

Machine Learning aplicado a las Ciencias Sociales

Clase 3. Fundamentos de clustering



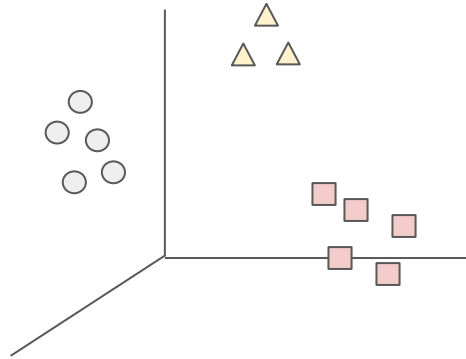
Clustering

- Amplia clase de métodos que buscan detectar grupos y subgrupos en los datos
 - Dado un conjunto de clientes, ¿es posible encontrar segmentos de clientes similares, según su historial de compras?
 - Agrupar películas en función de sus clasificaciones
 - Encontrar grupos de jurisdicciones en función de características sociodemográficas
 - Etc...



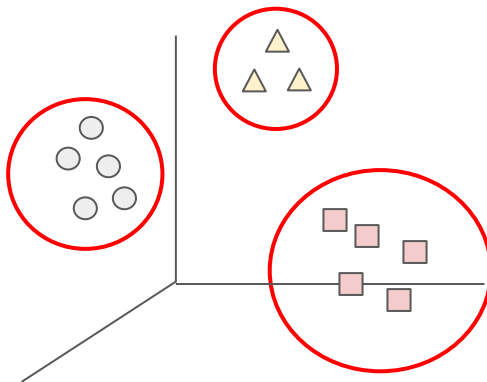
Clustering

Encontrar **subgrupos** (*clústers*) en los datos



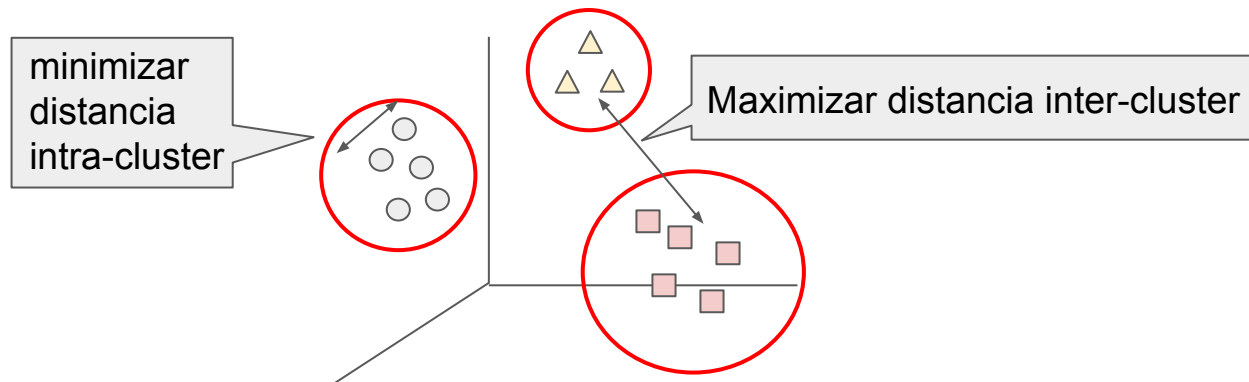
Clustering

Encontrar **subgrupos** (*clústers*) en los datos



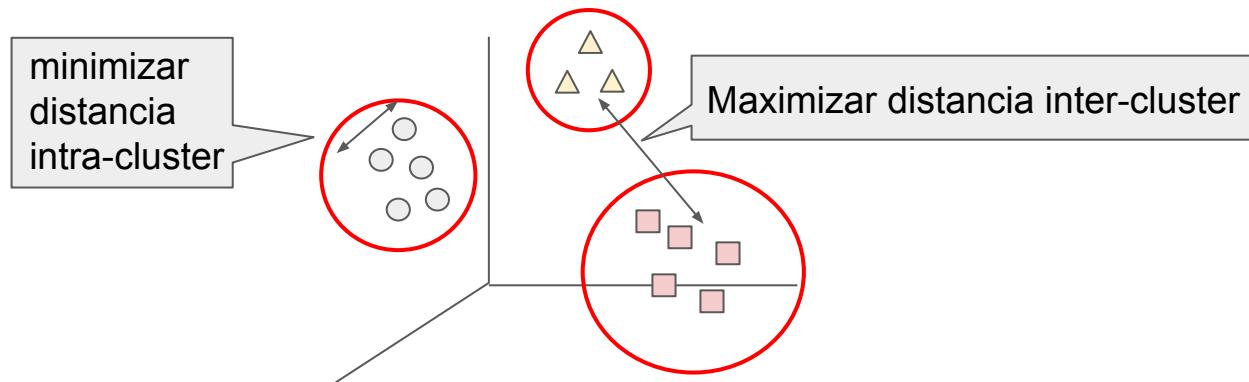
Clustering

Encontrar **subgrupos** (*clústers*) en los datos



Clustering

Encontrar **subgrupos** (*clústers*) en los datos

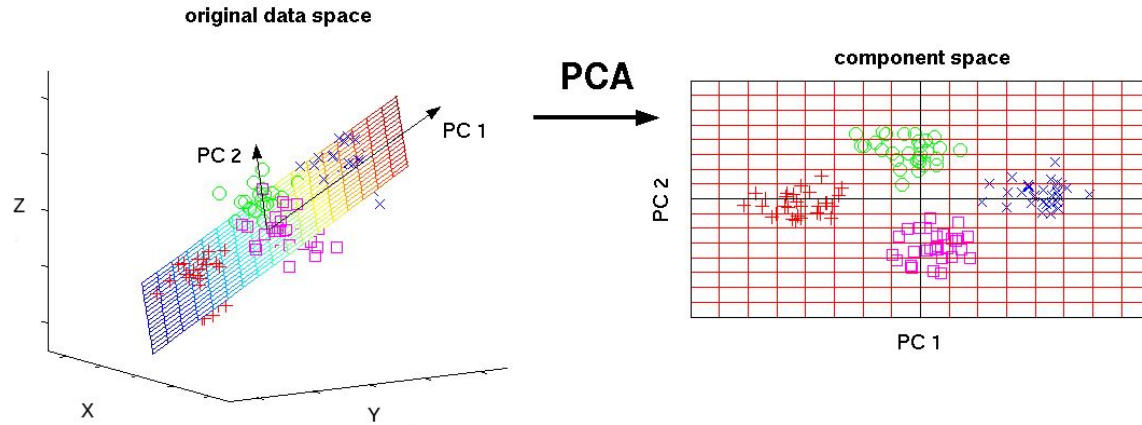


Observaciones dentro de un cluster **similares**

Observaciones entre clusters **no similares**



Clustering - Diferencia con PCA

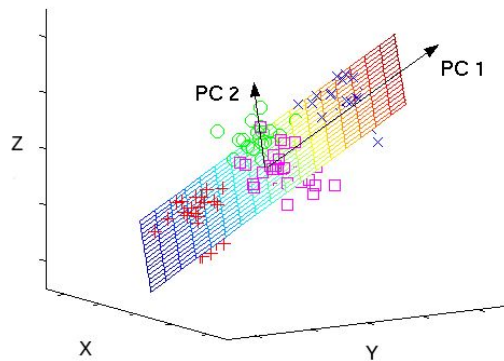


Reducir dimensión maximizando la varianza



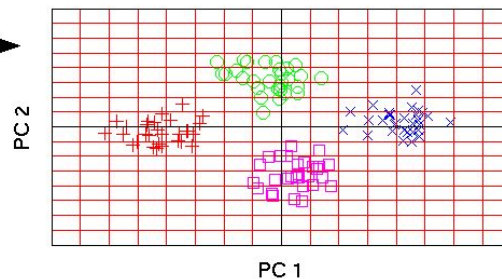
Clustering - Diferencia con PCA

original data space



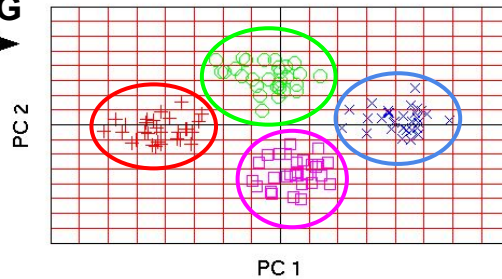
PCA

component space



CLUSTERING

component space

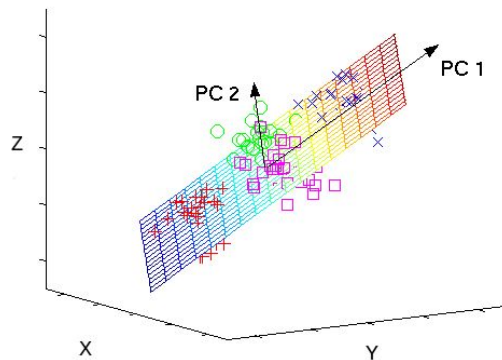


