UNIDAD DIDÁCTICA 3: Técnicas de agrupamiento
Tema 1: Análisis Clúster

TÉCNICAS ESTADÍSTICAS PARA EL APRENDIZAJE I

Máster Universitario en Estadística Computacional
y Ciencia de Datos para la Toma de Decisiones

Índice

Introducción

APRENDIZAJE NO SUPERVISADO

- Base de datos **p variables n observaciones** $\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$

- Separar las n observaciones en **grupos de individuos homogéneos** (similares)
- Queremos agruparlos en grupos de forma que los **grupos estén constituidos por individuos semejantes siendo los grupos lo más distintos posible entre sí**.
- Normalmente **se agrupan las observaciones**, pero **el análisis puede también aplicarse para agrupar variables**.
- Estos métodos también se conocen con el nombre de métodos de **clasificación automática o no supervisada** o de reconocimiento de patrones sin supervisión.

El análisis clúster estudia tres tipos de problemas: **Partición de datos**, **construcción de jerarquías** y **clasificación de variables.**

Partición de los datos

Disponemos de datos que sospechamos son heterogéneos y se desea dividirlos en un número de grupos prefijado, de manera que:

(1) Cada elemento pertenezca a uno y solo uno de los grupos.
(2) Todo elemento quede clasificado
(3) Cada grupo sea internamente homogéneo

Construcción de jerarquías

Deseamos estructurar los **elementos de un conjunto de forma jerárquica** por su similitud.

Una clasificación jerárquica implica que **los datos se ordenan en niveles**, de manera que **los niveles superiores contienen a los inferiores**.

Este tipo de clasificación es muy frecuente en **biología**, al clasificar animales, plantas etc. Estrictamente, **estos métodos no definen grupos, sino la estructura de asociación** en cadena que pueda existir entre los elementos. Sin embargo, como veremos, **la jerarquía construida permite obtener** también una partición de los datos en grupos.

Clasificación de variables

En problemas con muchas variables es interesante hacer un estudio exploratorio inicial para dividir las variables en grupos. Este estudio puede orientarnos para plantear los modelos formales para reducir la dimensión. **Las variables pueden clasificarse en grupos o estructurarse en una jerarquía**.

Los métodos de partición utilizan la matriz de datos, pero los algoritmos jerárquicos utilizan la matriz de distancias o similitudes entre elementos. **Para agrupar variables se parte de la matriz de relación entre variables**.

Agrupamiento por k-medias

Agrupamiento por k-medias

Este tipo de **algoritmo de aprendizaje no supervisado** es útil para **explorar, describir y resumir datos**. Utilizar este agrupamiento de datos nos puede servir para **confirmar (o rechazar) algún tipo de clasificación previa**. También nos **puede ayudar a descubrir patrones y relaciones** que desconocíamos.

Por ejemplo, podemos aplicar K-means en:

