UNIVERSIDAD EAFIT

Científico de Datos

Aliados:

Microsoft

idata
Advanced analytics for business

Vigilada Mineducación

UNIVERSIDAD EAFIT

APRENDIZAJE NO SUPERVISADO

Aliados:

Microsoft

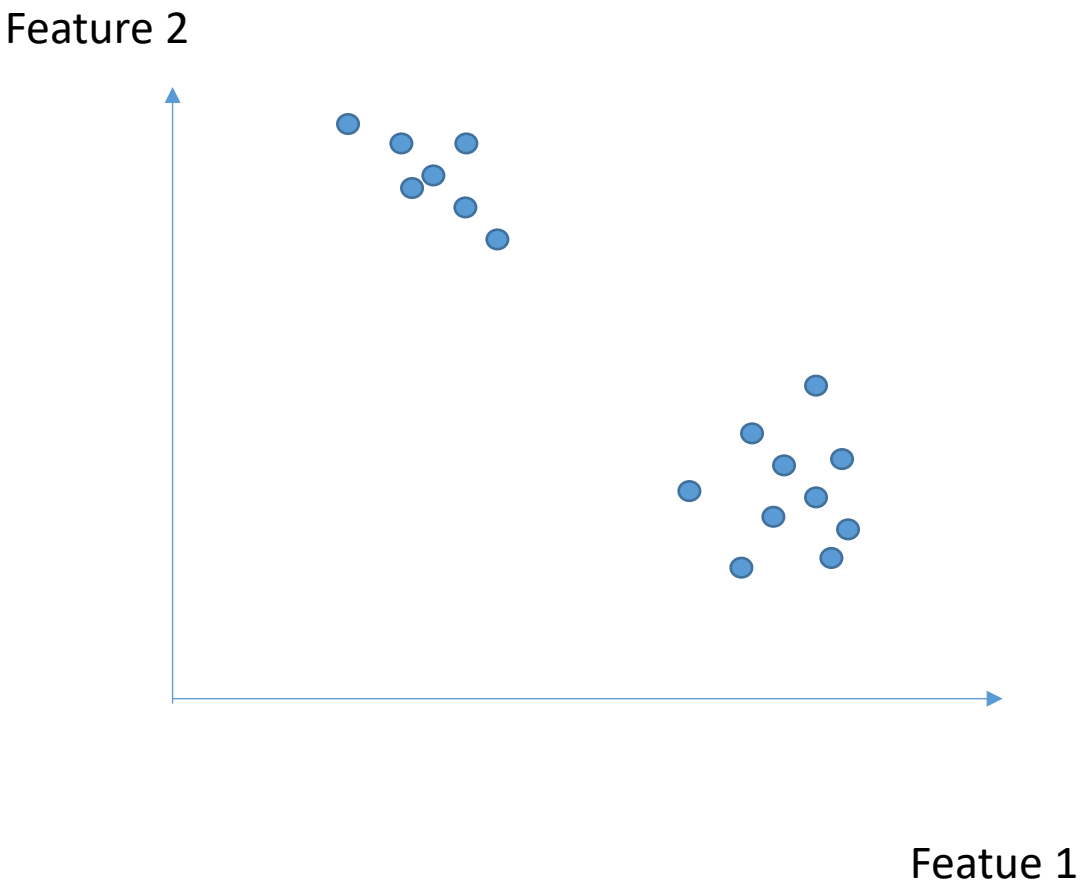
idata
Advanced analytics for business

Vigilada Mineducación

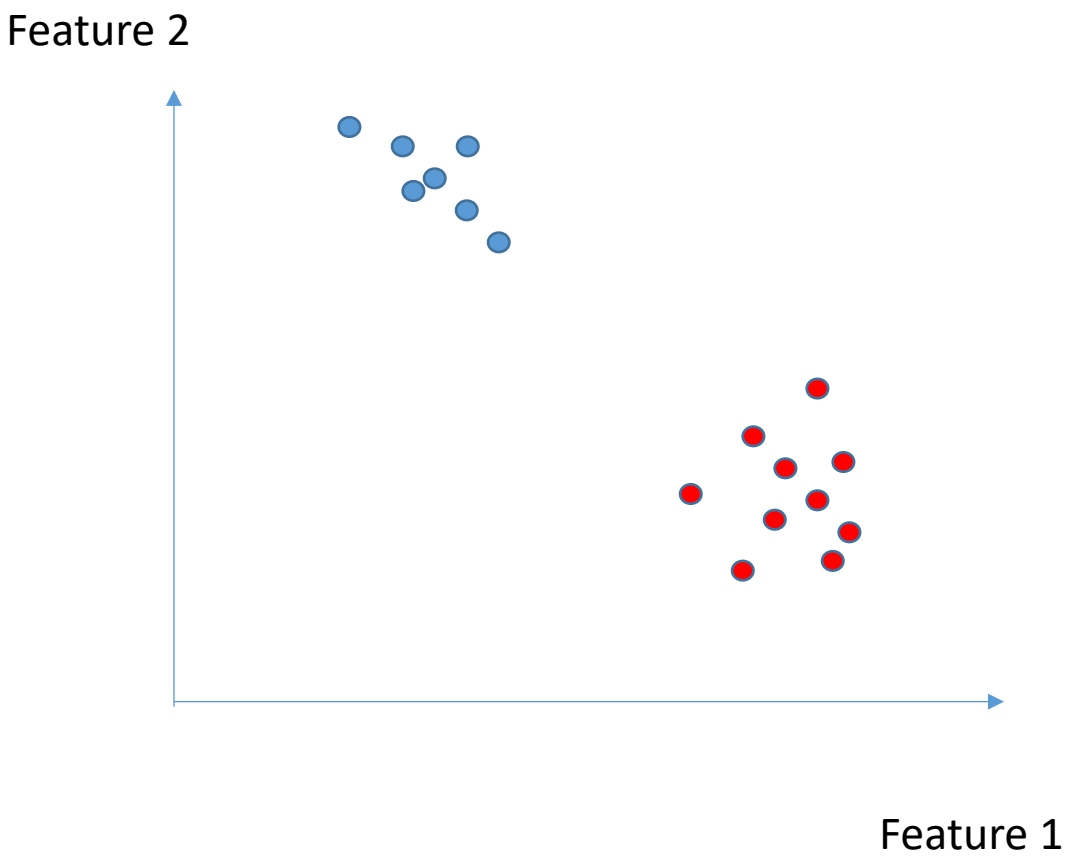
DEFINTION

- Encontrar etiquetas (labels) a objetos sin etiquetas
- Es el proceso de particionar un conjunto de objetos en subconjuntos.