Reducir dimensión maximizando la varianza



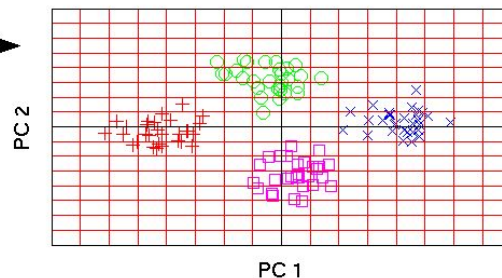
Clustering - Diferencia con PCA

original data space



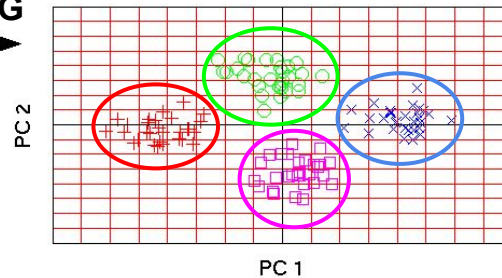
PCA

component space



CLUSTERING

component space



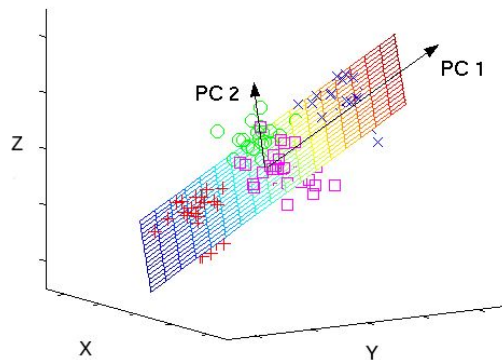
Reducir dimensión maximizando la varianza

Encontrar grupos homogéneos



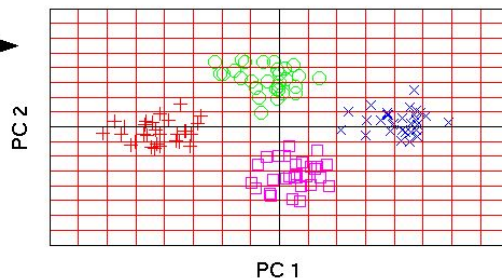
Clustering - Diferencia con PCA

original data space



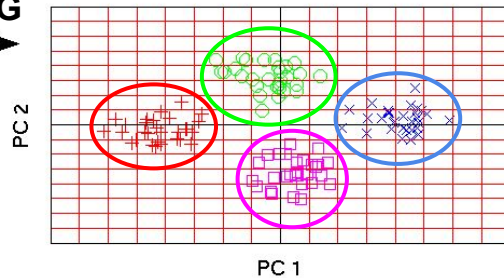
PCA

component space



CLUSTERING

component space



Reducir dimensión maximizando la varianza

Encontrar grupos homogéneos

Se puede encontrar grupos en el espacio de features original

Si son muchos

-> podría ser costoso computacionalmente

-> podrían esconderse las características que mejor agrupan los datos



Tipos de problemas

- Buscamos encontrar grupos de casos que sean parecidos entre sí
- Es necesario definir qué significa *similitud* o *diferencia*
- El conocimiento del dominio es fundamental en esta definición



Dos métodos (entre muchos otros)

- K-medias

Encontrar una partición de las observaciones según un número pre-definido de clusters

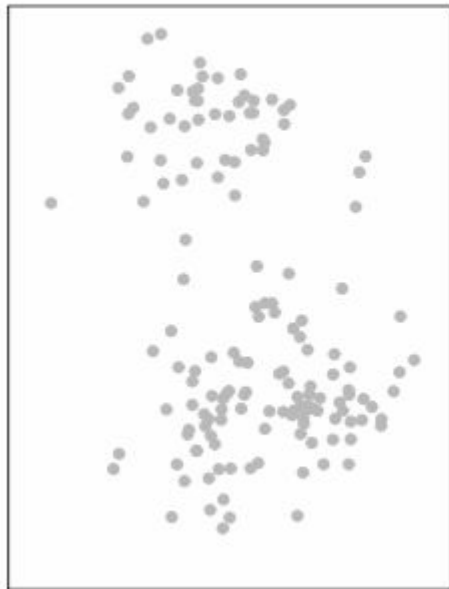


Dos métodos (entre muchos otros)

- Clustering jerárquico
 - No conocemos a priori el número de clusters
 - Usamos el *dendograma* para definirlo