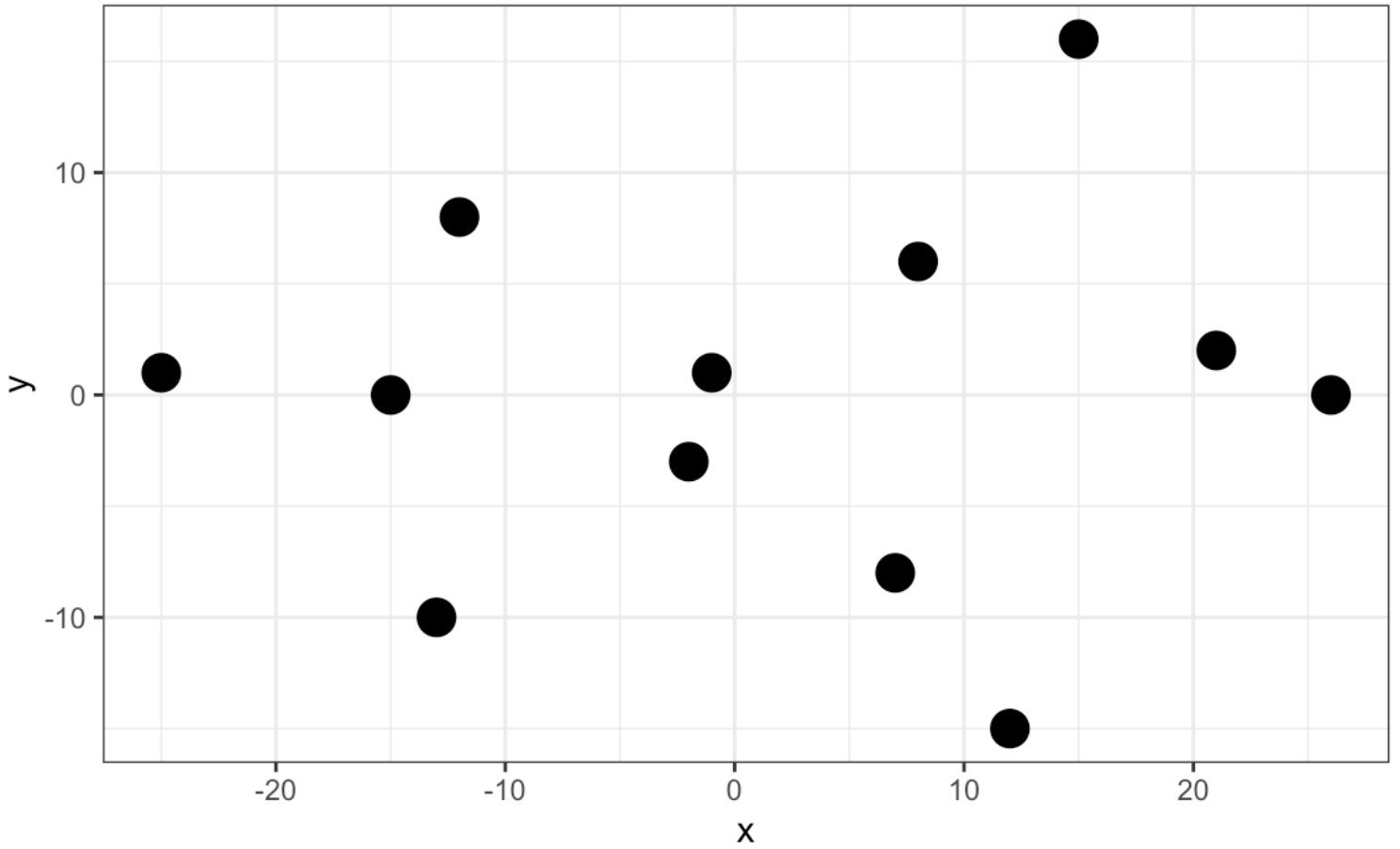
- Segmentación de clientes
- Agrupación de textos que hablan de temas similares
- Geoestadística
- Comunidades de redes sociales
- …

Ejemplo:

Agrupamiento con k = 2

Supongamos que estamos analizando la ubicación de tiendas de una cadena de supermercados y queremos agruparlas en dos regiones para optimizar la logística y el transporte.

Hemos obtenido las coordenadas geográficas aproximadas de cada tienda y las representamos como puntos negros en el plano.

