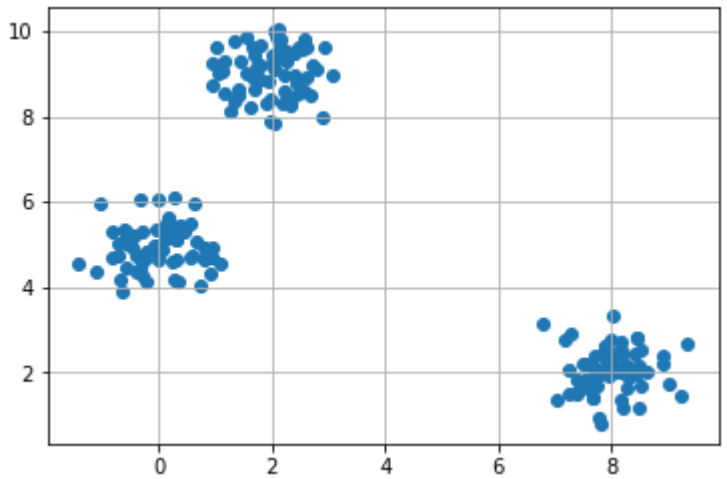
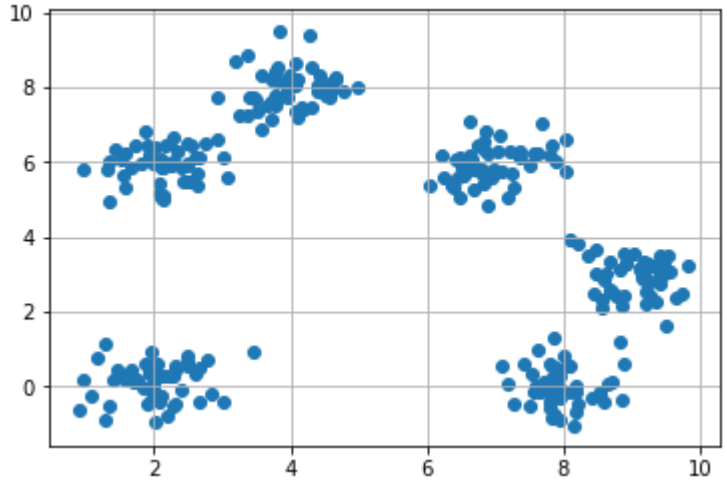
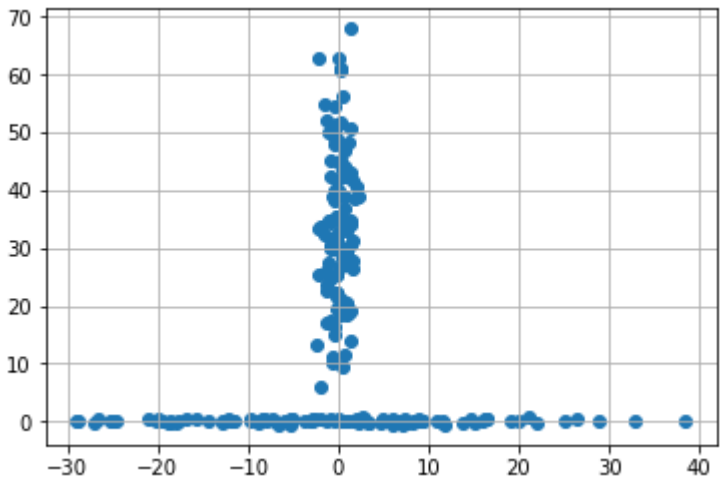
CLUSTERING JERÁRQUICO

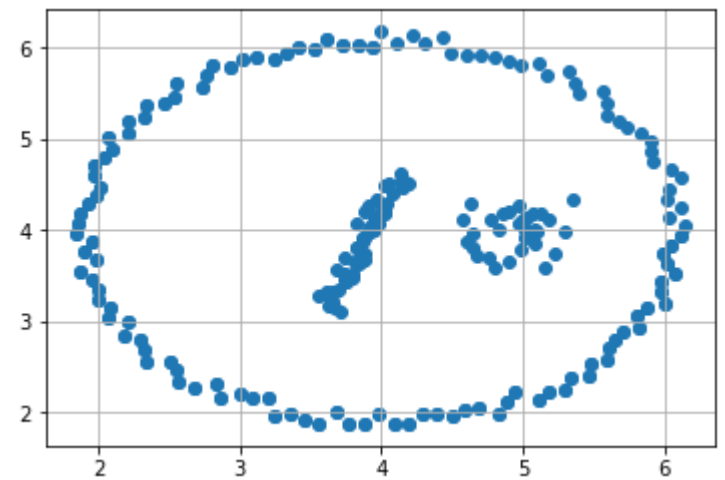
EXPERTO UNIVERSITARIO EN CIENCIA DE DATOS Y BIG DATA
MODULO 2 Ciencia Datos Técnicas Inteligentes
Extracción de conocimiento

Miguel Pagola


10
8
6
4
2
0
2
4
6
8


10
8
6
4
2
0
2
4
6
8
10


70
60
50
40
30
20
10
0
−30
−20
−10
0
10
20
30
40


6
5
4
3
2
2
3
4
5
6

EM y k-means

- En EM y k-means necesitamos saber K.
- Sólo sirven para clústeres esféricos (k-means) o elipsoides(EM).
- Parten de una inicialización aleatoria que puede acabar en mínimo local.
- Un elemento ruidoso (outlier) puede modificar la solución.
- Necesitamos poder calcular la distancia (euclídea y u otras) entre cada par de elementos.

