PASOS PARA SU APLICACIÓN

ANÁLISIS DE CONGLOMERADOS

MÉTODOS NO JERÁRQUICOS

Métodos no jerárquicos

Diseñados para clasificar individuos en *K* clústers. K se especifica *a priori* o bien se determina como una parte del proceso.

El funcionamiento general de estos métodos es elegir una partición inicial de individuos y después intercambiar los miembros de estos clústers para obtener una mejor partición.

La mayoría de las aplicaciones adoptan *métodos heurísticos*:

- *k-medias*: Cada clúster está representado por el valor medio de los objetos del clúster.
- *k-medianas* o PAM (Partition around medoids): Cada clúster está representado por uno de los objetos situados cerca del centro del clúster.

3. Elección de la técnica de agrupación

Elección de puntos semilla

Es necesario establecer un conjunto de *K* semillas que puedan emplearse como núcleo de los clústers sobre los cuales el conjunto de individuos puede agruparse. Algunos procedimientos son:

1. Elegir los primeros *K* individuos del conjunto de datos (McQueen, 1967). Cuidar que los individuos hayan sido incluidos aleatoriamente.

2. Etiquetar los casos de 1 a *m* y elegir aquellos etiquetados como

$$\left[\frac{m}{k}\right], \left[\frac{2m}{k}\right], \ldots, \left[\frac{(k-1)m}{k}\right], \mathrm{m}$$

3. Etiquetar los casos de 1 a *m* y elegir los casos correspondientes a *K* números aleatorios diferentes (McRae, 1971)

3. Elección de la técnica de agrupación

K – Medias de McQueen

Busca asignar cada individuo al clúster (de los *K* prefijados) con el centroide *más próximo*. El centroide es calculado a partir de los miembros del clúster tras cada asignación.

El algoritmo propuesto es el siguiente:

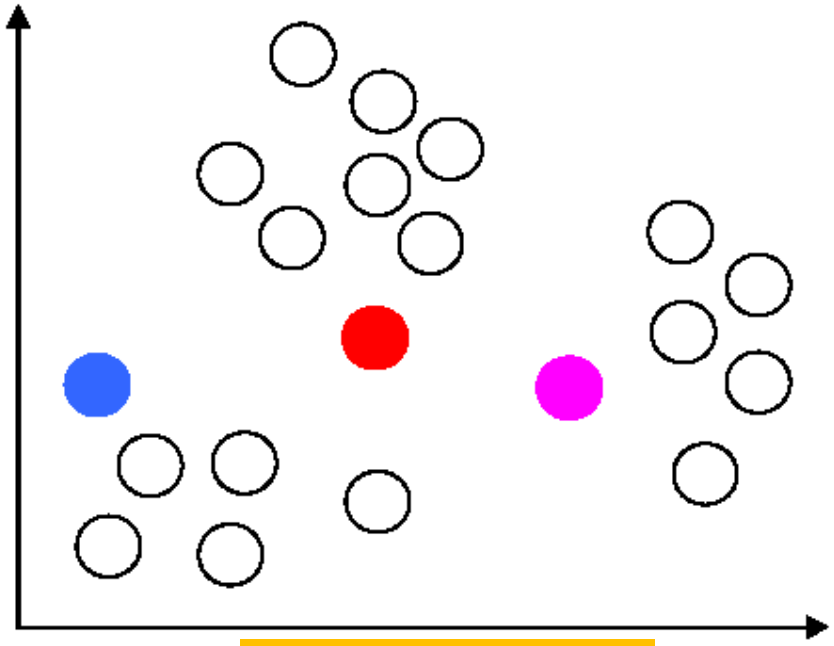
1. Tomar los *K* primeros casos como clústers unitarios.

2. Asignar cada uno de los *m-K* individuos restantes al clúster con el centroide más próximo. Después de cada asignación, recalcular el centroide del clúster obtenido.

3. Tras la asignación de todos los individuos en el paso anterior, tomar los centroides de los clústers existentes como puntos semilla fijos y hacer una pasada más sobre los datos asignados cada dato al punto semilla más cercano.