DEFINTION

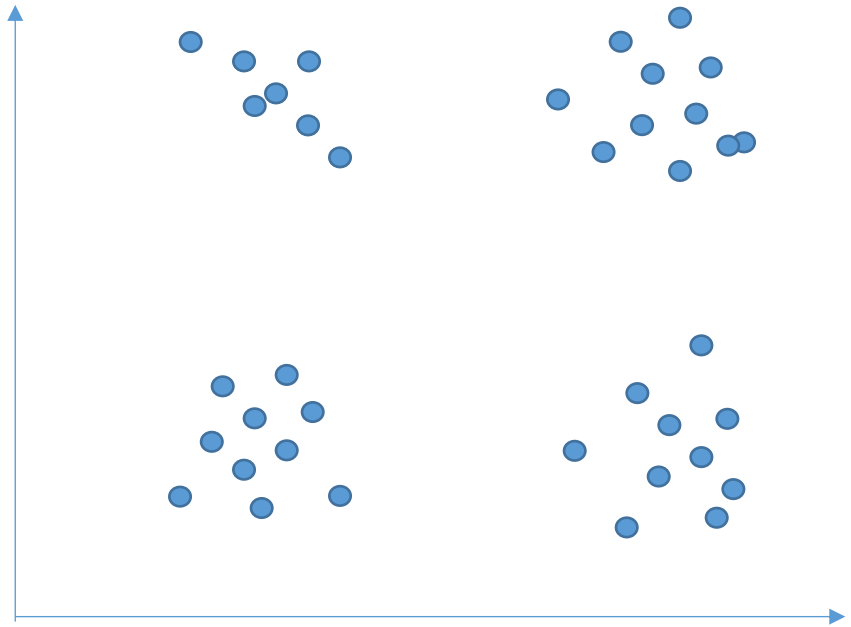


DEFINTION



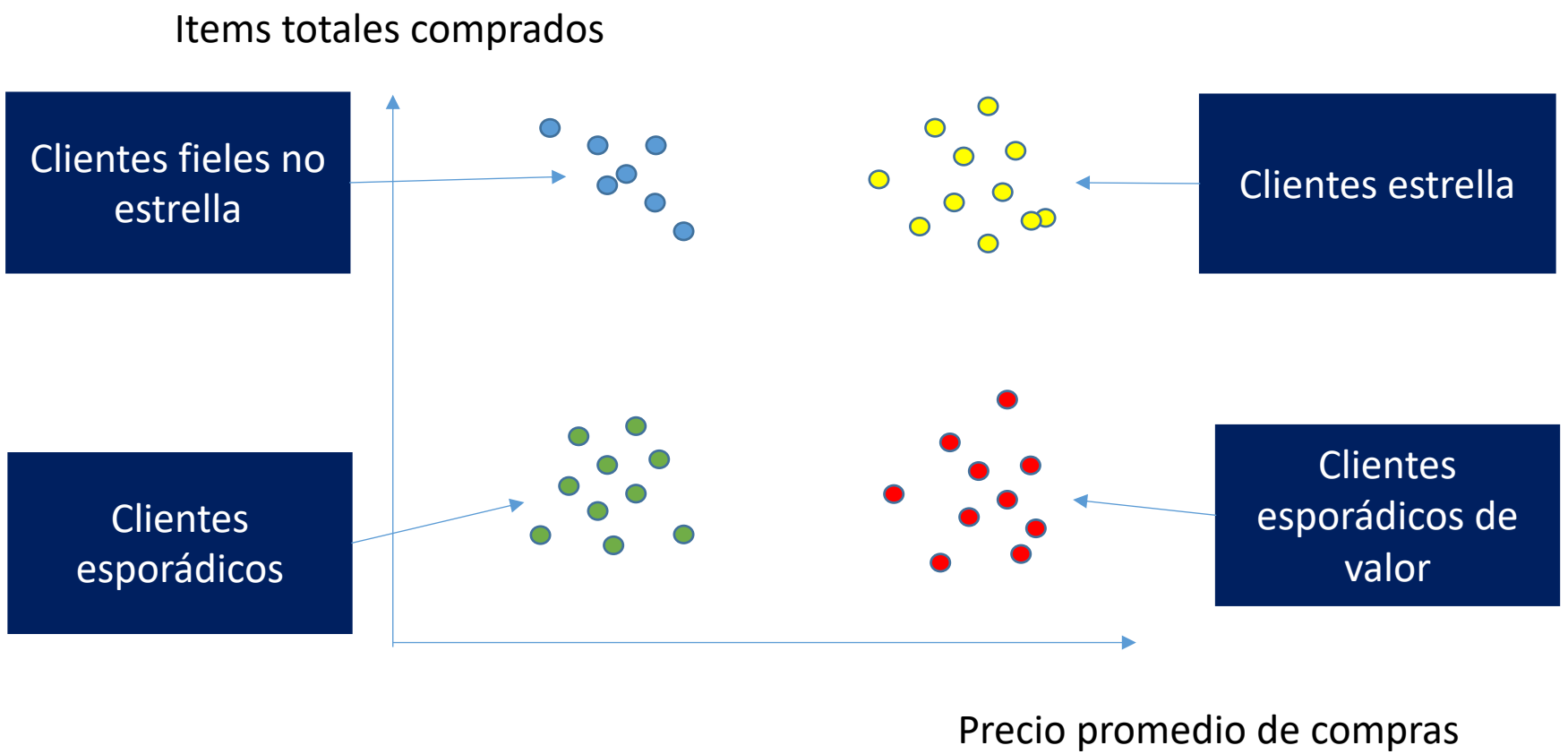
DEFINTION

Items totales comprados



Precio promedio de compras

DEFINTION


Items totales comprados
Clientes fieles no estrella
Clientes estrella
Clientes esporádicos
Clientes esporádicos de valor
Precio promedio de compras

DEFINTION


$\{x_1, x_2, x_3, \dots, x_n\}$
Algoritmo de clustering
$\{C_1, C_2, C_1, \dots, C_k\}$
$C_1 = \{x_1, x_3, ..\}$
$C_2 = \{x_2, ..\}$
$C_k = \{x_n, ..\}$