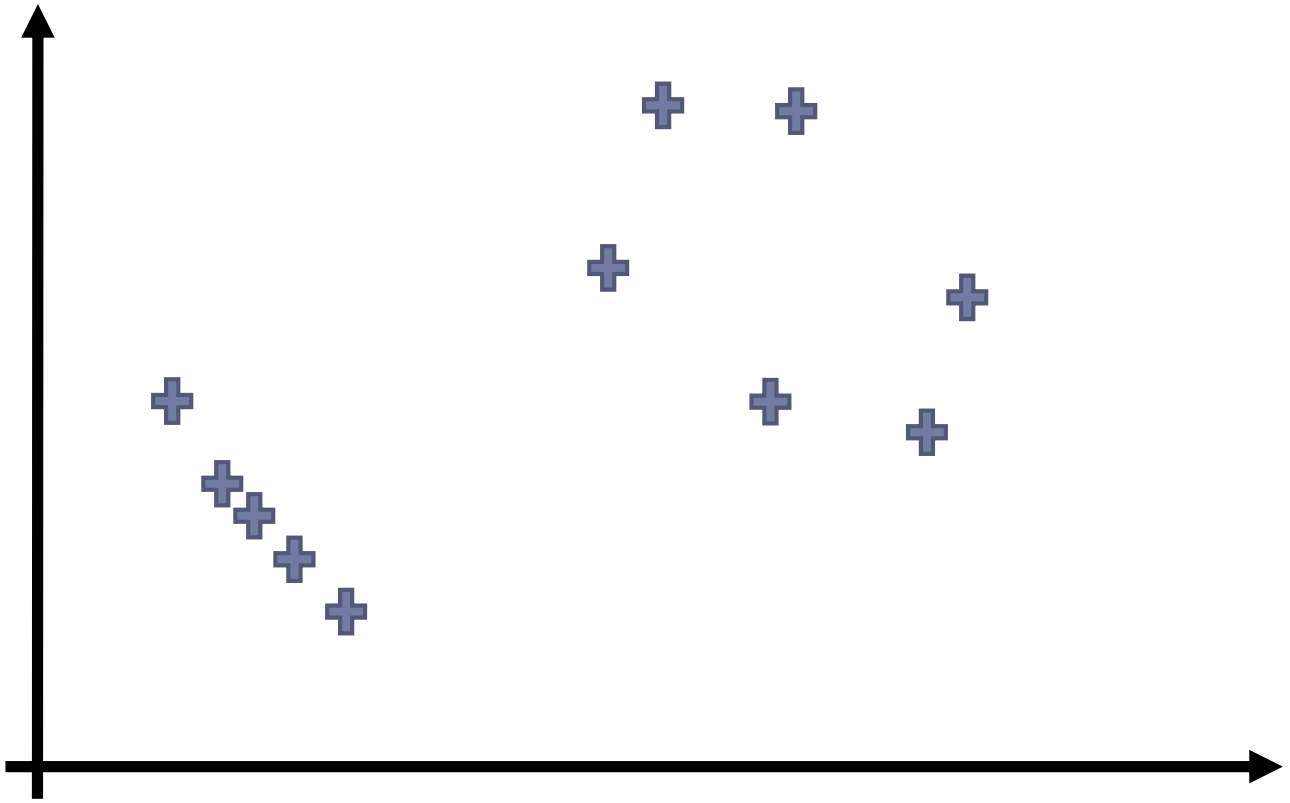
Clustering Jerárquico

- Algoritmo:

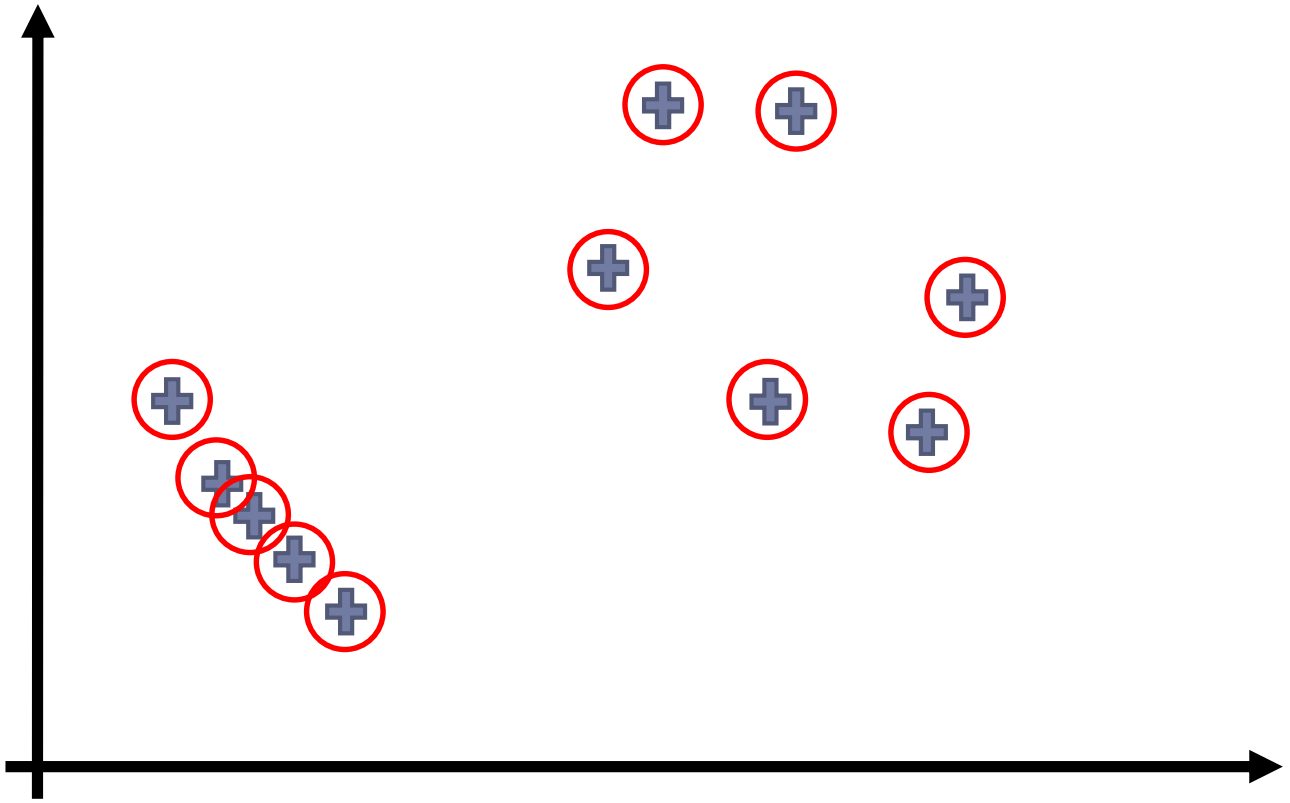
1. Cada elemento forma su propio cluster
2. Repetir:
   - **Juntar los dos clusteres “más cercanos”**