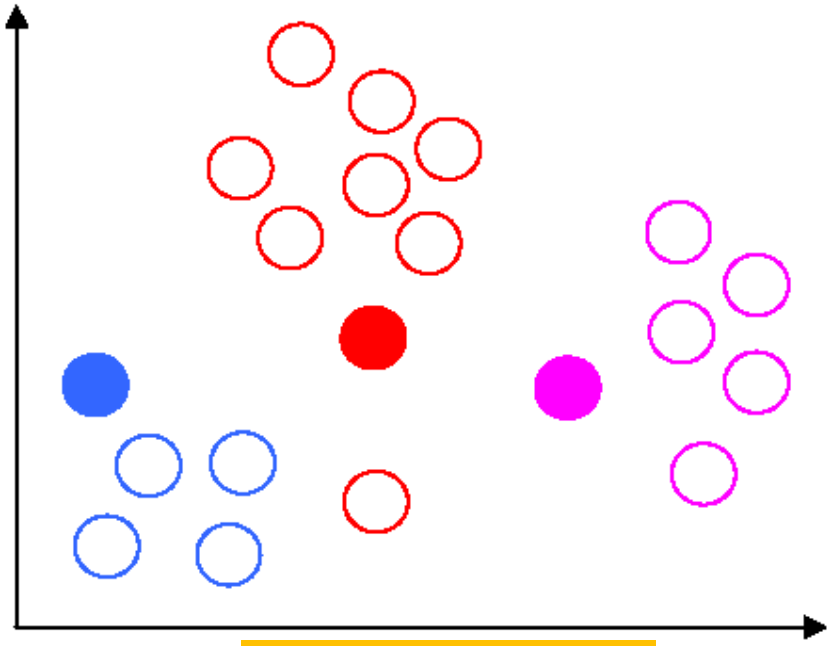
3. Elección de la técnica de agrupación

K – Medias de McQueen



PASO 1

Selección de las observaciones que serán las semillas para iniciar el algoritmo de agrupación

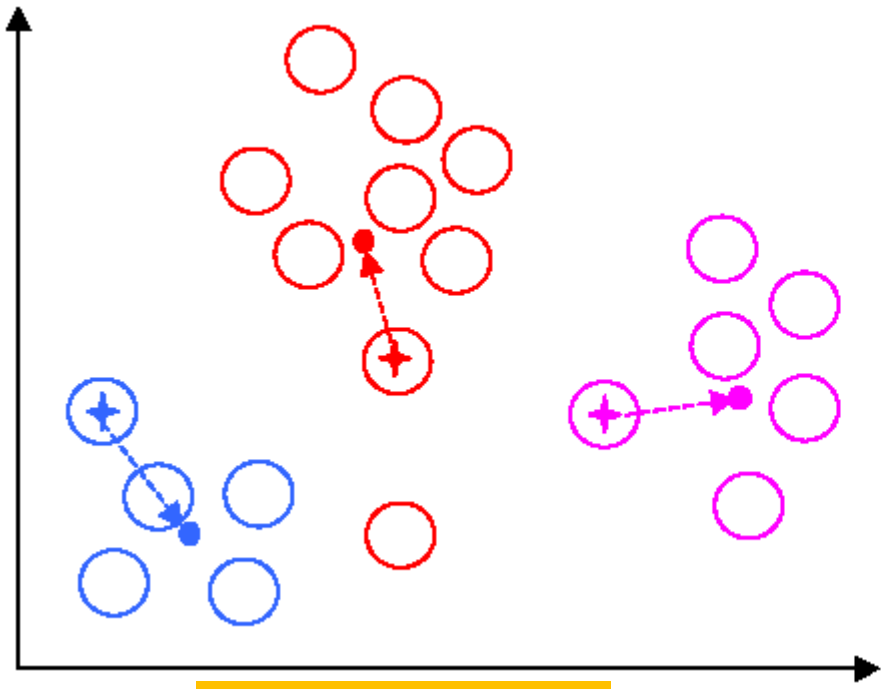


PASO 2

Agrupación inicial de los casos restantes tomando como centroide las observaciones consideradas como semillas