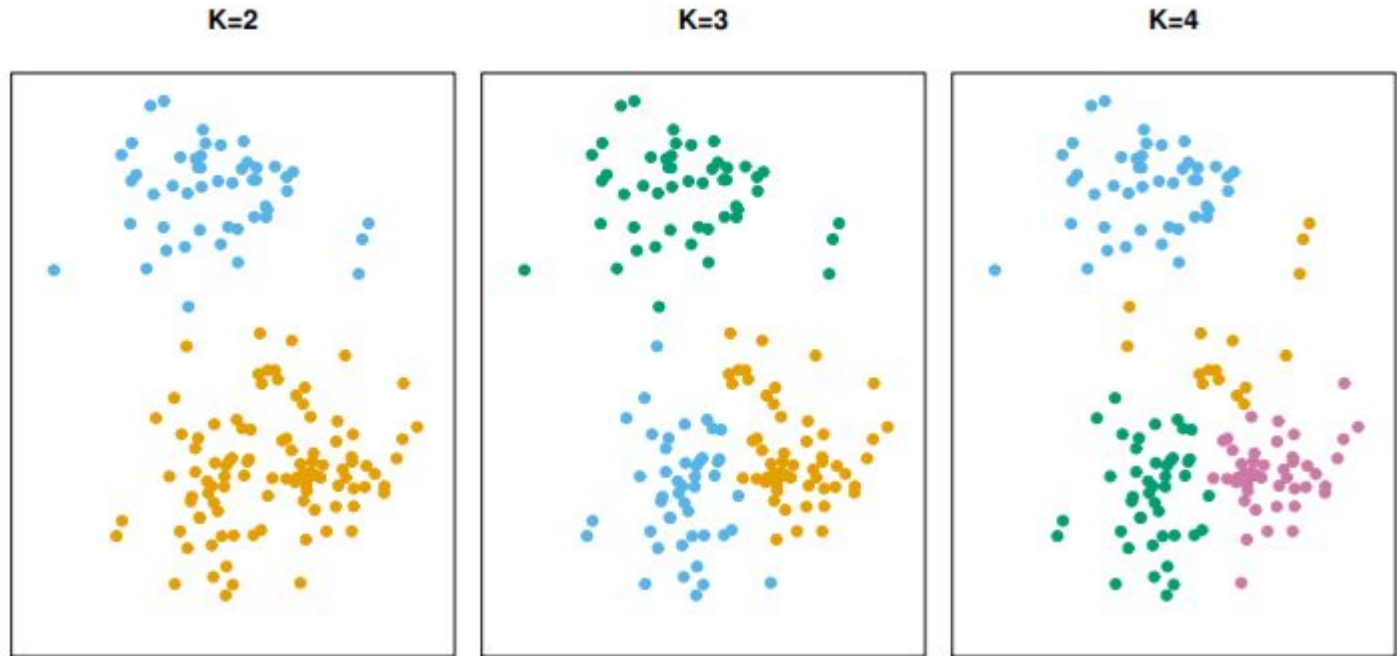
K-Medias



- Cada puntito es una persona, caracterizada por su experiencia laboral (eje X) y su ingreso (eje Y)
- ¿Cuántos grupos existen?



K-Medias



K-Medias

- Tenemos K grupos C_1, C_2, \dots, C_K
- Cada grupo tiene dos propiedades:
 - $C_1 \cup C_2 \cup \dots \cup C_K = \{1, 2, 3, \dots, n\}$, es decir que cada unidad pertenece a un cluster
 - $C_1 \cap C_2 \cap \dots \cap C_K = \emptyset$, es decir, que elemento queda clasificado en un solo cluster



K-Medias

- La idea es que un buen esquema de clustering implica que adentro del cluster existe la menor variabilidad posible
- Variabilidad intra-cluster $WCV(C_k)$: medida que indica cuánto se diferencian los elementos al interior de un cluster

$$\underset{C_1, C_2, \dots, C_K}{\text{minimize}} \left\{ \sum_{k=1}^K WCV(C_k) \right\}$$



K-Medias

- ¿Cómo definimos la medida de variación intra cluster -WCV-? Típicamente, la distancia euclidiana:

$$WCV(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^p \left(x_{i,j} - x_{i',j} \right)^2$$

- Entonces, queremos minimizar (**nuestra función objetivo**)

$$\text{minimize}_{C_1, C_2, \dots, C_k} \left\{ \sum_{k=1}^K \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^p \left(x_{i,j} - x_{i',j} \right)^2 \right\}$$



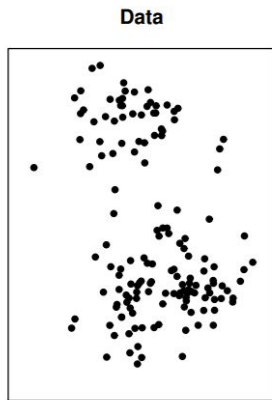
K-Medias - Algoritmo

1. Asignar aleatoriamente cada observación a un cluster (entre 1 y K). Esto sirve como un punto de inicialización
2. Repetir hasta que la asignación deje de cambiar:
 - 2.1 Para cada uno de los K clusters, computar el centroide (el vector que contiene el promedio de cada una de las variables a considerar)
 - 2.2 Computar la distancia euclidiana de cada caso a cada centroide y asignar cada caso al cluster con el centroide más cercano



K-Medias - Algoritmo

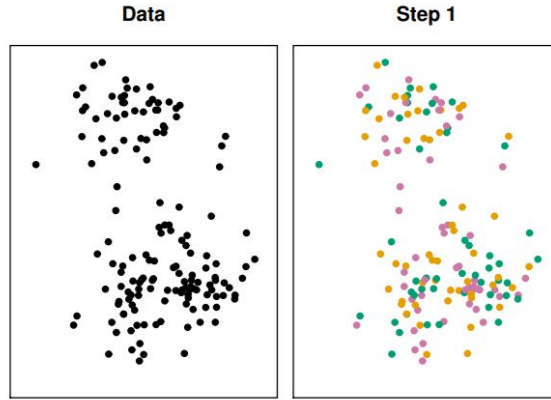
Definimos el número de clusters que buscamos (k)



K-Medias - Algoritmo

Inicializamos los
centroides de forma
aleatoria

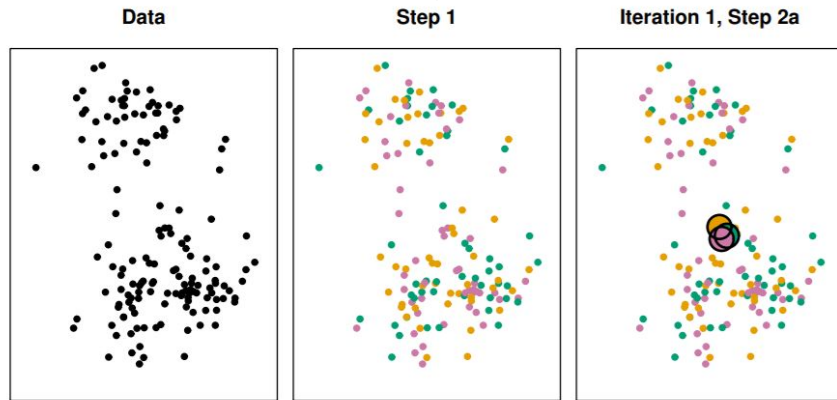
Definimos el
número de
clusters que
buscamos (k)