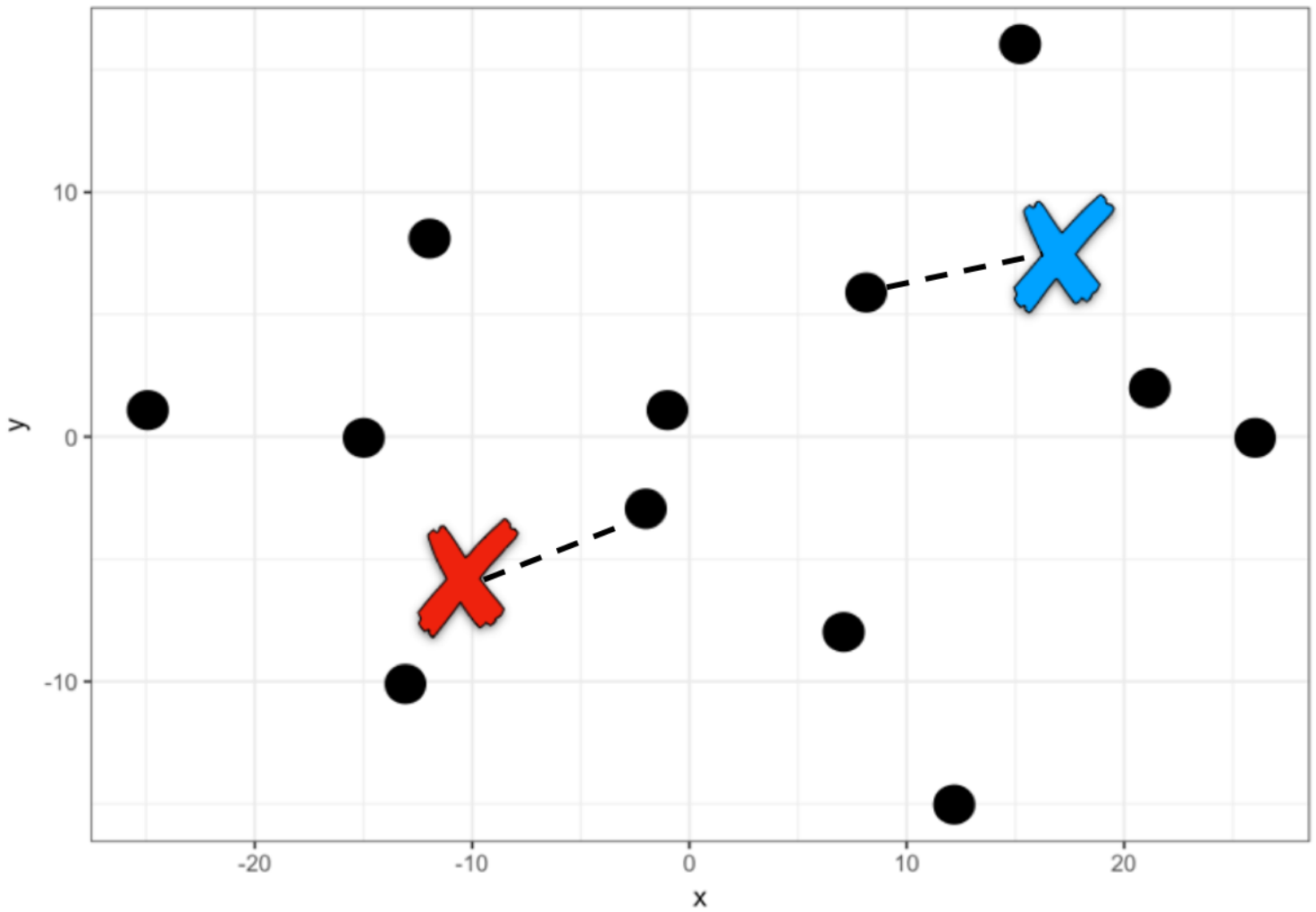


Agrupamiento por k-medias

Este método nos permite agrupar a partir de la definición de centroides.
Definiremos tantos centroides como grupos queremos obtener.
Sabemos que deben de haber dos regiones utilizaremos 2 centroidoes (k = 2).

El algoritmo k-medias coloca entonces en una primera iteración estos 2 puntos (centroides) de forma aleatoria en el plano.

