$$s_{ii'} = \frac{\sum_{j=1}^{p_1}\left(1 - \frac{|x_{ij} - x_{i'j}|}{r_j}\right) + a + d + \alpha}{p_1 + p_2 + p_3}$$

p_1 es el número de variables continuas
r_k rango de la *k*-ésima variable continua
p_2 número de variables binarias
a número de coincidencias en 1 de las variables binarias
d número de coincidencias en 0 de las variables binarias
p_3 número de variables cualitativas
α número de coincidencias de las variables cualitativas

Métodos de aglomeración jerárquicos – Distancias y similaridades

Variables cuantitativas (ordinal/razón) – Para agrupar variables

Se utilizan los coeficientes de correlación ρ_{ij} (pearson, spearman, ó kendall) como medida de similaridad ó simétricamente: $d^2_{ij} = 1 - \rho^2_{ij}$

¿Cómo elegir la métrica de distancia adecuada?

En función del objetivo del problema a resolver

Por ejemplo:

A = (1, 1, 0) B = (4, 4, 0) C = (0, 0, 1)

Distancia Euclídea: d(A,B) = 4,24 d(A,C) = 1,732

Distancia "de la cuerda":

e_A = (0,707; 0,707; 0) e_B = (0,707; 0,707; 0) e_C = (0; 0; 1)

y ahora d(A,B) = 0 d(A,C) = 1,414

Métodos no jerárquicos

Métodos no jerárquicos

Parten directamente de una matriz de individuos por variables o de individuos por coordenadas factoriales (siempre politéticos y divisivos).

Se siguen los siguientes pasos:

1. Se determina cuántos grupos se pretenden formar (*K*).
2. Se divide al grupo entero en *K* grupos. Al azar ó de acuerdo a cercanías respecto a *K* centros (elegidos arbitrariamente ó al azar).

Métodos de aglomeración no jerárquicos

3. Se mejora el agrupamiento de los individuos optimizando algún criterio:
   - cercanía a centros de grupos
   - minimización de traza de $\mathbf{S}_w$
   - minimización del determinante de $\mathbf{S}_w$
   - maximizar la traza de $\boldsymbol{S}_{\boldsymbol{w}}^{-1}\boldsymbol{B}$

siendo $\mathbf{S}_w$, de orden *p* x *p*, la matriz de variancias y covariancias intra-clusters

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\boldsymbol{x}_i^{(k)}-\bar{\boldsymbol{x}}^{(k)})(\boldsymbol{x}_i^{(k)}-\bar{\boldsymbol{x}}^{(k)})'$$

B, también de orden *p* x *p*, es la matriz de variancias y covariancias entre-clusters

$$\sum_{k=1}^{K}(\bar{\boldsymbol{x}}^{(k)}-\bar{\boldsymbol{x}})(\boldsymbol{x}_i^{(k)}-\bar{\boldsymbol{x}})'$$

Métodos de aglomeración no jerárquicos

4. Se recalculan los centros de grupo, la traza de **S**w, el determinante de **S**w, etc. Y se vuelve al paso 3.
5. Se detiene cuando no hay cambios que mejoren.

El procedimiento completo puede repetirse varias veces y finalmente se evalúa si los grupos formados son suficientemente parecidos mediante distintas medidas.