DEFINTION

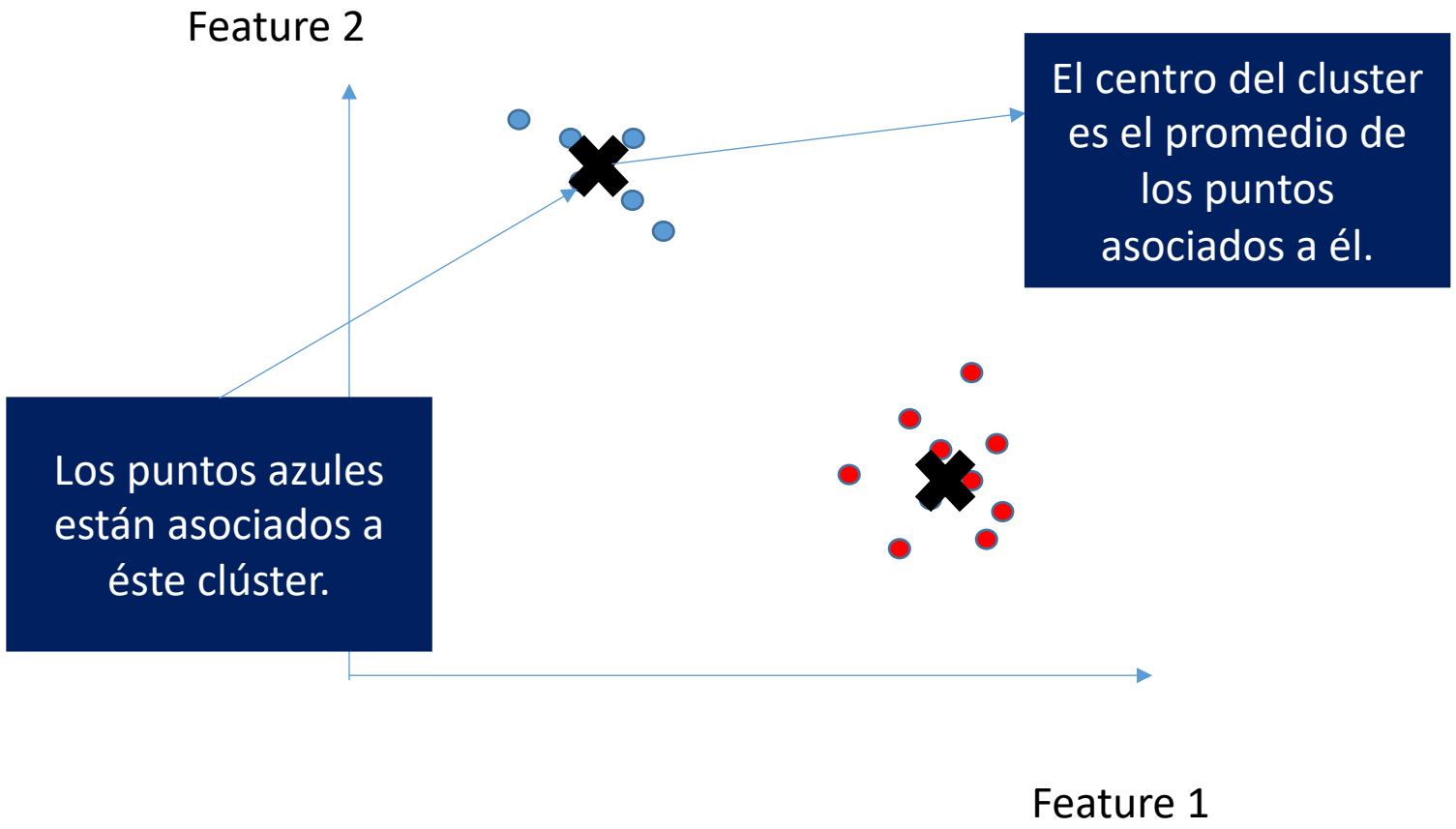
Method	General Characteristics
Partitioning methods	– Find mutually exclusive clusters of spherical shape – Distance-based – May use mean or medoid (etc.) to represent cluster center – Effective for small- to medium-size data sets
Hierarchical methods	– Clustering is a hierarchical decomposition (i.e., multiple levels) – Cannot correct erroneous merges or splits – May incorporate other techniques like microclustering or consider object "linkages"
Density-based methods	– Can find arbitrarily shaped clusters – Clusters are dense regions of objects in space that are separated by low-density regions – Cluster density: Each point must have a minimum number of points within its "neighborhood" – May filter out outliers

K- MEANS

- Método de clustering duro (hard) porque cada dato pertenence sólo a un cluster dado.

- Dos conceptos: **centroides y puntos**.

K-MEANS


Feature 2
El centro del cluster es el promedio de los puntos asociados a él.
Los puntos azules están asociados a éste clúster.
Feature 1

K-MEANS

Inicializar centros $k = \{\overrightarrow{c_1}, \overrightarrow{c_2}, \ldots, \overrightarrow{c_k}\}$

1. Para cada centro identificamos qué datos tiene asociados.
2. Calculamos el nuevo valor del centro como la media de todos los datos.
3. Evaluamos criterio de convergencia. Si no se cumple, volver al paso 1.

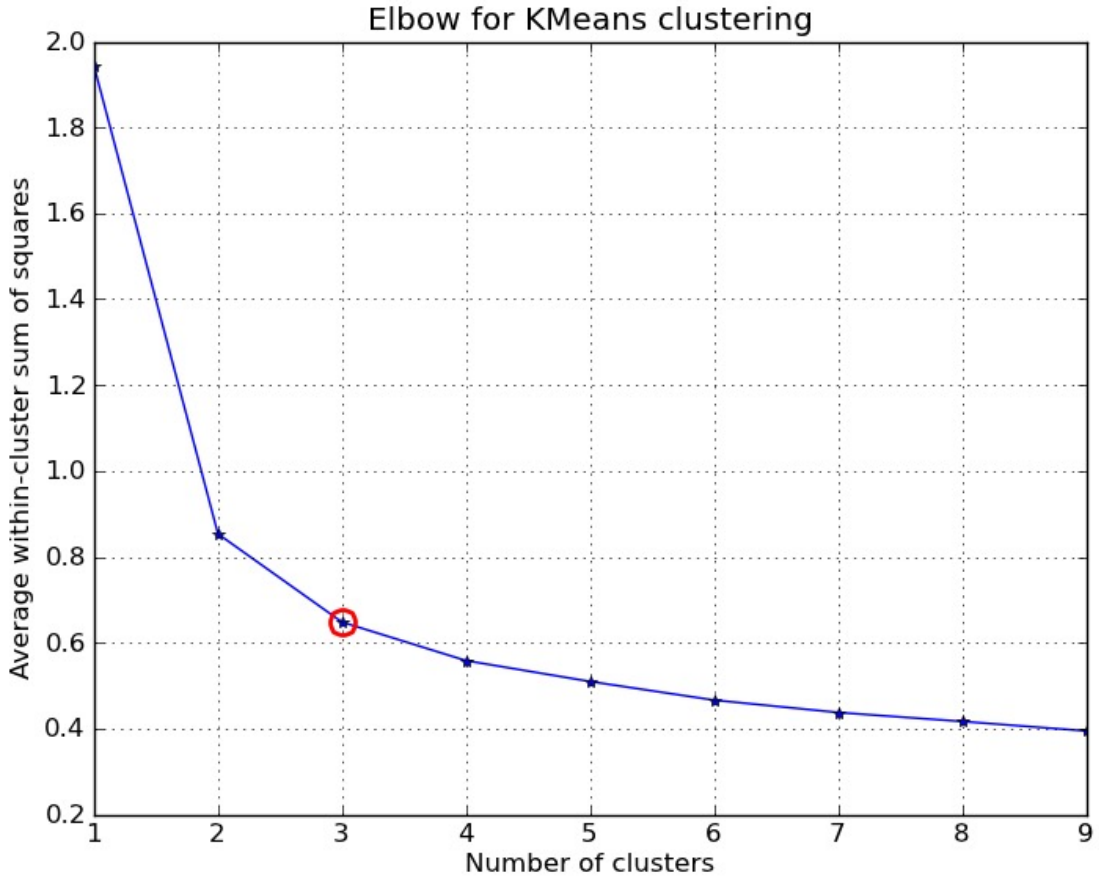
CONSIDERACIONES

- Asume clusters circulares
- Se debe determinar K: hay algunas heurísticas para elegir K.
- Se deben inicializar centros. Esto hace que pueda caer en mínimos locales. Se aconseja inicializar varias veces.
- Sensible a outliers.
- Problemas con muchas dimensiones.
- ¡Converge!