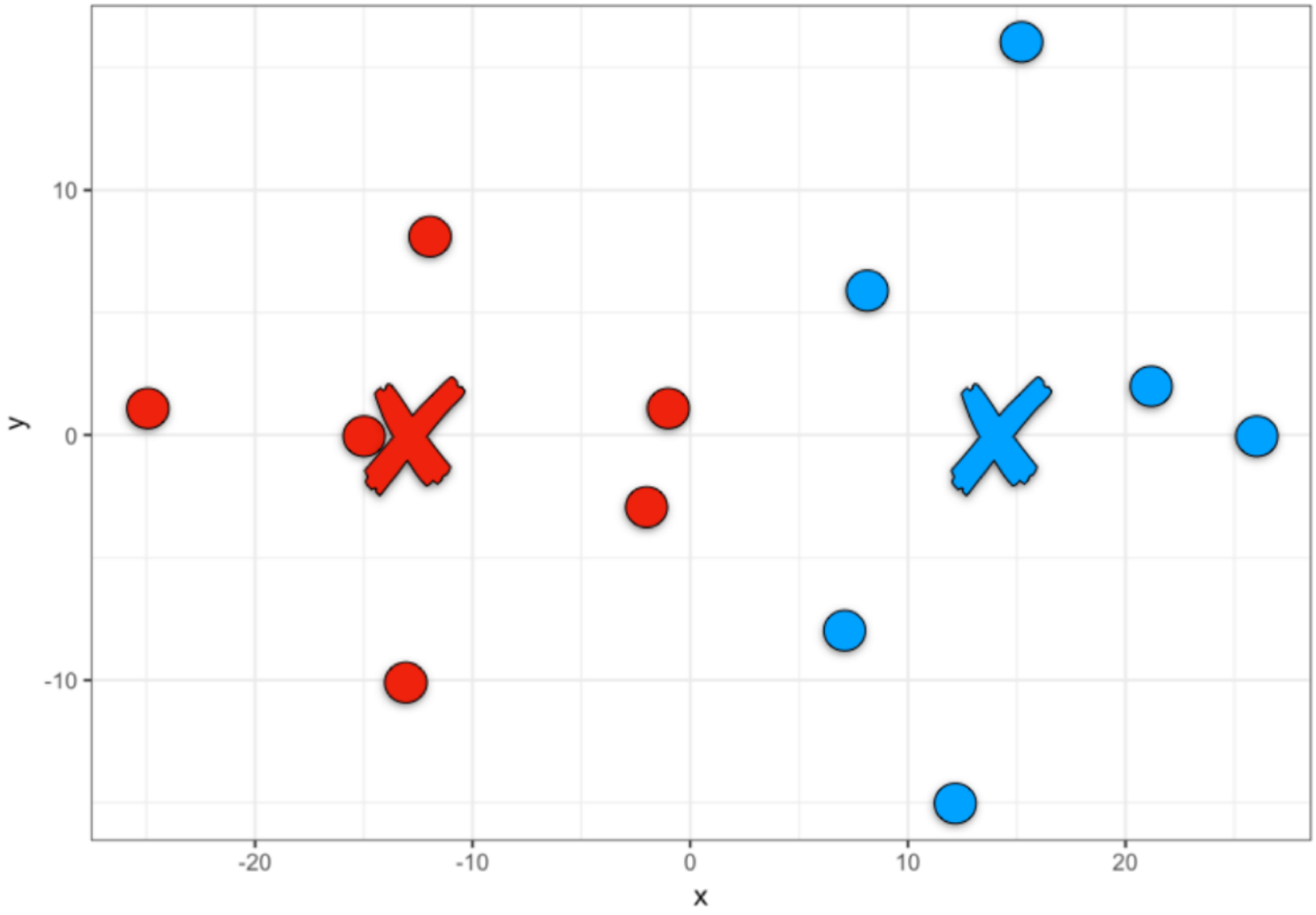
Luego, calcula la distancia entre cada centro y los puntos. Si está más cercano al centroide 1 entonces lo asigna al Clúster 1, sino al Clúster 2.