K-Medias - Algoritmo

Definimos el número de clusters que buscamos (k)

Inicializamos los centroides de forma aleatoria



Calculamos los **centroides** (centros) de cada cluster como el promedio de las features de sus samples

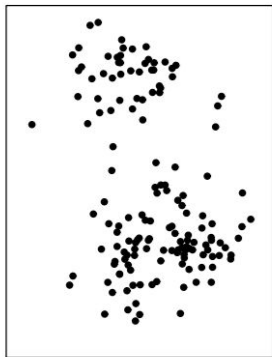


K-Medias - Algoritmo

Inicializamos los
centroides de forma
aleatoria

Definimos el
número de
clusters que
buscamos (k)

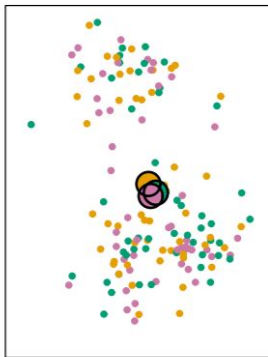
Data



Step 1

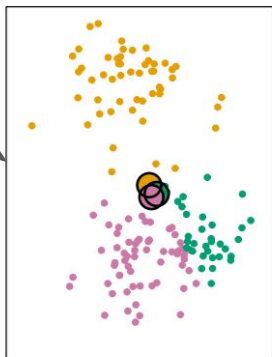


Iteration 1, Step 2a



Calculamos los
centroides
(centros) de cada
cluster como el
promedio de las
features de sus
samples

Iteration 1, Step 2b



Asignamos
cada caso al
cluster con
centroide
más cercano

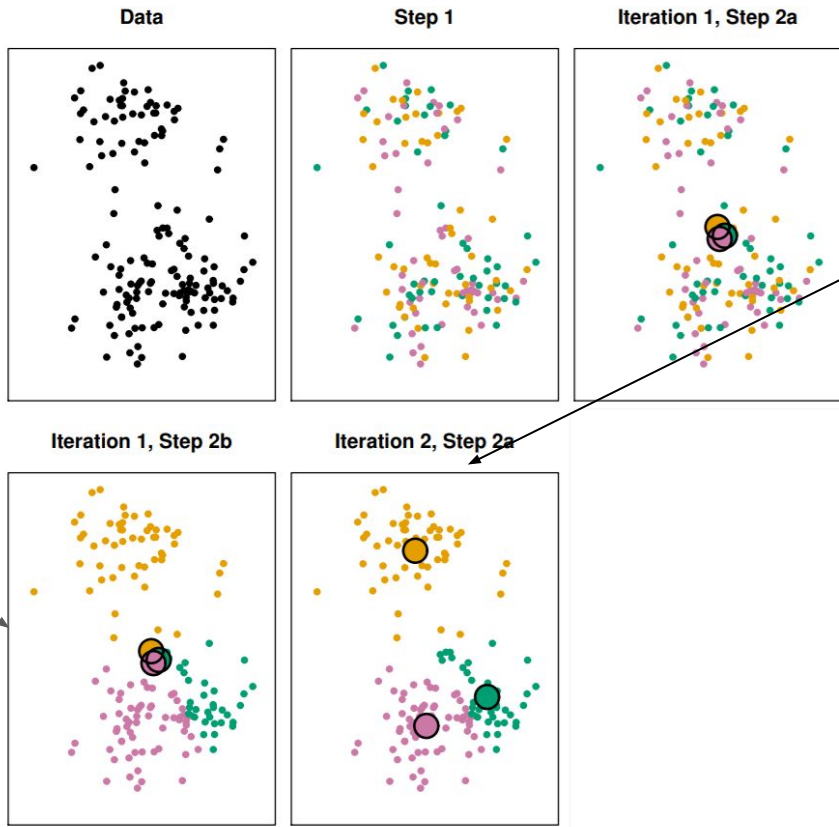


factor-data
EIDAEs_UNSAM

K-Medias - Algoritmo

Inicializamos los
centroides de forma
aleatoria

Definimos el
número de
clusters que
buscamos (k)



Calculamos los
centroides
(centros) de cada
cluster como el
promedio de las
features de sus
samples

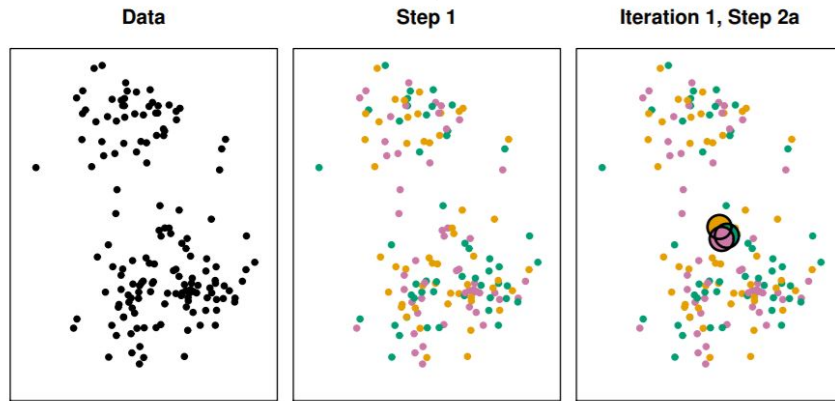
Asignamos
cada caso al
cluster con
centroide
más cercano



K-Medias - Algoritmo

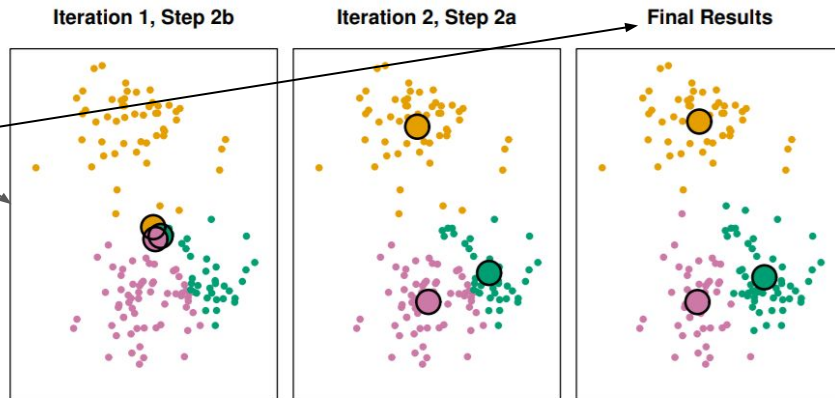
Inicializamos los centroides de forma aleatoria