   Hasta que todos los datos esten agrupados en un único cluster

Clustering jeráquico

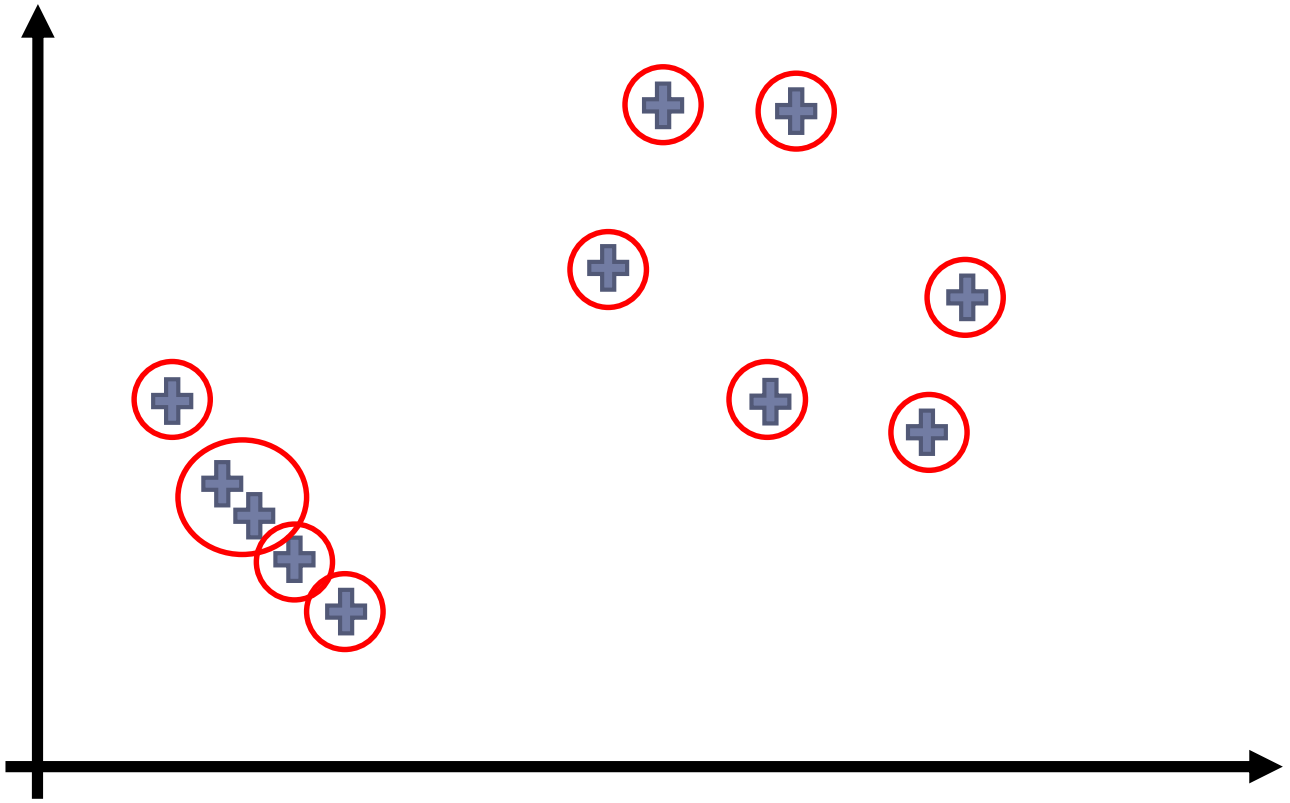


Clustering jerárquico



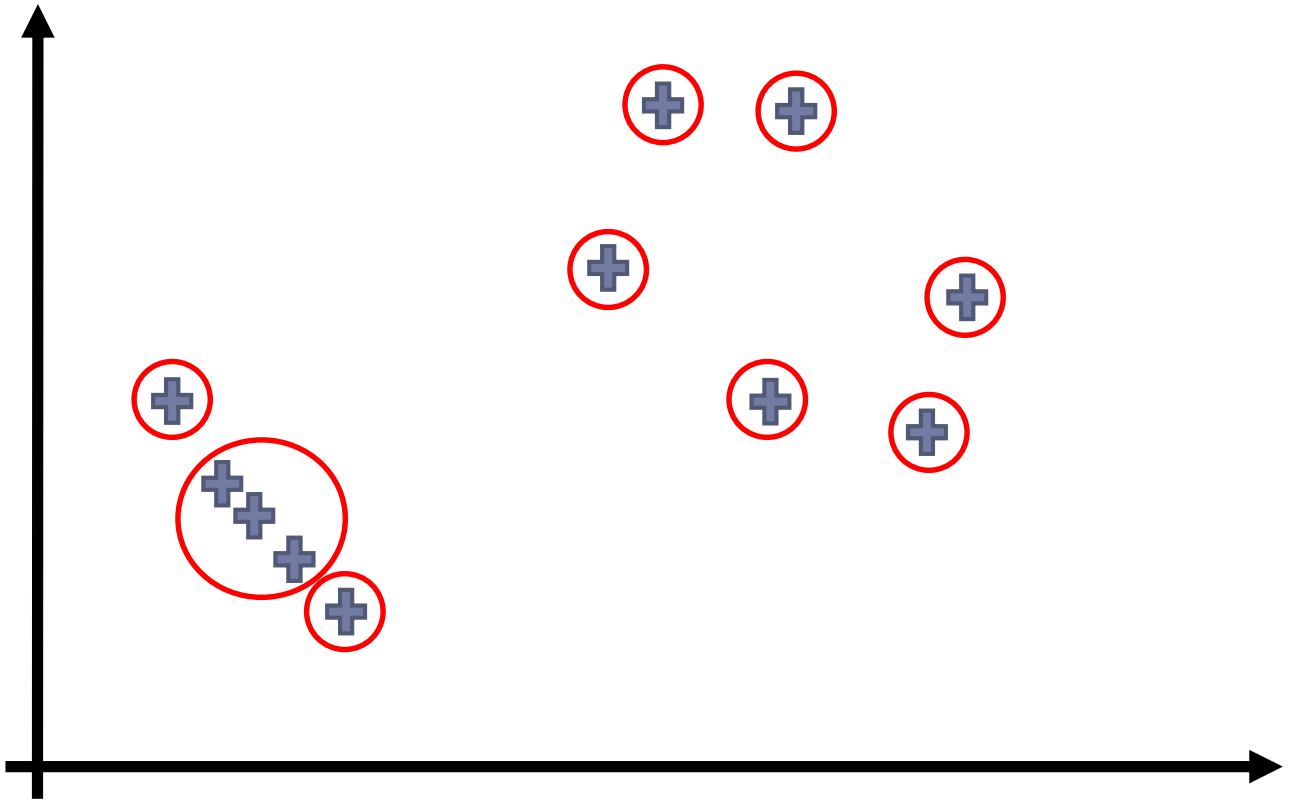
Primer paso: cada elemento forma su propio cluster

Clustering jerárquico



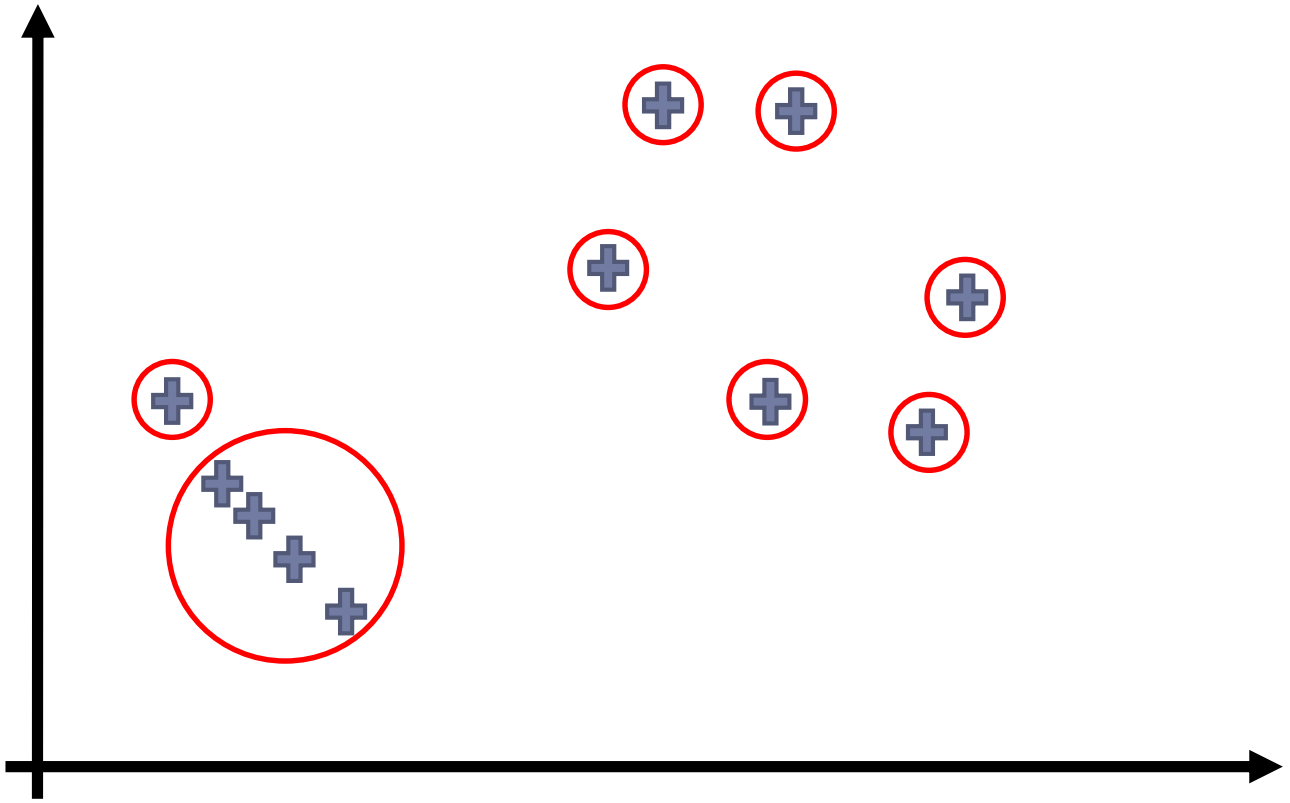
Repetir: juntar los dos clústeres más cercanos