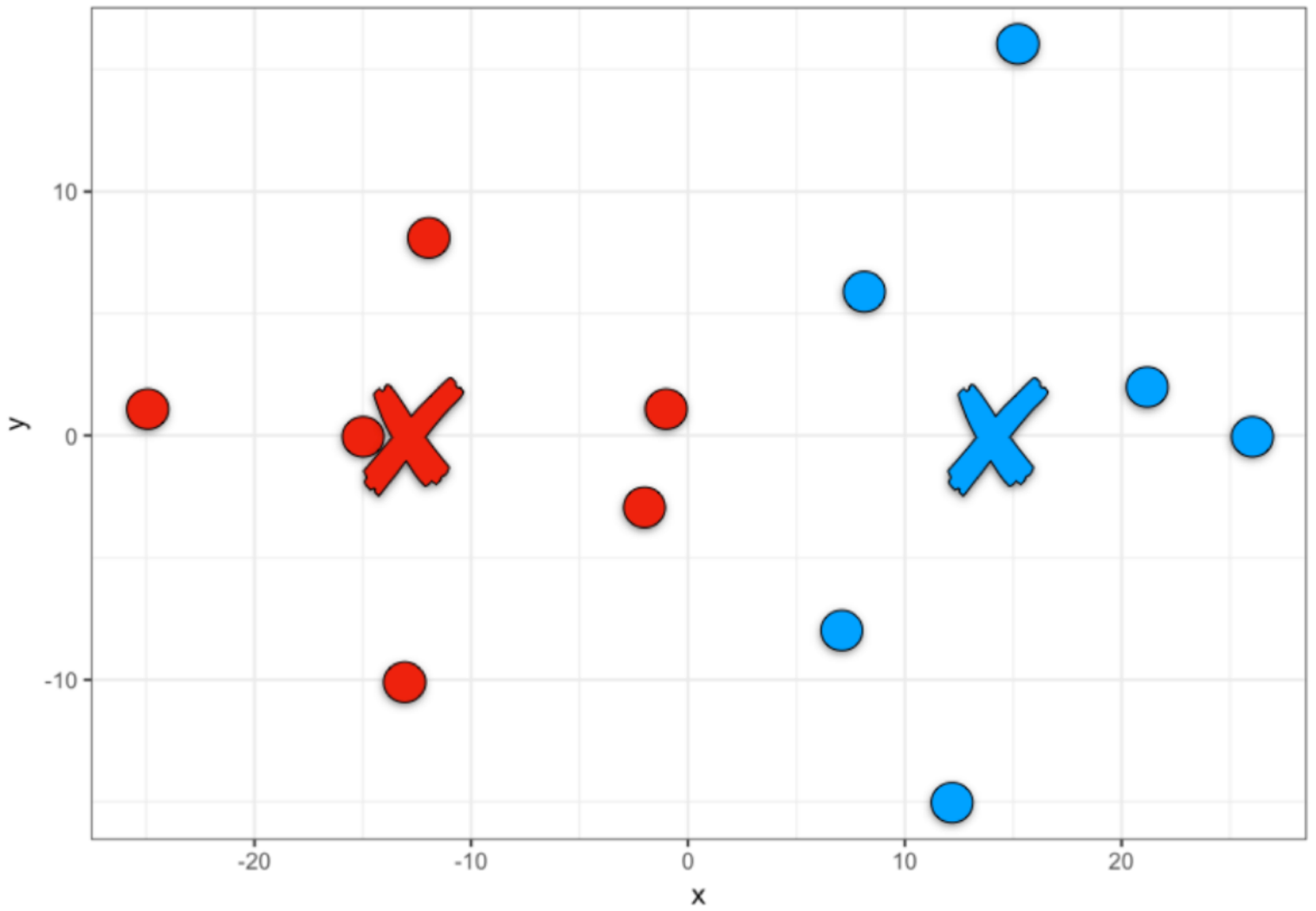
Agrupamiento por k-medias

Ya se ha realizado una primera agrupación. Ahora cada centroide dentro de cada grupo se ubica en la media de los demás puntos de su grupo y se da otra iteración para volver a asignar a todos los puntos.

Esta iteración se hace una y otra vez hasta que los centroides quedan fijos (o con una diferencia muy pequeña entre iteraciones).



Agrupamiento por k-medias



Caso general

El objetivo es dividir esta muestra en un número de grupos prefijado, G. El algoritmo de k-medias requiere las cuatro etapas siguientes :

(1) Seleccionar G puntos como centros de los grupos iniciales. Esto puede hacerse:

a) asignando **aleatoriamente** los centros de los G grupos;
b) **tomando como centros los G puntos más alejados** entre sí
c) construyendo los grupos con información a priori, o bien seleccionando los centros a priori.