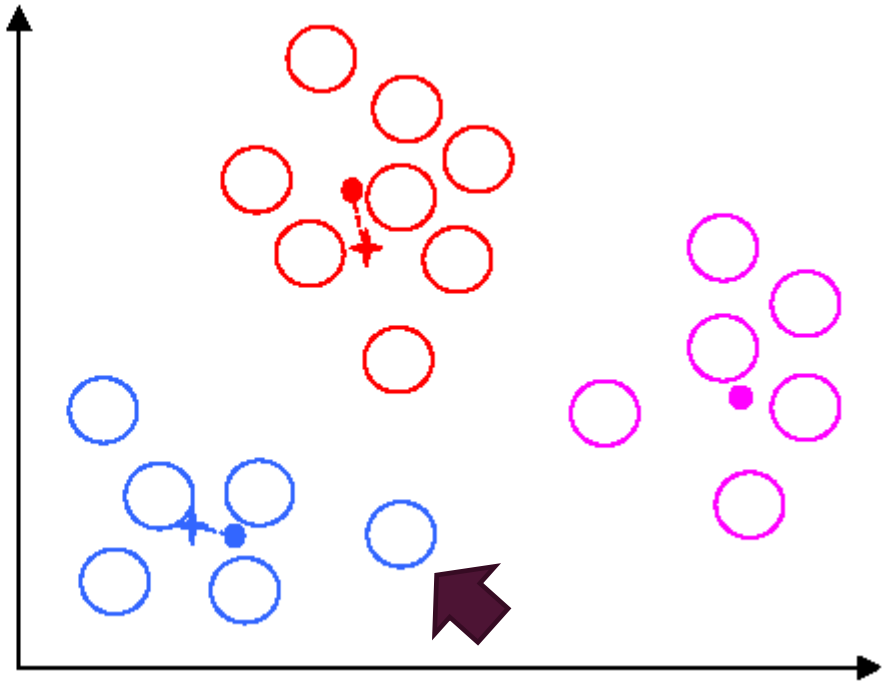
3. Elección de la técnica de agrupación

K – Medias de McQueen



PASO 3

Calcular el centriode de cada cluster.

