



PyData Granada

TERCER MEETUP

Patrocina: Novatec

Jueves 22 Junio
19:00 h

Paseo de la Bomba 5, 18008 Granada



Ponentes
Nuria Rico
Pablo Estévez

NUMFOCUS
OPEN CODE • BETTER SCIENCE



NOVATEC

Clasificación basada en datos

índice

- 1. Por qué queremos clasificar**
- 2. Cuánto se parecen dos cosas**
- 3. Métodos jerárquicos**
- 4. Métodos no jerárquicos**

índice

1. Por qué queremos clasificar

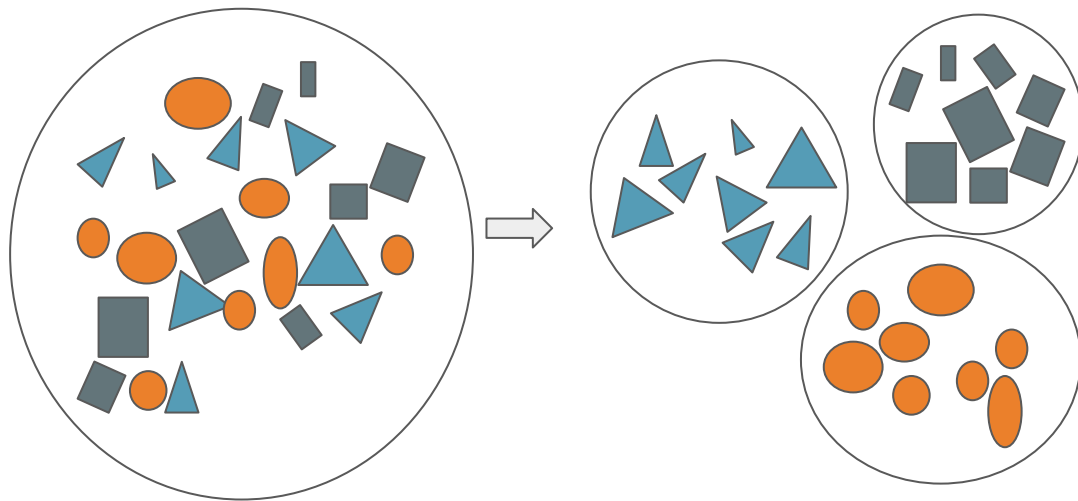
la vida es así, nos gusta unir las cosas y clasificar por colores, por formas, por nombre o por sabor es útil porque

- + ordena
- + permite optimizar
- + entendemos mejor



1. Por qué queremos clasificar

estamos hablando de clasificar en este sentido:



buscamos grupos exhaustivos y mutuamente excluyentes

1. Por qué queremos clasificar

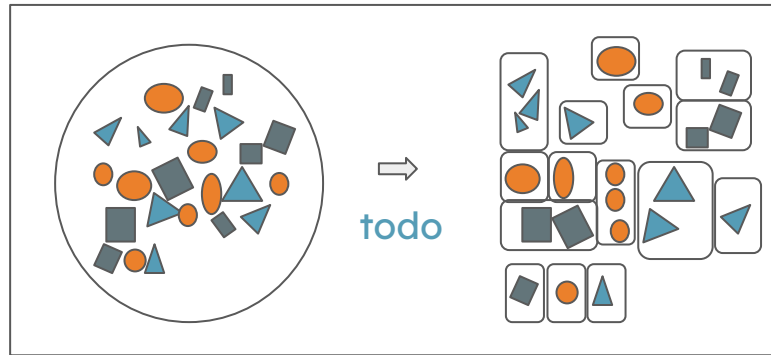
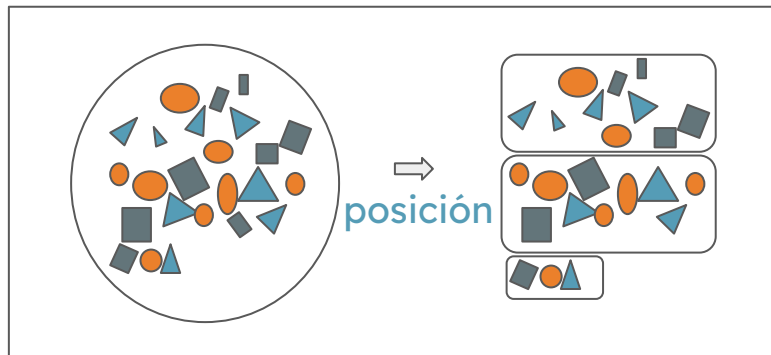
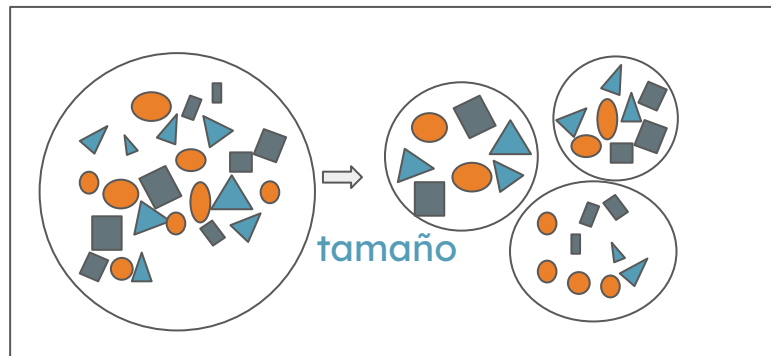
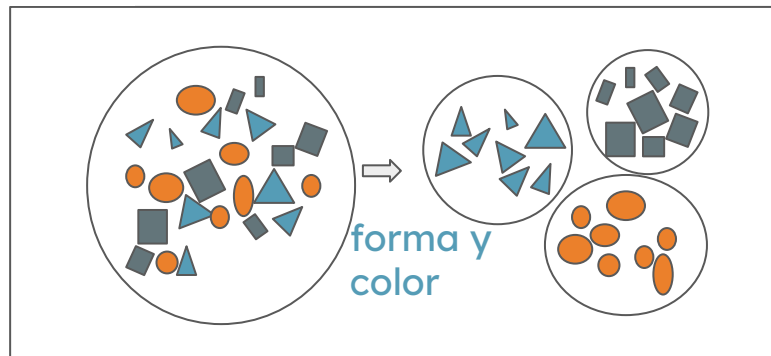
clasificar nos permite

- + identificar perfiles
- + establecer categorías
- + abstraer la realidad
- + conocer el entorno

aunque a veces nos lleva a

- + tener prejuicios
- + ver la parte como total
- + simplificar demasiado
- + obviar realidades