(2) Calcular las distancias euclídeas de cada elemento al centro de los G grupos, y **asignar cada elemento al grupo** más próximo. La asignación se realiza secuencialmente y al introducir un nuevo elemento en un grupo se recalculan las coordenadas de la nueva media de grupo.

(3) Definir un criterio de optimalidad y comprobar si reasignando uno a uno cada elemento de un grupo a otro mejora el criterio.

(4) Si no es posible mejorar el criterio de optimalidad, terminar el proceso.

Implementación del algoritmo

El criterio de homogeneidad que se utiliza en el algoritmo de k-medias es la suma de cuadrados dentro de los grupos (SCDG) para todas las variables, que es equivalente a la suma ponderada de las varianzas de las variables en los grupos:

$$SCDG = \sum_{g=1}^{G} \sum_{j=1}^{p} \sum_{i=1}^{n_g} (x_{ijg} - \overline{x}_{jg})^2$$

El criterio se escribe

$$\min SCDG = \min \sum_{g=1}^{G} \sum_{j=1}^{p} n_g s_{jg}^2$$

Este criterio requeriría calcularlo para todas las posibles particiones, labor claramente imposible, salvo para valores de n muy pequeños.

El algoritmo de k–medias busca la partición óptima con la restricción de que en cada iteración sólo se permite mover un elemento de un grupo a otro.

Algoritmo k-medias:

