

Aprendizaje no supervisado

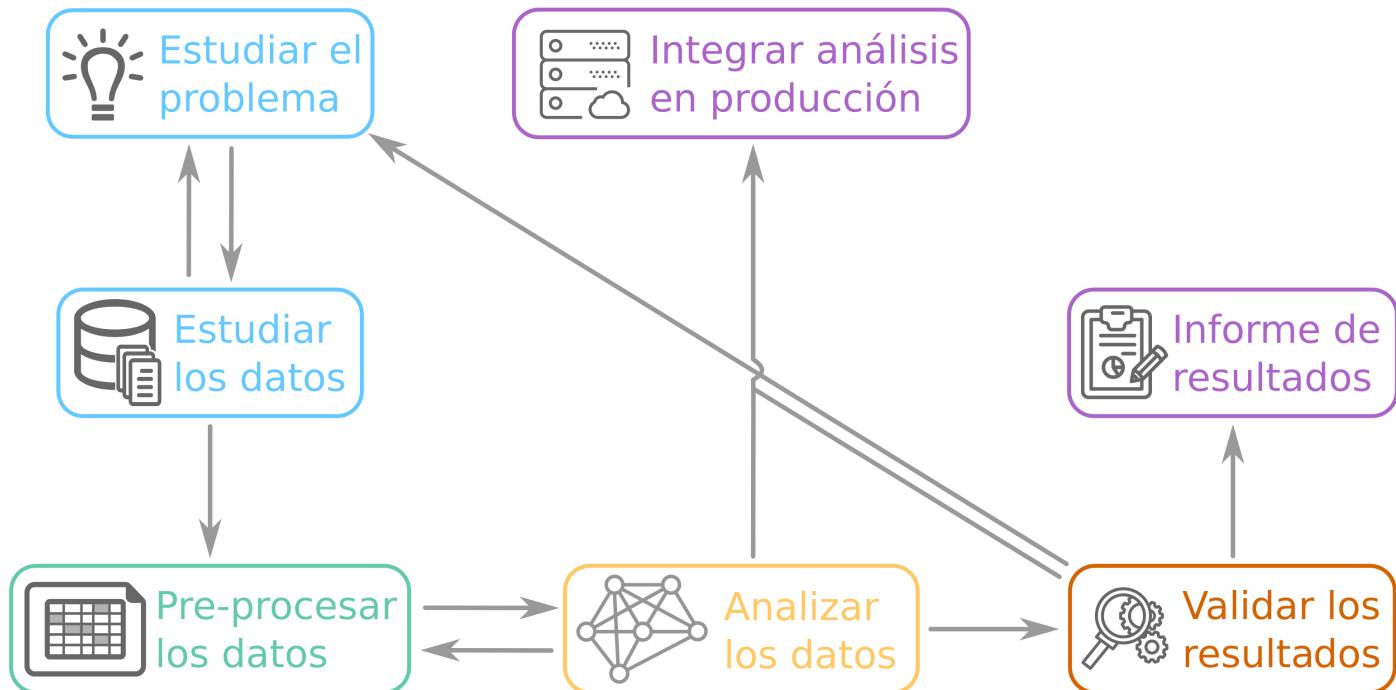
VC01: Introducción al aprendizaje automático

Rocío del Amor

mrocio.delamor@campusviu.es

Universidad Internacional de Valencia

Aprendizaje automático



Aprendizaje automático



**Aprendizaje
Supervisado**

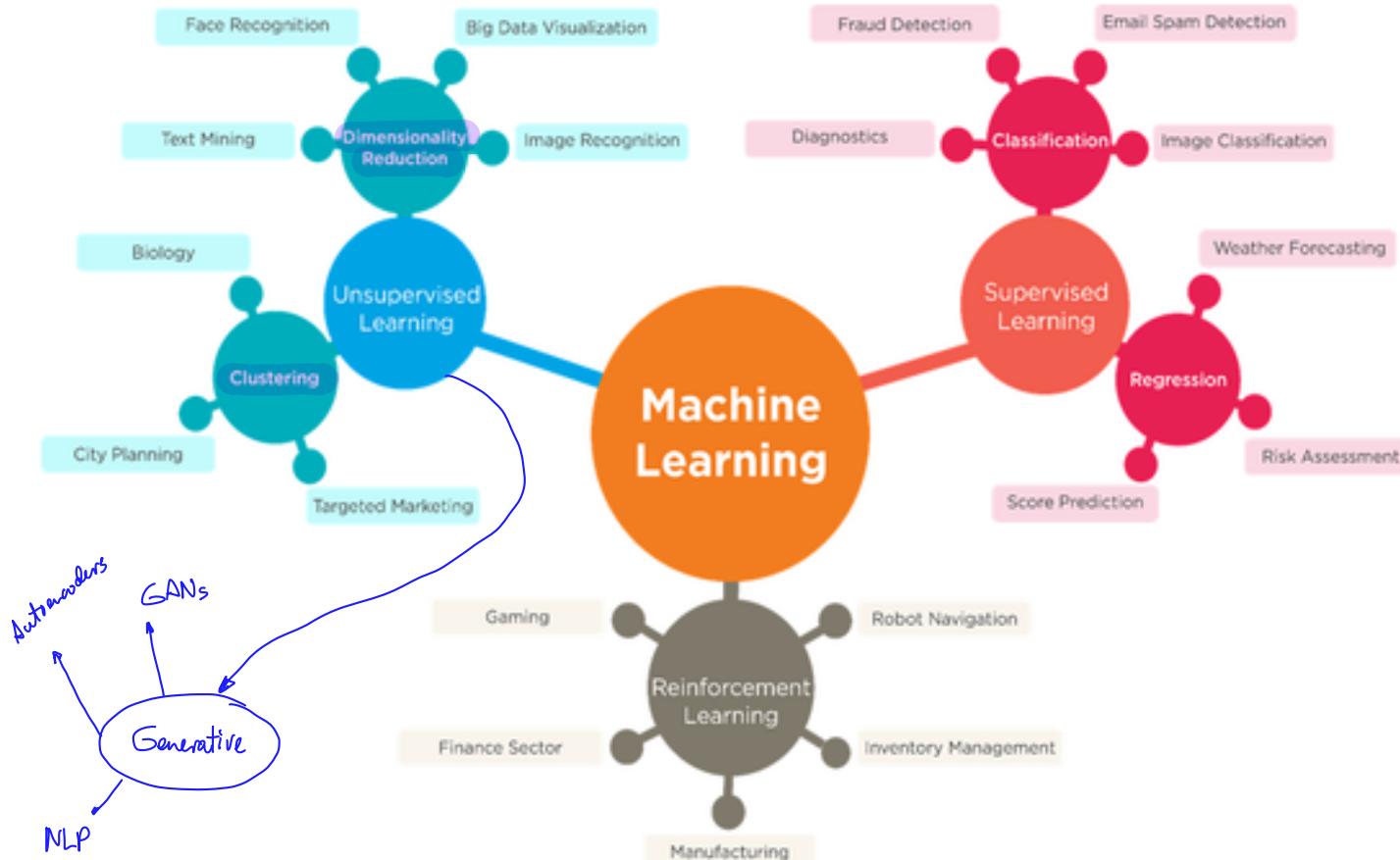


**Aprendizaje
No Supervisado**



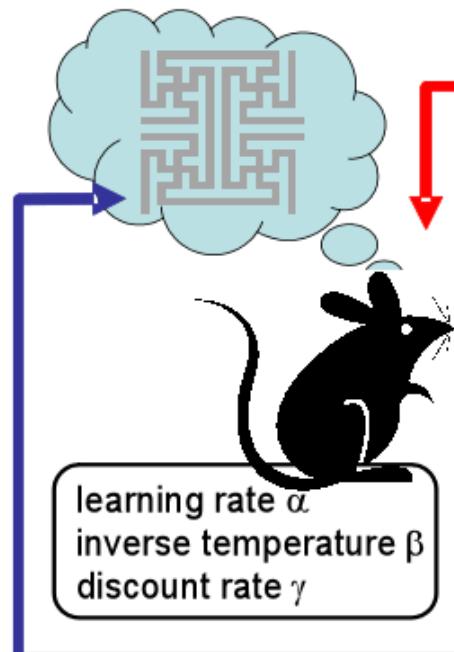
**Aprendizaje
Por Refuerzo**

Aprendizaje automático



Aprendizaje por refuerzo

internal state



reward

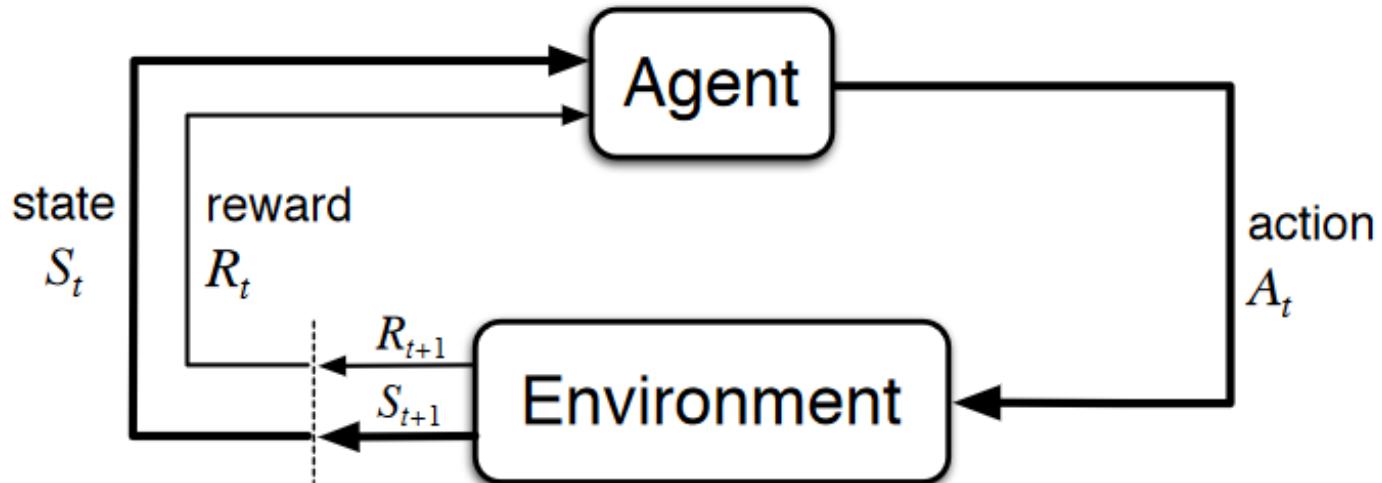
environment



action

observation

Aprendizaje por refuerzo



Aprendizaje por refuerzo

Ejemplo: Fábrica

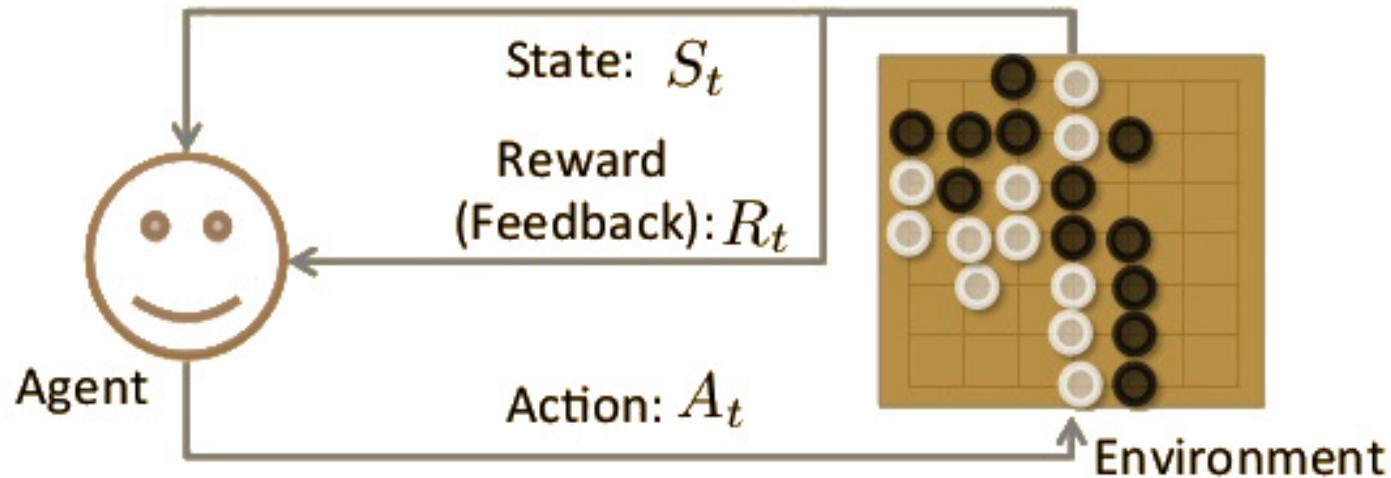
Configuración de un brazo robótico

¿Cuál es exactamente el mejor movimiento del brazo para optimizar la tarea?

Aprendizaje por refuerzo

https://www.youtube.com/watch?v=ZVIxt2rtl_4

Aprendizaje por refuerzo

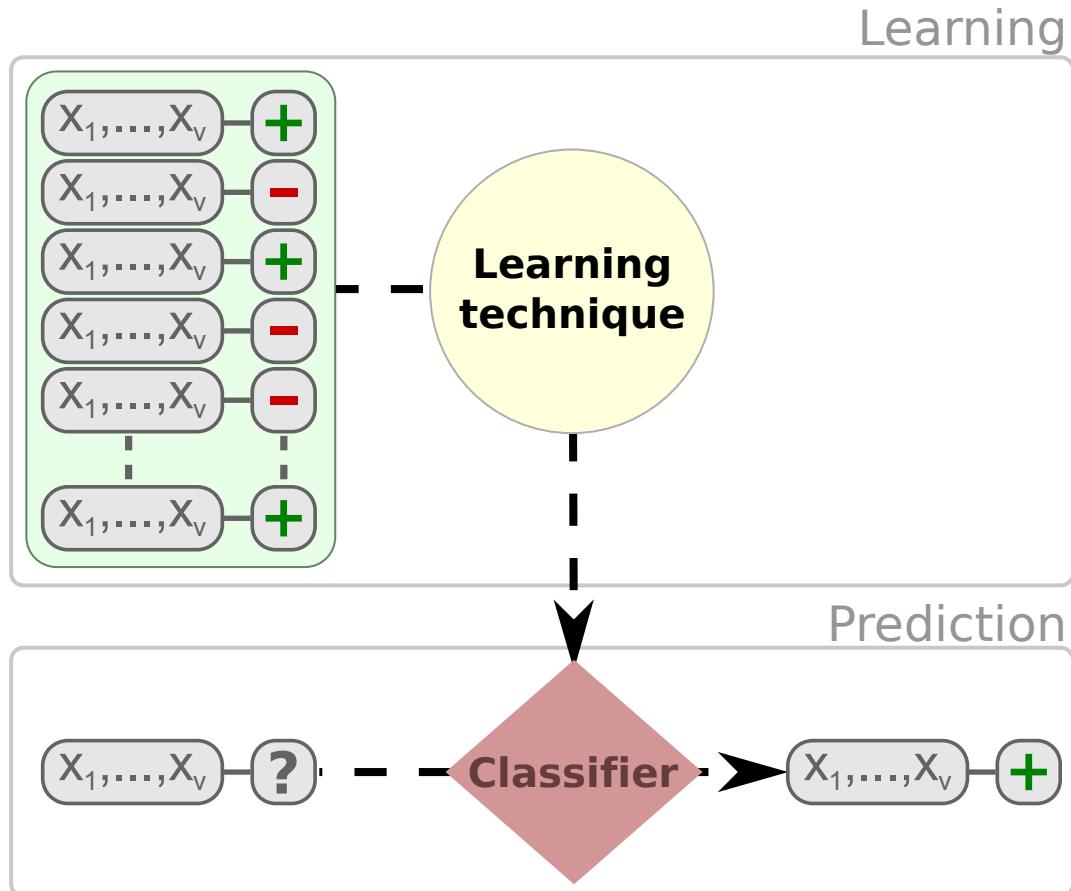


Aprendizaje por refuerzo

<https://www.youtube.com/watch?v=n9K3zJNf75Q>

Aprendizaje supervisado

Clasificación



Aprendizaje supervisado

Clasificación

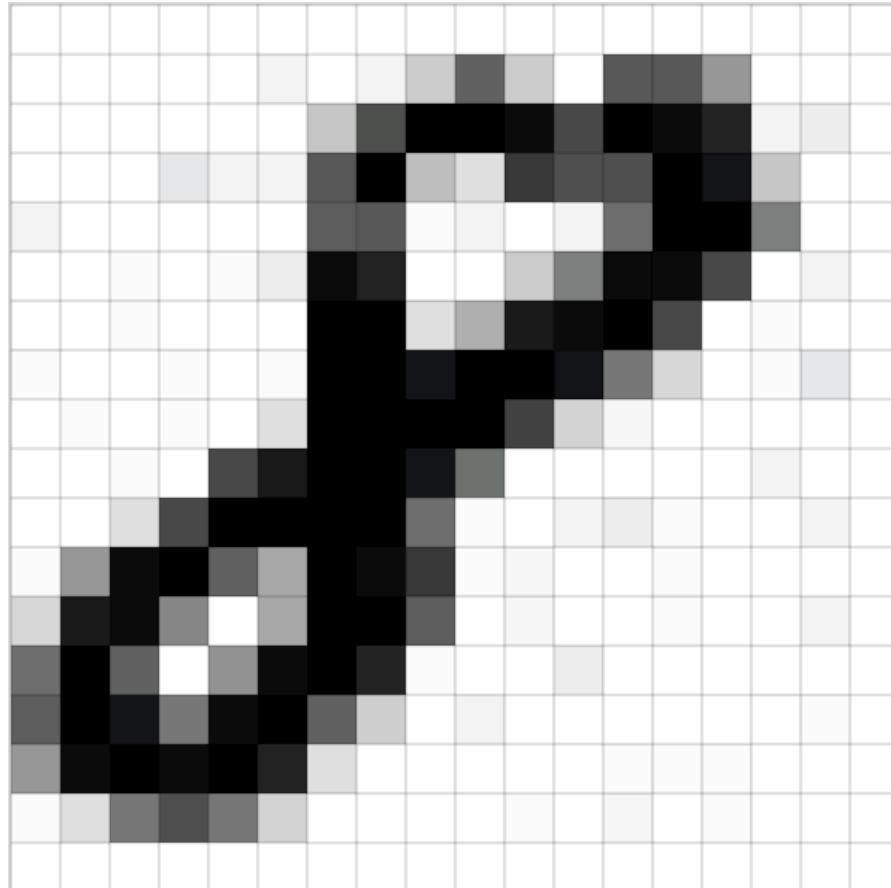
Ejemplo: Fábrica

Detección de piezas defectuosas

¿Es esta pieza defectuosa?

Aprendizaje supervisado

Clasificación

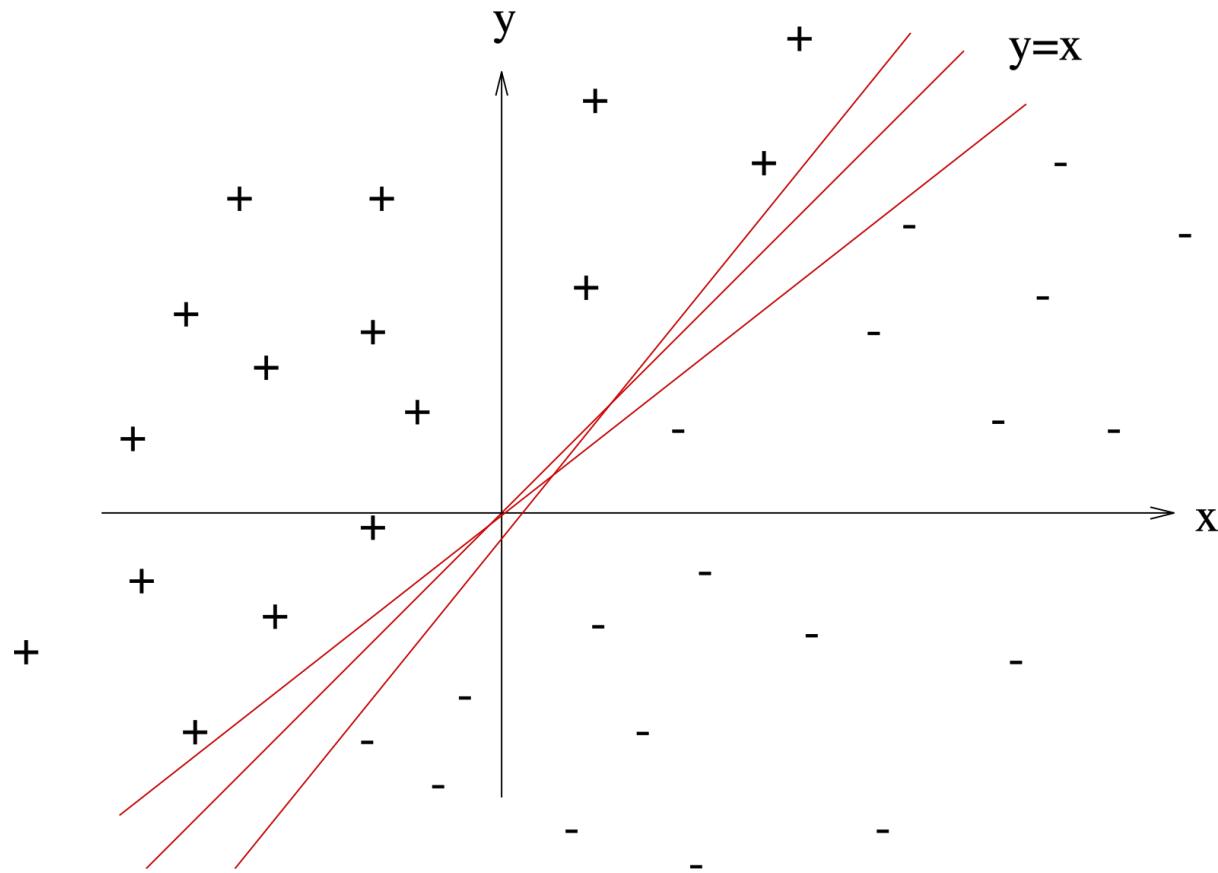


Aprendizaje supervisado

Clasificación

Aprendizaje supervisado

Clasificación



Aprendizaje supervisado

Clasificación

Otros ejemplos:

¿Este cliente se dará de baja de nuestro servicio?

¿Este correo es *spam*?

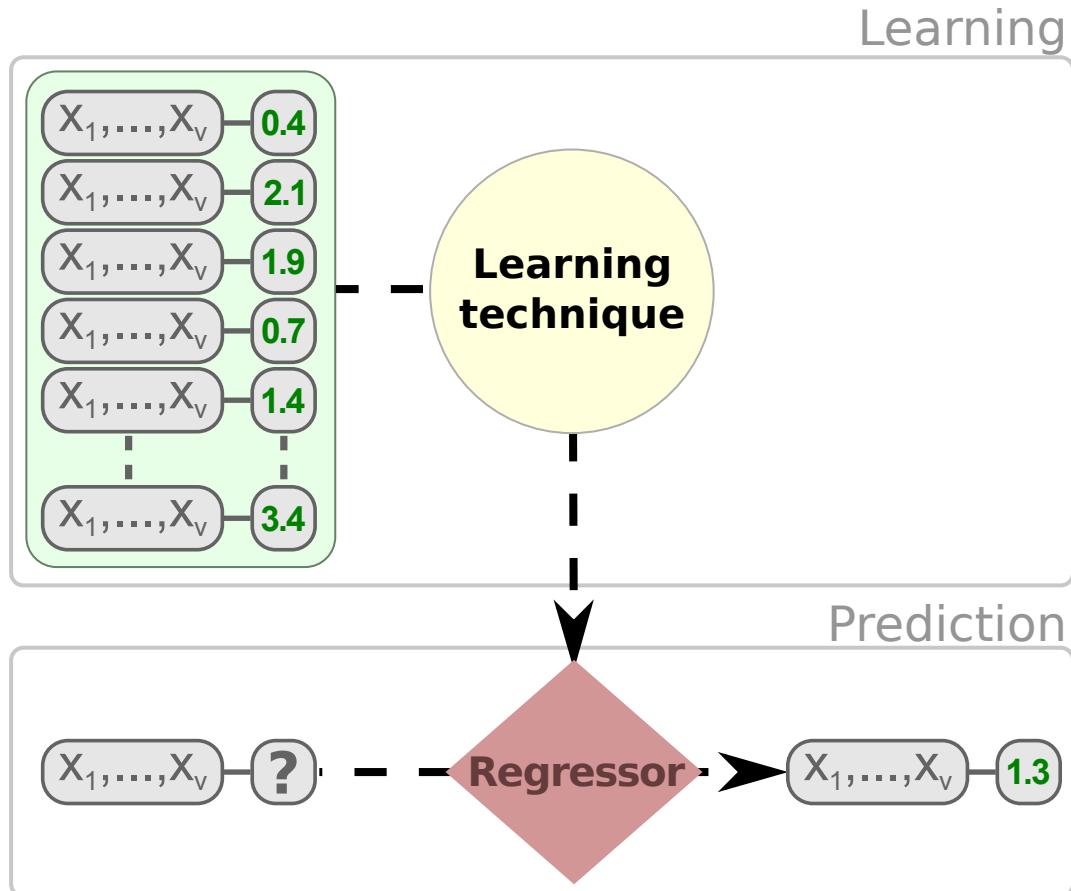
¿Lo que se ve en esta foto es un perro?

¿Esta persona tiene neumonía?

¿A qué partido político votará esta persona en las siguientes elecciones?

Aprendizaje supervisado

Regresión



Aprendizaje supervisado

Regresión

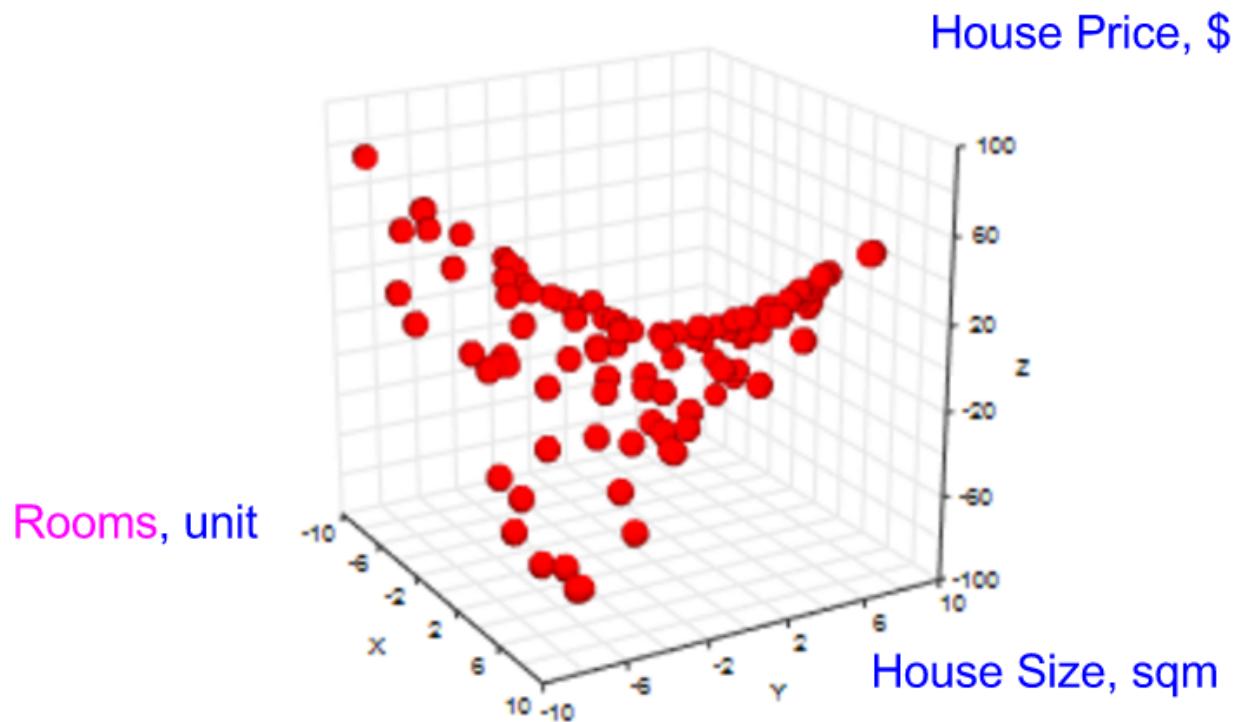
Ejemplo: Fábrica, piezas NO defectuosas

Predicción de la esperanza de vida de las piezas

¿Cuánto tiempo funcionará esta pieza hasta que necesite ser sustituida?

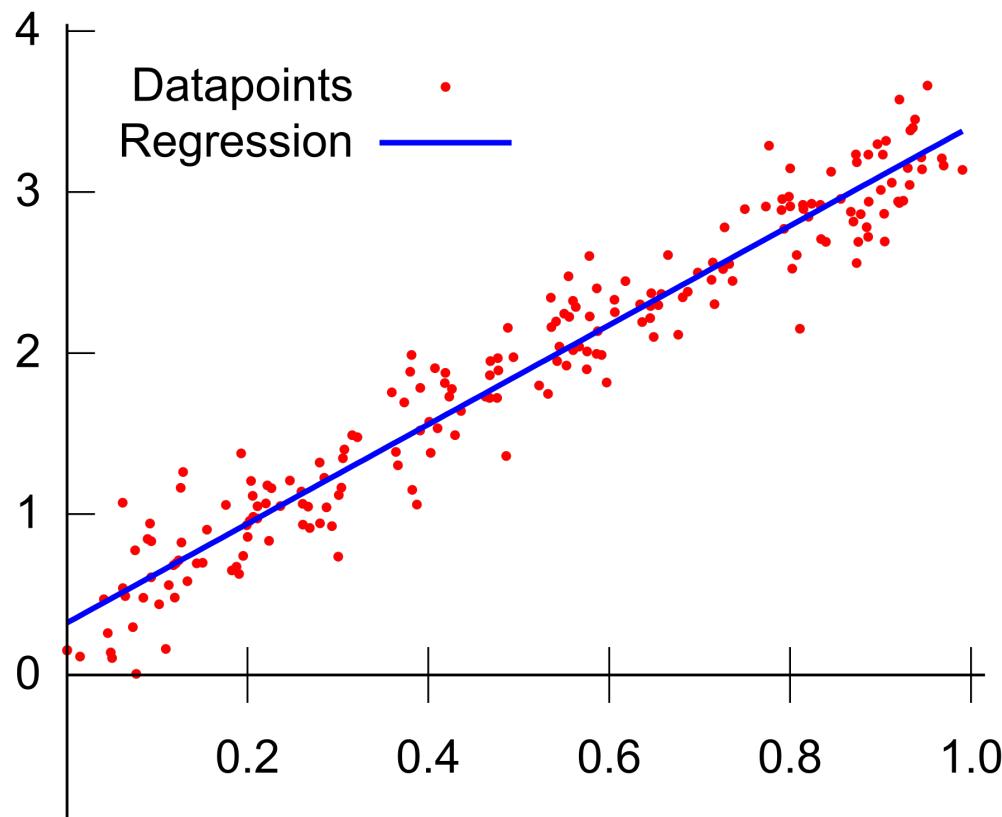
Aprendizaje supervisado

Regresión



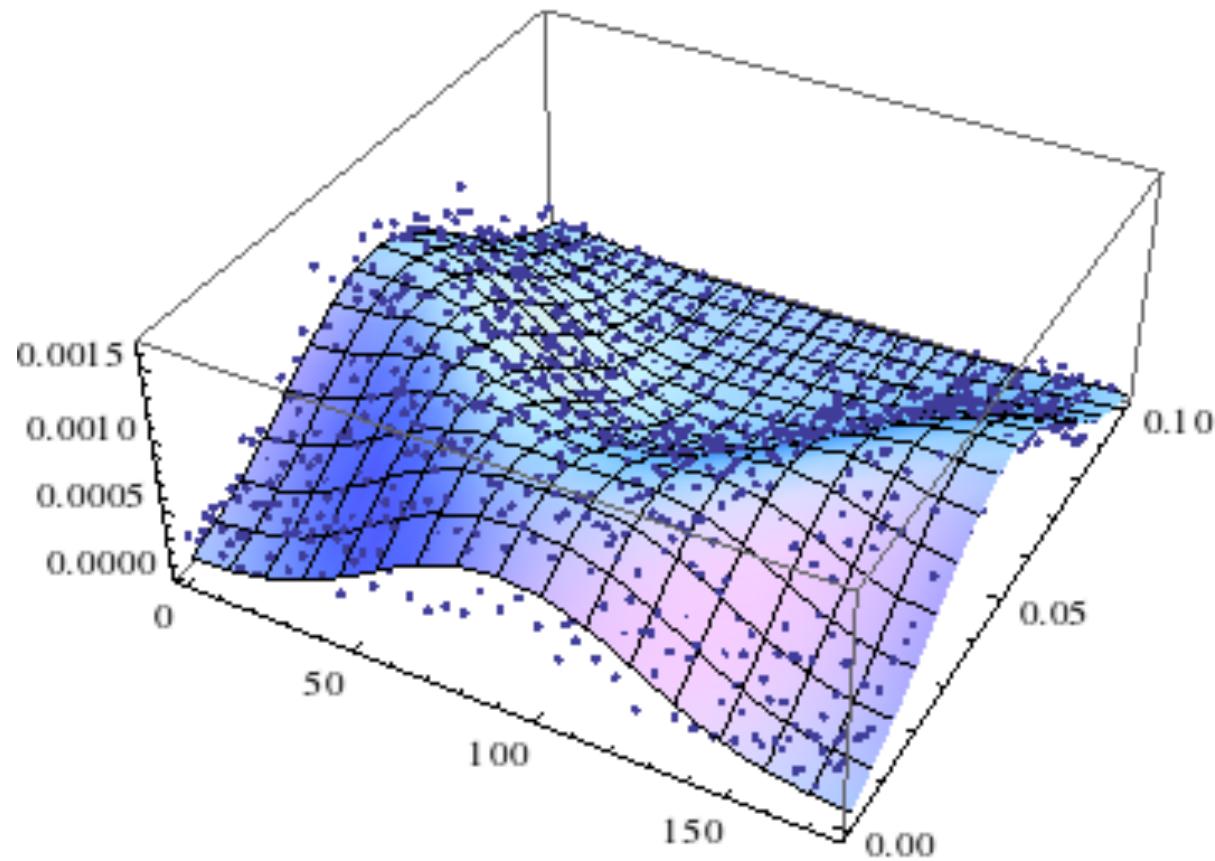
Aprendizaje supervisado

Regresión



Aprendizaje supervisado

Regresión



Aprendizaje supervisado

Regresión

Otros ejemplos:

¿Qué sueldo debería pagar a mis empleados/as?

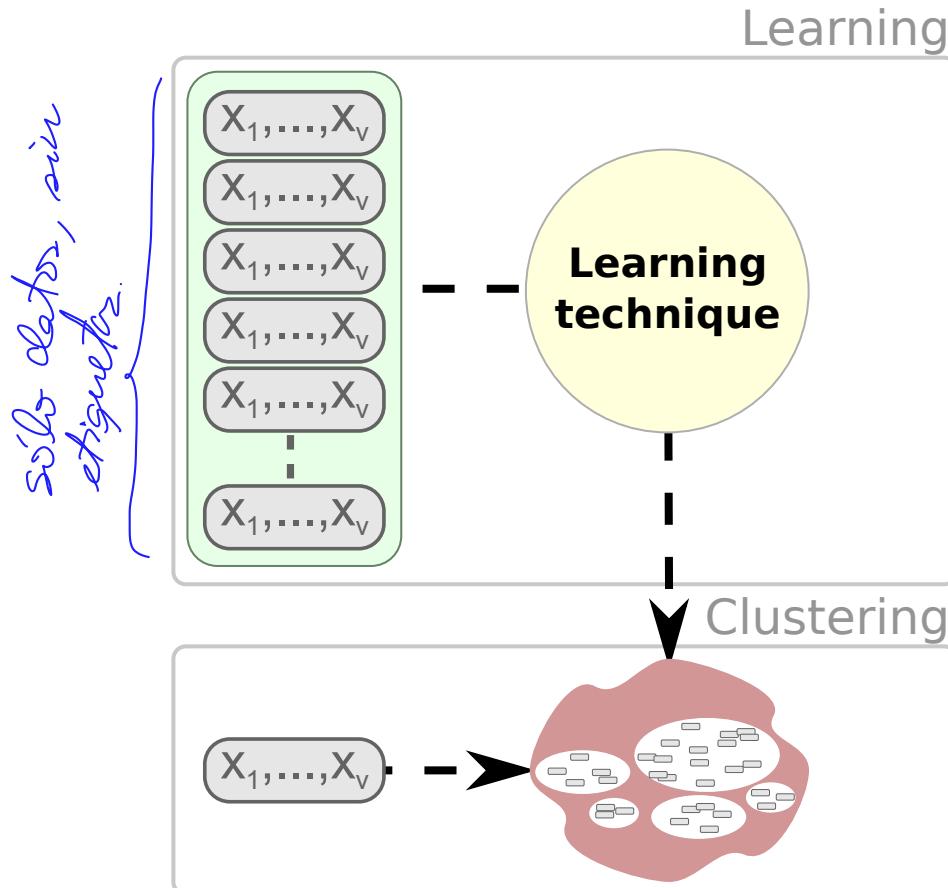
¿Cuál será la nota media de este alumno al acabar la carrera?

Esta persona con artritis, ¿cuánto podrá doblar el codo?

Esta niña, ¿cuánto medirá dentro de 10 años?

Aprendizaje no supervisado

Agrupamiento



Aprendizaje no supervisado

Agrupamiento

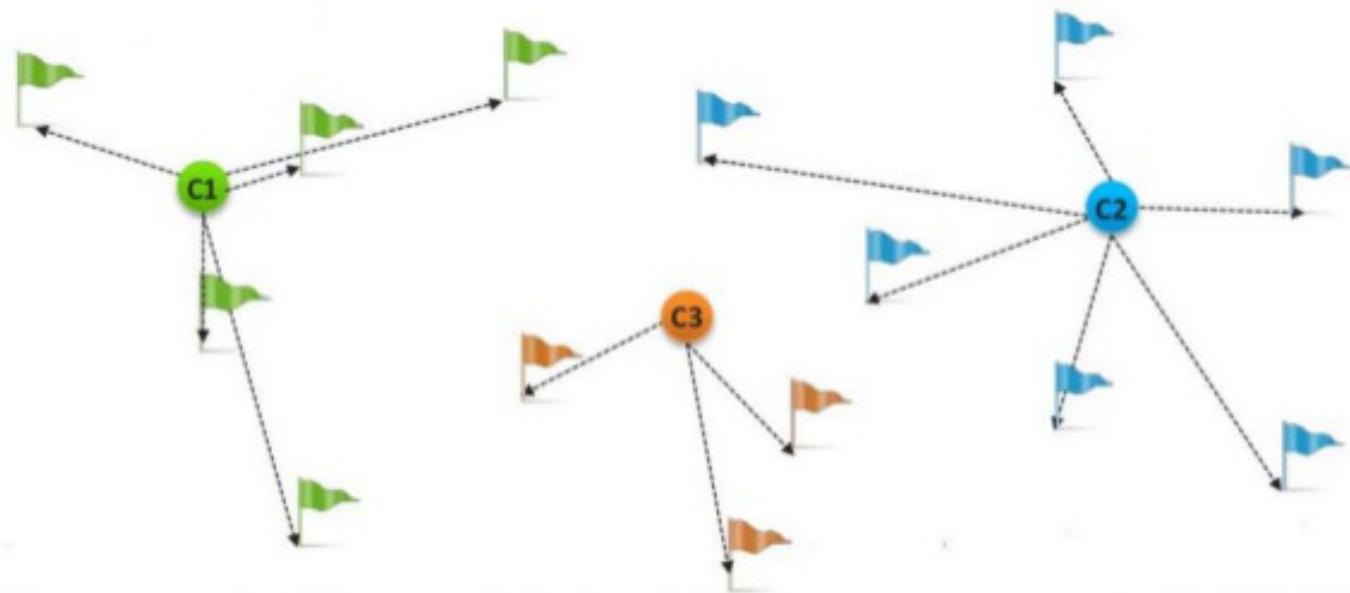
Ejemplo: Fábrica, piezas defectuosas

Detectar diferentes tipos de defectos

¿Existen subgrupos o diferentes tipos de defectos?

Aprendizaje no supervisado

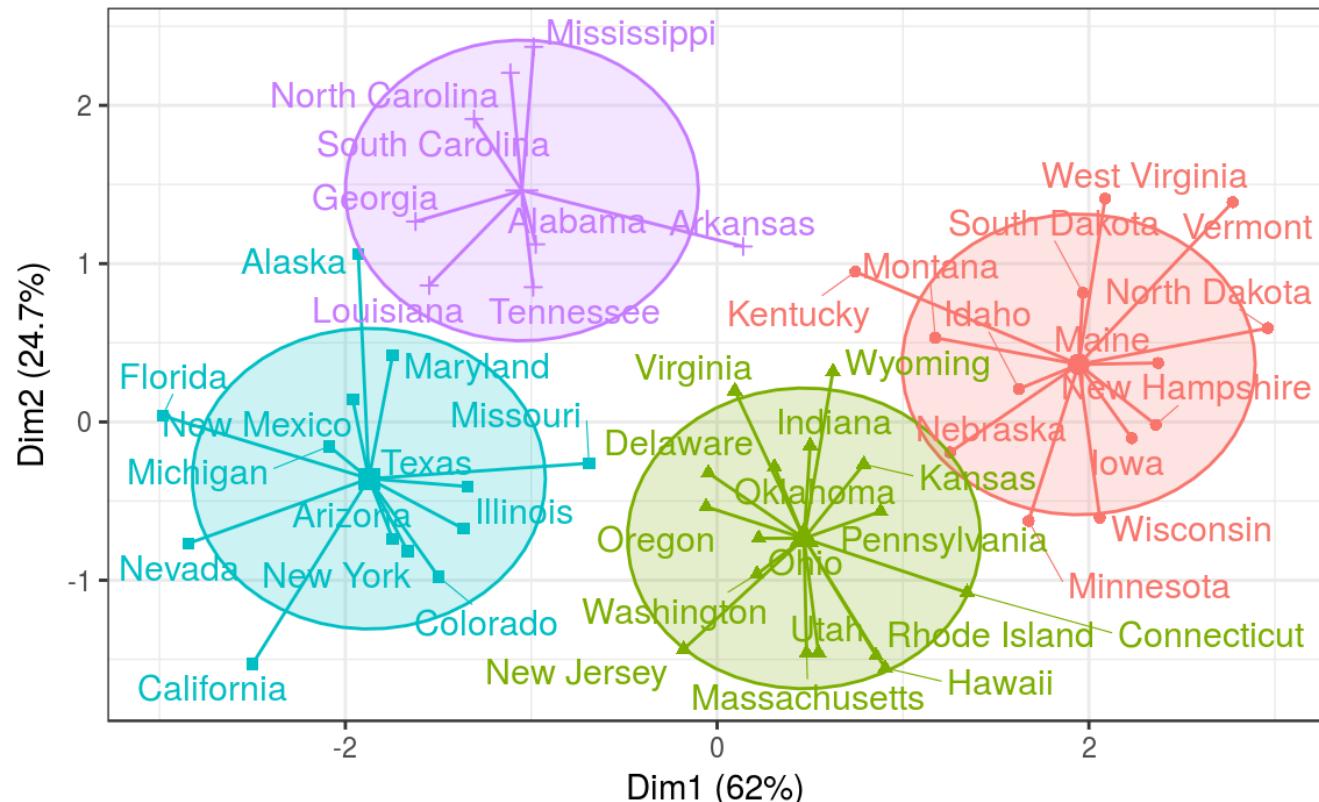
Agrupamiento



Aprendizaje no supervisado

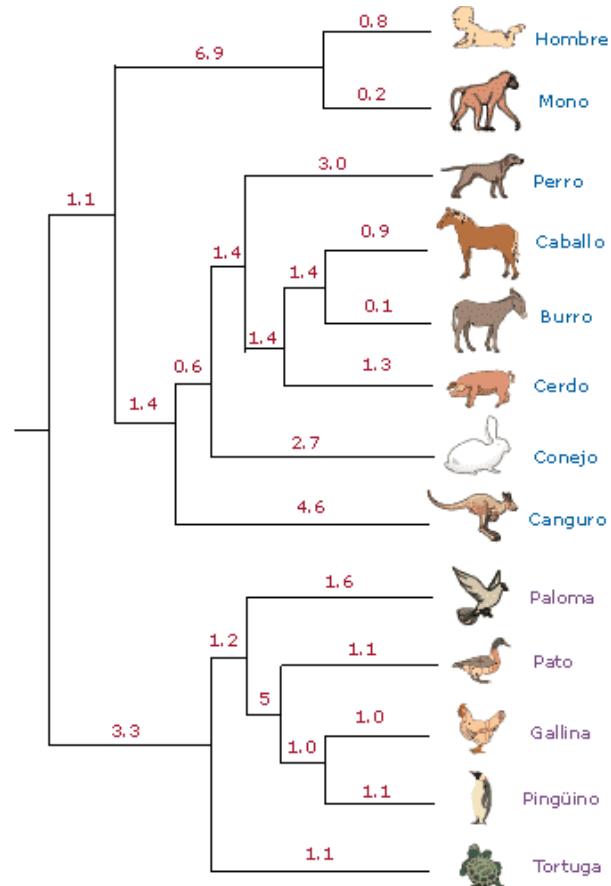
Agrupamiento

Resultados clustering K-means



Aprendizaje no supervisado

Agrupamiento



Aprendizaje no supervisado

Agrupamiento

Otros ejemplos:

¿Existen diferentes tipos de hogares en cuanto al consumo eléctrico?

¿Existen diferentes perfiles entre los consumidores de mi negocio?

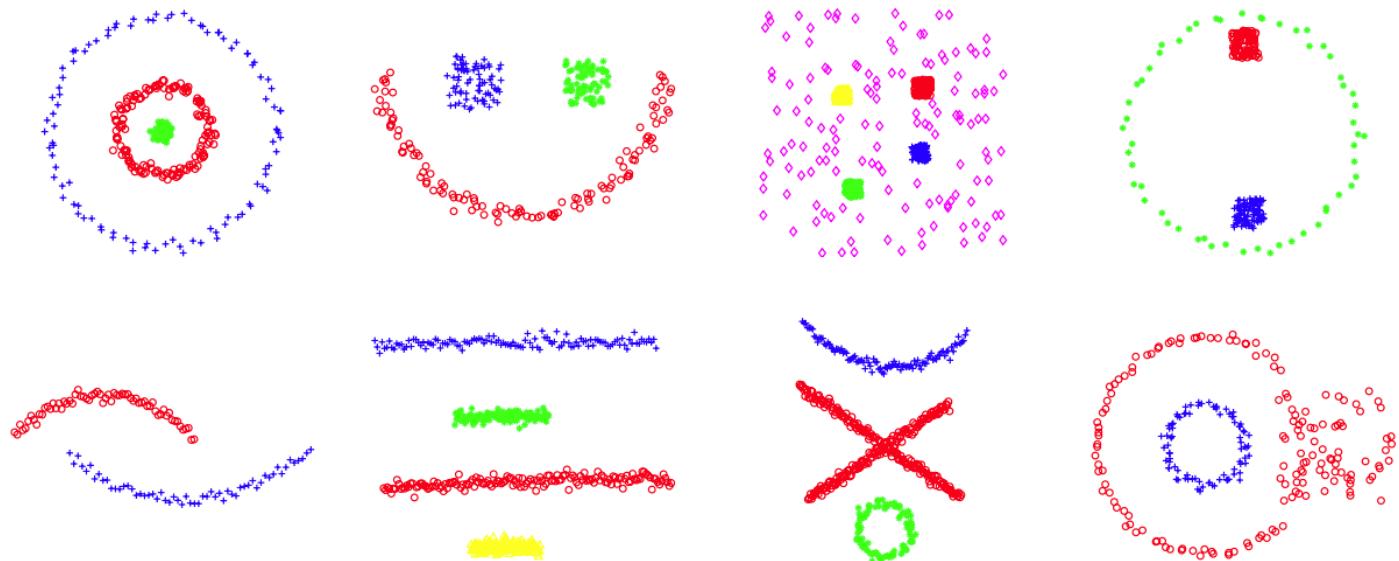
Todos estos objetos celestes, ¿cómo se agrupan? ¿son todos estrellas?

Las noticias de los periódicos de hoy, ¿tratan sobre los mismos temas?

Aprendizaje no supervisado

Agrupamiento

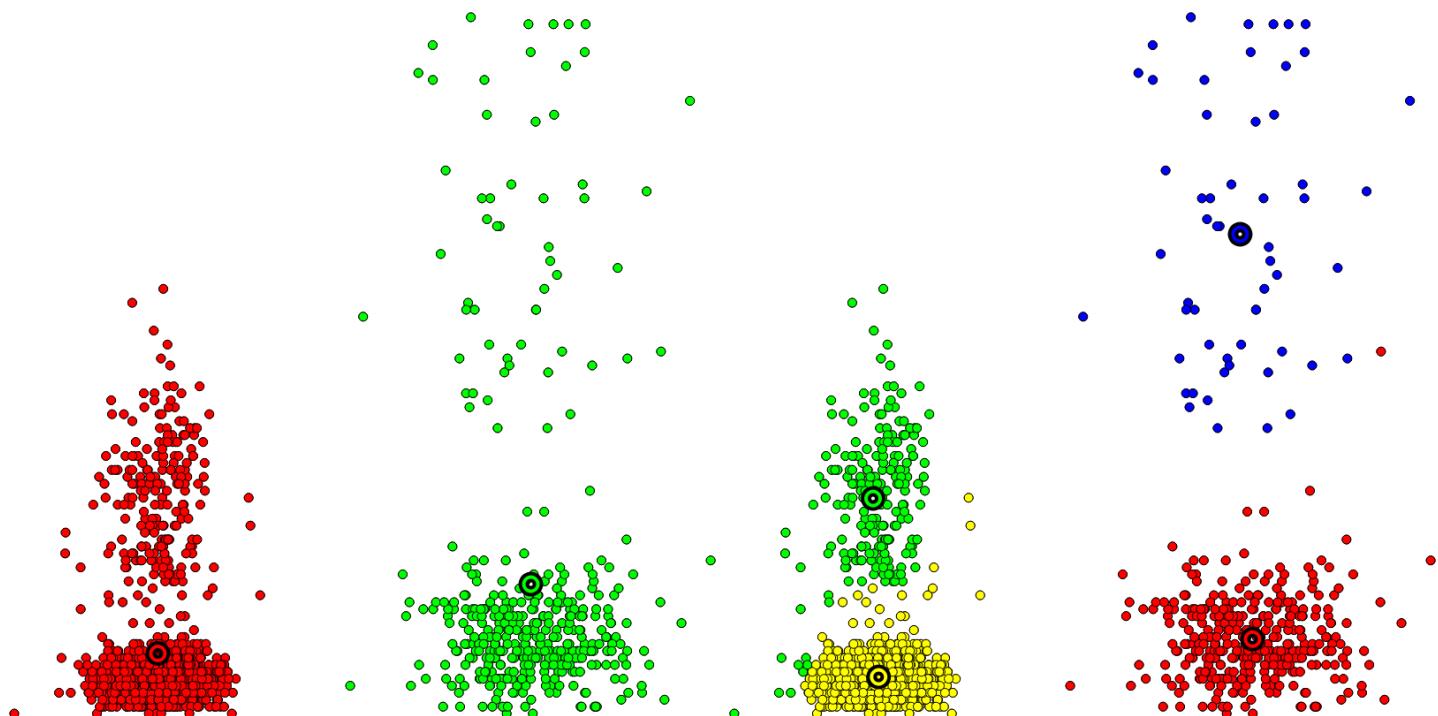
Importancia de la distancia



Aprendizaje no supervisado

Agrupamiento

Evaluación



Aprendizaje no supervisado

VC01: Medidas de distancia

Rocío del Amor

mrocio.delamor@campusviu.es

Universidad Internacional de Valencia

Similitud y disimilitud

Similitud: ¿Cuánto se parecen dos elementos?

Disimilitud: ¿Cuánto se diferencian dos elementos?

Similitud y disimilitud

Disimilitud: ¿Cuánto se diferencian dos elementos?

Distancia: ~ disimilitud, con una serie de condiciones: \rightarrow distancia \subseteq disimilitud

- ▶ No negatividad:

$$d(a, b) \geq 0, \forall a, b \in \mathbb{R}$$



- ▶ Simetricidad:

$$d(a, b) = d(b, a), \forall a, b \in \mathbb{R}$$

- ▶ Identidad de los indiscernibles:

$$d(a, b) = 0 \Leftrightarrow a = b, \forall a, b \in \mathbb{R}$$

- ▶ Desigualdad triangular:

$$d(a, b) \leq d(a, c) + d(c, b), \forall a, b, c \in \mathbb{R}$$

Similitud y disimilitud

Variables aleatorias:

$$\mathbf{X} = (X_1, X_2, \dots, X_v)$$

Similitud y disimilitud

Variables aleatorias:

$$\mathbf{X} = (X_1, X_2, \dots, X_v)$$

Valores de las variables aleatorias:

$$\mathbf{x} = (x_1, x_2, \dots, x_v)$$

Similitud y disimilitud

Variables aleatorias:

$$\mathbf{X} = (X_1, X_2, \dots, X_v)$$

Valores de las variables aleatorias:

$$\mathbf{x} = (x_1, x_2, \dots, x_v)$$

- ▶ Variable continua, X : valor numérico, $x \in \mathbb{R}$
- ▶ Variable categórica, X : valor discreto, $x \in \Omega_X$
con $\Omega_X = \{A, B, \dots, C\}$

Similitud y disimilitud

Variables continuas:

Una única variable

$$d(x_1, x_2) = |x_1 - x_2|$$

Similitud y disimilitud

Variables continuas:

Varias variables

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^v (x_{1j} - x_{2j})^2} = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T (\mathbf{x}_1 - \mathbf{x}_2)}$$

Distancia euclídea

Similitud y disimilitud

Variables continuas:

Minkowski
distance

$$d_p(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|_p = \left(\sum_{j=1}^v |x_{1j} - x_{2j}|^p \right)^{(1/p)}$$

Similitud y disimilitud

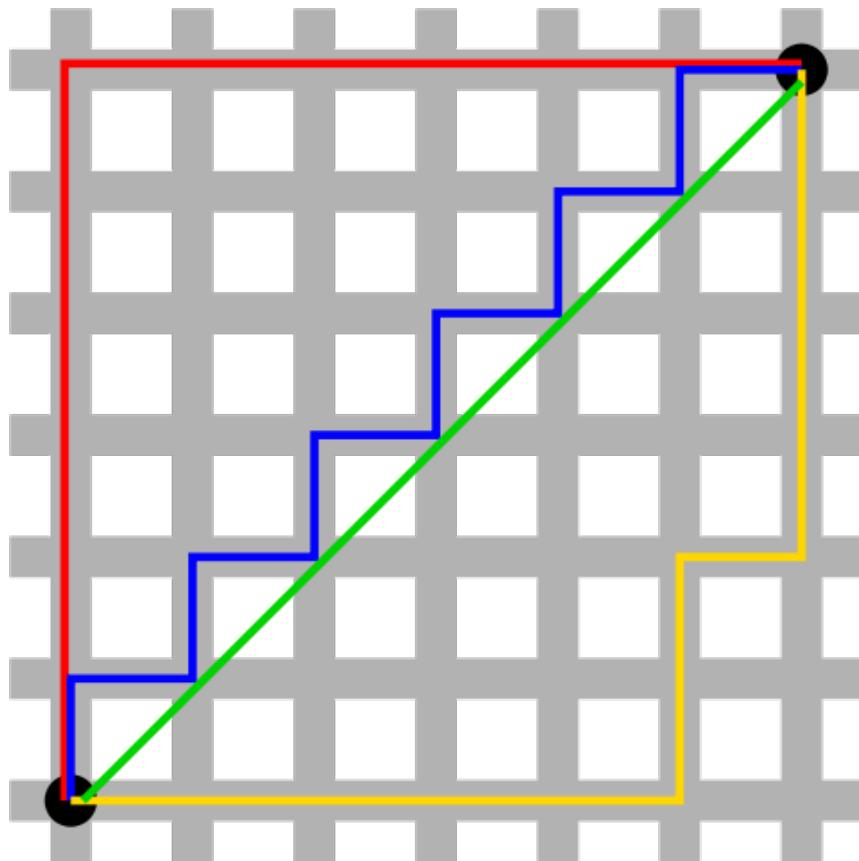
Variables continuas:

$$d_p(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|_p = \left(\sum_{j=1}^v |x_{1j} - x_{2j}|^p \right)^{(1/p)}$$

- Manhattan ($p = 1$):

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^v |x_{1j} - x_{2j}| = \|\mathbf{x}_1 - \mathbf{x}_2\|_1$$

Similitud y disimilitud



Similitud y disimilitud

Variables continuas:

$$d_p(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|_p = \left(\sum_{j=1}^v |x_{1j} - x_{2j}|^p \right)^{(1/p)}$$

- Manhattan ($p = 1$):

EXAMEN

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^v |x_{1j} - x_{2j}| = \|\mathbf{x}_1 - \mathbf{x}_2\|_1$$

- Euclíadiana ($p = 2$):

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^v (x_{1j} - x_{2j})^2} = \|\mathbf{x}_1 - \mathbf{x}_2\|_2$$



La más usada para variables continuas

Similitud y disimilitud

Variables continuas:

$$d_p(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|_p = \left(\sum_{j=1}^v |x_{1j} - x_{2j}|^p \right)^{(1/p)}$$

- Manhattan ($p = 1$):

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^v |x_{1j} - x_{2j}| = \|\mathbf{x}_1 - \mathbf{x}_2\|_1$$

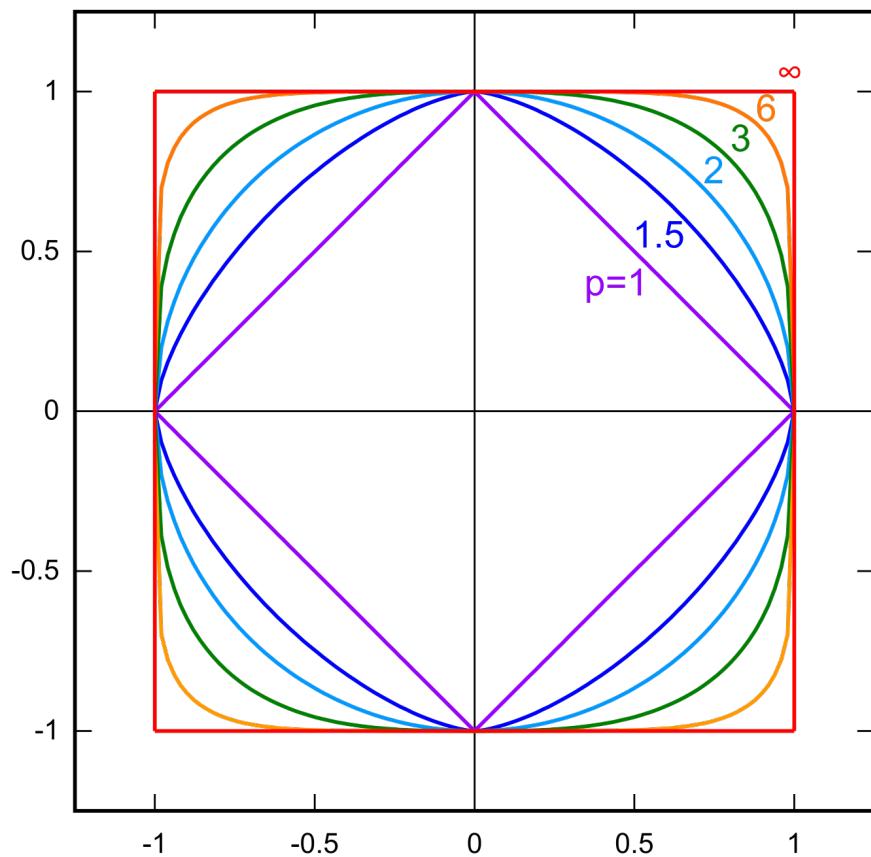
- Euclíadiana ($p = 2$):

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^v (x_{1j} - x_{2j})^2} = \|\mathbf{x}_1 - \mathbf{x}_2\|_2$$

- Máximo ($p = \infty$):

$$d(\mathbf{x}_1, \mathbf{x}_2) = \max_{j \in 1, \dots, v} |x_{1j} - x_{2j}| = \|\mathbf{x}_1 - \mathbf{x}_2\|_\infty$$

Similitud y disimilitud



Similitud y disimilitud

Variables continuas:

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T \Sigma^{-1} (\mathbf{x}_1 - \mathbf{x}_2)}$$

Distancia Mahalanobis

EXAMEN

Si se quiere tener en cuenta la dependencia entre las variables

se usa la matriz de covarianza.