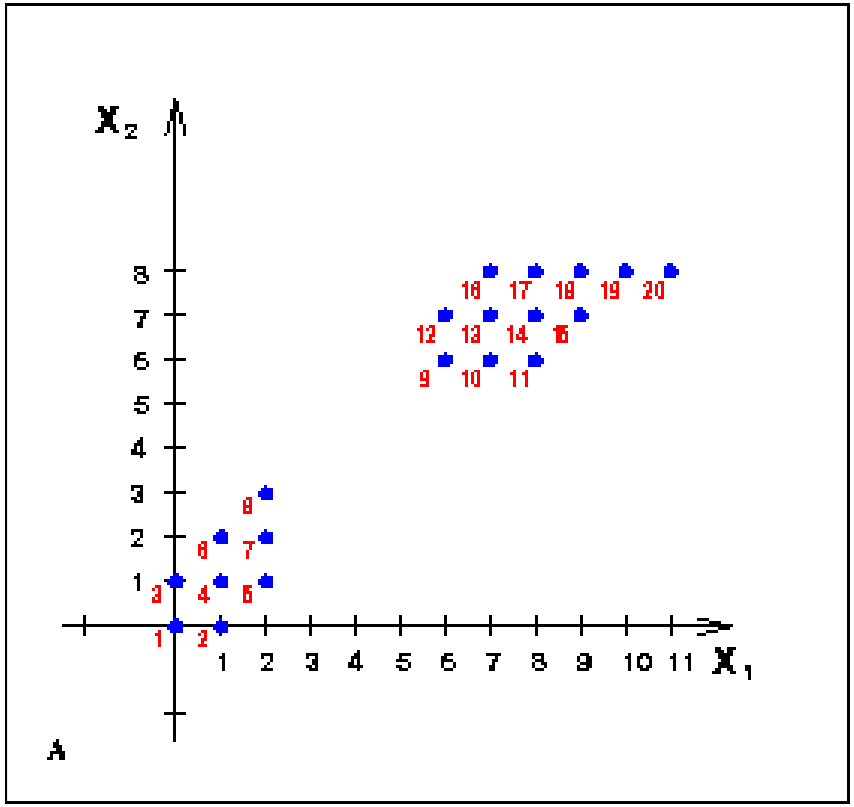


Mejor resultado encontrado

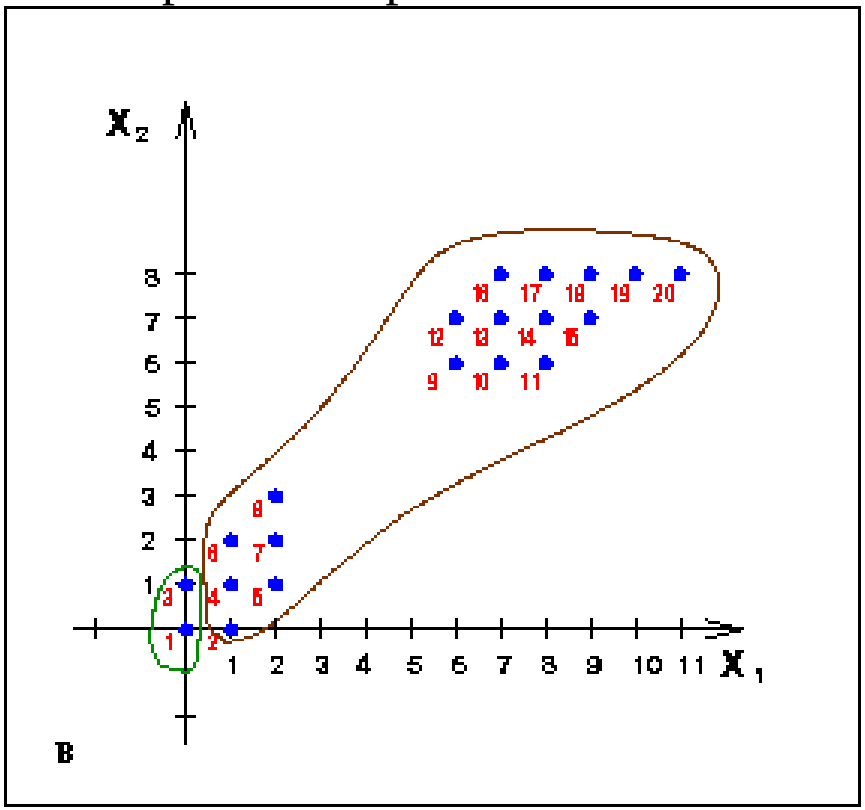
Se verifica si alguno de las observaciones esta más próxima a los nuevos centroides. Si alguna está más próxima se cambia de cluster y se recalcula el centroide. Repetir hasta que no haya cambios de cluster.

3. Elección de la técnica de agrupación

Situación inicial



Después de la primera iteración



Paso 1. $S_1(0) = \{X_1\}$ $Z_1(0) = (0, 0)$

$S_2(0) = \{X_2\}$ $Z_2(0) = (1, 0)$

Paso 2. $S_1(1) = \{X_1, X_3\}$ $Z_1(1) = (0, 0.5)$

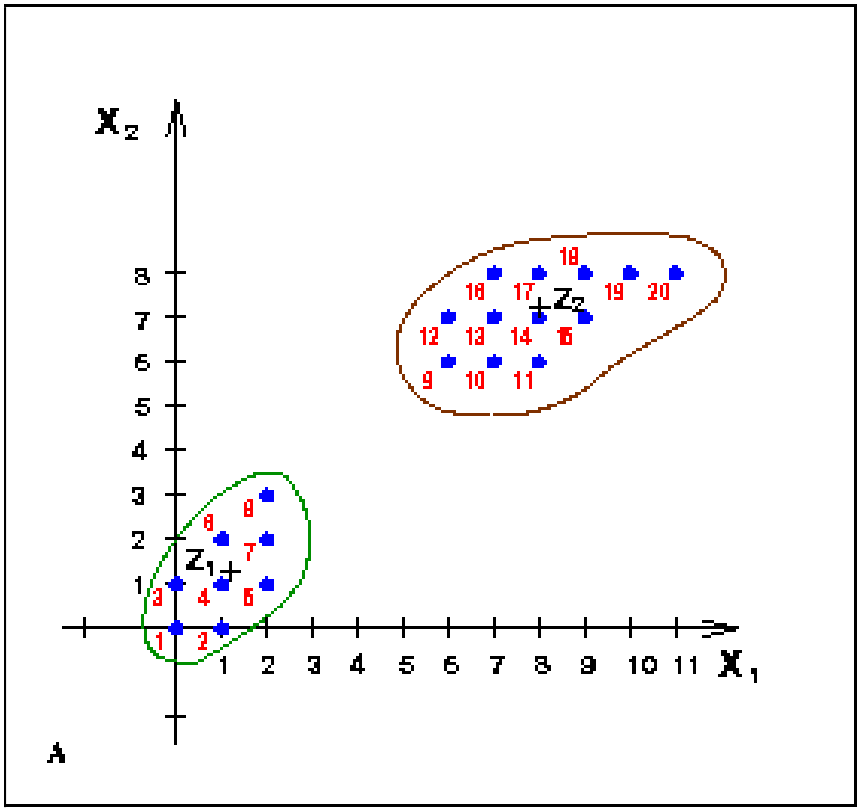
$S_2(1) = \{X_2,..., X_{20}\}$ $Z_2(1) = (5.8, 5.3)$