Clustering jerárquico



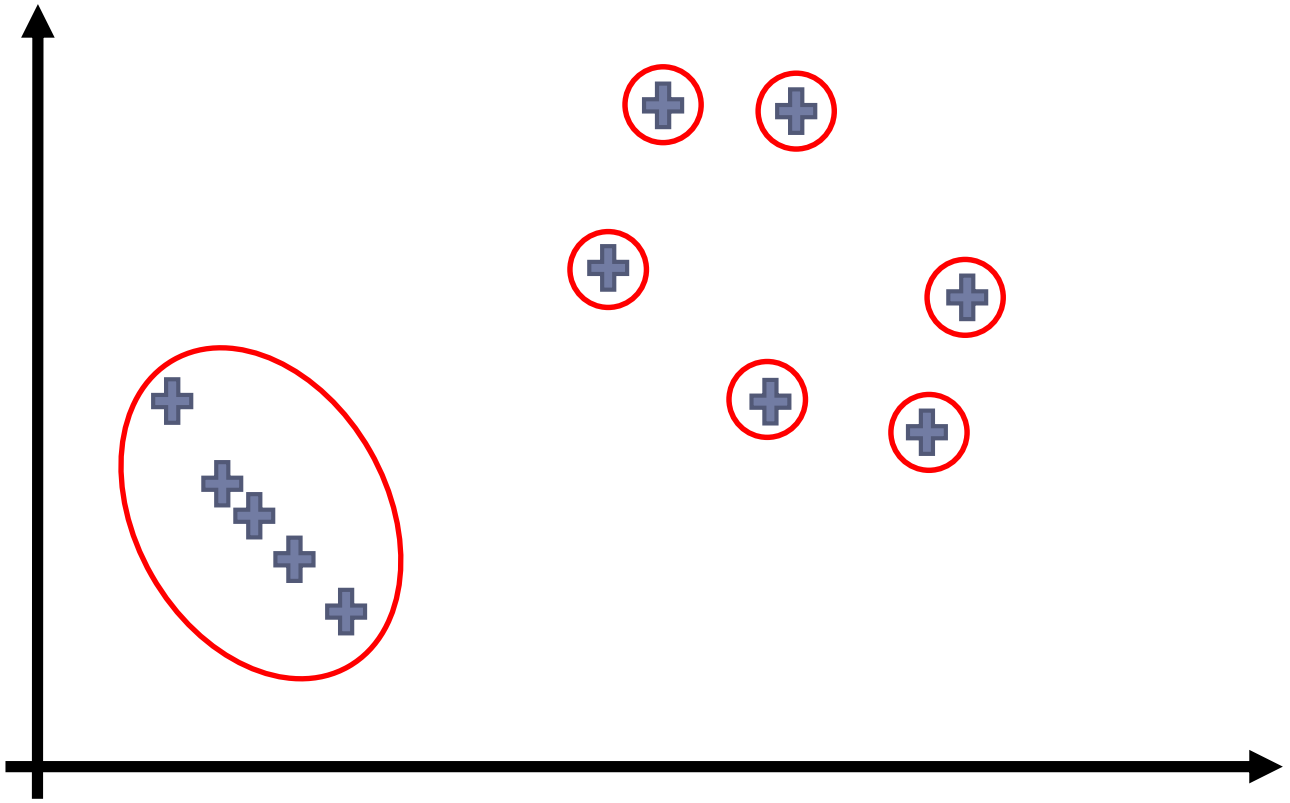
Repetir: juntar los dos clústeres más cercanos

Clustering jerárquico



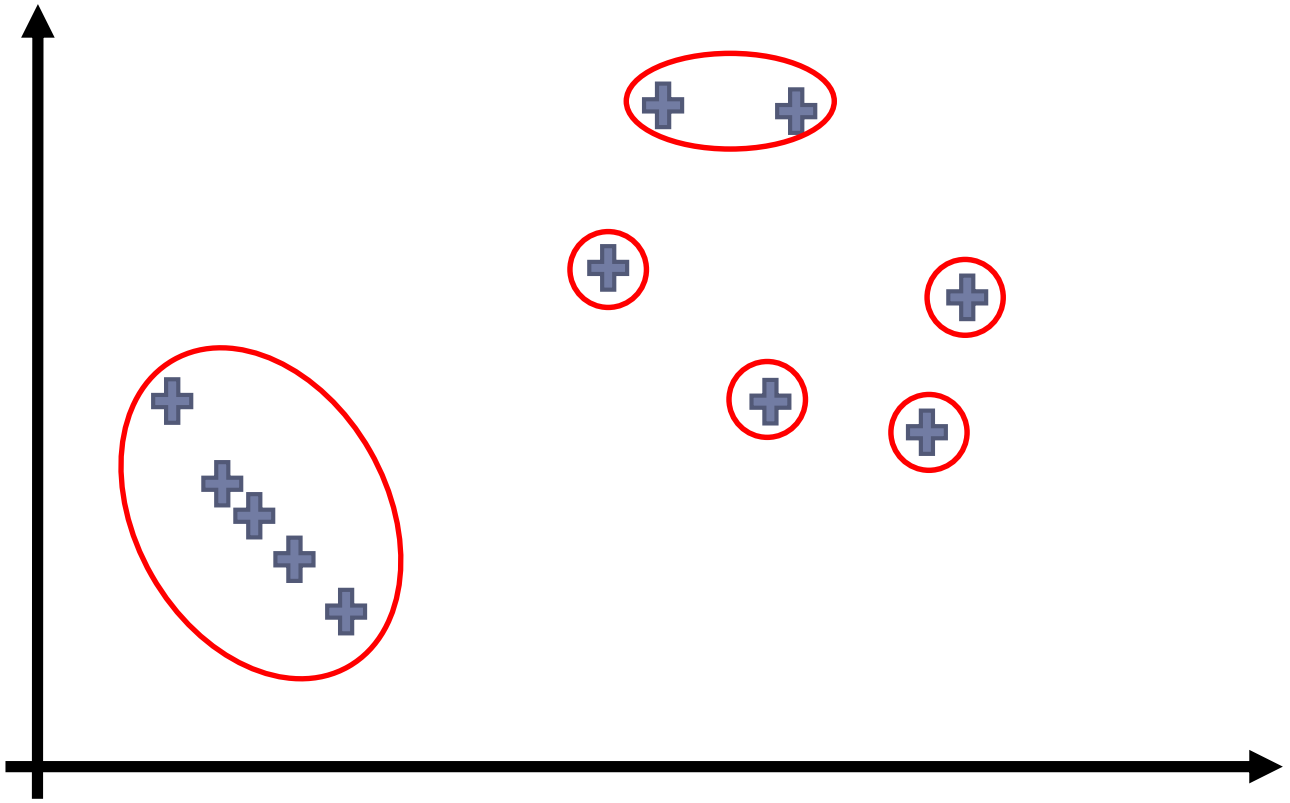
Repetir: juntar los dos clústeres más cercanos