Similitud y disimilitud

$$\Sigma = \begin{bmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & E[(X_1 - \mu_1)(X_n - \mu_n)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & E[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ E[(X_n - \mu_n)(X_1 - \mu_1)] & E[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & E[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}$$

$$\Sigma = E \left[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T \right]$$

$$\sigma^2 = \text{var}(X) = E \left[(X - E[X])^2 \right] = E [(X - E[X])(X - E[X])]$$

$$\text{cov}(X, Y) = E [(X - E[X])(Y - E[Y])]$$

Similitud y disimilitud

Variables continuas:

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T \Sigma^{-1} (\mathbf{x}_1 - \mathbf{x}_2)}$$

Distancia Mahalanobis

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^v \left(\frac{x_{1j} - x_{2j}}{\sigma_j} \right)^2} = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T S^{-1} (\mathbf{x}_1 - \mathbf{x}_2)}$$

Distancia euclídea estandarizada

variables independientes

Similitud y disimilitud

Variables continuas:

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T \Sigma^{-1} (\mathbf{x}_1 - \mathbf{x}_2)}$$

Distancia Mahalanobis

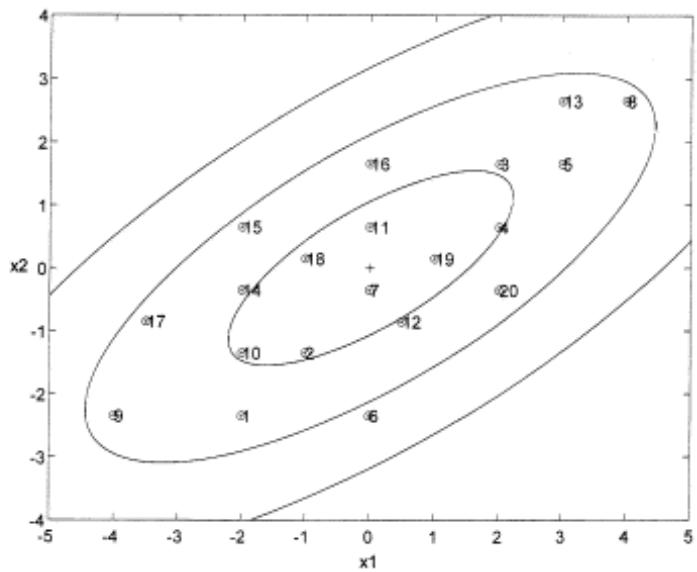
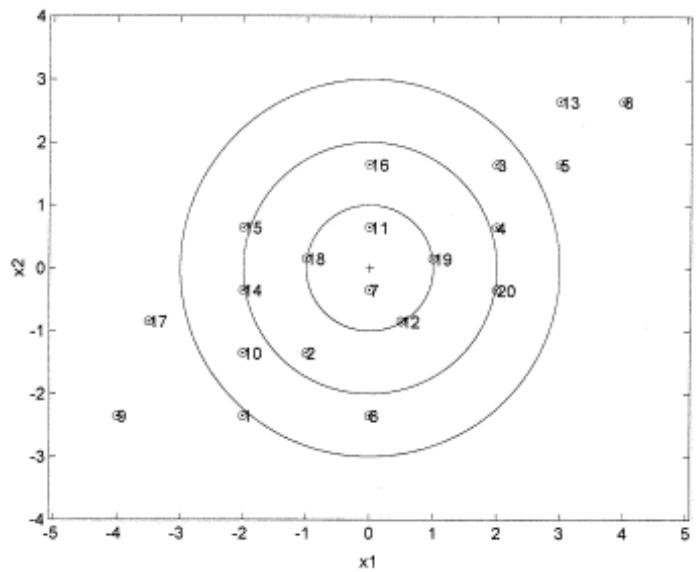
$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^v \left(\frac{x_{1j} - x_{2j}}{\sigma_j} \right)^2} = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T S^{-1} (\mathbf{x}_1 - \mathbf{x}_2)}$$

Distancia euclíadiana estandarizada

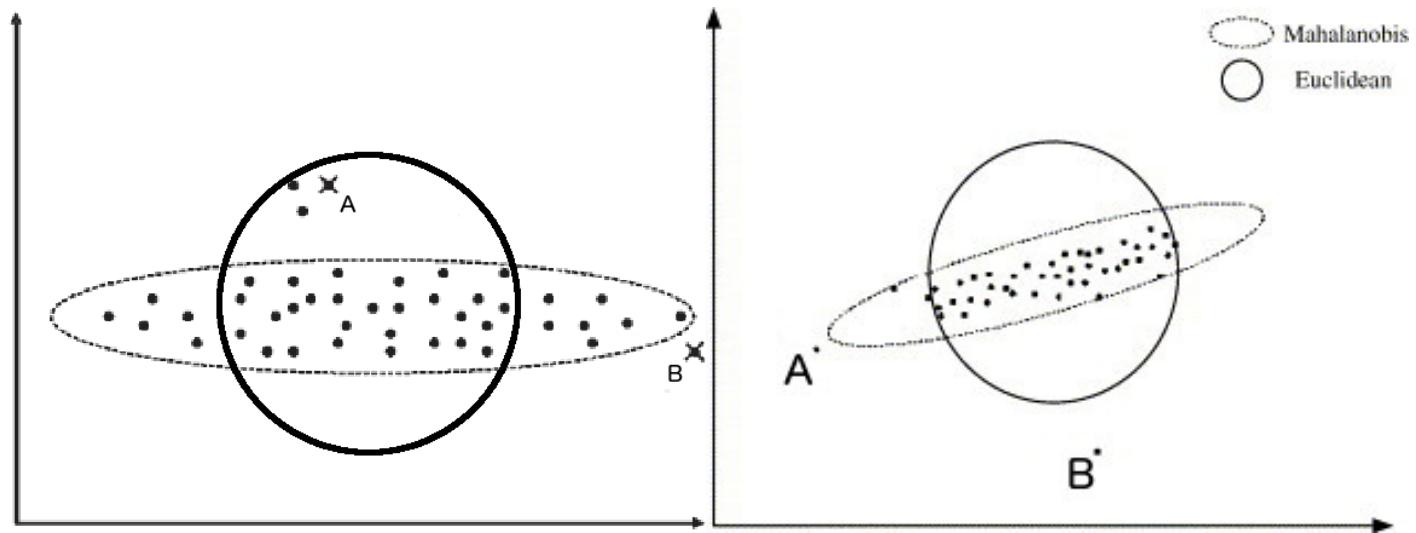
$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^v (x_{1j} - x_{2j})^2} = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T (\mathbf{x}_1 - \mathbf{x}_2)}$$

Distancia euclíadiana

Similitud y disimilitud



Similitud y disimilitud



Similitud y disimilitud

Variable continuas:

$$s(\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathbf{x}_1 \cdot \mathbf{x}_2}{\|\mathbf{x}_1\| \cdot \|\mathbf{x}_2\|} = \frac{\sum_{j=1}^v x_{1j} \cdot x_{2j}}{\sqrt{\sum_{j=1}^v x_{1j}^2} \sqrt{\sum_{j=1}^v x_{2j}^2}}$$

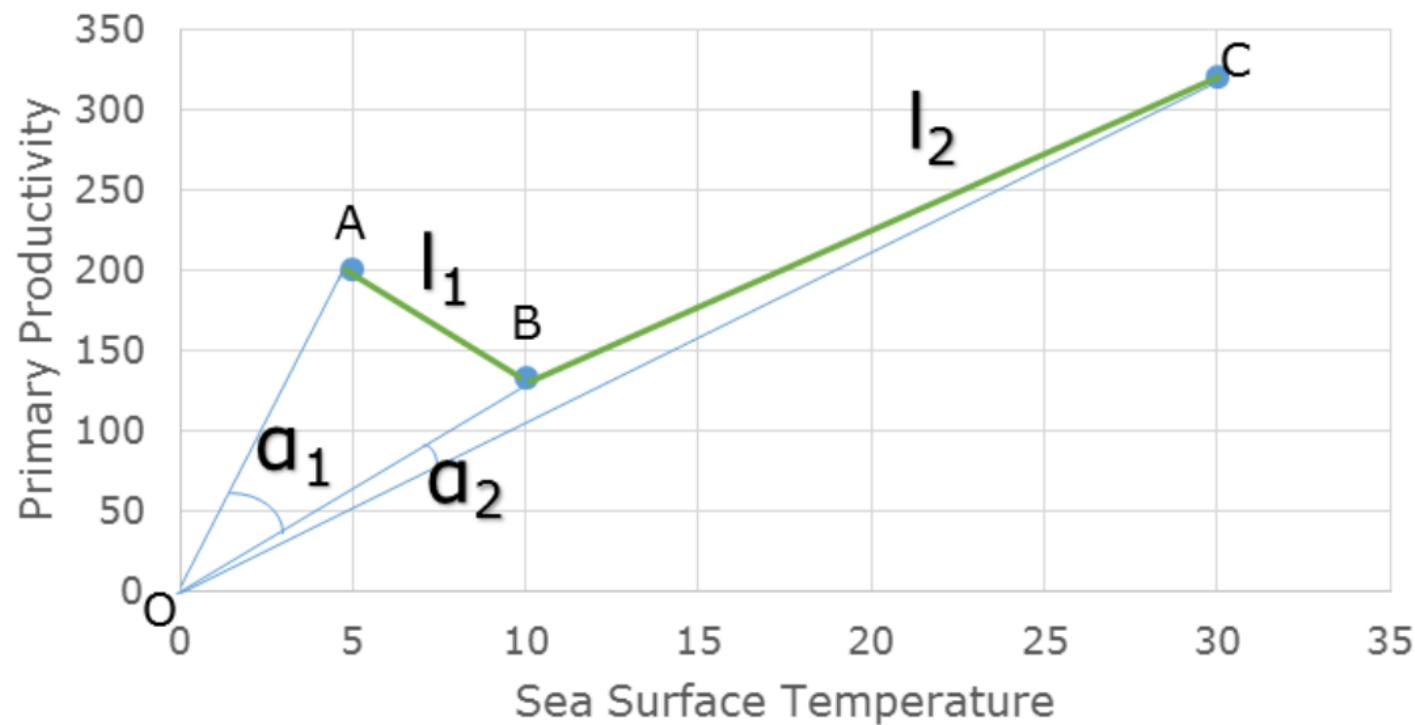
Similitud coseno

- entre -1 y 1
- ángulo menor \Rightarrow distancia mayor

Similitud y disimilitud

Variable continuas:

Similitud coseno



Similitud y disimilitud

Variable binarias:

$$d(\mathbf{x}_1, \mathbf{x}_2) = |x_{1j} = x_{2j}|_{j \in \{1, \dots, v\}}$$

Distancia de Hamming

$$s(\mathbf{x}_1, \mathbf{x}_2) = \frac{|x_{1j} = 1 \wedge x_{2j} = 1|_{j \in \{1, \dots, v\}}}{|x_{1j} = 1 \vee x_{2j} = 1|_{j \in \{1, \dots, v\}}}$$

Similitud de Jaccard

Similitud y disimilitud

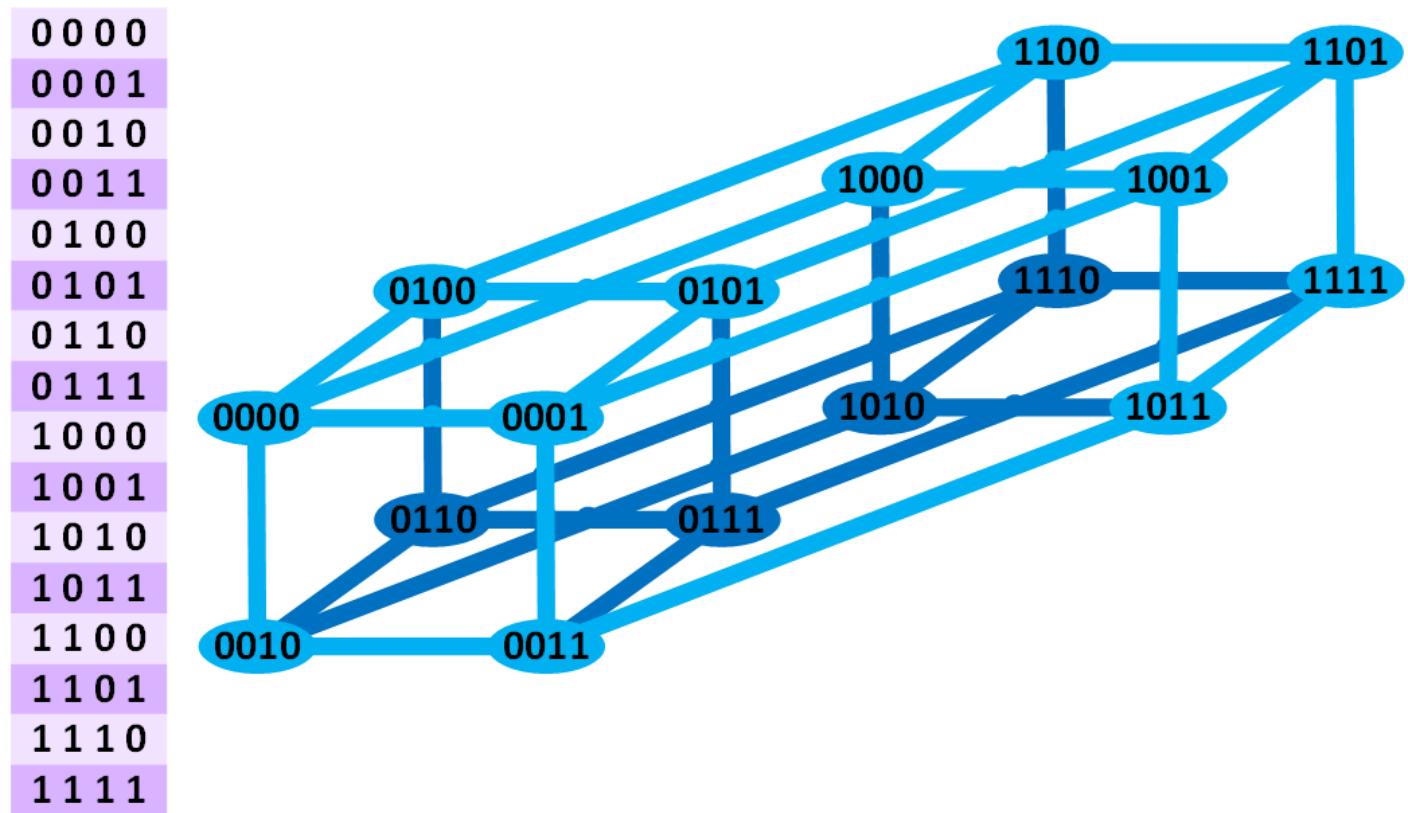
Variable binarias:

<i>A</i>	1	0	1	1	0	0	1	0	0	1
<i>B</i>	1	0	0	1	0	0	0	0	1	1

Distancia de Hamming

Similitud y disimilitud

Variable binarias:



Distancia de Hamming

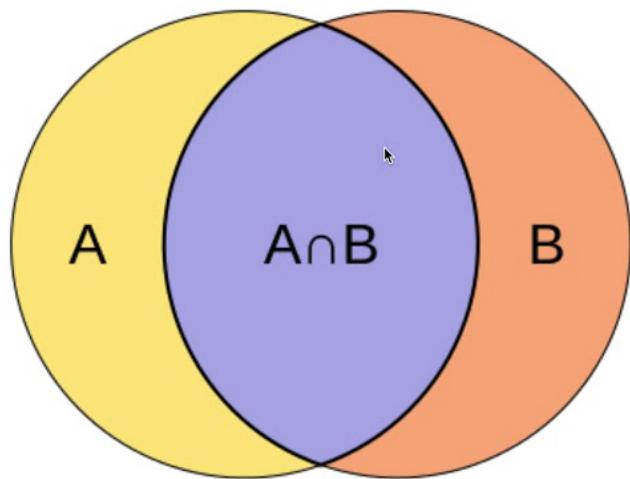
Similitud y disimilitud

Variable binarias:

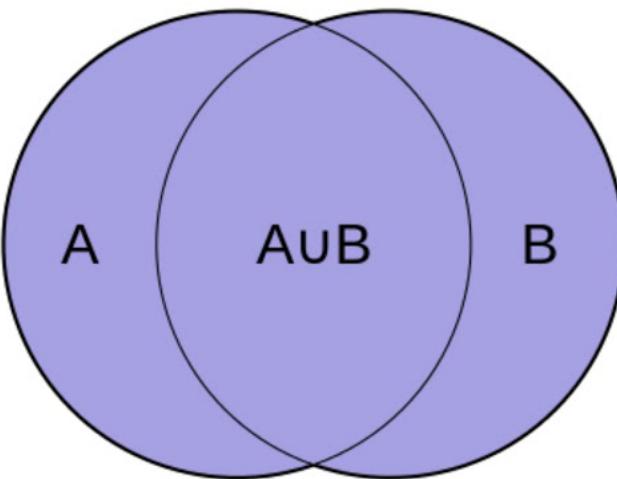
$$\frac{IOU}{h}$$

cuanto más se parezcan
los conjuntos
 \Leftrightarrow Int. ~ Union.

Intersection



Union



$$IOU = \frac{\text{Intersection}}{\text{Union}}$$

Similitud de Jaccard

Similitud y disimilitud

Variable categórica:

$$d_j(x_{1j}, x_{2j}) = \begin{cases} 1, & \text{si } x_{1j} \neq x_{2j} \\ 0, & \text{si } x_{1j} = x_{2j} \end{cases}$$

Combinar medidas por variable:

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^v d_j(x_{1j}, x_{2j})$$

Similitud y disimilitud

Combinar medidas por variable:

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^v d_j(x_{1j}, x_{2j})$$

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^v w_j \cdot d_j(x_{1j}, x_{2j}), \text{ con } \sum_{j=1}^v w_j = 1$$

Propuesta de Hastie et al. (2008):

$$w_j = 1/\hat{d}_j, \text{ con } \hat{d}_j = \frac{1}{n^2} \sum_{i=1}^n \sum_{i'=1}^n d_j(x_{ij}, x_{i'j})$$

Similitud y disimilitud

Combinar medidas por variable:

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^v d_j(x_{1j}, x_{2j})$$

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^v w_j \cdot d_j(x_{1j}, x_{2j}), \text{ con } \sum_{j=1}^v w_j = 1$$

Propuesta de Hastie et al. (2008):

$$w_j = 1/\hat{d}_j, \text{ con } \hat{d}_j = \frac{1}{n^2} \sum_{i=1}^n \sum_{i'=1}^n d_j(x_{ij}, x_{i'j})$$

Si $d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2$ para todo j , entonces: $w_j = 1/(2\text{var}_j)$

Similitud y disimilitud

Transformar matriz de ejemplos D ($n \times v$) en...

matriz de distancias, M ($n \times n$), tal que:

$$M_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$$

y ésta, a su vez, en una matriz de similitudes, S ($n \times n$):

$$S_{ij} = \exp(-M_{ij}^2/c)$$

Aprendizaje no supervisado

VC01: Medidas de evaluación de agrupamientos

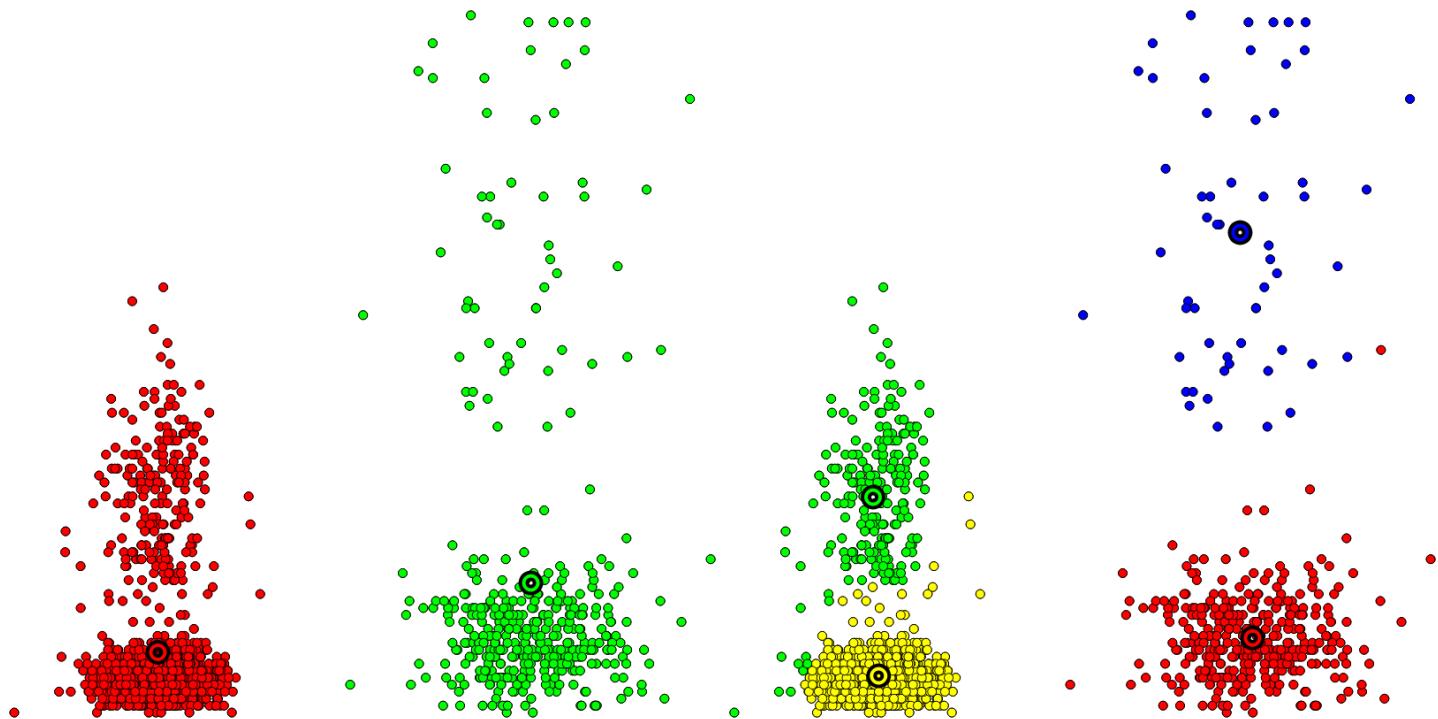
Rocío del Amor del Amor

mrocio.delamor@campusviu.es

Universidad Internacional de Valencia

Evaluación

Extrínseca vs. Intrínseca



Evaluación extrínseca

- ▶ Se conoce el agrupamiento real (**verdad básica**) *ground truth*.
- ▶ Se compara el resultado del algoritmo con la verdad básica
- ▶ No existe el concepto de etiqueta:
búsqueda de la **correspondencia** entre **clúster real** y **predicho**

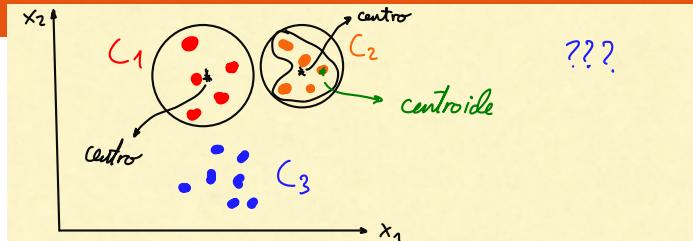
Evaluación intrínseca

- ▶ No se conoce el agrupamiento real (**verdad básica**), ni se sabe si existe
- ▶ Se mide la congruencia del agrupamiento
- ▶ Diferentes criterios posibles

Evaluación

$1 \dots k \approx$ número de clases

$\{B_I\}_{I=1}^{K'}:$ Verdad básica
 \hookrightarrow_l



$\{C_k\}_{k=1}^K:$ Agrupamiento resultante de un algoritmo de *clustering*

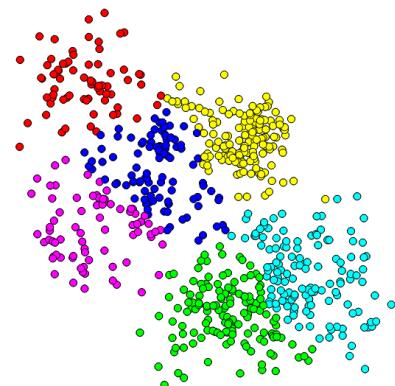
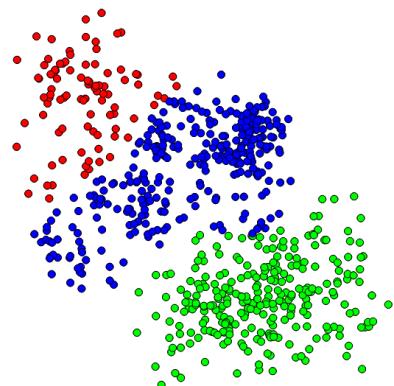
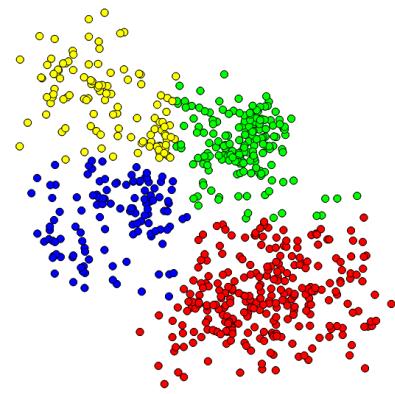
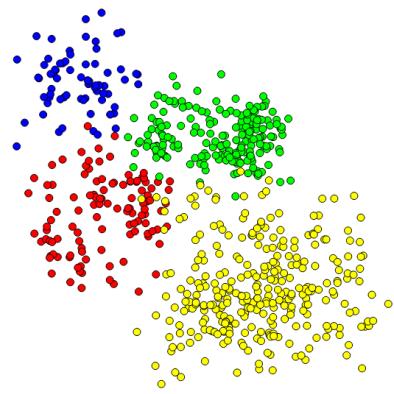
$\{\mathbf{c}_k\}_{k=1}^K:$ Centro(ide)s de los clústeres resultantes

$n_I = |B_I|:$ Tamaño de un clúster verdadero

$n_k = |C_k|:$ Tamaño de un clúster resultante

$n_{kI} = |C_k \cap B_I|:$ Número de ejemplos que comparten un clúster resultante y otro verdadero

Evaluación

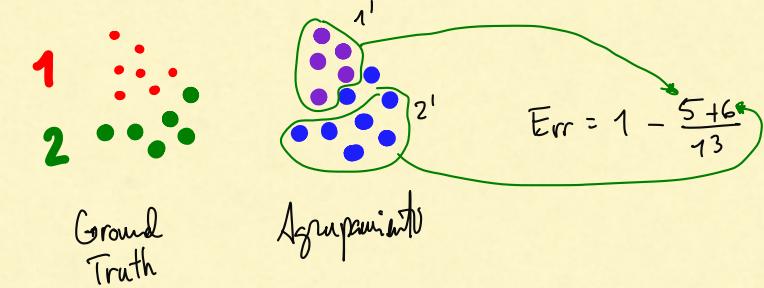


Evaluación

Extrínseca

Extrínseca

Error:

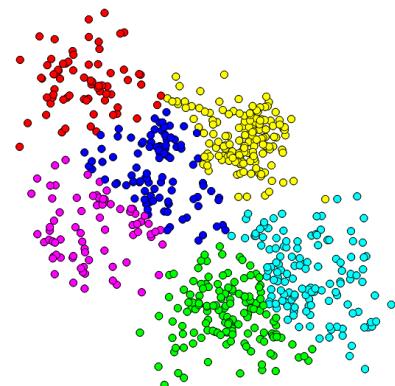
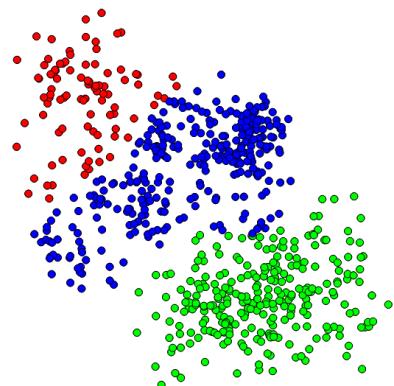
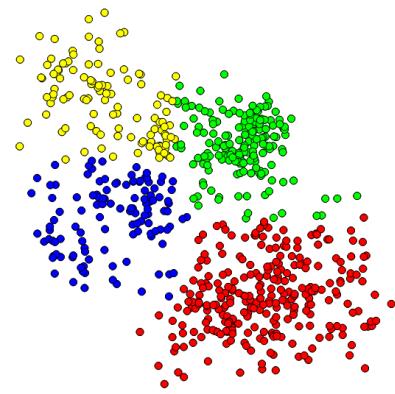
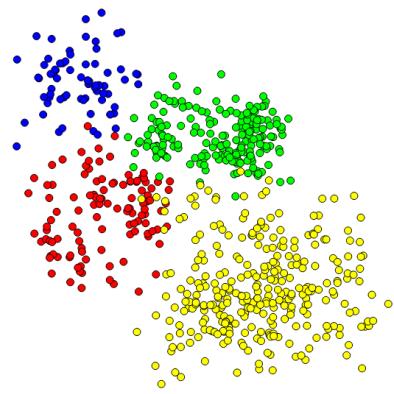


$$E = 1 - \frac{1}{n} \max_{\sigma} \sum_{l=1}^{K'} n_{\sigma(l)} l$$

donde σ es una función de $\sigma : \{1, \dots, K'\} \rightarrow \{1, \dots, K\}$

- ▶ Recorrido sobre los clústeres reales
- ▶ Máximo (optimista) para identificar la correspondencia $C-B$

Evaluación



Evaluación

Extrínseca

Precisión:

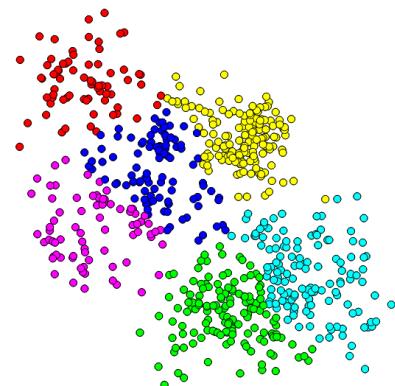
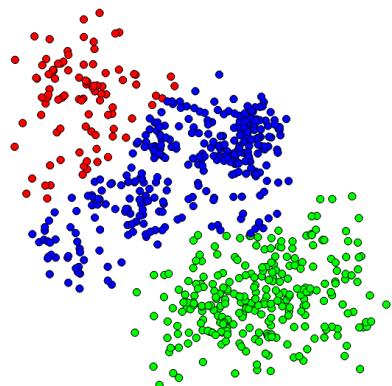
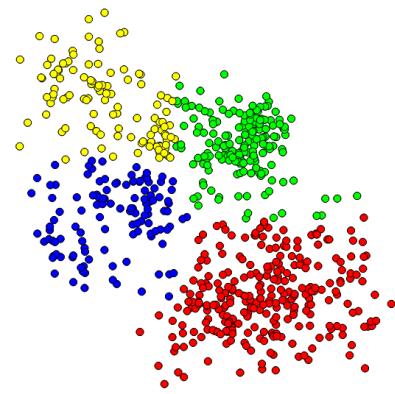
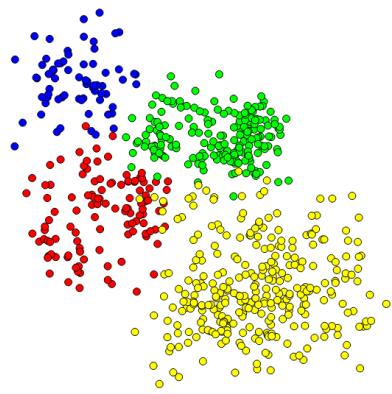
$$P_{kl} = \frac{n_{kl}}{n_k.}$$

Recall:

$$R_{lk} = \frac{n_{kl}}{n_.l}$$

- ▶ Medidas entre un clúster real y otro resultante
- ▶ Precisión: ¿Cuántos de los elementos del clúster resultante k lo son también del clúster real /?
- ▶ Recall: ¿Cuántos de los elementos del clúster real / lo son también del clúster resultante k ?

Evaluación



Evaluación

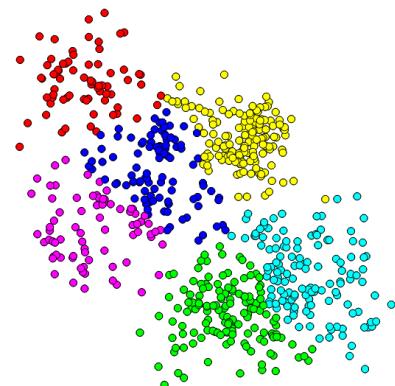
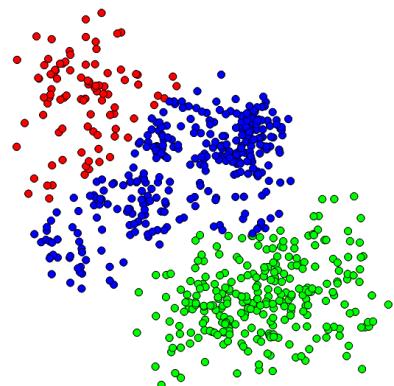
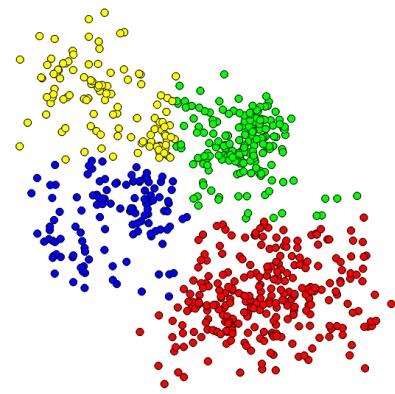
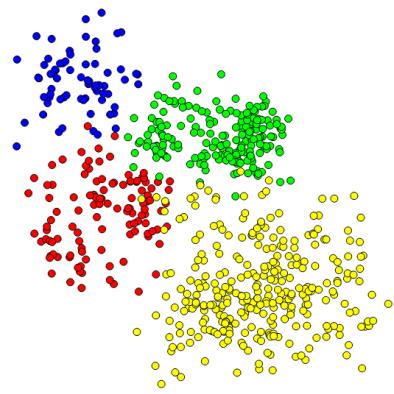
Extrínseca

Pureza:

$$Pu = \sum_{k=1}^K \frac{n_{k\cdot}}{n} \max_{l \in \{1, \dots, K'\}} P_{kl}$$

- ▶ Media ponderada de la precisión
- ▶ Recorrido sobre los clústeres resultantes
- ▶ Máximo (optimista) para identificar la correspondencia C-B

Evaluación



Evaluación

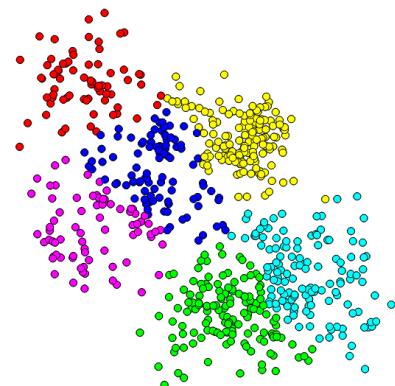
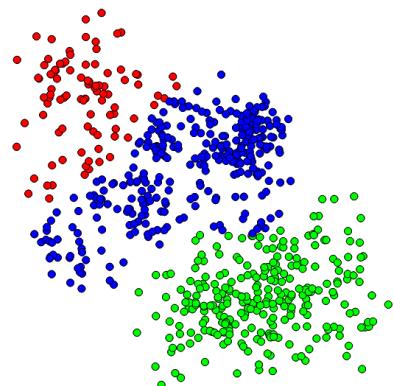
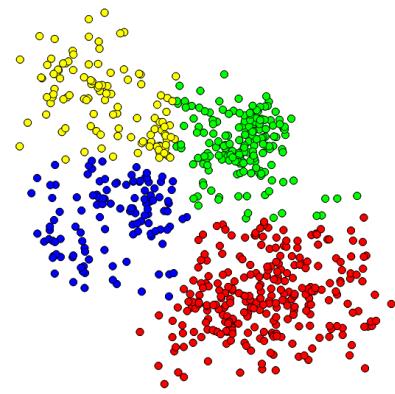
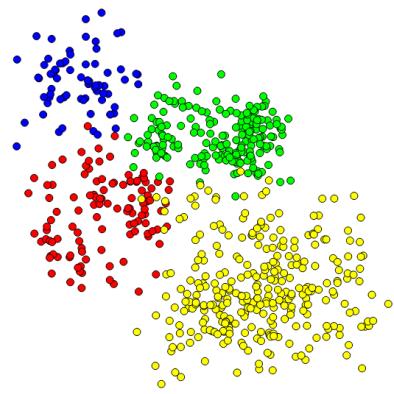
Extrínseca

Medida F:

$$F1 = \sum_{l=1}^{K'} \frac{n_{\cdot l}}{n} \max_{k \in \{1, \dots, K\}} \left(\frac{2P_{kl}R_{lk}}{P_{kl} + R_{lk}} \right)$$

- ▶ Media ponderada de la media harmónica de la precisión y el recall
- ▶ Recorrido sobre los clústeres reales
- ▶ Máximo (optimista) para identificar la correspondencia C-B

Evaluación



Evaluación

Extrínseca

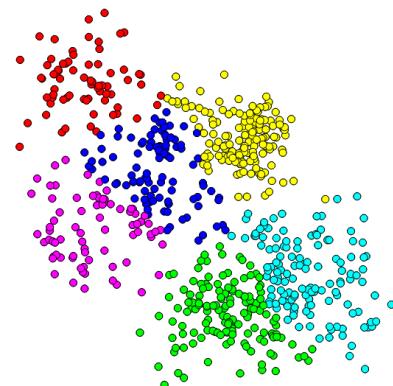
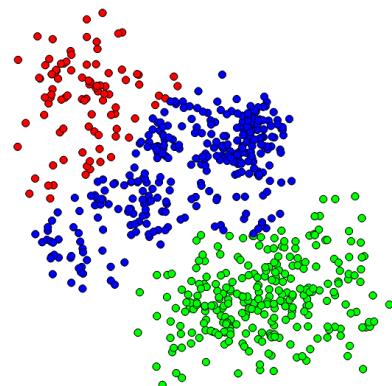
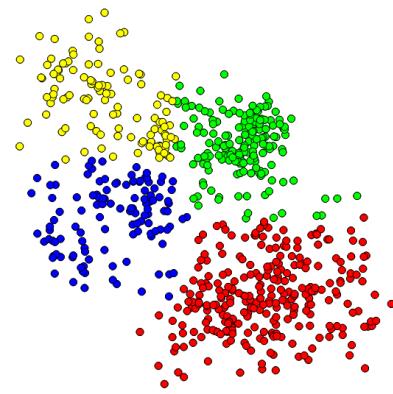
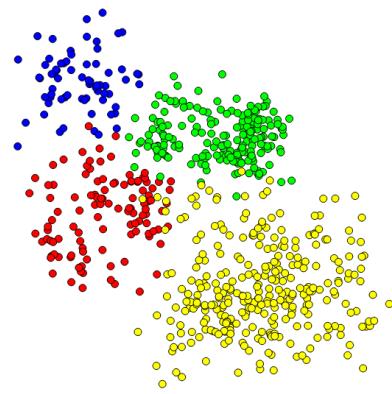
Entropía:

$$H = - \sum_{k=1}^K \frac{n_{k\cdot}}{n} \sum_{l=1}^{K'} \frac{n_{kl}}{n_{k\cdot}} \log \frac{n_{kl}}{n_{k\cdot}}$$

- ▶ Media ponderada de la entropía de cada clúster resultante
- ▶ Entropía: mide cómo se distribuyen los ejemplos de un clúster resultante entre los clústeres reales (crece a mayor desorden)
- ▶ Recorrido (principal) los clústeres resultantes

Entropía $\downarrow \downarrow \Rightarrow$ buenos, clusters homogéneos

Evaluación



Evaluación

Extrínseca

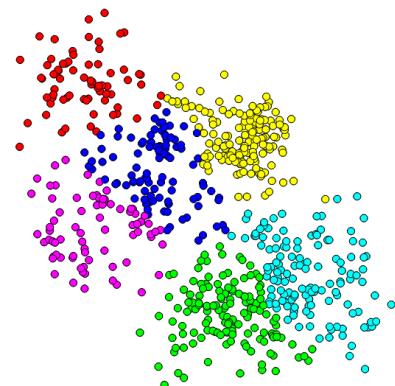
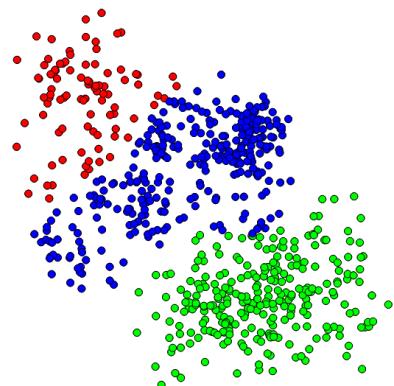
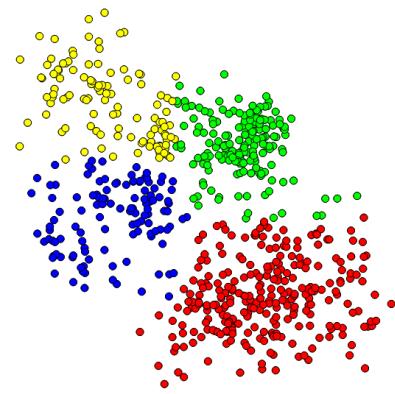
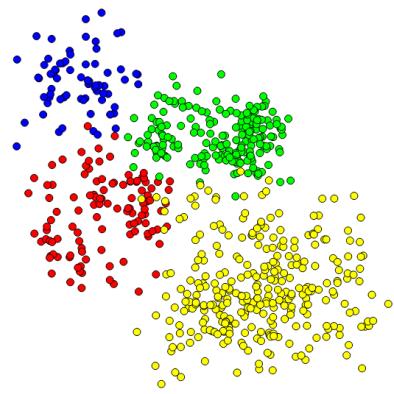
Información mutua:

$$I = \sum_{k=1}^K \sum_{l=1}^{K'} \frac{n_{kl}}{n} \log \frac{n \cdot n_{kl}}{n_{k\cdot} \cdot n_{\cdot l}}$$

- ▶ Información mutua entre dos agrupamientos (real y resultante)
- ▶ I : mide cómo se explican mutuamente ambos agrupamientos

Información mutua $\uparrow\uparrow\uparrow \Rightarrow$ bueno, clusters explican perfectamente el Ground Truth.

Evaluación



Evaluación

Intrínseca

¿Es razonable asumir la existencia de la verdad básica?

¿Para qué queremos entonces un algoritmo de *clustering*?

Evaluación

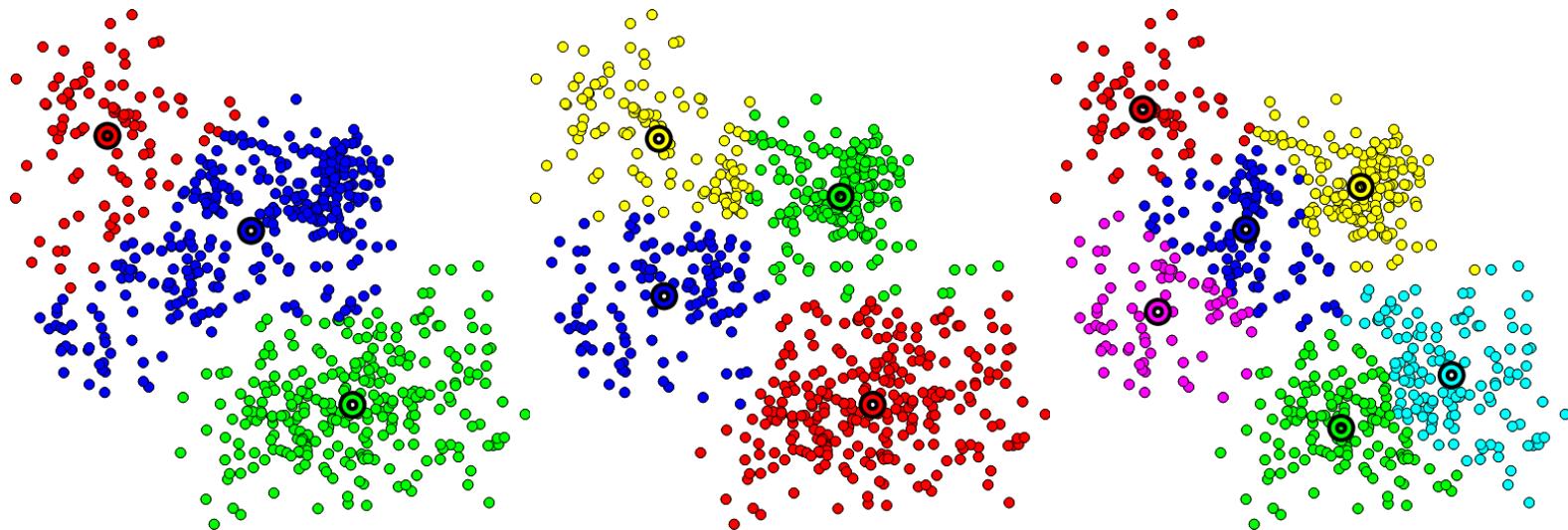
Intrínseca

La raíz del cuadrado de la media de la desviación típica:

$$RMSSTD = \sqrt{\frac{\sum_{k=1}^K \sum_{x_i \in C_k} ||x_i - c_k||^2}{v \cdot \sum_{k=1}^K (|C_k| - 1)}}$$

- ▶ Mide la heterogeneidad de los clústeres
- ▶ Se reduce fácilmente aumentando el número de clústeres resultantes K

Evaluación



Evaluación

Intrínseca

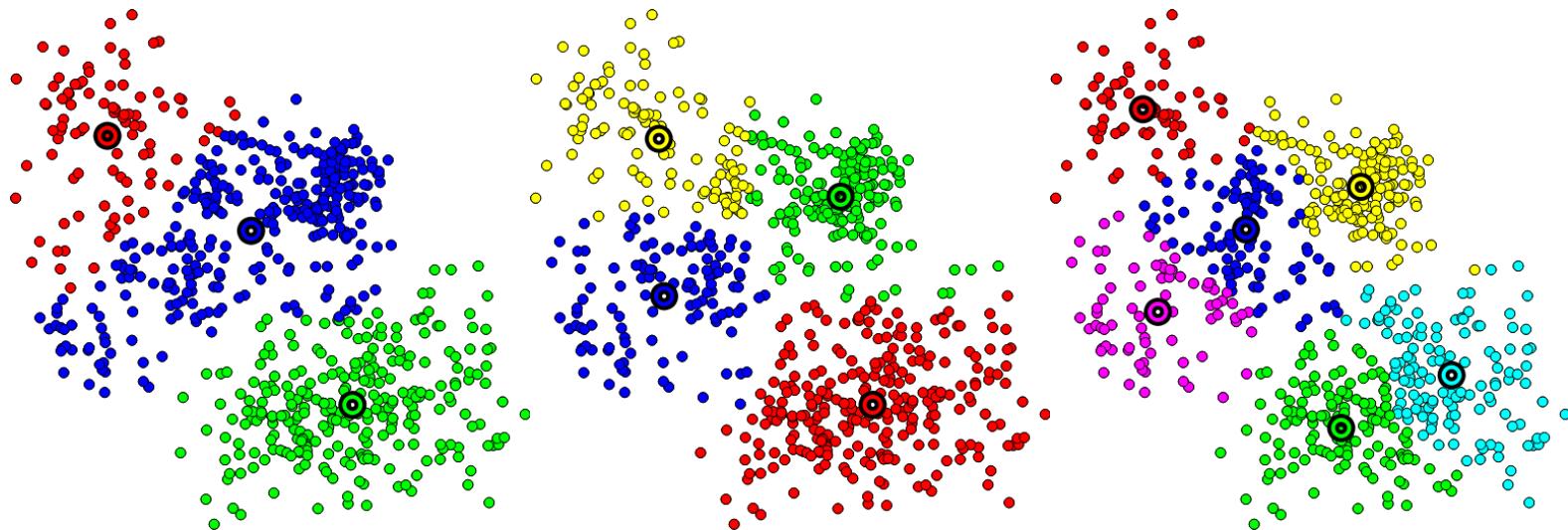
Medida R -cuadrado

$$R^2 = \frac{\sum_{\mathbf{x}_i} \|\mathbf{x}_i - \mathbf{c}\|^2 - \sum_{k=1}^K \sum_{\mathbf{x}_{i'} \in C_k} \|\mathbf{x}_{i'} - \mathbf{c}_k\|^2}{\sum_{\mathbf{x}_i} \|\mathbf{x}_i - \mathbf{c}\|^2}$$

donde \mathbf{c} es el centro de todo el dataset.

- ▶ Mide la homogeneidad de los clústeres
- ▶ Acotada entre 0 (sólo un clúster) y 1 ($K = n$)
- ▶ Se incrementa fácilmente aumentando el número de clústeres resultantes K

Evaluación



Evaluación

Intrínseca

Silueta

$$S = \frac{1}{n} \sum_{x_i} \frac{b_k(x_i) - a_k(x_i)}{\max\{b_k(x_i), a_k(x_i)\}}$$

$$\Rightarrow S < 0$$

↓

bueno.

donde

Inter →

$$a_k(x_i) = \frac{1}{n_k - 1} \sum_{x_j \in C_k: x_j \neq x_i} d(x_i, x_j)$$

y

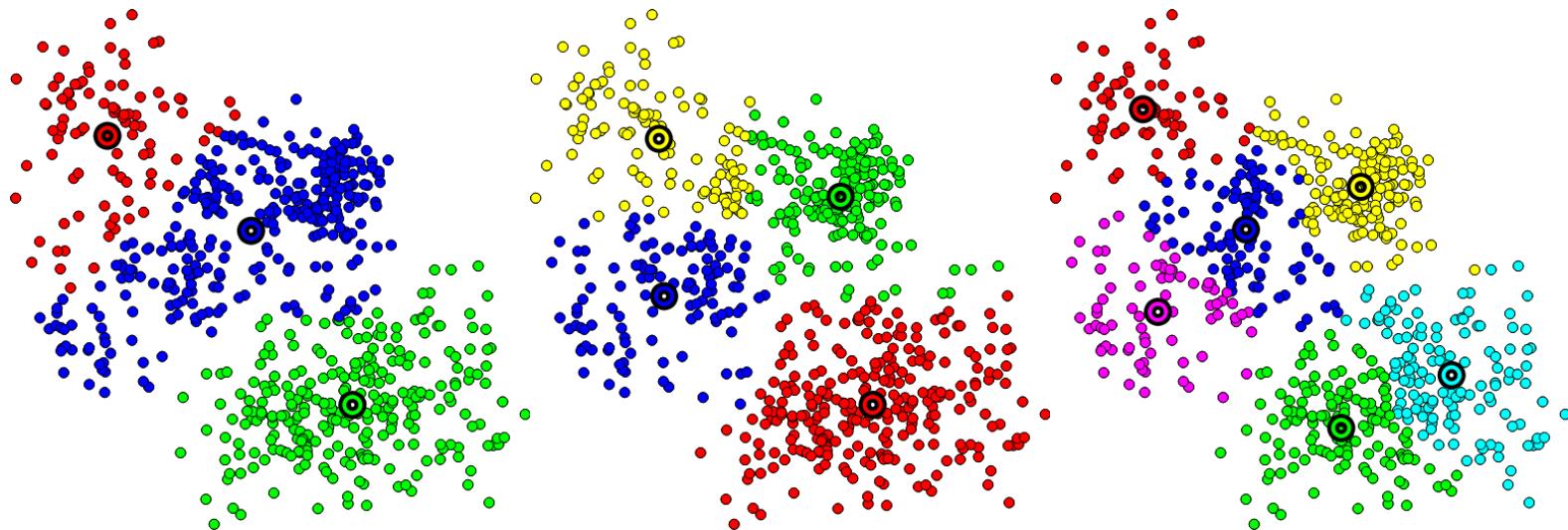
Intra →

$$b_k(x_i) = \min_{h \neq k} \frac{1}{n_h} \sum_{x_j \in C_h} d(x_i, x_j)$$

- ▶ Diferencia normalizada entre la distancia intraclúster y la interclúster
- ▶ Acotada entre -1 y 1

Busco { Intracluster ↑↑
 Intercluster ↓↓

Evaluación



Evaluación

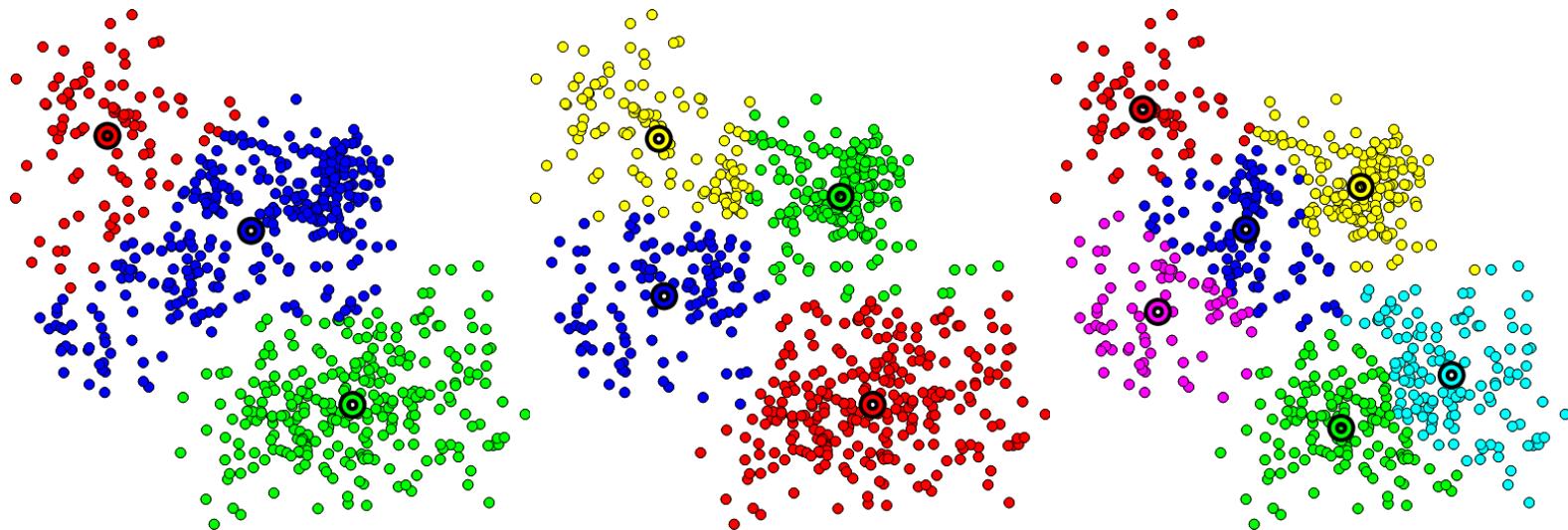
Intrínseca

Índice Calinski-Harabasz:

$$CH = \frac{(n - K) \sum_{k=1}^K n_k \cdot d(\mathbf{c}_k, \mathbf{c})^2}{(K - 1) \sum_{k=1}^K \sum_{\mathbf{x}_i \in C_k} d(\mathbf{x}_i, \mathbf{c}_k)^2}$$

- ▶ Suma promedio de las distancias inter e intraclúster al cuadrado
- ▶ A mayor valor, mejor agrupamiento

Evaluación



Evaluación

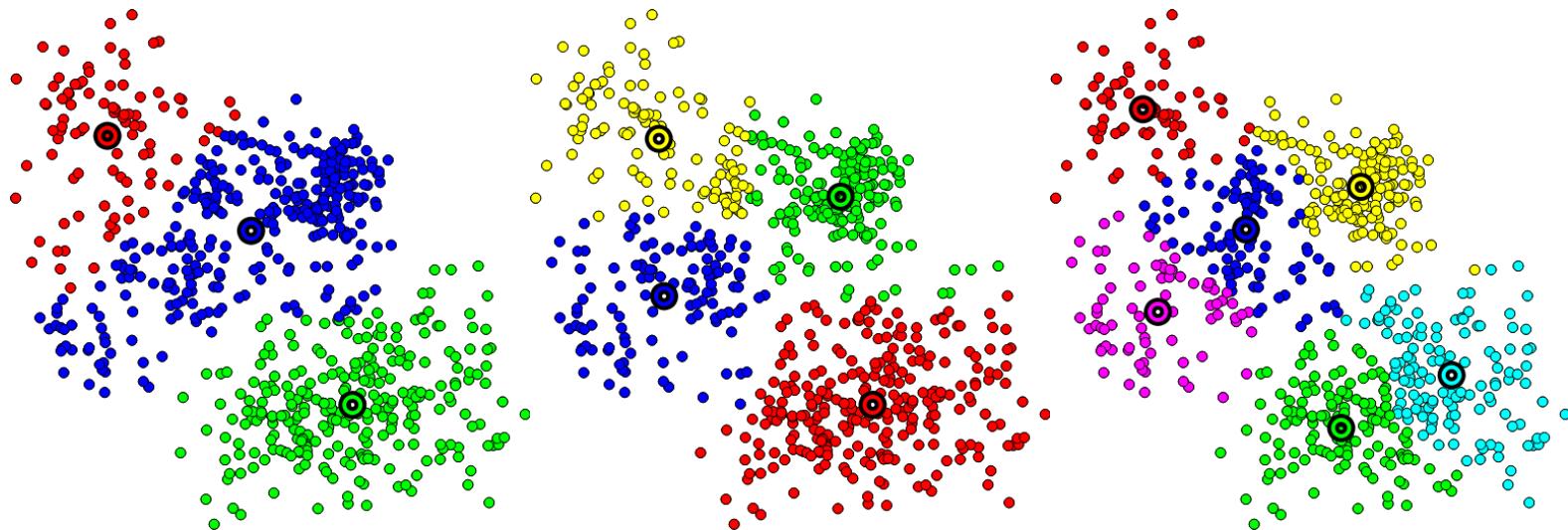
Intrínseca

Índice I:

$$I = \left(\frac{\sum_{\mathbf{x}_i} d(\mathbf{x}_i, \mathbf{c})}{K \cdot \sum_{k=1}^K \sum_{\mathbf{x}_i \in C_k} d(\mathbf{x}_i, \mathbf{c}_k)} \cdot \max_{i,j \in \{1, \dots, K\}} d(\mathbf{c}_i, \mathbf{c}_j) \right)^p$$

- ▶ Mide la separación interclúster con respecto a la homogeneidad intraclúster
- ▶ A mayor valor, mejor agrupamiento

Evaluación



Aprendizaje no supervisado

VC02: Inicialización de K-means

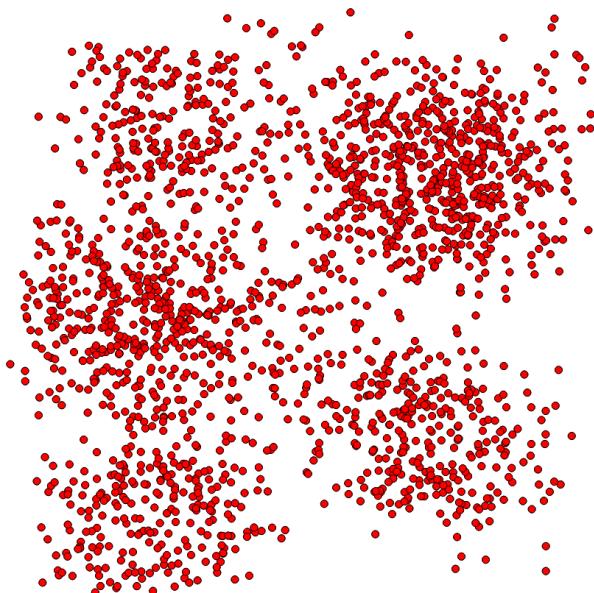
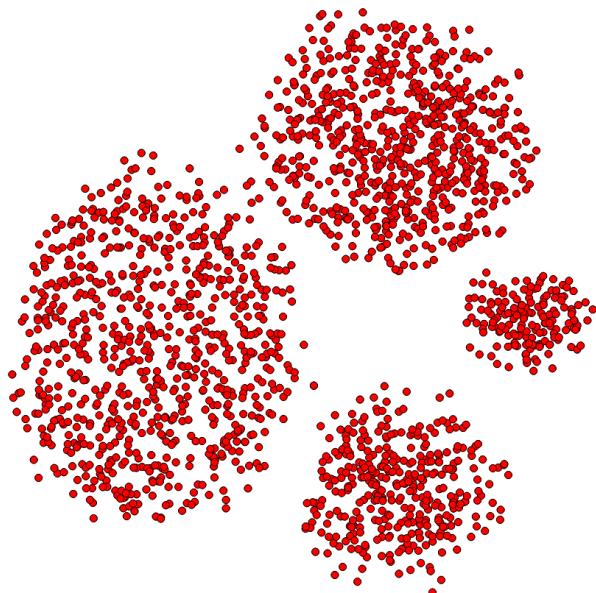
Rocío del Amor del Amor
mrocio.delamor@campusviu.es

Universidad Internacional de Valencia

Agrupamiento

Definición

Dado un conjunto de datos, el agrupamiento trata de identificar subgrupos homogéneos de ejemplos que manifiestan diferencias relevantes con los otros subgrupos que se formen.



Agrupamiento

Definición

Dado un conjunto de datos, el agrupamiento trata de identificar subgrupos homogéneos de ejemplos que manifiestan diferencias relevantes con los otros subgrupos que se formen.

→ *características extraídas de los datos.*

- ▶ Vectores descriptores de los ejemplos
- ▶ Conjunto de datos
- ▶ No existe ninguna variable “especial” respuesta (*ground truth*)
- ▶ Formar grupos:
 - * No se conoce el número de grupos
 - * No se conocen las pertenencias de ejemplos a grupos

Agrupamiento

Dos instrucciones:

Definición

Dado un conjunto de datos, el agrupamiento trata de identificar **subgrupos homogéneos** de ejemplos que manifiestan **diferencias relevantes con los otros subgrupos** que se formen.

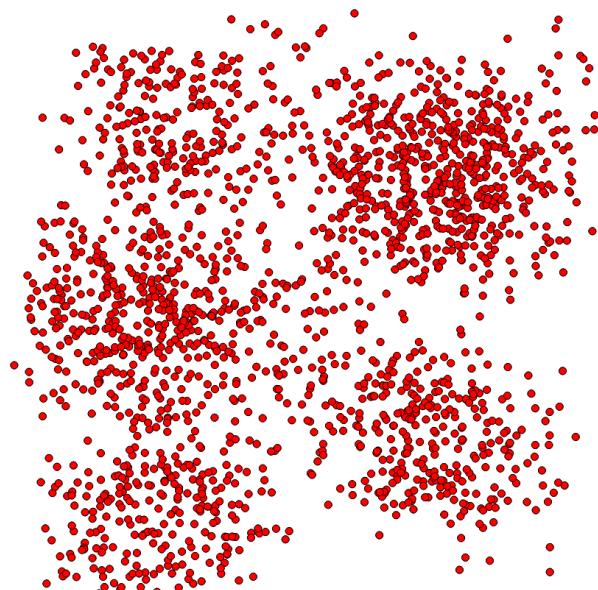
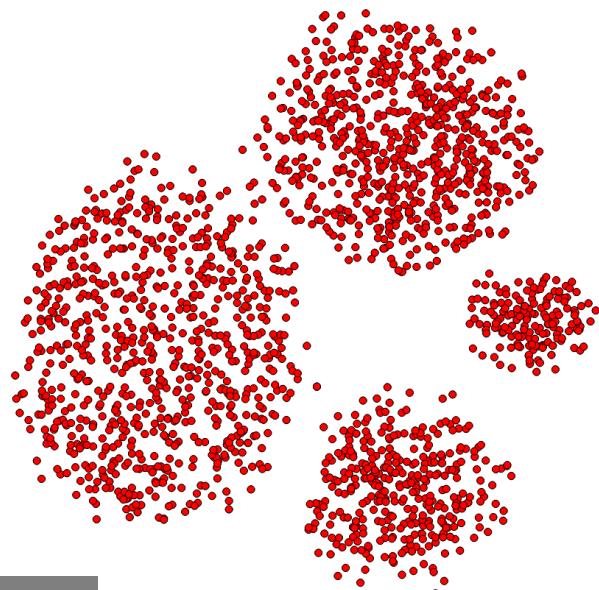
- ▶ **Dispersión intraclúster**
- ▶ **Dispersión interclúster**

Agrupamiento

Dos instrucciones:

Definición

Dado un conjunto de datos, el agrupamiento trata de identificar **subgrupos homogéneos** de ejemplos que manifiestan **diferencias relevantes con los otros subgrupos** que se formen.



Agrupamiento

Objetivo

Encontrar un agrupamiento que maximice la dispersión interclúster y minimice la dispersión intraclúster:

- Dispersión intraclúster *lor de dentro de un cluster*

$$I(C) = \frac{1}{2} \sum_{k=1}^K \sum_{i:C(x_i)=k} \sum_{i':C(x_{i'})=k} d(x_i, x_{i'})$$

- Dispersión interclúster *respecto a todos los otros (^{todos los} clusters)*

$$O(C) = \frac{1}{2} \sum_{k=1}^K \sum_{i:C(x_i)=k} \sum_{i':C(x_{i'}) \neq k} d(x_i, x_{i'})$$

Agrupamiento

Objetivo

Encontrar un agrupamiento que maximice la dispersión interclúster
y minimice la dispersión intraclúster:

¡Ambos objetivos son equivalentes!

Agrupamiento

Objetivo

Encontrar un agrupamiento que maximice la dispersión interclúster
y minimice la dispersión intraclúster:

$$T = \frac{1}{2} \sum_{i=1}^n \sum_{i'=1}^n d(x_i, x_{i'})$$

= constante

Agrupamiento

Objetivo

Encontrar un agrupamiento que maximice la dispersión interclúster y minimice la dispersión intraclúster:

$$T = \frac{1}{2} \sum_{k=1}^K \sum_{i:C(x_i)=k} \left(\sum_{i':C(x_{i'})=k} d(x_i, x_{i'}) + \sum_{i':C(x_{i'}) \neq k} d(x_i, x_{i'}) \right)$$

Agrupamiento

Objetivo

Encontrar un agrupamiento que maximice la dispersión interclúster y minimice la dispersión intraclúster:

$$T = I(C) + O(C)$$

$$\arg \min_C I(C) = \arg \min_C T - O(C) = \arg \max_C O(C)$$



Agrupamiento

El objetivo es buscar el mejor agrupamiento C que maximiza (minimiza) la dispersión intraclúster (interclúster)

$$\arg \min_C I(C) = \arg \min_C T - O(C) = \arg \max_C O(C)$$

Agrupamiento

El objetivo es buscar el mejor agrupamiento C que maximiza (minimiza) la dispersión intraclúster (interclúster)

$$\arg \min_C I(C) = \arg \min_C T - O(C) = \arg \max_C O(C)$$

¡Número inabordable de posibles combinaciones!

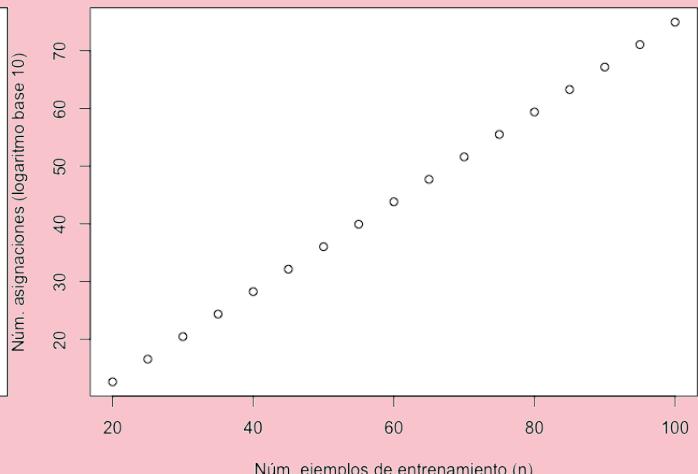
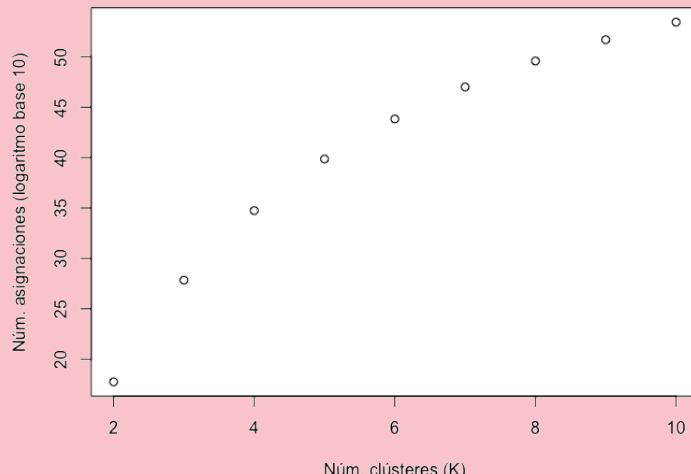
$$S(n, K) = \frac{1}{K!} \sum_{k=1}^K (-1)^{K-k} \binom{K}{k} k^n$$

Agrupamiento

El objetivo es buscar el mejor agrupamiento C que maximiza (minimiza) la dispersión intraclúster (interclúster)

$$\arg \min_C I(C) = \arg \min_C T - O(C) = \arg \max_C O(C)$$

¡Número inabordable de posibles combinaciones!



Agrupamiento

¡Necesario recurrir a heurísticas de búsqueda!

¡Necesario recurrir a heurísticas de búsqueda!

Heurística

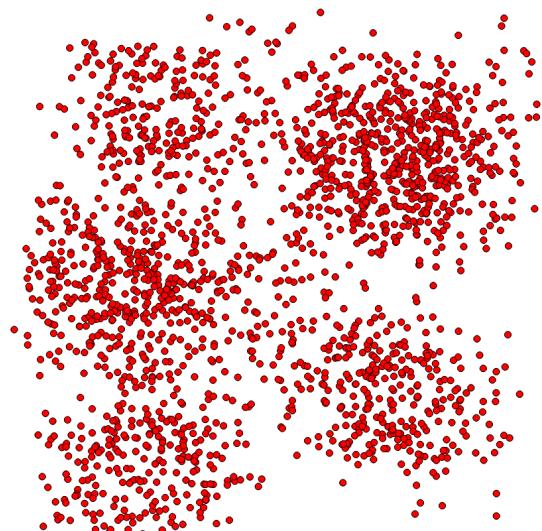
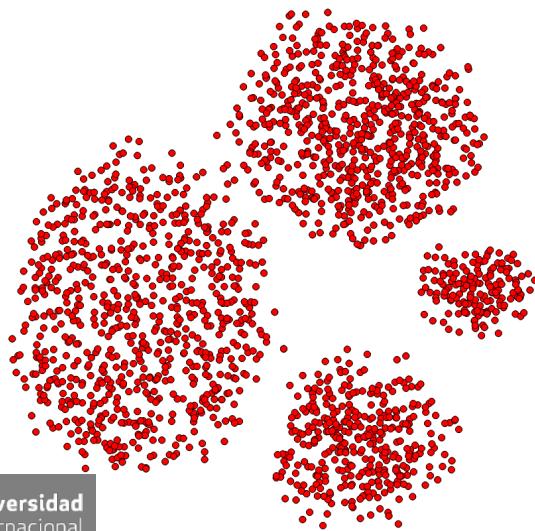
En informática, se trata de técnicas diseñadas para resolver un problema de manera rápida cuando la aproximación exhaustiva es muy lenta y/o para encontrar una solución aproximada cuando encontrar la solución exacta es muy difícil o imposible.

Se puede expresar como un *trade-off (balance)* entre velocidad y optimalidad-completitud.

Agrupamiento

Heurísticas de búsqueda del mejor agrupamiento

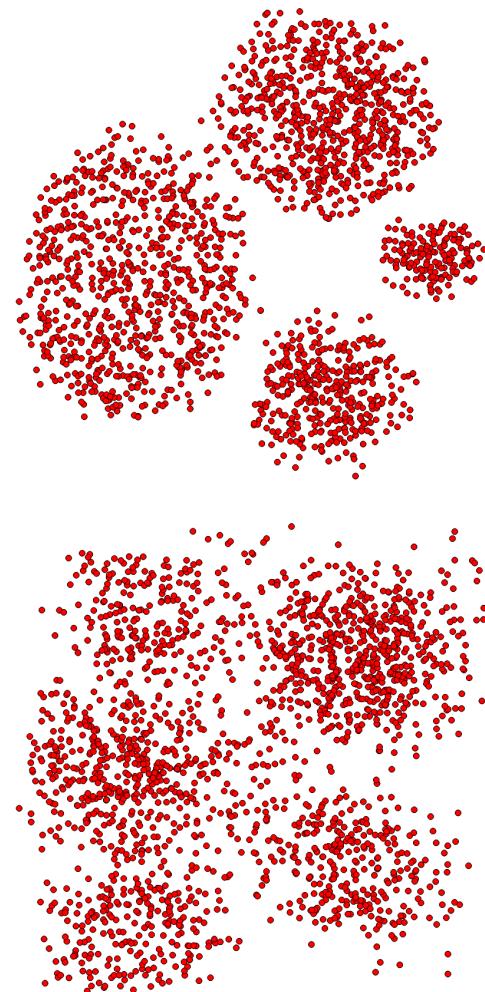
1. Encontrar un agrupamiento válido
2. Plantear diferentes alternativas a ese agrupamiento
3. Escoger la mejor alternativa
4. Volver al paso 2



Agrupamiento

Tipos de algoritmos de agrupamiento

- ▶ Basados en particiones
- ▶ Jerárquicos
- ▶ Espectrales
- ▶ Basados en densidad
- ▶ Probabilísticos



Agrupamiento basado en particiones

Búsqueda de la mejor partición de los datos

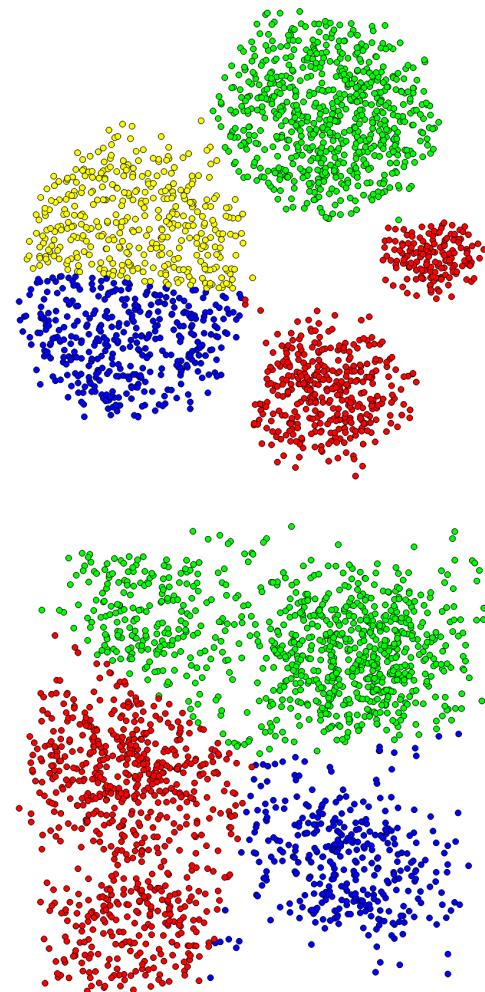
Se partitiona el dataset según criterios basados en distancia

** Uso de centro(ide)s

** ¿Fijar el número de clústeres (K)?

Algoritmos:

- ▶ K -means
- ▶ K -medoids



Agrupamiento basado en particiones

K-means

Intuición

Los clústeres homogéneos se agrupan alrededor de un centro. Por lo tanto, se puede calcular:

1. **Centro**: El centro de un clúster es la media de los elementos que pertenecen al él
2. **Pertenencia**: Un ejemplo pertenece al clúster cuyo centro le es más cercano

** La combinación de ambos conceptos permite construir el agrupamiento

Agrupamiento basado en particiones

K-means

Dispersión intraclúster

$$I(C) = \sum_{k=1}^K N_k \cdot \sum_{x_i: C(x_i)=k} ||x_i - \bar{x}_k||^2$$

Objetivo a minimizar

$$\arg \min_{C: (\bar{x}_1, \dots, \bar{x}_K)} I(C)$$

Agrupamiento basado en particiones

K-means

Heurística

Partiendo de un conjunto de centros aleatorio, buscar la pertenencia más probable de los ejemplos a los clústeres y obtener un nuevo conjunto de centros (agrupamiento)

¡Naturaleza iterativa!