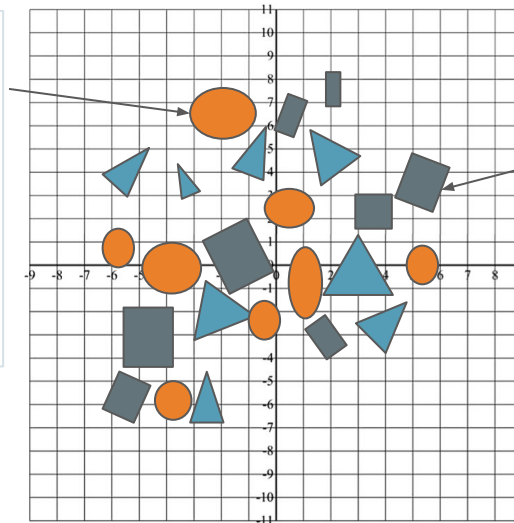
2. Cuánto se parecen dos cosas



2. Cuánto se parecen dos cosas

- + elegir el significado de la diferencia
- + saber qué y cómo observar (← datos)

color: “naranja”
forma: “óvalo”
ángulos: 0
tamaño: “grande”
área: 15.49
coordenadas: (-3,7)
cuadrante: segundo



color: “gris”
forma: “cuadrado”
ángulos: 4
tamaño: “mediano”
área: 13.92
coordenadas: (6,4)
cuadrante: primero

2. Cuánto se parecen dos cosas

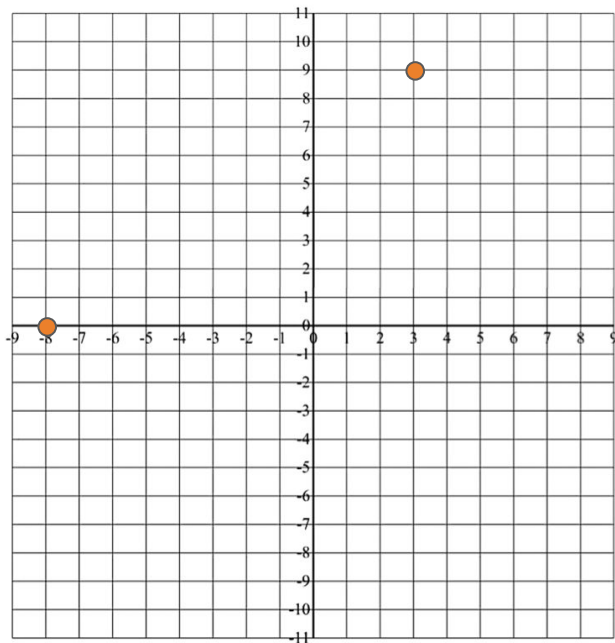
- + elegir el significado de la diferencia
- + saber qué y cómo observar (← datos)
- + normalizar las medidas si es preciso
(eliminar unidades de medida, centrar, tipificar)

2. Cuánto se parecen dos cosas

- + elegir el significado de la diferencia
- + saber qué y cómo observar (← datos)
- + normalizar las medidas si es preciso
- + calcular la distancia o la similaridad

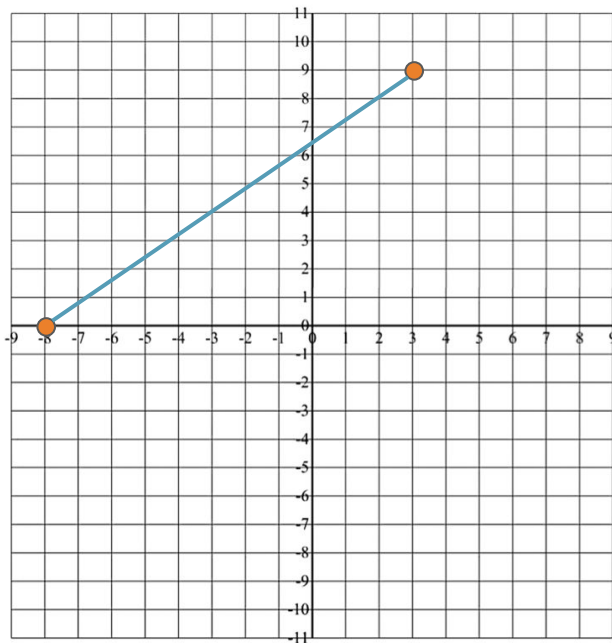
2. Cuánto se parecen dos cosas

+ distancia



2. Cuánto se parecen dos cosas

+ distancia

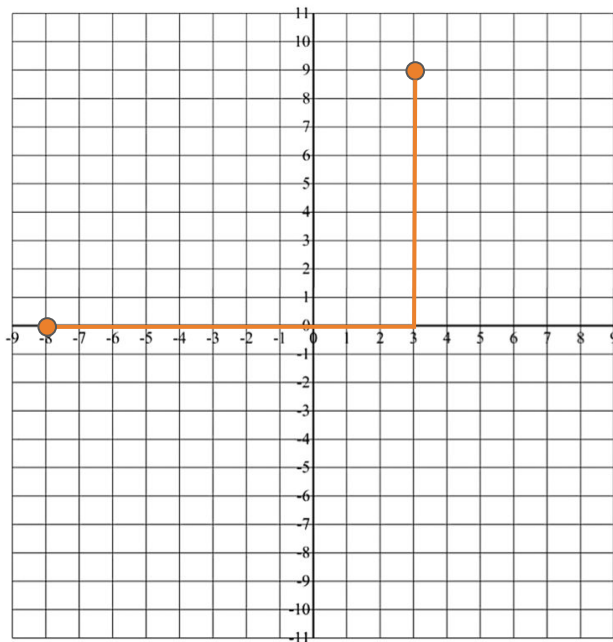


distancia euclídea

$$d_2(\mathbf{x}_r, \mathbf{x}_s) = \sqrt{\sum_{h=1}^p (x_{rh} - x_{sh})^2}$$

2. Cuánto se parecen dos cosas

+ distancia



distancia euclídea

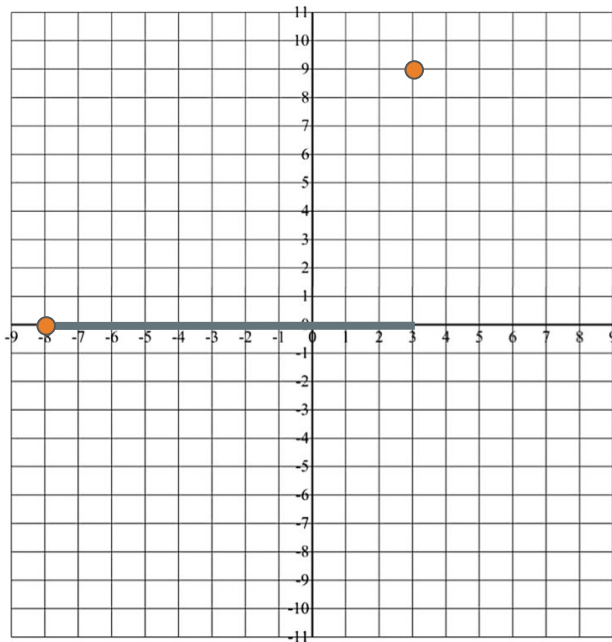
$$d_2(\mathbf{x}_r, \mathbf{x}_s) = \sqrt{\sum_{h=1}^p (x_{rh} - x_{sh})^2}$$

distancia Manhattan

$$d_1(\mathbf{x}_r, \mathbf{x}_s) = \sum_{h=1}^p |x_{rh} - x_{sh}|$$

2. Cuánto se parecen dos cosas

+ distancia



distancia euclídea

$$d_2(\mathbf{x}_r, \mathbf{x}_s) = \sqrt{\sum_{h=1}^p (x_{rh} - x_{sh})^2}$$

distancia Manhattan

$$d_1(\mathbf{x}_r, \mathbf{x}_s) = \sum_{h=1}^p |x_{rh} - x_{sh}|$$

distancia del máximo

$$d_\infty(\mathbf{x}_r, \mathbf{x}_s) = \max_{h=1, \dots, p} \{|x_{rh} - x_{sh}|\}$$