Clustering jerárquico



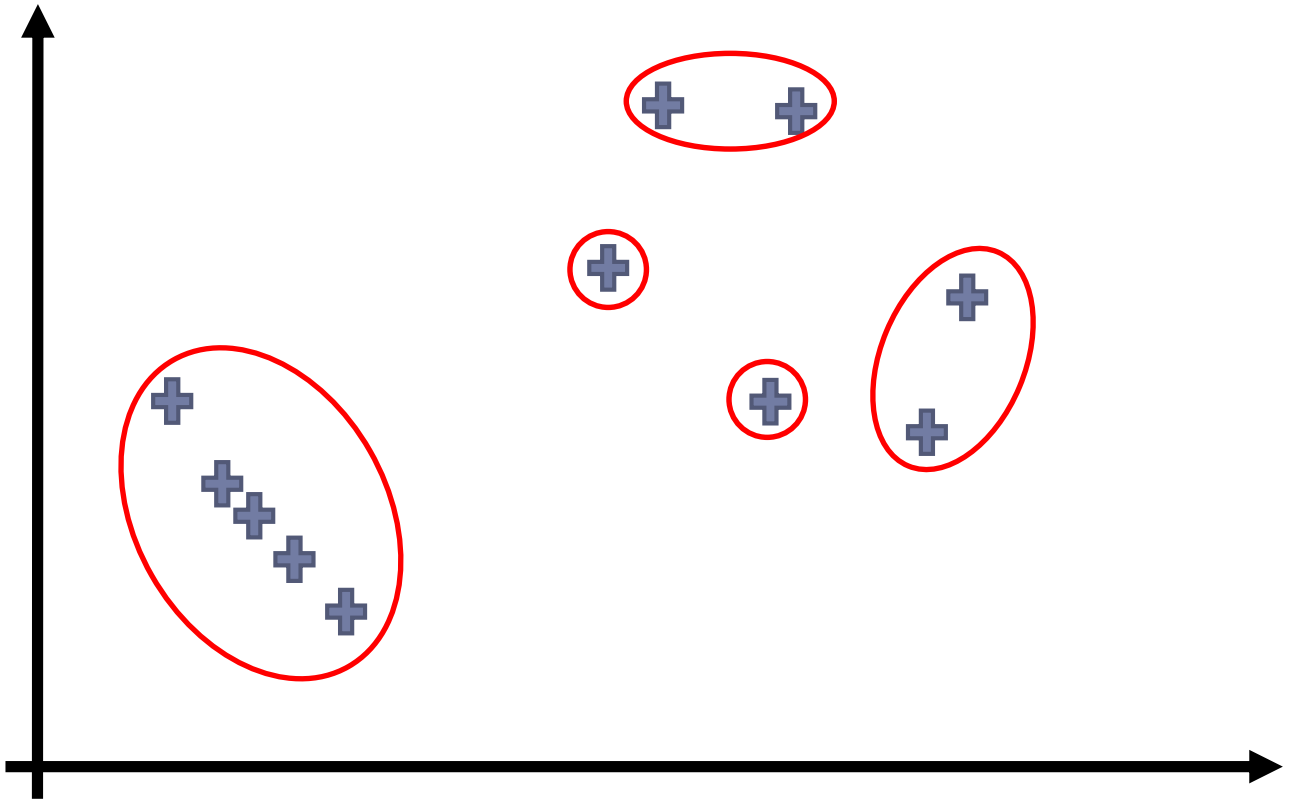
Repetir: juntar los dos clústeres más cercanos

Clustering jerárquico



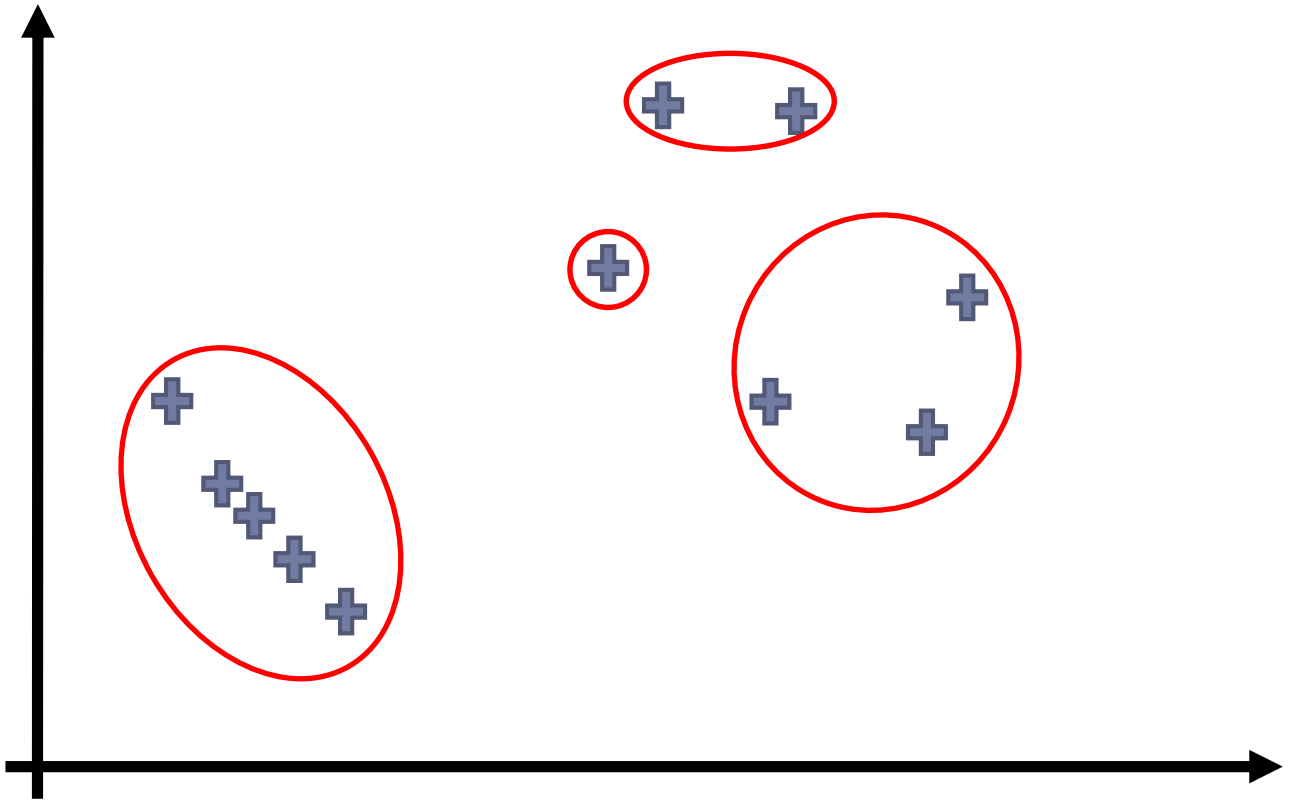
Repetir: juntar los dos clústeres más cercanos