Métodos de aglomeración no jerárquicos

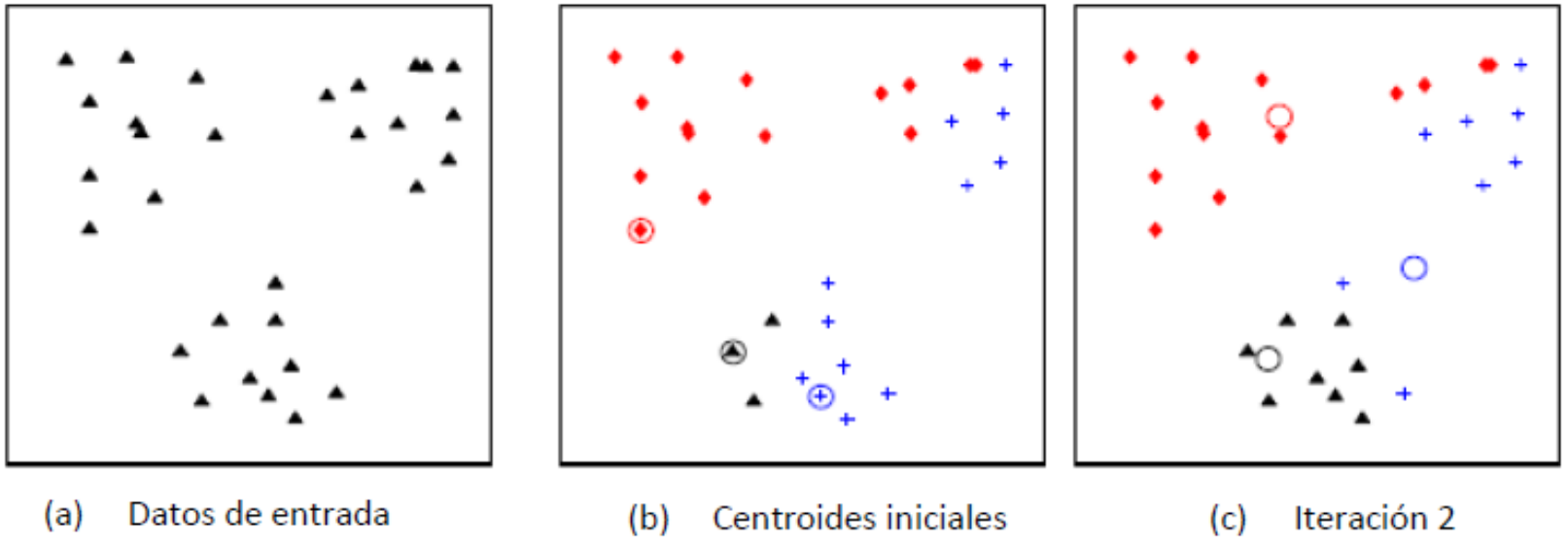
Cuando en el paso 3 se utiliza la cercanía a centros de grupos para mejorar el agrupamiento de los individuos y como centro de los grupos se utilizan los *K* vectores de medias, tenemos el conocido **método de agrupamiento *k-means***.

En *k–means* se busca minimizar la suma de errores al cuadrado (SCE) respecto de los promedios de los puntos de cada cluster.

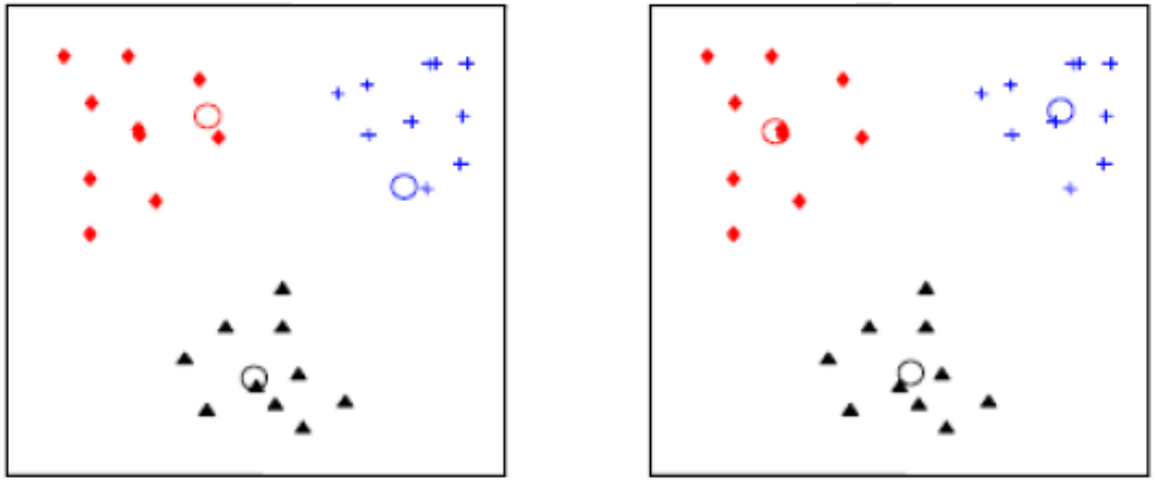
$$\sum_{k=1}^{K}\sum_{j=1}^{n_k}(\boldsymbol{x}_{kj} - \overline{\boldsymbol{x}}^{(k)})'(\boldsymbol{x}_{kj} - \overline{\boldsymbol{x}}^{(k)})$$

Métodos de aglomeración no jerárquicos

Ilustración algoritmo *k-means* en un conjunto de datos bivariados



(a) Datos de entrada (b) Centroides iniciales (c) Iteración 2



(d) Iteración 3 (e) Agrupamiento final

Ciardullo (2020). "Estudio comparativo de métodos de clasificación no supervisada en contextos de grandes bases de datos", adaptado de Jain, A. K. (2010) "Data Clustering: 50 Years Beyond K-means. Machine Learning and Knowledge Discovery in Databases".