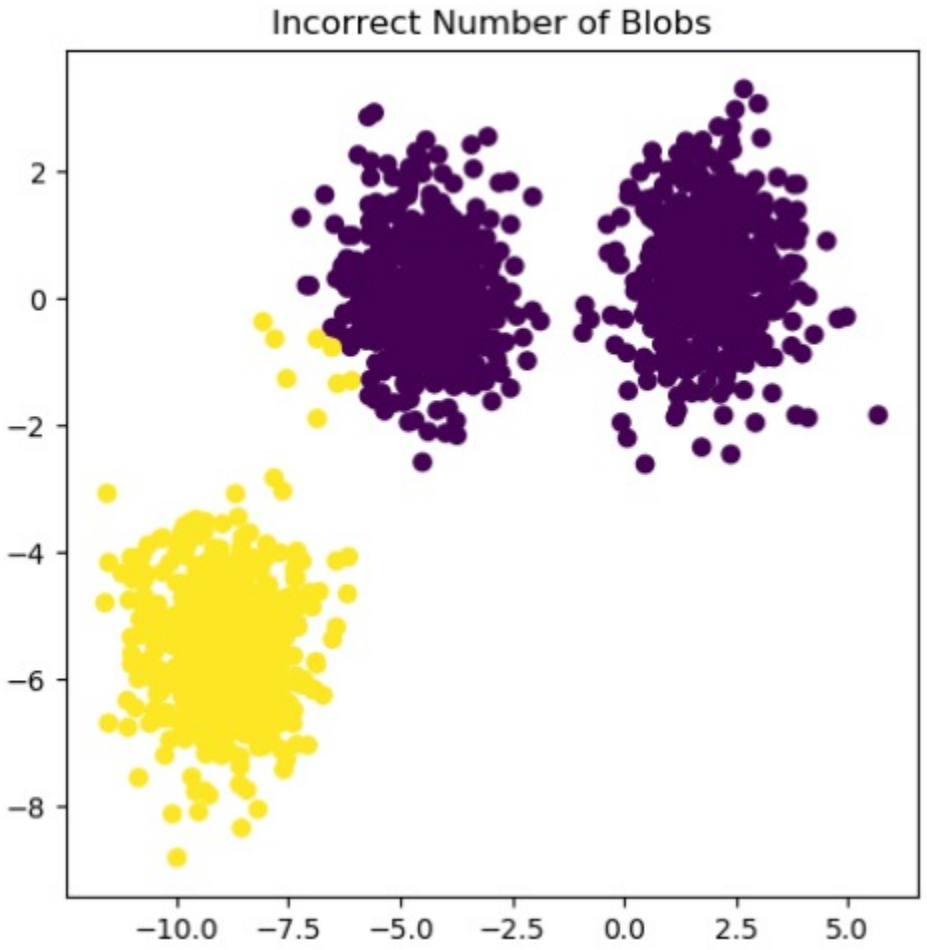
CÓMO ELEGIR EL NÚMERO DE CLUSTERS (K)

- Heurística: Criterio del codo (elbow criterion).
- Correr el algoritmo K-Means de 1 hasta un valor máximo. Evaluar la calidad del clústering.
- Seleccionar k cuando la calidad del clustering comience a disminuir poco.

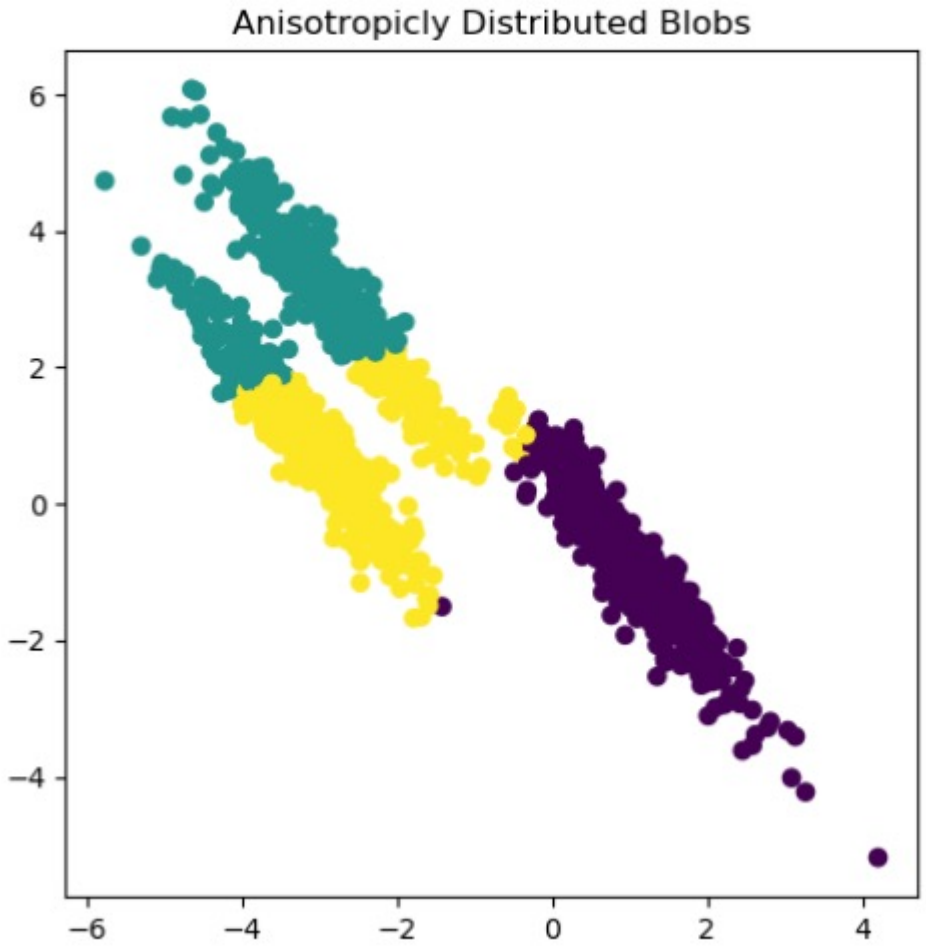
CÓMO ELEGIR EL NÚMERO DE CLUSTERS (K)