Clustering jerárquico



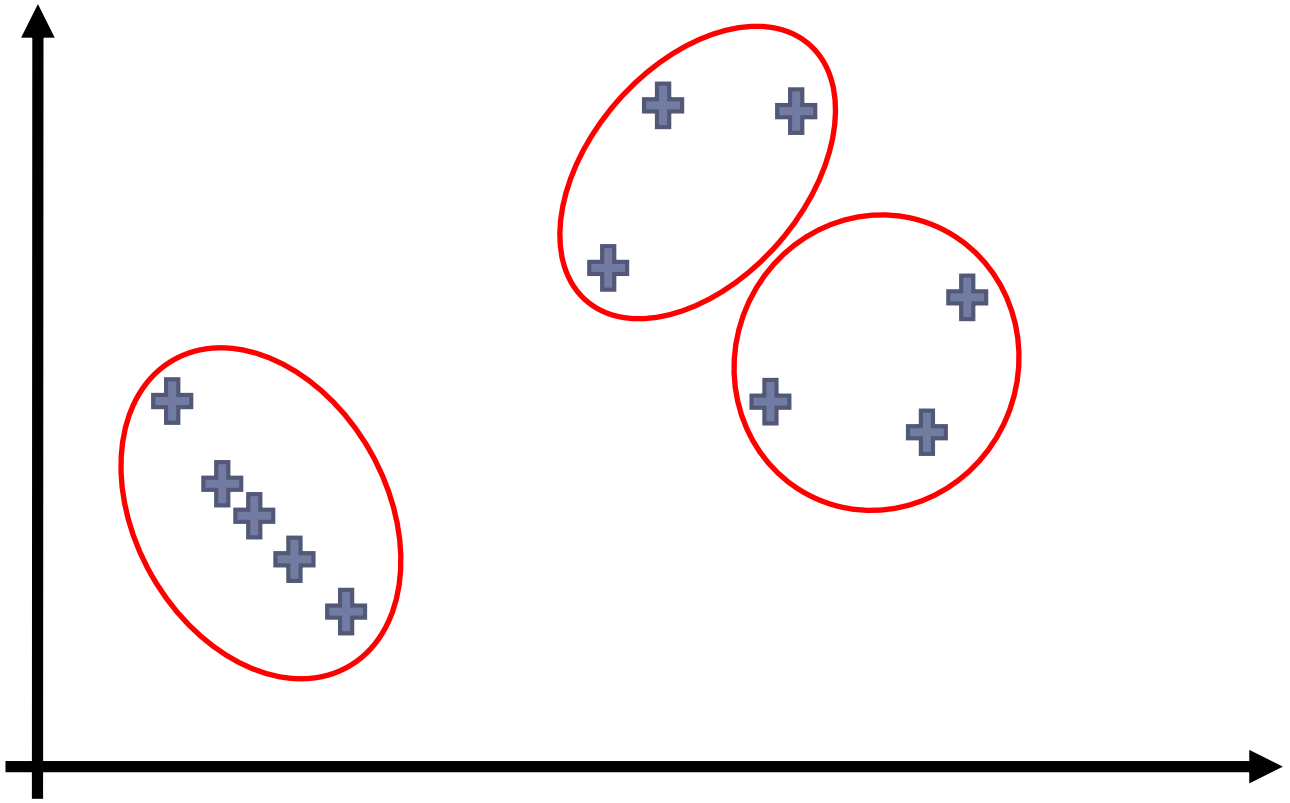
Repetir: juntar los dos clústeres más cercanos

Clustering jerárquico



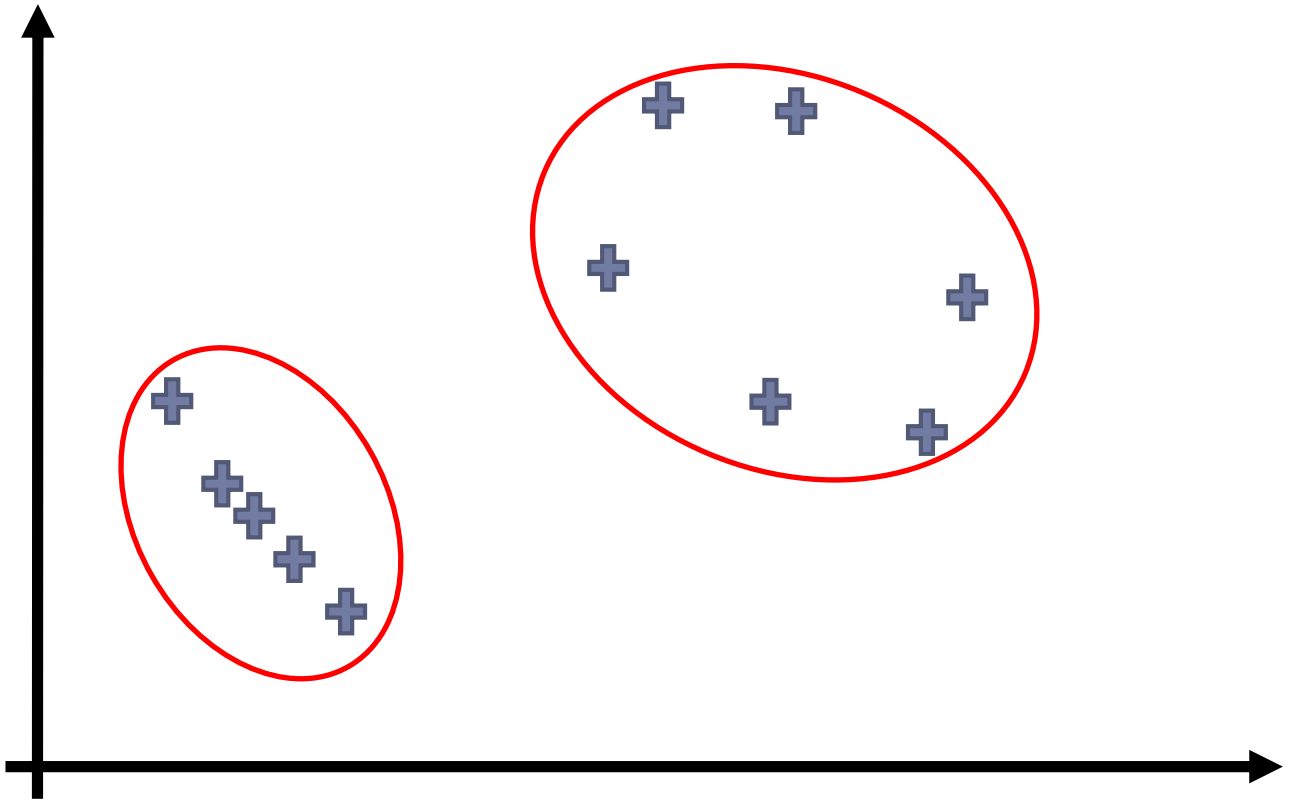
Repetir: juntar los dos clústeres más cercanos

Clustering jerárquico



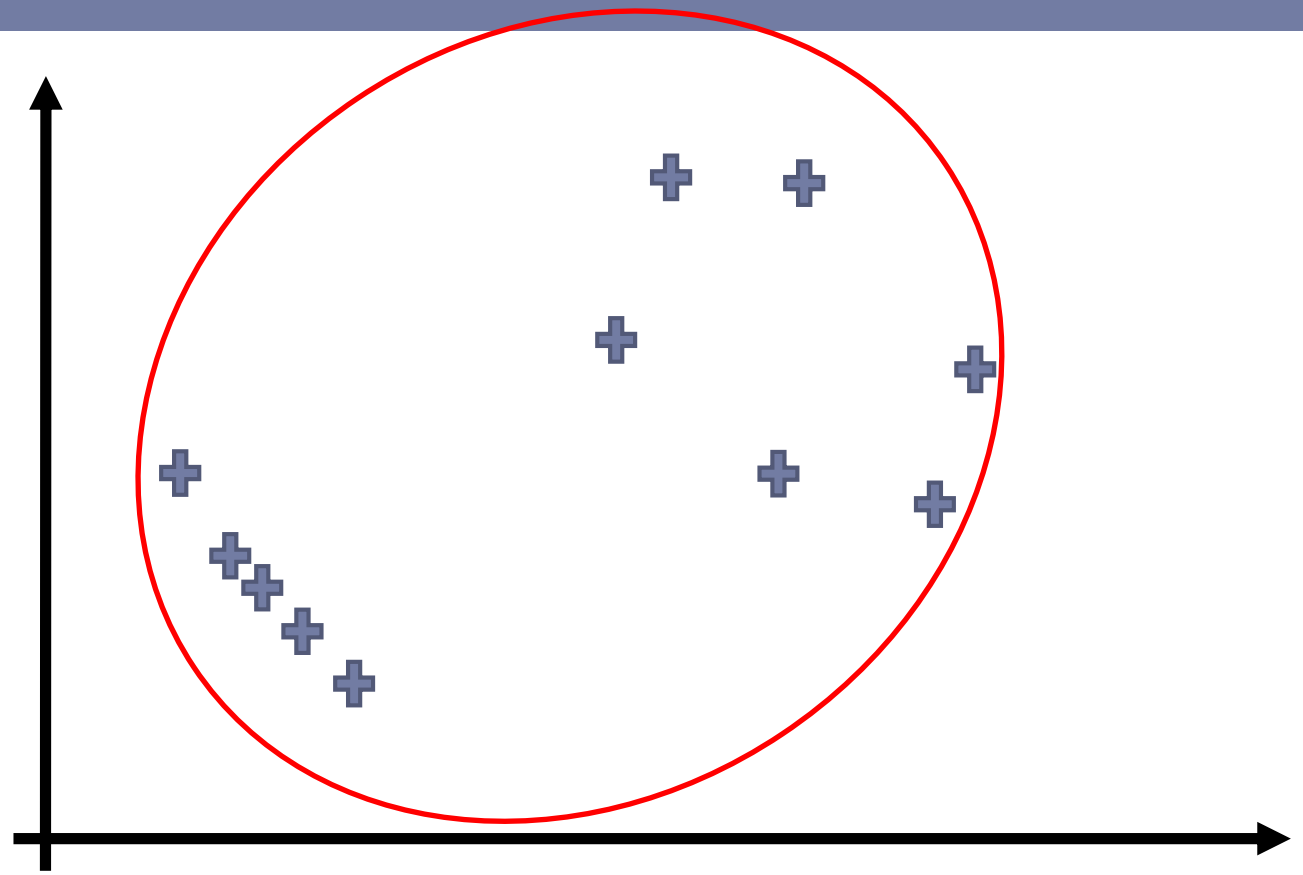
Repetir: juntar los dos clústeres más cercanos