2. Cuánto se parecen dos cosas

+ similaridad

$\mathbf{x_r} / \mathbf{x_s}$	Presencia (1)	Ausencia (0)	Total
Presencia (1)	a	b	$a + b$
Ausencia (0)	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d = m$

2. Cuánto se parecen dos cosas

+ similaridad

$\mathbf{x_r} / \mathbf{x_s}$	Presencia (1)	Ausencia (0)	Total
Presencia (1)	a	b	$a + b$
Ausencia (0)	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d = m$

Russell y Rao $Sim(\mathbf{x_r}, \mathbf{x_s}) = \frac{a}{m}$

2. Cuánto se parecen dos cosas

+ similaridad

$\mathbf{x}_r / \mathbf{x}_s$	Presencia (1)	Ausencia (0)	Total
Presencia (1)	a	b	$a + b$
Ausencia (0)	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d = m$

Russell y Rao $Sim(\mathbf{x}_r, \mathbf{x}_s) = \frac{a}{m}$

Parejas simples $Sim(\mathbf{x}_r, \mathbf{x}_s) = \frac{a + d}{m}$

2. Cuánto se parecen dos cosas

+ similaridad

$\mathbf{x}_r / \mathbf{x}_s$	Presencia (1)	Ausencia (0)	Total
Presencia (1)	a	b	$a + b$
Ausencia (0)	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d = m$

Russell y Rao $Sim(\mathbf{x}_r, \mathbf{x}_s) = \frac{a}{m}$

Parejas simples $Sim(\mathbf{x}_r, \mathbf{x}_s) = \frac{a + d}{m}$

Jaccard $Sim(\mathbf{x}_r, \mathbf{x}_s) = \frac{a}{a + b + c}$

2. Cuánto se parecen dos cosas

+ matriz de distancias o similaridades

$$\begin{pmatrix} d_{11} & d_{12} & d_{13} & \dots & d_{1k} \\ d_{21} & d_{22} & d_{23} & \dots & d_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{k1} & d_{k2} & d_{k3} & \dots & d_{kk} \end{pmatrix}$$

simétrica

2. Cuánto se parecen dos cosas

+ matriz de distancias o similaridades

$$\begin{pmatrix} d_{11} & d_{12} & d_{13} & \dots & d_{1k} \\ d_{21} & d_{22} & d_{23} & \dots & d_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{k1} & d_{k2} & d_{k3} & \dots & d_{kk} \end{pmatrix}$$

simétrica
diagonal

2. Cuánto se parecen dos cosas

+ matriz de distancias o similaridades

$$\begin{pmatrix} d_{11} & d_{12} & d_{13} & \dots & d_{1k} \\ d_{21} & d_{22} & d_{23} & \dots & d_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{k1} & d_{k2} & d_{k3} & \dots & d_{kk} \end{pmatrix}$$

simétrica
diagonal
cuadrada

- + método de agrupación
 - + jerárquico - no se conoce el número de grupos; se obtiene un árbol (dendrograma)
 - + no jerárquico - se basa en una partición inicial, se debe fijar el número de grupos

3. Métodos jerárquicos

realizan la clasificación paso a paso;
bien van haciendo grupos
cada vez más numerosos (aglomerativos)
o bien van partiendo grupos numerosos para
tener cada vez más grupos (disociativos)

3. Métodos jerárquicos

realizan la clasificación paso a paso;
bien van haciendo grupos
cada vez más numerosos (aglomerativos)
o bien van partiendo grupos numerosos para
tener cada vez más grupos (disociativos)

3. Métodos jerárquicos

$$\begin{pmatrix} d_{11} & d_{12} & d_{13} & \dots & d_{1k} \\ d_{21} & d_{22} & d_{23} & \dots & d_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{k1} & d_{k2} & d_{k3} & \dots & d_{kk} \end{pmatrix}$$

mínimo

los objetos 2 y 3 son los más parecidos entre sí:

2 y 3 se **unen** en un grupo

3. Métodos jerárquicos

$$\begin{pmatrix} d_{11} & d_{12} & d_{13} & \dots & d_{1k} \\ d_{21} & d_{22} & d_{23} & \dots & d_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{k1} & d_{k2} & d_{k3} & \dots & d_{kk} \end{pmatrix}$$

mínimo

los objetos 2 y 3 son los más parecidos entre sí:

2 y 3 se **unen** en un grupo

3. Métodos jerárquicos

$$\begin{pmatrix} d_{11} & d_{12} & d_{13} & \dots & d_{1k} \\ d_{21} & d_{22} & d_{23} & \dots & d_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{k1} & d_{k2} & d_{k3} & \dots & d_{kk} \end{pmatrix} \quad \begin{pmatrix} d_{11} & d_{1(23)} & d_{14} & \dots & d_{1k} \\ d_{(23)1} & d_{(23)(23)} & d_{(23)4} & \dots & d_{(23)k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{k1} & d_{k(23)} & d_{k4} & \dots & d_{kk} \end{pmatrix}$$



uniendo los elementos 2 y 3

3. Métodos jerárquicos

$$\begin{pmatrix} d_{11} & d_{1(23)} & d_{14} & \dots & d_{1k} \\ d_{(23)1} & d_{(23)(23)} & d_{(23)4} & \dots & d_{(23)k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{k1} & d_{k(23)} & d_{k4} & \dots & d_{kk} \end{pmatrix}$$

mínimo

los objetos 1 y k son los más parecidos entre sí:

1 y k se **unen** en un grupo

3. Métodos jerárquicos


$$\begin{pmatrix} d_{11} & d_{1(23)} & d_{14} & \dots & d_{1k} \\ d_{(23)1} & d_{(23)(23)} & d_{(23)4} & \dots & d_{(23)k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{k1} & d_{k(23)} & d_{k4} & \dots & d_{kk} \end{pmatrix} \quad \begin{pmatrix} d_{(1k)(1k)} & d_{(1k)(23)} & d_{(1k)4} & \dots & d_{(1k)(k-1)} \\ d_{(23)(1k)} & d_{(23)(23)} & d_{(23)4} & \dots & d_{(23)(k-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{(k-1)(1k)} & d_{(k-1)(23)} & d_{(k-1)4} & \dots & d_{(k-1)(k-1)} \end{pmatrix}$$