Agrupamiento basado en particiones

K-means

EXAMEN

K-means

→ input

Recibe: Conjunto de entrenamiento, $\{x_1, \dots, x_n\}$; número de clústeres, K

1. Elección (aleatoria) de K puntos del conjunto de entrenamiento como centros, $\{\bar{x}_1, \dots, \bar{x}_K\}$.

2. Asignar cada ejemplo x_i al clúster del centro más cercano:
 $C(x_i) = \operatorname{argmin}_{k \in \{1, \dots, K\}} \|x_i - \bar{x}_k\|^2$

3. Para cada clúster k , recalcular su centro: $\bar{x}_k = \operatorname{argmin}_x \sum_{x_i: C(x_i)=k} \|x_i - x\|^2$

4. Los pasos 2 y 3 se iteran hasta que los centros no cambian.

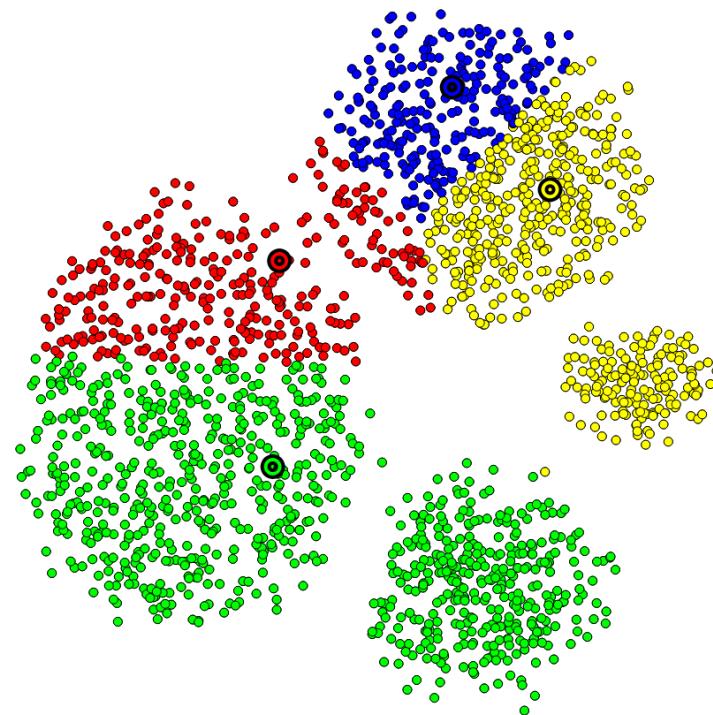
Devuelve: Conjunto de centros, $\{\bar{x}_1, \dots, \bar{x}_K\}$; Asignación, C

Lo que hace el Algor.

→ output

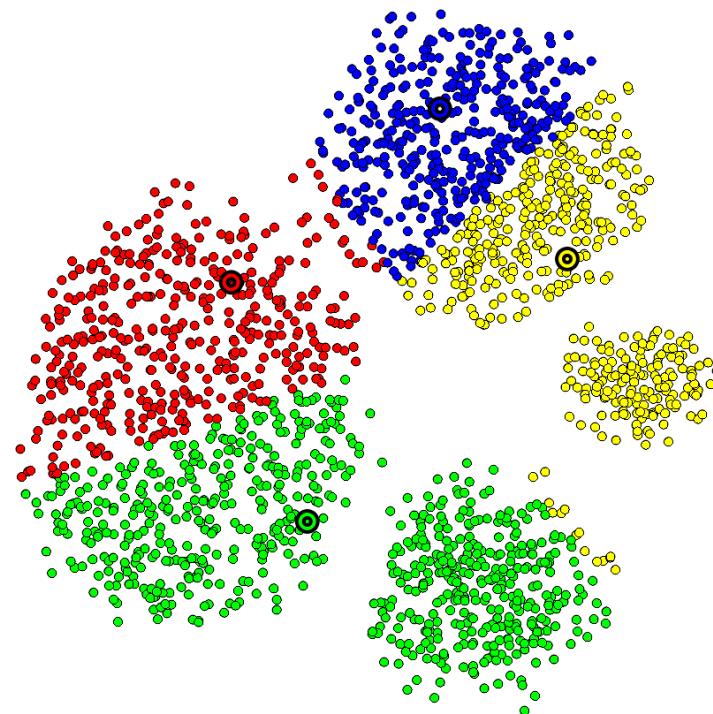
Agrupamiento basado en particiones

K-means



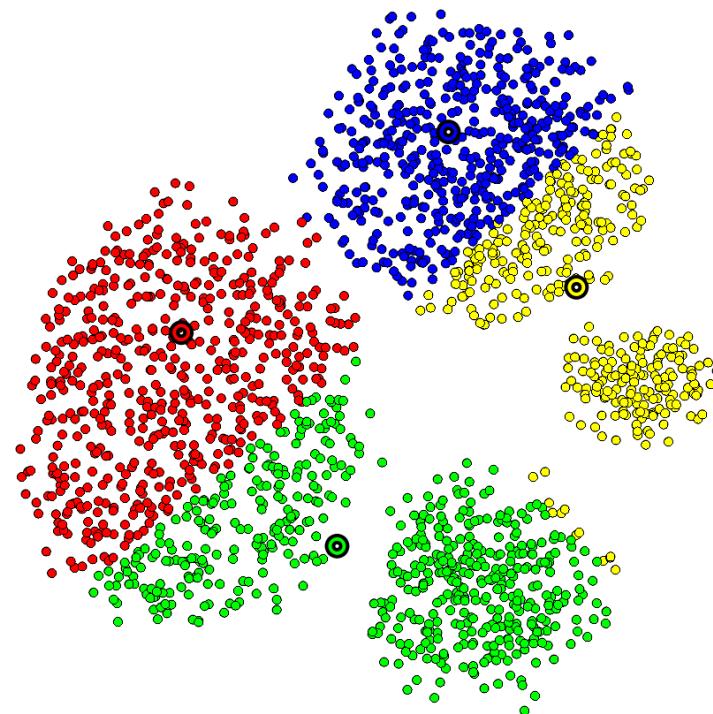
Agrupamiento basado en particiones

K-means



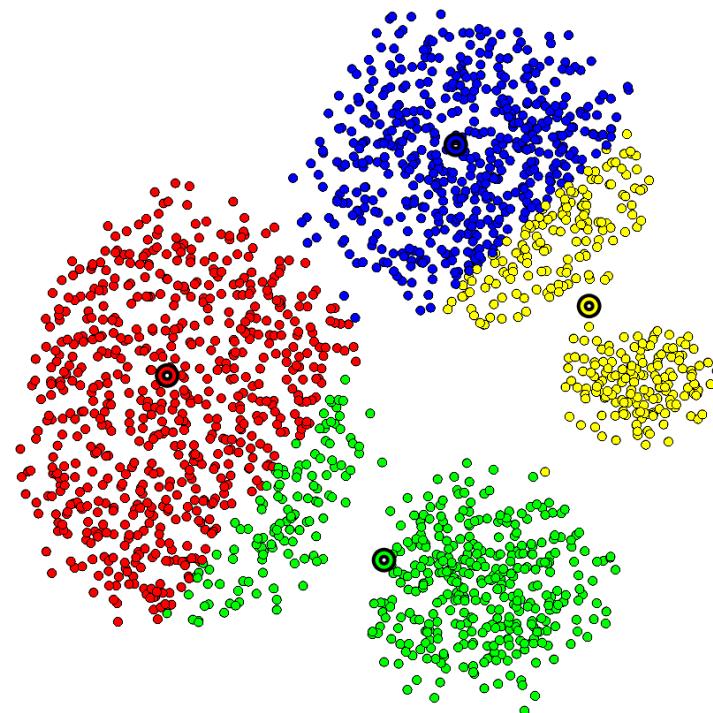
Agrupamiento basado en particiones

K-means



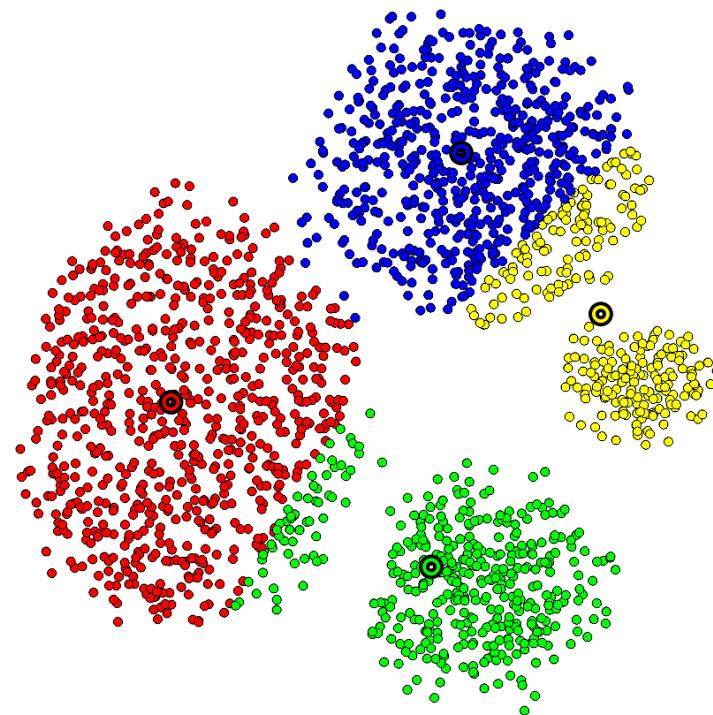
Agrupamiento basado en particiones

K-means



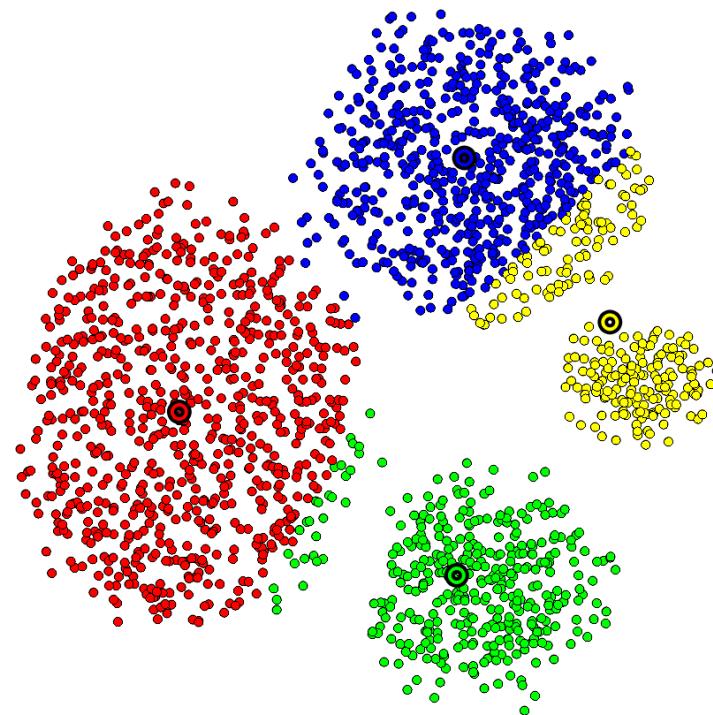
Agrupamiento basado en particiones

K-means



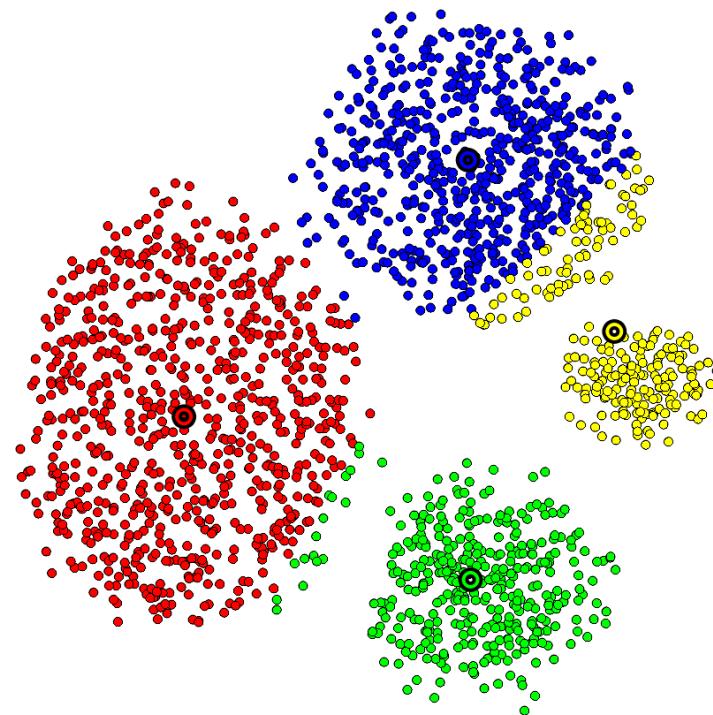
Agrupamiento basado en particiones

K-means



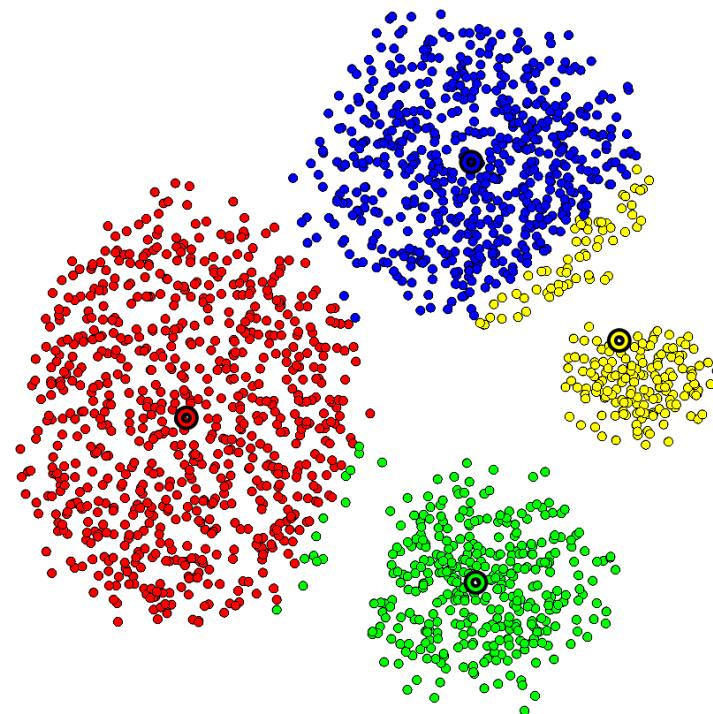
Agrupamiento basado en particiones

K-means



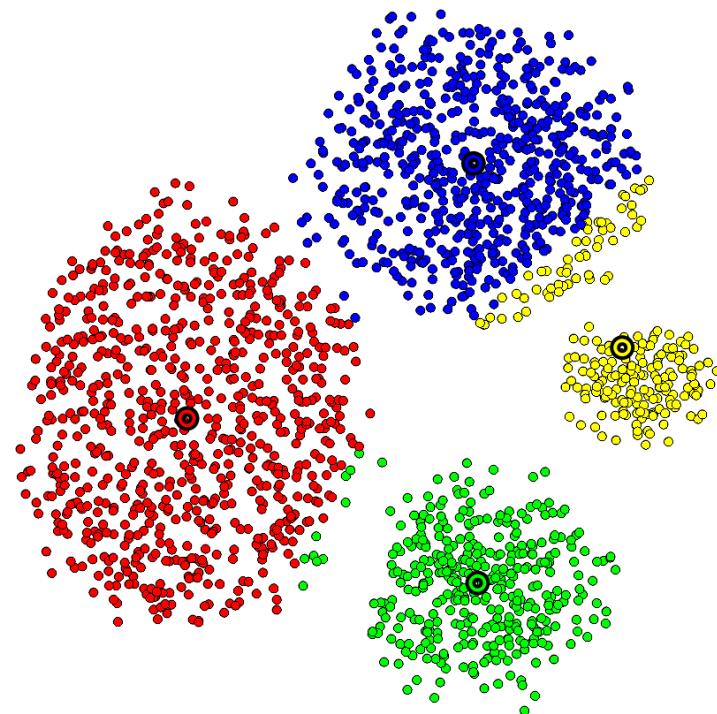
Agrupamiento basado en particiones

K-means



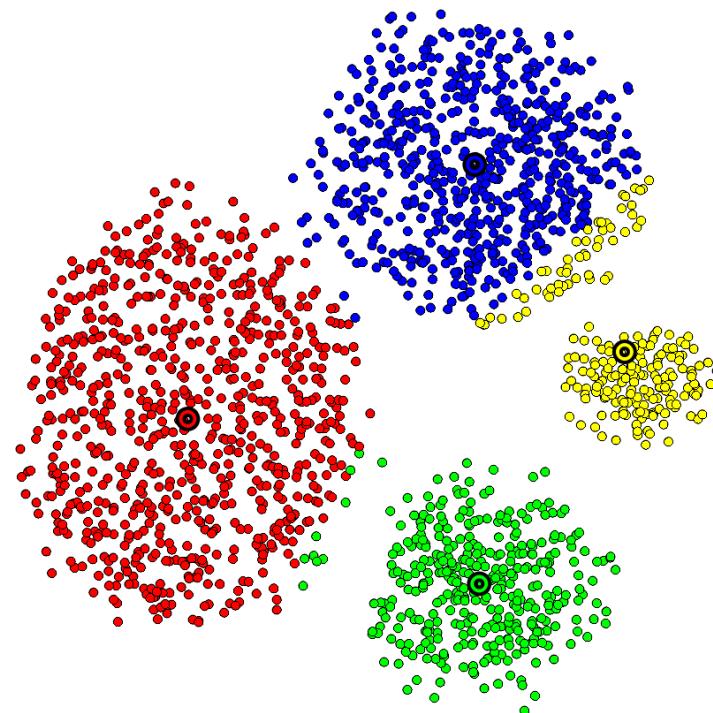
Agrupamiento basado en particiones

K-means



Agrupamiento basado en particiones

K-means



Agrupamiento basado en particiones

K-means

Ventajas

- ▶ Intuitivo
- ▶ Rápido
- ▶ Sencillo
- ▶ Mejorable → no es una ventaja como tal.

¡Algoritmo de *clustering* más popular!

Agrupamiento basado en particiones

K-means

Desventajas

- ▶ El número de clústeres es un parámetro (K)
- ▶ Dependencia de la inicialización
- ▶ Dependencia de los outliers *→ hacen que se mueva el centro*
- ▶ Funciona con variables descriptivas continuas *→ se hace la media*
- ▶ Dificultades cuando los clústeres son de diferente tamaño y/o densidad, o no son convexos *→ ver ejemplo de diapositivas anteriores*

Agrupamiento basado en particiones

K-means

Desventajas

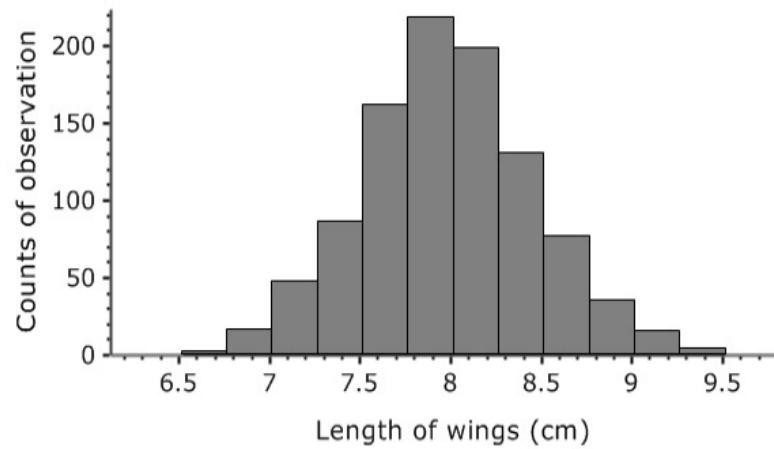
- ▶ El número de clústeres es un parámetro (K)
Validación cruzada
- ▶ Dependencia de la inicialización
Múltiples ejecuciones del algoritmo, K-means++
- ▶ Dependencia de los outliers
Preproceso
- ▶ Funciona con variables descriptivas continuas
K-medoids
- ▶ Dificultades cuando los clústeres son de diferente tamaño y/o densidad, o no son convexos

↳ otros algoritmos más potentes

Agrupamiento basado en particiones

K-medoids

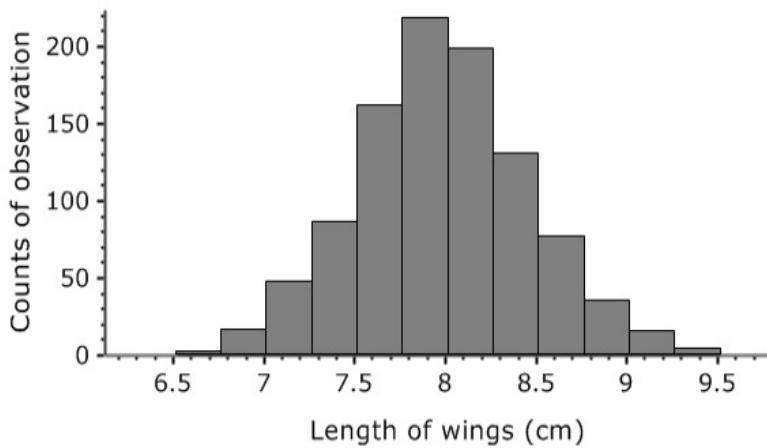
Histogram



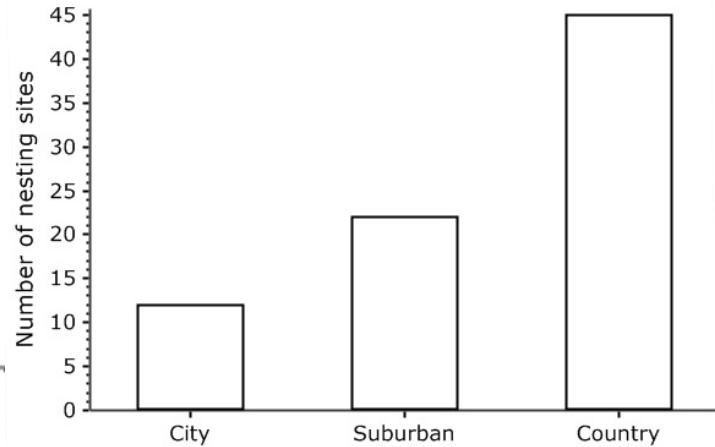
Agrupamiento basado en particiones

K-medoids

Histogram



Bar graph



Agrupamiento basado en particiones

K-medoids

Intuición

La idea iterativa de K-means

Se cambia el centro por el centriode:

Los centros son, en todo momento, ejemplos del conjunto de entrenamiento

→ punto de la región del cluster que existe en los datos
!!! (punto del conjunto de entrenamiento)

Agrupamiento basado en particiones

K-medoids

K-means

Recibe: Conjunto de entrenamiento, $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$; número de clústeres, K

-
1. Elección (aleatoria) de K puntos del conjunto de entrenamiento como centros, $\{\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_K\}$.
 2. Asignar cada ejemplo \mathbf{x}_i al clúster del centro más cercano:
 $C(\mathbf{x}_i) = \operatorname{argmin}_{k \in \{1, \dots, K\}} \|\mathbf{x}_i - \bar{\mathbf{x}}_k\|^2$
 3. Para cada clúster k , recalcular su centro: $\bar{\mathbf{x}}_k = \operatorname{argmin}_{\mathbf{x}} \sum_{\mathbf{x}_i: C(\mathbf{x}_i)=k} \|\mathbf{x}_i - \mathbf{x}\|^2$
 4. Los pasos 2 y 3 se iteran hasta que los centros no cambian.

Devuelve: Conjunto de centros, $\{\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_K\}$; Asignación, C

Agrupamiento basado en particiones

K-medoids

EXAMEN

K-medoids

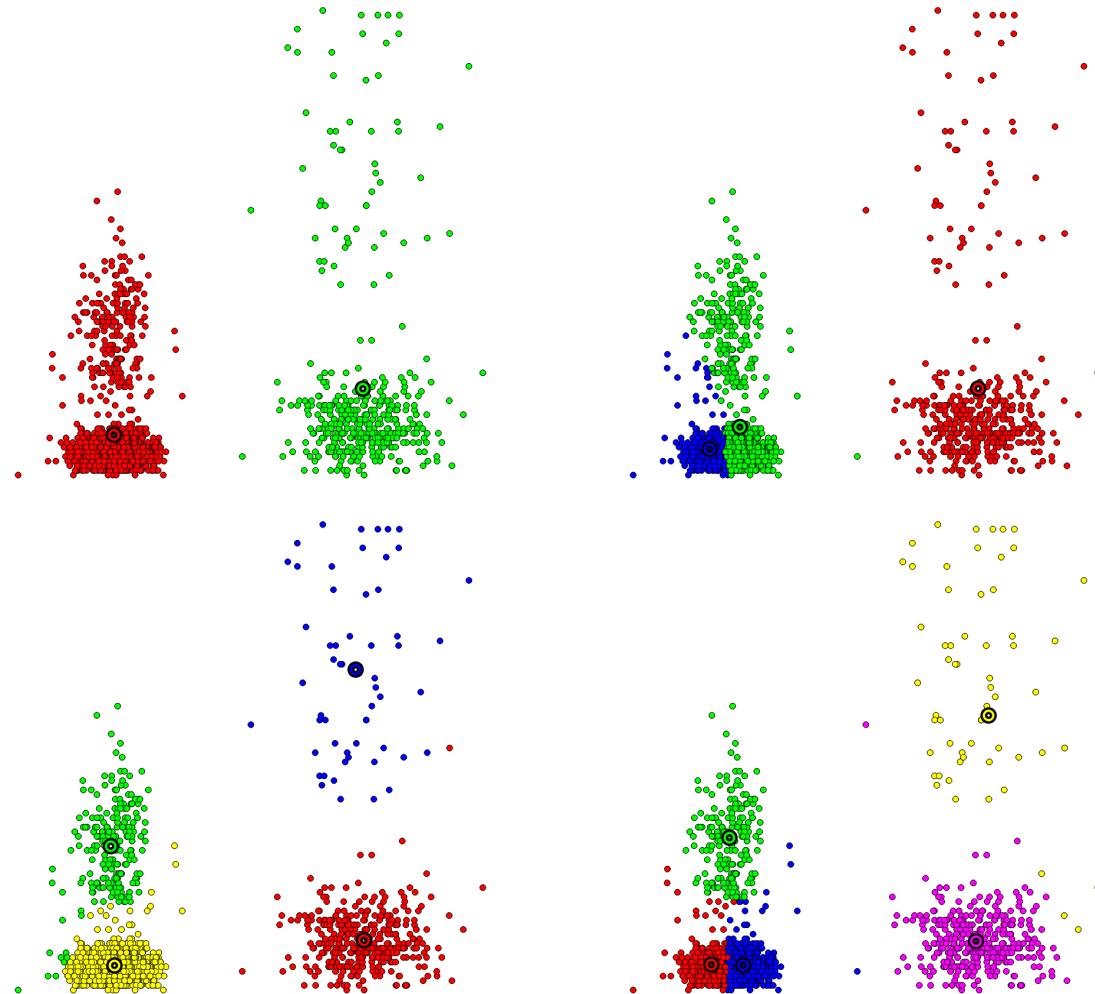
Recibe: Conjunto de entrenamiento, $\{x_1, \dots, x_n\}$; número de clústeres, K

1. Elección (aleatoria) de K puntos del conjunto de entrenamiento como medoides, $\{\tilde{x}_1, \dots, \tilde{x}_K\}$.
2. Asignar cada ejemplo x_i al clúster del medoide más cercano:
 $C(x_i) = \operatorname{argmin}_{k \in \{1, \dots, K\}} d(x_i, \tilde{x}_k)$ *ya no puede ser la euclídea*
3. Para cada clúster k , recalcular su medoide: $\tilde{x}_k = \operatorname{argmin}_{x: C(x)=k} \sum_{x_i: C(x_i)=k} d(x_i, x)$ *medoides*
4. Los pasos 2 y 3 se iteran hasta que los centros no cambian.

Devuelve: Conjunto de centros, $\{\tilde{x}_1, \dots, \tilde{x}_K\}$; Asignación, C

Agrupamiento basado en particiones

Elegir el número de clústeres (K)

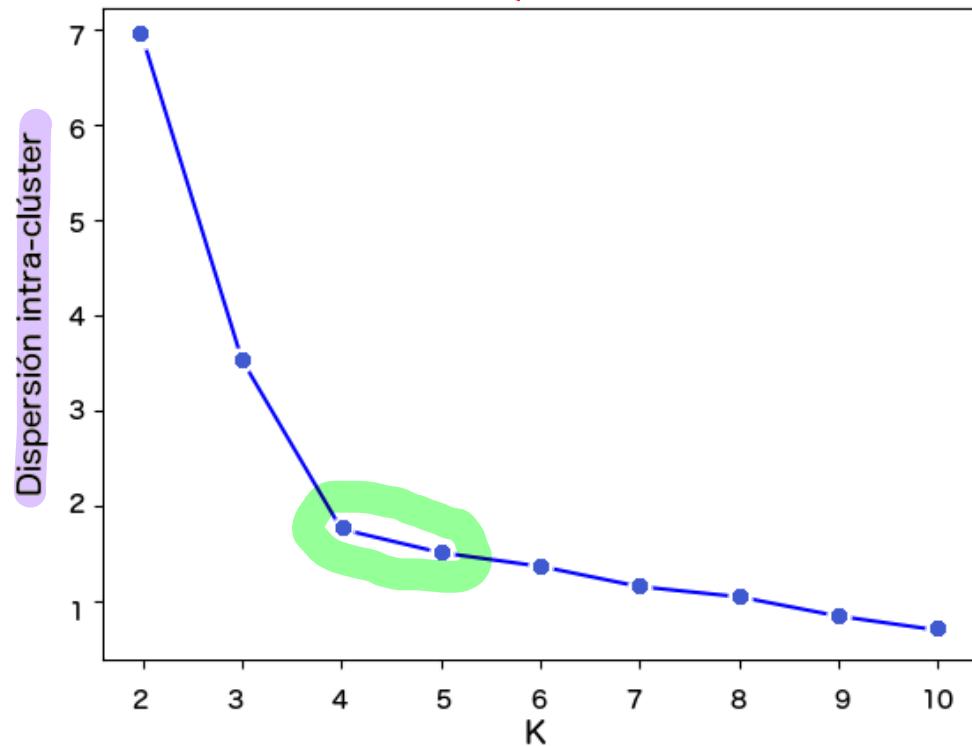


Agrupamiento basado en particiones

Elegir el número de clústeres (K)

EXAMEN

"Técnica del codo"



Agrupamiento basado en particiones

Elegir el número de clústeres (K)

Ideas:

- ▶ La dispersión intraclúster siempre se reduce
- ▶ Elegir el punto donde el cambio de tendencia es más pronunciado

Aprendizaje no supervisado

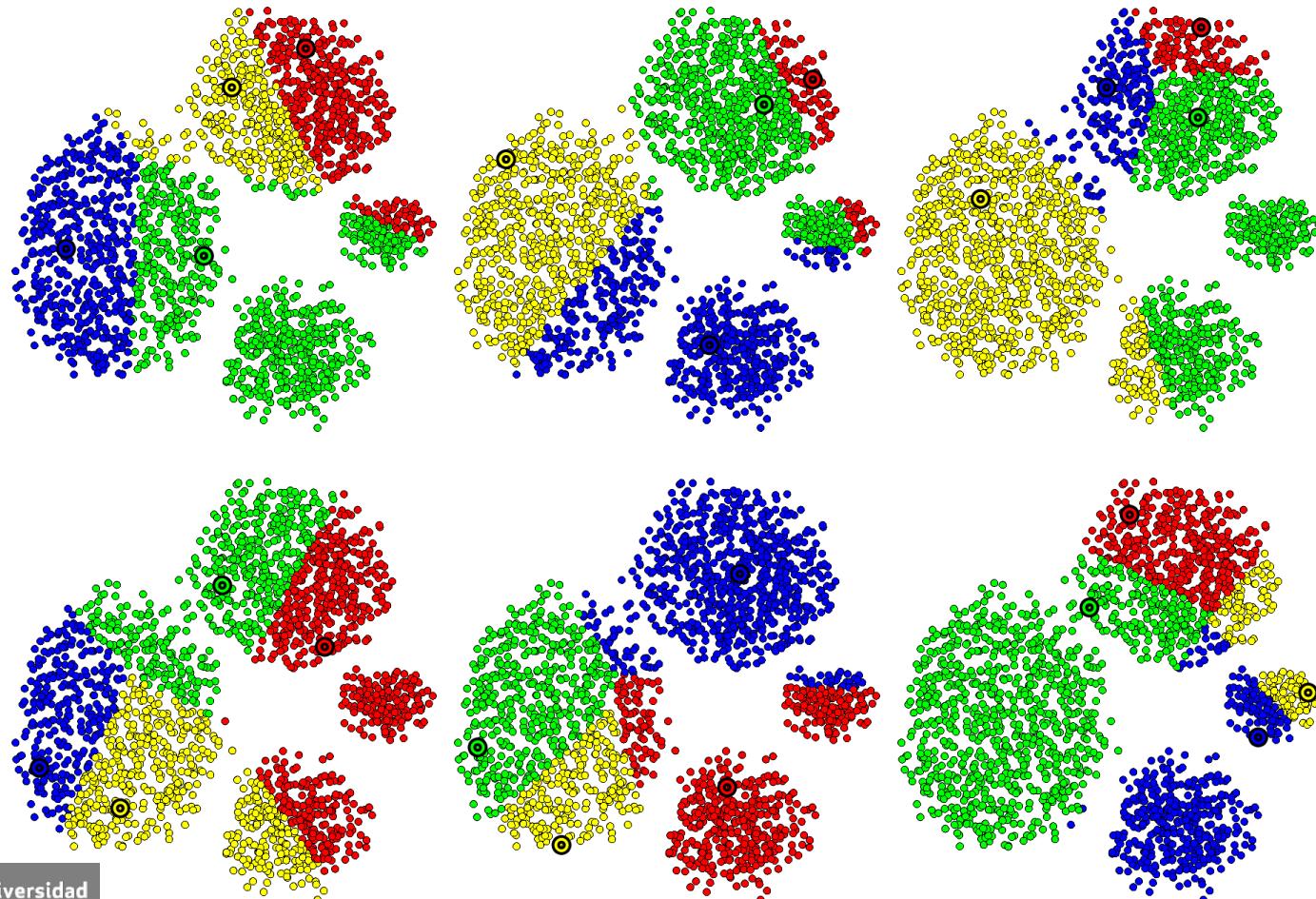
VC02: Inicialización avanzada de K-means

Rocío del Amor del Amor
mrocio.delamor@campusviu.es

Universidad Internacional de Valencia

Agrupamiento basado en particiones

Iniciar K-means



Agrupamiento basado en particiones

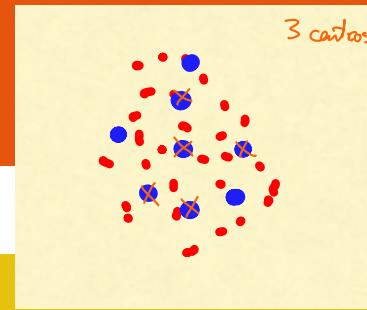
Iniciar K-means

Método basado en repeticiones

- ▶ Inicializar aleatoriamente y ejecutar K -means R veces
- ▶ Medir la bondad de los R diferentes agrupamientos
- ▶ Devolver como resultado el mejor agrupamiento

Agrupamiento basado en particiones

Iniciar K-means



K-means++

Inicialización avanzada del método K -means favoreciendo la separación de los centros iniciales

Intuición

Elección individual (y dependiente) de los centros

- ▶ **Centros:** Muestreo aleatorio no uniforme
- ▶ **Probabilidad:** Un ejemplo tiene mayor probabilidad de ser escogido como centro (inicial) cuanto mayor sea su distancia con los centros

Agrupamiento basado en particiones

K-means++

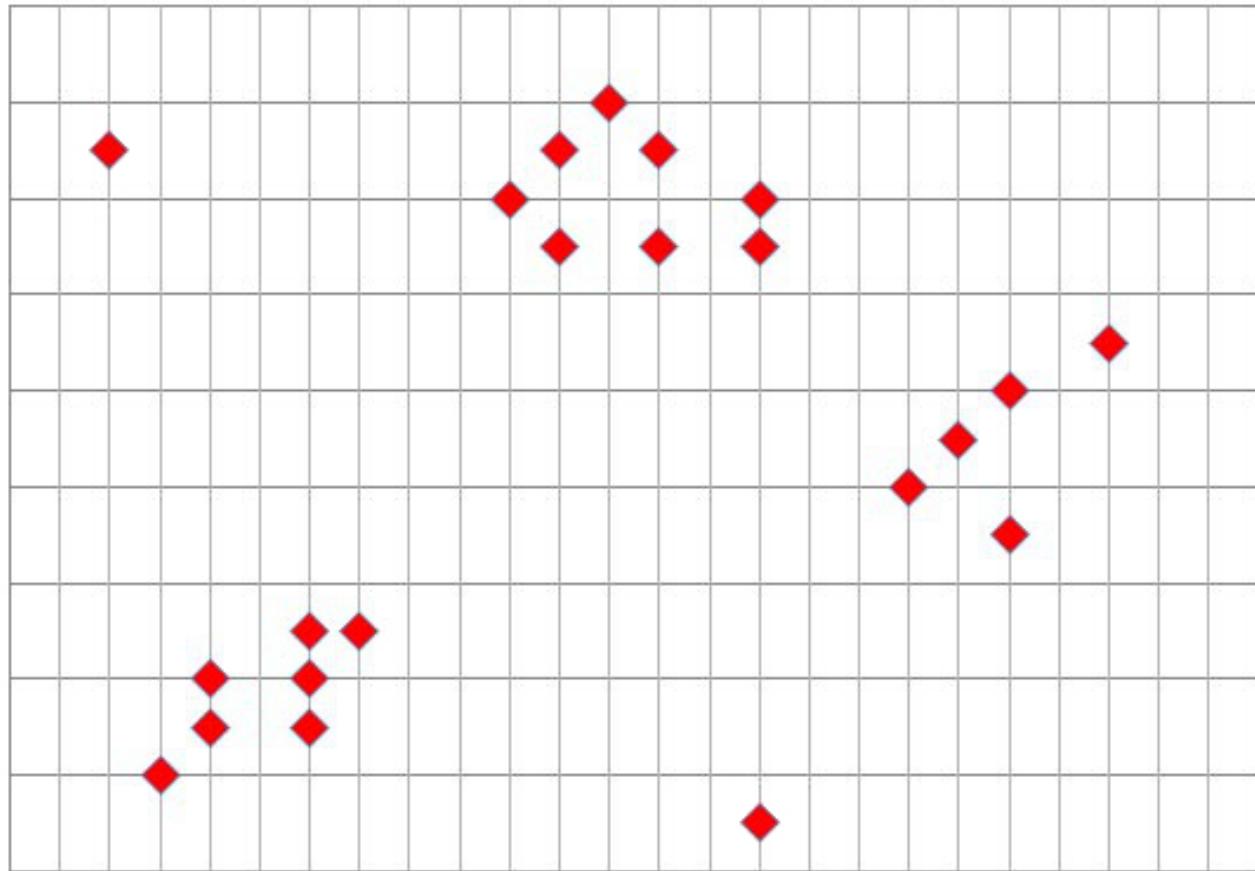
La probabilidad de un ejemplo es proporcional al cuadrado de su distancia mínima a un centro ya incluido, $D(\mathbf{x}, S)$

$$D(\mathbf{x}, S) = \min_{k \in \{1, \dots, |S|\}} \|\mathbf{x} - \bar{\mathbf{x}}_k\|^2$$

- Calculo todas las dist. a los centroides existentes y para cada punto me quedo con la mínima
 - Luego una vez finalizadas escijo la máxima de todas.

Agrupamiento basado en particiones

K-means++



Agrupamiento basado en particiones

K-means++

K-means++ (inicialización)

Recibe: Conjunto de entrenamiento, $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$; número de clústeres, K

1. Elección (aleatoria) de 1 punto del conjunto de entrenamiento como primer centro,
 $S = \{\bar{\mathbf{x}}_1\}$.
2. Mientras $|S| < K$, repetir
 - 2.1. Para todos los ejemplos de entrenamiento, calcular $D(\mathbf{x}_i, S)$, la distancia al centro más cercano: $D(\mathbf{x}_i, S) = \min_{k \in \{1, \dots, |S|\}} \|\mathbf{x}_i - \bar{\mathbf{x}}_k\|^2$
 - 2.2. Muestrear un nuevo caso \mathbf{x}' del conjunto de entrenamiento, donde el caso \mathbf{x} tiene probabilidad $D(\mathbf{x}, S)^2 / \sum_{i=1}^n D(\mathbf{x}_i, S)^2$ y añadir a S : $S = S \cup \{\mathbf{x}'\}$

Devuelve: Conjunto de centros, $\{\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_K\}$

Aprendizaje no supervisado

VC03: Agrupamiento jerárquico: Aglomerativo

Rocío del Amor del Amor

mrocio.delamor@campusviu.es

Universidad Internacional de Valencia

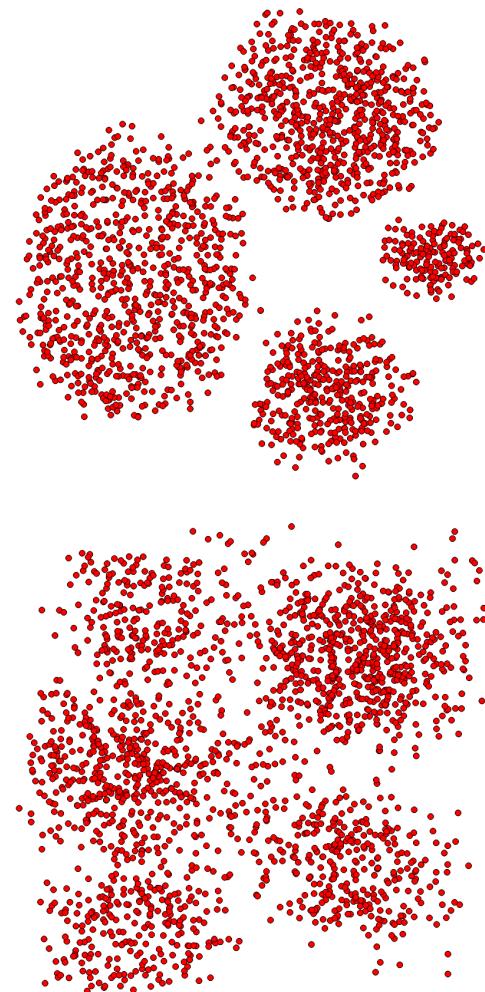
Cuando tenemos mezcla de variables categóricas y continuas podemos usar distancias distintas para ambos.

Ej. de implementación K-Prototypes

Agrupamiento

Tipos de algoritmos de agrupamiento

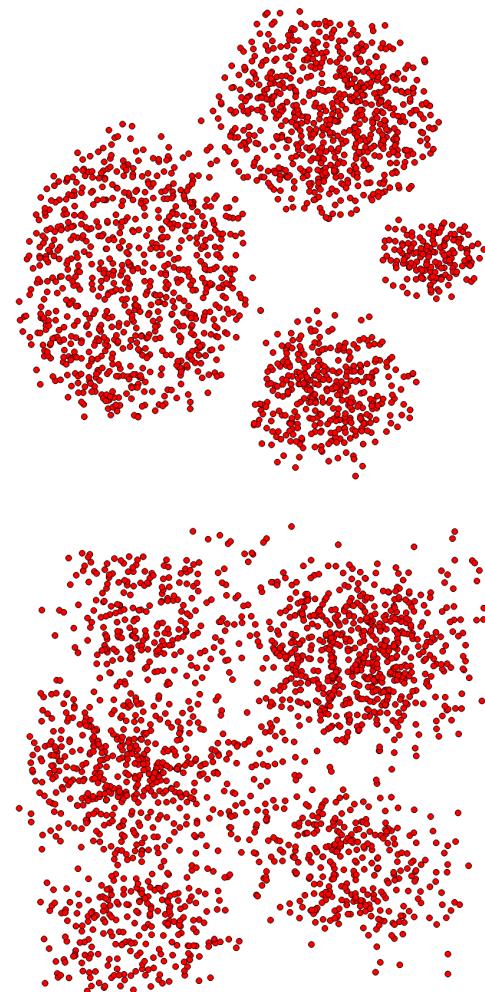
- ▶ Basados en particiones
- ▶ Jerárquicos
- ▶ Espectrales
- ▶ Basados en densidad
- ▶ Probabilísticos



Agrupamiento

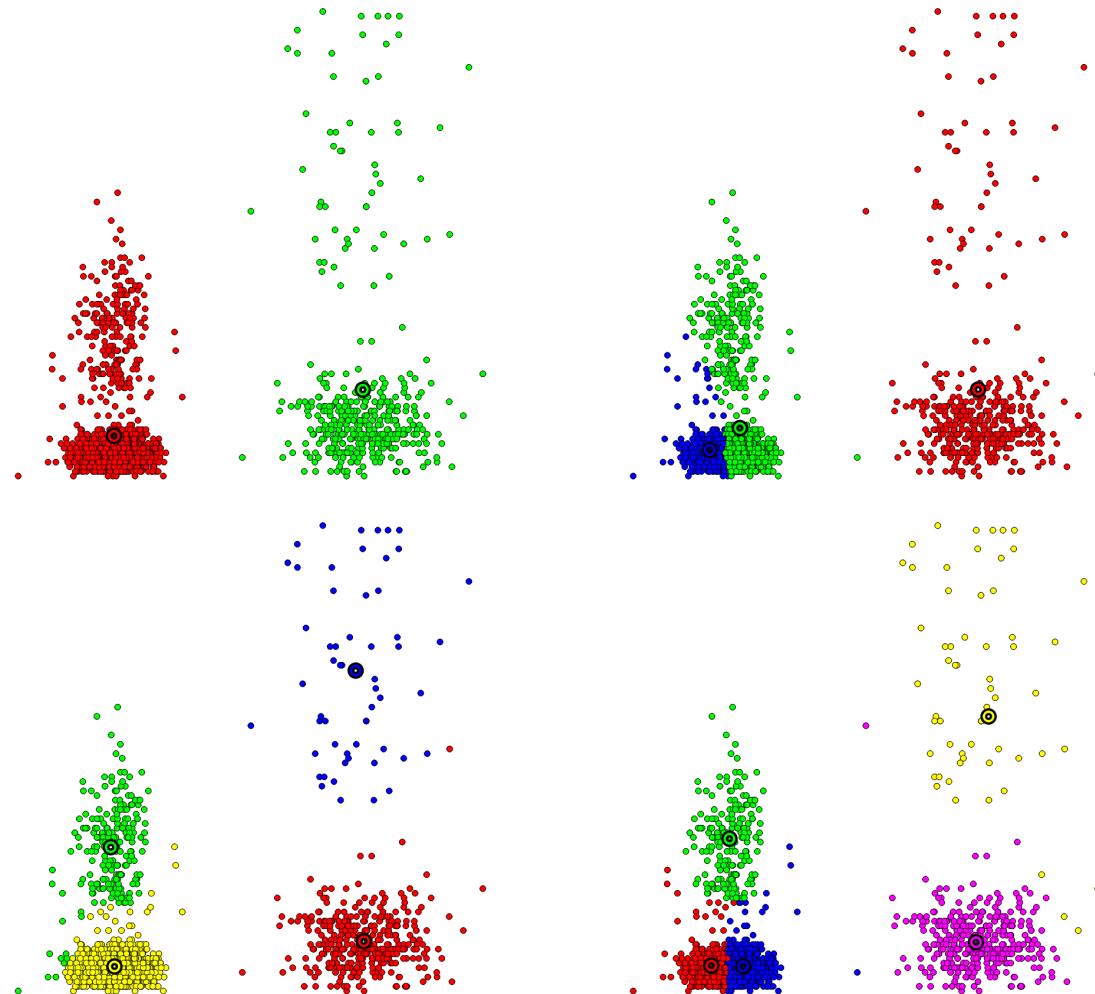
Tipos de algoritmos de agrupamiento

- ▶ Basados en particiones
- ▶ **Jerárquicos**
- ▶ Espectrales
- ▶ Basados en densidad
- ▶ Probabilísticos



Agrupamiento

Elegir el número de clústeres (K)



Agrupamiento Jerárquico

Un continuo de particiones de los datos

Se partitiona el dataset desde $K = 1$ hasta $K = n$

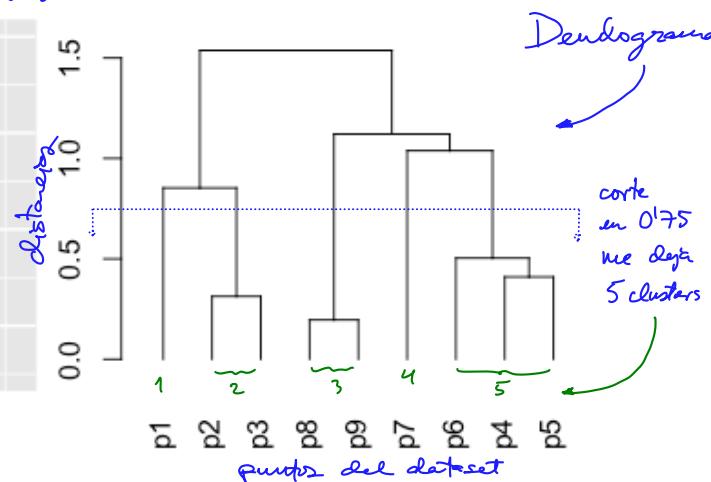
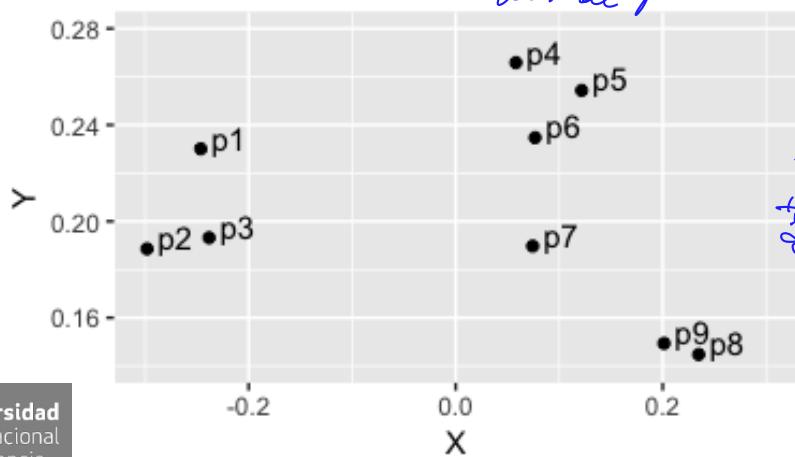
** ¿Cuál es la mejor partición?

Algoritmos:

- Aglomerativo
- Divisivo

parto de $K=N$ clusters (tanto como puntos)
y al final tengo 1 → luego veo dónde partitiono

parto de $k=1$ y llego a $k=N$ y luego
veo dónde partitiono

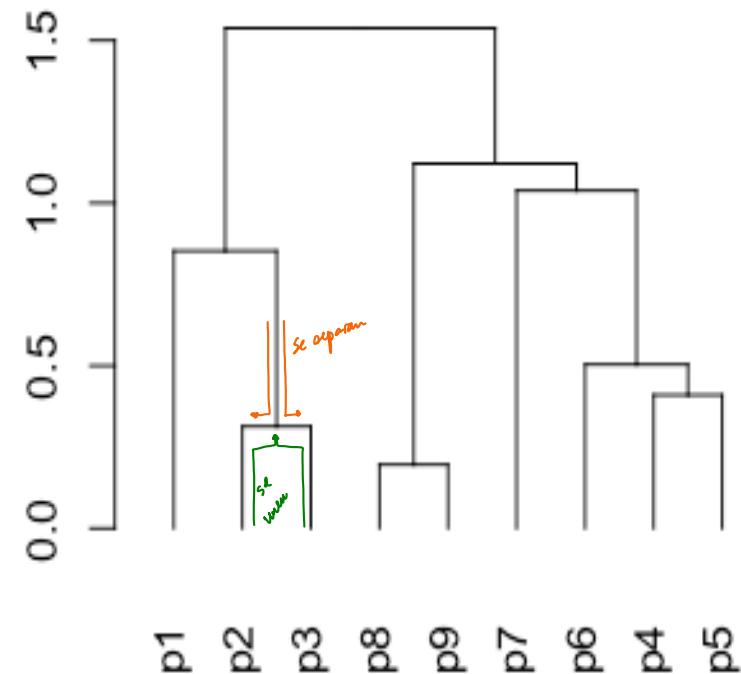


Agrupamiento Jerárquico

Dendrograma

Representación gráfica de un agrupamiento jerárquico

- ▶ Cada nodo, es un conjunto de ejemplos (clúster)
- ▶ Los clústeres se van uniendo/separando según criterios de distancia
- ▶ La longitud de las líneas verticales indica la distancia entre los clústeres que se unen/separan



Agrupamiento Jerárquico

Intuición

Si no conozco cuántos grupos/clústeres hay, de entrada no voy a elegir el número K

Los clústeres se forman de ejemplos que están cercanos entre ellos

El concepto de cercanía puede ser relativo:

1. **Términos absolutos:** La similitud entre estos dos clústeres es...
2. **Términos relativos:** Los dos clústeres más similares entre sí son...

** De manera equivalente, podemos hablar de lejanía/diferencia

Agrupamiento Jerárquico

Aglomerativo

Aglomeración

Partiendo de $K = n$, se van uniendo iterativamente pares de clústeres hasta $K = 1$ de manera voraz

0. Al principio, cada ejemplo tiene su propio clúster
1. Tras la primera unión, existen $K = n - 1$ clústeres
(todos unitarios, menos uno clúster que tiene 2 elementos)
- ...
- i. Tras la i -ésima unión, existen $K = n - i$ clústeres
- ...
- n-1. El algoritmo acaba cuando $K = 1$
(se unen los dos últimos clústeres en un clúster con todos los ejemplos)

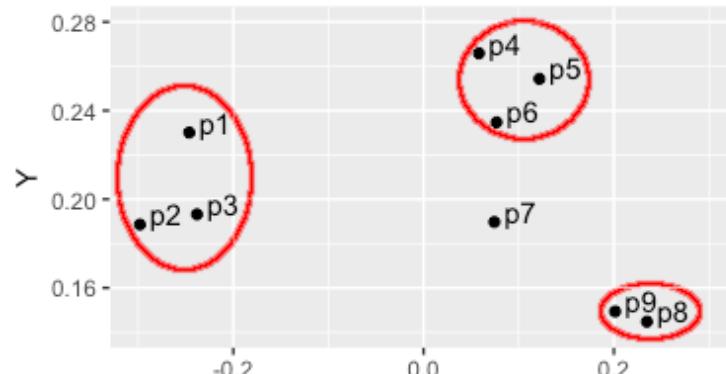
Agrupamiento Jerárquico

Aglomerativo

Dos cuestiones

A medida que avanza el algoritmo...

¿qué dos clústeres se deben unir en cada paso?



Agrupamiento Jerárquico

Aglomerativo

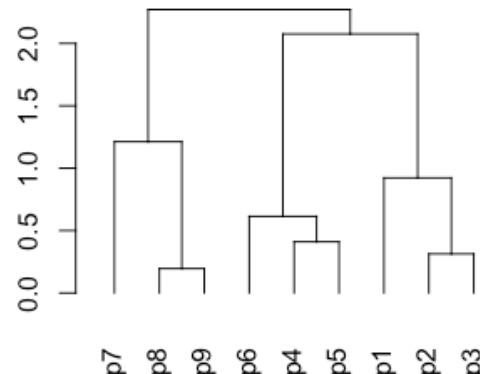
Dos cuestiones

A medida que avanza el algoritmo...

¿qué dos clústeres se deben unir en cada paso?

Al final del algoritmo, si queremos un partición concreta,

¿con qué partición nos quedamos?



Agrupamiento Jerárquico

Aglomerativo

Primera cuestión

A medida que avanza el algoritmo...

¿qué dos clústeres se deben unir en cada paso?

El par de clústeres, S_A^* y S_B^* , con menor disimilitud interclúster:

$$\{S_A^*, S_B^*\} = \arg \min_{\{S_A, S_B\}} d(S_A, S_B)$$

Agrupamiento Jerárquico

Aglomerativo

Primera cuestión

A medida que avanza el algoritmo...

¿qué dos clústeres se deben unir en cada paso?

El par de clústeres, S_A^* y S_B^* , con menor disimilitud interclúster:

$$\{S_A^*, S_B^*\} = \arg \min_{\{S_A, S_B\}} d(S_A, S_B)$$

¿cómo se mide la disimilitud interclúster?

Agrupamiento Jerárquico

Aglomerativo

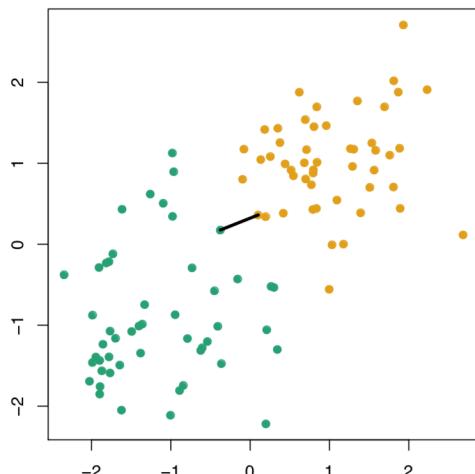
Criterios de unión

$$d(S_A, S_B) = \min_{x_a \in S_A; x_b \in S_B} d(x_a, x_b)$$

Disimilitud mínima

El mínimo de los mínimos

- Genera clusters meno
homogéneos (más sensible
a outliers)



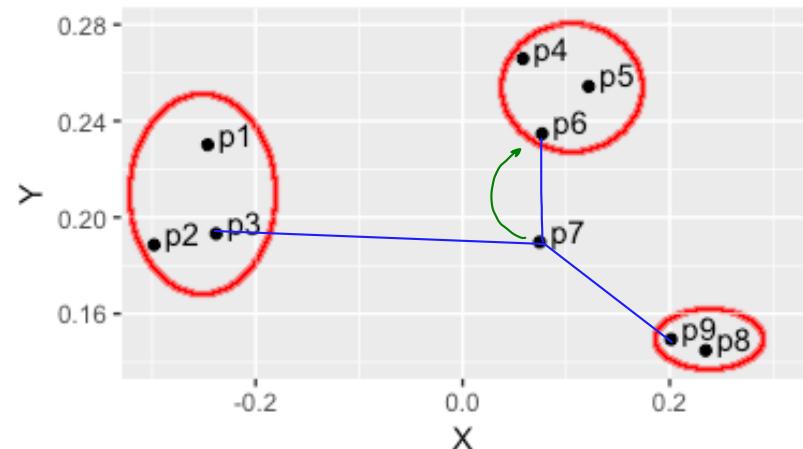
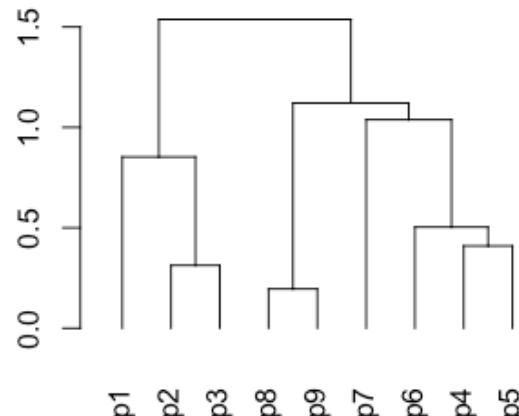
Agrupamiento Jerárquico

Aglomerativo

Criterios de unión

$$d(S_A, S_B) = \min_{x_a \in S_A; x_b \in S_B} d(x_a, x_b)$$

Disimilitud mínima



Agrupamiento Jerárquico

Aglomerativo

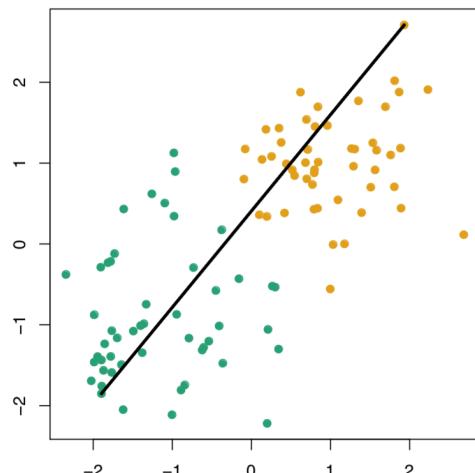
Criterios de unión

$$d(S_A, S_B) = \max_{x_a \in S_A; x_b \in S_B} d(x_a, x_b)$$

Disimilitud máxima

El mínimo de las máximas

- Genera clusters más homogéneos (menos sensible a outliers)



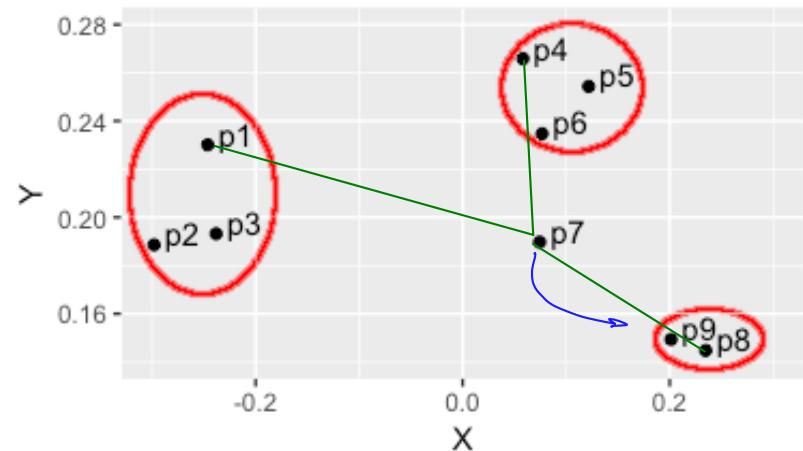
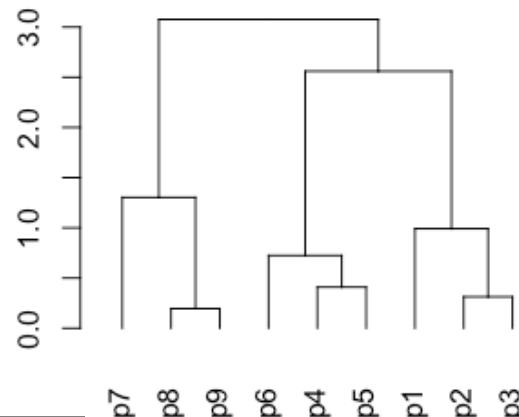
Agrupamiento Jerárquico

Aglomerativo

Criterios de unión

$$d(S_A, S_B) = \max_{x_a \in S_A; x_b \in S_B} d(x_a, x_b)$$

Disimilitud máxima



Agrupamiento Jerárquico

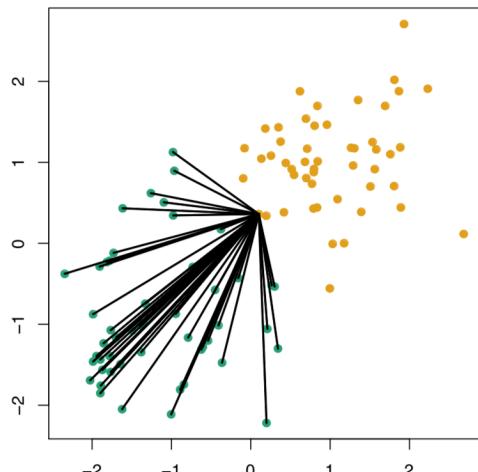
Aglomerativo

Criterios de unión

$$d(S_A, S_B) = \frac{1}{|S_A| \cdot |S_B|} \sum_{x_a \in S_A} \sum_{x_b \in S_B} d(x_a, x_b)$$

Disimilitud media

El mínimo de las medias.