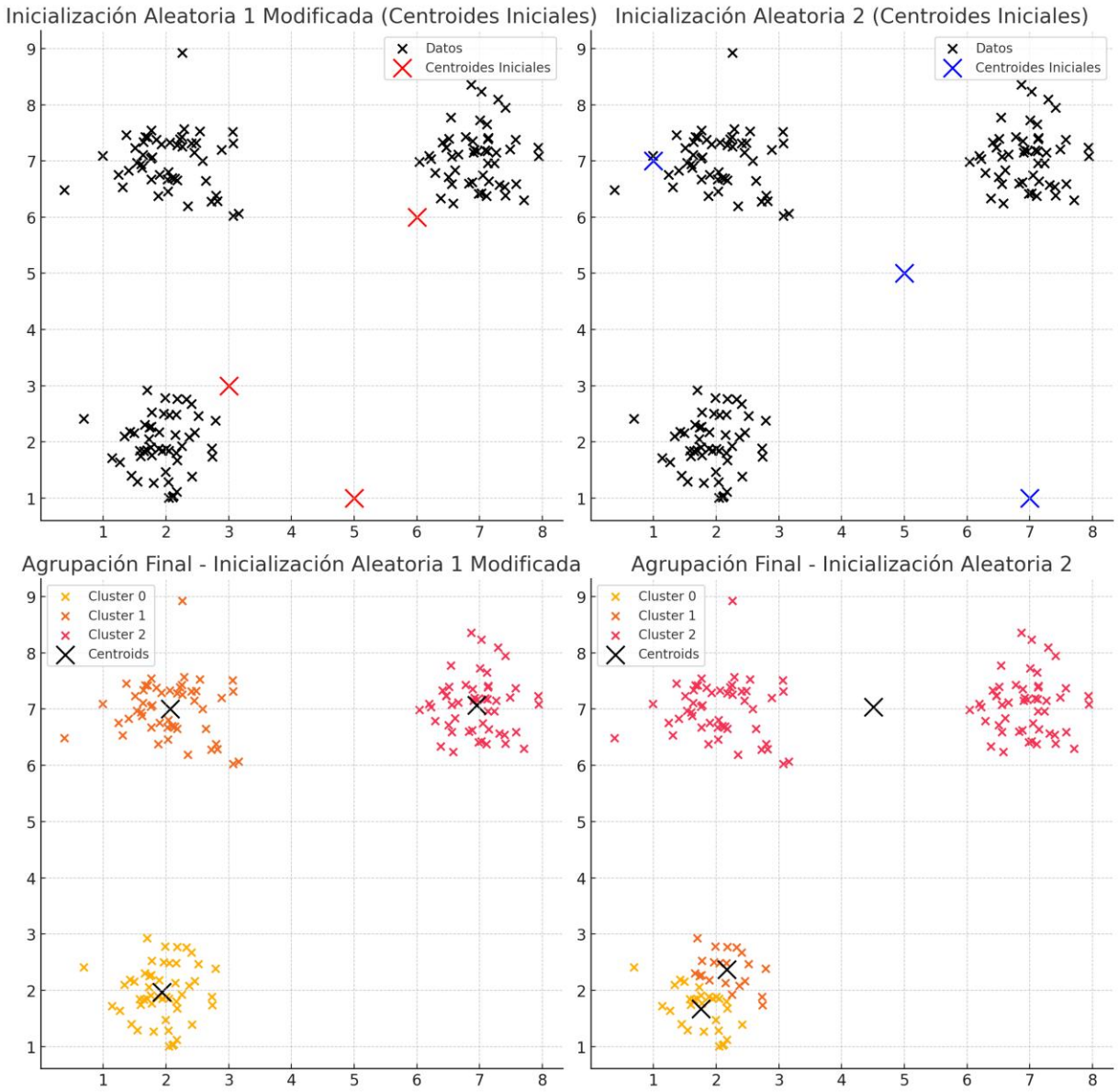
(1) Partir de una asignación inicial

(2) Comprobar si moviendo algún elemento se reduce *SCDG*

(3) Si es posible reducir SCDG al mover el elemento, recalcular las medias de los dos grupos afectados por el cambio y volver a (2). Si no es posible reducir *SCDG* terminar.

En consecuencia**, el resultado del algoritmo puede depender de la asignación inicial** y del orden de los elementos. Conviene siempre repetir el algoritmo desde distintos valores iniciales y permutando los elemento de la muestra.

Agrupamiento por k-medias



Número de clústers:

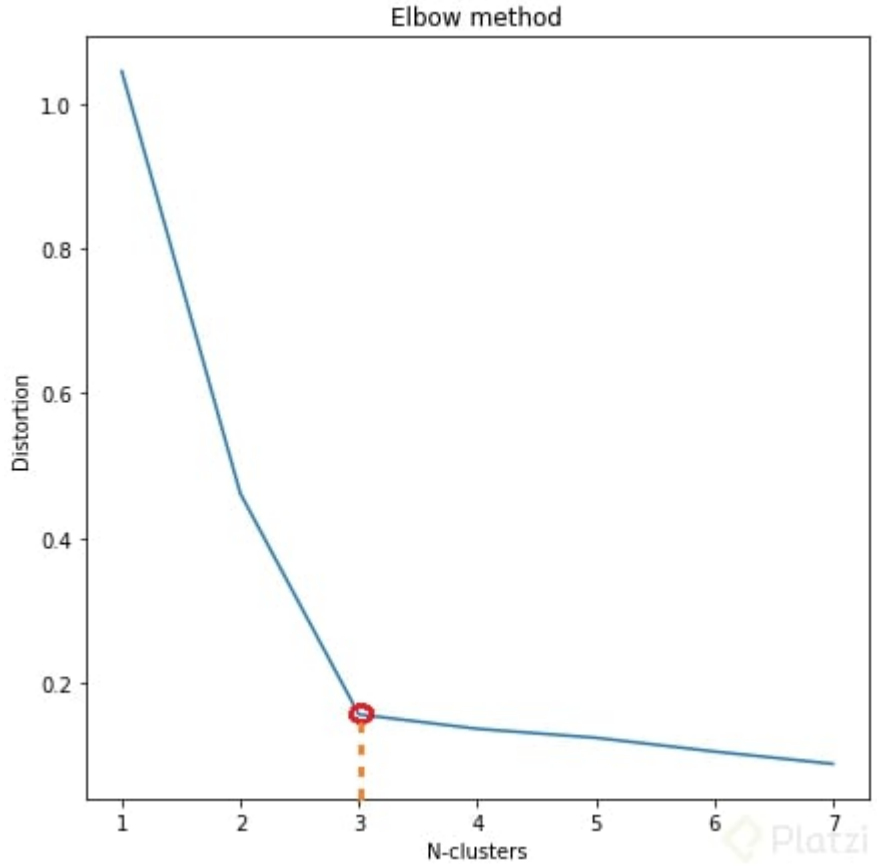
No existe un criterio objetivo ni ampliamente válido para la elección de un número óptimo de Clusters; pero tenemos que tener en cuenta, que una mala elección de los mismos puede dar lugar a realizar agrupaciones de datos muy heterogéneos (pocos Clusters); o datos, que siendo muy similares unos a otros los agrupemos en Clusters diferentes (muchos Clusters).