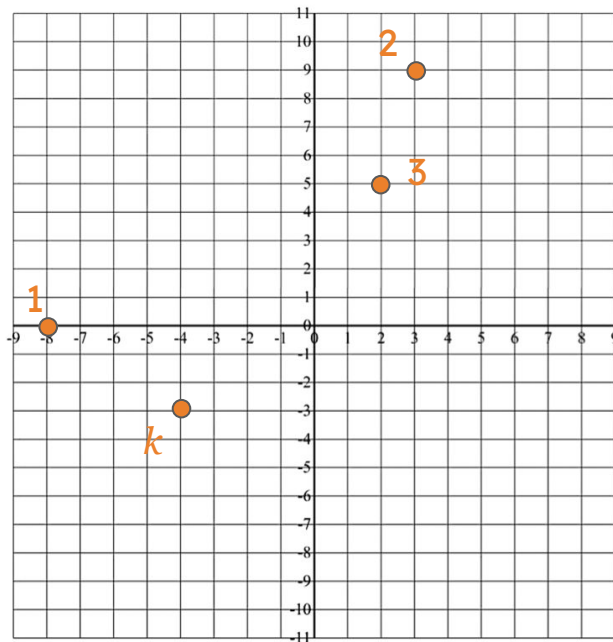
uniendo los elementos 1 y k

3. Métodos jerárquicos

- + En cada paso se unen los grupos más cercanos
- + Se obtiene una matriz de distancias de una dimensión menos
- + Se deben calcular distancias entre grupos

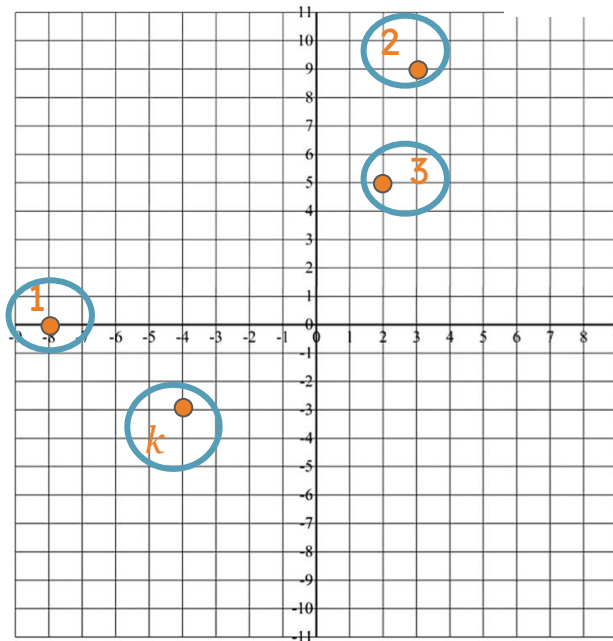
$$\begin{pmatrix} d_{11} & d_{12} & d_{13} & \dots & d_{1k} \\ d_{21} & d_{22} & d_{23} & \dots & d_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{k1} & d_{k2} & d_{k3} & \dots & d_{kk} \end{pmatrix} \rightarrow \begin{pmatrix} d_{11} & d_{1(23)} & d_{14} & \dots & d_{1k} \\ d_{(23)1} & d_{(23)(23)} & d_{(23)4} & \dots & d_{(23)k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{k1} & d_{k(23)} & d_{k4} & \dots & d_{kk} \end{pmatrix}$$


3. Métodos jerárquicos



3. Métodos jerárquicos

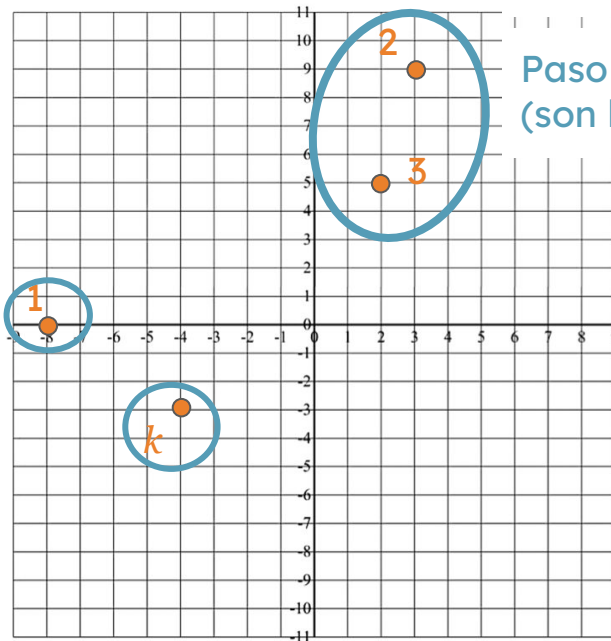
Paso 0: cada elemento es un grupo



3. Métodos jerárquicos

Paso 0: cada elemento es un grupo

Paso 1: unimos los elementos 2 y 3
(son los más cercanos)



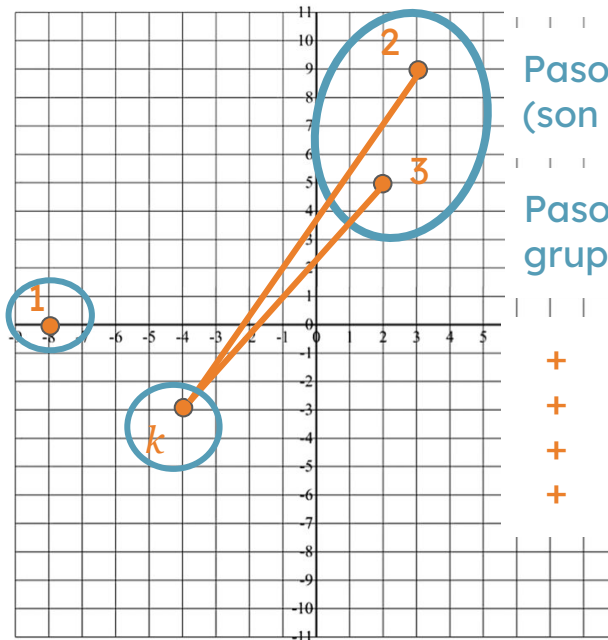
3. Métodos jerárquicos

Paso 0: cada elemento es un grupo

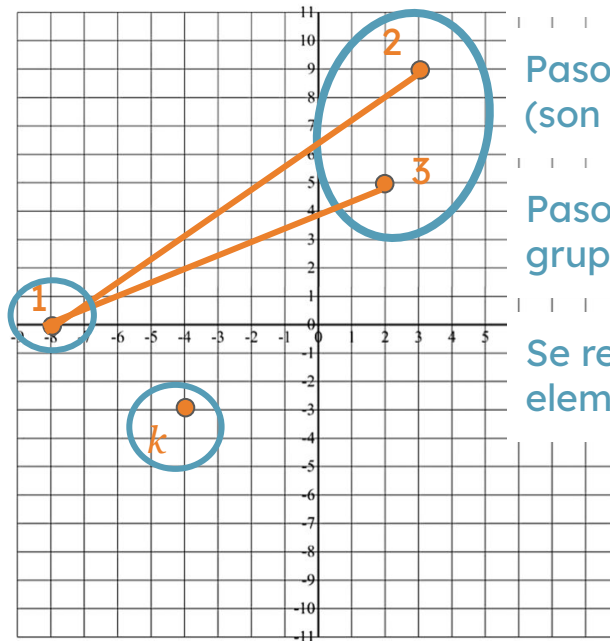
Paso 1: unimos los elementos 2 y 3
(son los más cercanos)