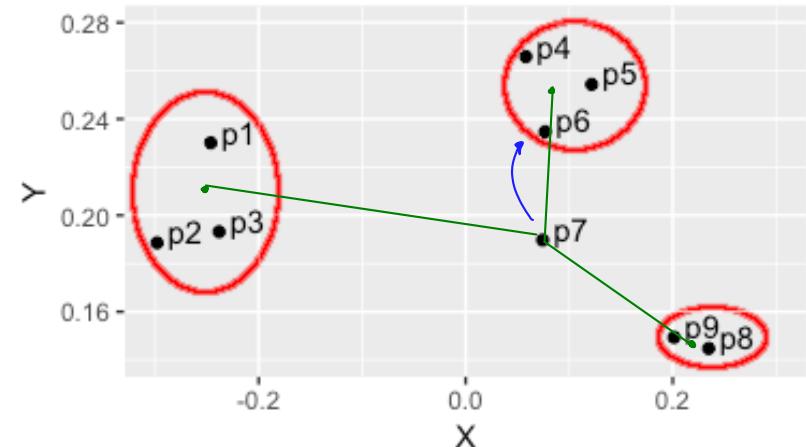
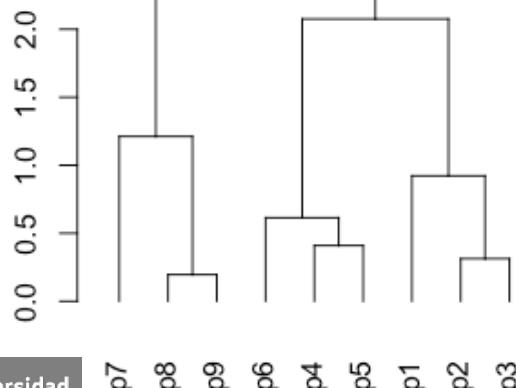
Agrupamiento Jerárquico

Aglomerativo

Criterios de unión

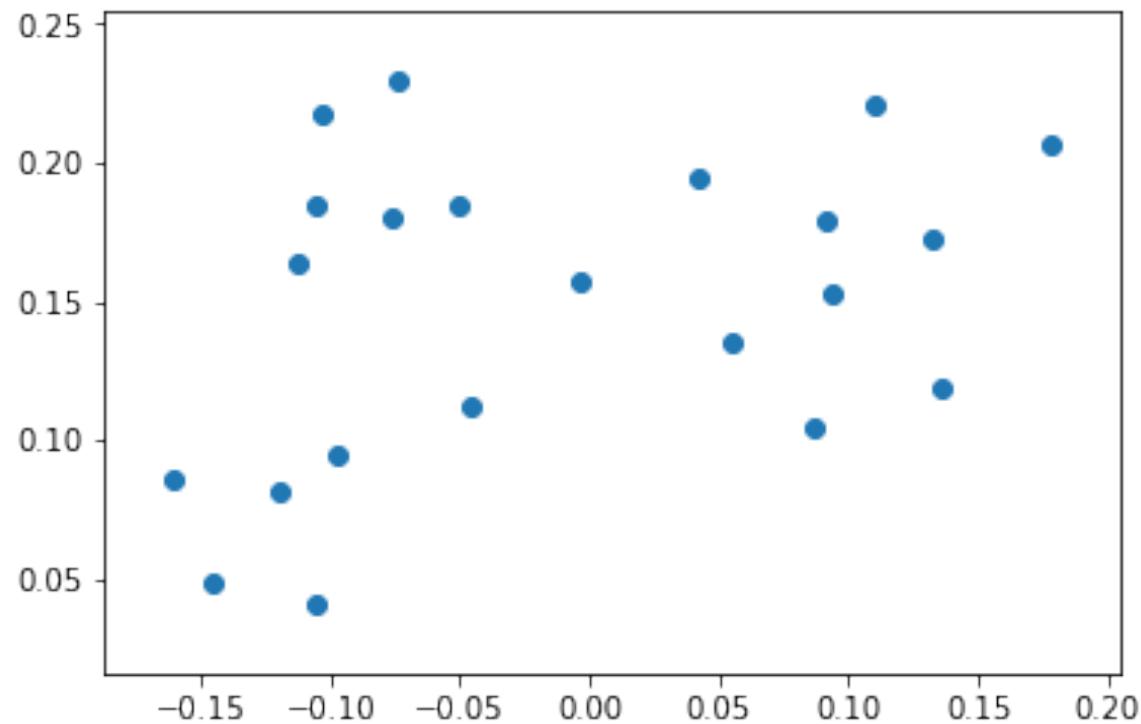
$$d(S_A, S_B) = \frac{1}{|S_A| \cdot |S_B|} \sum_{x_a \in S_A} \sum_{x_b \in S_B} d(x_a, x_b)$$