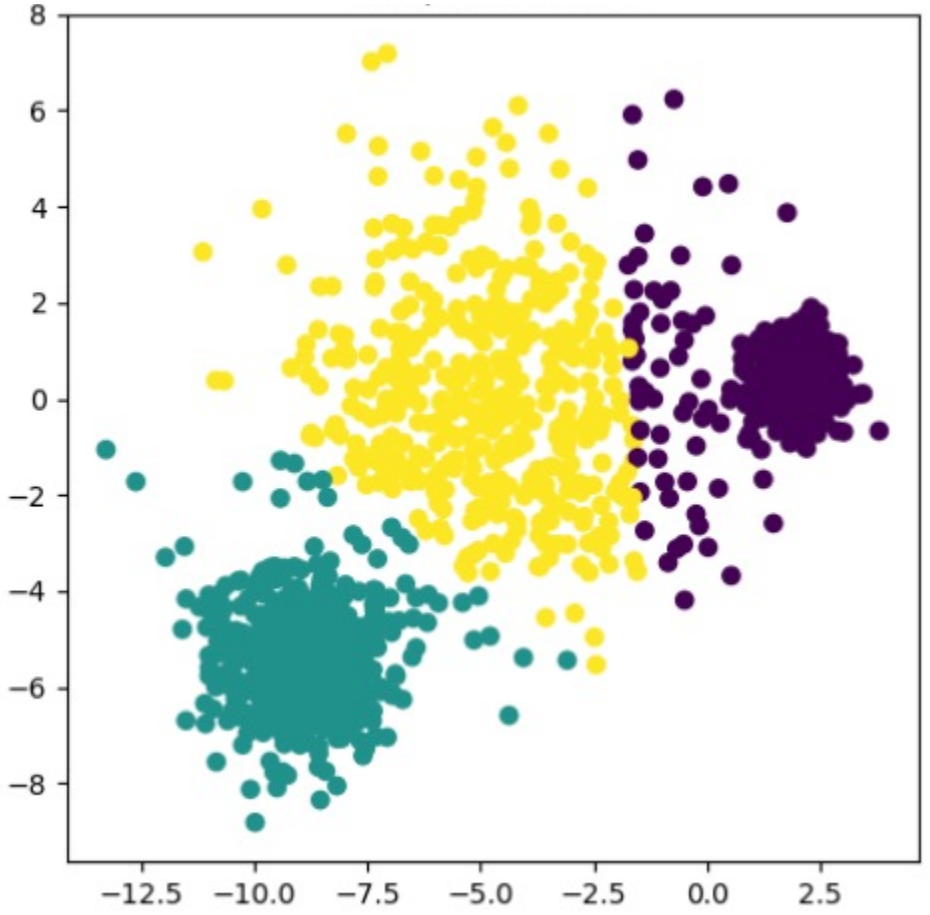
NÚMERO INCORRECTO DE CLUSTERS (K)