Paso 2: calculamos la distancia del
grupo 23 al elemento k

- + la menor posible (simple)
- + la mayor posible (completo)
- + media aritmética (promedio)
- + distancia entre centroides



3. Métodos jerárquicos



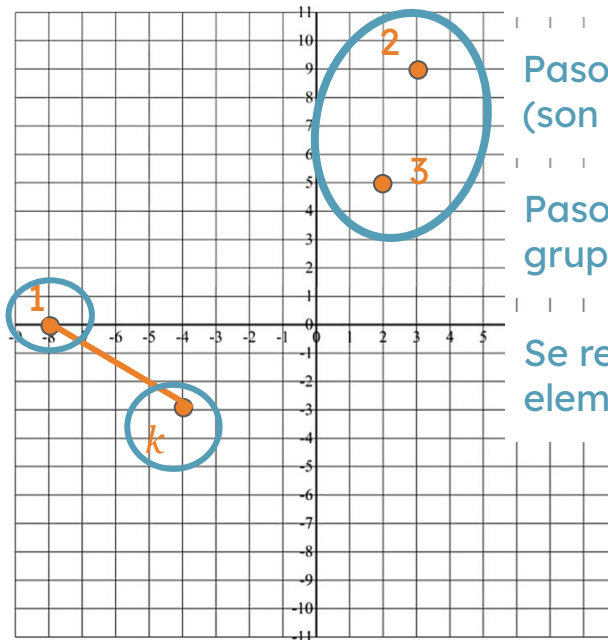
Paso 0: cada elemento es un grupo

Paso 1: unimos los elementos 2 y 3
(son los más cercanos)

Paso 2: calculamos la distancia del
grupo 23 al elemento k

Se repite paso 2 con todos los
elementos para obtener la matriz

3. Métodos jerárquicos



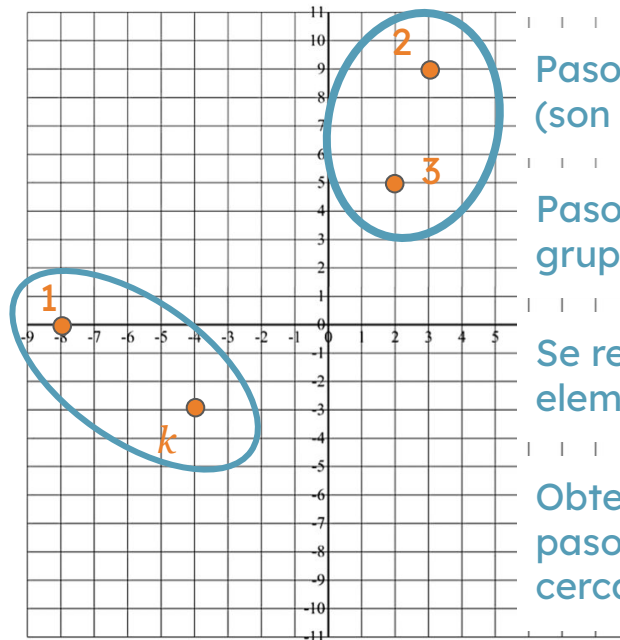
Paso 0: cada elemento es un grupo

Paso 1: unimos los elementos 2 y 3
(son los más cercanos)

Paso 2: calculamos la distancia del
grupo 23 al elemento k

Se repite paso 2 con todos los
elementos para obtener la matriz

3. Métodos jerárquicos



Paso 0: cada elemento es un grupo

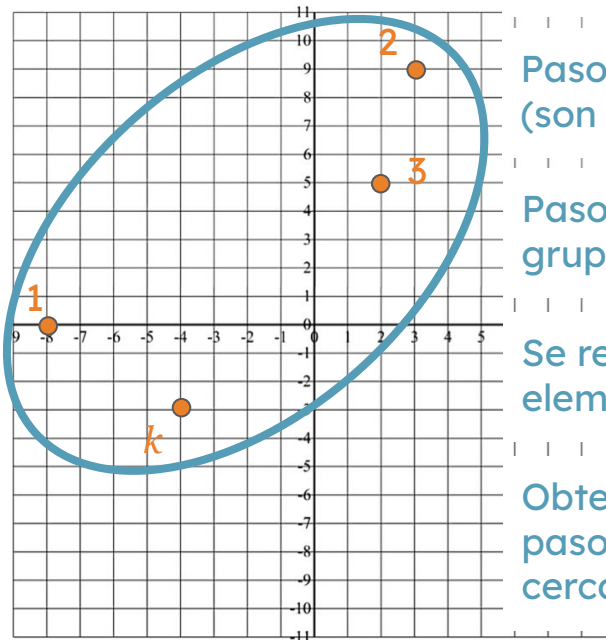
Paso 1: unimos los elementos 2 y 3
(son los más cercanos)

Paso 2: calculamos la distancia del
grupo 23 al elemento k

Se repite paso 2 con todos los
elementos para obtener la matriz

Obtenida la matriz, se vuelve al
paso 1 para unir los elementos más
ceranos

3. Métodos jerárquicos



Paso 0: cada elemento es un grupo

Paso 1: unimos los elementos 2 y 3
(son los más cercanos)