Disimilitud media



Agrupamiento Jerárquico

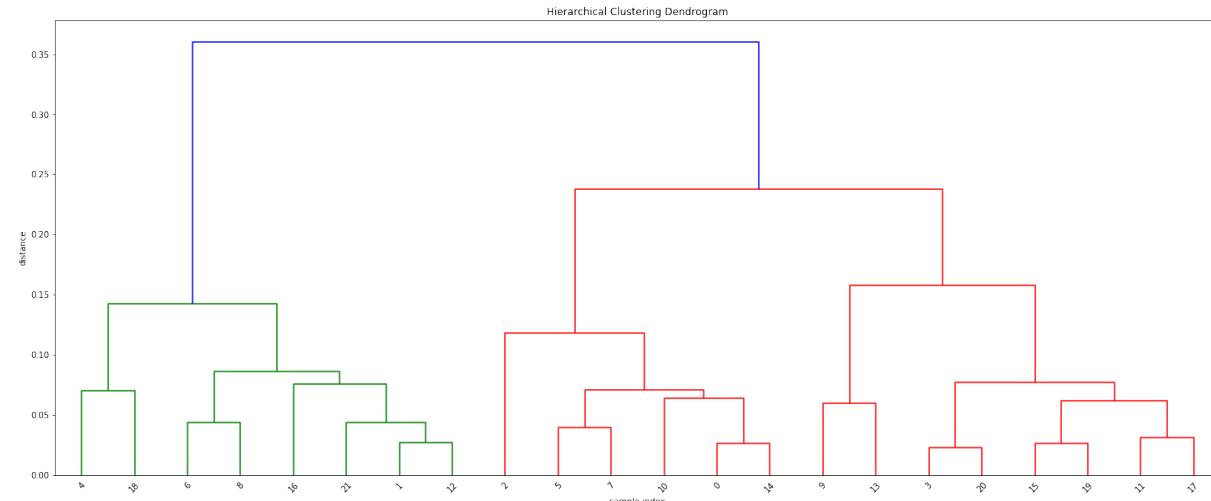
Aglomerativo



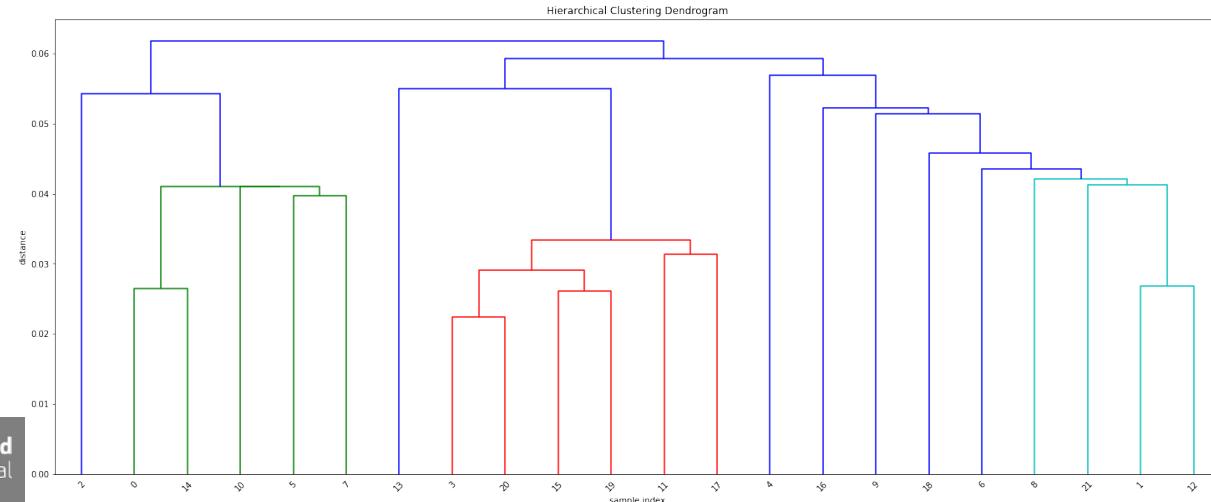
Agrupamiento Jerárquico

Aglomerativo

Máximo



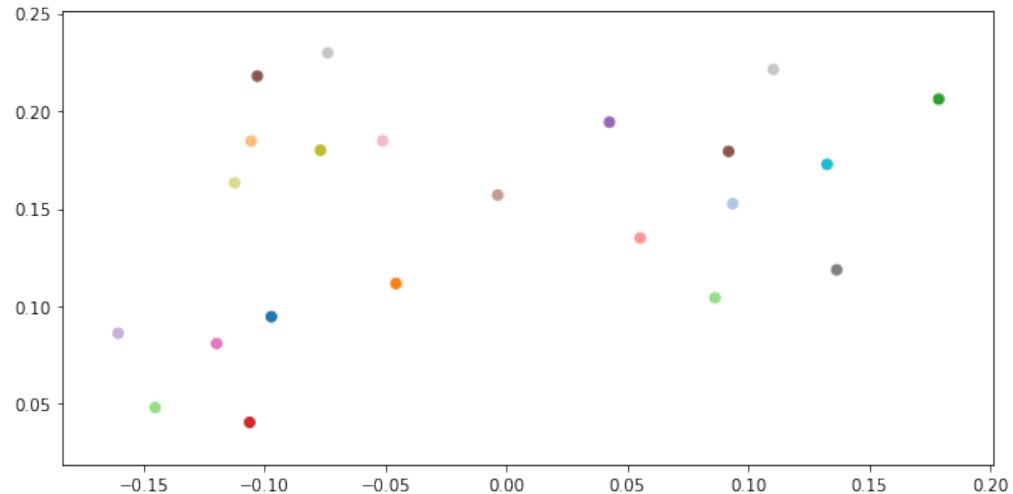
Mínimo



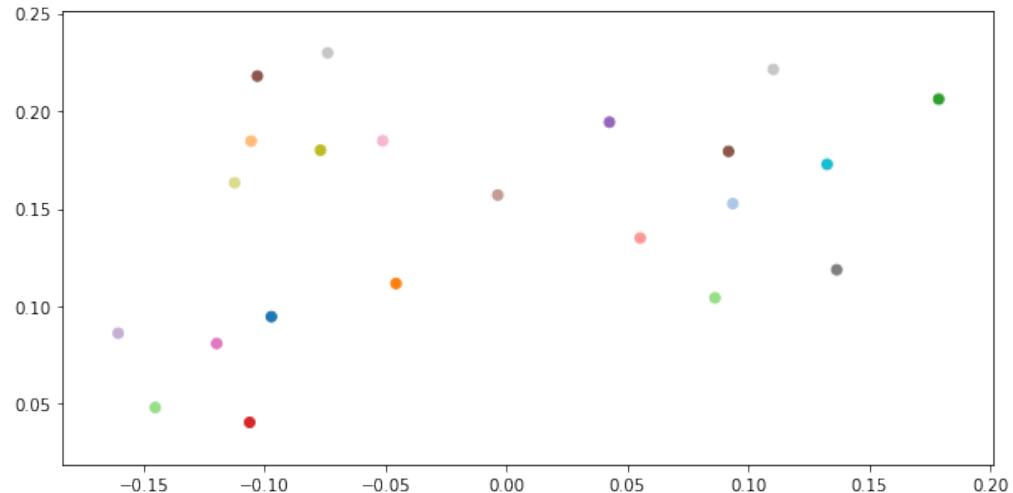
Agrupamiento Jerárquico

Aglomerativo

Máximo



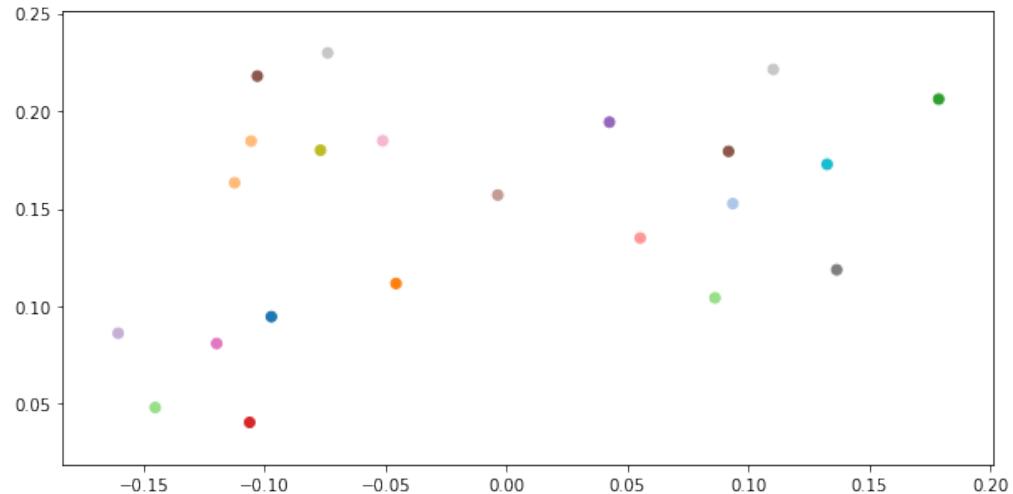
Mínimo



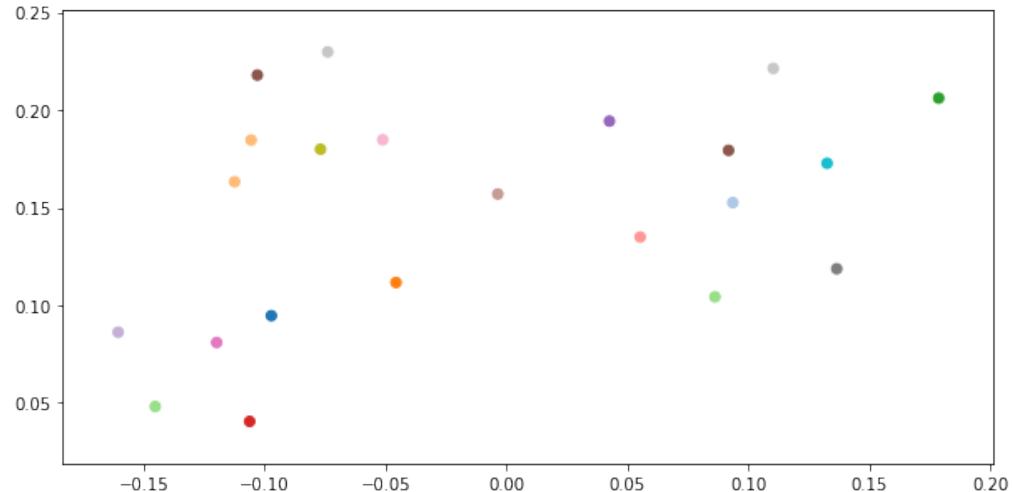
Agrupamiento Jerárquico

Aglomerativo

Máximo



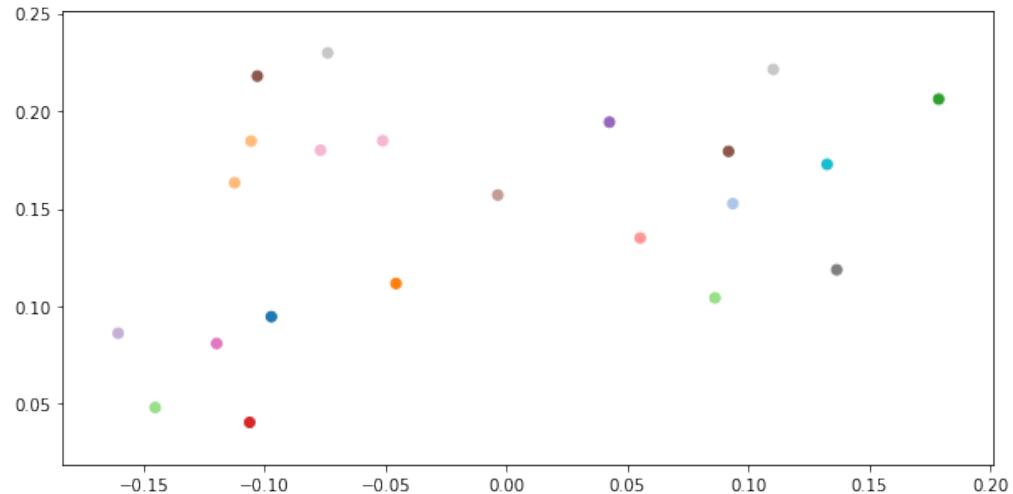
Mínimo



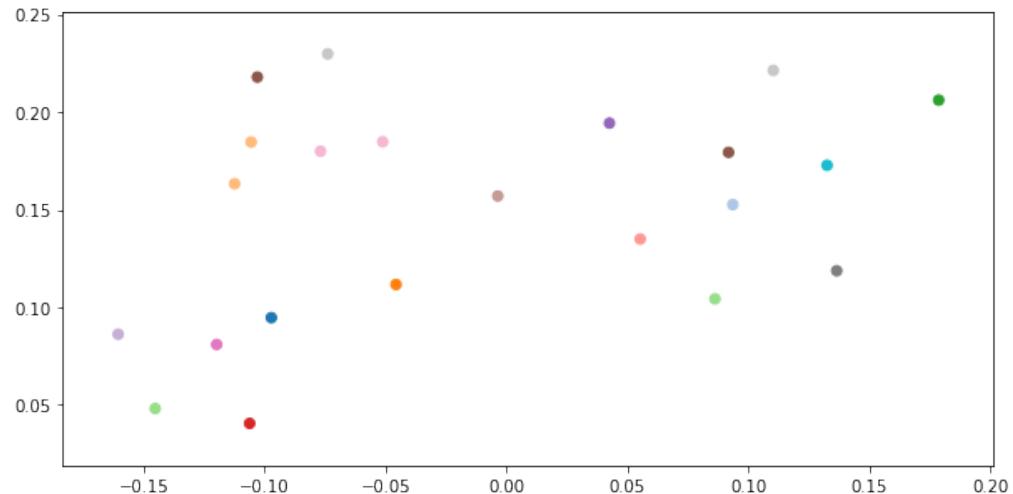
Agrupamiento Jerárquico

Aglomerativo

Máximo



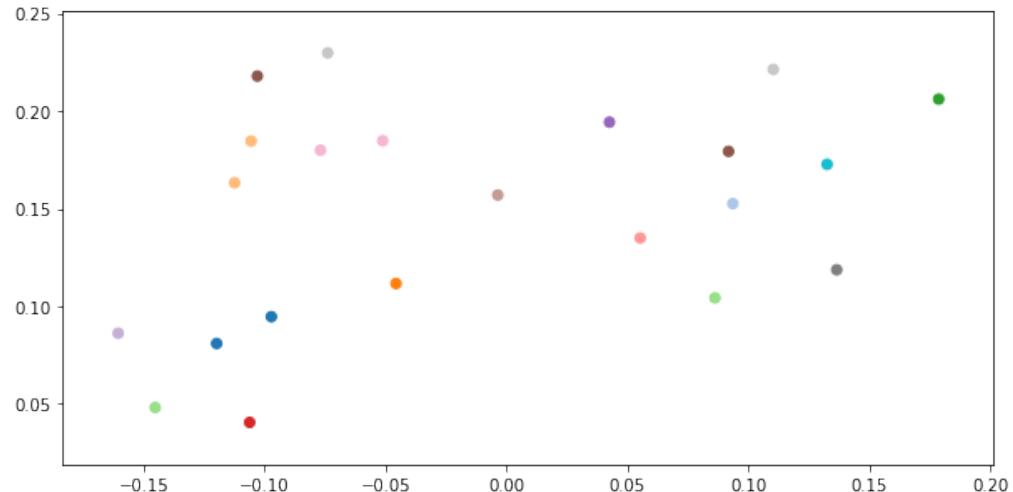
Mínimo



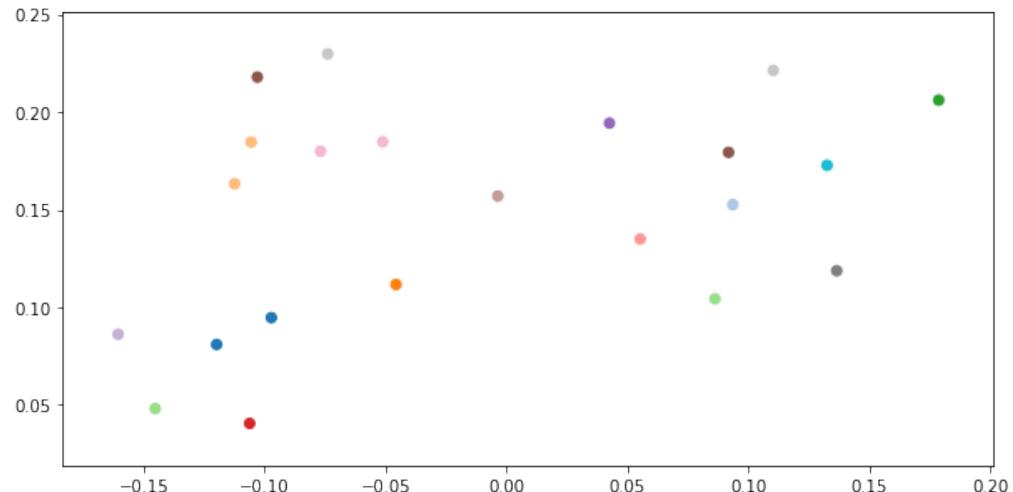
Agrupamiento Jerárquico

Aglomerativo

Máximo



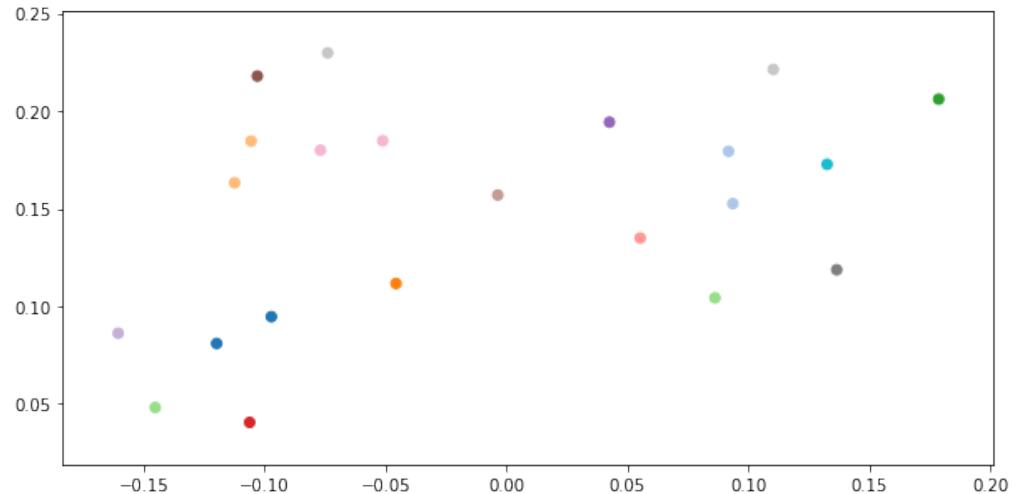
Mínimo



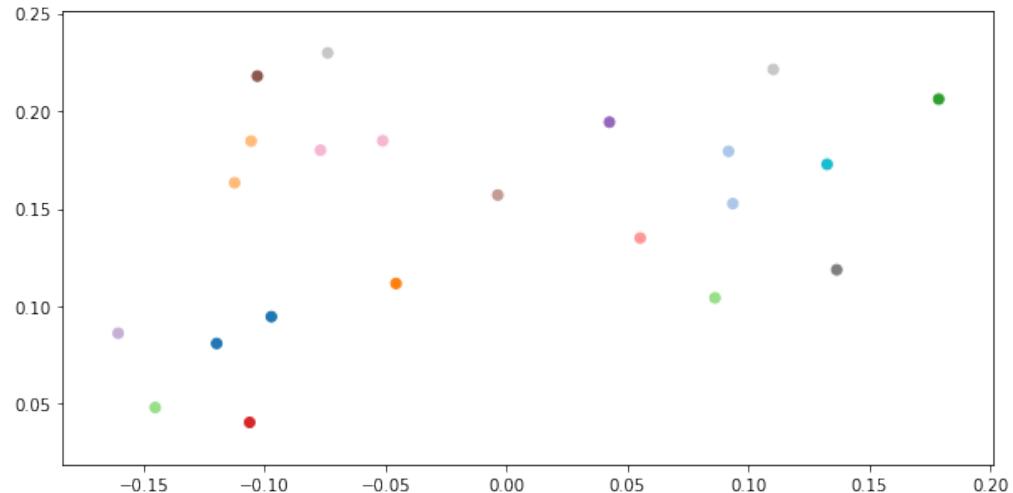
Agrupamiento Jerárquico

Aglomerativo

Máximo



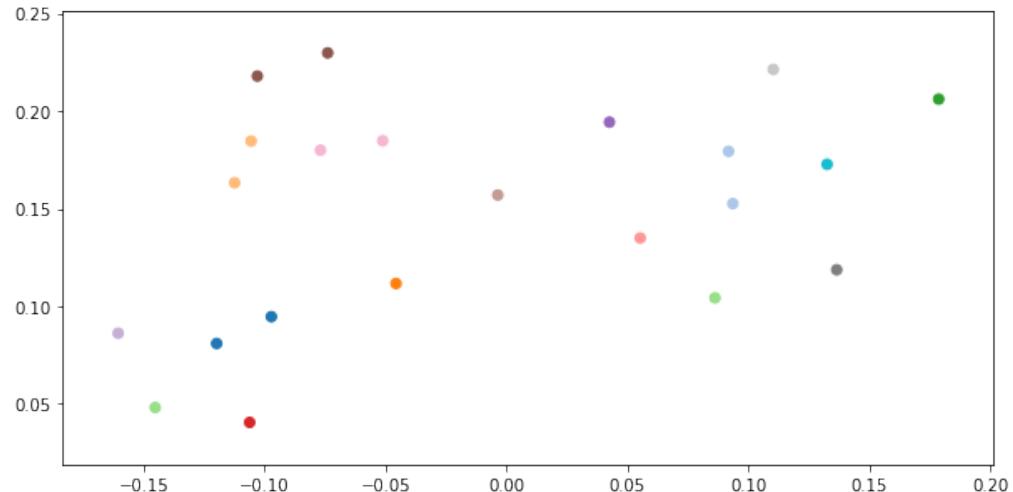
Mínimo



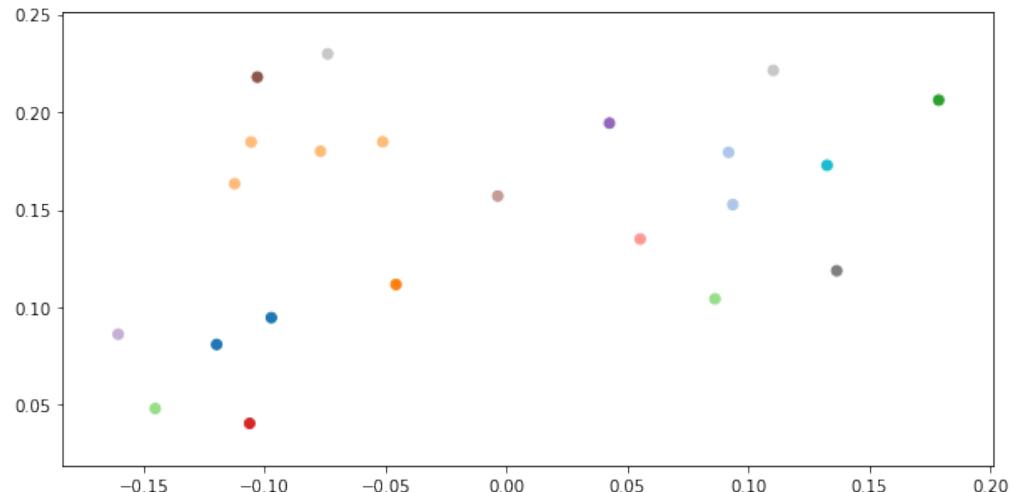
Agrupamiento Jerárquico

Aglomerativo

Máximo



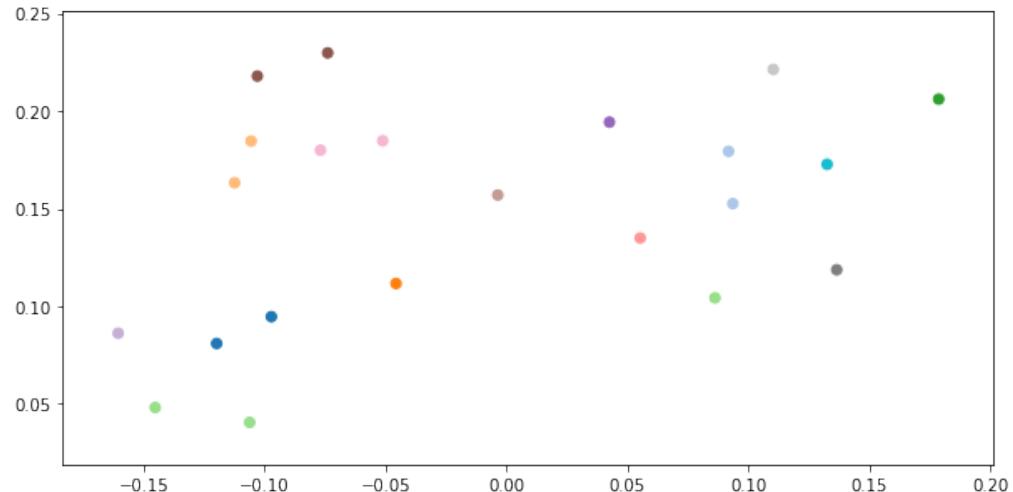
Mínimo



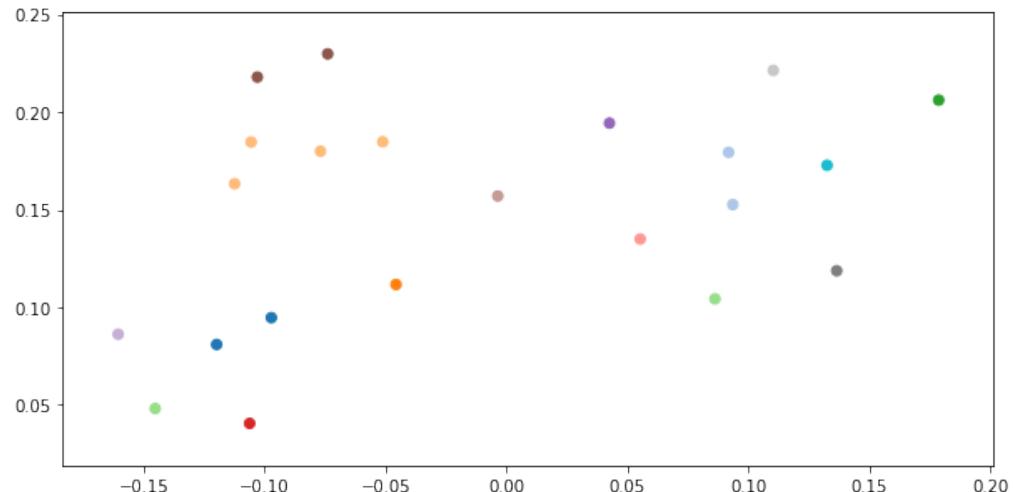
Agrupamiento Jerárquico

Aglomerativo

Máximo



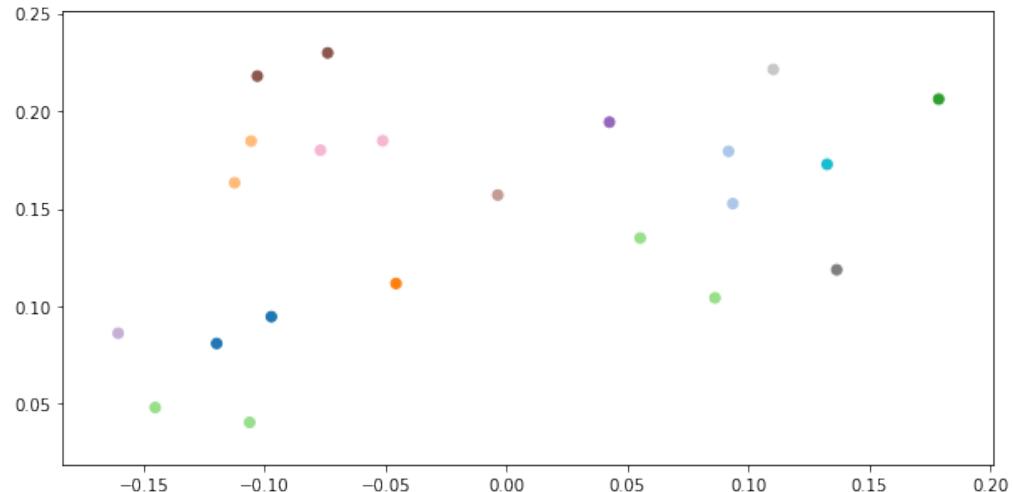
Mínimo



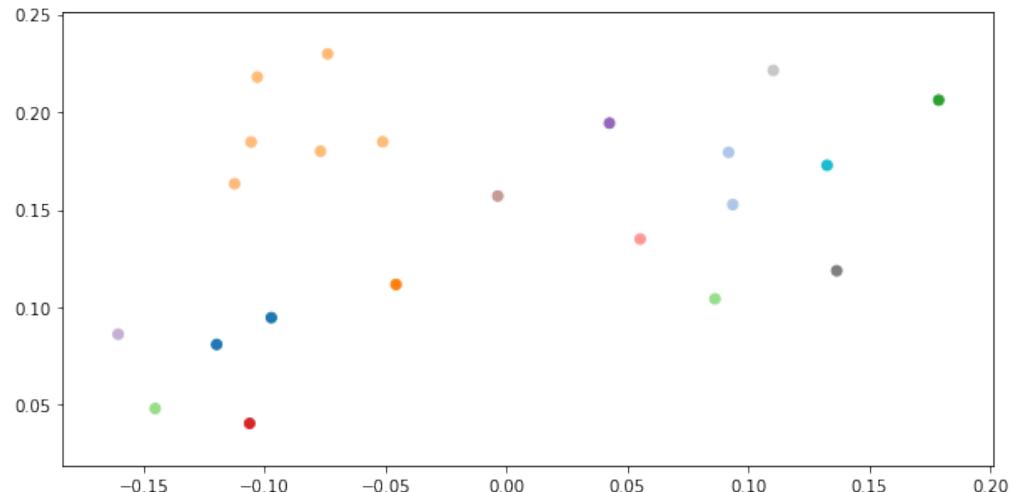
Agrupamiento Jerárquico

Aglomerativo

Máximo



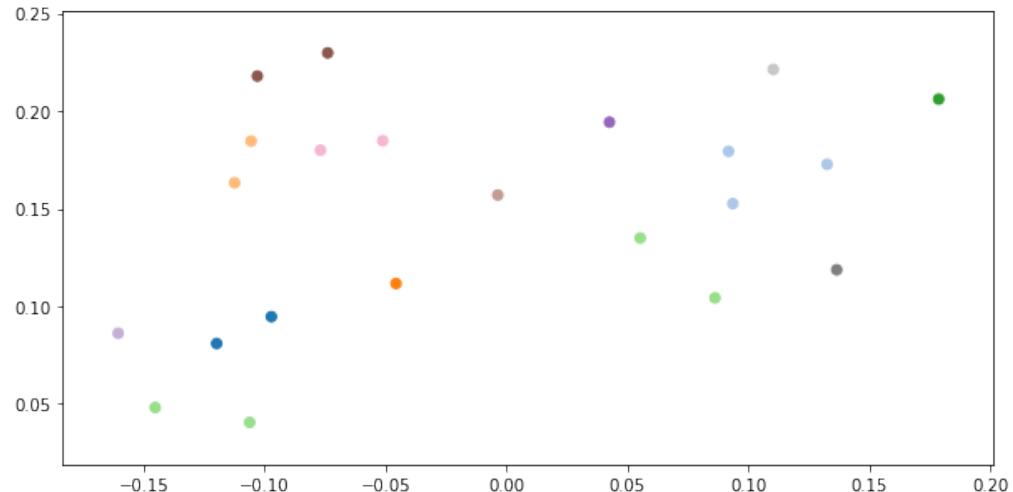
Mínimo



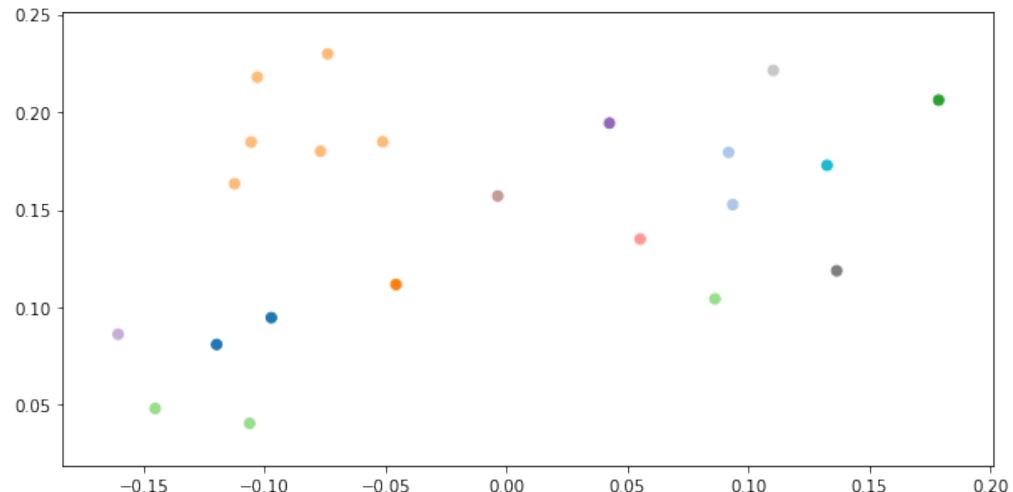
Agrupamiento Jerárquico

Aglomerativo

Máximo



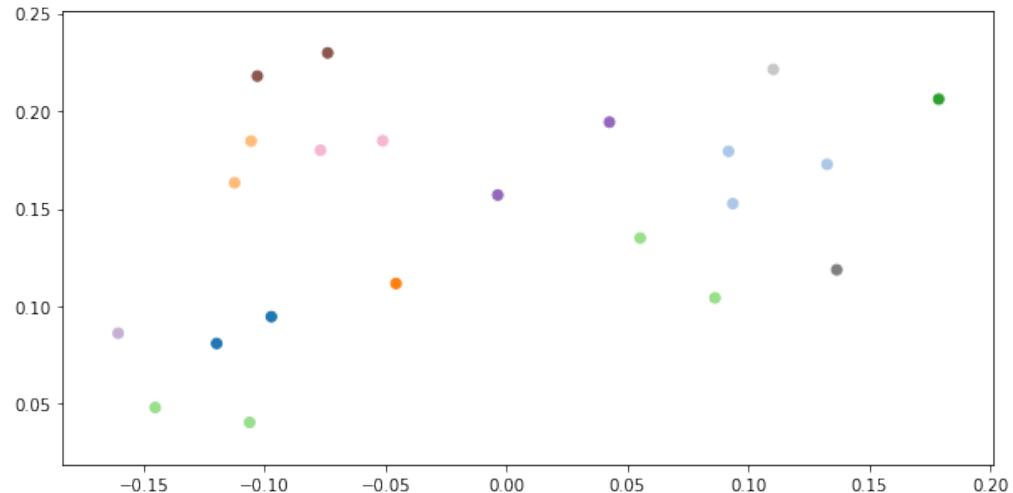
Mínimo



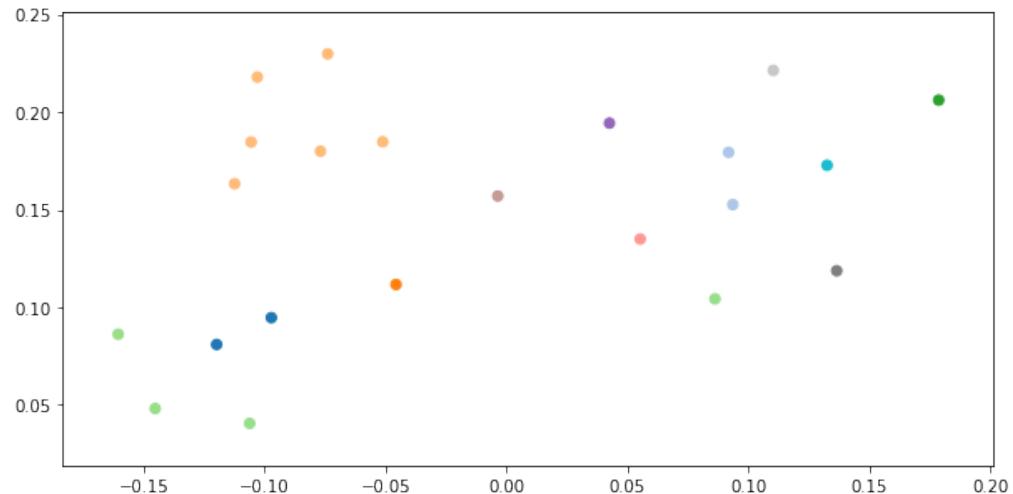
Agrupamiento Jerárquico

Aglomerativo

Máximo



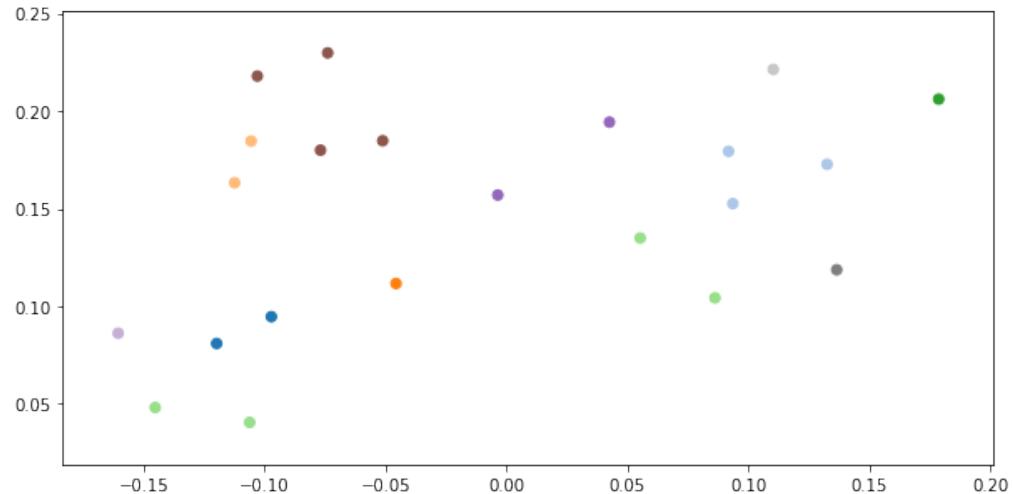
Mínimo



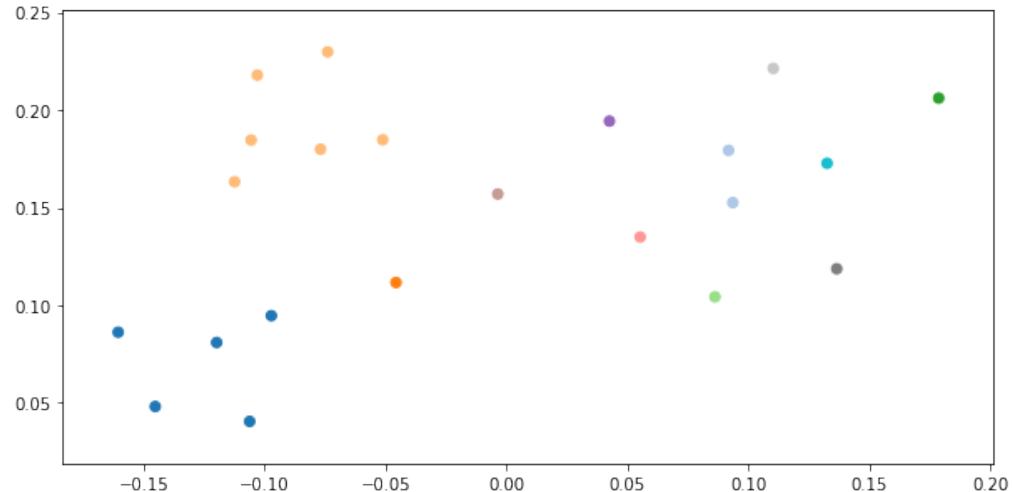
Agrupamiento Jerárquico

Aglomerativo

Máximo



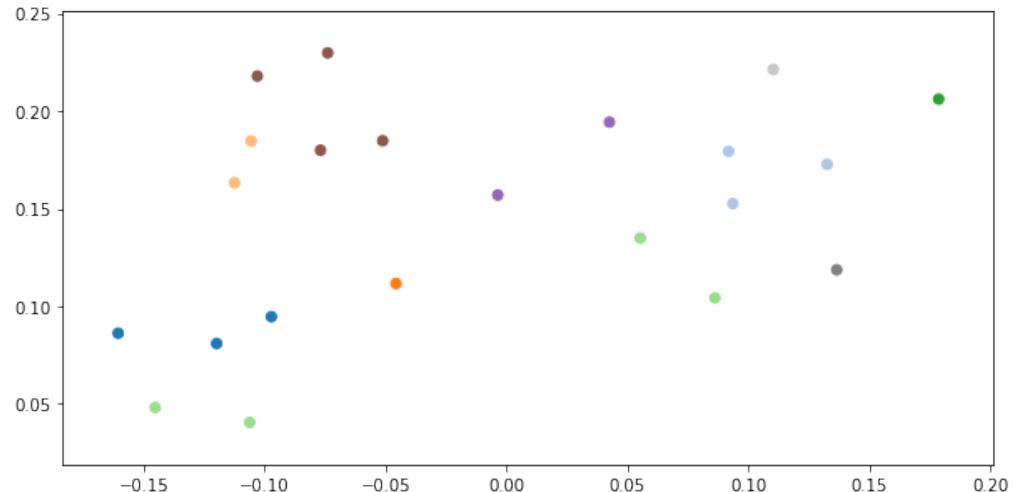
Mínimo



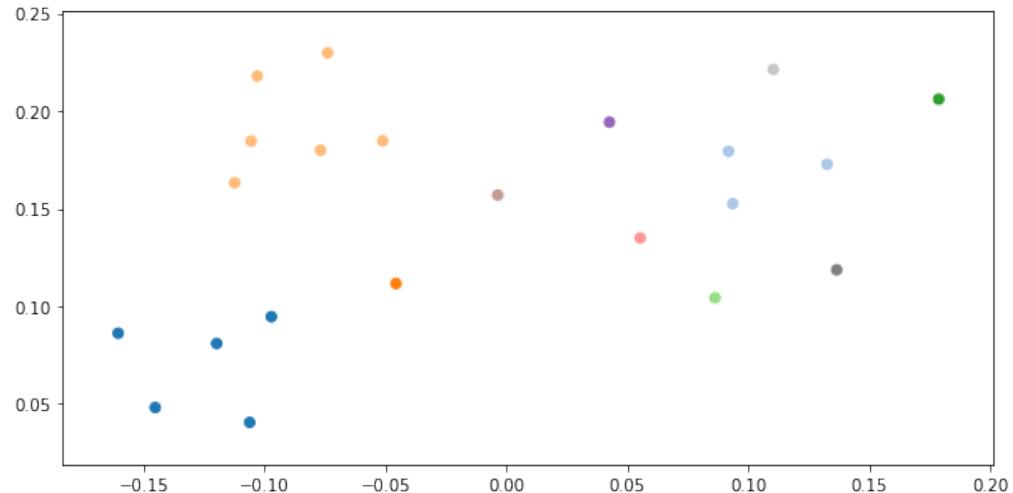
Agrupamiento Jerárquico

Aglomerativo

Máximo



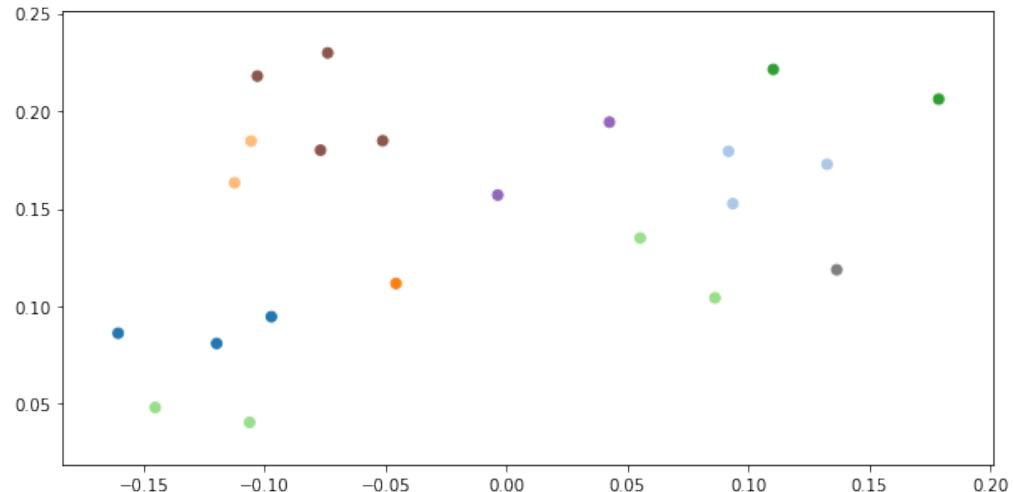
Mínimo



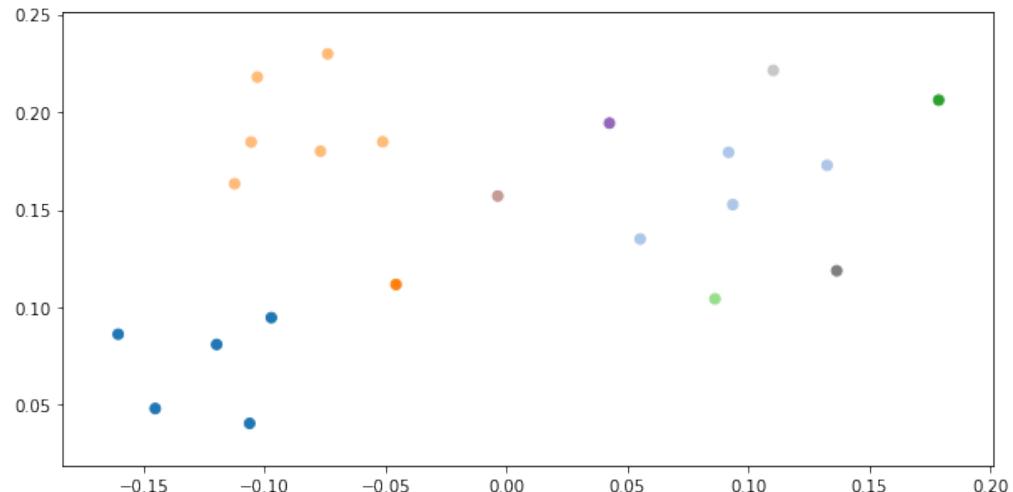
Agrupamiento Jerárquico

Aglomerativo

Máximo



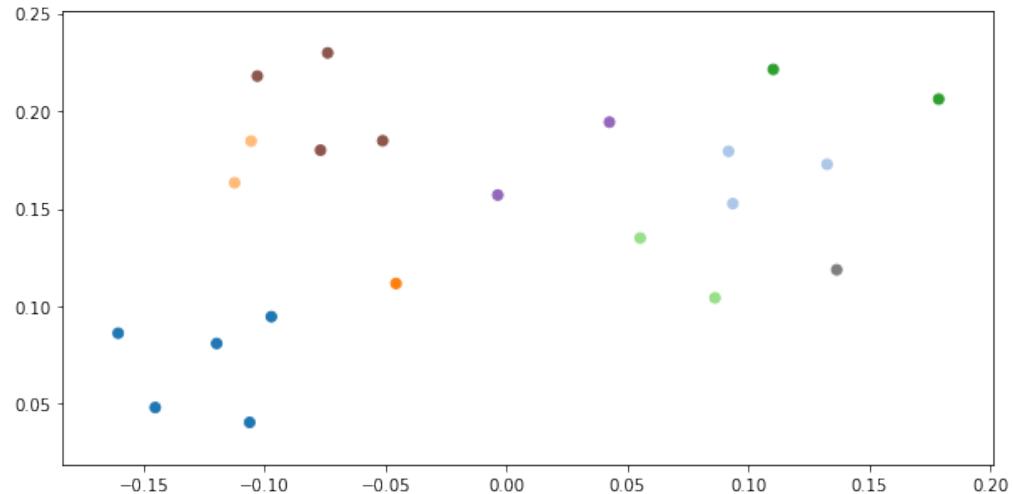
Mínimo



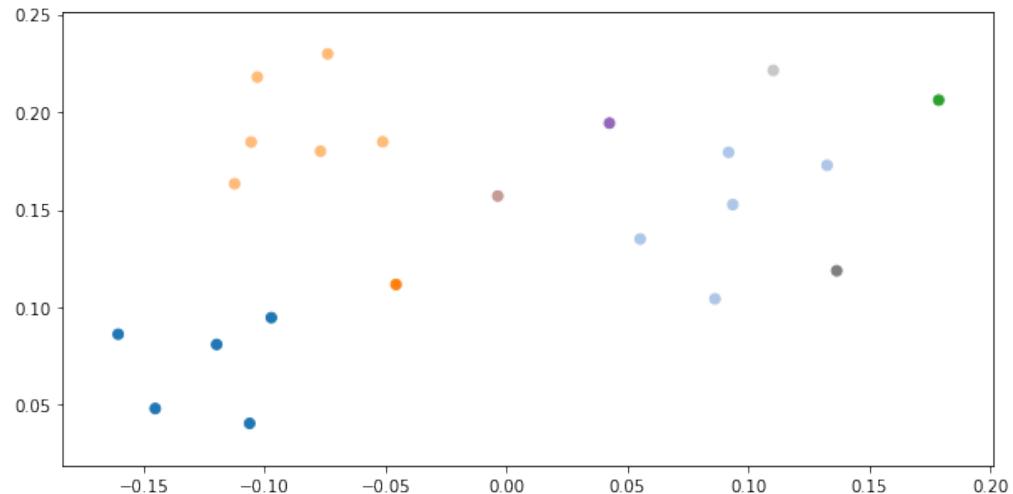
Agrupamiento Jerárquico

Aglomerativo

Máximo



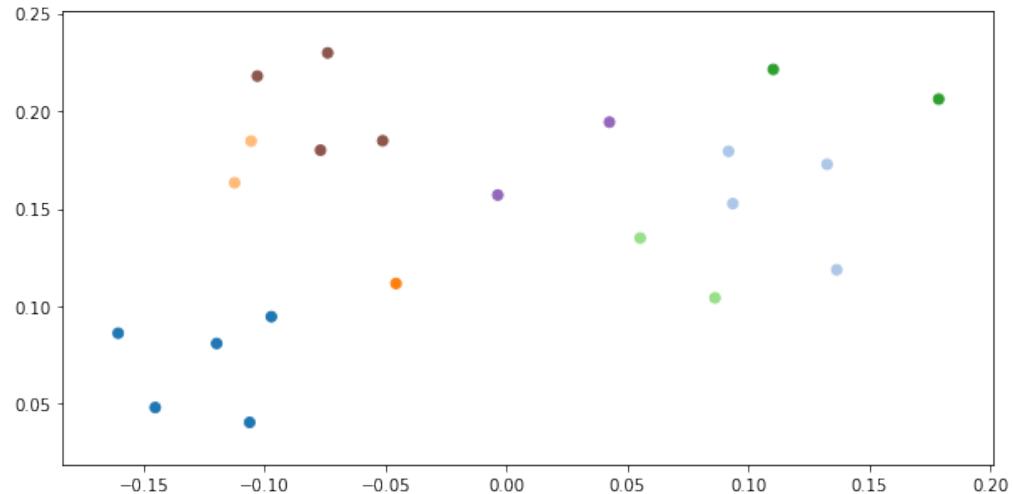
Mínimo



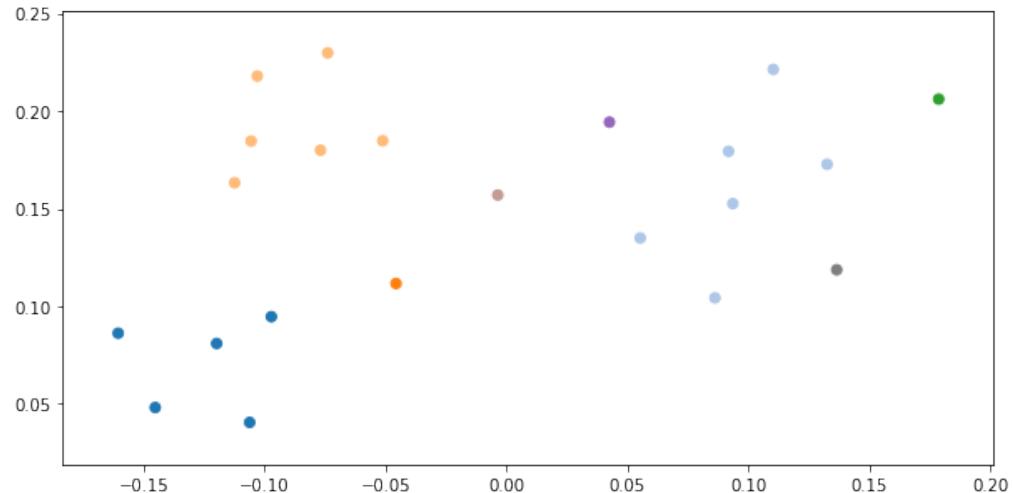
Agrupamiento Jerárquico

Aglomerativo

Máximo



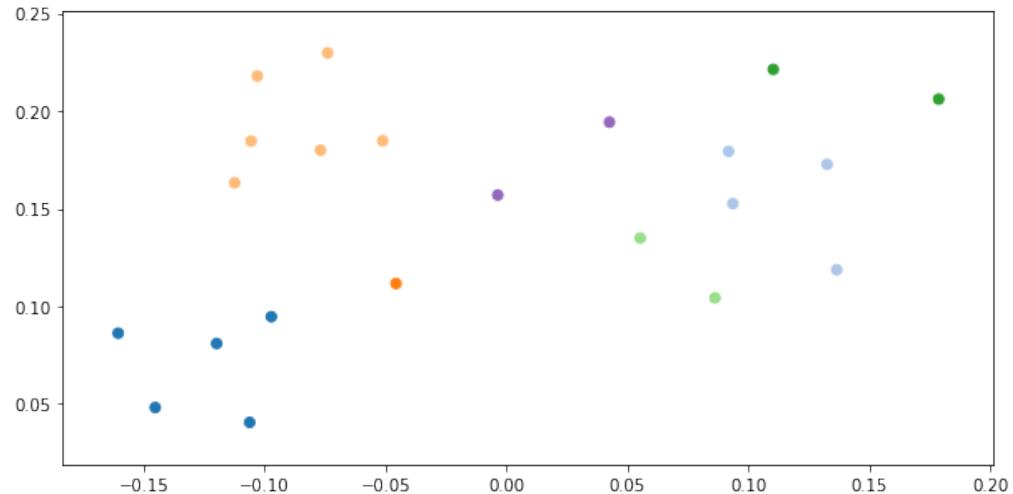
Mínimo



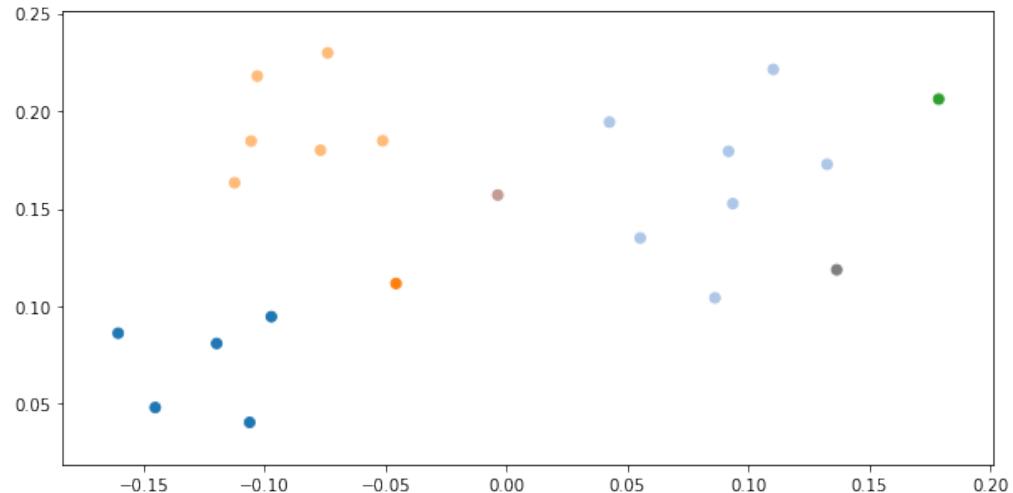
Agrupamiento Jerárquico

Aglomerativo

Máximo



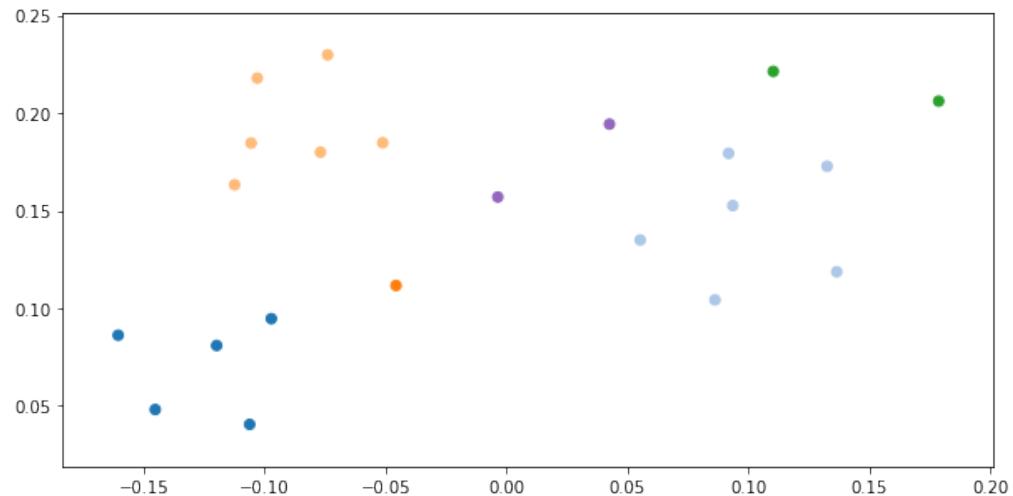
Mínimo



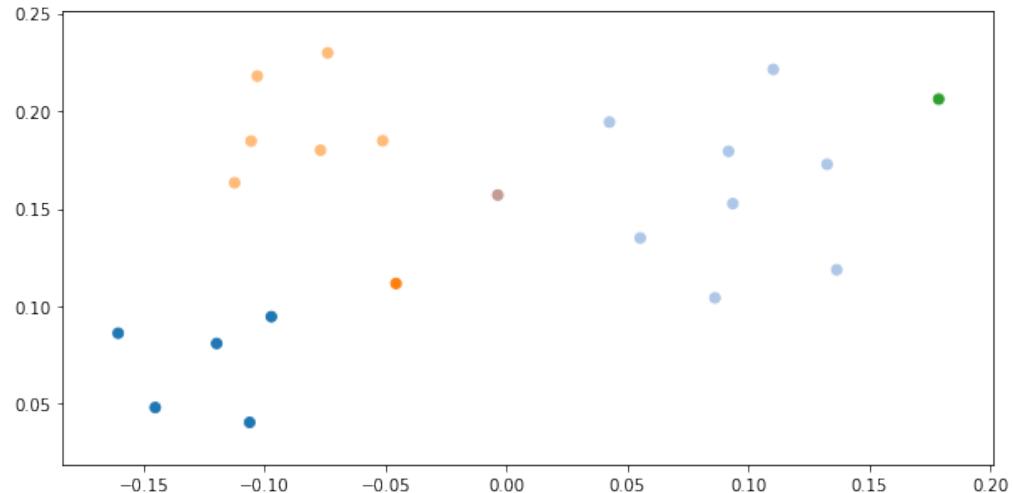
Agrupamiento Jerárquico

Aglomerativo

Máximo



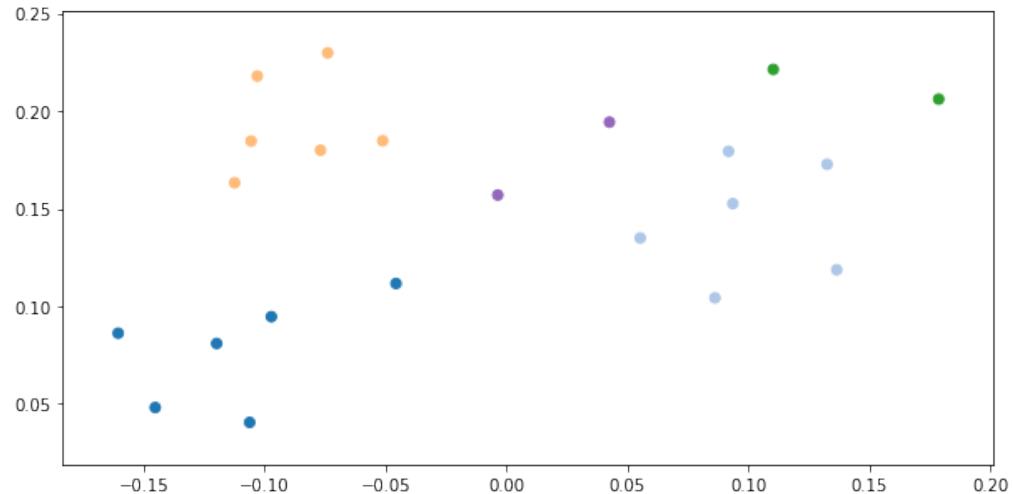
Mínimo



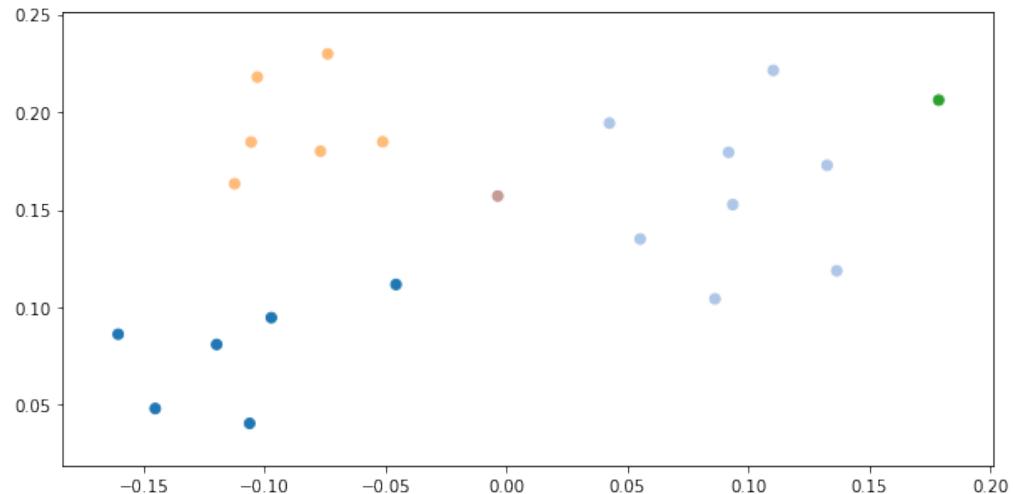
Agrupamiento Jerárquico

Aglomerativo

Máximo



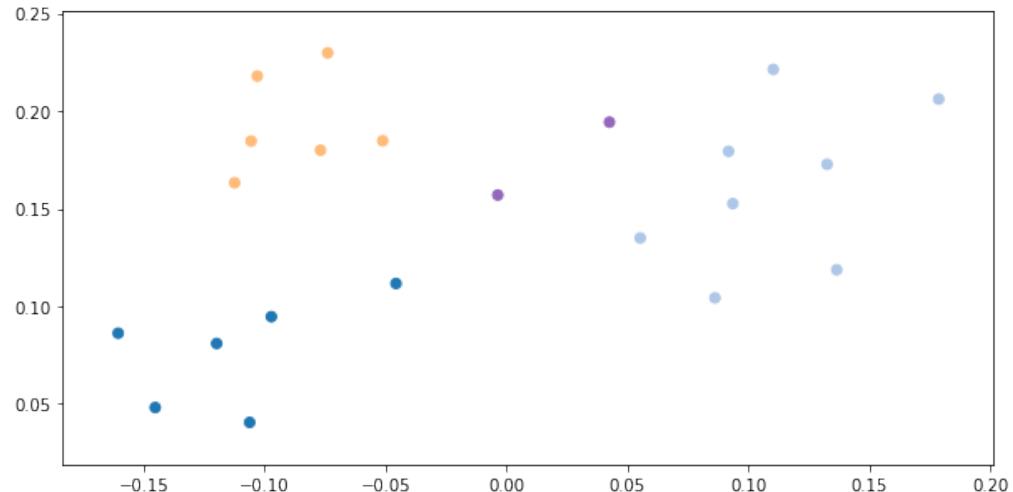
Mínimo



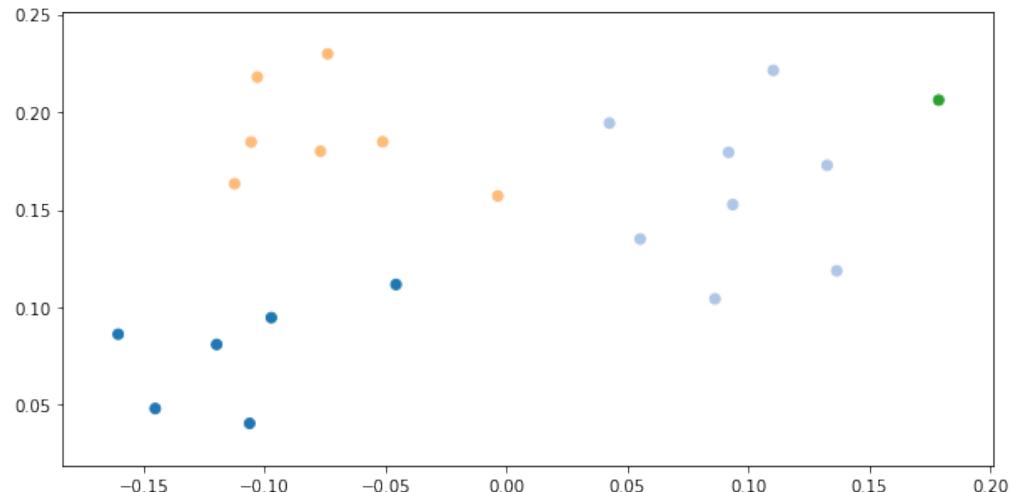
Agrupamiento Jerárquico

Aglomerativo

Máximo



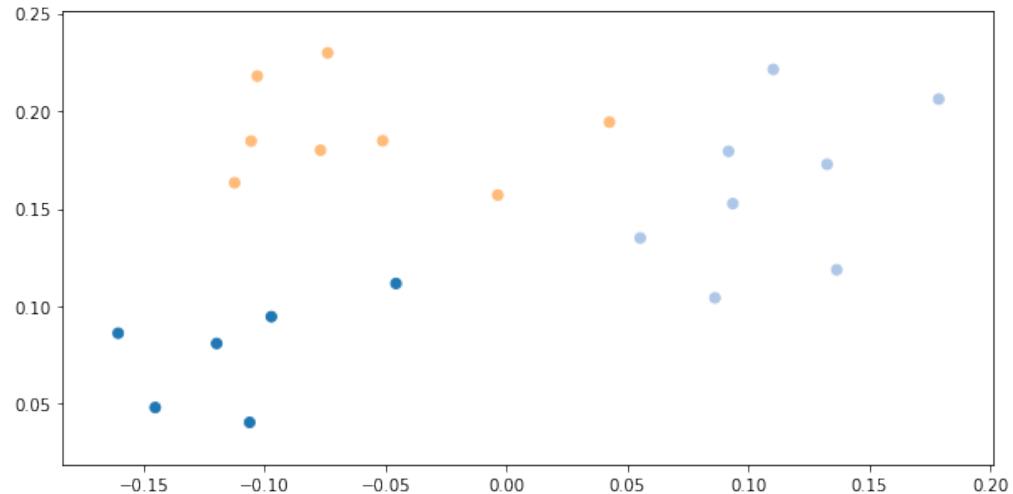
Mínimo



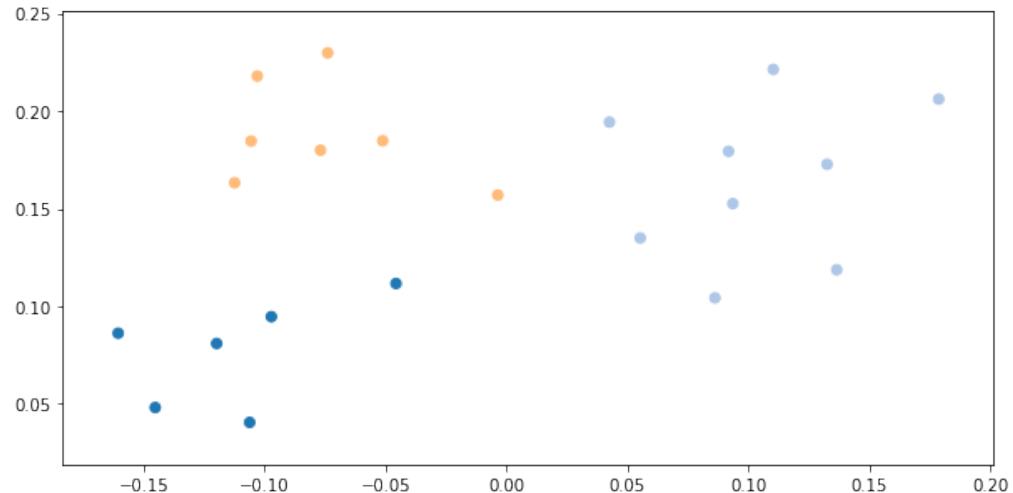
Agrupamiento Jerárquico

Aglomerativo

Máximo



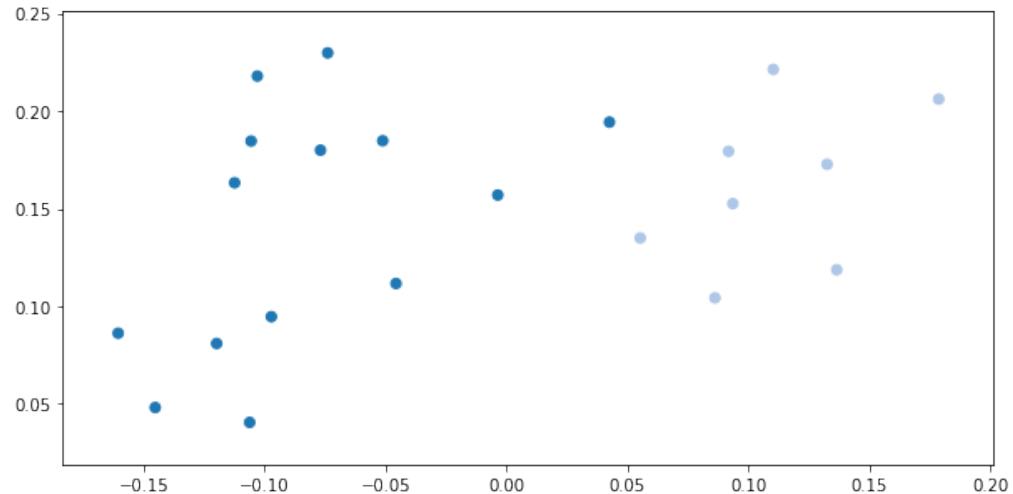
Mínimo



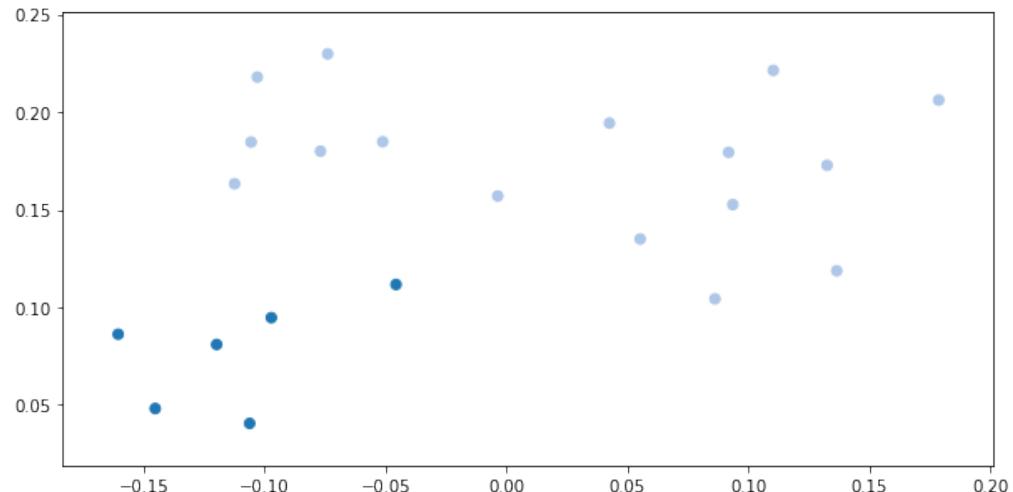
Agrupamiento Jerárquico

Aglomerativo

Máximo



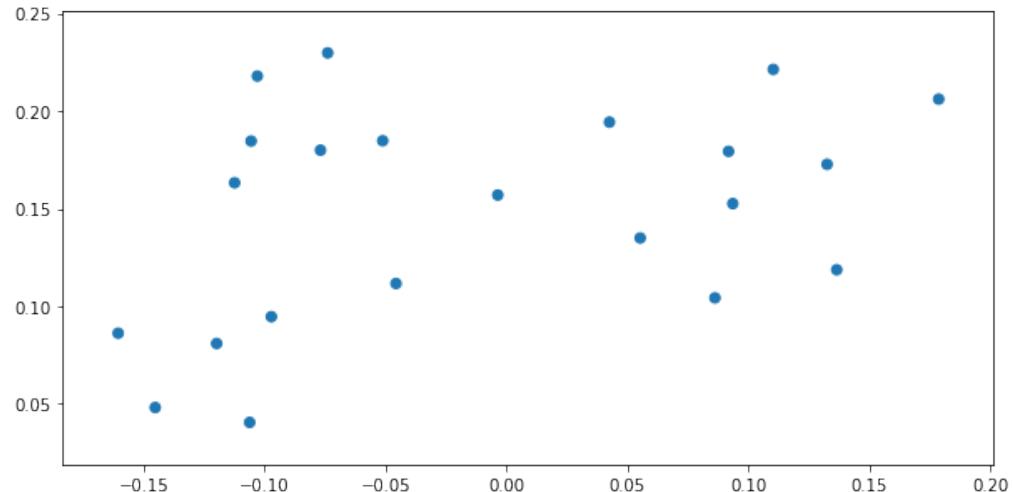
Mínimo



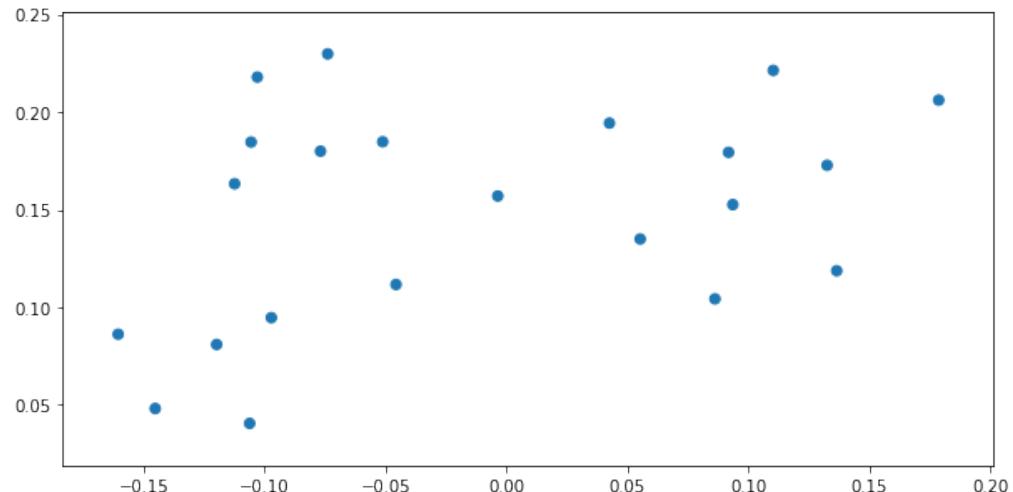
Agrupamiento Jerárquico

Aglomerativo

Máximo



Mínimo



Agrupamiento Jerárquico

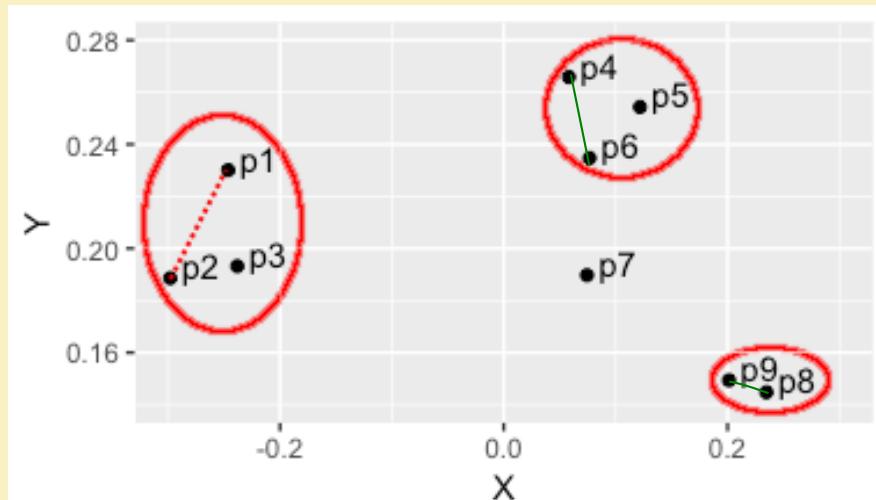
Aglomerativo

Tipos de clústeres obtenidos según criterio de unión

Definamos el concepto de **diámetro** de un clúster, S_K :

$$d(S_K) = \max_{x_i, x_j \in S_K} d(x_i, x_j)$$

Disimilitud máxima entre dos elementos del clúster S_K
(distancia)



Agrupamiento Jerárquico

Aglomerativo

Examen [21:30]

Diámetro mayor de clusters finales por criterio mínimo.

Tipos de clústeres obtenidos según criterio de unión

Disimilitud mínima:

$$d(S_A, S_B) = \min_{x_a \in S_A; x_b \in S_B} d(x_a, x_b)$$

Disimilitud máxima:

$$d(S_A, S_B) = \max_{x_a \in S_A; x_b \in S_B} d(x_a, x_b)$$

Disimilitud media:

$$d(S_A, S_B) = \frac{1}{|S_A| + |S_B|} \sum_{x_a \in S_A} \sum_{x_b \in S_B} d(x_a, x_b)$$

Agrupamiento Jerárquico

Aglomerativo

Tipos de clústeres obtenidos según criterio de unión

Disimilitud mínima:

$$d(S_A, S_B) = \min_{x_a \in S_A; x_b \in S_B} d(x_a, x_b)$$

- ▶ Clústeres de ejemplos similares que pueden no formar una unidad compacta
Idea de la cadena
- ▶ El diámetro puede salir perjudicado

Disimilitud máxima:

$$d(S_A, S_B) = \max_{x_a \in S_A; x_b \in S_B} d(x_a, x_b)$$

Disimilitud media:

$$d(S_A, S_B) = \frac{1}{|S_A| + |S_B|} \sum_{x_a \in S_A} \sum_{x_b \in S_B} d(x_a, x_b)$$

Agrupamiento Jerárquico

Aglomerativo

Tipos de clústeres obtenidos según criterio de unión

Disimilitud mínima:

$$d(S_A, S_B) = \min_{x_a \in S_A; x_b \in S_B} d(x_a, x_b)$$

Disimilitud máxima:

$$d(S_A, S_B) = \max_{x_a \in S_A; x_b \in S_B} d(x_a, x_b)$$

- ▶ Clústeres compactos con diámetro reducido
- ▶ Se minimiza el diámetro, precisamente
La disimilitud máxima intraclúster es, tras la unión, el diámetro del nuevo clúster
- ▶ Puede separar en clústeres diferentes a ejemplos *muy* similares

Disimilitud media:

$$d(S_A, S_B) = \frac{1}{|S_A| + |S_B|} \sum_{x_a \in S_A} \sum_{x_b \in S_B} d(x_a, x_b)$$

Agrupamiento Jerárquico

Aglomerativo

Tipos de clústeres obtenidos según criterio de unión

Disimilitud mínima:

$$d(S_A, S_B) = \min_{x_a \in S_A; x_b \in S_B} d(x_a, x_b)$$

Disimilitud máxima:

$$d(S_A, S_B) = \max_{x_a \in S_A; x_b \in S_B} d(x_a, x_b)$$

Disimilitud media:

$$d(S_A, S_B) = \frac{1}{|S_A| + |S_B|} \sum_{x_a \in S_A} \sum_{x_b \in S_B} d(x_a, x_b)$$

- ▶ Escenario intermedio
- ▶ Clústeres relativamente compactos
- ▶ Junta elementos no necesariamente muy similares

EXAMEN

Ventajas

- ▶ Intuitivo
- ▶ Conceptualmente sencillo
- ▶ Funciona con clústeres de diferente tamaño
- ▶ Una decisión de entrenamiento: criterio de unión
- ▶ Diferentes criterios
- ▶ Puede funcionar con diferentes medidas de distancia

EXAMEN

Desventajas

- ▶ Lento
- ▶ Problemas al lidiar con clústeres de diferente densidad
- ▶ ¿Qué partición elegir?

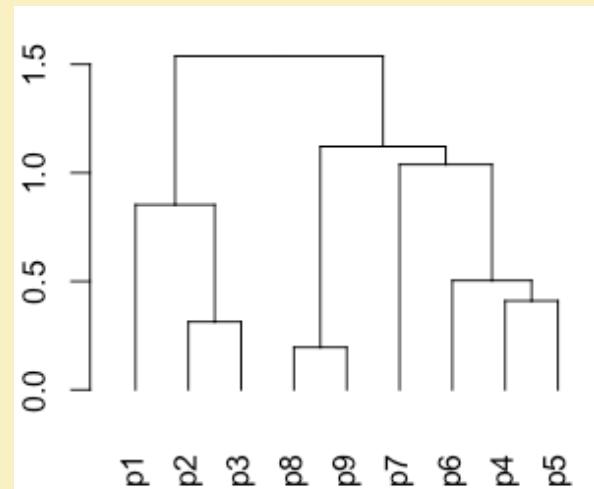
Agrupamiento Jerárquico

Aglomerativo

Elección de una partición

Elegir una altura en la jerarquía donde cortar

- ▶ Número de clústeres concreto (fijando K)
- ▶ Máxima distancia en la unión de clústeres



Aprendizaje no supervisado

VC03: Agrupamiento jerárquico: Divisivo

Rocío del Amor del Amor

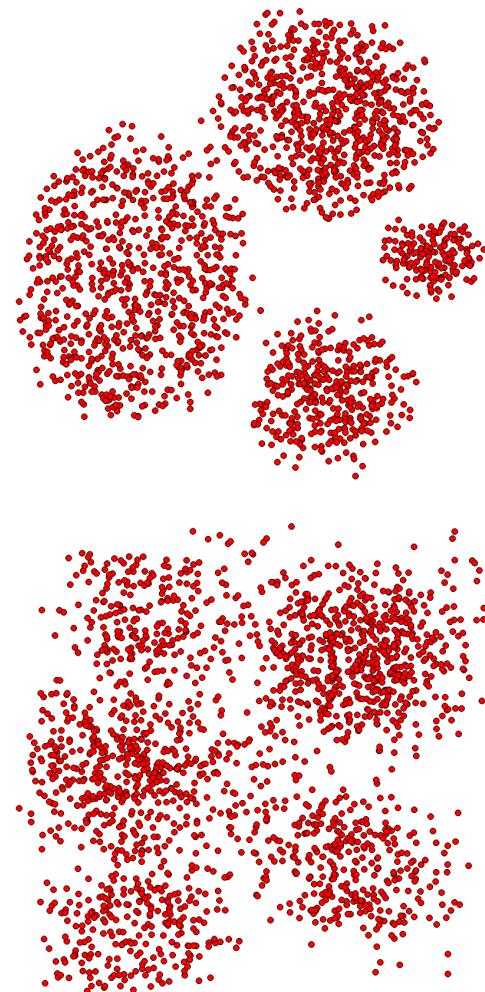
Mrocio.delamor@campusviu.es

Universidad Internacional de Valencia

Agrupamiento

Tipos de algoritmos de agrupamiento

- ▶ Basados en particiones
- ▶ **Jerárquicos**
- ▶ Espectrales
- ▶ Basados en densidad
- ▶ Probabilísticos



Agrupamiento Jerárquico

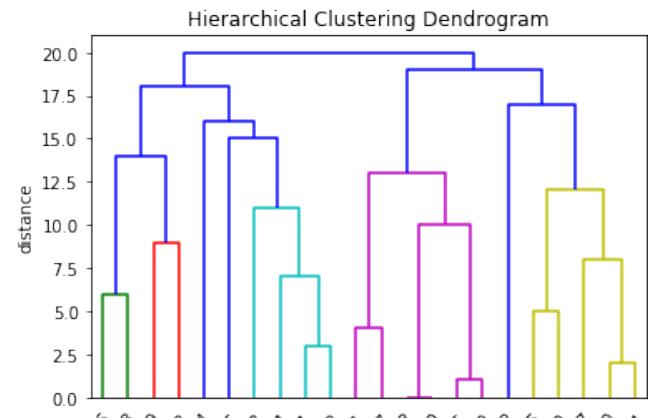
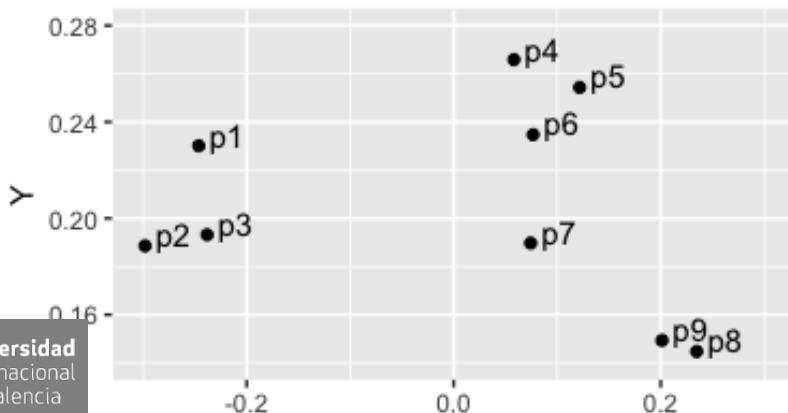
Un continuo de particiones de los datos

Se partitiona el dataset desde $K = 1$ hasta $K = n$

** ¿Cuál es la mejor partición?

Algoritmos:

- ▶ Aglomerativo
- ▶ **Divisivo**

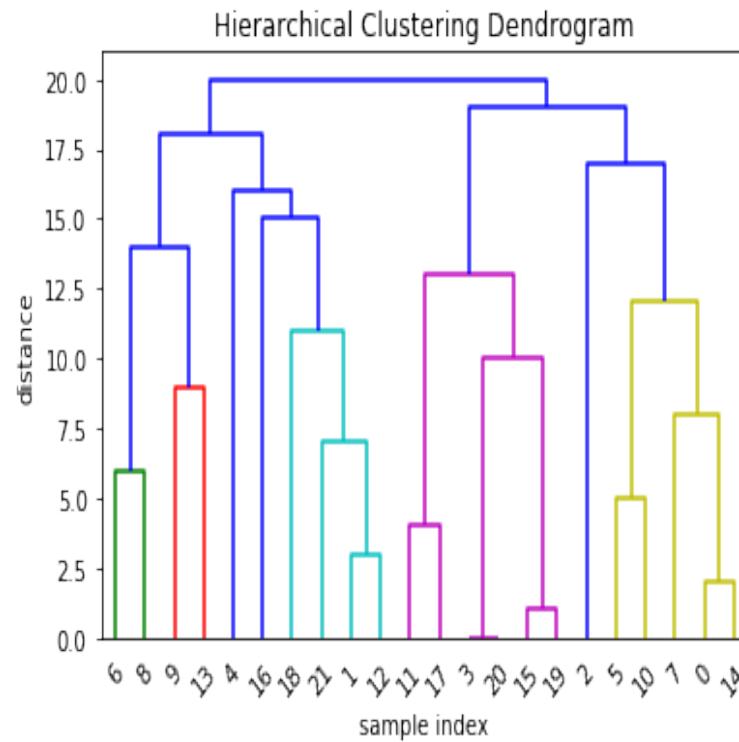


Agrupamiento Jerárquico

Dendrograma

Representación gráfica de un agrupamiento jerárquico

- ▶ Cada nodo, es un conjunto de ejemplos (clúster)
- ▶ Los clústeres se van uniendo/separando según criterios de distancia
- ▶ La longitud de las líneas verticales indica la distancia entre los clústeres que se unen/separan



Agrupamiento Jerárquico

Divisivo

División

Partiendo de $K = 1$, se va dividiendo (en dos) iterativamente un clúster hasta $K = n$, de manera voraz

0. Al principio, sólo existe un clúster ($K = 1$) con todos los ejemplos
1. Tras la primera división, existen $K = 2$ clústeres
(se reparten entre los dos todos los ejemplos)
- ...
- i. Tras la i -ésima división, existen $K = i + 1$ clústeres
- ...
- $n-1$. El algoritmo acaba cuando $K = n$
(se divide el único clúster que no es unitario y cada ejemplo tiene su propio clúster)

Agrupamiento Jerárquico

Divisivo

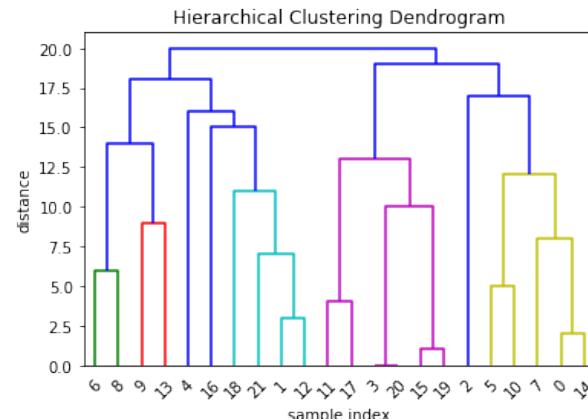
Tres cuestiones

A medida que avanza el algoritmo...

¿qué clúster se debe dividir en cada paso?

Al final del algoritmo, si queremos una partición concreta,

¿con qué partición nos quedamos?



Agrupamiento Jerárquico

Divisivo

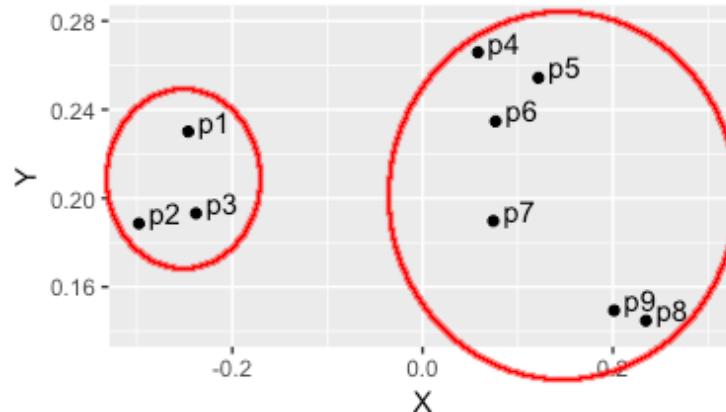
Tres cuestiones

A medida que avanza el algoritmo...

¿qué clúster se debe dividir en cada paso?

Al final del algoritmo, si queremos una partición concreta,

¿con qué partición nos quedamos?



Agrupamiento Jerárquico

Divisivo

Tres cuestiones

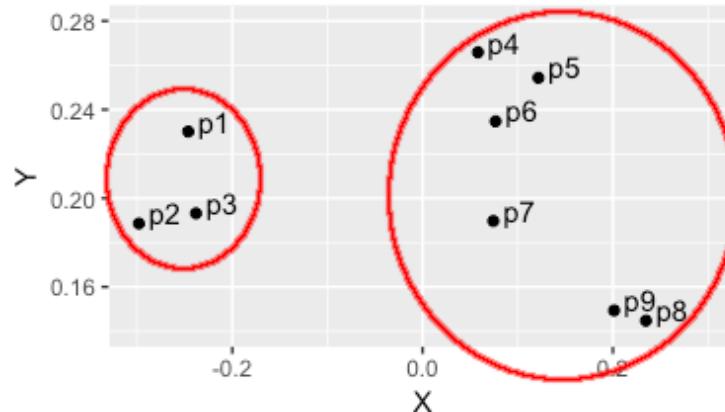
A medida que avanza el algoritmo...

¿qué clúster se debe dividir en cada paso?

¿cómo se divide el clúster seleccionado?

Al final del algoritmo, si queremos una partición concreta,

¿con qué partición nos quedamos?



Agrupamiento Jerárquico

Divisivo

Primera cuestión

A medida que avanza el algoritmo...

¿qué clúster se debe dividir en cada paso?

El clúster, S_K^* , con mayor disimilitud intraclúster:

$$S_K^* = \operatorname{arg\max}_{S_K} d(S_K)$$

Agrupamiento Jerárquico

Divisivo

Primera cuestión

A medida que avanza el algoritmo...

¿qué clúster se debe dividir en cada paso?

El clúster, S_K^* , con mayor disimilitud intraclúster:

$$S_K^* = \operatorname{arg\max}_{S_K} d(S_K)$$

¿cómo se mide la disimilitud intraclúster?

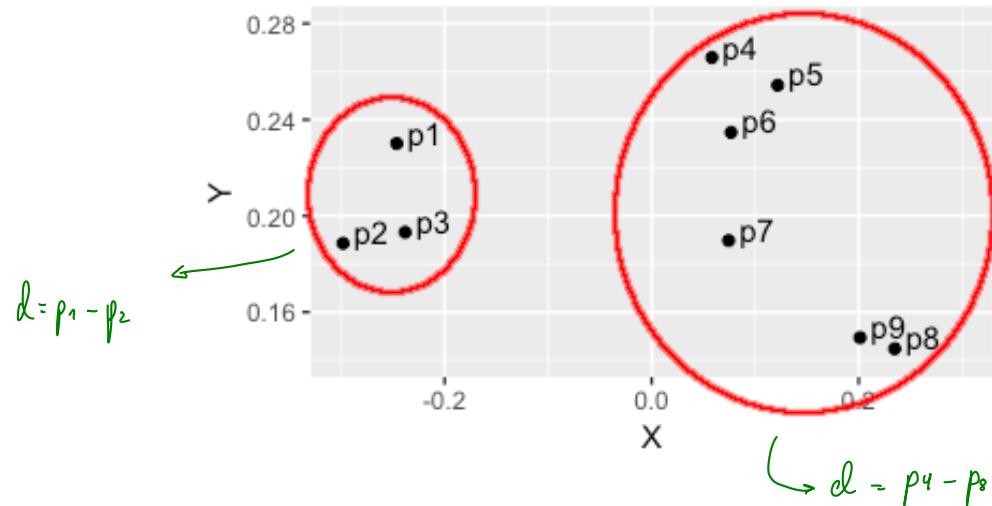
Agrupamiento Jerárquico

Divisivo

Criterios de división

$$d(S_K) = \max_{x_i, x_j \in S_K} d(x_i, x_j)$$

Diámetro



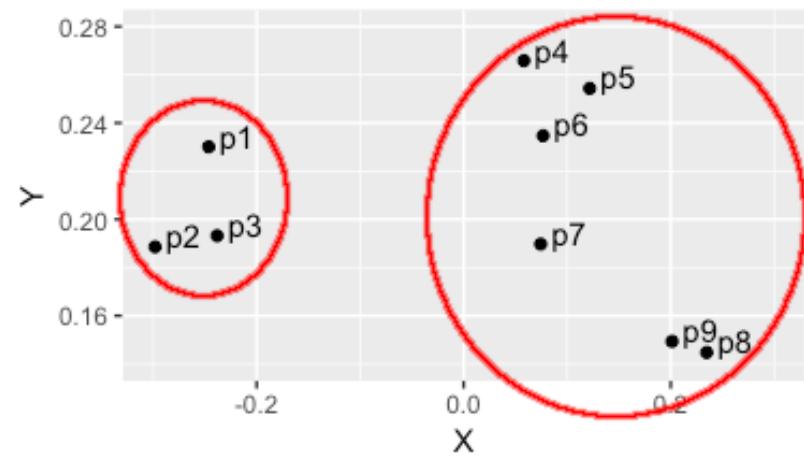
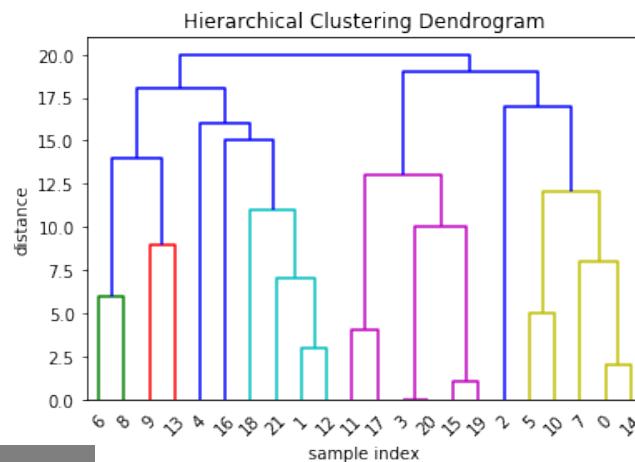
Agrupamiento Jerárquico

Divisivo

Criterios de división

$$d(S_K) = \max_{x_i, x_j \in S_K} d(x_i, x_j)$$

Diámetro



Agrupamiento Jerárquico

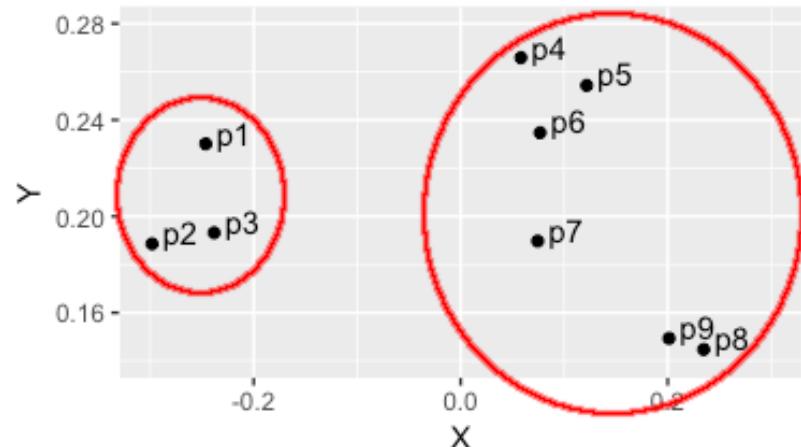
Divisivo

Criterios de división

$$d(S_K) = \frac{1}{|S_K|^2} \sum_{x_i \in S_K} \sum_{x_j \in S_K} d(x_i, x_j)$$

Disimilitud media

Mido todas las distancias entre puntos del cluster y tomo la media.



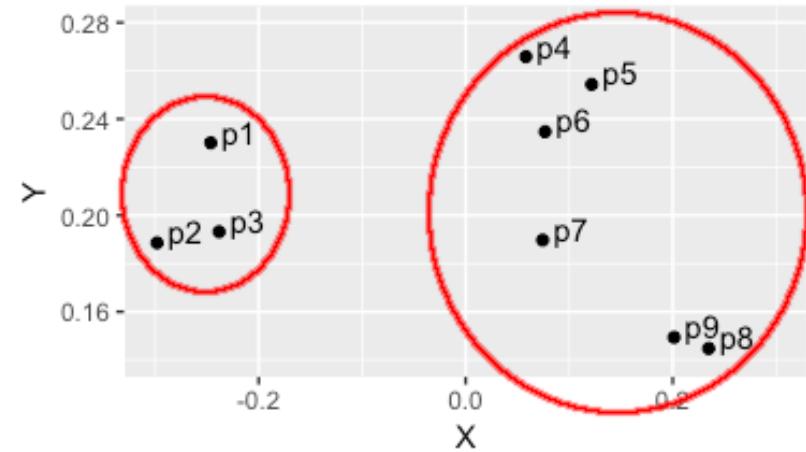
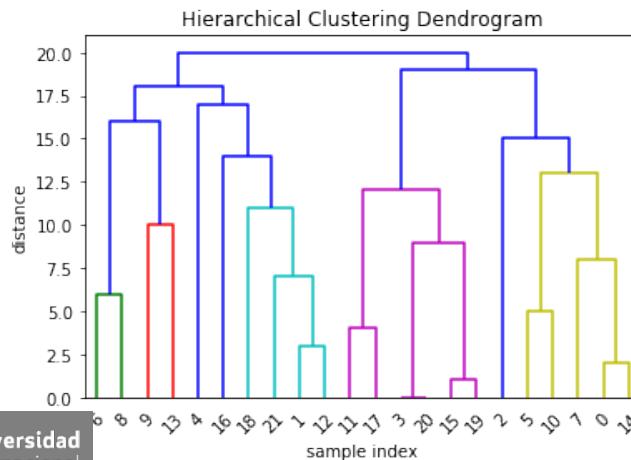
Agrupamiento Jerárquico

Divisivo

Criterios de división

$$d(S_K) = \frac{1}{|S_K|^2} \sum_{x_i \in S_K} \sum_{x_j \in S_K} d(x_i, x_j)$$

Disimilitud media



Agrupamiento Jerárquico

Divisivo

Segunda cuestión

A medida que avanza el algoritmo, una vez se ha elegido un clúster a dividir...

¿cómo se divide el clúster seleccionado?

La mejor separación del conjunto de datos:

$$\{S_A^*, S_B^*\} = \arg \max_{S_A, S_B \subset S_K} \text{sep}(S_A, S_B)$$

Se separa primero el de \uparrow similitud intracluster

Agrupamiento Jerárquico

Divisivo

Segunda cuestión

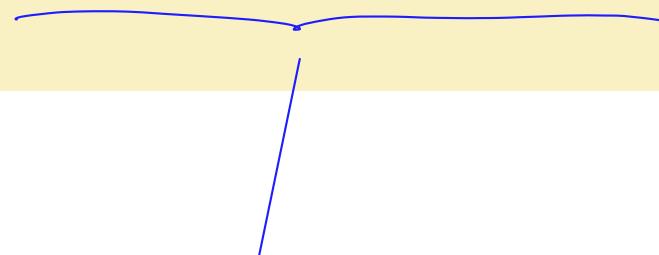
A medida que avanza el algoritmo, una vez se ha elegido un clúster a dividir...

¿cómo se divide el clúster seleccionado?

La mejor separación del conjunto de datos:

$$\{S_A^*, S_B^*\} = \arg \max_{S_A, S_B \subset S_K} \text{sep}(S_A, S_B)$$

¿cómo se encuentra la mejor separación o división?



Agrupamiento Jerárquico

Divisivo

Criterios de división

K -means

Aplicar el popular algoritmo K -means con $K = 2$



Agrupamiento Jerárquico

Divisivo

Criterios de división

K -means

Aplicar el popular algoritmo K -means con $K = 2$

Agrupamiento Jerárquico

Divisivo

Cierre 28/09/23 min 25:19

Criterios de división

Técnica de Macnaughton-Smith

- ▶ Elegir el ejemplo x que en media está más lejano del resto del clúster, S_K
- ▶ $S_B = \underline{S_K \setminus \{x\}}$; $S_A = \underline{\{x\}}$
 todos los puntos menos esa x la x más lejana
- ▶ Ir añadiendo casos x^*
$$x^* = \operatorname{argmáx}_{x_i \in S_B} \left(\frac{1}{|S_B|-1} \sum_{x_{i'} \in S_B: x_i \neq x_{i'}} d(x_i, x_{i'}) - \frac{1}{|S_A|} \sum_{x_j \in S_A} d(x_i, x_j) \right)$$

el siguiente más cercano a
lo que ya se tienen en S_A
- ▶ Hasta que el valor de la función argmax sea inferior a 0

↳ hasta que la distancia intraclúster sea mayor

en S_A que en S_B

Agrupamiento Jerárquico

Divisivo

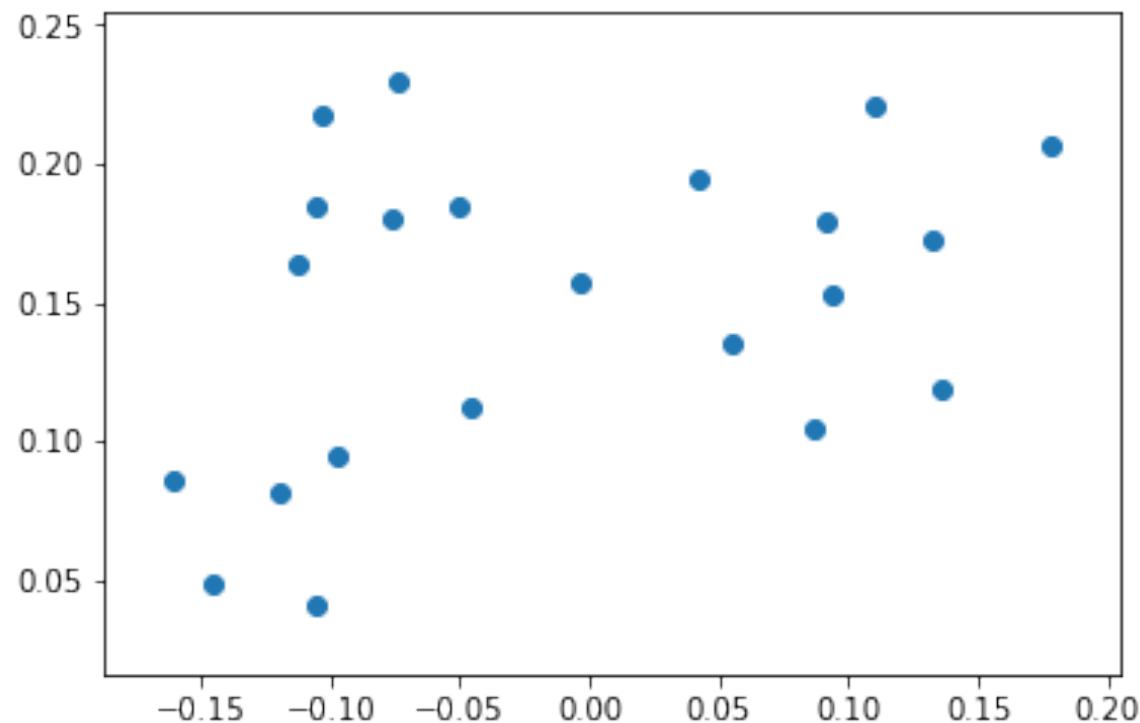
Criterios de división

Técnica de Macnaughton-Smith

- ▶ Elegir el ejemplo x que en media está más lejano del resto del clúster, S_K
- ▶ $S_B = S_K \setminus \{x\}; S_A = \{x\}$
- ▶ Ir añadiendo casos x^*
$$x^* = \operatorname{argmáx}_{x_i \in S_B} \left(\frac{1}{|S_B|-1} \sum_{x_{i'} \in S_B: x_i \neq x_{i'}} d(x_i, x_{i'}) - \frac{1}{|S_A|} \sum_{x_j \in S_A} d(x_i, x_j) \right)$$
- ▶ Hasta que el valor de la función *argmax* sea inferior a 0

Agrupamiento Jerárquico

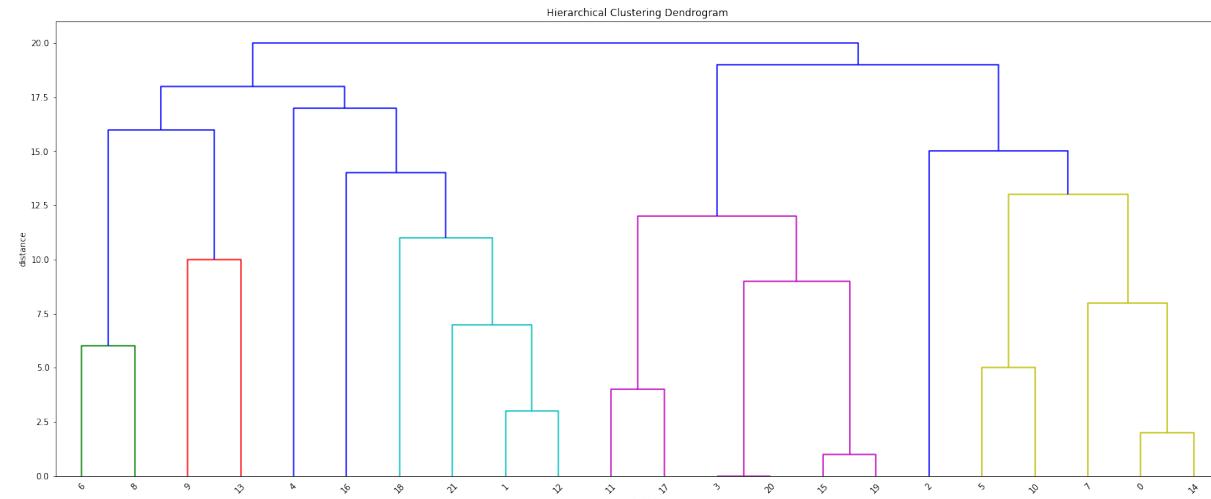
Divisivo



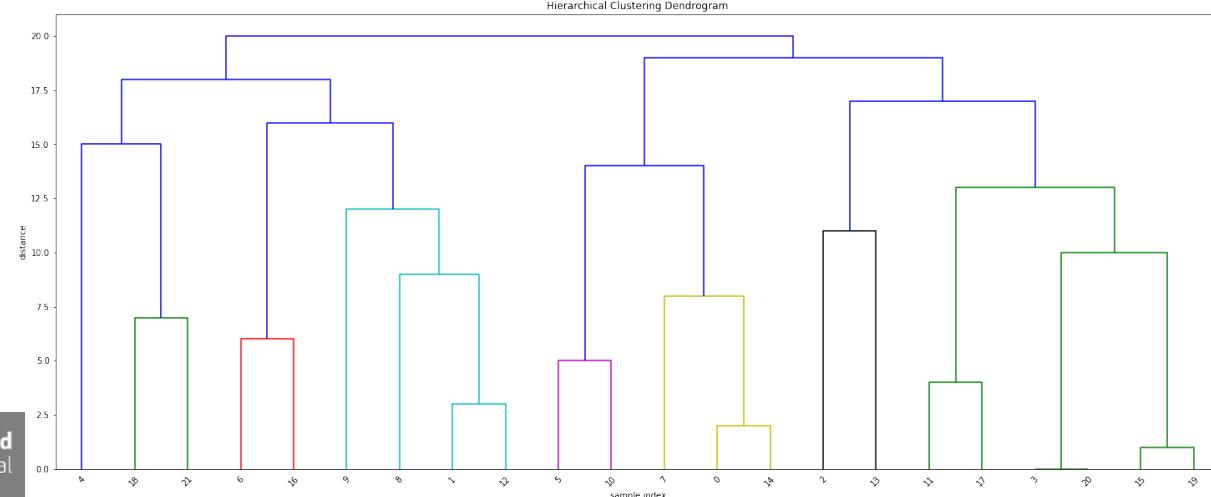
Agrupamiento Jerárquico

Divisivo

Mac.-Smith

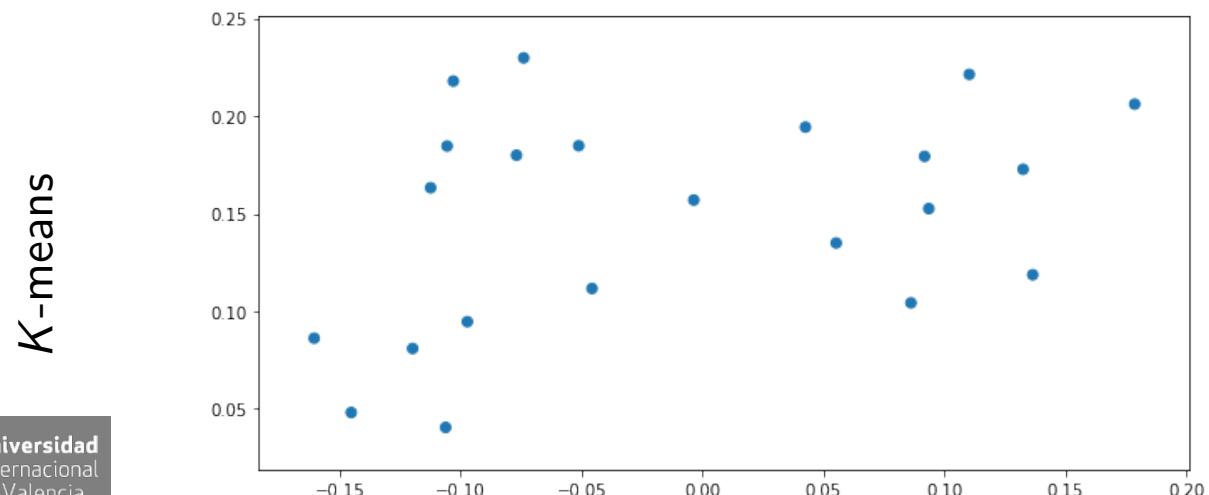
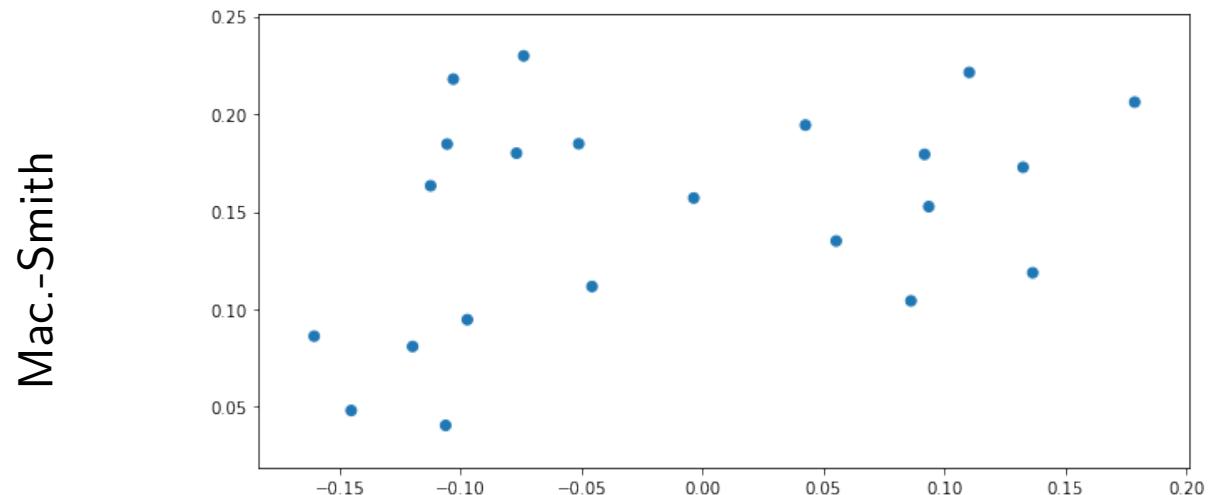