Clustering jerárquico



Repetir: juntar los dos clústeres más cercanos

Clustering jerárquico



Repetir: juntar los dos clústeres más cercanos

Dendogramas

- El clustering jerárquico también se le suele llama aglomerativo.
- La similitud (distancia) entre los clústeres que se juntan es decreciente (creciente) monótono con respecto al nivel (iteración) de unión.
- Provee una visualización interpretable del algoritmo y de los datos.
- Dendograma: dibujar cada union de los clusteres
- Es una herramienta útil para la interpretación, por eso el clustering jerárquico es tan popular.

Clustering jeráquico

- Cada nivel del árbol resultante es una segmentación / partición de los datos.
- Queda en la mano del usuario decidir cual es el mejor clustering

- Es igual que el algoritmo de KRUSKAL!.
- Elementos→ vértices, distancias → coste del arco, pares de puntos → arcos.

Distancia entre clusteres

- Dadas las distancia entre todos los pares de elementos, existen diferentes opciones para definir la distancia entre los clústeres.
- Single-link:

$$d_{SL}(A,B) = \min_{i \in A, j \in B} \{d(i,j)\}$$

- Complete-link:

$$d_{CL}(A,B) = \max_{i \in A, j \in B} \{d(i,j)\}$$

- Average-link:

$$d_{AL}(A,B) = \frac{1}{m_A m_B} \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} d(i,j)$$

Single Link

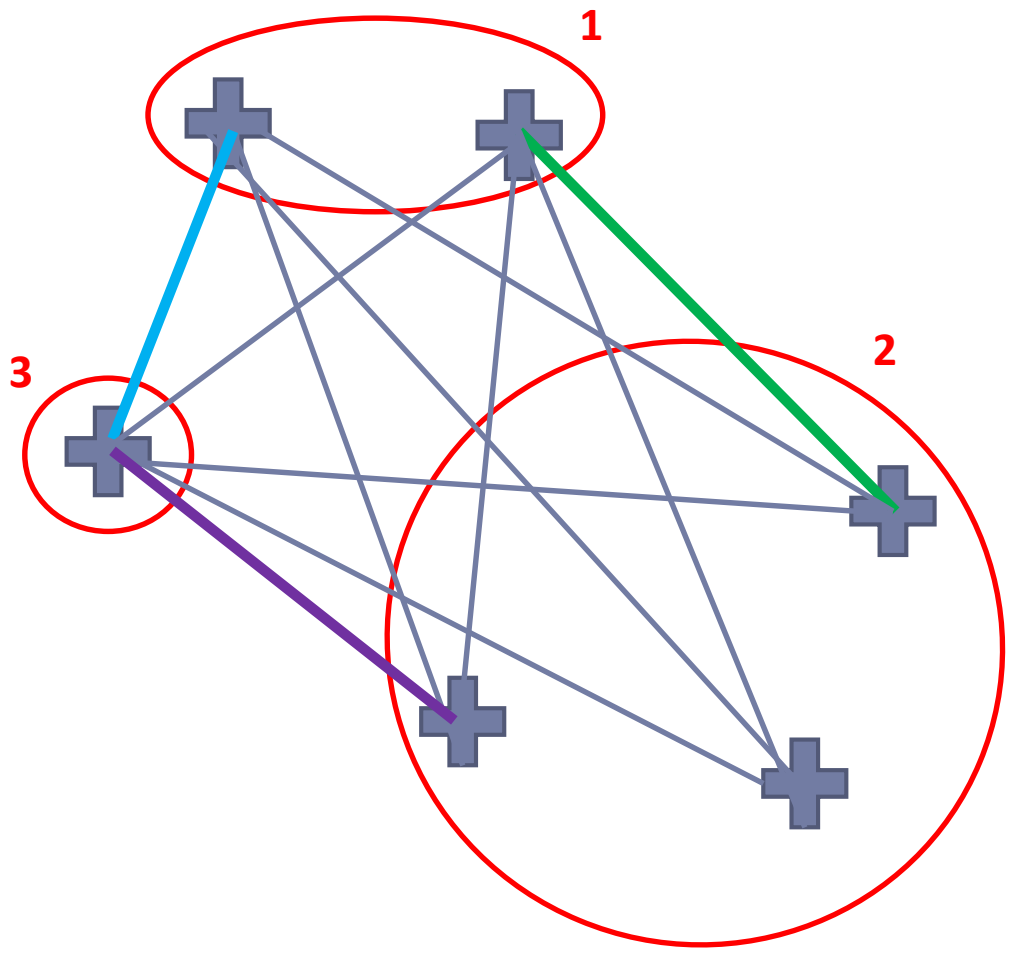
Dadas todas las distancias entre los pares de elementos entre dos clústeres la distancia single-link es la menor.

Distancia cluster 1-2

Distancia cluster 1-3

Distancia cluster 2-3

El paso del algoritmo continua siendo el mismo, uno los dos clústeres que más juntos estén (1-3).