Paso 2: calculamos la distancia del
grupo 23 al elemento k

Se repite paso 2 con todos los
elementos para obtener la matriz

Obtenida la matriz, se vuelve al
paso 1 para unir los elementos más
ceranos

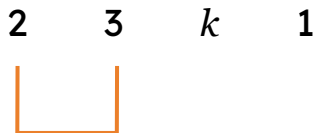
3. Métodos jerárquicos

dendrograma

2 3 k 1

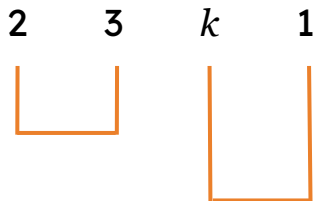
3. Métodos jerárquicos

dendrograma



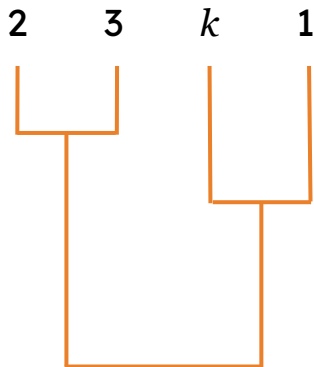
3. Métodos jerárquicos

dendrograma



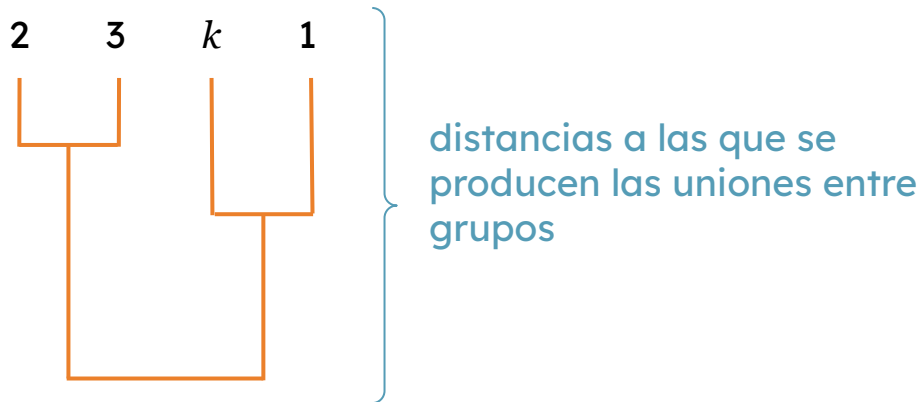
3. Métodos jerárquicos

dendrograma



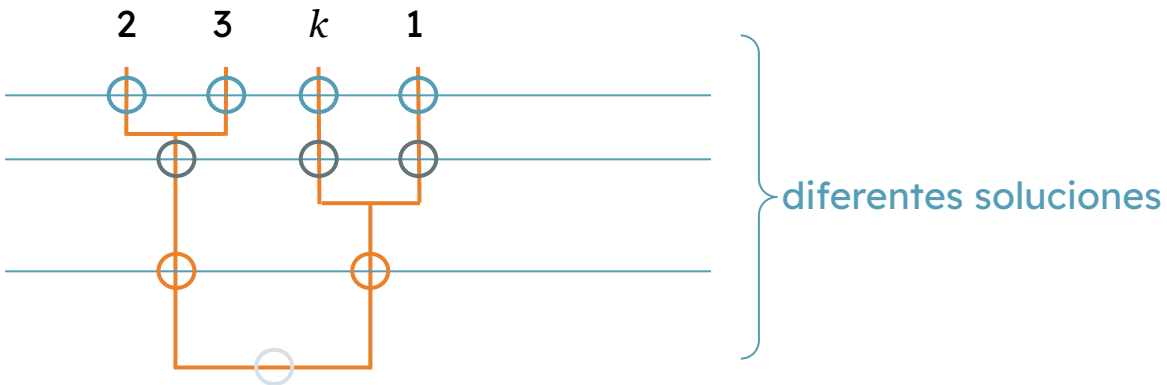
3. Métodos jerárquicos

dendrograma



3. Métodos jerárquicos

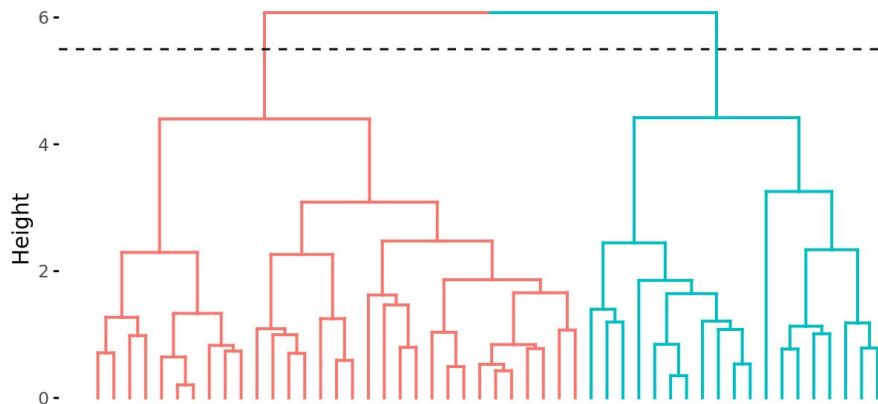
dendrograma



3. Métodos jerárquicos

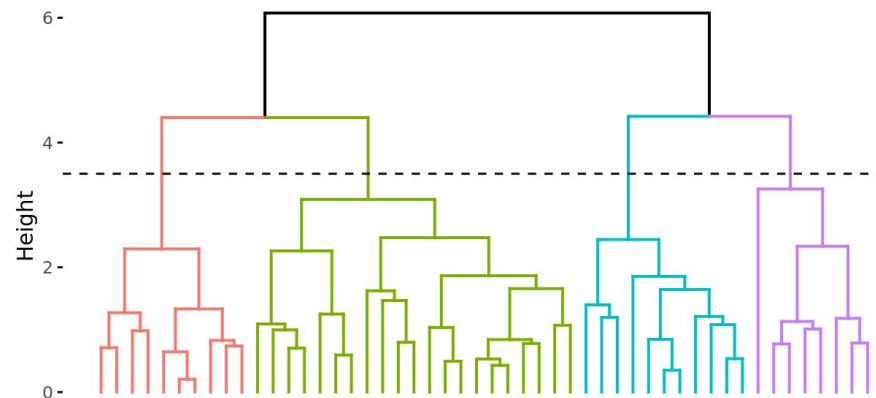
Herarchical clustering

Distancia euclídea, Linkage complete, K=2



Herarchical clustering

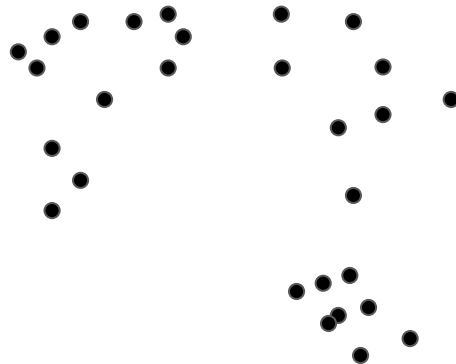
Distancia euclídea, Linkage complete, K=4



4. Métodos no jerárquicos

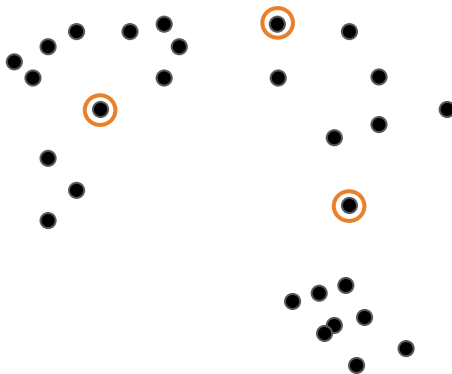
los métodos no jerárquicos hacen una partición
y después se itera el procedimiento
para obtener una partición mejor