K-means



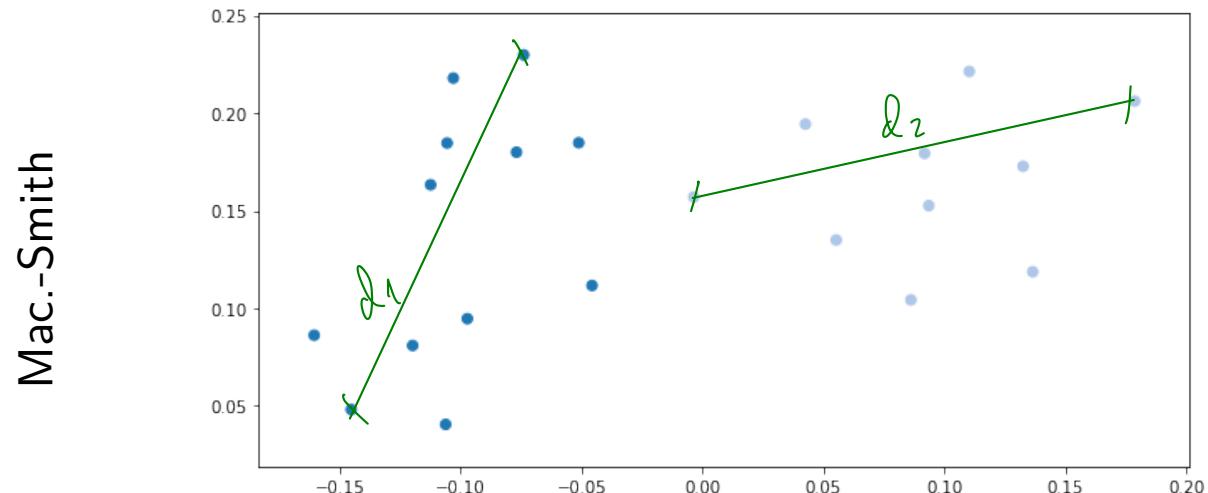
Agrupamiento Jerárquico

Divisivo

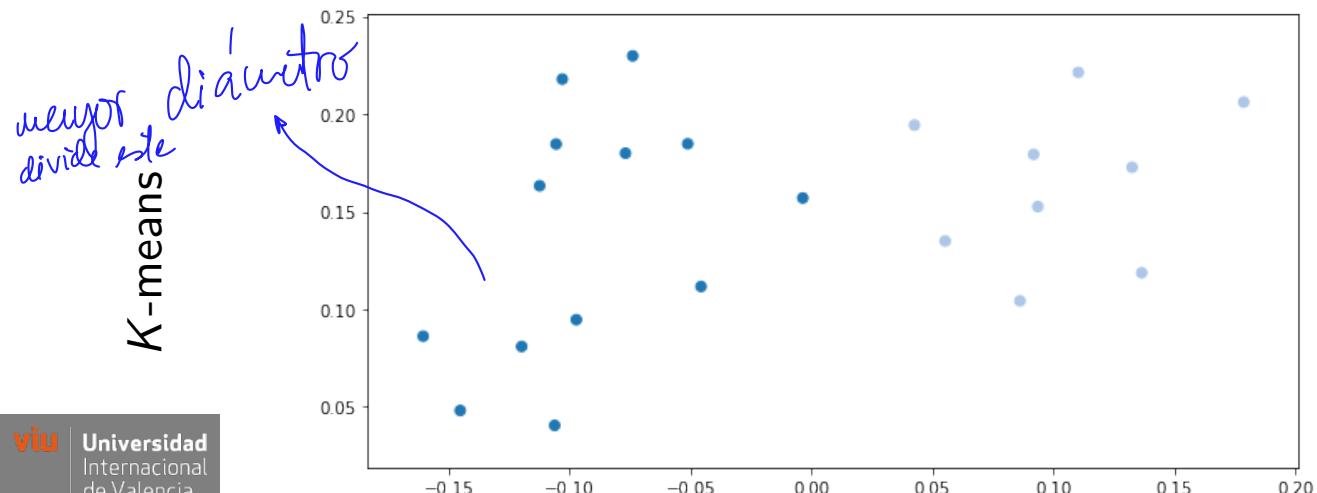


Agrupamiento Jerárquico

Divisivo



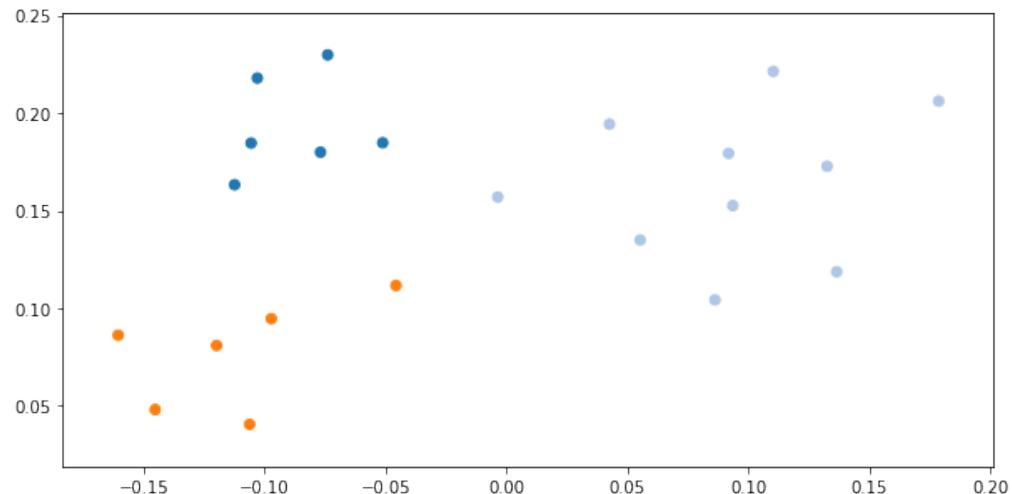
$d_2 > d_1$
siguiente
iteración



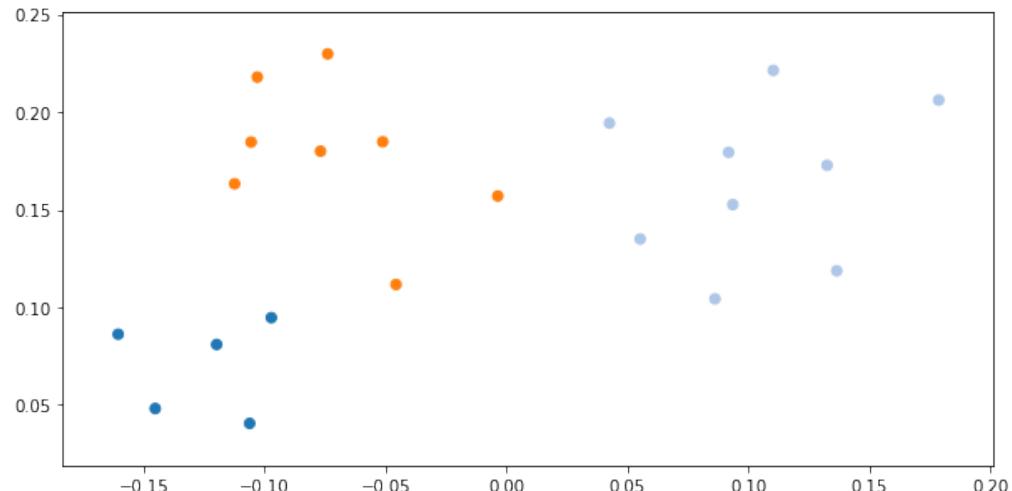
Agrupamiento Jerárquico

Divisivo

Mac.-Smith



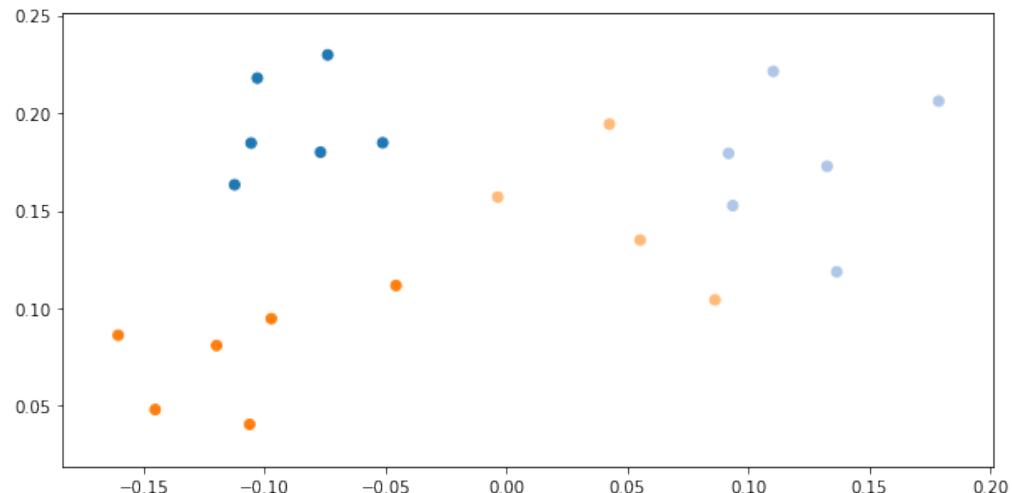
K-means



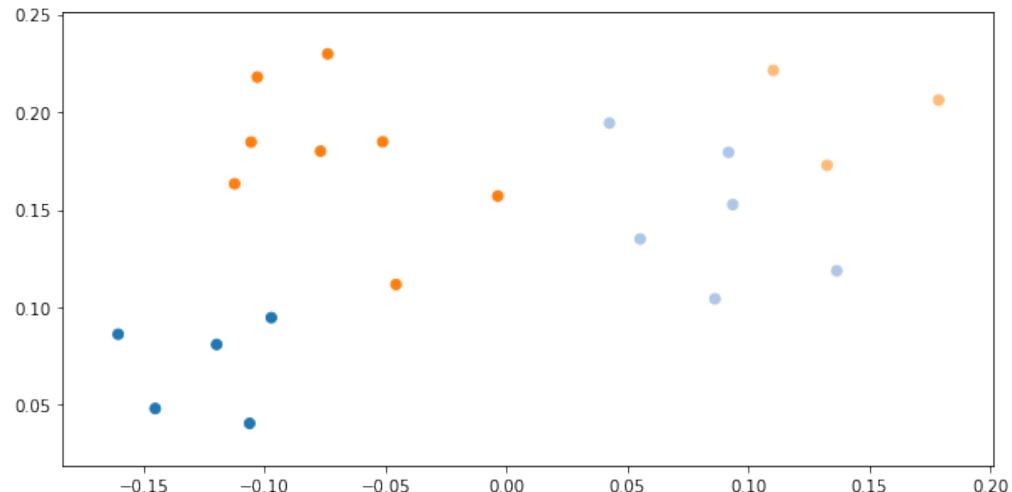
Agrupamiento Jerárquico

Divisivo

Mac.-Smith



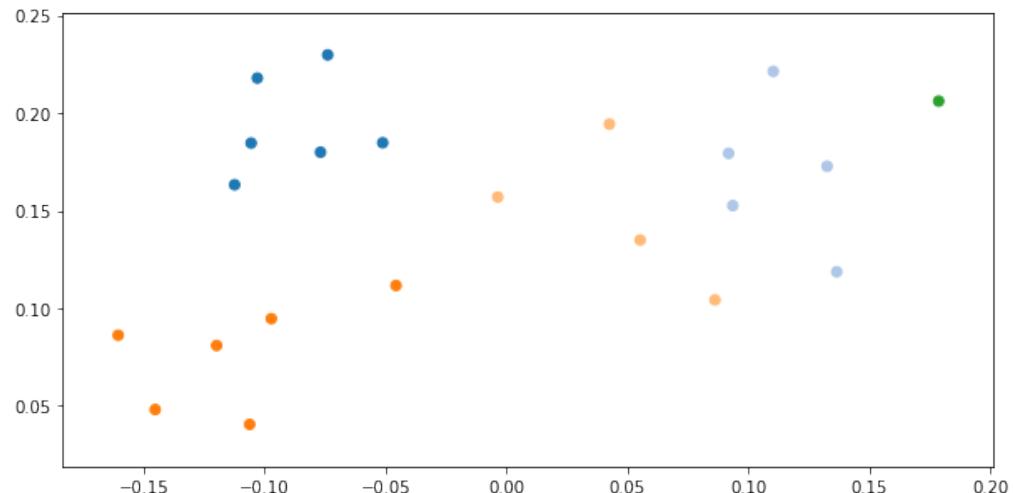
K-means



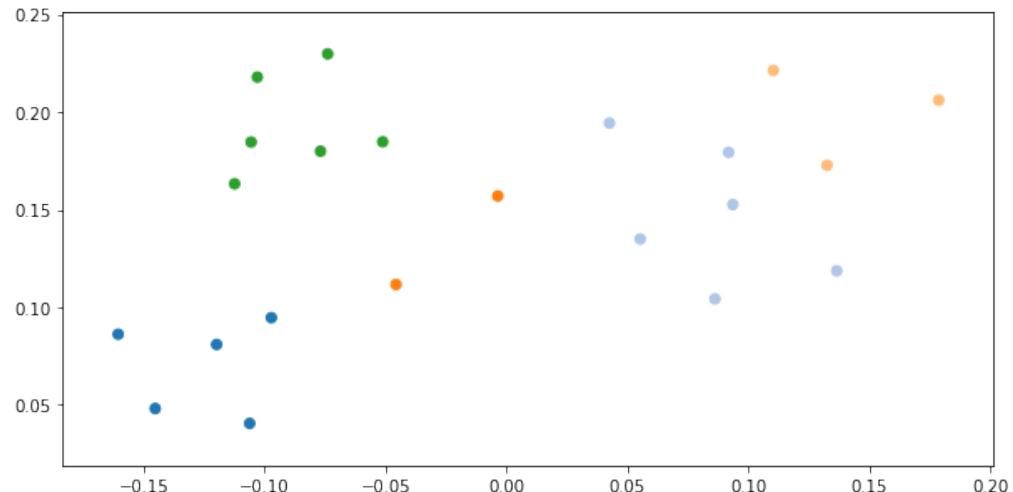
Agrupamiento Jerárquico

Divisivo

Mac.-Smith



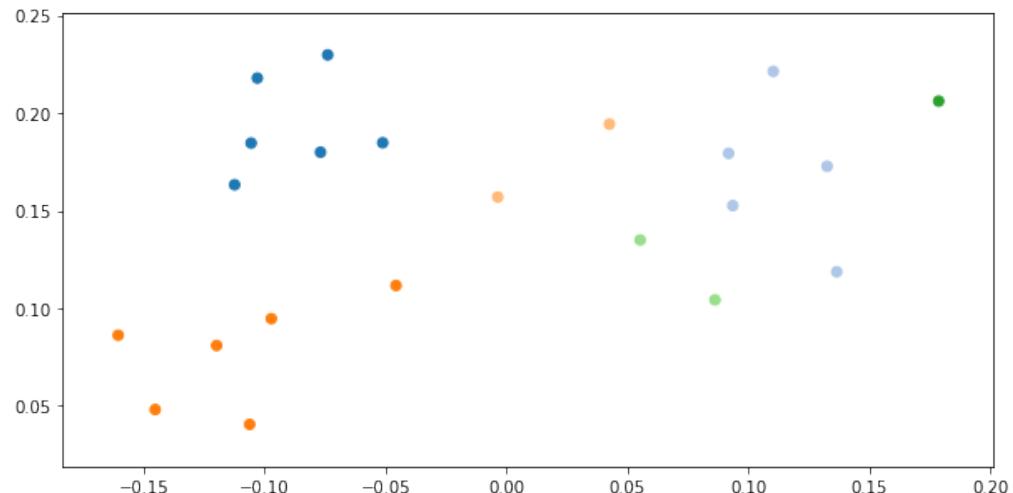
K-means



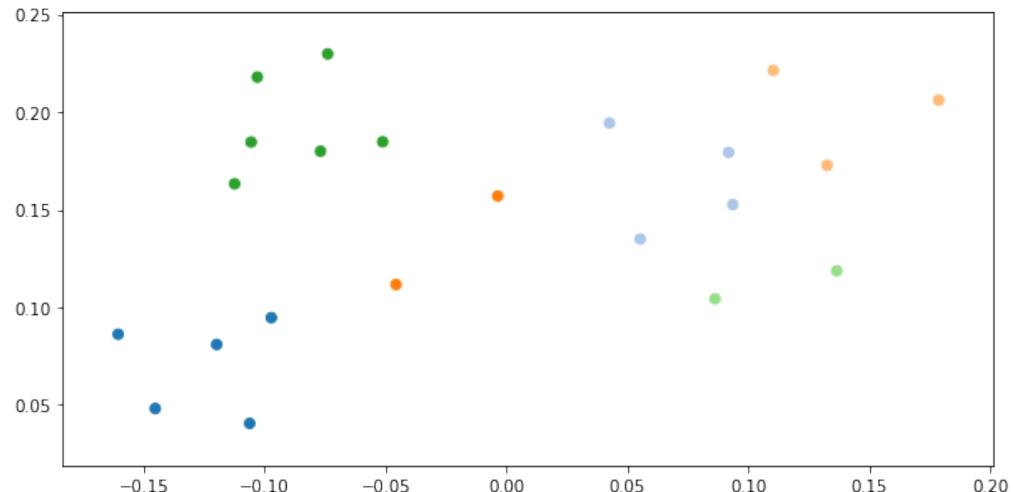
Agrupamiento Jerárquico

Divisivo

Mac.-Smith



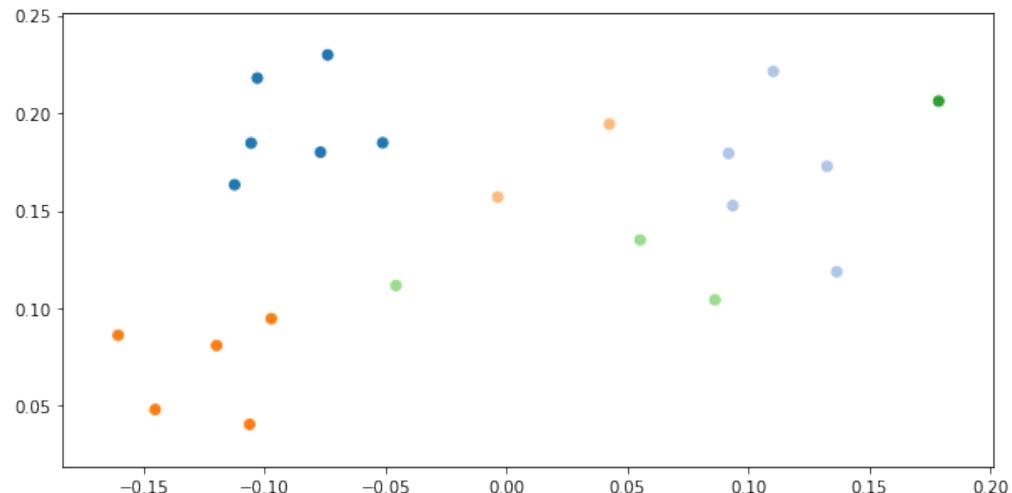
K-means



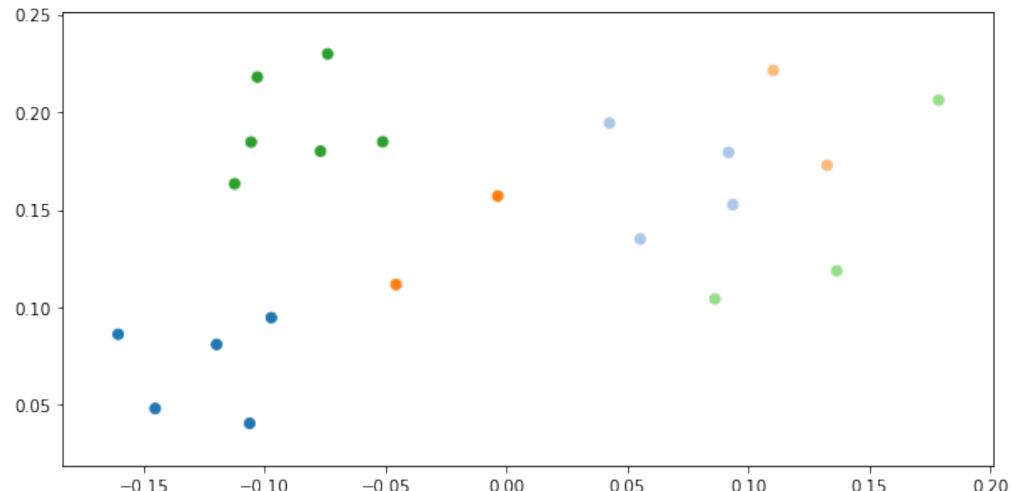
Agrupamiento Jerárquico

Divisivo

Mac.-Smith



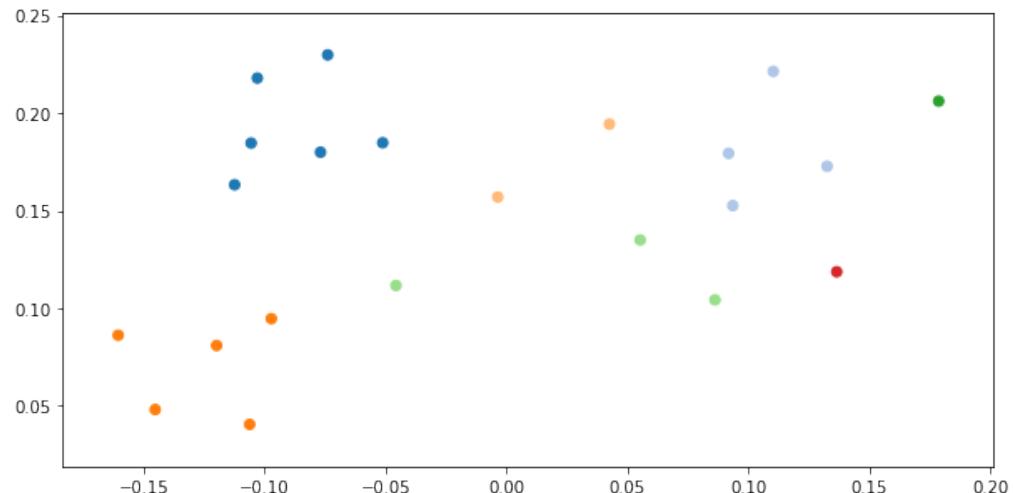
K-means



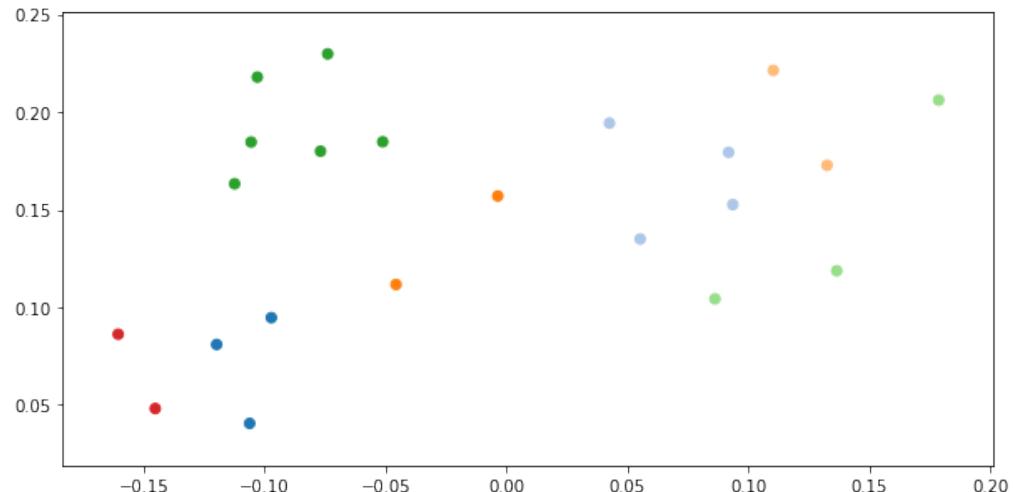
Agrupamiento Jerárquico

Divisivo

Mac.-Smith



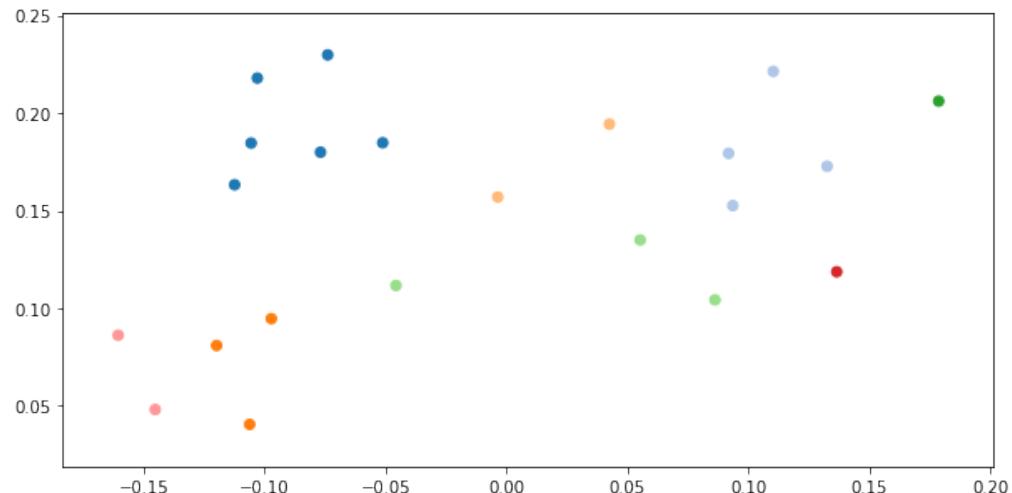
K-means



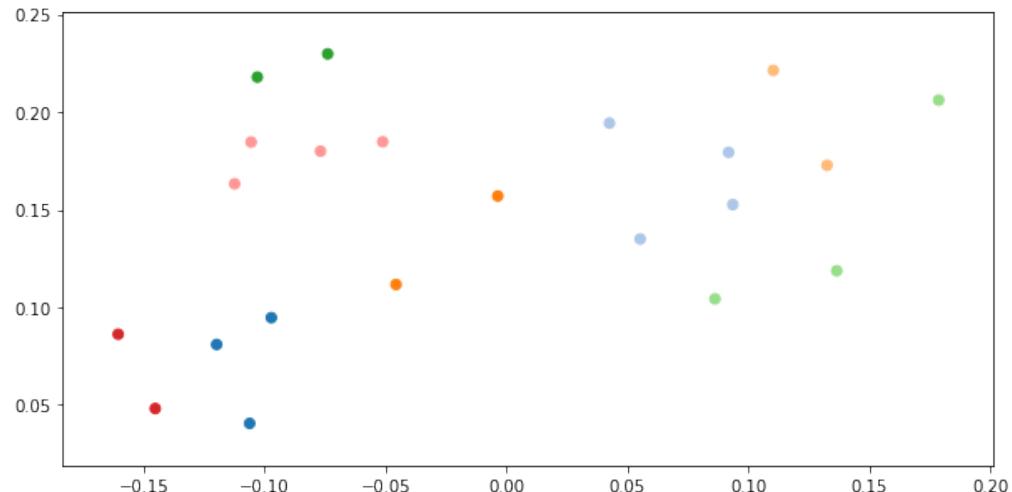
Agrupamiento Jerárquico

Divisivo

Mac.-Smith



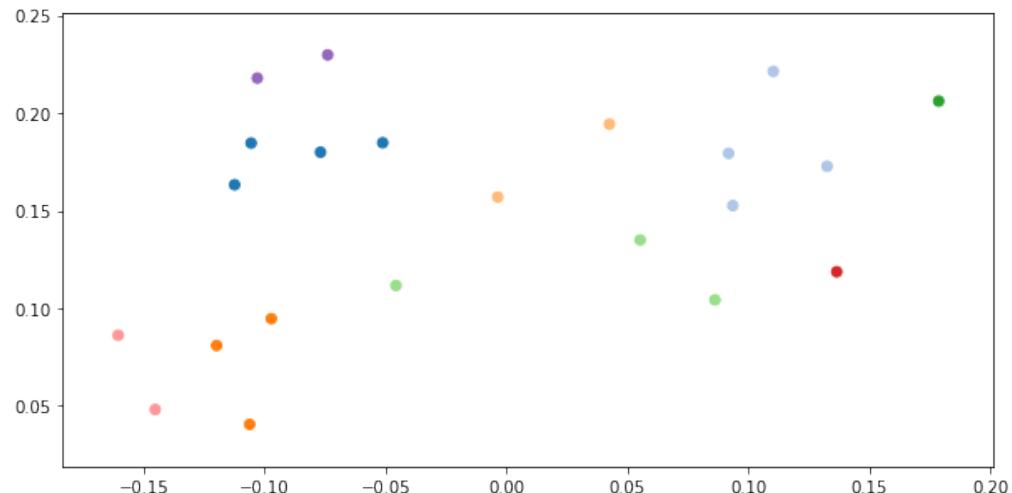
K-means



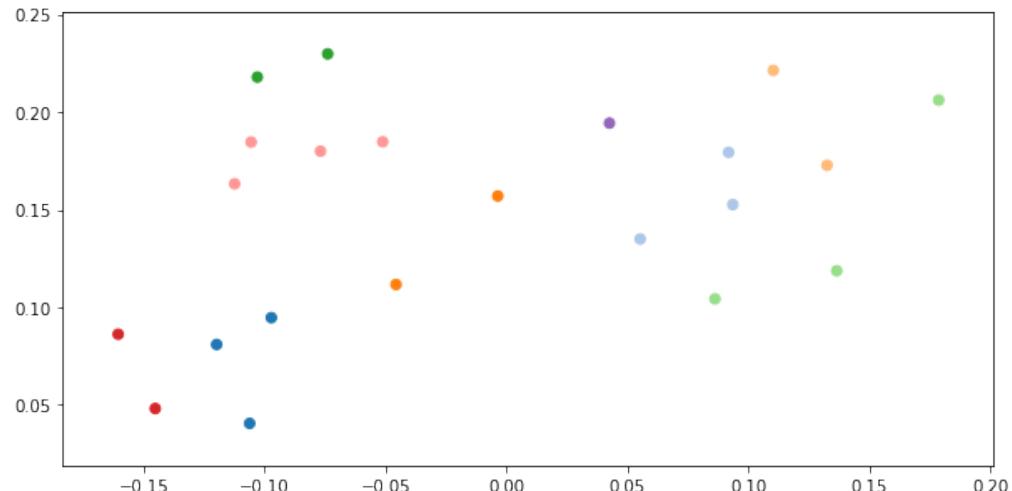
Agrupamiento Jerárquico

Divisivo

Mac.-Smith



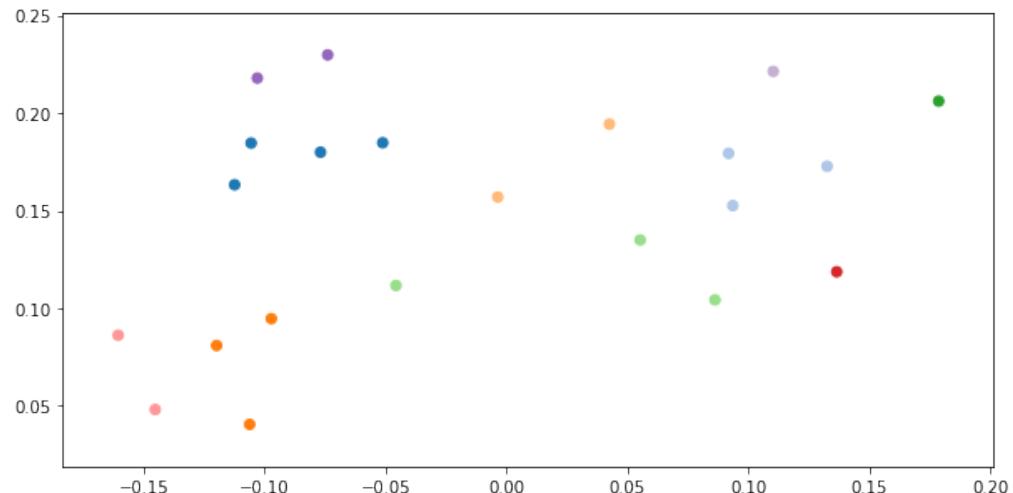
K-means



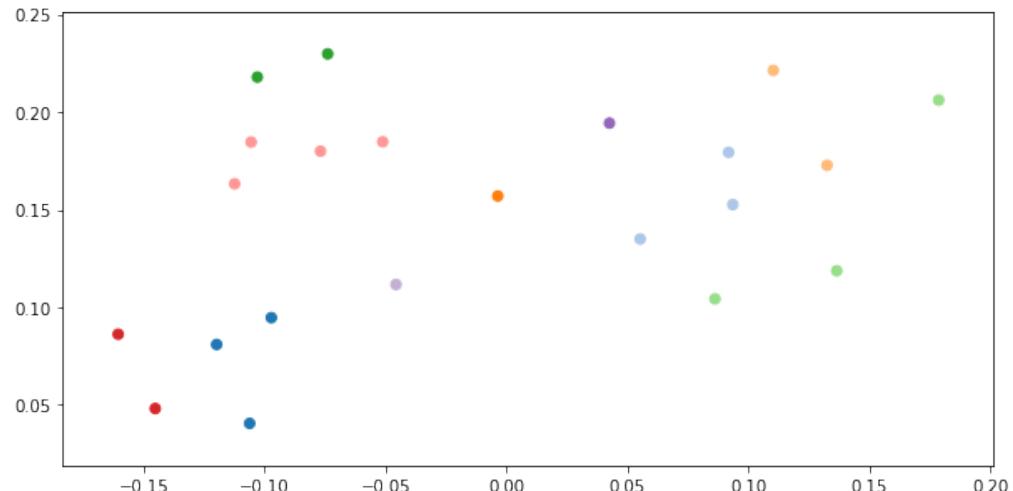
Agrupamiento Jerárquico

Divisivo

Mac.-Smith



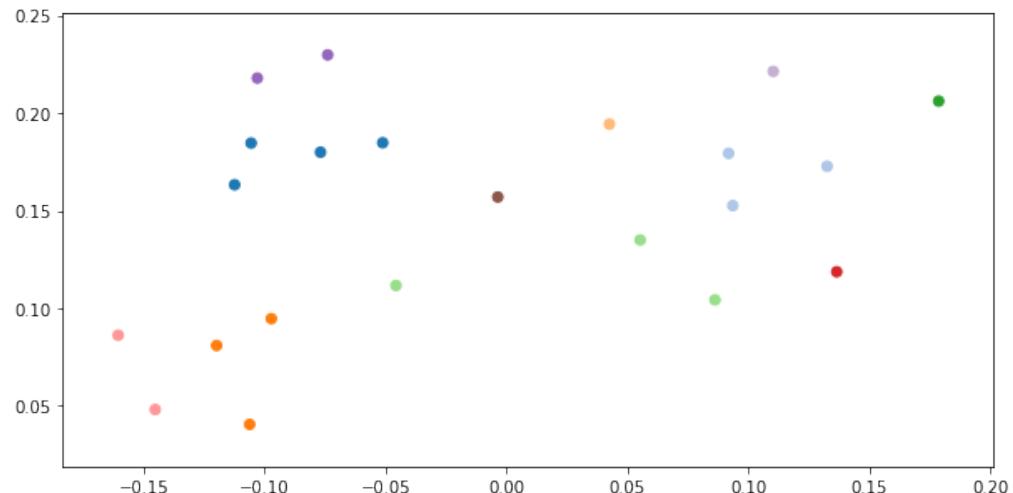
K-means



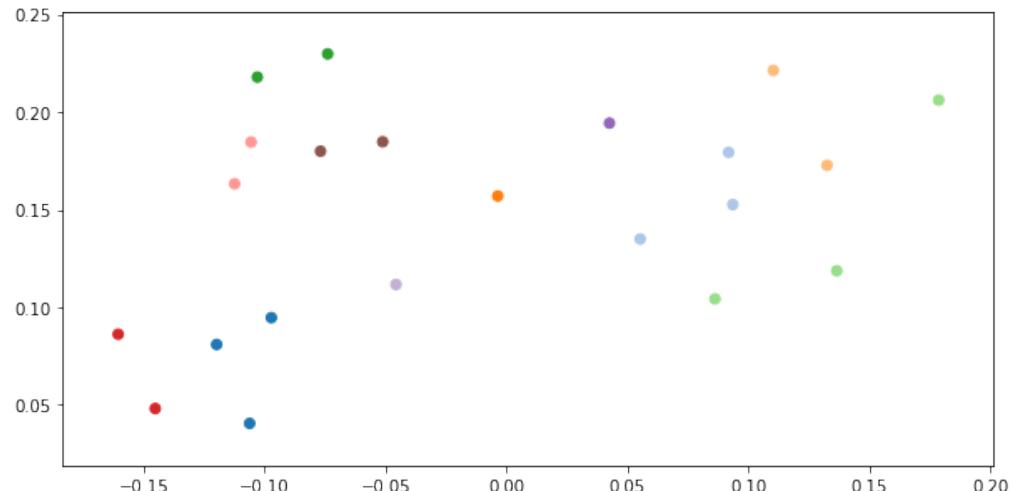
Agrupamiento Jerárquico

Divisivo

Mac.-Smith



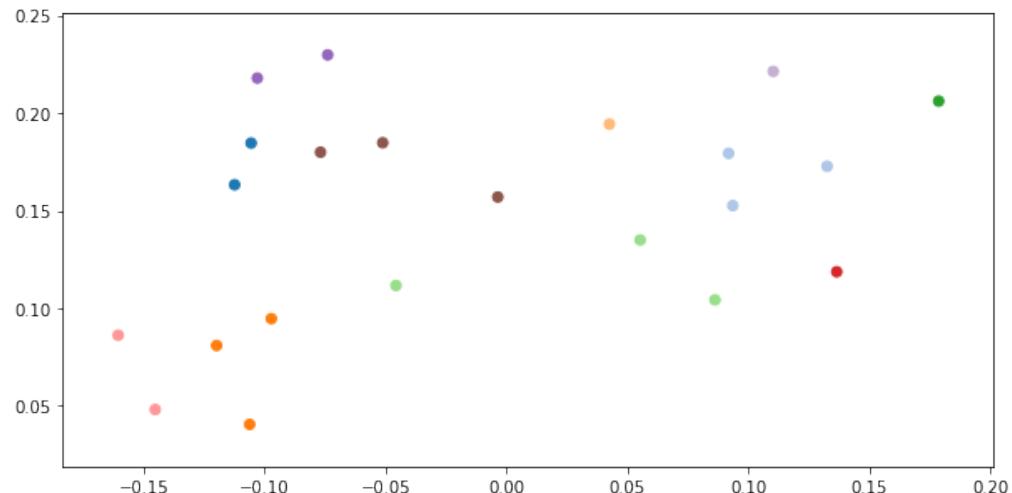
K-means



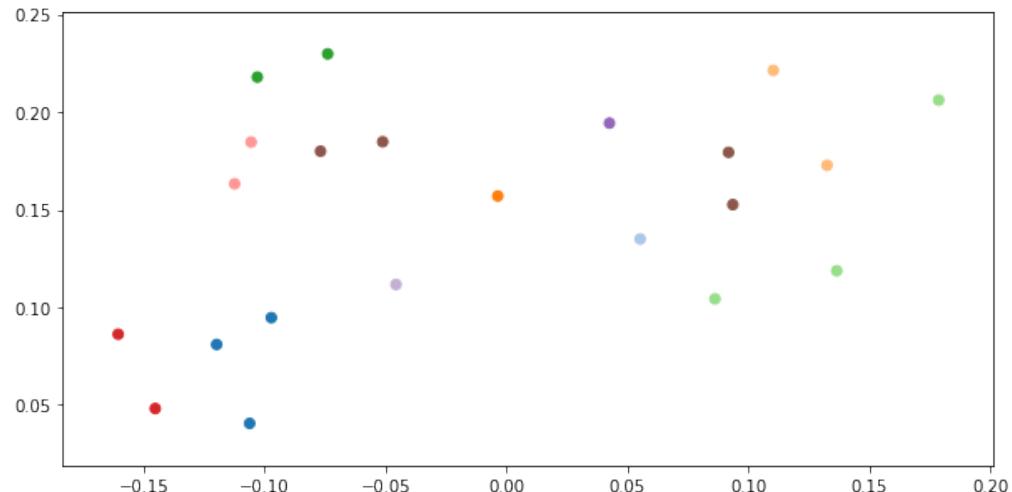
Agrupamiento Jerárquico

Divisivo

Mac.-Smith



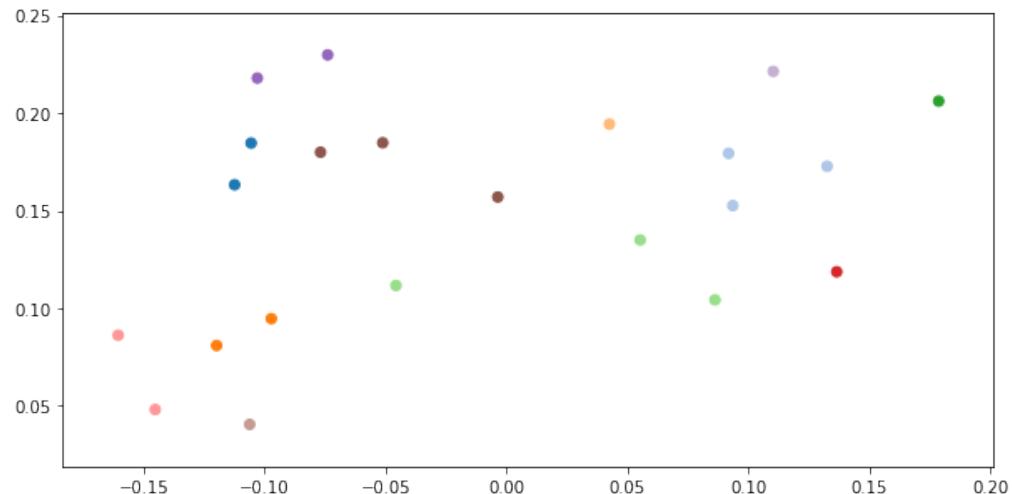
K-means



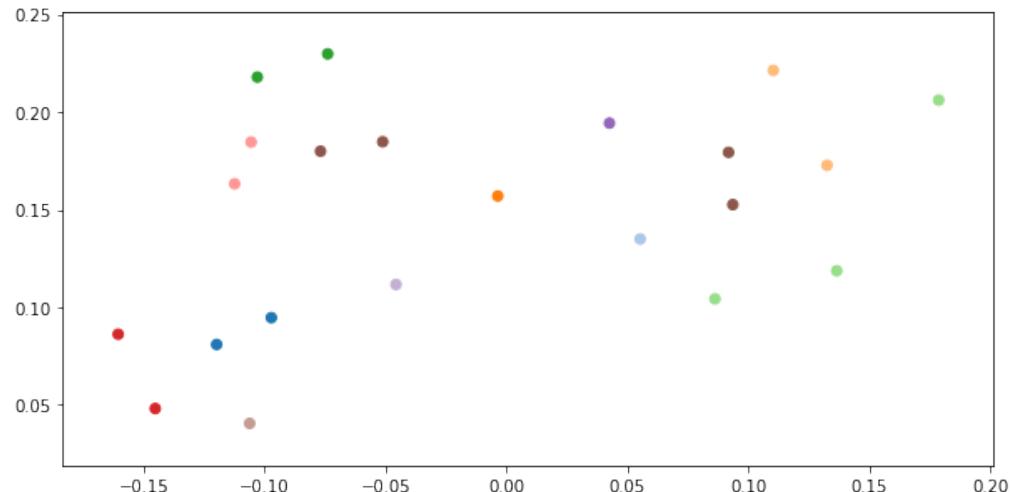
Agrupamiento Jerárquico

Divisivo

Mac.-Smith



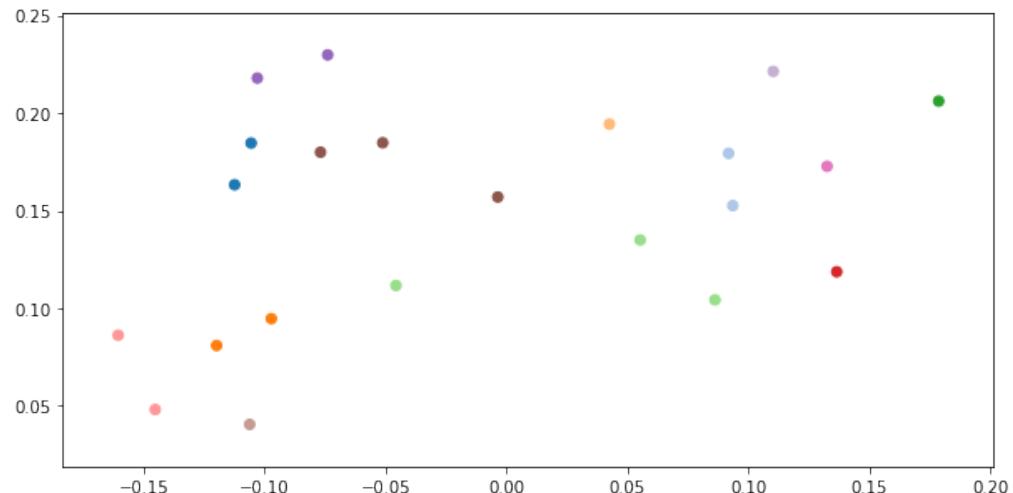
K-means



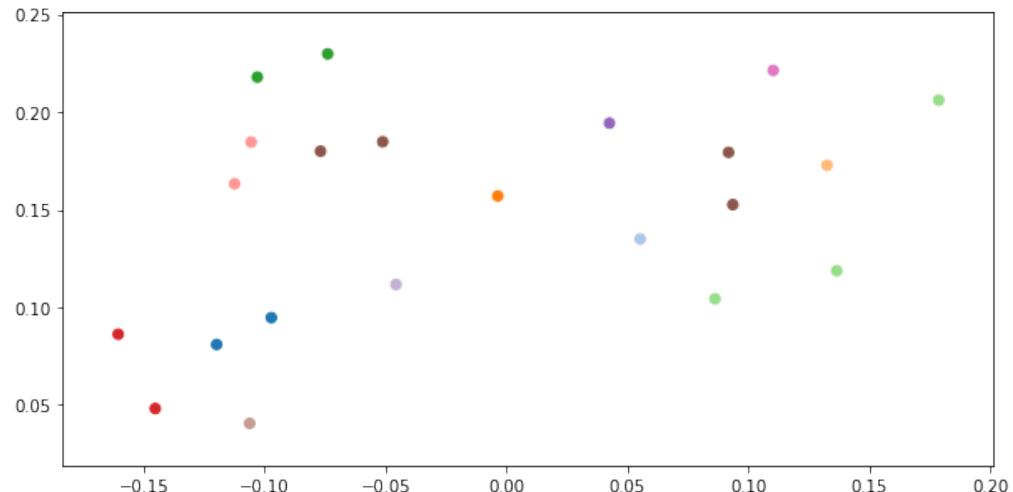
Agrupamiento Jerárquico

Divisivo

Mac.-Smith

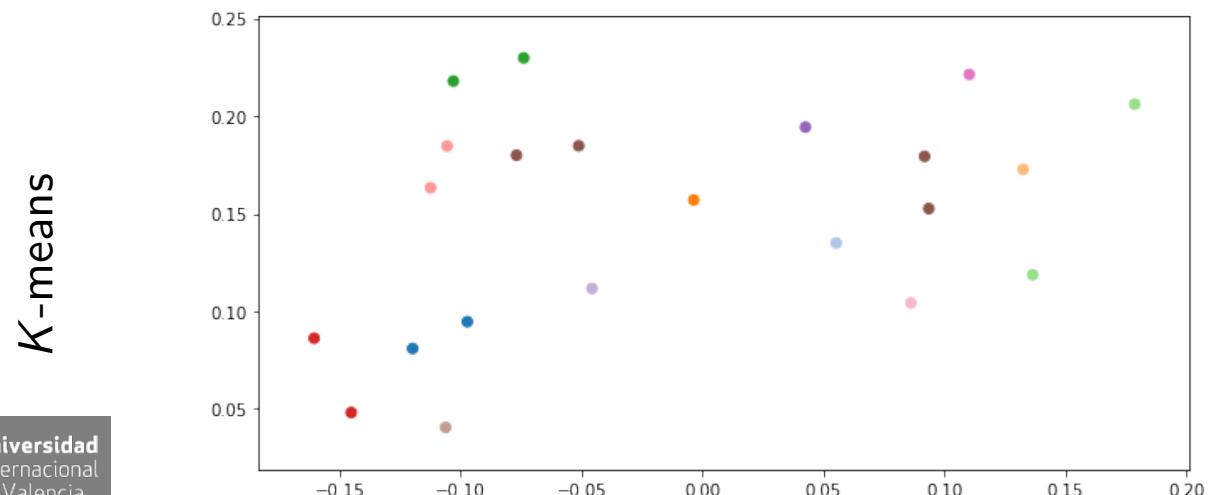
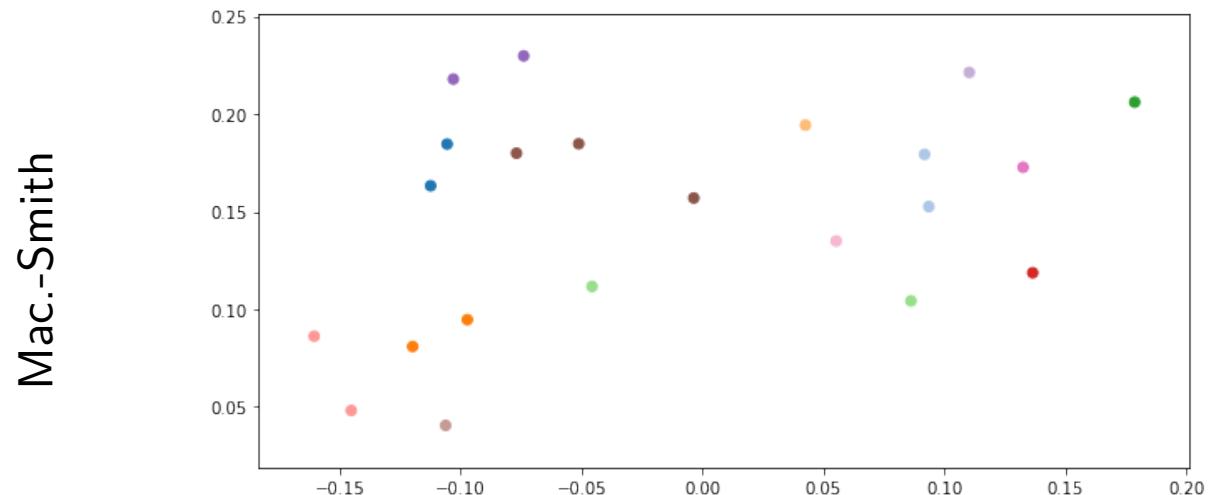


K-means



Agrupamiento Jerárquico

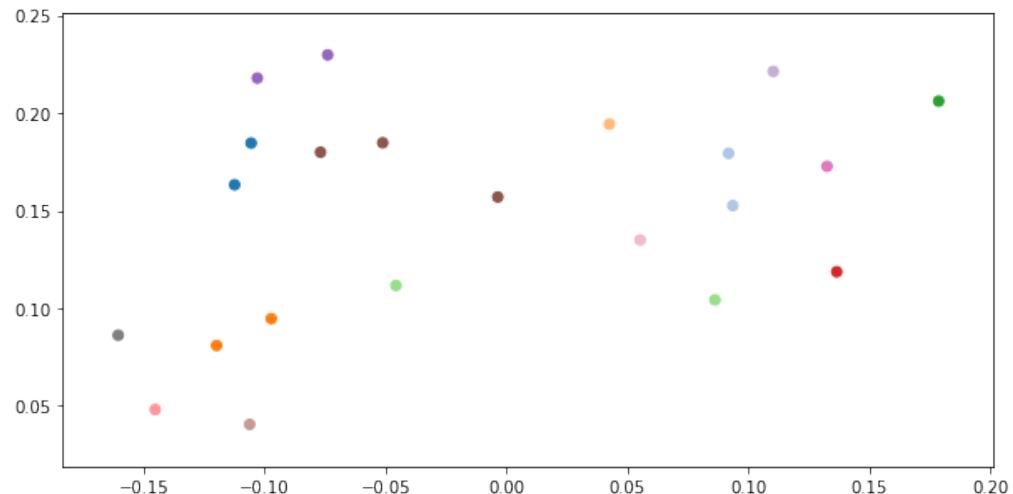
Divisivo



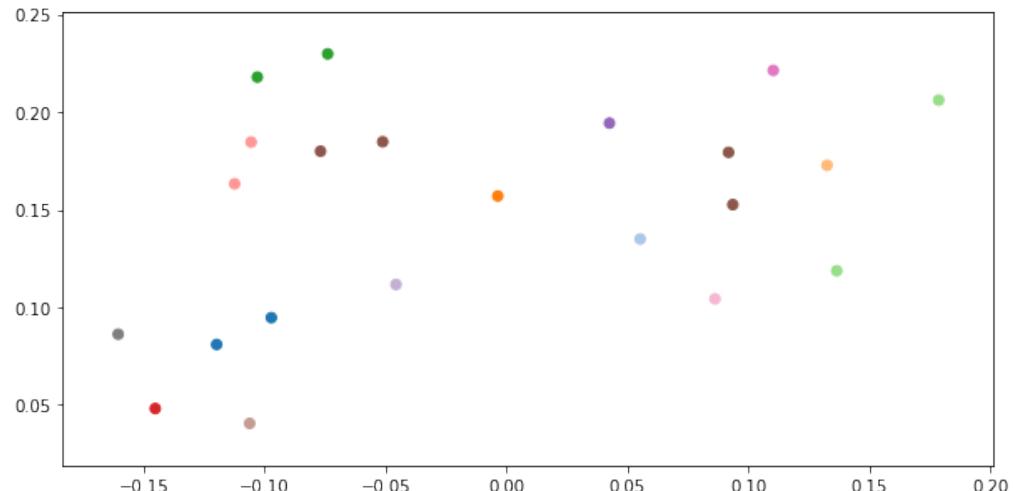
Agrupamiento Jerárquico

Divisivo

Mac.-Smith



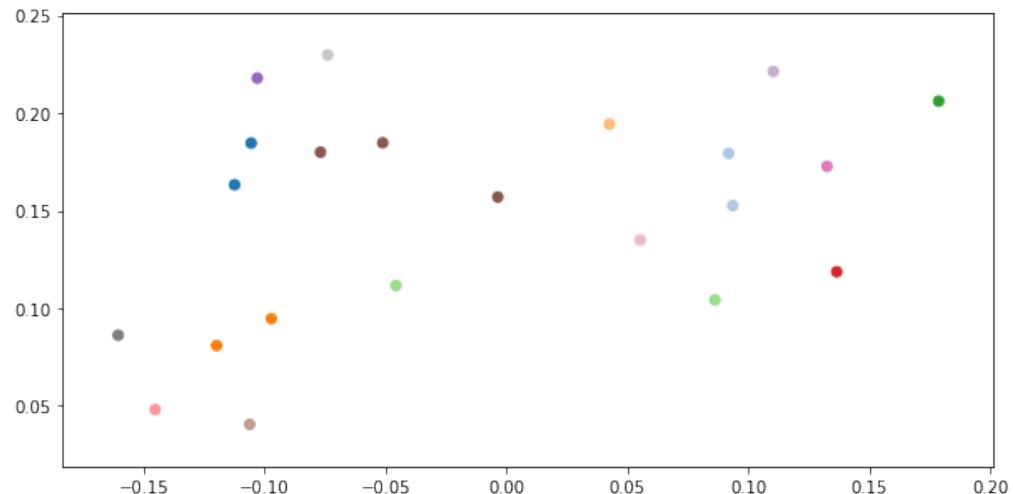
K-means



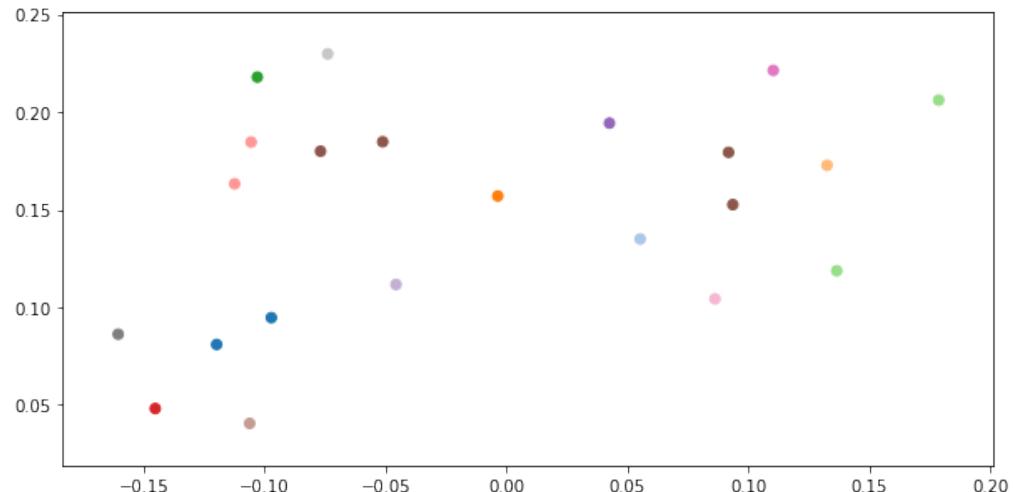
Agrupamiento Jerárquico

Divisivo

Mac.-Smith



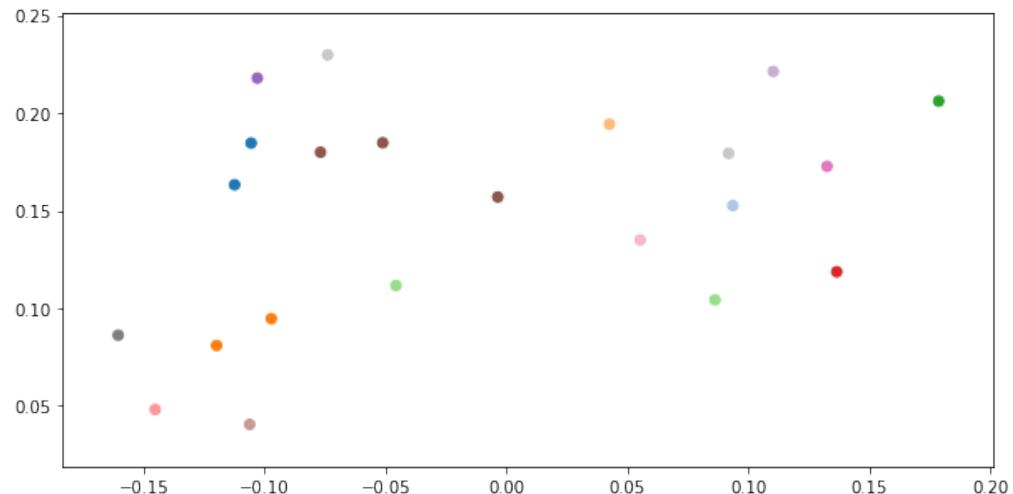
K-means



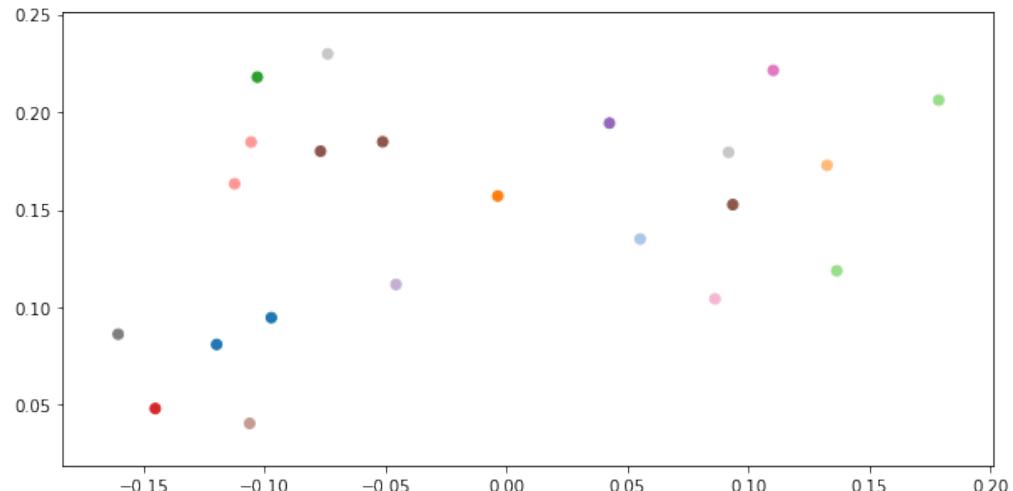
Agrupamiento Jerárquico

Divisivo

Mac.-Smith



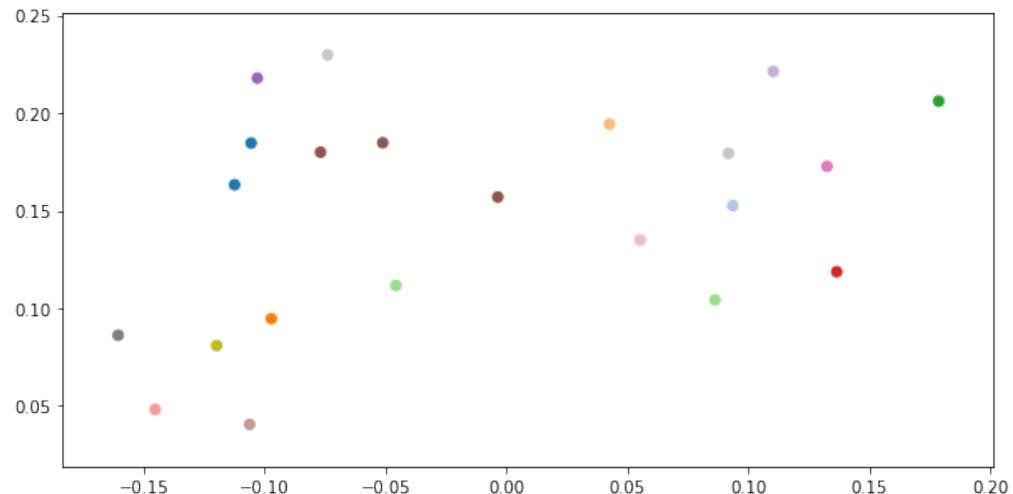
K-means



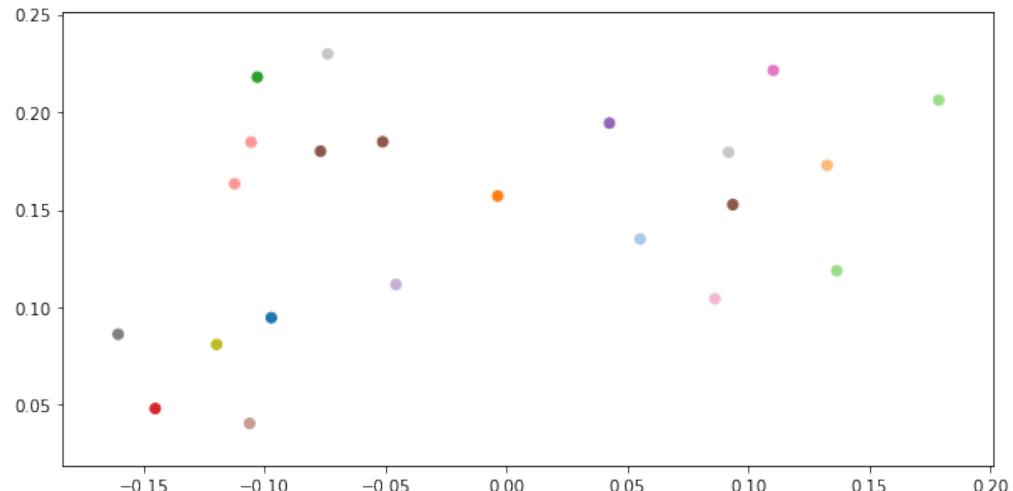
Agrupamiento Jerárquico

Divisivo

Mac.-Smith



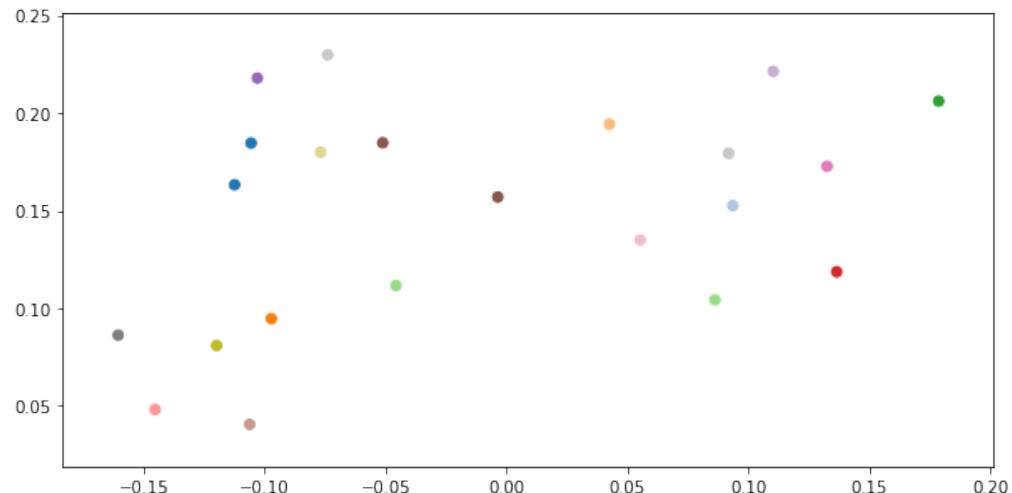
K-means



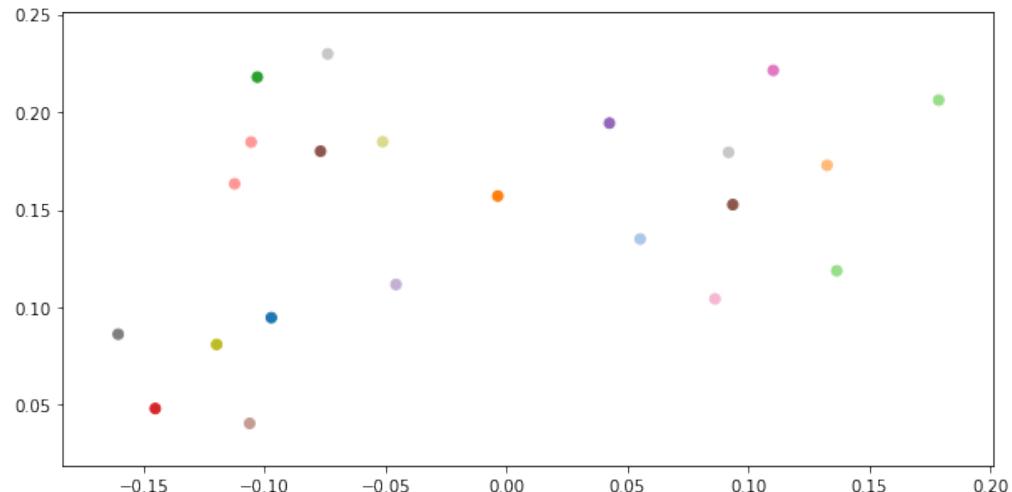
Agrupamiento Jerárquico

Divisivo

Mac.-Smith



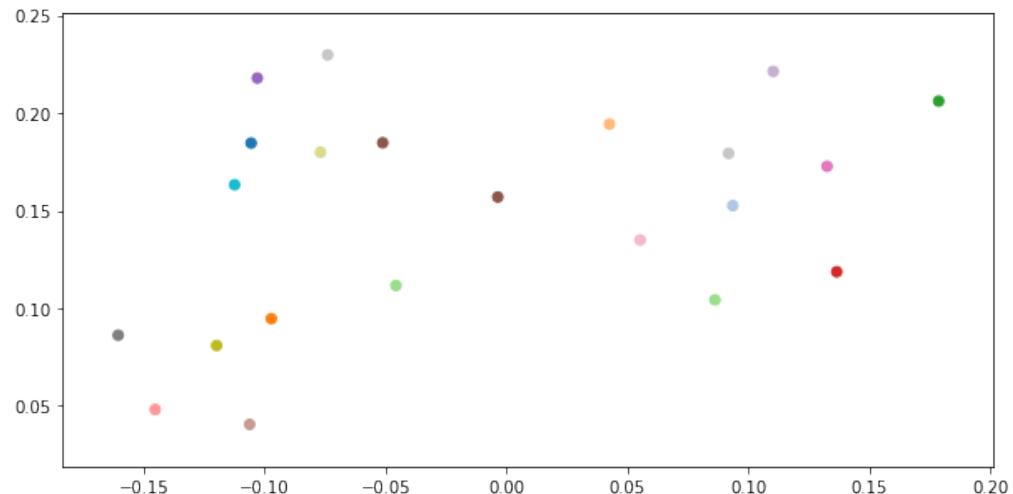
K-means



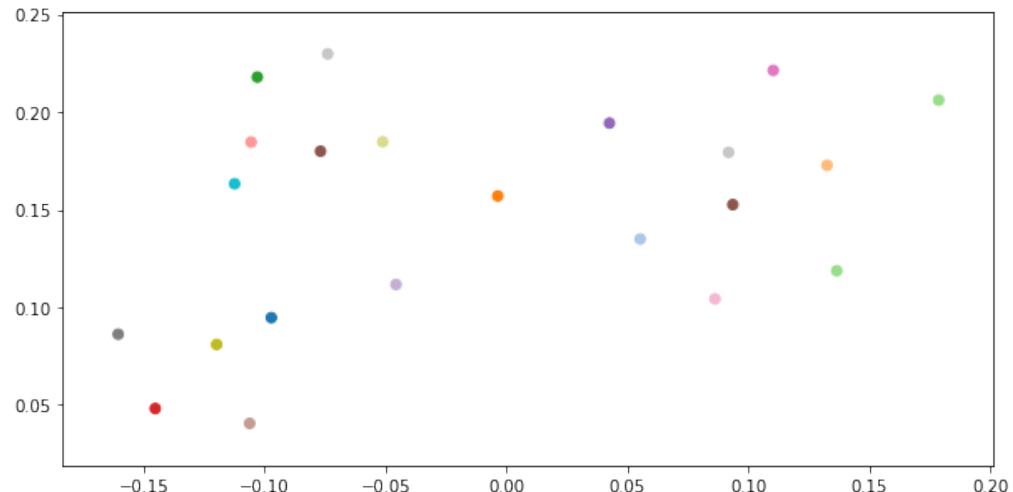
Agrupamiento Jerárquico

Divisivo

Mac.-Smith



K-means



Agrupamiento Jerárquico

Divisivo

[EXAMEN]

¿Qué algoritmo es más complejo, el aglomerativo o el divisivo?

El divisivo, en cada paso hay dos decisiones ¿cuál divido? y ¿cómo lo divido? mientras que en el aglomerativo es sólo ¿cuál uno?

Ventajas

- ▶ Intuitivo
- ▶ Funciona con clústeres de diferente tamaño
- ▶ Diferentes criterios y maneras de dividir
- ▶ Puede funcionar con diferentes medidas de distancia

Agrupamiento Jerárquico

Divisivo

Desventajas

- ▶ Muy lento
- ▶ Problemas al lidiar con clústeres de diferente densidad
- ▶ Dos decisiones de división

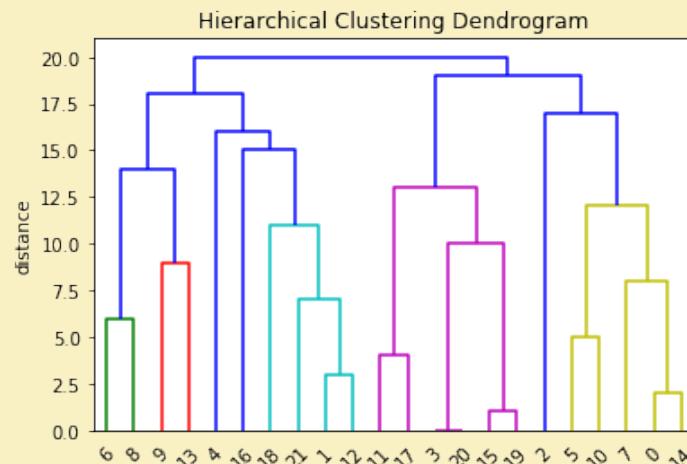
Agrupamiento Jerárquico

Divisivo

Desventajas

28/09/2023 notebook min 42:14

- ▶ Muy lento
- ▶ Problemas al lidiar con clústeres de diferente densidad
- ▶ Dos decisiones de división
- ▶ ¿Qué partición elegir?
 - ▶ Número de clústeres concreto (fijando K)
 - ▶ Máxima distancia en la unión de clústeres



28/09/2023 - min 1:09:50

Aprendizaje no supervisado

VC04: Agrupamiento espectral – Conocimientos básicos

Rocío del Amor del Amor

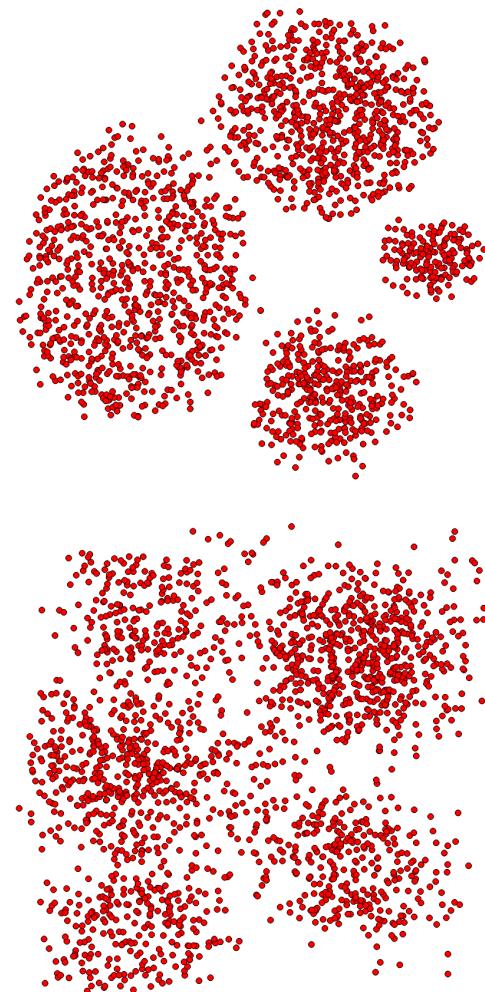
mrocio.delamor@campusviu.es

Universidad Internacional de Valencia

Agrupamiento

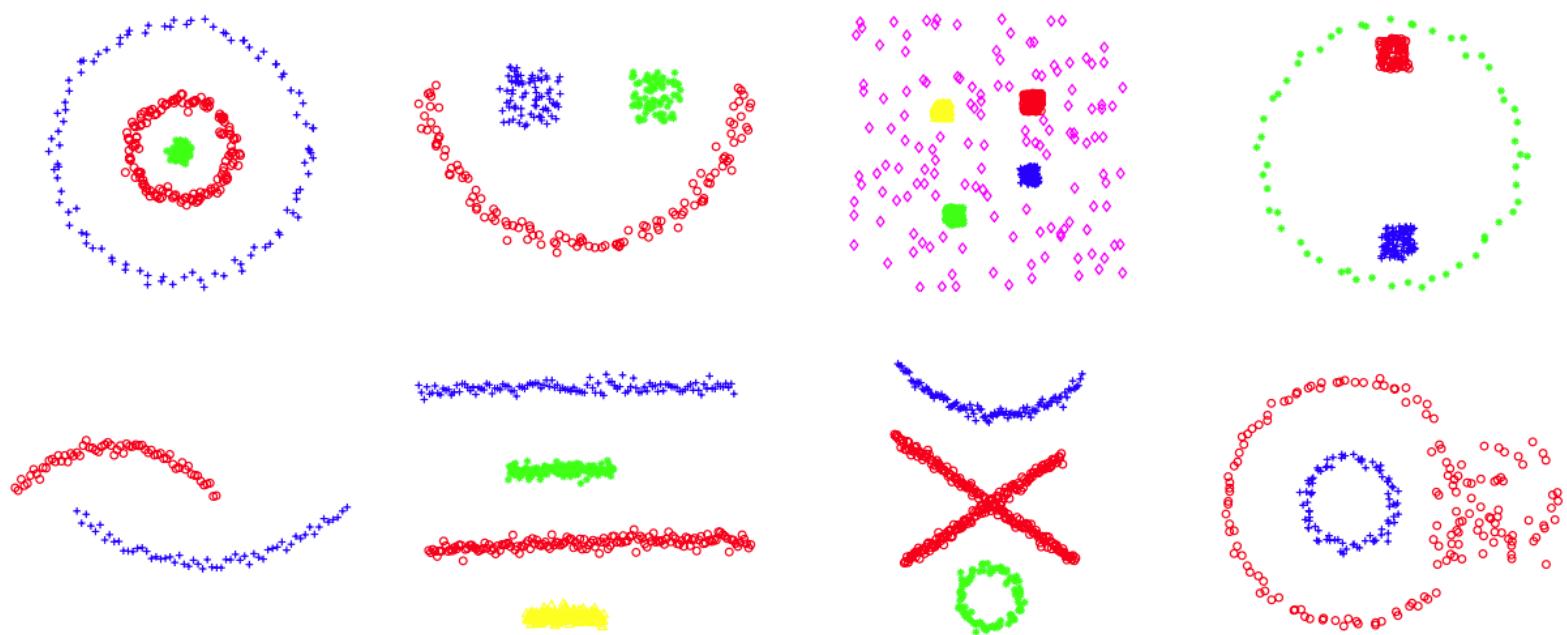
Tipos de algoritmos de agrupamiento

- ▶ Basados en particiones
- ▶ Jerárquicos
- ▶ **Espectrales**
- ▶ Basados en densidad
- ▶ Probabilísticos



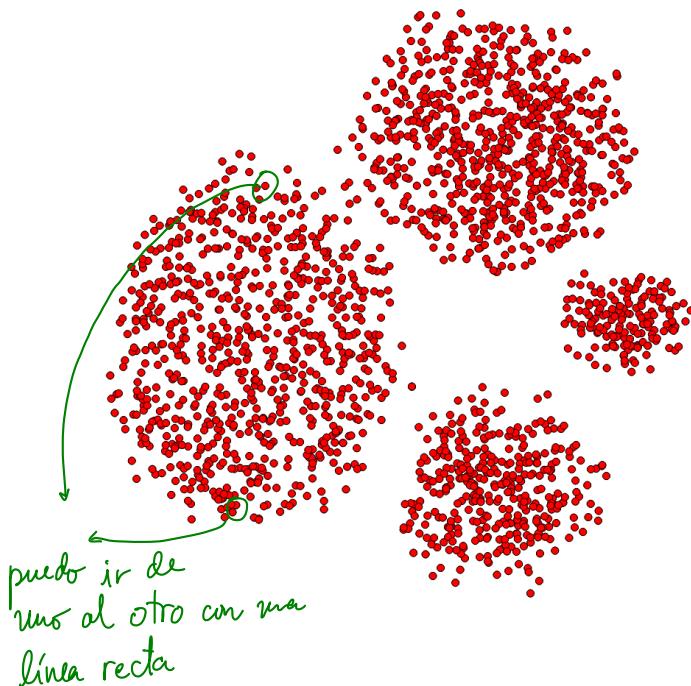
Agrupamiento

Clústeres de formas diversas

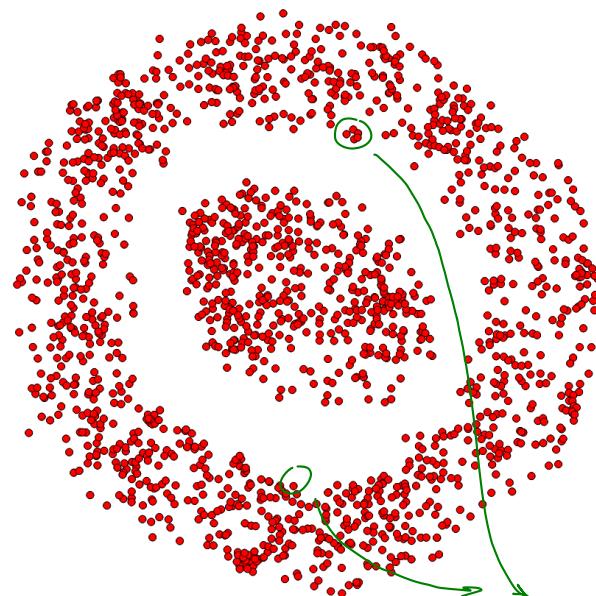


Agrupamiento

Tipos de clústeres



Grupos compactos
K-means



Convexos
Grupos conexos
Espectral

no puedo ir de este
dato al otro con una
línea recta sin salirme
del aglomerado de
punto.

Agrupamiento

Definición

Definición

Dado un conjunto de datos, el agrupamiento trata de identificar **subgrupos homogéneos** de ejemplos que manifiestan **diferencias relevantes con los otros subgrupos** que se formen.

Buscar el agrupamiento que maximiza la **dispersión interclúster**:

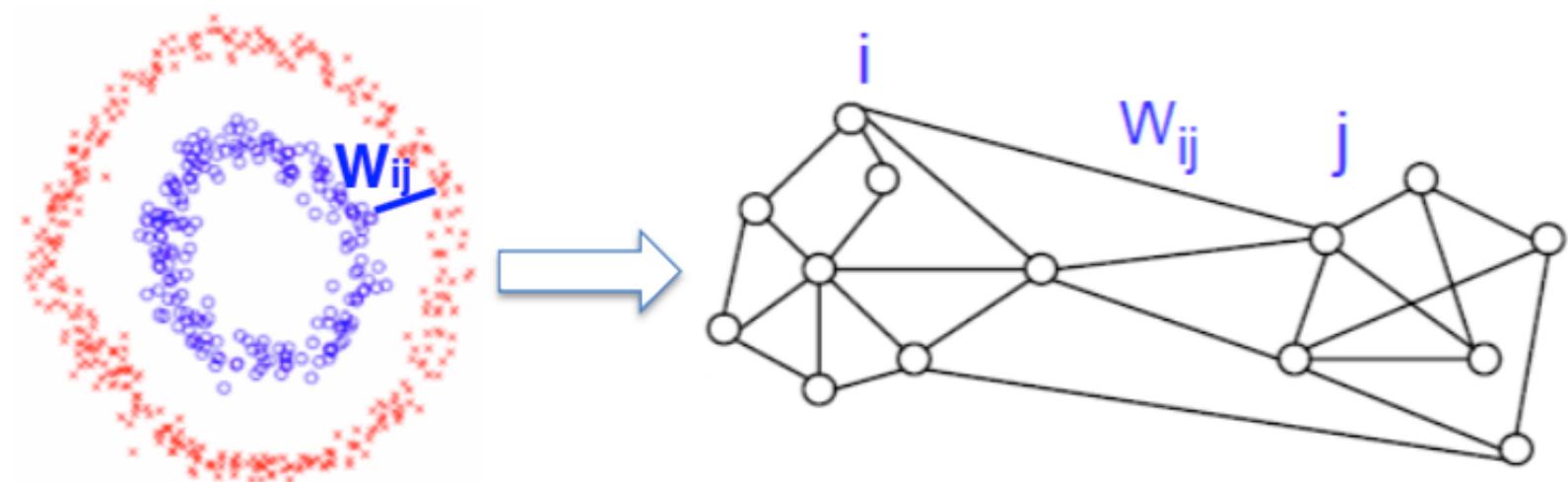
$$O(C) = \frac{1}{2} \sum_{k=1}^K \sum_{i:C(x_i)=k} \sum_{i':C(x_{i'}) \neq k} d(x_i, x_{i'})$$

y minimiza la **dispersión intraclúster**:

$$I(C) = \frac{1}{2} \sum_{k=1}^K \sum_{i:C(x_i)=k} \sum_{i':C(x_{i'})=k} d(x_i, x_{i'})$$

Agrupamiento espectral

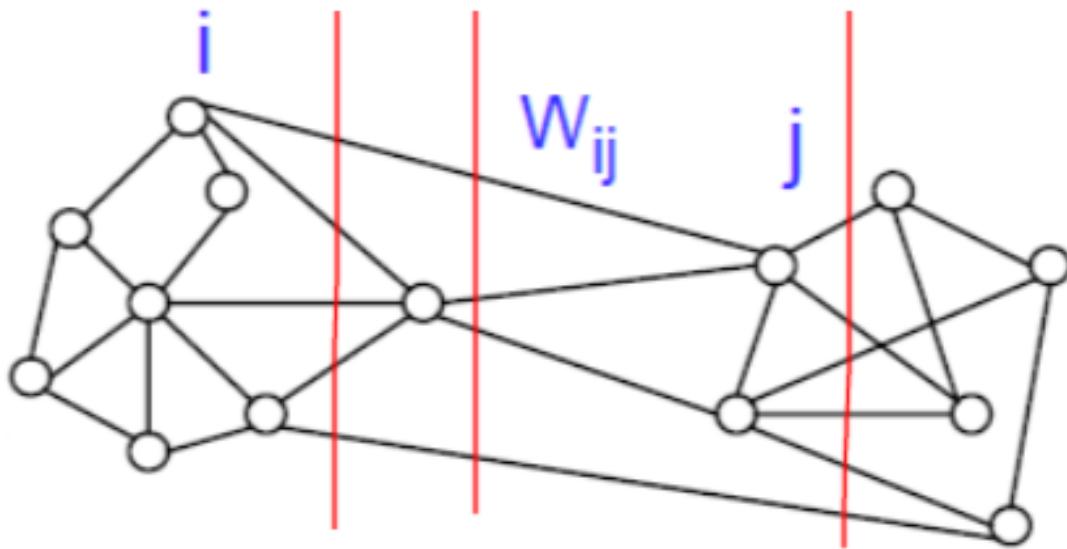
Dataset a grafo



El peso W_{ij} será mayor cuanto más parecidos sean
dos elementos *más cercanos*.