Definimos el número de clusters que buscamos (k)



Calculamos los **centroides** (centros) de cada cluster como el promedio de las features de sus samples

Asignamos cada caso al cluster con centroe más cercano



K-Medias - Diferentes inicializaciones

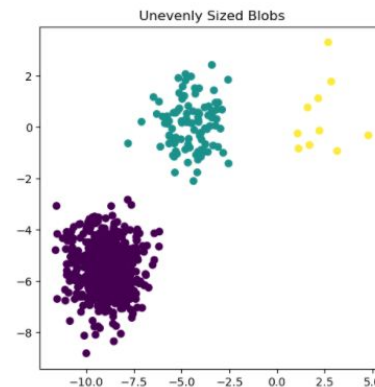
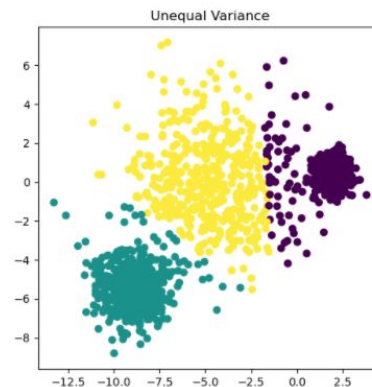
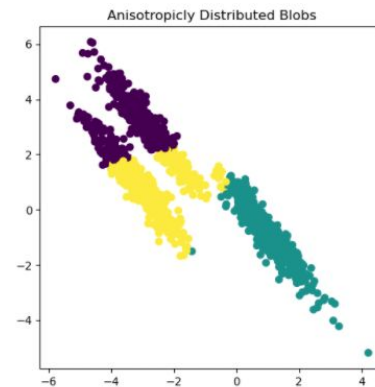
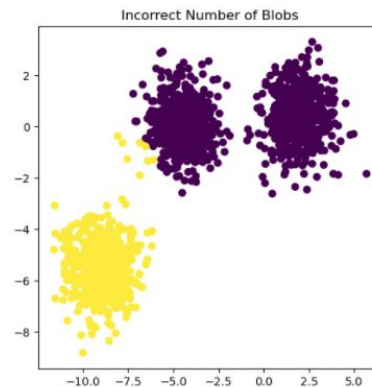
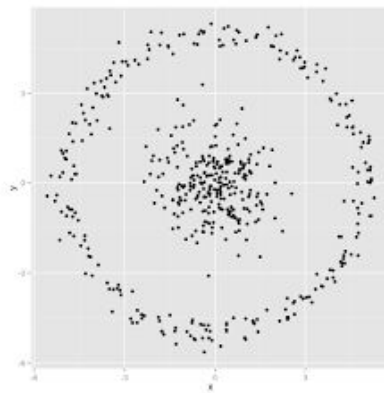


- Inicialización aleatoria: modelo no determinista
- Los resultados dependen de esa inicialización



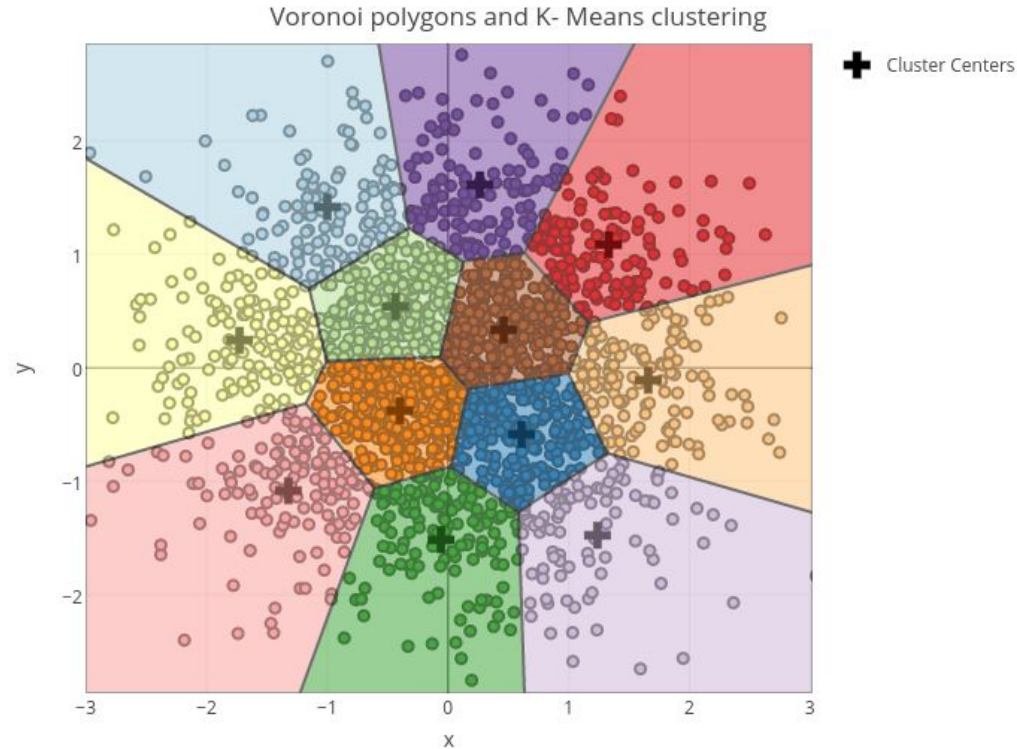
K-Medias - Ventajas y límites

- + Simple y Fácil de implementar
- + Orden del algoritmo es lineal
- Depende de la inicialización
- Tiende a caer en un mínimo local
- Sensible a outliers
- Los clusters tienen que tener forma esférica
- No se puede aplicar a data categórica



K-Medias - Ventajas y límites

- Da un clustering de los datos aún si los datos no están “clusterizados”