4. Métodos no jerárquicos



4. Métodos no jerárquicos

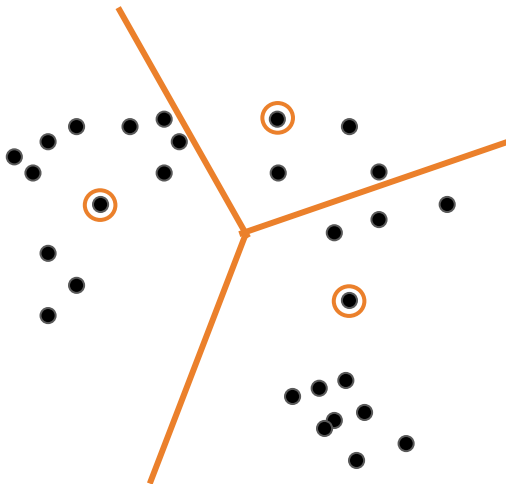
0. se proponen k medias



4. Métodos no jerárquicos

0. se proponen k medias

1. se asigna cada elemento
al centro más cercano

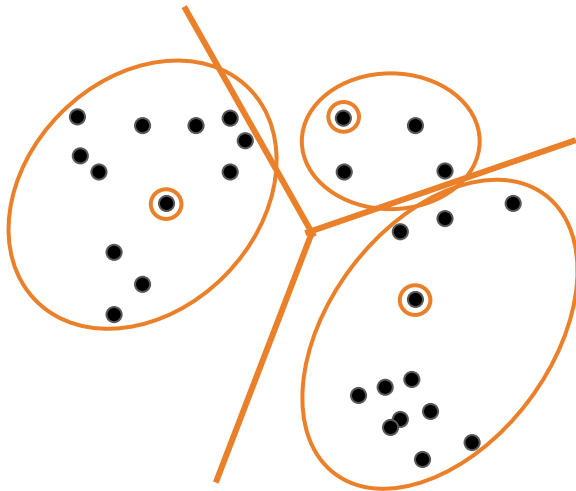


4. Métodos no jerárquicos

0. se proponen k medias

1. se asigna cada elemento al centro más cercano

2. así se forman los grupos



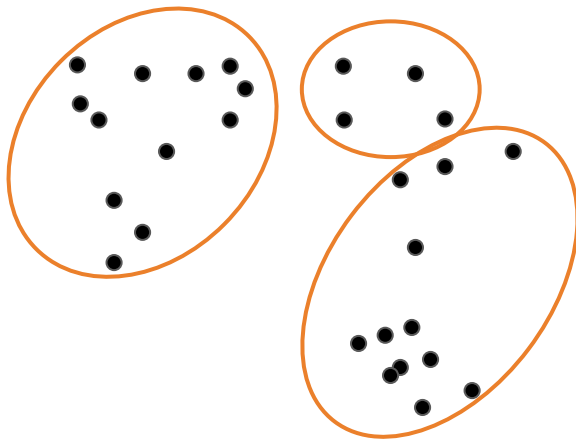
4. Métodos no jerárquicos

0. se proponen k medias

1. se asigna cada elemento al centro más cercano

2. así se forman los grupos

3. olvidamos los centros iniciales



4. Métodos no jerárquicos

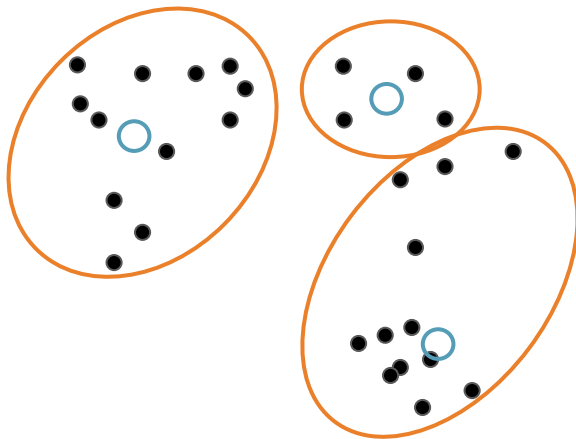
0. se proponen k medias

1. se asigna cada elemento al centro más cercano

2. así se forman los grupos

3. olvidamos los centros iniciales

4. calculamos los centros de los grupos



4. Métodos no jerárquicos

0. se proponen k medias

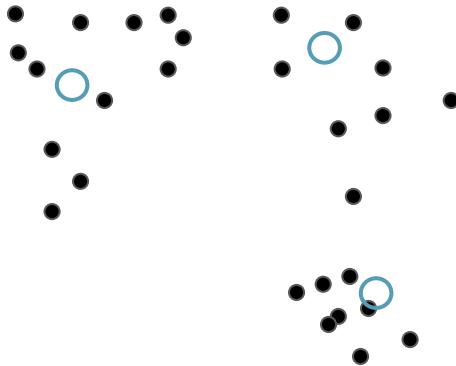
1. se asigna cada elemento al centro más cercano

2. así se forman los grupos

3. olvidamos los centros iniciales

4. calculamos los centros de los grupos

5. volvemos al paso 1



4. Métodos no jerárquicos

0. se proponen k medias

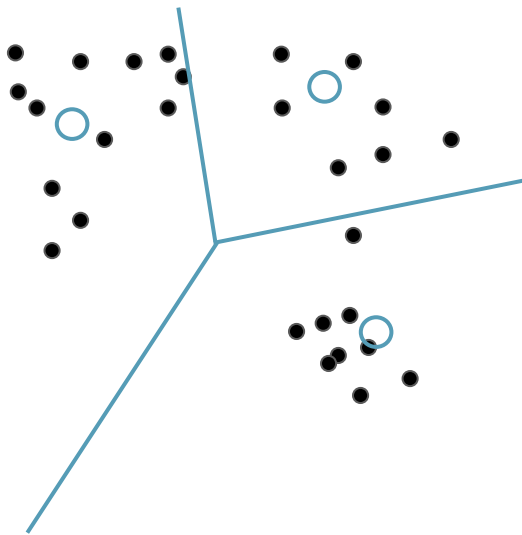
1. se asigna cada elemento al centro más cercano

2. así se forman los grupos

3. olvidamos los centros iniciales

4. calculamos los centros de los grupos

5. volvemos al paso 1



4. Métodos no jerárquicos

0. se proponen k medias

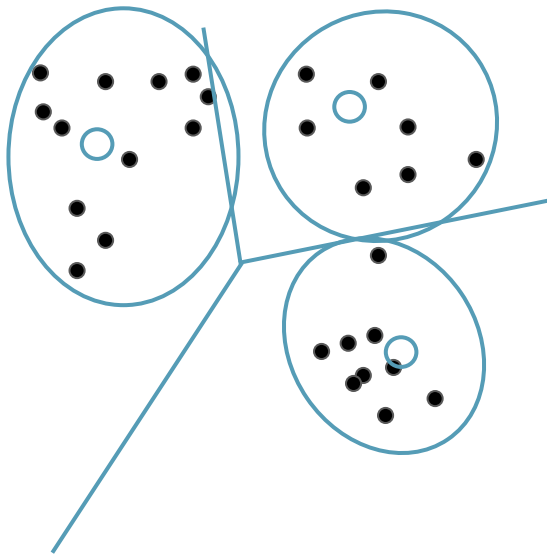
1. se asigna cada elemento al centro más cercano

2. así se forman los grupos

3. olvidamos los centros iniciales

4. calculamos los centros de los grupos

5. volvemos al paso 1



4. Métodos no jerárquicos

0. se proponen k medias

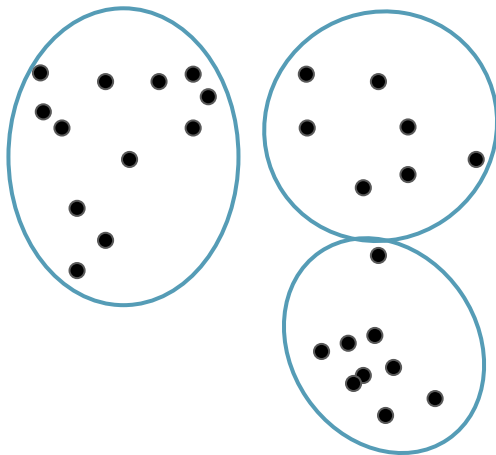
1. se asigna cada elemento al centro más cercano