Agrupamiento espectral

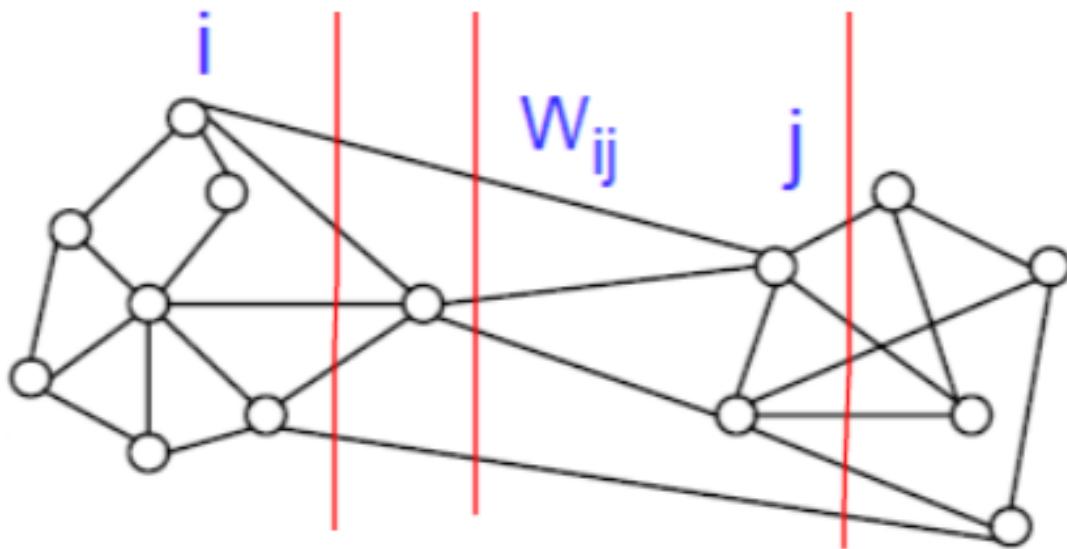
Corte mínimo de un grafo



Separar en dos el grafo de tal manera que se
eliminen el mínimo número de aristas

Agrupamiento espectral

Corte mínimo de un grafo



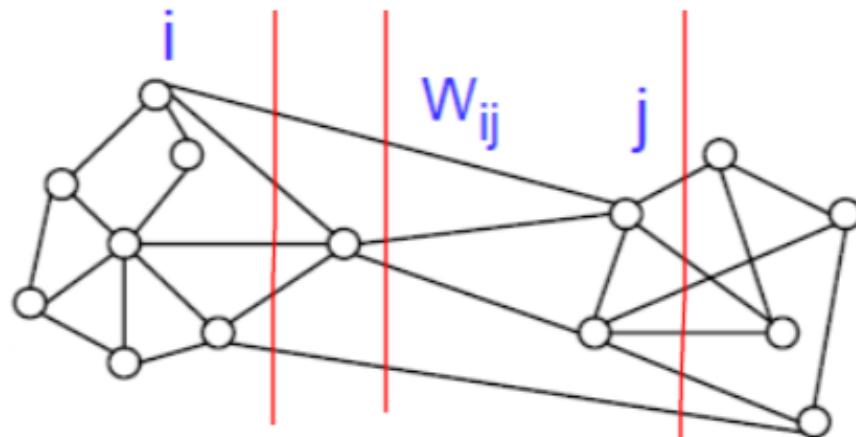
Separar en dos el grafo de tal manera que la suma de los pesos de las aristas eliminadas sea mínima

Agrupamiento espectral

Corte mínimo de un grafo

Separar en dos el grafo de tal manera que la suma de los pesos de las aristas eliminadas sea mínima

$$\arg \min_{\{A,B\}} \text{corte}(A, B) = \arg \min_{\{A,B\}} \sum_{i \in A} \sum_{j \in B} W_{ij}$$

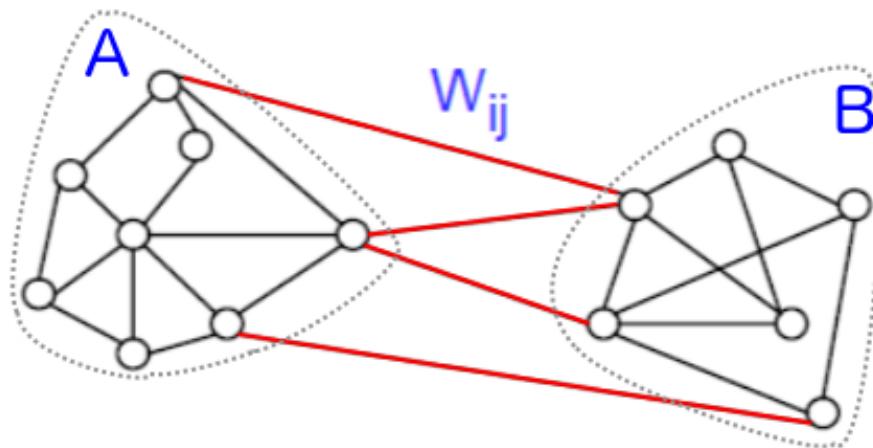


Agrupamiento espectral

Corte mínimo de un grafo

Separar en dos el grafo de tal manera que la suma de los pesos de las aristas eliminadas sea mínima

$$\arg \min_{\{A,B\}} \text{corte}(A, B) = \arg \min_{\{A,B\}} \sum_{i \in A} \sum_{j \in B} W_{ij}$$

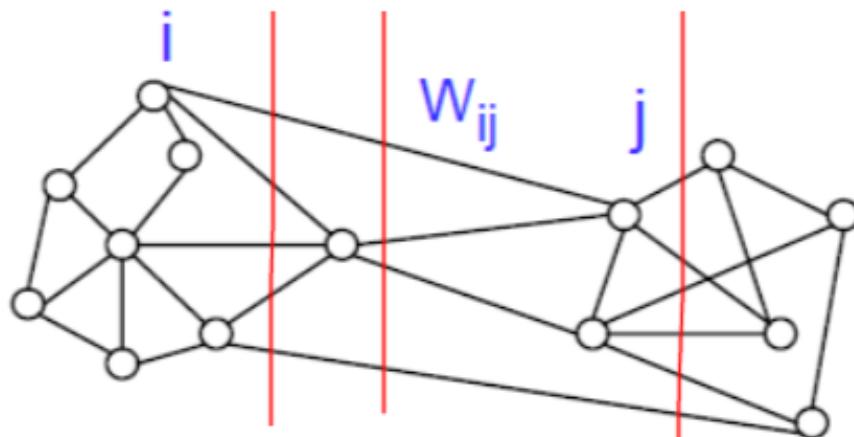


Agrupamiento espectral

Corte mínimo de un grafo

Separar en dos el grafo de tal manera que la suma de los pesos de las aristas eliminadas sea mínima

$$\arg \min_{\{A,B\}} \text{corte}(A, B) = \arg \min_{\{A,B\}} \sum_{i \in A} \sum_{j \in B} W_{ij}$$

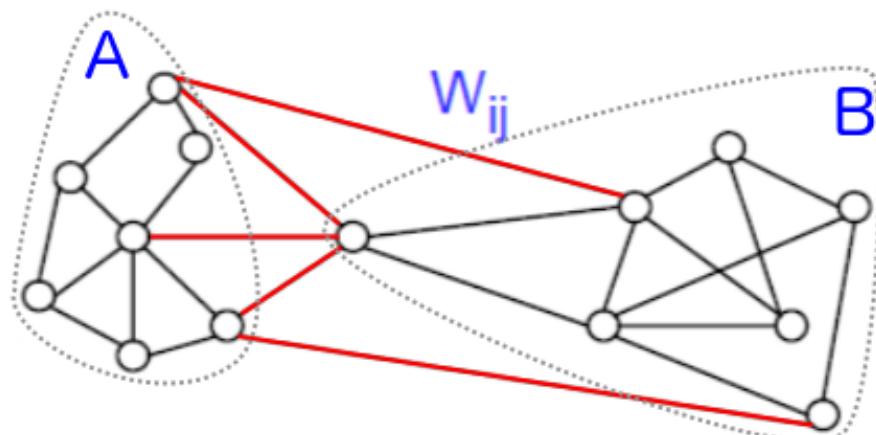


Agrupamiento espectral

Corte mínimo de un grafo

Separar en dos el grafo de tal manera que la suma de los pesos de las aristas eliminadas sea mínima

$$\arg \min_{\{A,B\}} \text{corte}(A, B) = \arg \min_{\{A,B\}} \sum_{i \in A} \sum_{j \in B} W_{ij}$$

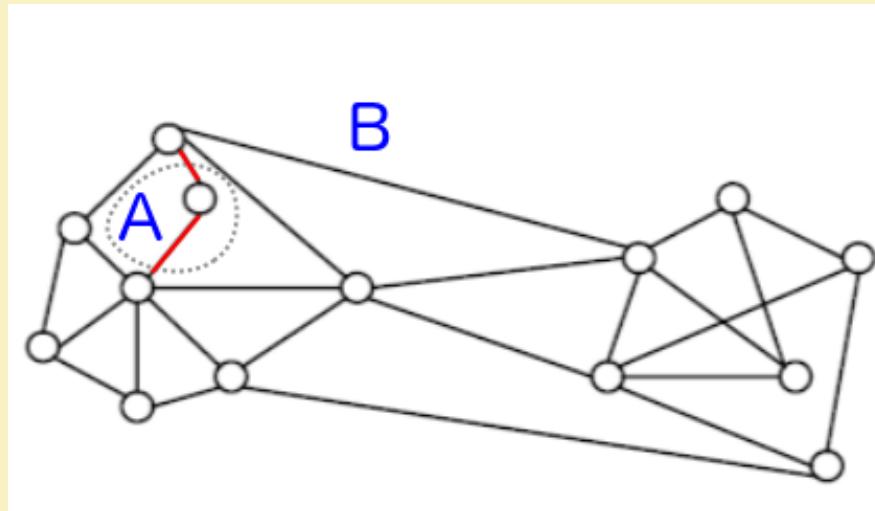


Agrupamiento espectral

Corte mínimo de un grafo

Características

- ▶ Fácil de obtener en tiempo razonable
- ▶ Cortes mínimos, no esperados



Agrupamiento espectral

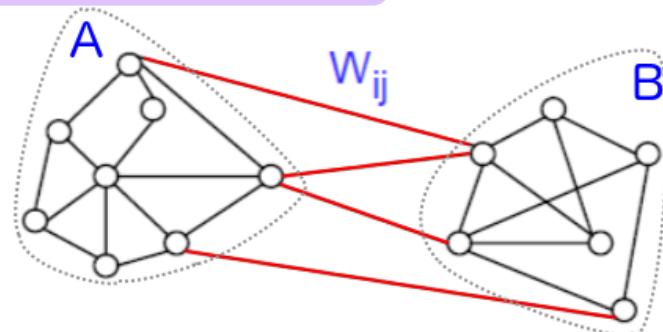
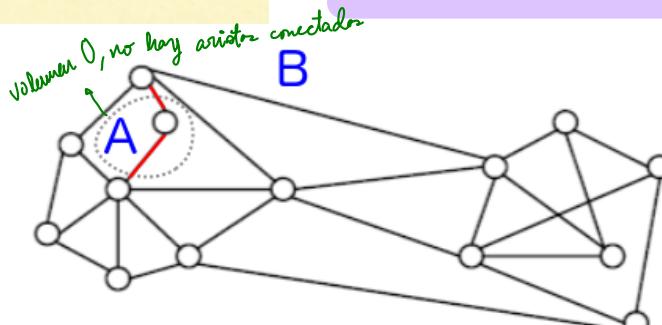
Corte **normalizado** mínimo de un grafo

Separar en dos el grafo de tal manera que la suma de los pesos de las aristas eliminadas sea mínima y los grupos resultantes sean de tamaño similar

Se hará el corte normalizado mínimo del grafo

$$\arg \min_{\{A,B\}} \left(\frac{1}{vol(A)} + \frac{1}{vol(B)} \right) corte(A, B)$$

$$vol(A) = \sum_{i \in A} grado_i = \sum_{i \in A} \sum_{j \neq i} W_{ij}$$



Agrupamiento espectral

Corte **normalizado** mínimo de un grafo

Problema:

$$\arg \min_{\{A,B\}} \left(\frac{1}{\sum_{i \in A} \sum_{j \neq i} W_{ij}} + \frac{1}{\sum_{i \in B} \sum_{j \neq i} W_{ij}} \right) \sum_{i \in A} \sum_{i \in B} W_{ij}$$

Características

- ▶ Los casos aislados no serán detectados como cortes mínimos
- ▶ No se puede resolver en un tiempo razonable

El agrupamiento espectral aproxima esta optimización mediante una transformación de los datos a partir de la matriz de adyacencias del grafo

Agrupamiento espectral

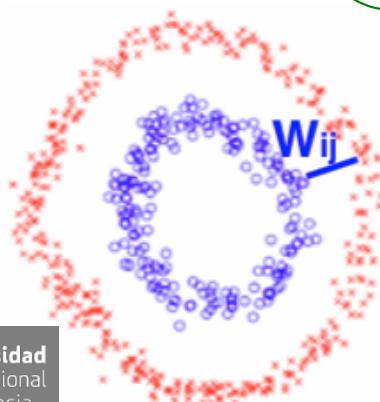
Interpretación de **camino aleatorio** en un grafo

Idea

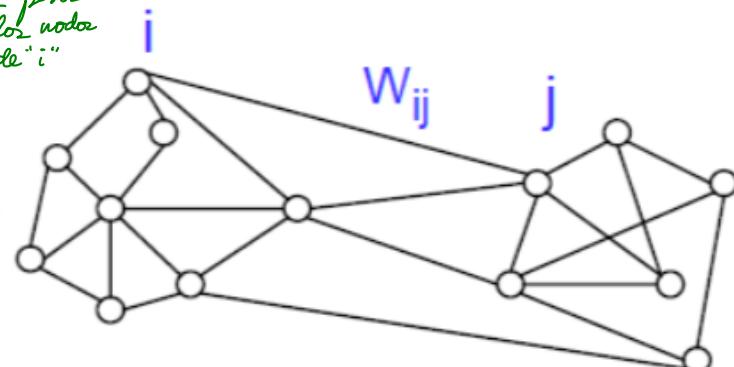
Probabilidad de alcanzar un nodo del grafo transitando por él de manera aleatoria.

En cada momento (nodo i), se selecciona el siguiente nodo j de manera aleatoria según el **peso** de las aristas de i

$$P_{ij} = \begin{cases} \frac{W_{ij}}{\sum_{j' \neq i} W_{ij'}} & , \text{ si el nodo } j' \text{ está conectado con el } i \\ 0 & , \text{ si no} \end{cases}$$



suma de los pesos
del resto de los nodos
que salgan de "i"
excepto "j"



Agrupamiento espectral

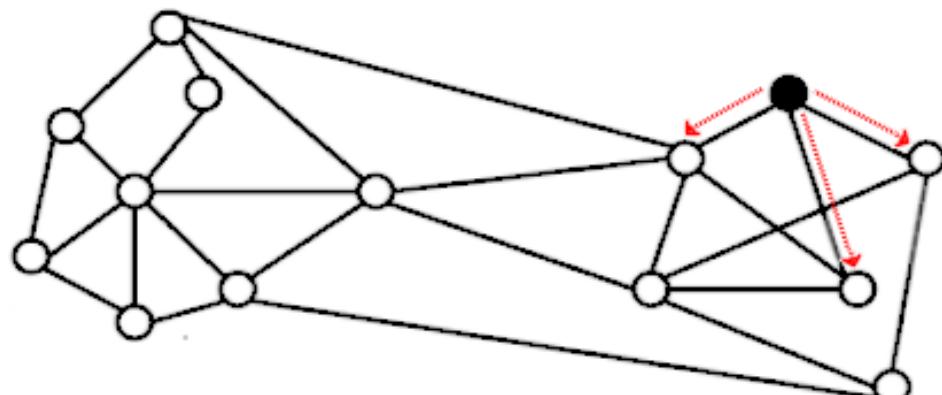
Interpretación de camino aleatorio en un grafo

Idea

Probabilidad de alcanzar un nodo del grafo transitando por él de manera aleatoria.

En cada momento (nodo i), se selecciona el siguiente nodo j de manera aleatoria según el peso de las aristas de i

$$P_{ij} = \begin{cases} \frac{W_{ij}}{\sum_{j' \neq i} W_{ij'}} & , \text{ si el nodo } j' \text{ está conectado con el } i \\ 0 & , \text{ si no} \end{cases}$$



Agrupamiento espectral

Interpretación de camino aleatorio en un grafo

De manera natural, se construye una matriz de transiciones P_{ij}

La n -ésima potencia de una matriz de transiciones, P^n , recoge en cada celda P_{ij}^n la probabilidad de llegar del nodo i al nodo j en n pasos tomados aleatoriamente

P	1	2	3	4	5	6
1	0.00	0.00	0.37	0.00	0.31	0.32
2	0.35	0.36	0.00	0.00	0.00	0.28
3	0.00	0.00	0.40	0.27	0.33	0.00
4	0.21	0.32	0.16	0.31	0.00	0.00
5	0.52	0.00	0.00	0.18	0.30	0.00
6	0.17	0.27	0.00	0.23	0.14	0.18

P^2	1	2	3	4	5	6
1	0.21	0.08	0.15	0.23	0.26	0.06
2	0.18	0.21	0.13	0.07	0.15	0.27
3	0.23	0.09	0.20	0.25	0.23	0.00
4	0.18	0.22	0.19	0.14	0.12	0.16
5	0.19	0.06	0.22	0.11	0.25	0.16
6	0.25	0.22	0.10	0.14	0.12	0.16

Agrupamiento espectral

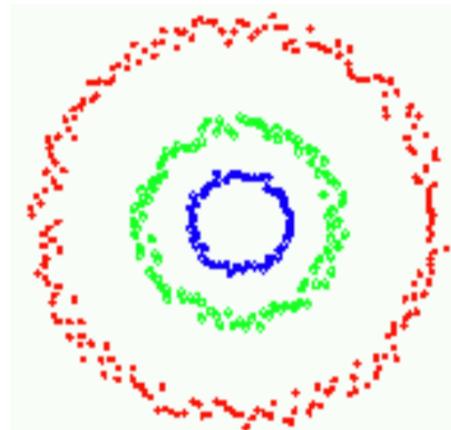
Interpretación de camino aleatorio en un grafo

Resultado

La probabilidad de transitar entre nodos de distintos clústeres, si están separados, será menor que la de transitar entre los nodos de un mismo clúster

Si los diferentes clústeres están conectados, la matriz es única

Si están desconectados, la matriz tiene una forma...



Agrupamiento espectral

Interpretación de camino aleatorio en un grafo

Resultado

La probabilidad de transitar entre nodos de distintos clústeres, si están separados, será menor que la de transitar entre los nodos de un mismo clúster

Si los diferentes clústeres están conectados, la matriz es única

$$\begin{bmatrix} L_1 & & & \\ & \ddots & & 0 \\ & & L_2 & \\ & \ddots & & \\ 0 & & & L_3 \end{bmatrix}$$

matriz laplaciana.

Aprendizaje no supervisado

VC04: Agrupamiento espectral

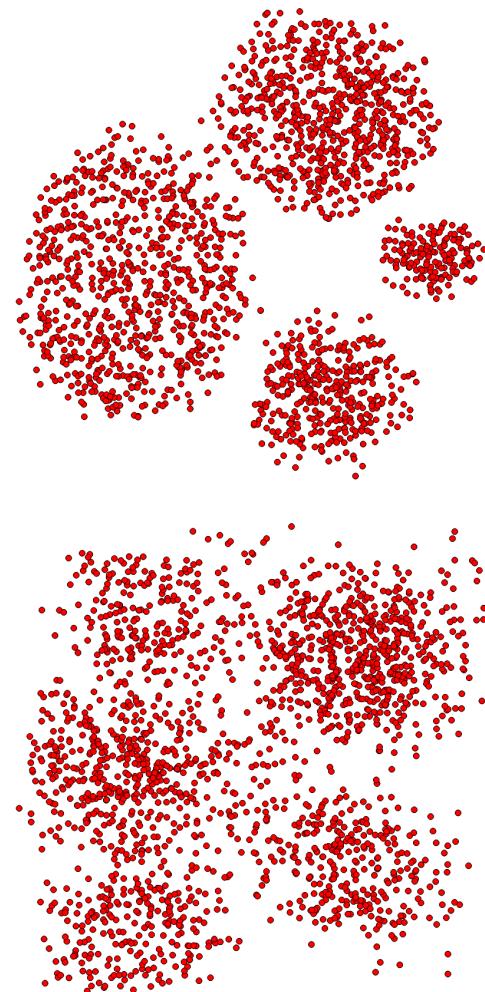
Rocío del Amor del Amor
mrocio.delamor@campusviu.es

Universidad Internacional de Valencia

Agrupamiento

Tipos de algoritmos de agrupamiento

- ▶ Basados en particiones
- ▶ Jerárquicos
- ▶ **Espectrales**
- ▶ Basados en densidad
- ▶ Probabilísticos



Agrupamiento espectral

[Examen] Puntos básicos del agrupamiento espectral
y desarrollar un poco.

Puntos básicos

1. Obtener un grafo y su matriz de adyacencias
2. Obtener una representación alternativa de los datos
3. Aplicar un algoritmo de agrupamiento estándar (K -means)

Puntos básicos

1. **Obtener un grafo y su matriz de adyacencias**
2. Obtener una representación alternativa de los datos
3. Aplicar un algoritmo de agrupamiento estándar (K -means)

Agrupamiento espectral

Consideraciones previas

Matriz de similitud

- El algoritmo no funciona con la matriz original
- Se usa la **matriz de similitud**

① Se transforma la matriz de ejemplos D ($n \times v$) en... matriz de distancias, M ($n \times n$), tal que:

simétrica

$$M_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$$

② y ésta, a su vez, en una matriz de similitudes, S ($n \times n$):

$$S_{ij} = \exp(-M_{ij}^2 / 2 \cdot \sigma^2)$$

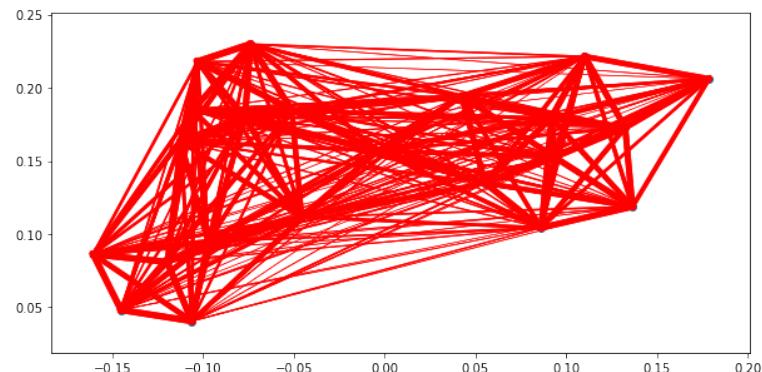
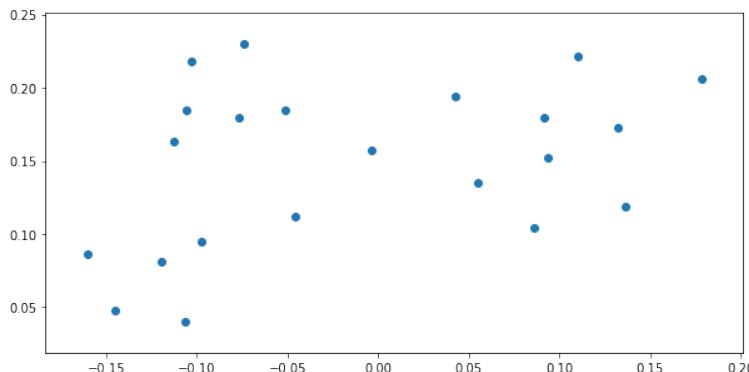
Agrupamiento espectral

Obtener un grafo y su matriz de adyacencias

Procedimientos de generación de un grafo

- ▶ Cada ejemplo de D es un nodo del grafo
- ▶ Todos los nodos están conectados con todos
- ▶ El peso de la arista entre dos nodos es su similitud, $W_{ij} = S_{ij}$

Grafo completo



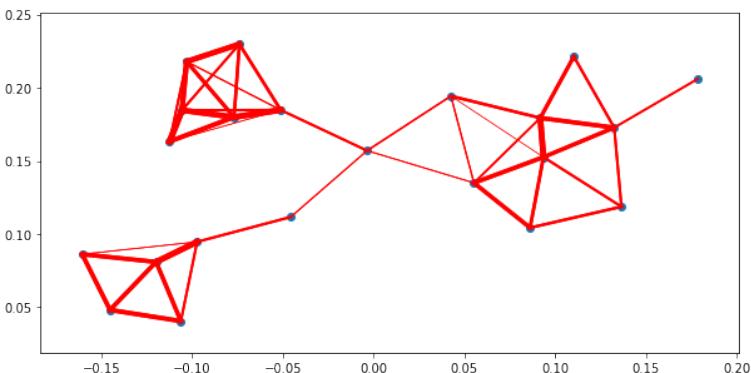
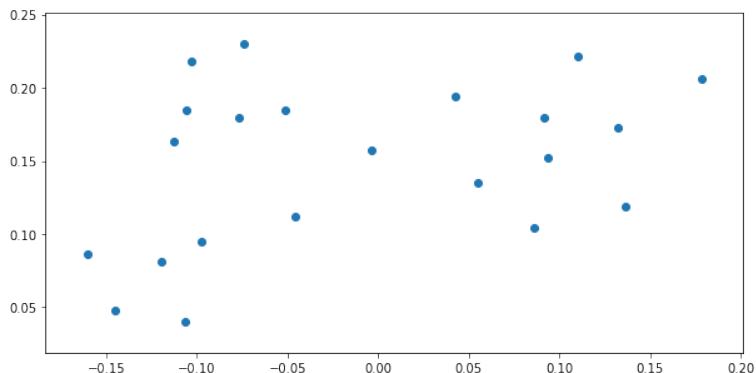
Agrupamiento espectral

Obtener un grafo y su matriz de adyacencias

Procedimientos de generación de un grafo

- ▶ Cada ejemplo de D es un nodo del grafo
- ▶ Existe una arista entre dos nodos si hay un mínimo de similitud entre ambos, $S_{ij} > \epsilon$ → hiperparámetro
- ▶ El peso de la arista entre dos nodos es su similitud, $W_{ij} = S_{ij}$

Grafo umbral



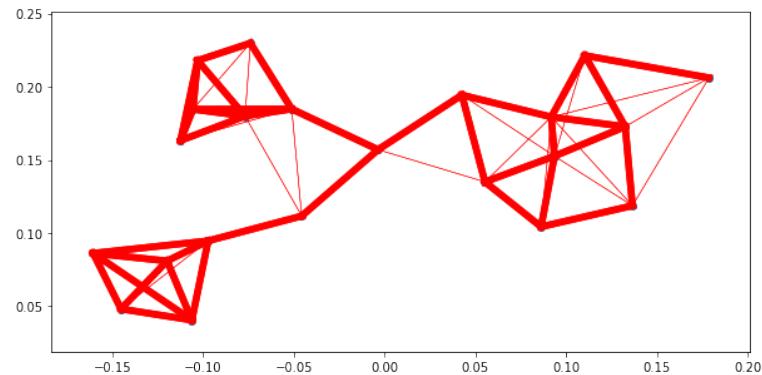
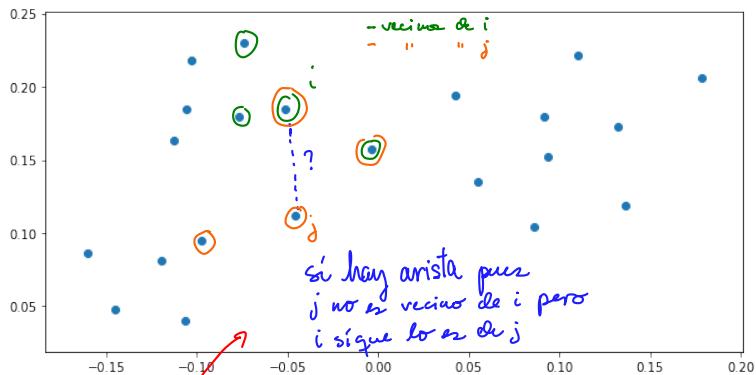
Agrupamiento espectral

Obtener un grafo y su matriz de adyacencias

Procedimientos de generación de un grafo

- ▶ Cada ejemplo de D es un nodo del grafo
- ▶ Un nodo i tiene una arista con otro nodo j si j es uno de los K 'vecinos' más similares de i (o viceversa)
- ▶ El peso de la arista entre dos nodos es su similitud, $W_{ij} = S_{ij}$

Grafo KNN



[Examen]

Close 28/09/23

min 1:45:46

Agrupamiento

[Examen]



Clase 03/10/2023

Puntos b)

Diferencias entre matriz de similitud y matriz de adyacencia
Para construir el grafo

- Construyo la matriz de similitud $\begin{bmatrix} 0 & s_{12} & \dots & s_{1n} \\ s_{21} & 0 & \dots & \vdots \\ \vdots & \vdots & \ddots & 0 \end{bmatrix}$ simétrica
- De esta matriz me construyo el grafo (por ejemplo, con umbral T) \Rightarrow hay nodos que no se conectan si al usar un umbral o KNN
- Una vez que se tiene el grafo se construye la matriz de adyacencia, es decir, será similar a la de similitud pero tendrá ceros donde los nodos no están conectados

1. Obtener un grafo y su matriz de adyacencias
2. **Obtener una representación alternativa de los datos**
3. Aplicar un algoritmo de agrupamiento estándar (K -means)

Agrupamiento espectral

Obtener una representación alternativa de los datos

Procedimientos de la matriz Laplaciana

$$L = D - W$$

- W : matriz representativa del grafo (paso anterior)
- D : matriz diagonal con $D_{ii} = \sum_j W_{ij}$

(matriz de grados)

Matriz Laplaciana básica

Cuentas
conexiones tiene
con otros nodos?

$$D_{11} = d_1 = 0 + \underbrace{w_{12} + w_{13} + \dots + w_{1n}}_{\text{algunas pueden ser } 0}$$

porque no están conectados
 $\Rightarrow w_{11} = 0$

$$\begin{bmatrix} L_1 & & & \\ & \ddots & & 0 \\ & & L_2 & \\ & & & \ddots \\ 0 & & & & L_3 \end{bmatrix}$$

Agrupamiento espectral

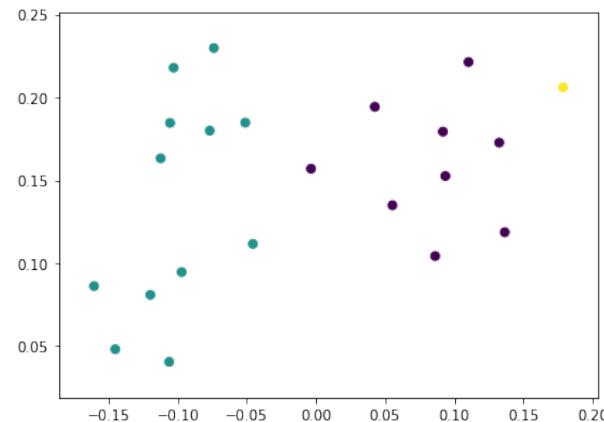
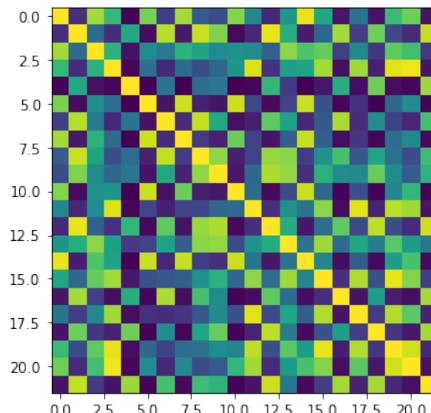
Obtener una representación alternativa de los datos

Procedimientos de la matriz Laplaciana

$$L = D - W$$

- ▶ W : matriz representativa del grafo (paso anterior)
- ▶ D : matriz diagonal con $D_{ii} = \sum_j W_{ij}$

Matriz Laplaciana básica



Agrupamiento espectral

Obtener una representación alternativa de los datos

Procedimientos de la matriz Laplaciana

$$L = I - D^{-1}W$$

- ▶ W : matriz representativa del grafo (paso anterior)
- ▶ D : matriz diagonal con $D_{ii} = \sum_j W_{ij}$
- ▶ I : matriz identidad

Matriz Laplaciana normalizada (RW)

$$\begin{bmatrix} L_1 & & & \\ & \ddots & & 0 \\ & & L_2 & \\ \ddots & & & \ddots \\ 0 & & & L_3 \end{bmatrix}$$

Agrupamiento espectral

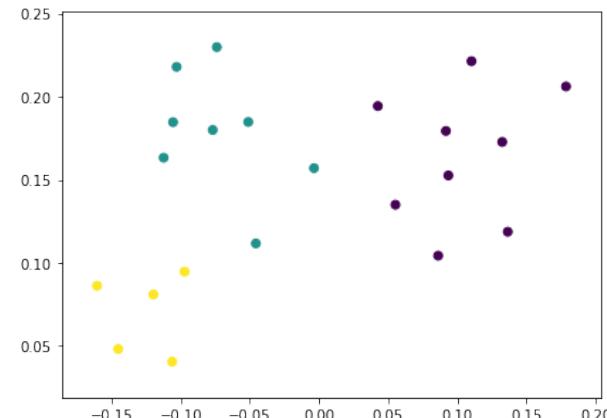
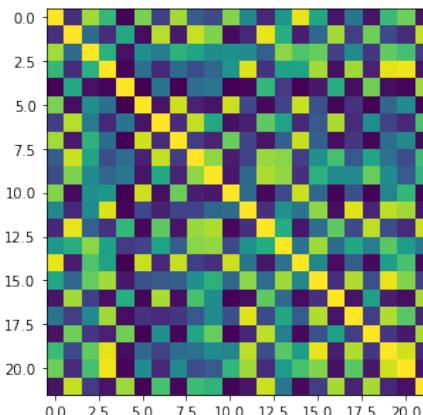
Obtener una representación alternativa de los datos

Procedimientos de la matriz Laplaciana

$$L = I - D^{-1}W$$

- ▶ W : matriz representativa del grafo (paso anterior)
- ▶ D : matriz diagonal con $D_{ii} = \sum_j W_{ij}$
- ▶ I : matriz identidad

Matriz Laplaciana normalizada (RW)



Agrupamiento espectral

Obtener una representación alternativa de los datos

Procedimientos de la matriz Laplaciana

$$L = I - D^{-1/2}WD^{-1/2}$$

- ▶ W : matriz representativa del grafo (paso anterior)
- ▶ D : matriz diagonal con $D_{ii} = \sum_j W_{ij}$
- ▶ I : matriz identidad

Matriz Laplaciana normalizada simétrica

$$\begin{bmatrix} L_1 & & & \\ & \ddots & & 0 \\ & & L_2 & \\ & \ddots & & L_3 \\ 0 & & & \ddots \end{bmatrix}$$

Agrupamiento espectral

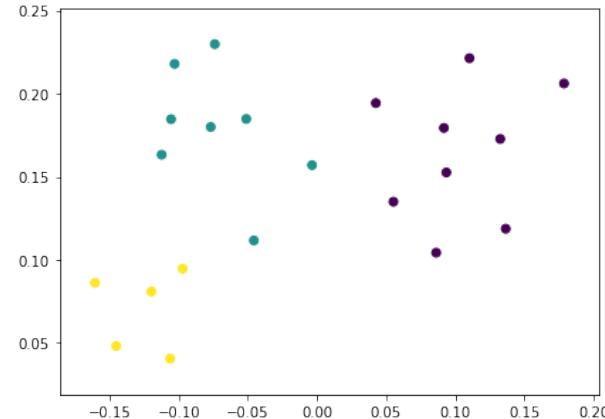
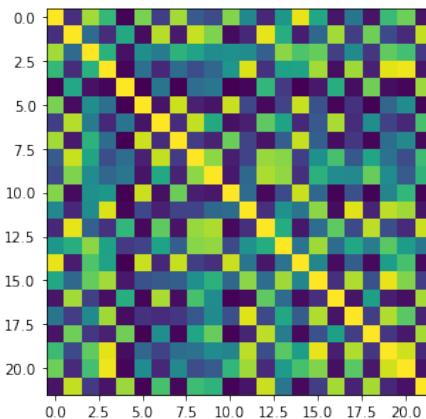
Obtener una representación alternativa de los datos

Procedimientos de la matriz Laplaciana

$$L = I - D^{-1/2}WD^{-1/2}$$

- ▶ W : matriz representativa del grafo (paso anterior)
- ▶ D : matriz diagonal con $D_{ii} = \sum_j W_{ij}$
- ▶ I : matriz identidad

Matriz Laplaciana normalizada simétrica



Agrupamiento espectral

Obtener una representación alternativa de los datos

2º Hiperparámetro: Forma de construir la matriz Laplaciana

Obtener datos transformados

1. Descomponer la matriz L en vectores propios
2. Ordenar los vectores propios según el valor propio correspondiente (ascendente)
3. Seleccionar los primeros K vectores propios

número de componentes conexas \Rightarrow número de clusters separables



cuyos valores propios son 0 o cercanos a 0

Cada vector propio es una “variable” en el dataset transformado, que tiene n filas (tantas como ejemplos) y K variables (vectores propios):

$$D(n \times v) \rightarrow M(n \times K)$$

ejemplos
variables

Dataset

(500, 700)

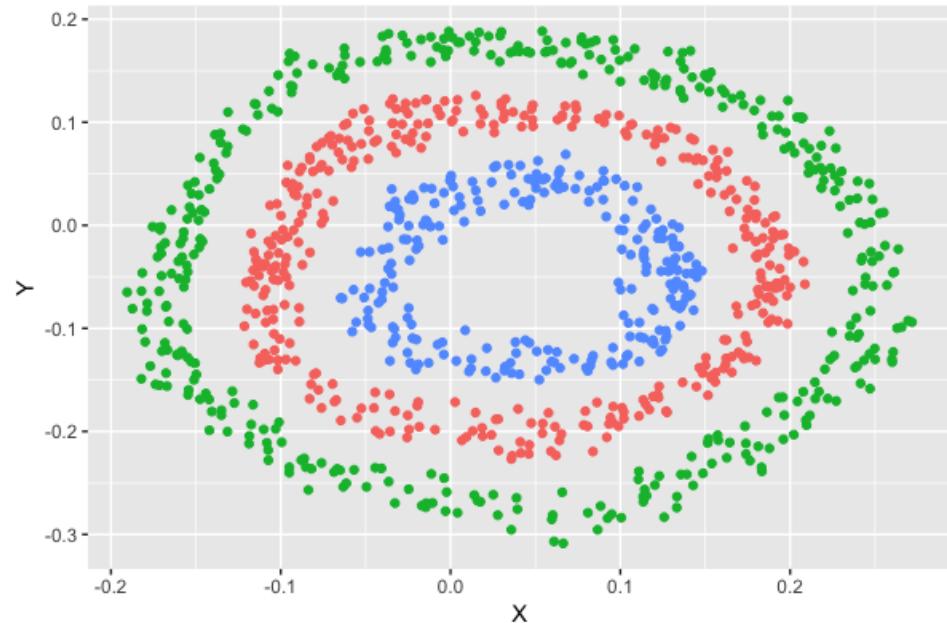
→ nuevas dimensiones (500, k)

las nuevas características del dataset van a ser los k vectores propios

Reducción de dimensionalidad

Agrupamiento espectral

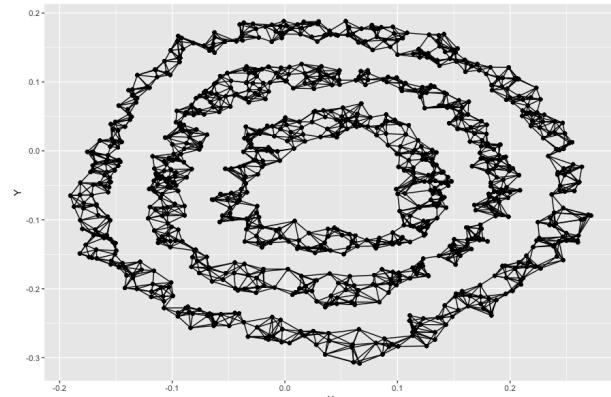
Efecto de la K en KNN (generación del grafo)



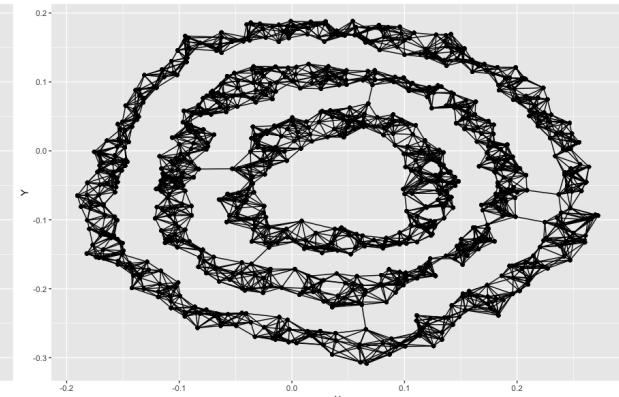
Agrupamiento espectral

Efecto de la K en KNN (generación del grafo)

$k=2$



$k=4$

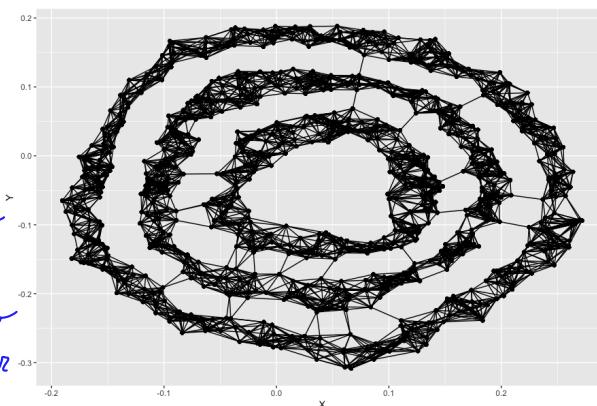


[Examen]

¿Cómo influye K en el número de aristas que se forman en los grafos?

Cuanto menor sea K , menor número de aristas se crean en los grafos.

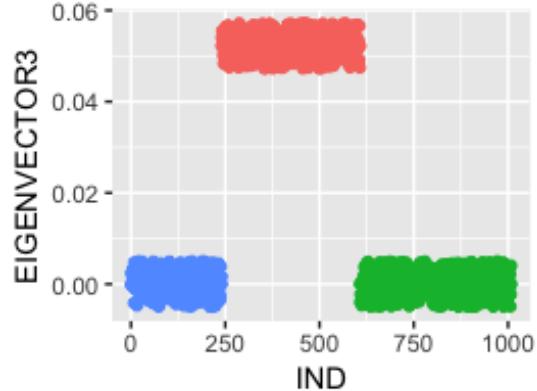
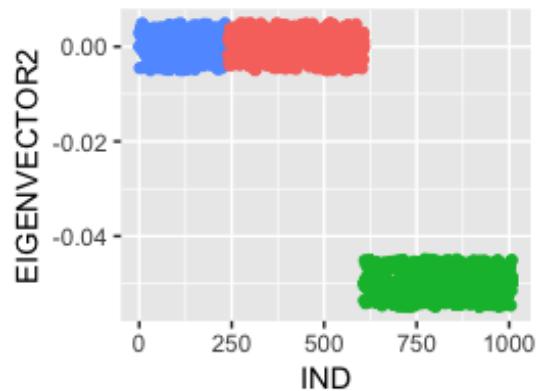
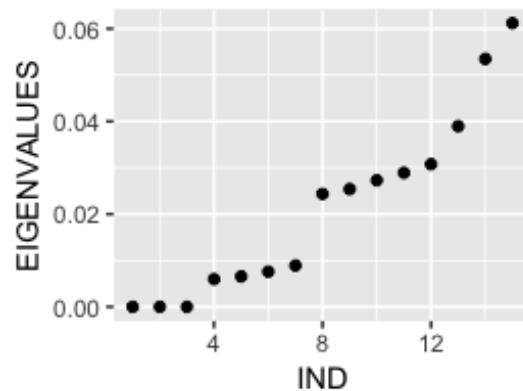
Cuanto mayor sea K más complicado es encontrar un espacio donde separar los grafos.



$k=8$

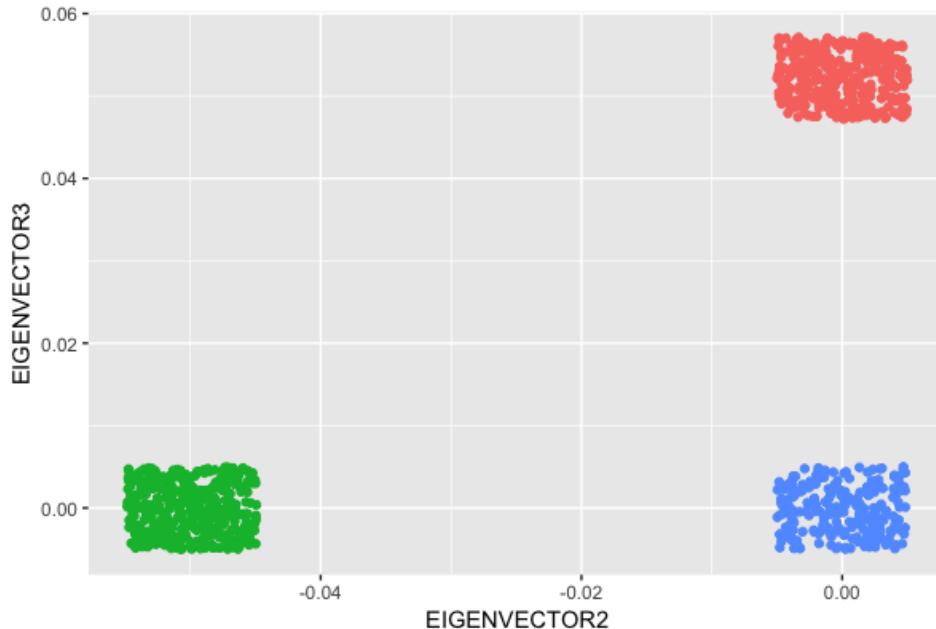
Agrupamiento espectral

Efecto de la K en KNN (generación del grafo)



Agrupamiento espectral

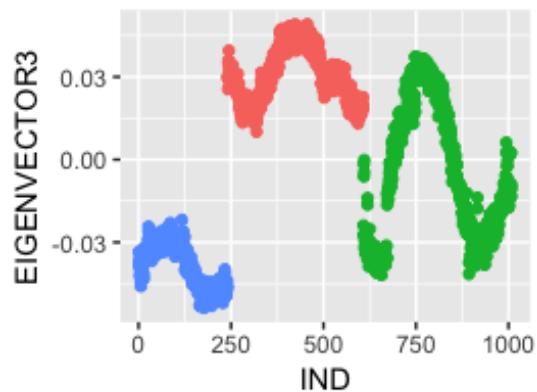
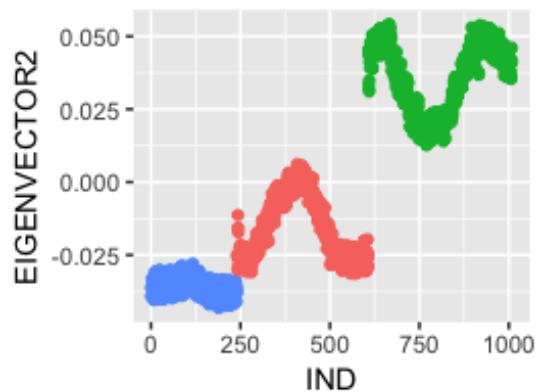
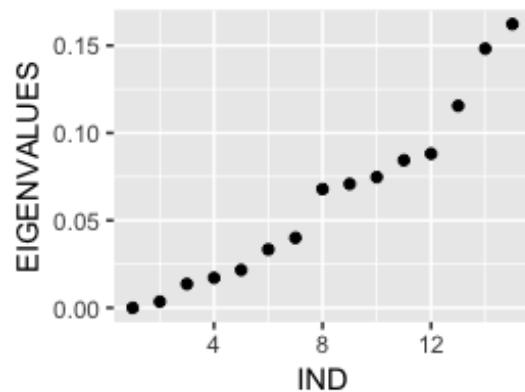
Efecto de la K en KNN (generación del grafo)



Agrupamiento espectral

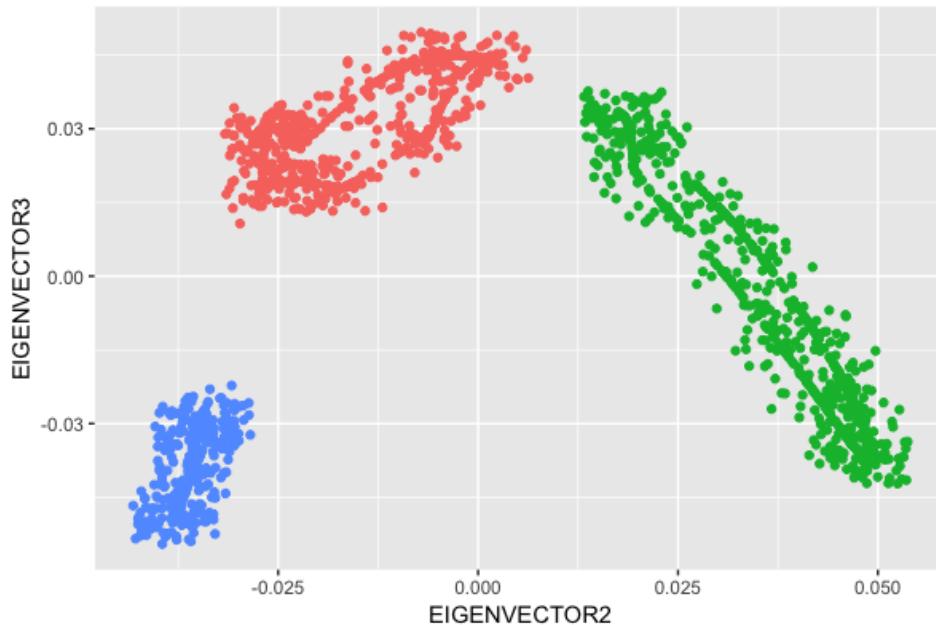
Efecto de la K en KNN (generación del grafo)

$$\underline{\underline{k=4}}$$



Agrupamiento espectral

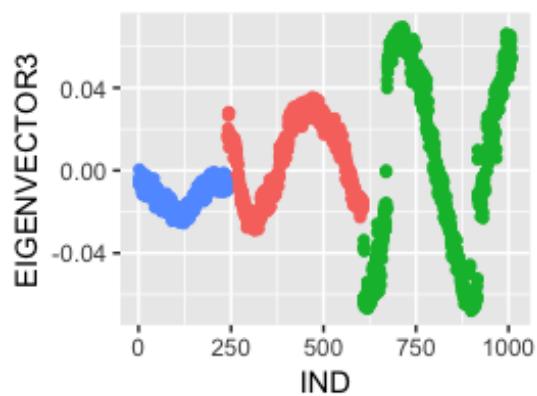
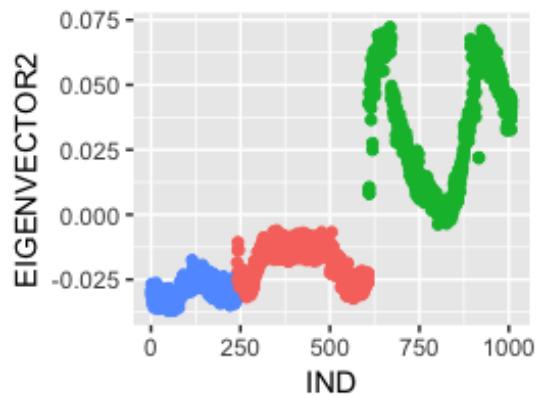
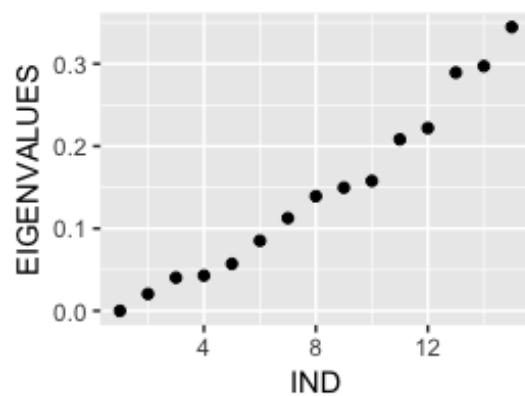
Efecto de la K en KNN (generación del grafo)



Agrupamiento espectral

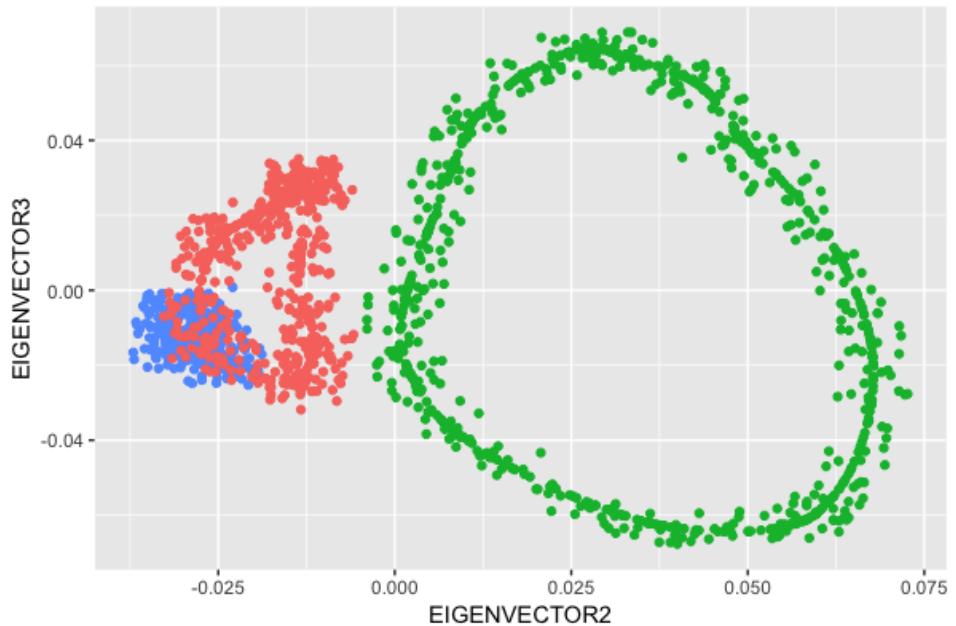
Efecto de la K en KNN (generación del grafo)

$$\underline{k = 8}$$



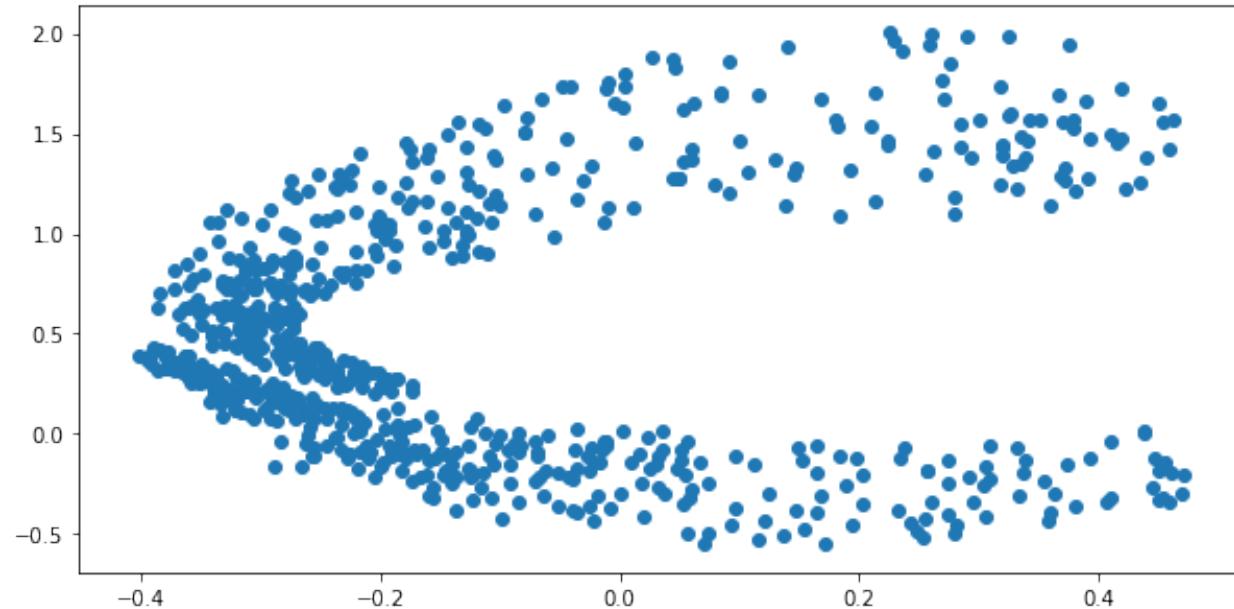
Agrupamiento espectral

Efecto de la K en KNN (generación del grafo)



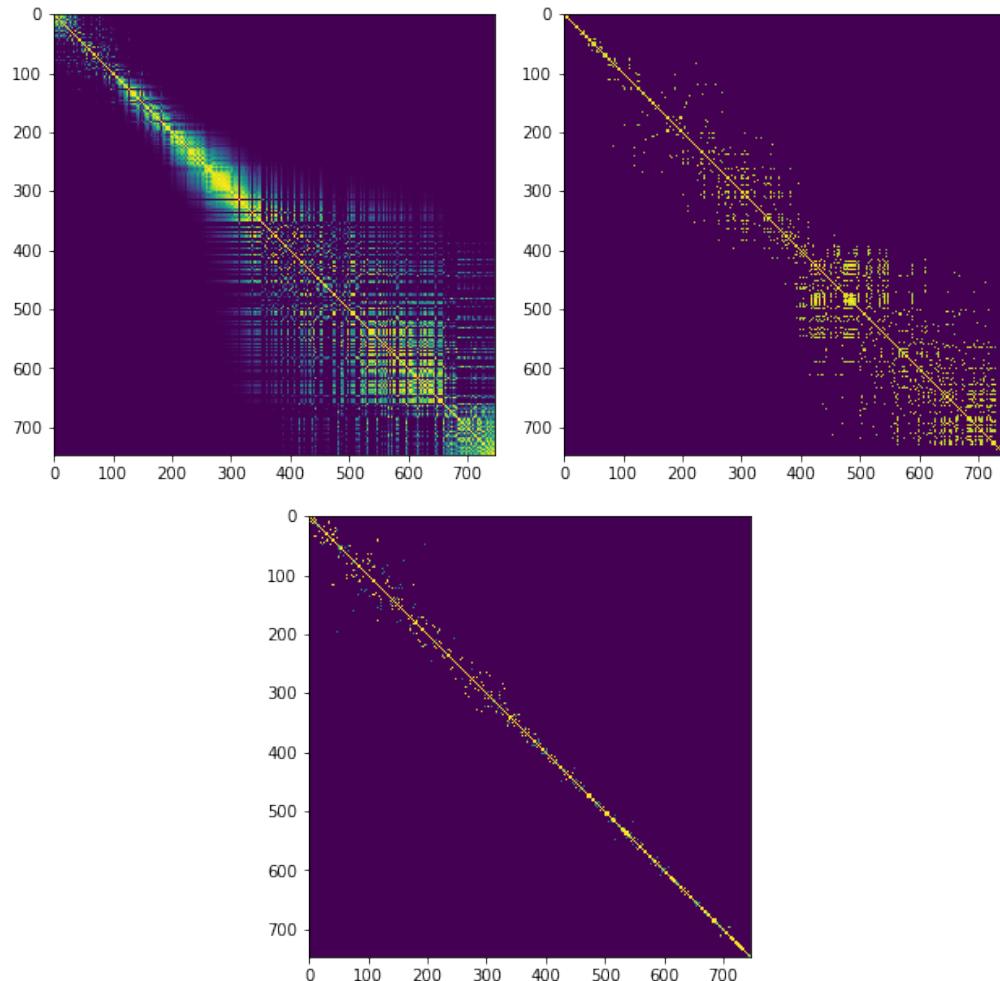
Agrupamiento espectral

Diferentes generaciones del grafo



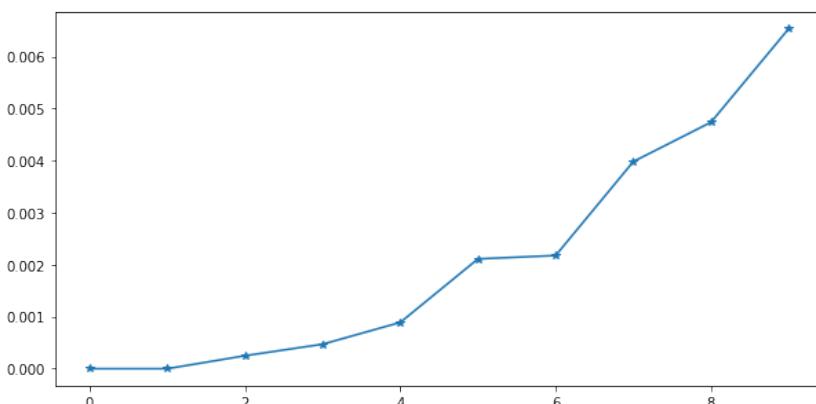
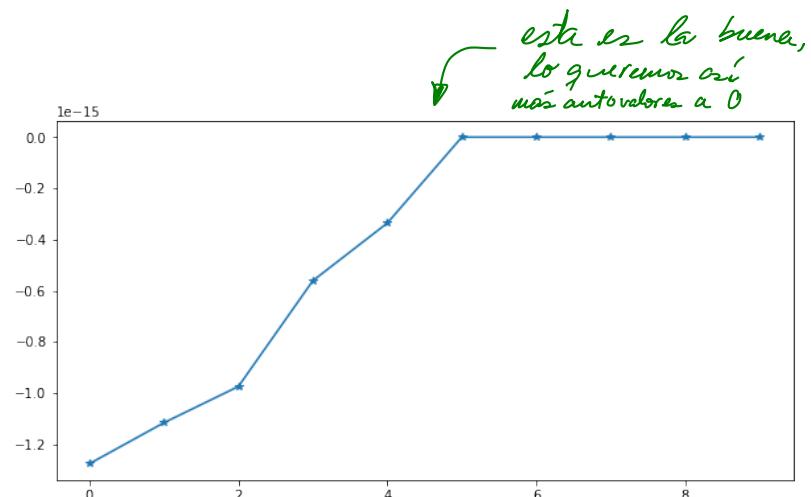
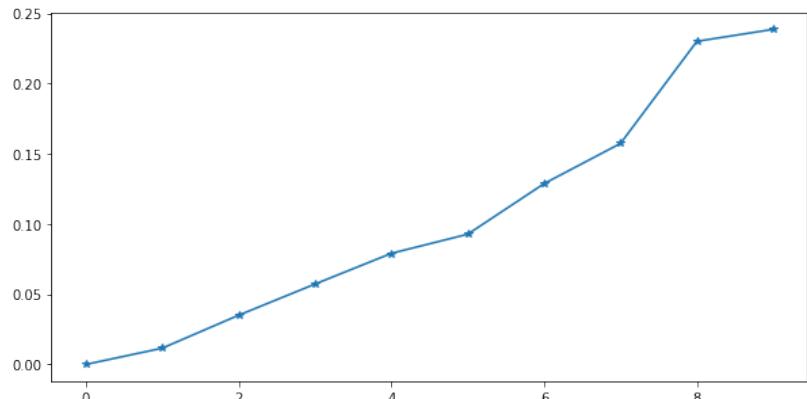
Agrupamiento espectral

Diferentes generaciones del grafo



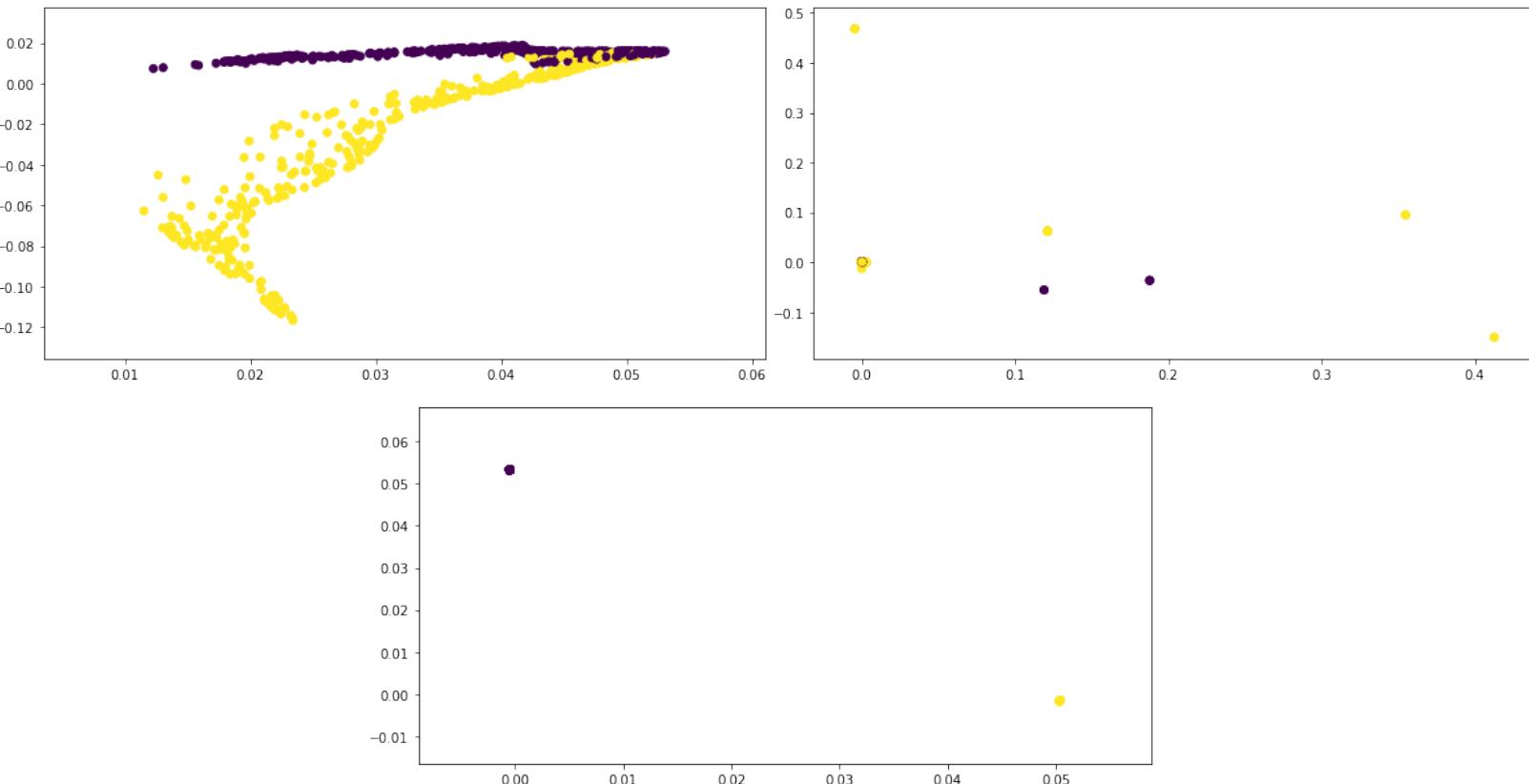
Agrupamiento espectral

Diferentes generaciones del grafo



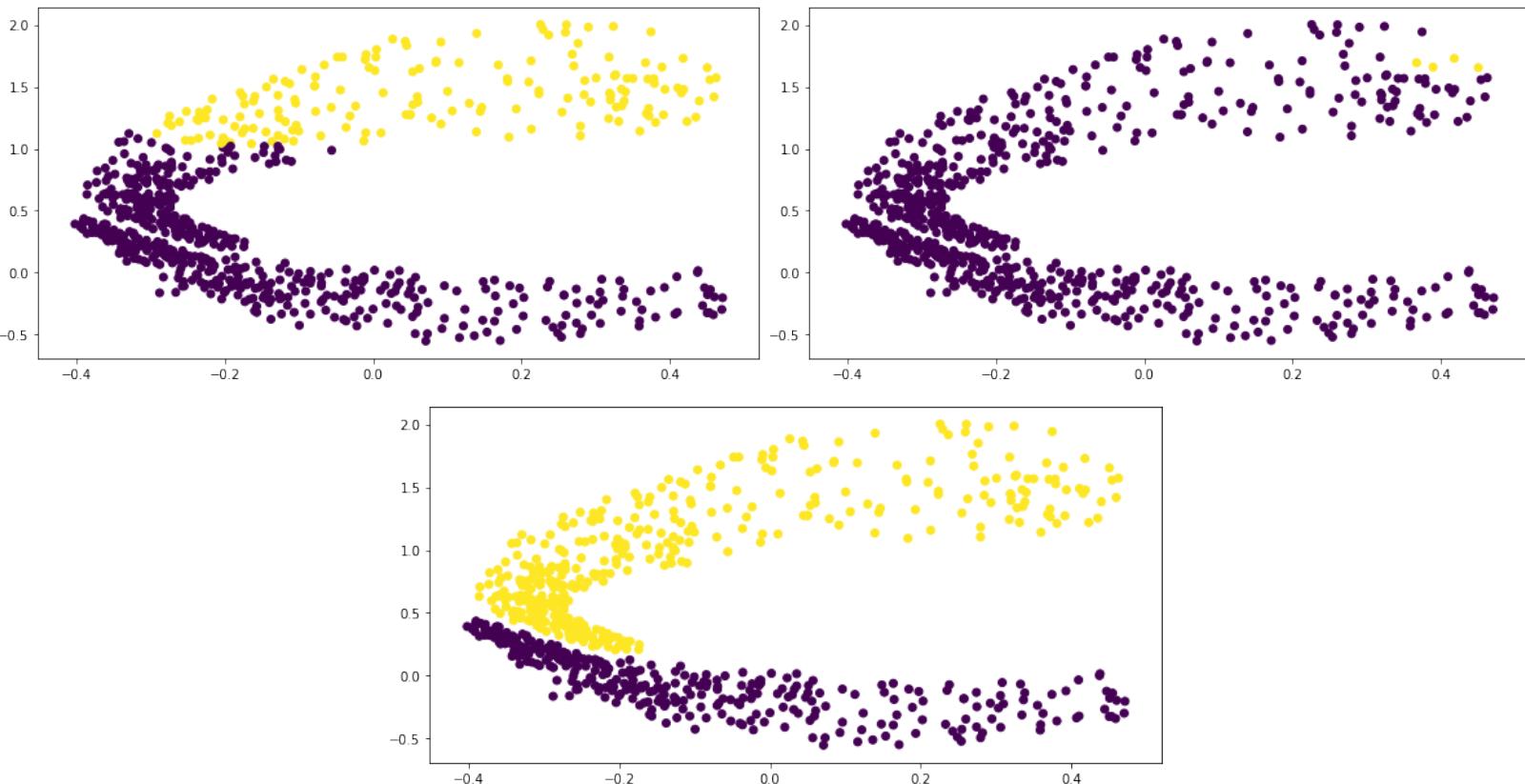
Agrupamiento espectral

Diferentes generaciones del grafo



Agrupamiento espectral

Diferentes generaciones del grafo



Agrupamiento espectral

Puntos básicos

1. Obtener un grafo y su matriz de adyacencias
2. Obtener una representación alternativa de los datos
3. **Aplicar un algoritmo de agrupamiento estándar (K -means)**

Agrupamiento espectral

Ventajas

- ▶ Sólida base matemática
- ▶ Funciona con clústeres de diversa forma
- ▶ Diferentes criterios y maneras de crear el grafo de similitudes
- ▶ Puede funcionar con diferentes medidas de distancia
- ▶ Se pueden usar diferentes algoritmos sobre la matriz transformada

Agrupamiento espectral

Problemas

Obtener datos transformados

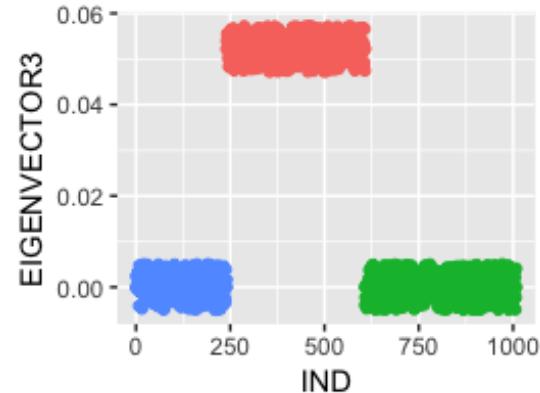
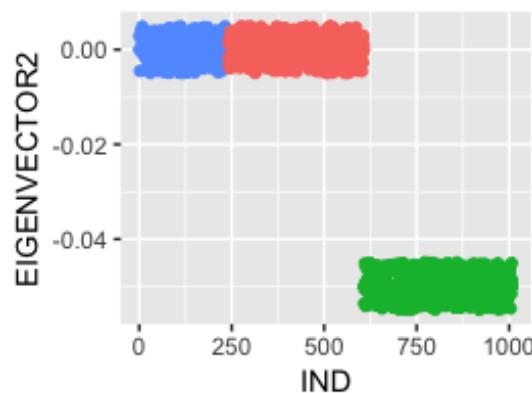
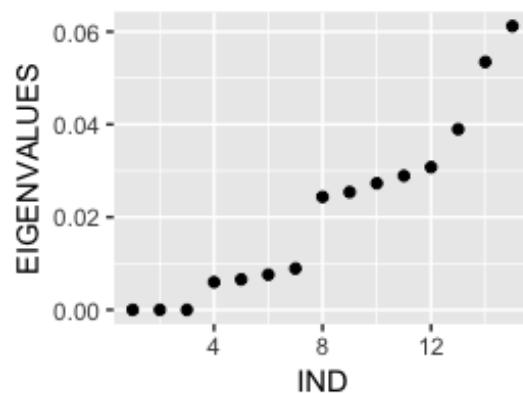
- ▶ Elegir el tipo de grafo (y K en KNN, o ϵ en umbral)
- ▶ Elegir el tipo de matriz Laplaciana
- ▶ Elegir el número de vectores propios

Agrupamiento espectral

Problemas

Obtener datos transformados

- ▶ Elegir el tipo de grafo (y K en KNN, o ϵ en umbral)
- ▶ Elegir el tipo de matriz Laplaciana
- ▶ Elegir el número de vectores propios
 - ▶ En la práctica, el número de clústeres
 - ▶ Salto máximo entre dos valores propios consecutivos



Aprendizaje no supervisado

VC05: Agrupamiento basado en densidad - DBSCAN

Rocío del Amor del Amor

mrocio.delamor@campusviu.es

Universidad Internacional de Valencia

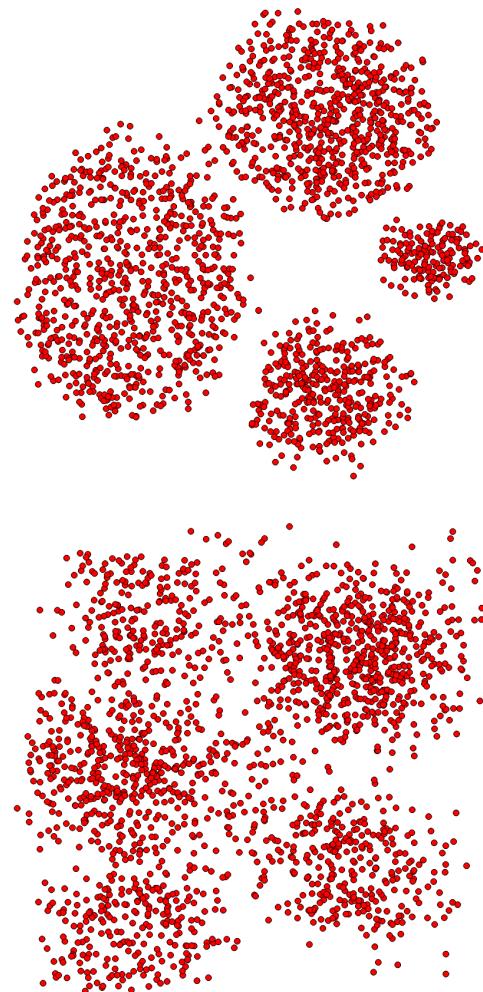
Agrupamiento

[Examen]

Tipos de algoritmos de agrupamiento

¿ Se necesita saber el número de clusters a priori ?