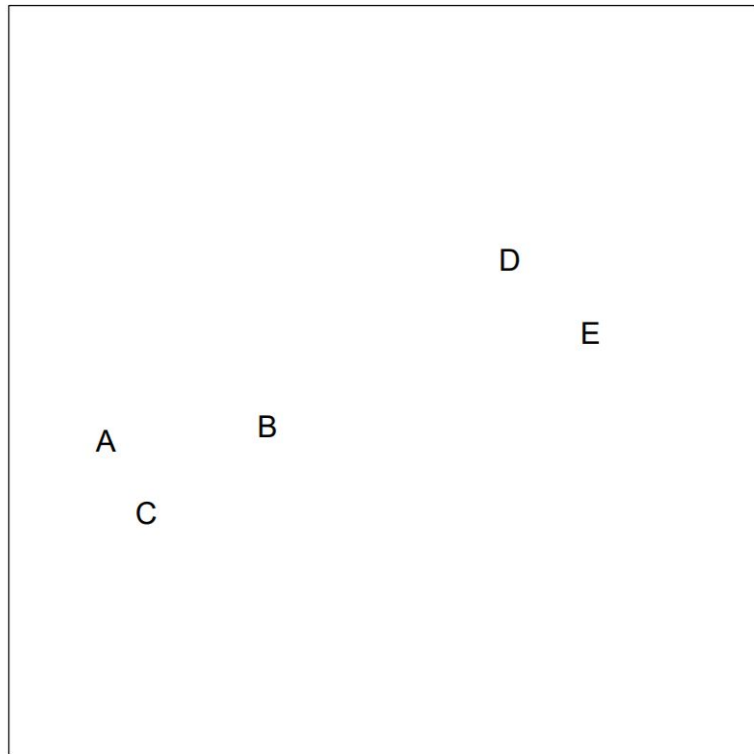


Clustering jerárquico

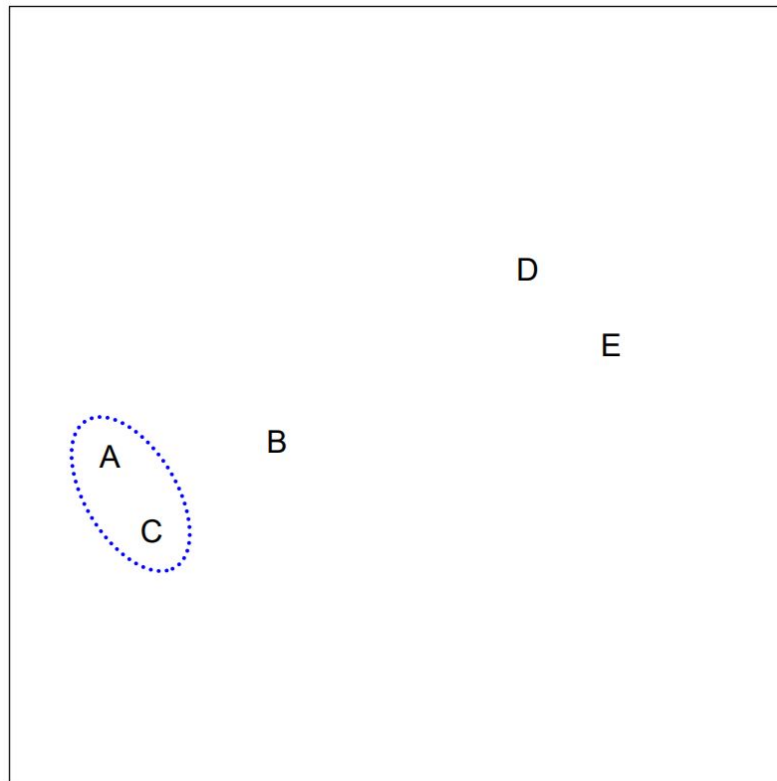
- K-medias requiere especificar de antemano la cantidad de clusters buscados
- El clustering jerárquico ofrece un método para tratar de estimarlo
- Veremos el método más común de clustering jerárquico: bottom-up o “aglomerativo”