Vamos a describir tres de los métodos que más se utilizan:

- Método del codo
- Método de la silueta

Método del codo (Elbow method):

Calcular la distorsión promedia de los clústers, que es la distancia promedia del centroide a todos los puntos del clúster. Así, cuando se va de una situación en la que el número de clústers es inferior al correcto a una situación en la que el número es el adecuado, el valor de la dispersión disminuye bruscamente, mientras que si aumenta el número de clústers al adecuado, el valor de la dispersión se reducirá más lentamente, formando un codo en la gráfica.



Método de la silueta promedio:

El coeficiente de la silueta para x está dado por:

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

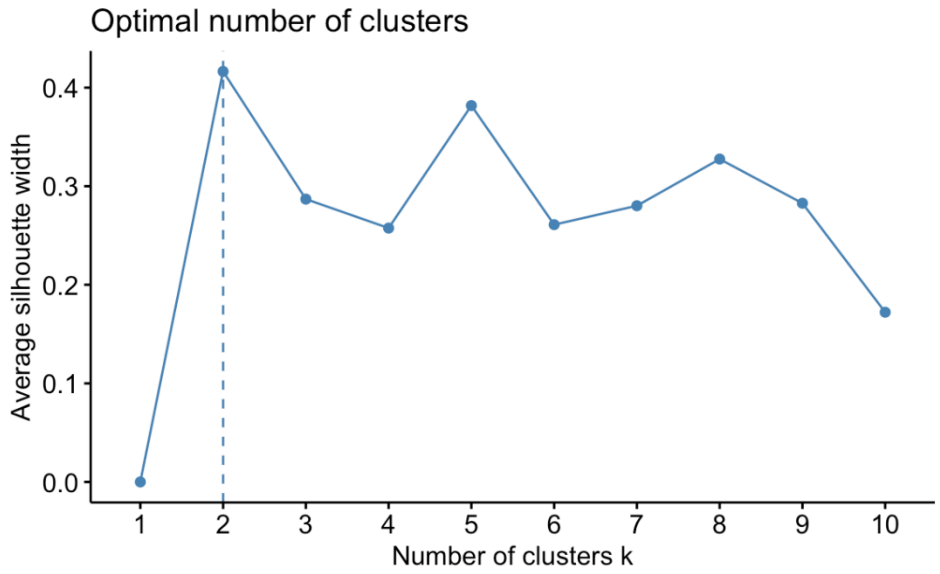
a(x)= distancia promedio de x a todos los demás puntos en el mismo clúster.
b(x)= distancia promedio de x a todos los demás puntos en el clúster más cercano.

s(x) puede variar entre -1 y 1

- Valores cercanos a -1 indica un agrupamiento incorrecto.
- Valores cercanos a +1 indica un agrupamiento correcto.

El número de clúster óptimo es el k con mayor coeficiente silueta promedio.

$$SC = \frac{1}{n}\sum_{i=1}^{n} s(x)$$



Ventajas y desventajas de este método:

K-medias es uno de los métodos de clustering más utilizados. Destaca por la sencillez y velocidad de su algoritmo, sin embargo, presenta una serie de limitaciones que se deben tener en cuenta.

- Requiere que se indique de antemano el número de clústers que se van a crear.
- Las agrupaciones resultantes pueden variar dependiendo de la asignación aleatoria inicial de los centroides.
- Presenta problemas de robustez frente a outliers.