DISTRIBUCIÓN NO ADECUADA PARA K-MEANS



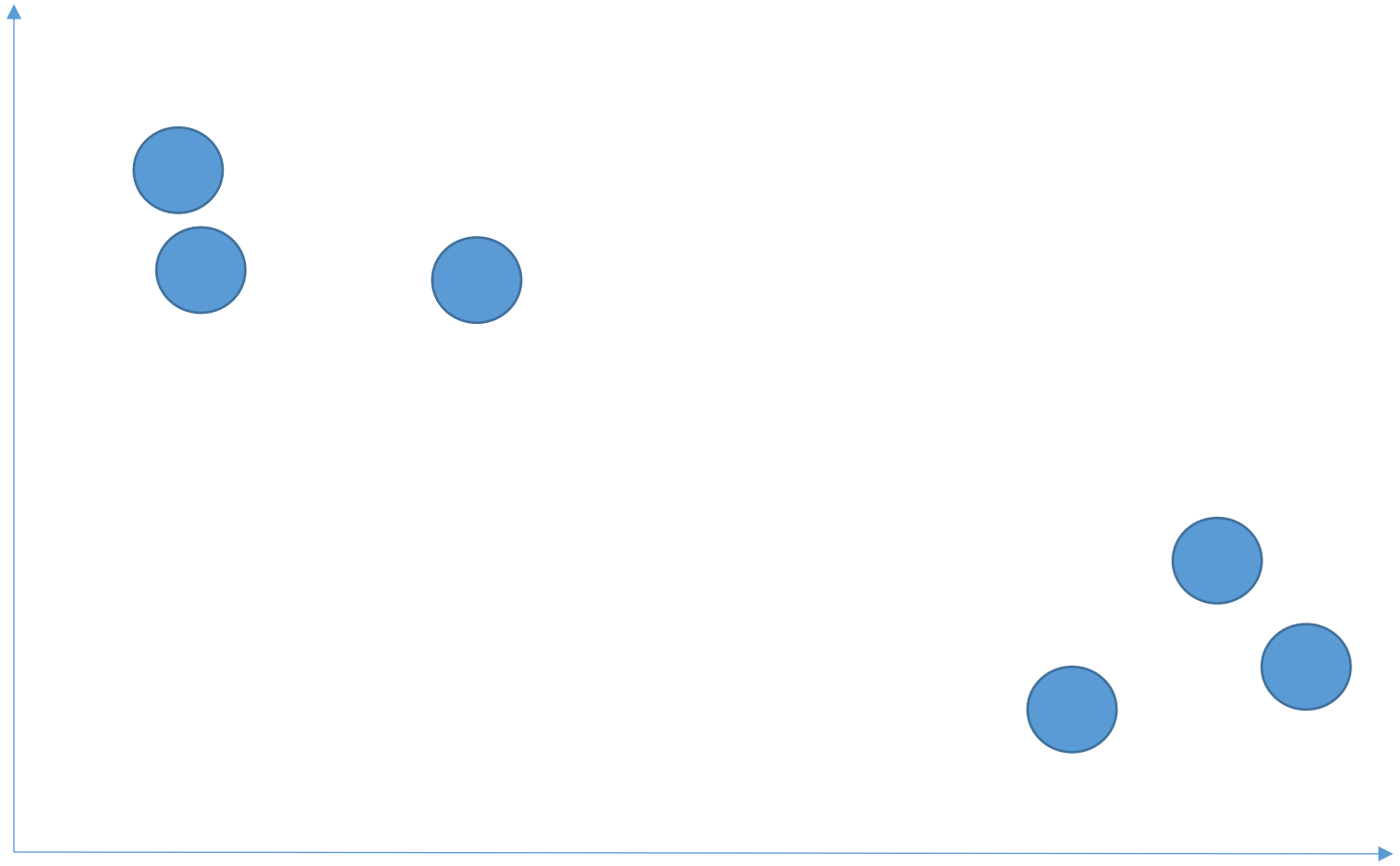
K-MEANS



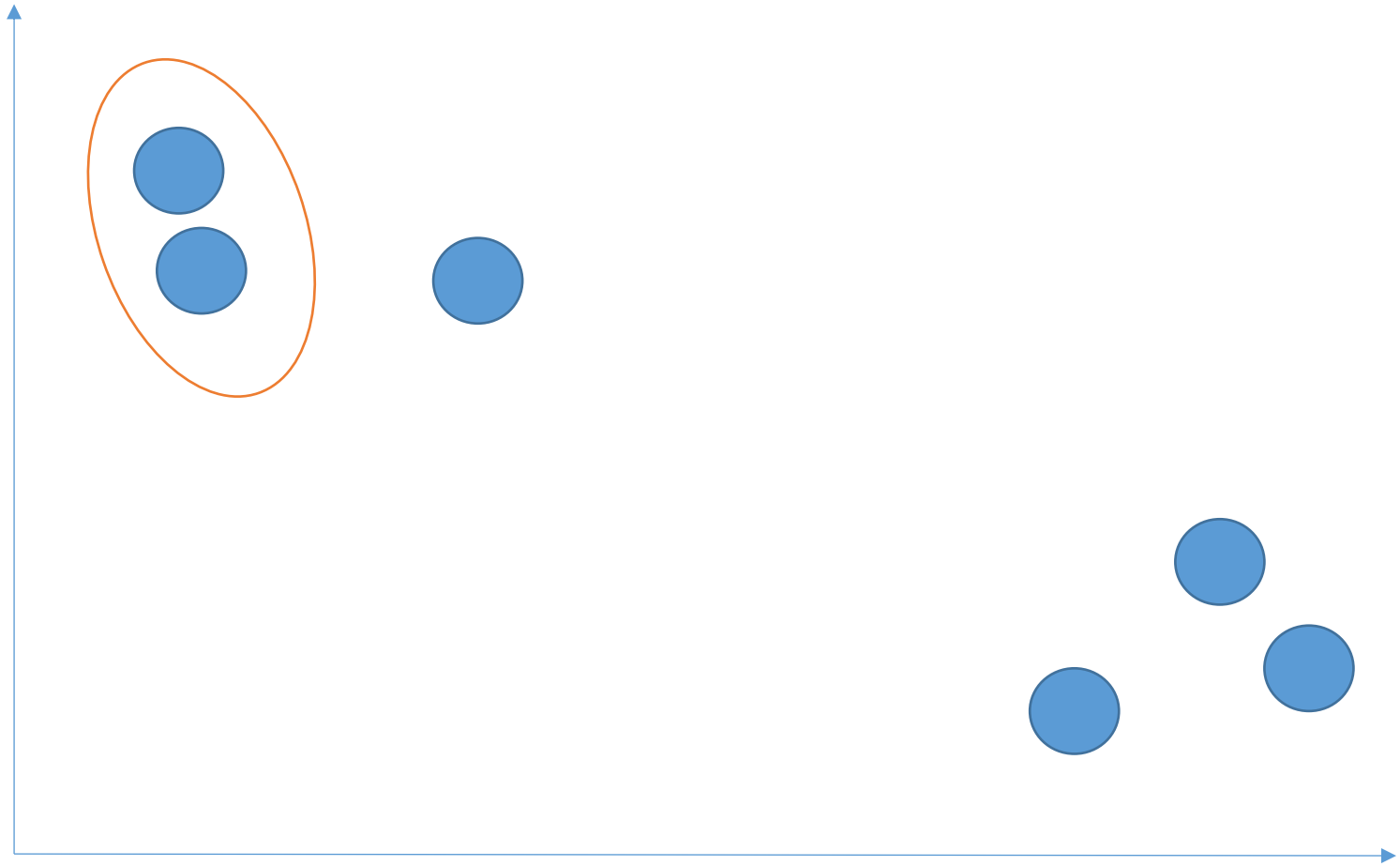
Fuzzy C-Means

- Permite tener un grado de pertenencia a cada clúster por muestra. El grado de pertenencia es un número entre 0 y 1.
- Tiene problemas similares al K-Means: se debe determinar el número de clúster y tiene problemas en altas dimensionalidades.
- Es muy conveniente cuando no se desea asignar cada muestra a un clúster específico.