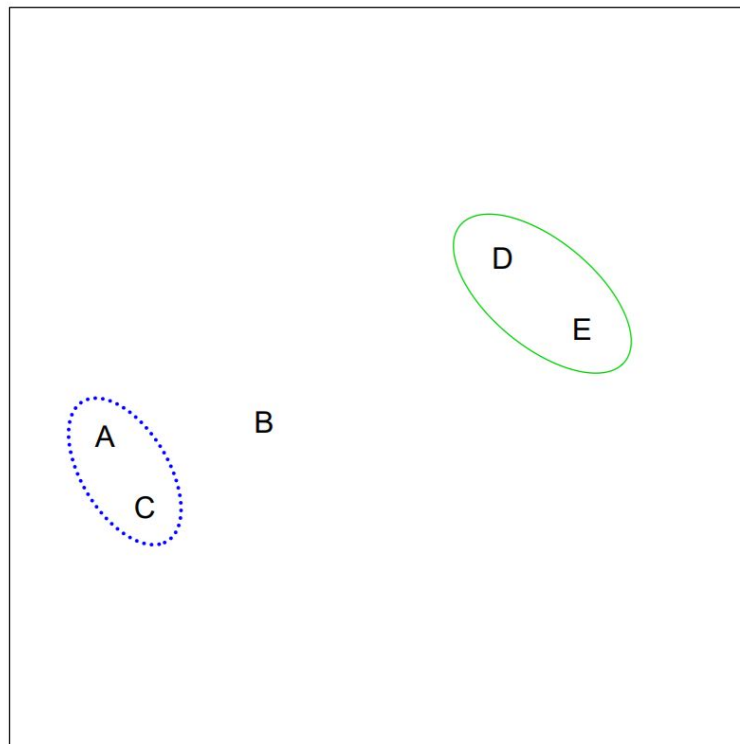
Clustering jerárquico



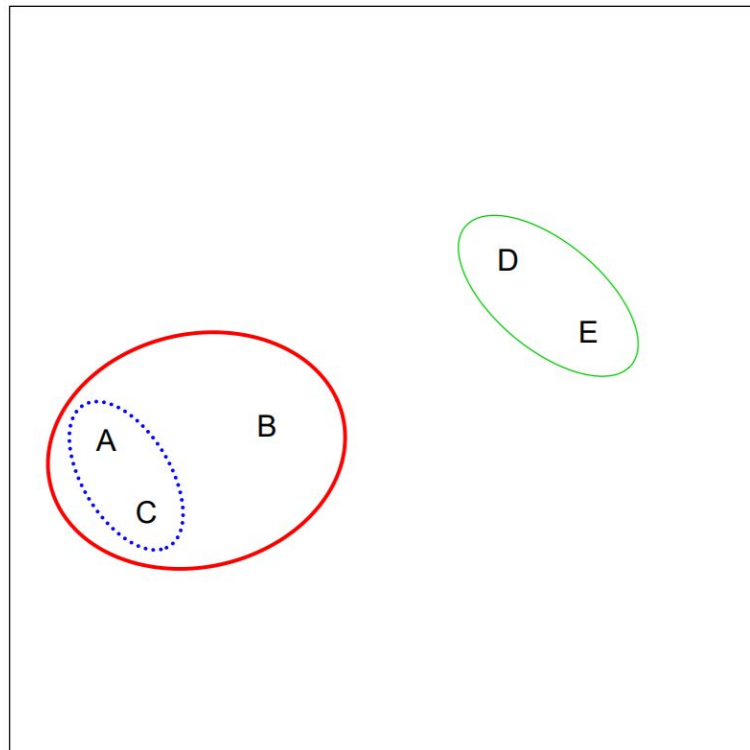
Clustering jerárquico



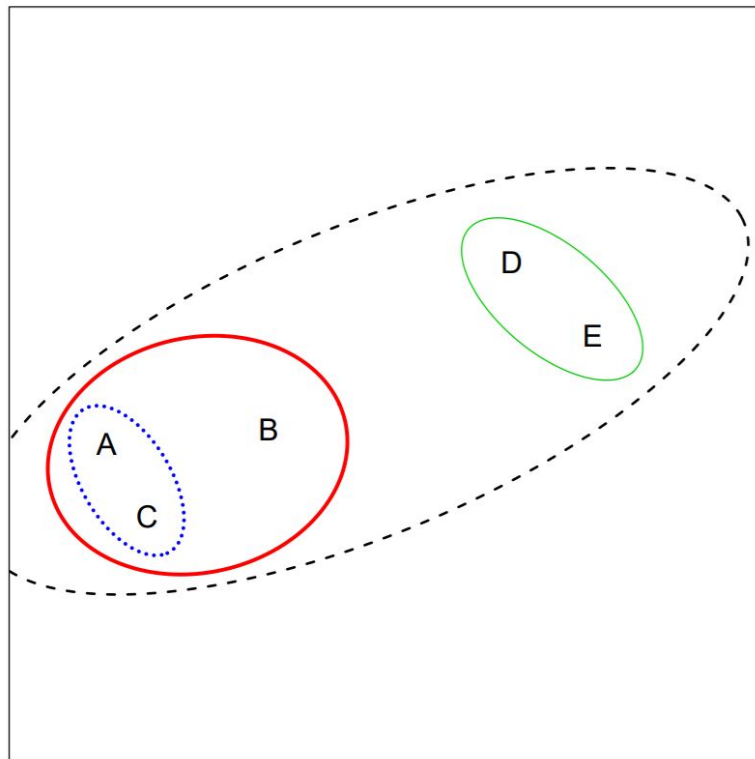
Clustering jerárquico



Clustering jerárquico



Clustering jerárquico



Clustering jerárquico

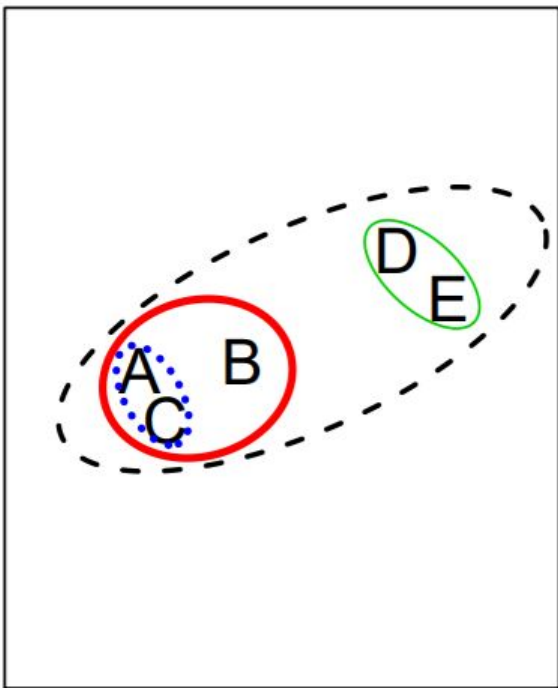
- Empieza con cada caso como un cluster único
- Identifica los clusters más cercanos y los une
- Repite el procedimiento
- Finaliza cuando todos los puntos han sido asignados a un único cluster