Métodos de aglomeración no jerárquicos

Una variante al algoritmo *k–means* que tiene la ventaja de ser robusta a valores extremos es ***k-mediods*** que minimiza la suma de disimilaridades entre cada uno de los puntos de un cluster y el centro del mismo, que también es un punto perteneciente al cluster.

Existen infinidad de otros algoritmos de clusterización no jerárquica.

¿Cómo evaluar los grupos formados?

Homogeneidad en los grupos formados

Entre las distintas medidas para evaluar homogeneidad en los grupos formados, tenemos:

- **Índice Silhouette**

El índice Silhouette se define para cada individuo como:

$$s_k(i) = \frac{d^*_{x_i^{kk\prime}} - \bar{d}_{x_i^k}}{\max\{d^*_{x_i^{kk\prime}}, \bar{d}_{x_i^k}\}}$$

donde

$\bar{d}_{x_i^k} = \frac{1}{n_k}\sum_{j=1}^{n_k} d(x_i^k, x_j^k), j = 1, \dots, n_k$, es la distancia promedio entre el individuo $i-$é$simo$ y todos los otros individuos dentro del mismo cluster $k-$é$simo$.

$d^*_{x_i^{kk\prime}} = \min_{k'=1,\dots,K;k'\neq k}\left\{\frac{1}{n_{k'}}\sum_{j=1}^{n_{k'}} d(x_i^k, x_j^{k'})\right\}, j = 1, \dots, n_{k'}$, es la menor distancia promedio entre el individuo $i-$é$simo$ y todos los puntos de cualquier otro *cluster* k'=1,…,K con $k' \neq k$,

Homogeneidad en los grupos formados

Por su definición el índice Silhouette varía entre −1 y 1.

$$s_k(i) = \frac{d^*_{x_i^{kk'}} - \bar{d}_{x_i^k}}{\max\{d^*_{x_i^{kk'}}, \bar{d}_{x_i^k}\}}$$

- El Silhouette de un individuo x_i será cercano a 1 cuando la disimilaridad intra sea mucho menor que la disimilaridad entre ($d^*_{x_i^{kk'}} > \bar{d}_{x_i^k}$).

- Si el Silhouette es cercano a cero $d^*_{x_i^{kk'}}$ y $\bar{d}_{x_i^k}$ son muy cercanas, lo cual indica que el individuo x_i podría estar igualmente bien clasificado en el cluster vecino.

- Si el Silhouette toma valores cercanos a −1, $\bar{d}_{x_i^k}$ es mucho mayor que $d^*_{x_i^{kk'}}$, por lo que x_i esta mucho más cerca a los individuos del clúster vecino que a los de su propio clúster. En esta situación el individuo x_i está mal clasificado.

Homogeneidad en los grupos formados

- **<u>RMSSTD</u> (Root Mean Square Standard Deviation)**

Es la raíz cuadrada del promedio de las variancias de las *p* variables en el nuevo cluster formado

RMSSTD = $\sqrt{\frac{1}{p}\sum_{j=1}^{p} s_j^2}$ donde s_j^2 es la variancia de la variable j en el nuevo cluster formado, $j = 1, \ldots, p$.

RMSSTD es menor a mayor homogeneidad en los clusters.

Homogeneidad en los grupos formados

- **RS (R-squared)**

Es es el cociente entre la suma de cuadrados entre clusters y la suma de cuadrados total.

$$\text{RS} = \frac{SCB}{SCT} = \frac{\sum_{k=1}^{K} (\bar{x}^{(k)} - \bar{x})'(x_i^{(k)} - \bar{x})}{\sum_{k=1}^{K}\sum_{l=1}^{n_k} (x_{kl} - \bar{x})'(x_{kl} - \bar{x})} = 1 - \frac{SCW}{SCT}$$

RS es mayor para clusters más diferentes (entre sí).

- **distancias entre clusters**

Será mayor cuando los clusters están más separados.

Muchas gracias.

www.austral.edu.ar

UNIVERSIDAD AUSTRAL | Valores que inspiran