Paso 3. $Z_1(1) \neq Z_1(0)$ y $Z_2(1) \neq Z_2(0)$

Volver al paso 2

3. Elección de la técnica de agrupación

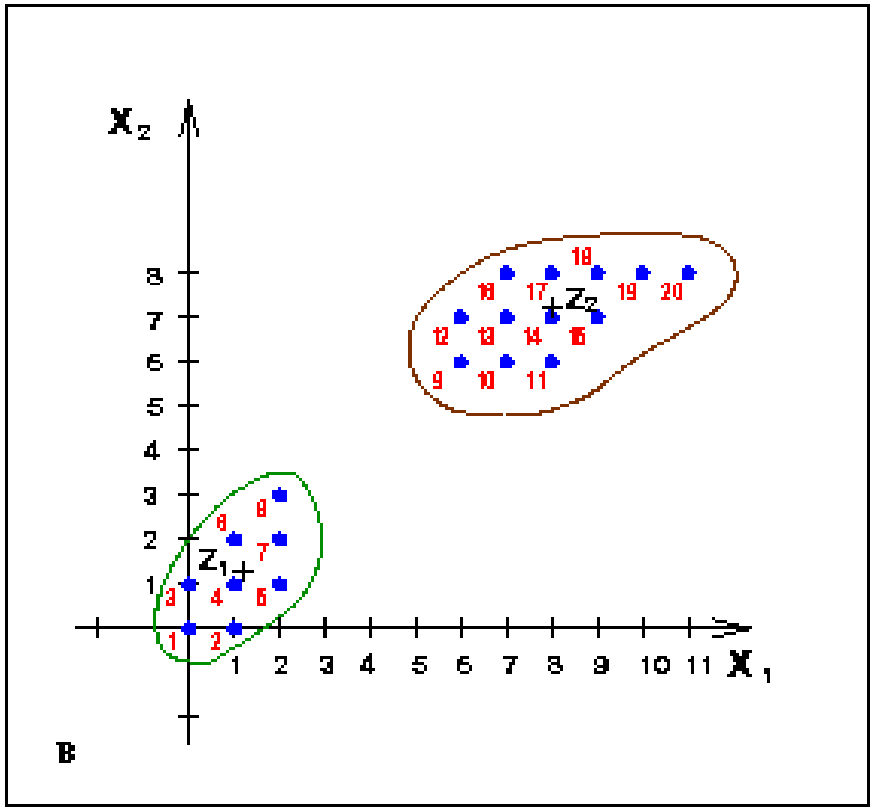
Segunda iteración



Paso 2. $S_1(2) = \{X_1,..., X_8\}$ $Z_1(2) = (1.1, 1.3)$
$S_2(2) = \{X_9,..., X_{20}\}$ $Z_2(2) = (8.0, 7.2)$

Paso 3. $Z_1(2) \neq Z_1(1)$ y $Z_2(2) \neq Z_2(1)$
Volver al paso 2

Tercera iteración



Paso 2. $S_1(3) = \{X_1,..., X_8\}$ $Z_1(3) = (1.1, 1.3)$
$S_2(3) = \{X_9,..., X_{20}\}$ $Z_2(3) = (8.0, 7.2)$

Paso 3. $Z_1(3) = Z_1(2)$ y $Z_2(3) = Z_2(2)$
FIN

3. Elección de la técnica de agrupación

Ventajas:

- Algoritmo es muy sencillo.
- Funciona bien para encontrar clústers con forma esférica.

Inconvenientes:

- El resultado final depende del valor de *K* y de la inicialización de los centros.
- Sólo para datos a los que se les puede aplicar la media.
- No adecuado para formas no convexas o clústers de diferentes tamaños.
- Sensible a ruidos y *outliers*.

3. Elección de la técnica de agrupación

Otros métodos no jerarquicos

- **K-modes**: Para datos cualitativos, reemplazando las medias por **modas**. Usando el total de *discordancias* entre dos objetos: *mientras más pequeño este número, más similar ambos objetos.*

$$d(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j)$$

$$\delta(x_j, y_j) = \begin{cases} 0 & x_j = y_j \\ 1 & x_j \neq y_j \end{cases}$$

- **K-prototypes**: Integración de K-medias y K-modes para datos cualitativos y cuantitativos.

3. Elección de la técnica de agrupación

- **K-medianas:** Desarrollado por Kaufman y Rousseeuw en 1987.
  Soluciona la sensibilidad del K-medias frente a los *outliers*.

Se toma como punto de referencia el objeto situado en el centro del clúster, en vez de tomar el valor medio. Más robusto que K-medias. Su procesamiento es más costoso.