Complete Link

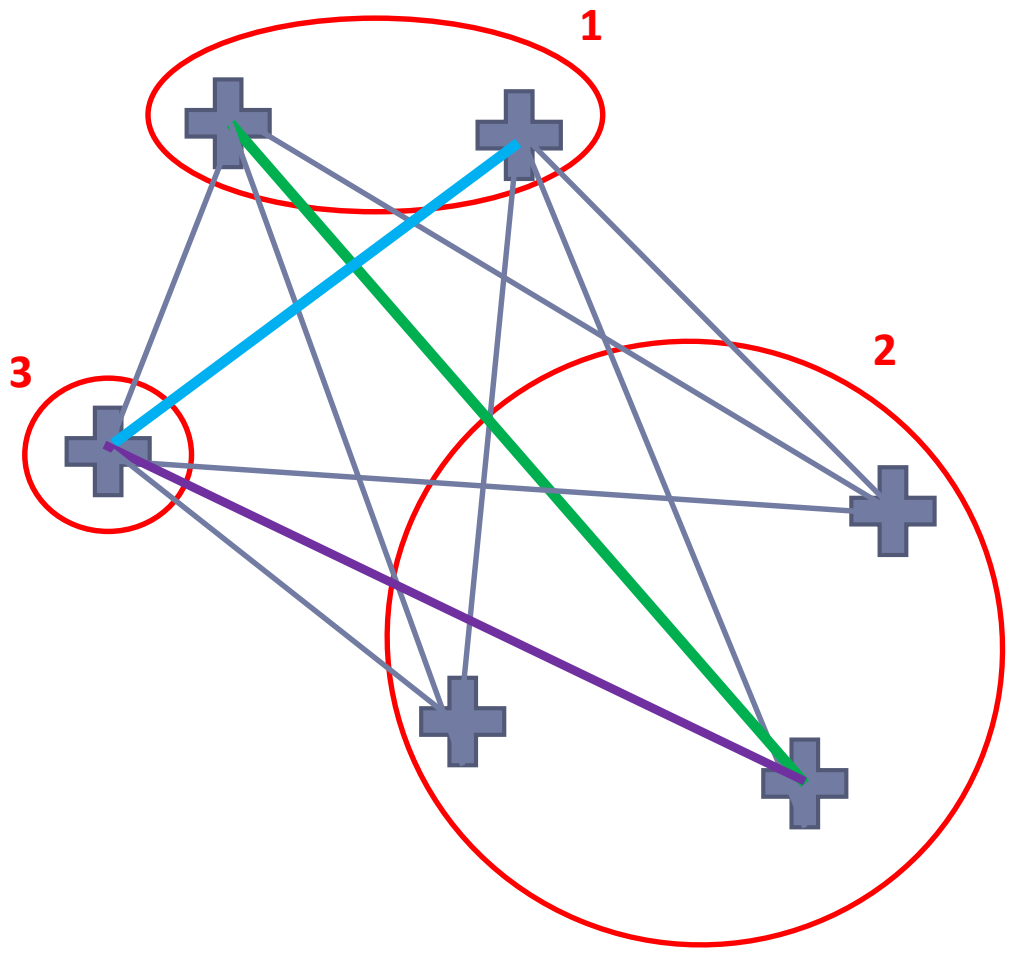
Dadas todas las distancias entre los pares de elementos entre dos clústeres la distancia complete-link es la mayor.

Distancia cluster 1-2

Distancia cluster 1-3

Distancia cluster 2-3

El paso del algoritmo continua siendo el mismo, uno los dos clústeres que más juntos estén (1-3).



Average Link

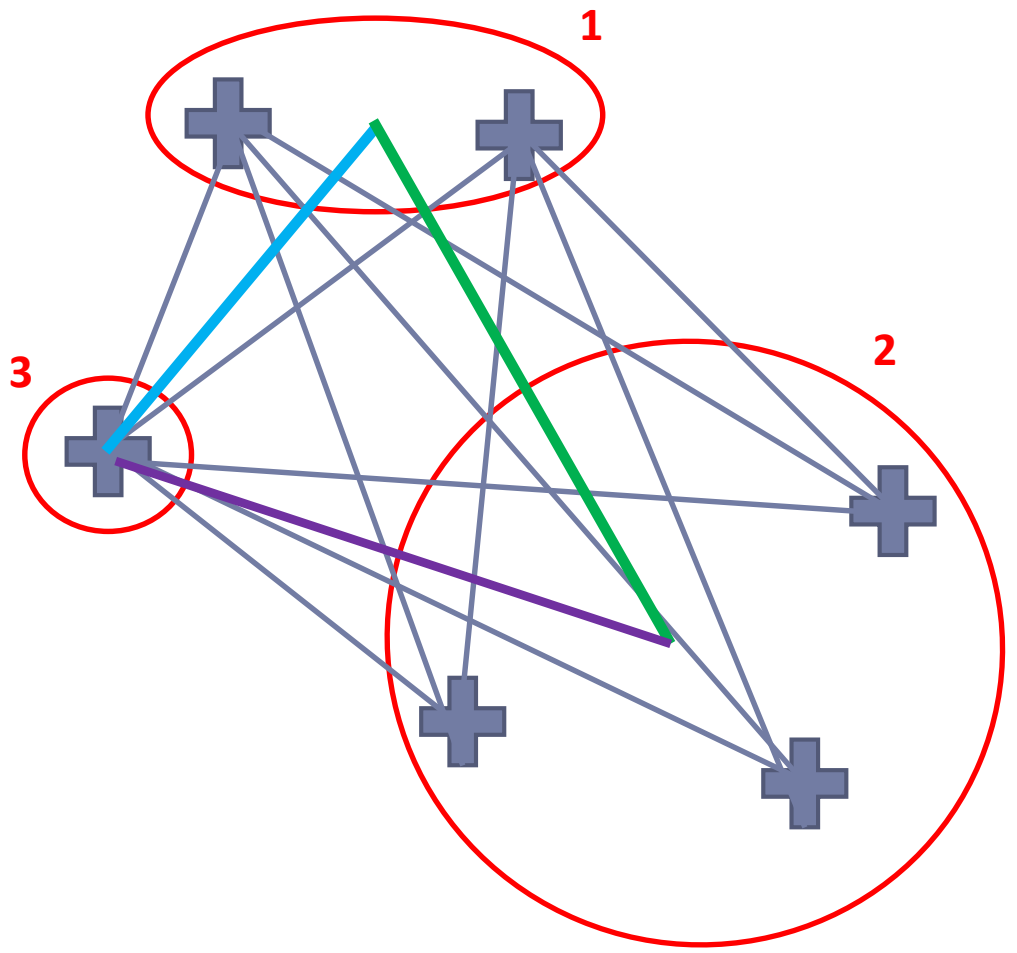
Dadas todas las distancias entre los pares de elementos entre dos clústeres la distancia average-link es la media de todas las distancias calculadas.

Distancia cluster 1-2

Distancia cluster 1-3

Distancia cluster 2-3

El paso del algoritmo continua siendo el mismo, uno los dos clústeres que más juntos estén (1-3).



Propiedades de los diferentes “linkages”

- **Single-link:** produce “encadenamientos”, en donde secuencias de puntos encadenados pueden unirse rápidamente.
- **Complete-link:** tiene justo el problema contrario, puede que clusteres que estén juntos no se unan porque tienen puntos entre sí que están bastante separados.
- **Average-link**: es un compromiso entre single-link y complete-link.

Clustering en datos no numéricos

- En K-means y EM necesitamos calcular la distancia entre dos pares de elementos del dataset, o poder calcular la media de varios elementos.
- Existen datos en los que tenemos diferentes tipos de variables o incluso datos que no estén representados mediante un vector.
- Opiniones subjetivas o valoraciones de preferencia.
- En estos casos utilizaremos el clustering jerárquico ya que SOLO necesitamos la matriz de disimilaridad (distancias).

Clustering en datos no numéricos

- Vamos a suponer que trabajamos en una agencia de viajes y queremos agrupar la publicidad según grupos de ciudades que consideramos parecidas.
- El criterio para agrupar no tiene que ser territorial, sino de tipo de actividades que ofrece, tipología de turistas que van a ellas, etc.
- Vamos a construir una matriz de SIMILITUD

Clustering en datos no numéricos

	Atenas	Londres	Paris	Nueva York	Roma	Melbourne	Dubai
Atenas					**0,9**		
Londres							
Paris							
Nueva York							
Roma							**0,2**
Melbourne							
Dubai							

Cuanto más parecidas las ciudades damos un valor más alto, y cuando más diferentes un valor más bajo (siempre entre 0 y 1)

Para ejecutar el algoritmo necesitamos la matriz de distancias D = 1- S