2. así se forman los grupos

3. olvidamos los centros iniciales

4. calculamos los centros de los grupos

5. volvemos al paso 1



4. Métodos no jerárquicos

0. se proponen k medias

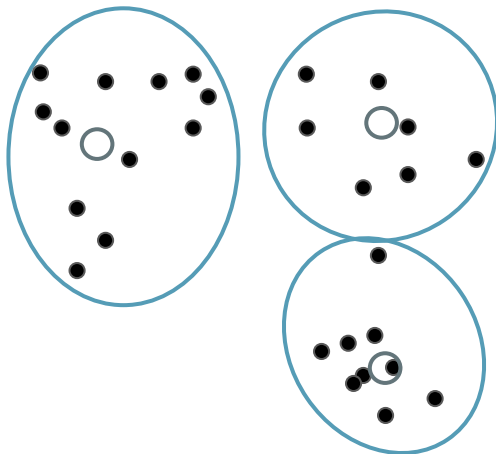
1. se asigna cada elemento al centro más cercano

2. así se forman los grupos

3. olvidamos los centros iniciales

4. calculamos los centros de los grupos

5. volvemos al paso 1



4. Métodos no jerárquicos

0. se proponen k medias

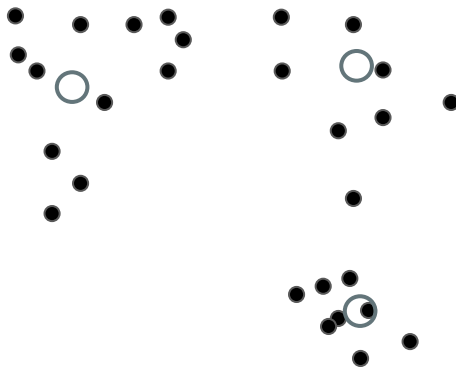
1. se asigna cada elemento al centro más cercano

2. así se forman los grupos

3. olvidamos los centros iniciales

4. calculamos los centros de los grupos

5. volvemos al paso 1



4. Métodos no jerárquicos

0. se proponen k medias

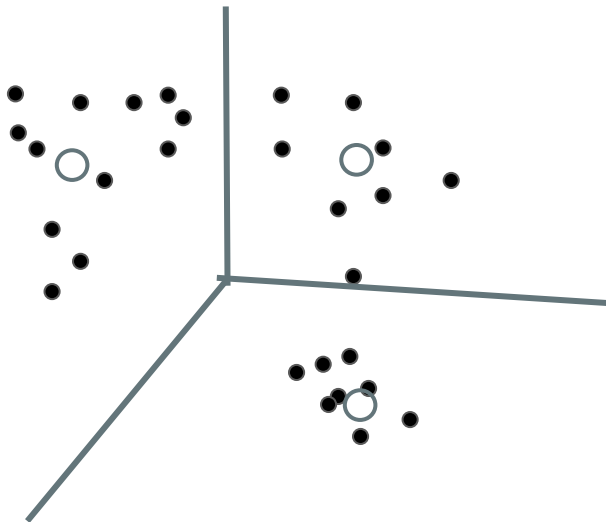
1. se asigna cada elemento al centro más cercano

2. así se forman los grupos

3. olvidamos los centros iniciales

4. calculamos los centros de los grupos

5. volvemos al paso 1



4. Métodos no jerárquicos

0. se proponen k medias

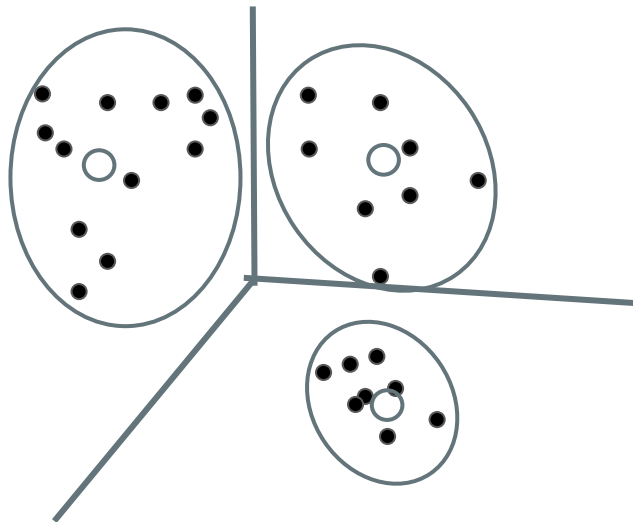
1. se asigna cada elemento al centro más cercano

2. así se forman los grupos

3. olvidamos los centros iniciales

4. calculamos los centros de los grupos

5. volvemos al paso 1



4. Métodos no jerárquicos

0. se proponen k medias

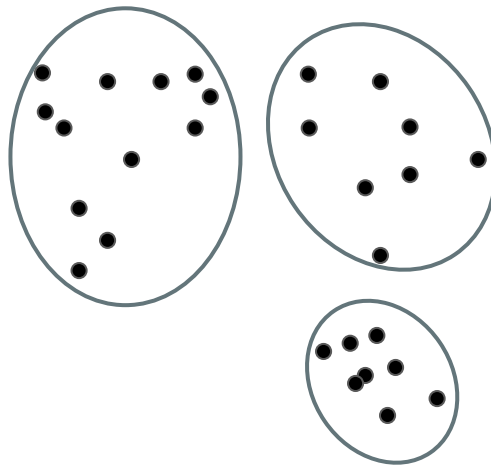
1. se asigna cada elemento al centro más cercano

2. así se forman los grupos

3. olvidamos los centros iniciales

4. calculamos los centros de los grupos

5. volvemos al paso 1



4. Métodos no jerárquicos

0. se proponen k medias

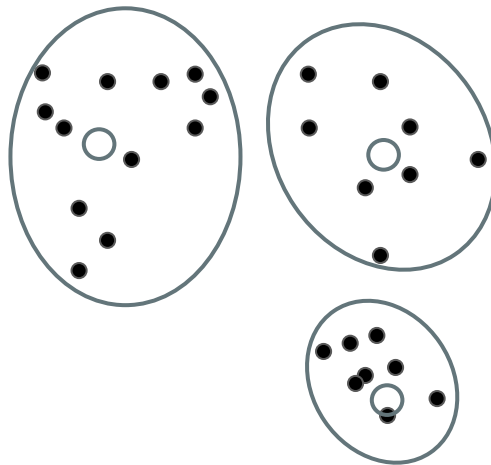
1. se asigna cada elemento al centro más cercano

2. así se forman los grupos

3. olvidamos los centros iniciales

4. calculamos los centros de los grupos

5. volvemos al paso 1



6. hasta que los centros no cambien o se alcance un criterio de parada (tolerancia o número máximo de iteraciones)

4. Métodos no jerárquicos

- + variaciones
 - + no usar la media sino el medoide
 - + recalcular tras cada asignación
 - + elegir los primeros centros según método
 - + ...

FIN



PyData
Granada

gracias