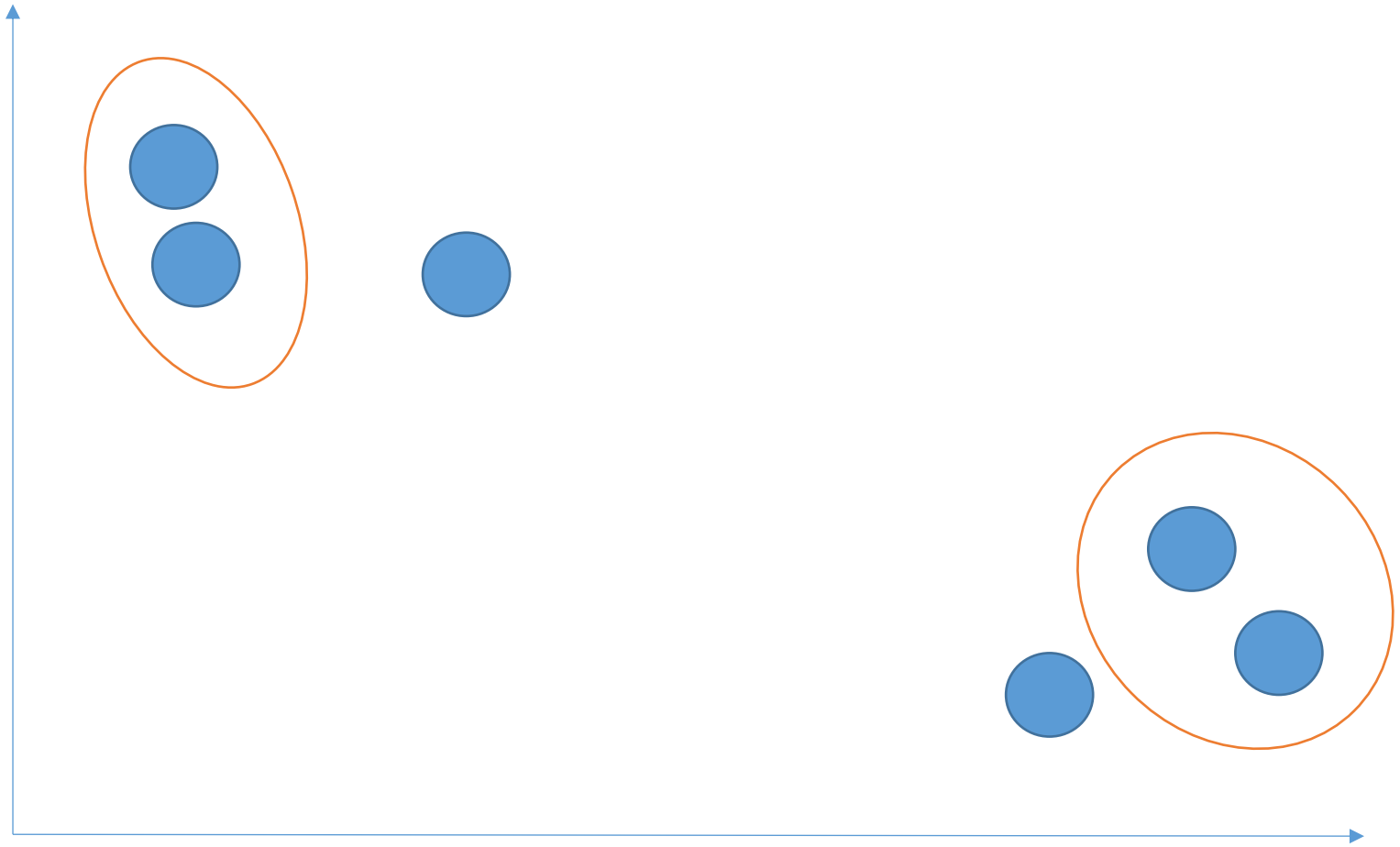
CLUSTERING JERÁRQUICO



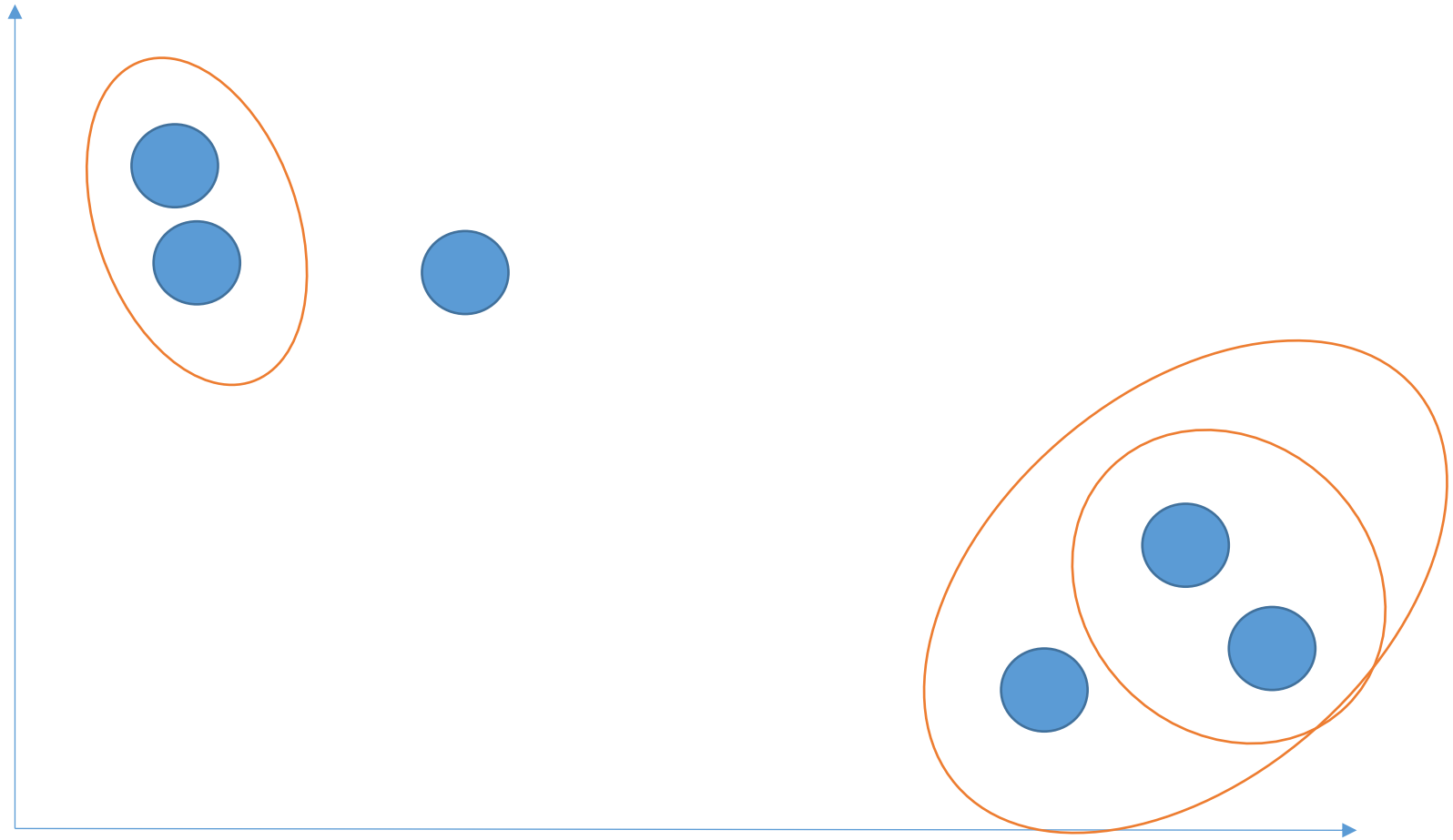
CLUSTERING JERÁRQUICO



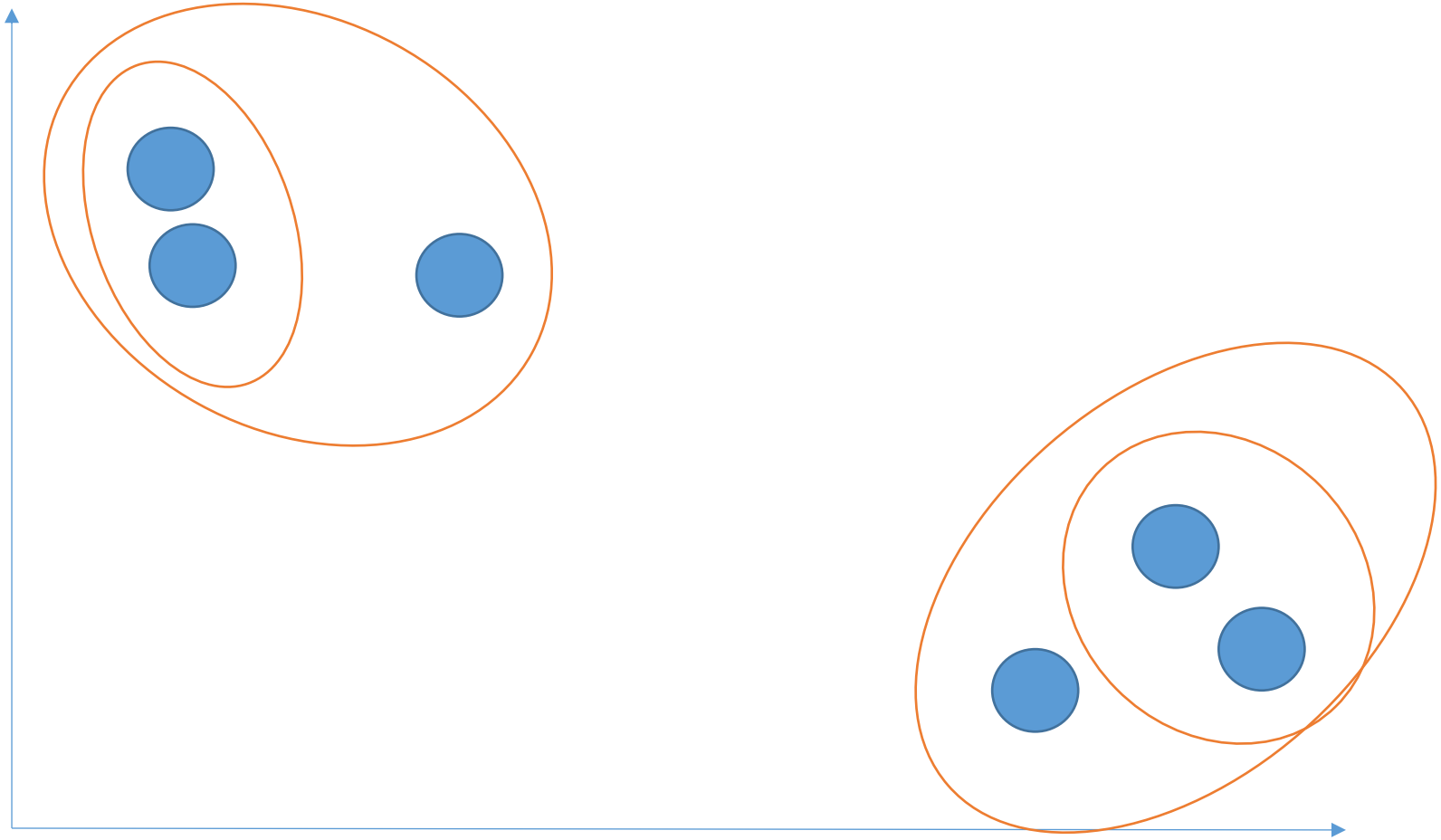
CLUSTERING JERÁRQUICO



CLUSTERING JERÁRQUICO



CLUSTERING JERÁRQUICO



CLUSTERING JERÁRQUICO

- Agrupa datos de manera consecutiva mientras construye una jerarquía.
- No hay que especificar un número de grupos deseados, pero hay que especificar un punto de corte.

CLUSTERING JERÁRQUICO

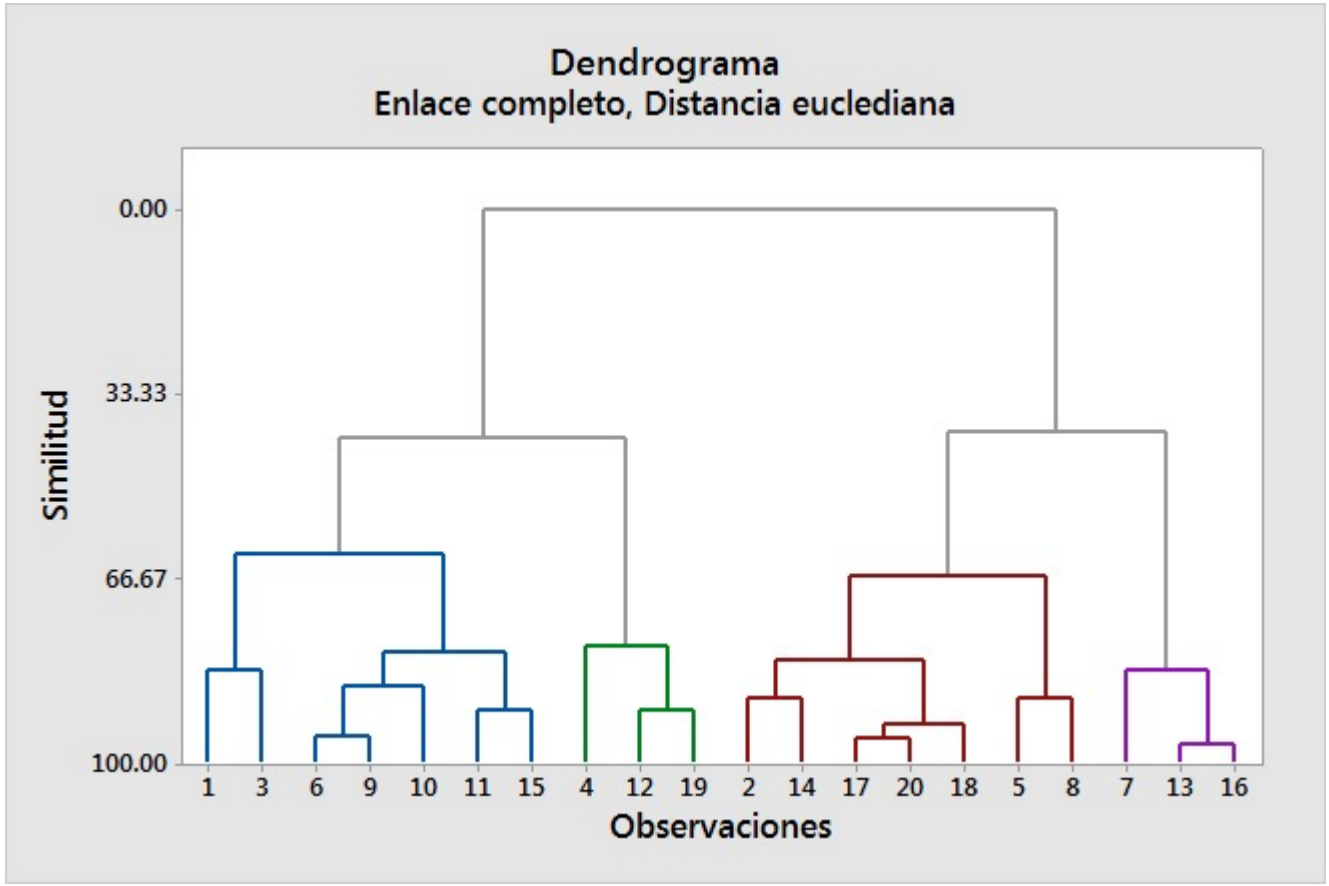
1. Empieza con clusters para cada dato: $C_n = \{\{x_1\}, \{x_2\}, \dots, \{x_n\}\}$

2. Busca la mínima distancia (u otro criterio) entre los cluster.
3. Se unen los dos clusters más cercanos en un nuevo cluster.

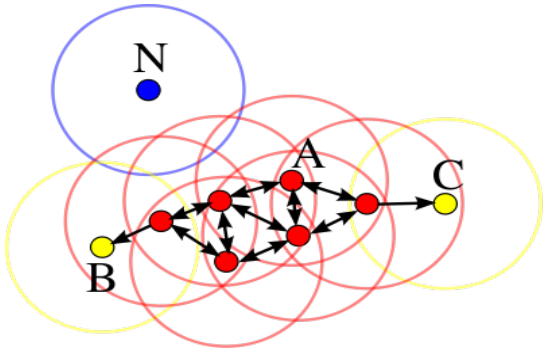
CRÍTERIOS DE ENLACE

- Distancia mínima (Single Linkage)
- Distancia máxima (Complete Linkage)
- Distancia promedio (average linkage)

CLUSTERING JERÁRQUICO



DBSCAN



- No necesita especificar el número de clusters
- Permite encontrar clusters con formas geométricas arbitrarias
- Es robusto detectando outliers

Índice Silhoutte

- **a** es el promedio de las disimilitudes (o distancias) de la observación i con las demás observaciones del cluster al que pertenece i

- **b** es la distancia mínima a otro cluster que no es el mismo en el que está la observación i. Ese cluster es la segunda mejor opción para i y se lo denomina vecindad de i.

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

- $s(i) \approx 1$, la observación *i* está bien asignada a su *cluster*
- $s(i) \approx 0$, la observación *i* está entre dos *cluster*
- $s(i) \approx -1$, la observación *i* está mal asignada a su *cluster*

UNIVERSIDAD EAFIT

¡Gracias!

Aliados:

Microsoft

idata
Advanced analytics for business

Vigilada Mineducación