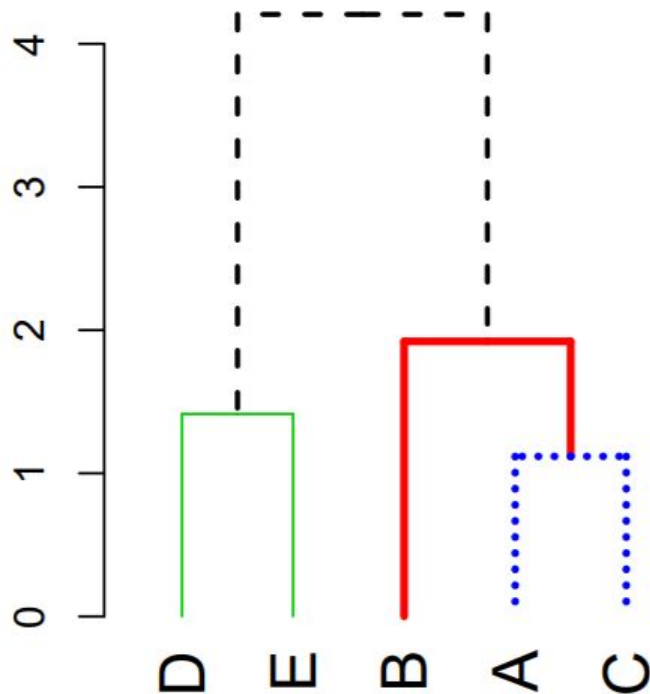
Clustering jerárquico



fa
EI

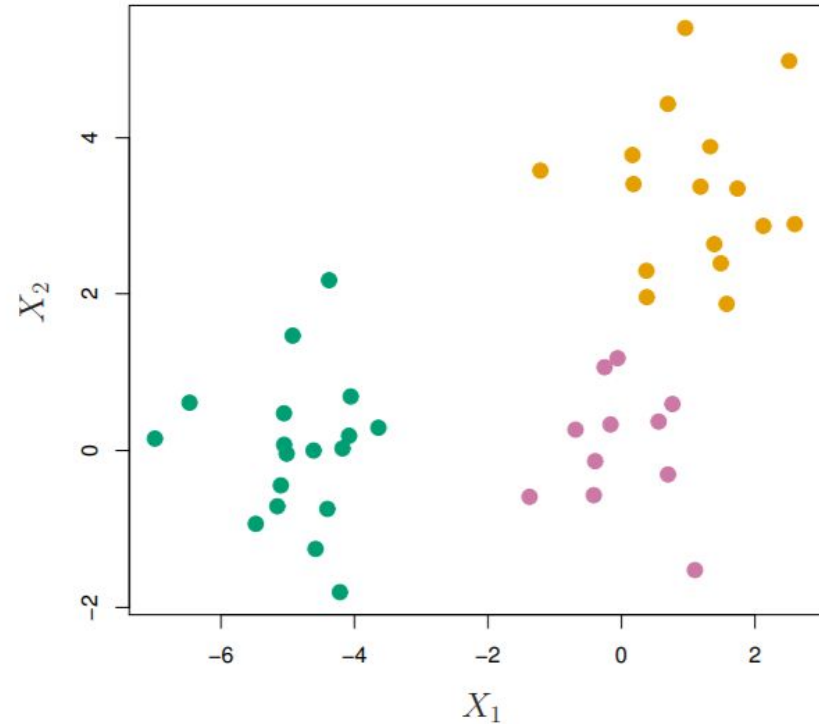


Dendrogram

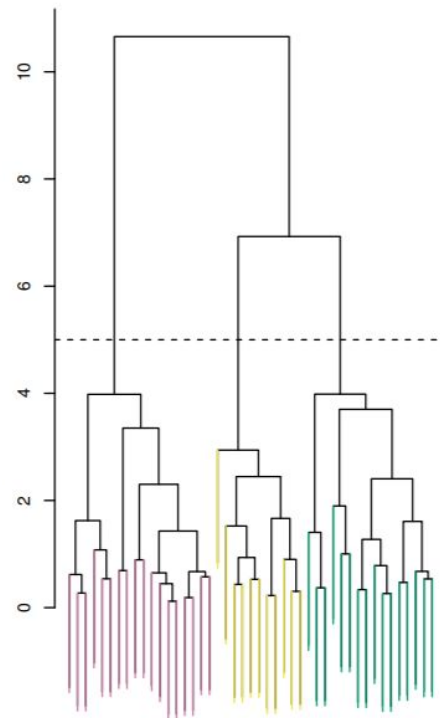
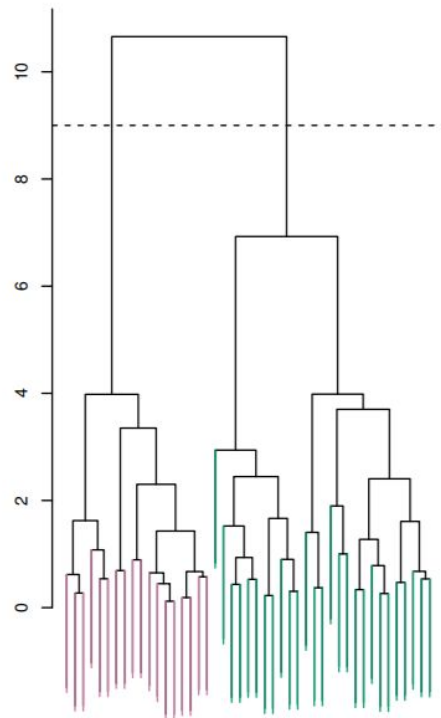
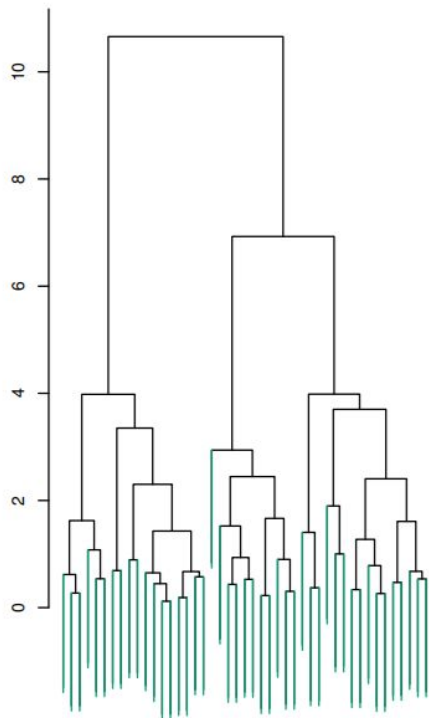


Clustering jerárquico

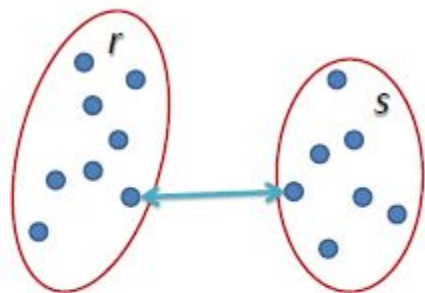
- Otro ejemplo



Clustering jerárquico

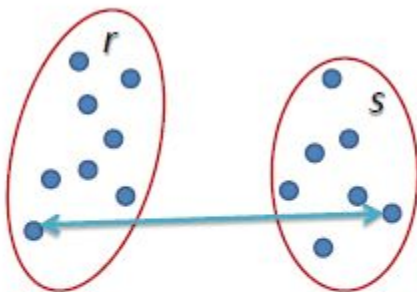


¿Cómo definir el método de aglomeración?



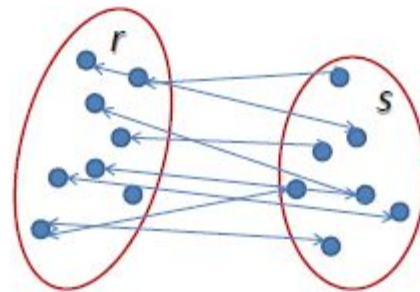
$$L(r, s) = \min(D(x_{ri}, x_{sj}))$$

Single Linkage



$$L(r, s) = \max(D(x_{ri}, x_{sj}))$$

Complete Linkage



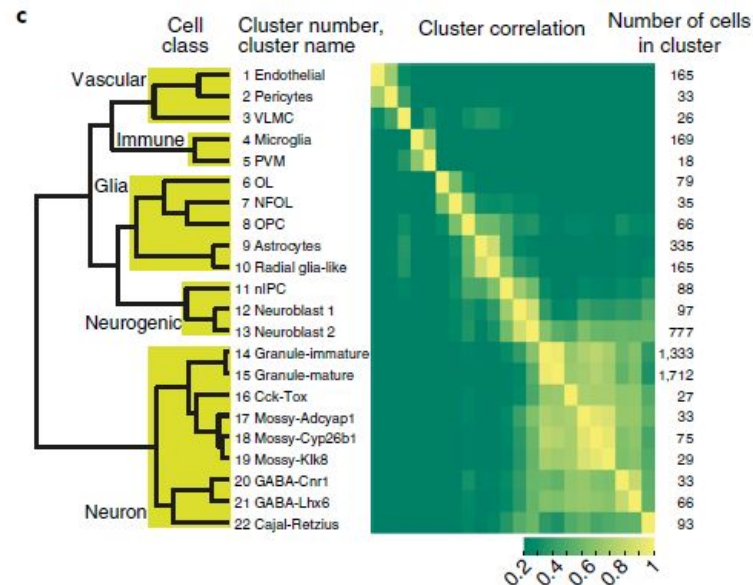
$$L(r, s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

Average Linkage



Clustering jerárquico. Ventajas y limitaciones

- + Pueden revelar detalles finos en la relación de los datos
- + Proveen un dendrograma interpretable
- + Son determinísticos - producen el mismo resultado si se corre el mismo modelo con el mismo input
- Son computacionalmente costosos



Problemas prácticos

- ¿Es necesario normalizar las escalas de las variables?
- ¿Qué métrica de similaridad usar?
- ¿Qué método de aglomeración?
- ¿Cuál es el número óptimo de clusters a utilizar?