- ▶ Basados en particiones SI
- ▶ Jerárquicos NO
- ▶ Espectrales NO
- ▶ Basados en densidad NO
- ▶ Probabilísticos ?



Agrupamiento basado en densidad

Conceptos

Densidad

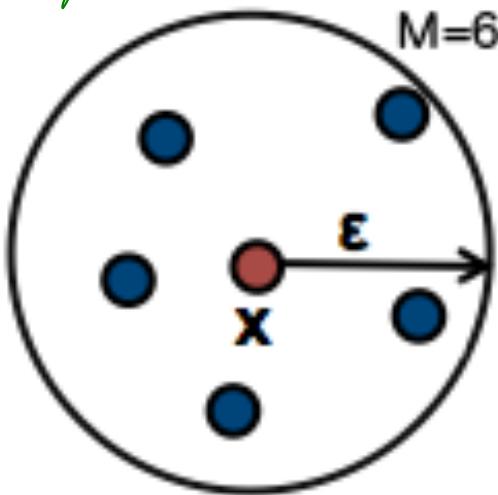
¿Cómo definir densidad?

- ▶ Vecindario (parámetro ϵ)
- ▶ Punto nuclear (parámetros ϵ , M)

distancia máxima que define el vecindario de un punto

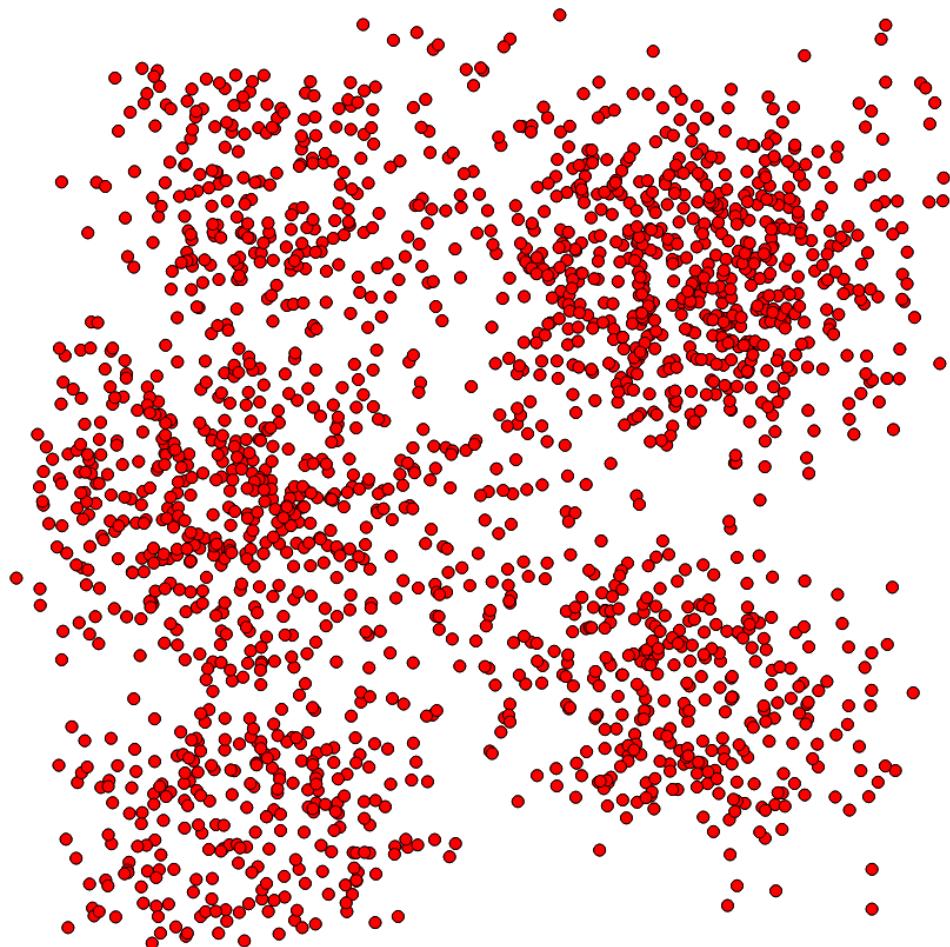
número de puntos contenido en un vecindario dado $\geq M$

Vecindario de X



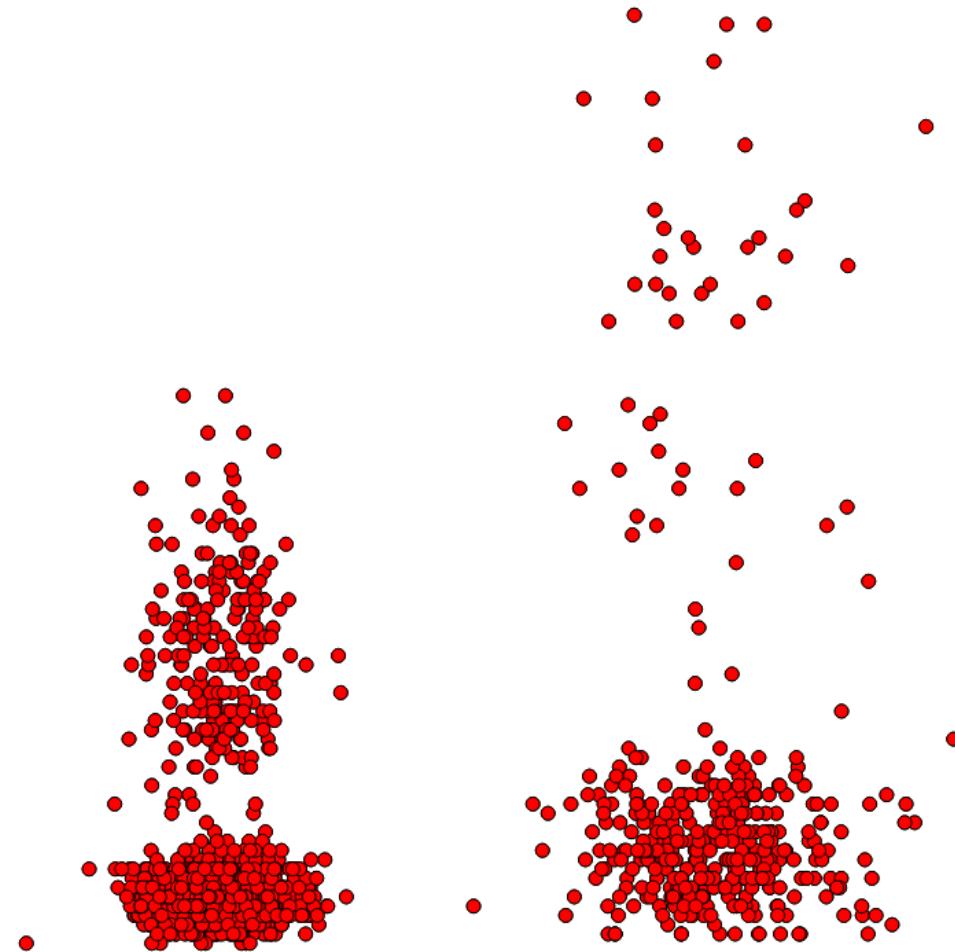
Agrupamiento basado en densidad

Conceptos



Agrupamiento basado en densidad

Conceptos



Agrupamiento basado en densidad

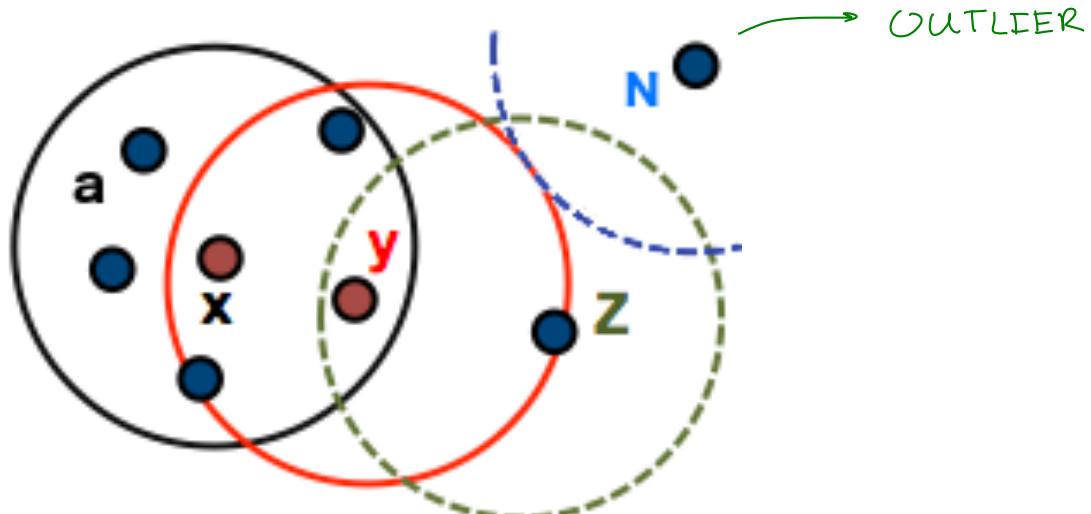
DBSCAN

Densidad

En base al valor de los parámetros ϵ y M , se definen:

- ▶ Punto nuclear (x o y)
 - { • comparten vecindario
• x tiene que ser nuclear }
- ▶ Punto directamente denso-alcanzable (y desde x , z desde y)
- ▶ Punto borde (z) → punto dentro del vecindario de un punto nuclear pero que no es punto nuclear
- ▶ Punto ruido (n) → no está en el vecindario de ningún punto nuclear

Buen algoritmo
para detectar
outliers



Agrupamiento basado en densidad

DBSCAN

Densidad

En base al valor de los parámetros ϵ y M , se definen:

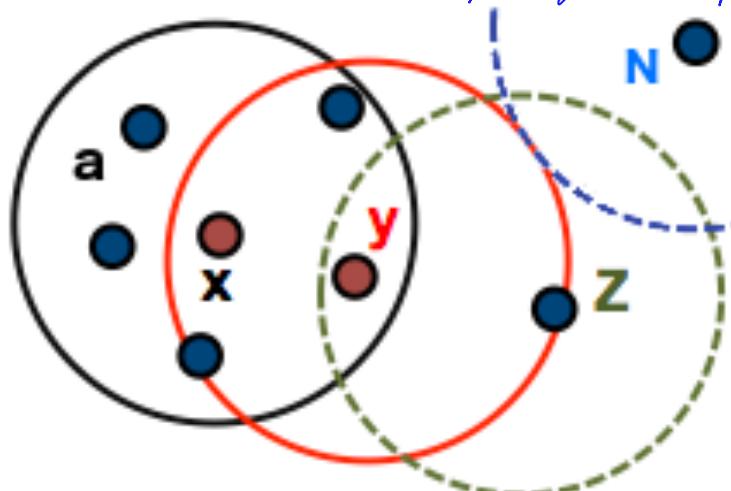
- ▶ Punto **directamente denso-alcanzable** (y desde x , z desde y)
- ▶ Punto **denso-alcanzable** (z desde x)
 - ** Relación **asimétrica** **
- ▶ Puntos **denso-conectados** (a y z)
 - ** Relación **simétrica** **

en el camino los saltos se hacen entre puntos que comparten vecindario

↳ { · no están en el mismo vecindario
· puedo encontrar un camino de puntos nucleares que me lleve de $x \rightarrow z$
· x es punto nuclear

↳ { · no están en el mismo vecindario
· puedo encontrar un camino de puntos " .. " .. " .. "
pero "a" y "z" no tienen por qué ser nucleares

$a \leftrightarrow z$



Agrupamiento basado en densidad

DBSCAN

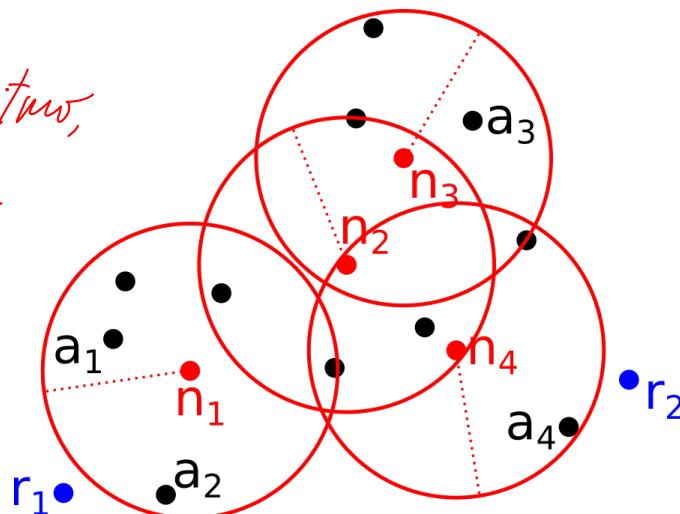
Clúster

Conjunto de puntos nucleares denso-conectados y el resto de puntos (directamente) denso-alcanzables desde ellos.

[Examen]

Problema del algoritmo,
muy dependiente de
la inicialización.

Los puntos borde
dependiendo de esto
caen en un cluster
o en otro



Agrupamiento basado en densidad

DBSCAN

DBSCAN

→ número de clusters

1. $C = 1$

[Examen]

2. Para todo ejemplo, x_i

2.1. Si x_i ya está asignado, continuar (volver a 2)

2.2. Calcular vecindario V de x_i (dado ϵ)

→ cardinalidad de V

2.3. Si $|V| < M$, asignar x_i como ruido y continuar (volver a 2)

2.4. Crear clúster número C y asignarle x_i

2.5. Para todo ejemplo $x_j \in V$

2.5.1. Si x_j está asignado como ruido, asignarlo al clúster C y continuar (volver a 2.5)

2.5.2. Si x_j tiene otra asignación, continuar (volver a 2.5)

2.5.3. Asignar x_j al clúster C → por si el punto no tiene ninguna asignación

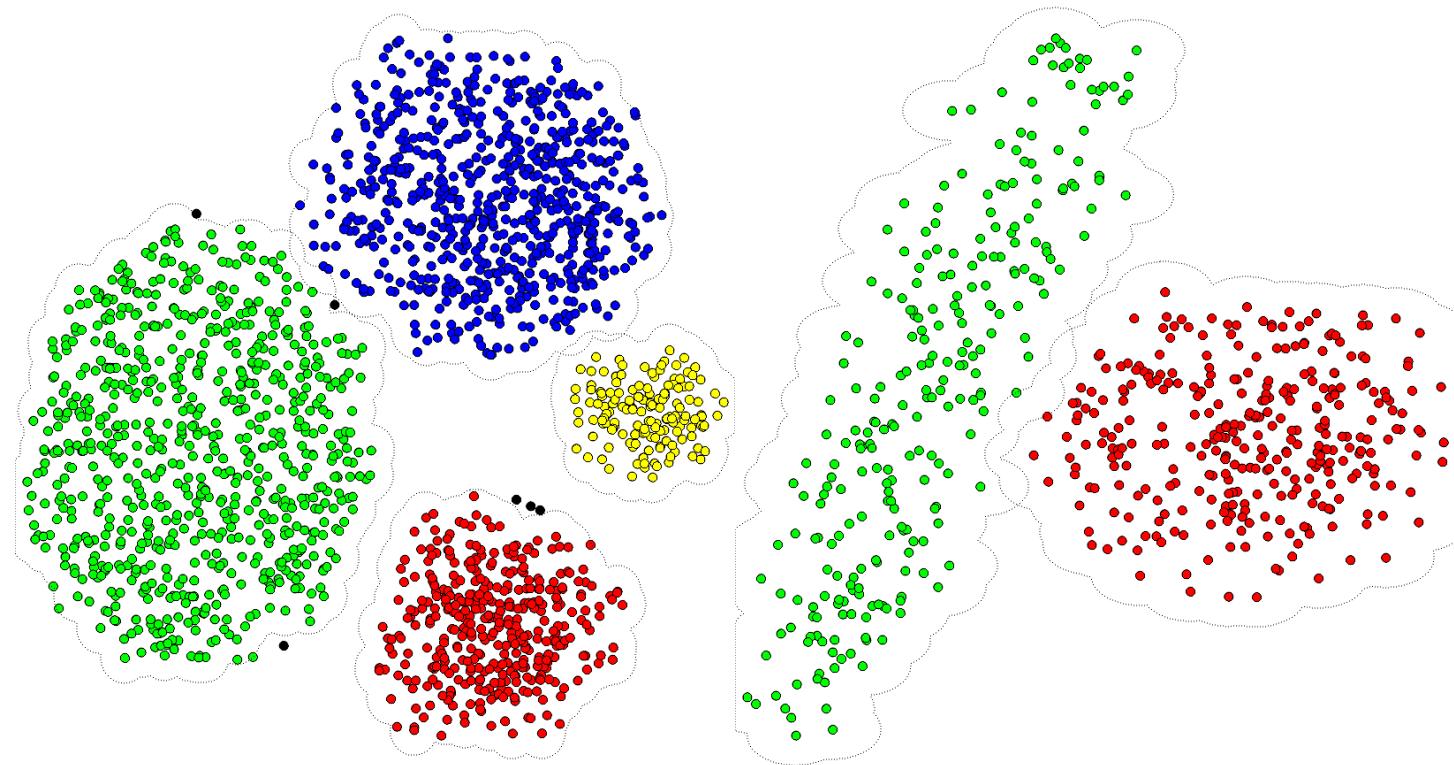
2.5.4. Calcular vecindario V' de x_j (dado ϵ)

2.5.5. Si $|V'| \geq M$; $V = V \cup V'$

2.6. $C = C + 1$

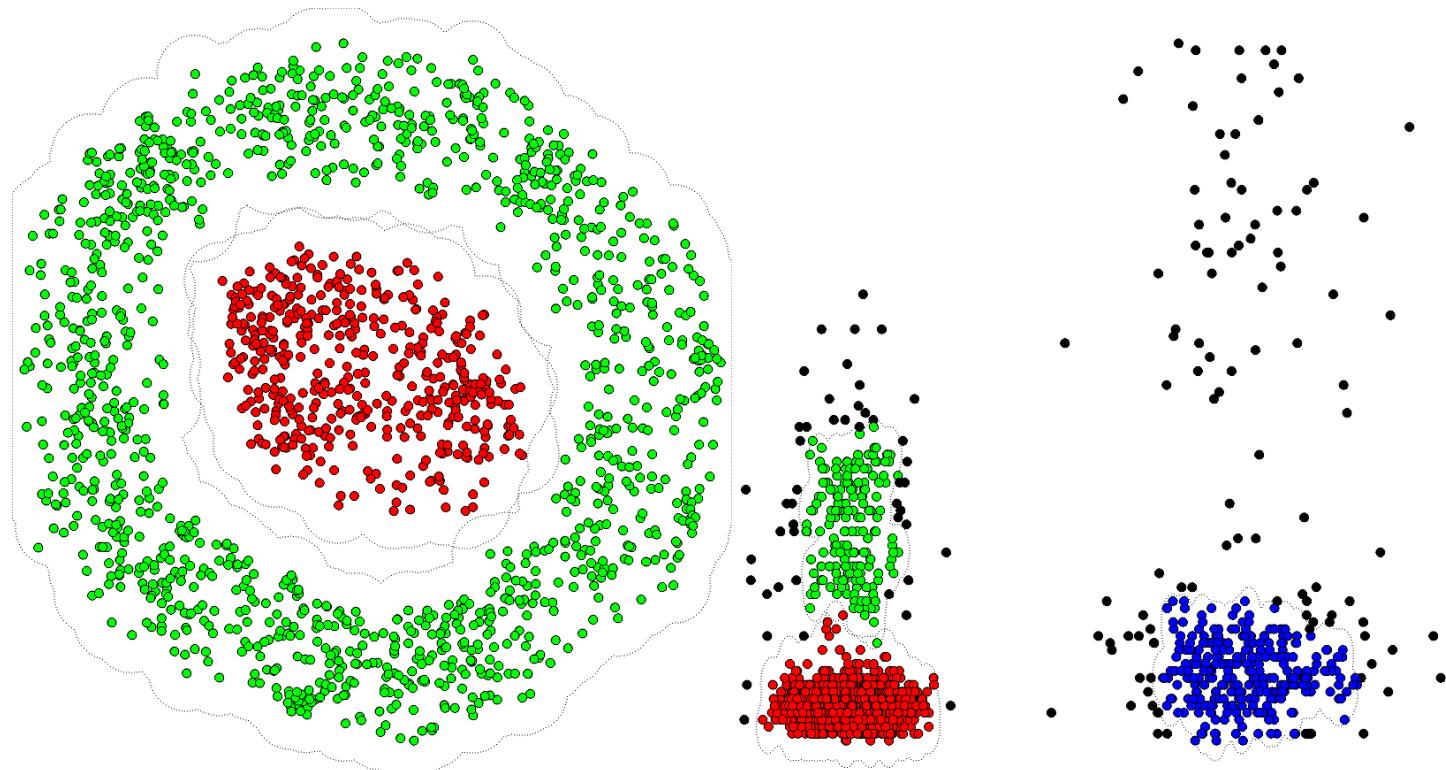
Agrupamiento basado en densidad

DBSCAN



Agrupamiento basado en densidad

DBSCAN



Agrupamiento basado en densidad

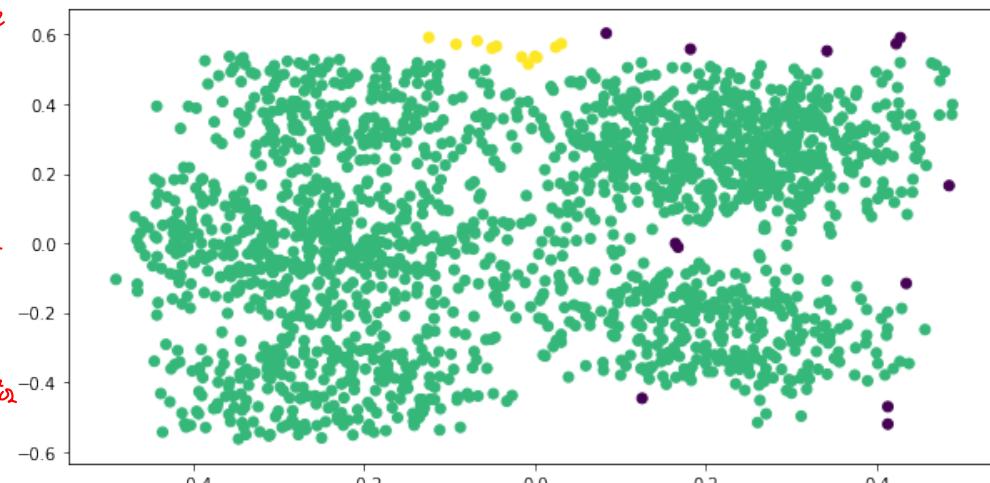
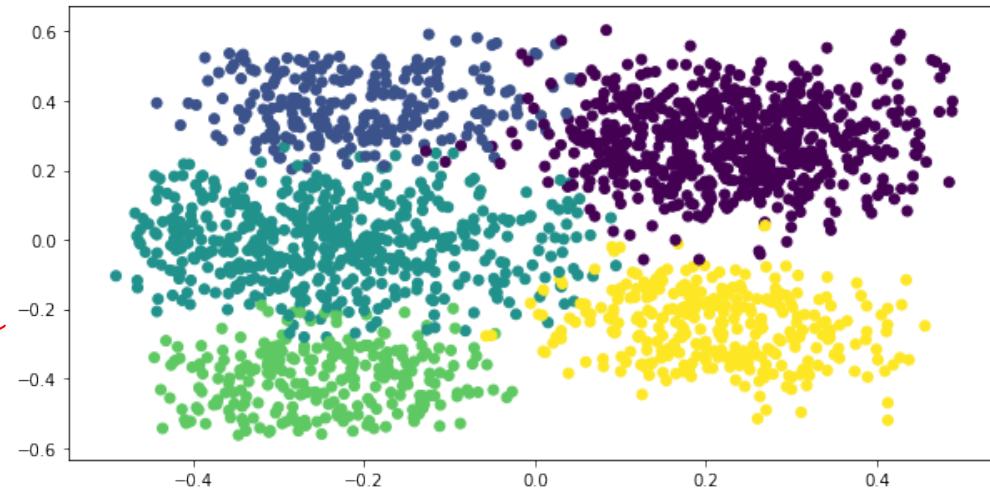
DBSCAN: Efecto de M

[Examen]

Si aumentamos M el número de clusters aumenta, porque un cluster se forma de muchos puntos nucleares unidos. Esto implica que a mayor M , menor número de puntos nucleares y por tanto mayor separación entre puntos nucleares.

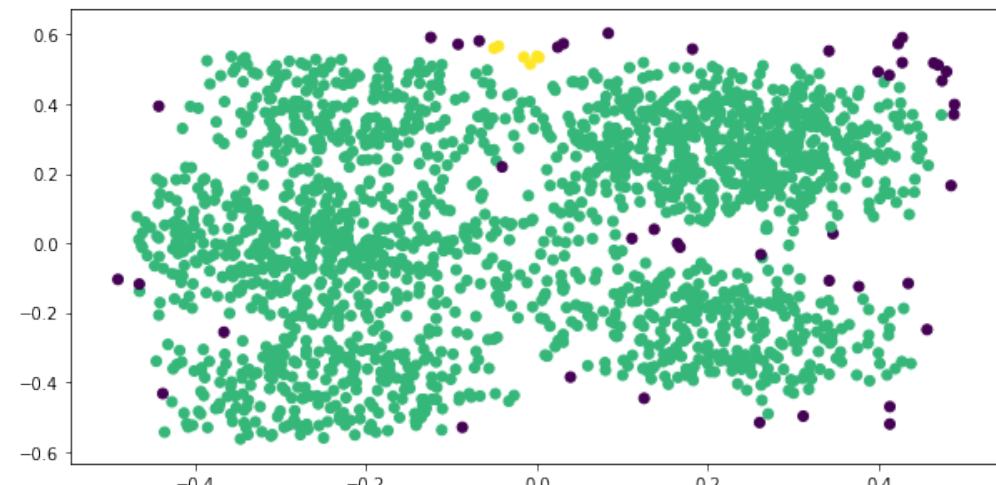
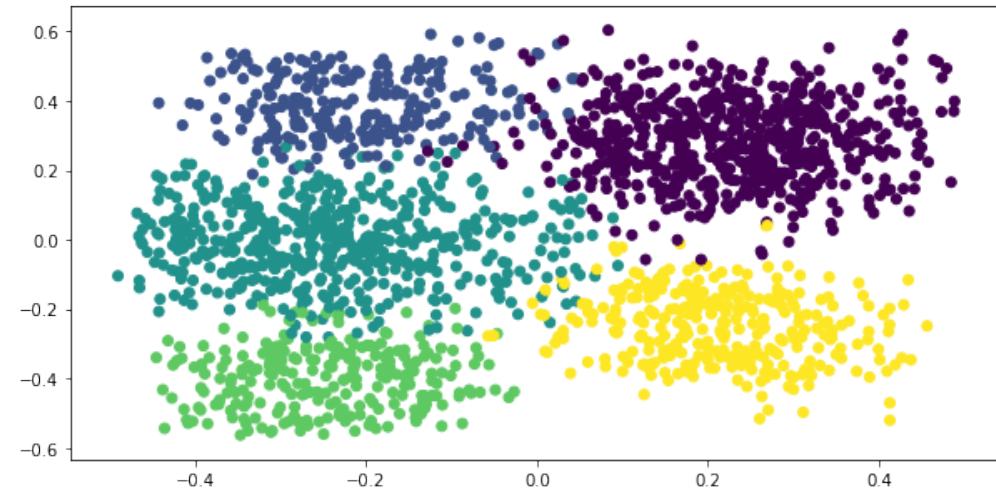


mayor número de clusters y más puntos ruidos



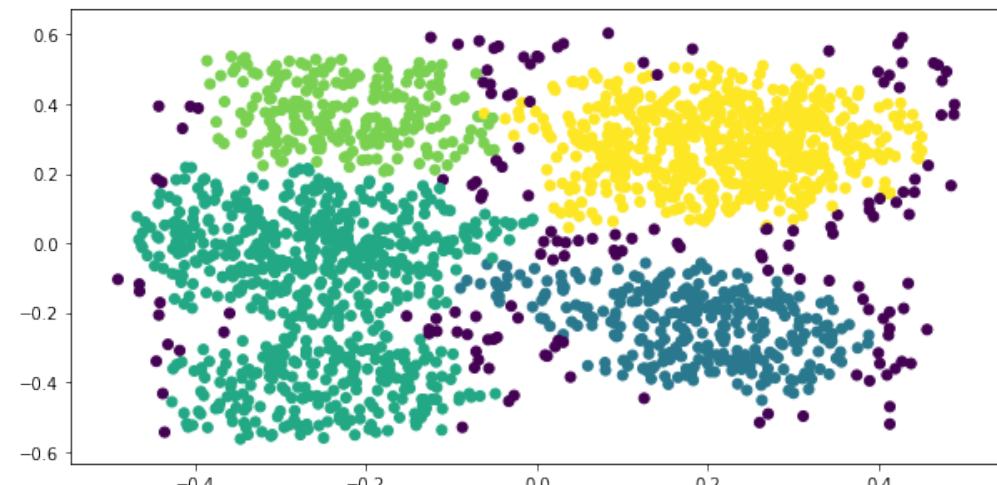
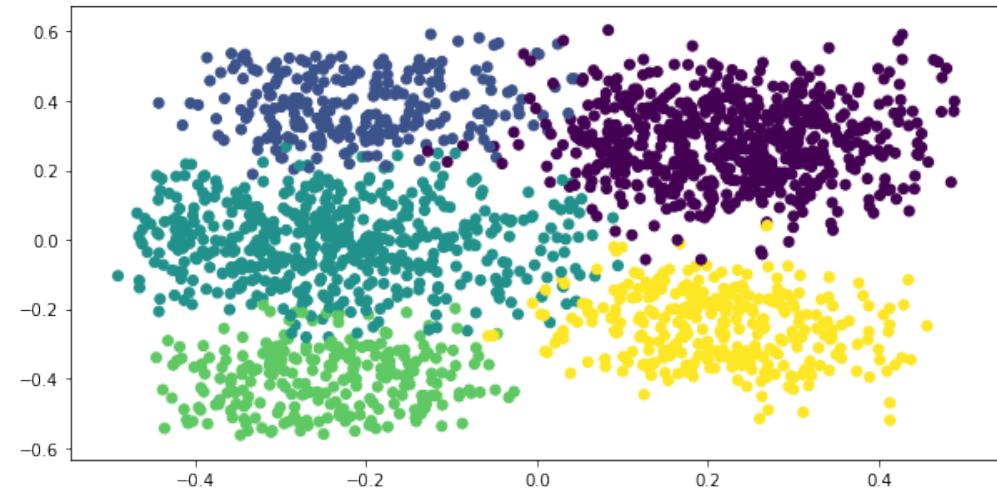
Agrupamiento basado en densidad

DBSCAN: Efecto de M



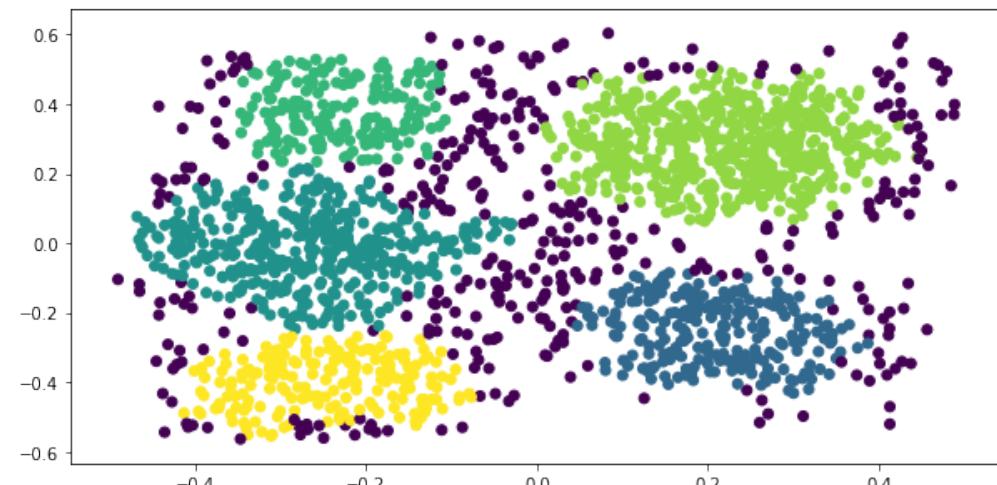
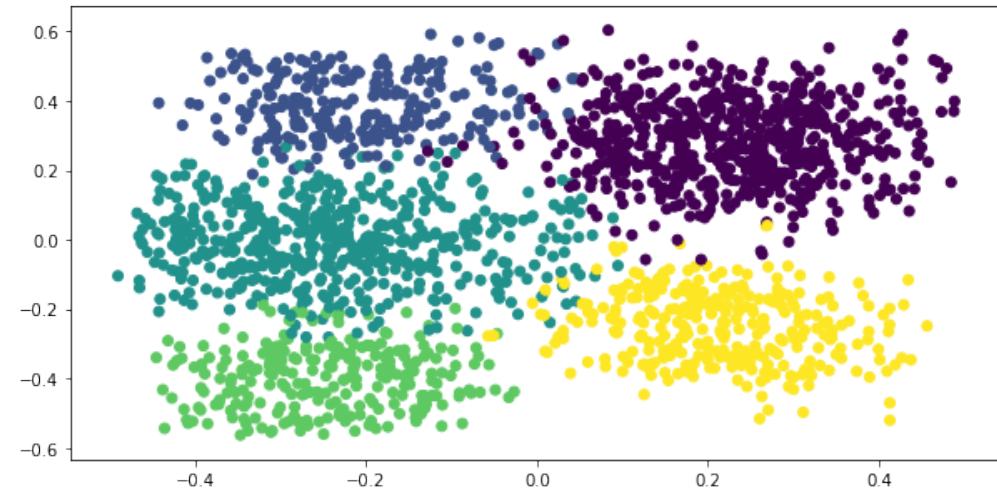
Agrupamiento basado en densidad

DBSCAN: Efecto de M



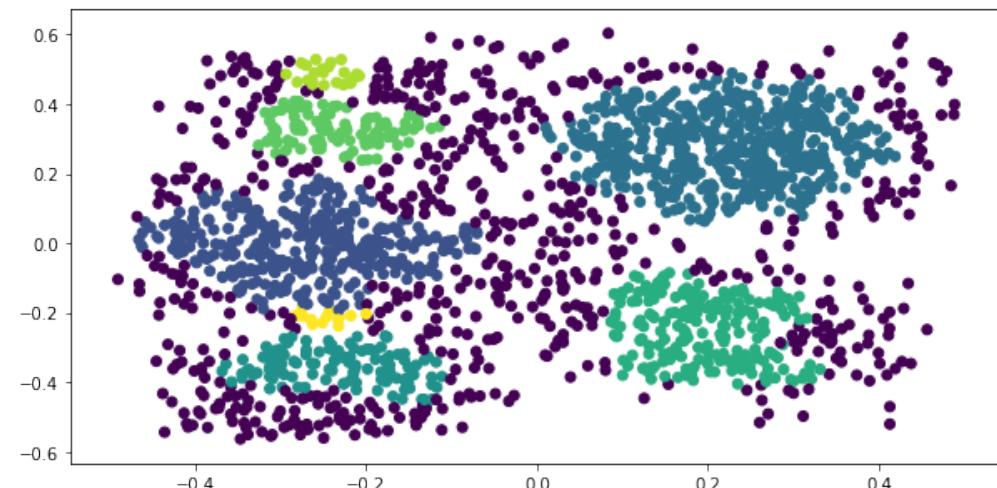
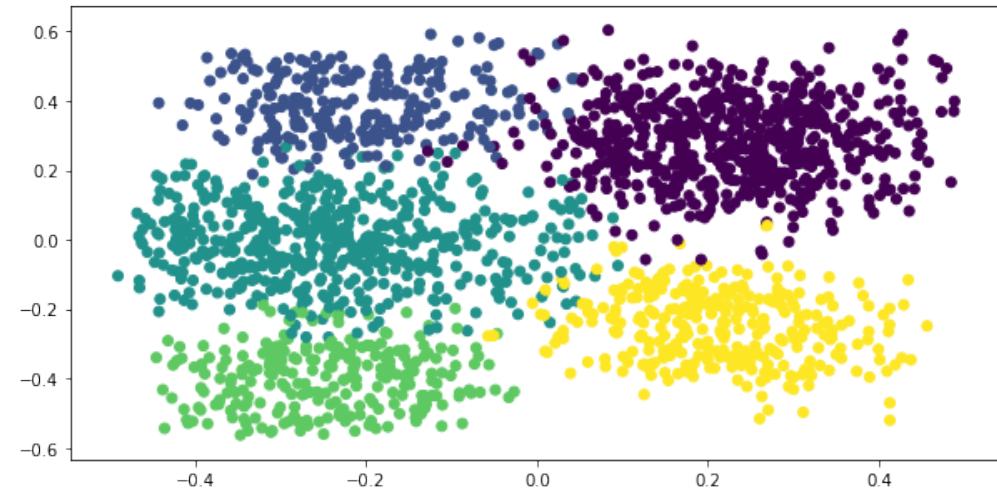
Agrupamiento basado en densidad

DBSCAN: Efecto de M



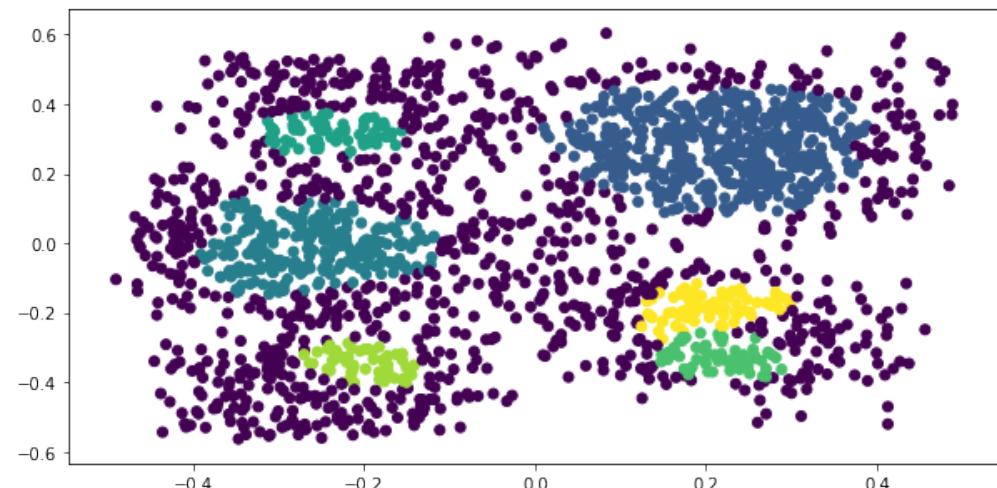
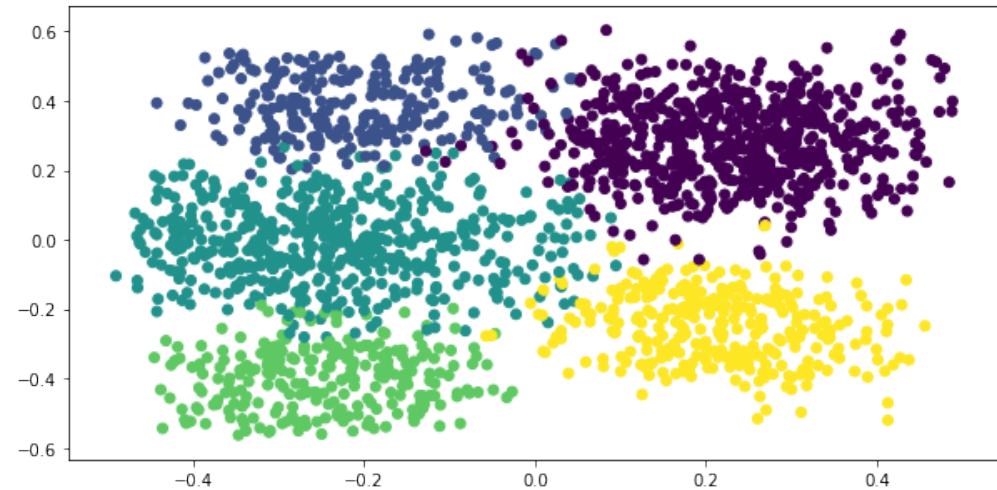
Agrupamiento basado en densidad

DBSCAN: Efecto de M



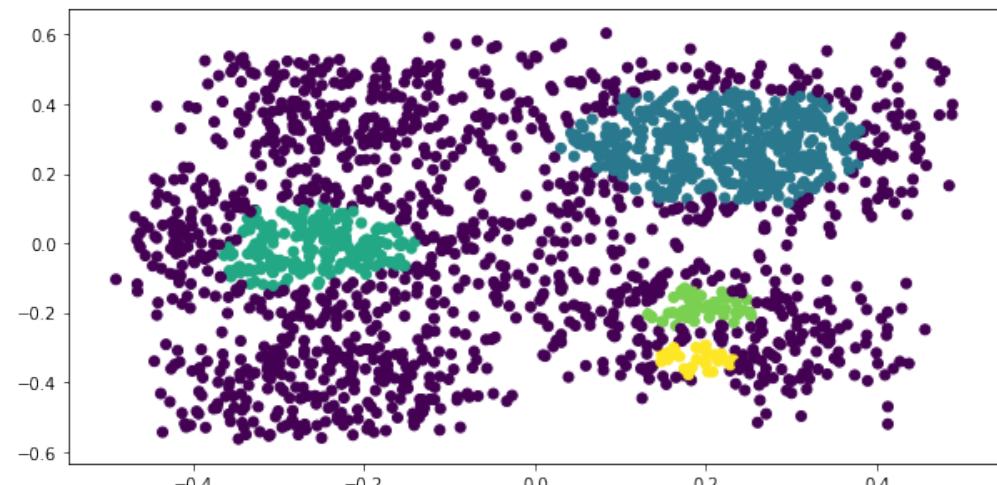
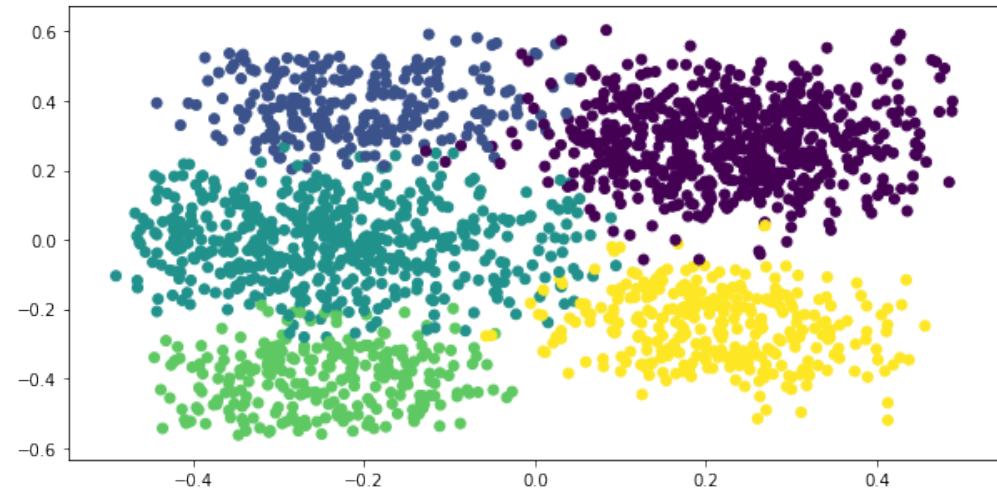
Agrupamiento basado en densidad

DBSCAN: Efecto de M



Agrupamiento basado en densidad

DBSCAN: Efecto de M



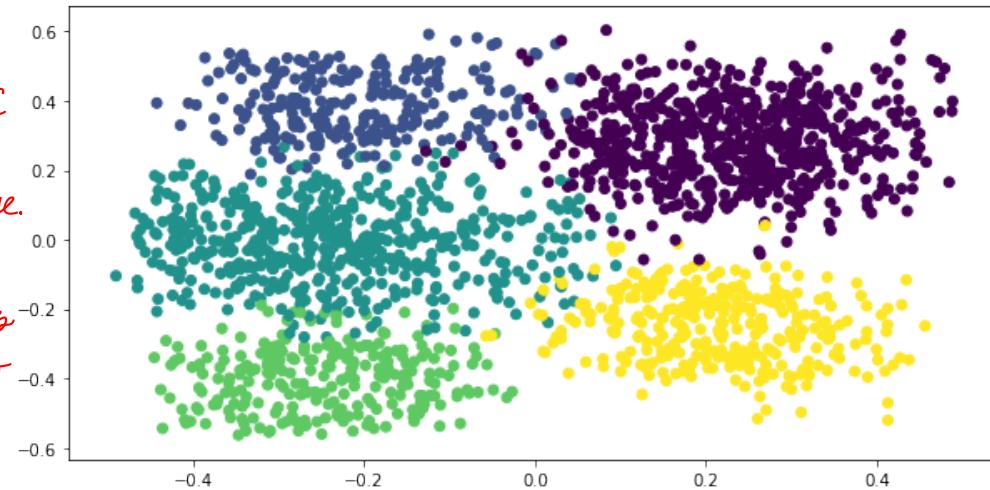
Agrupamiento basado en densidad

DBSCAN: Efecto de ϵ

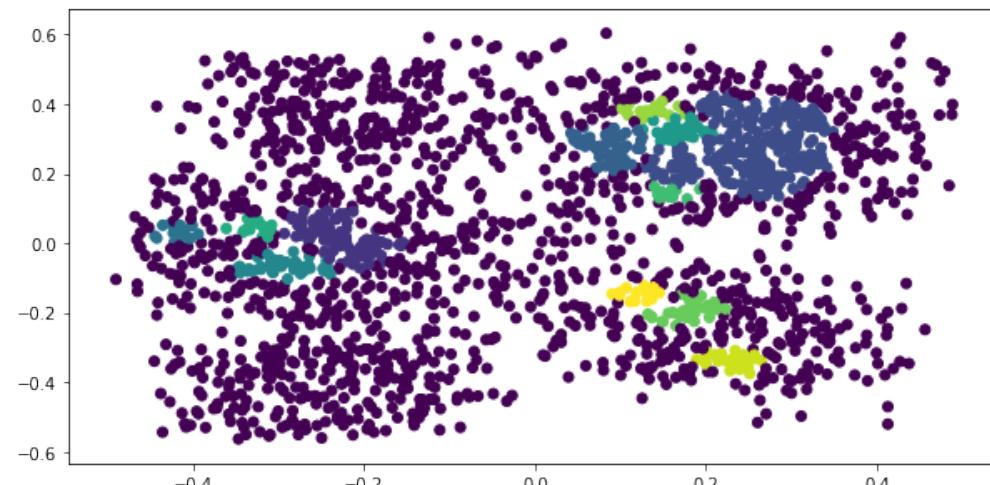
[Examen]

Si aumentamos ϵ
el número de
clusters disminuye.

Al $\uparrow \epsilon \Rightarrow$ 1 minor
de puntos nucleares
porque el vecindario
es más grande.

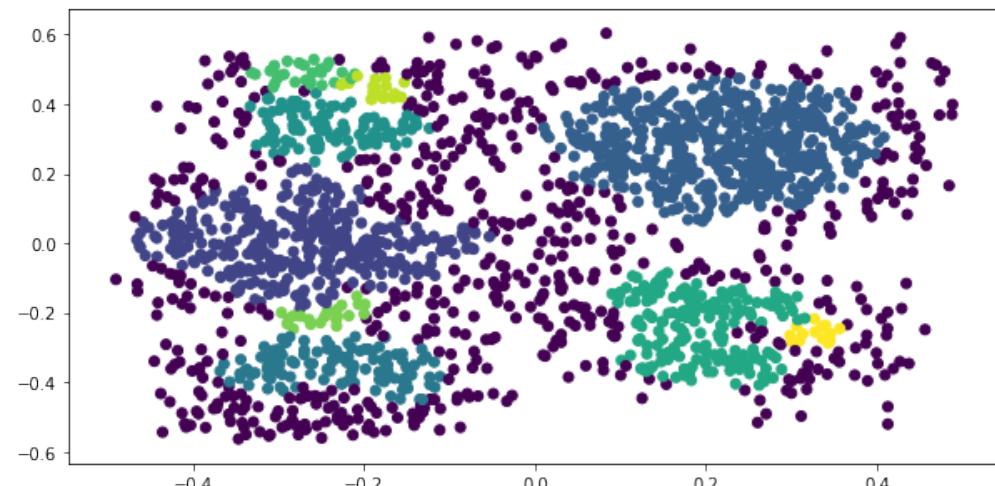
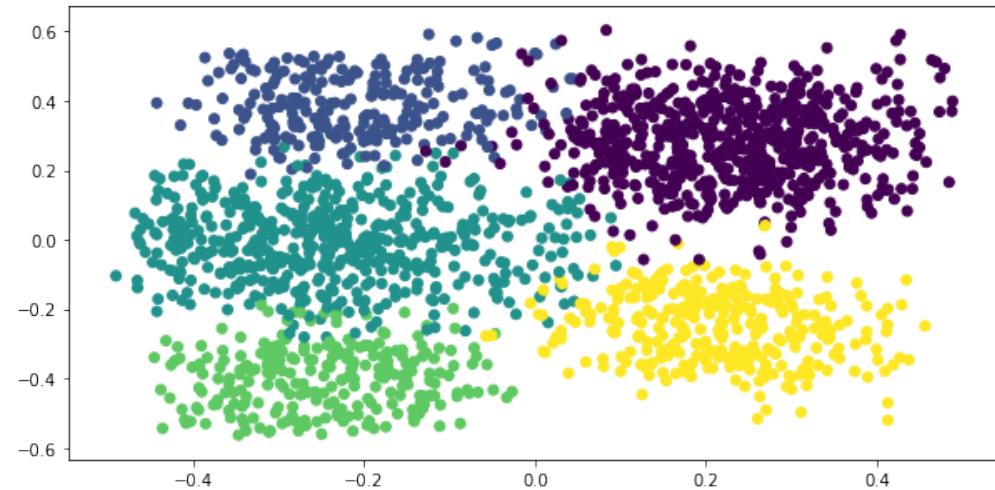


Disminuye el
número de clusters
y los puntos ruido



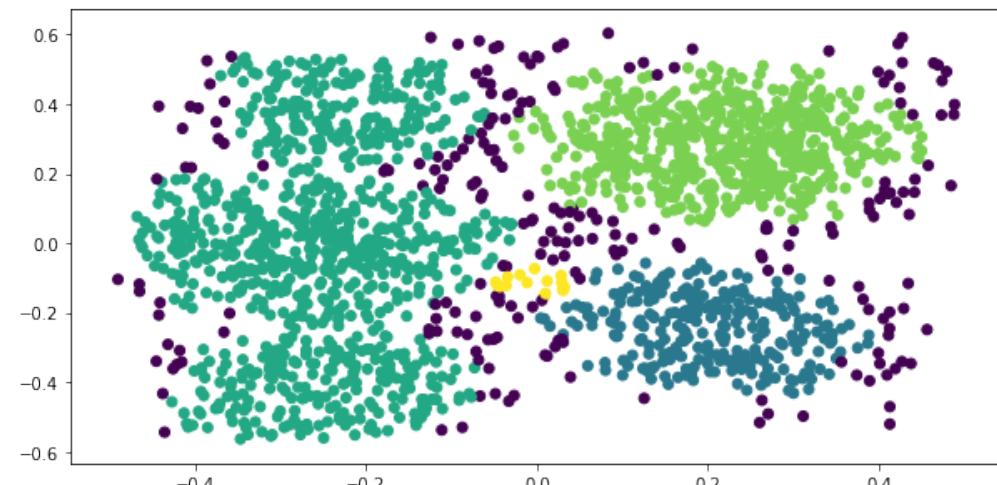
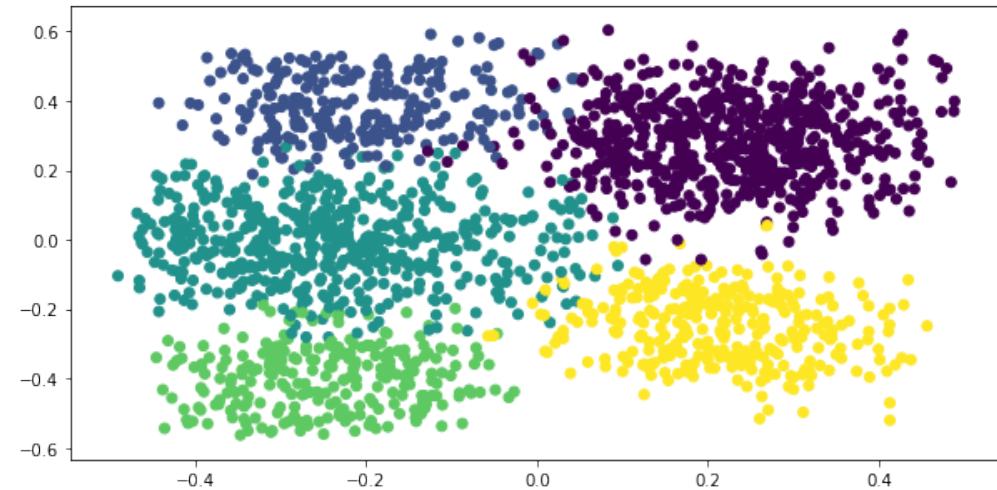
Agrupamiento basado en densidad

DBSCAN: Efecto de ϵ



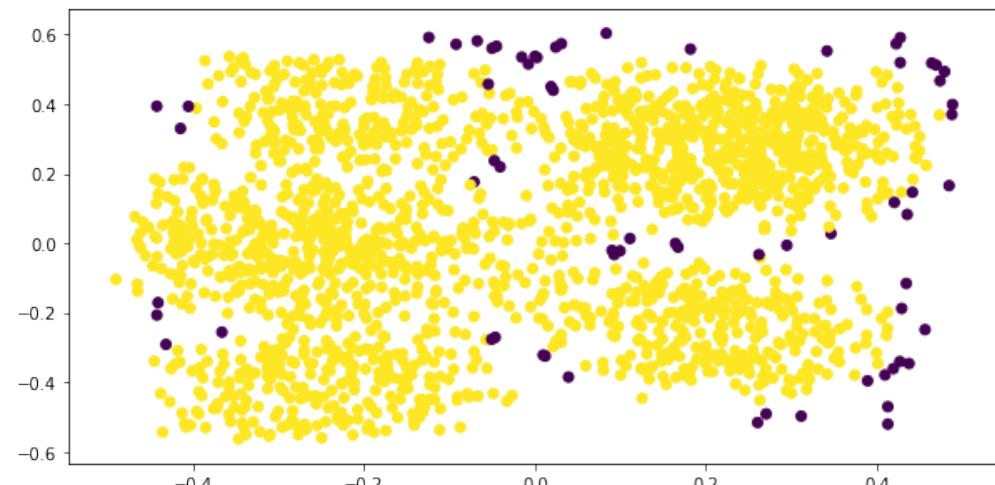
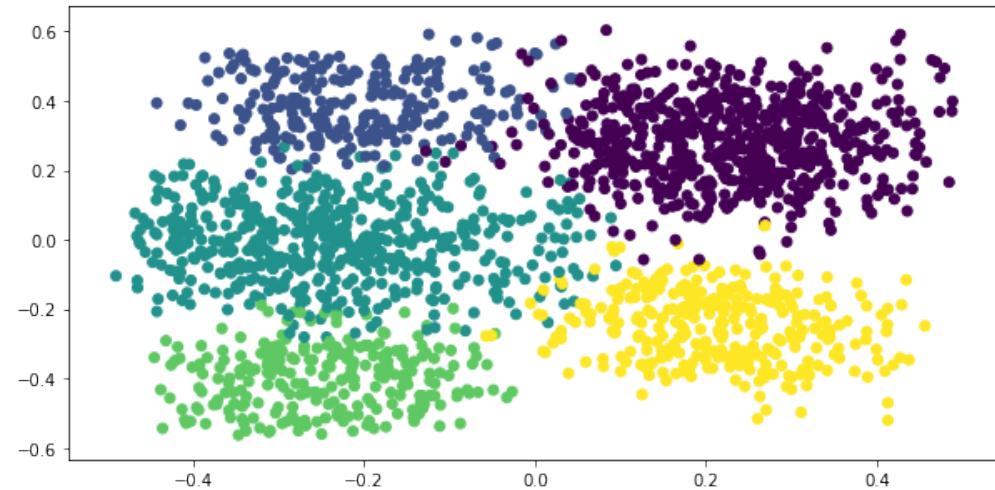
Agrupamiento basado en densidad

DBSCAN: Efecto de ϵ



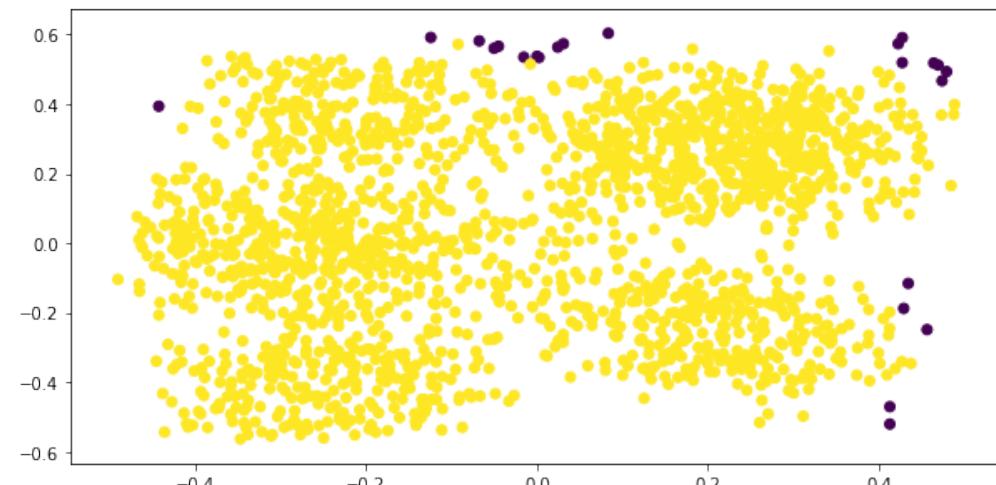
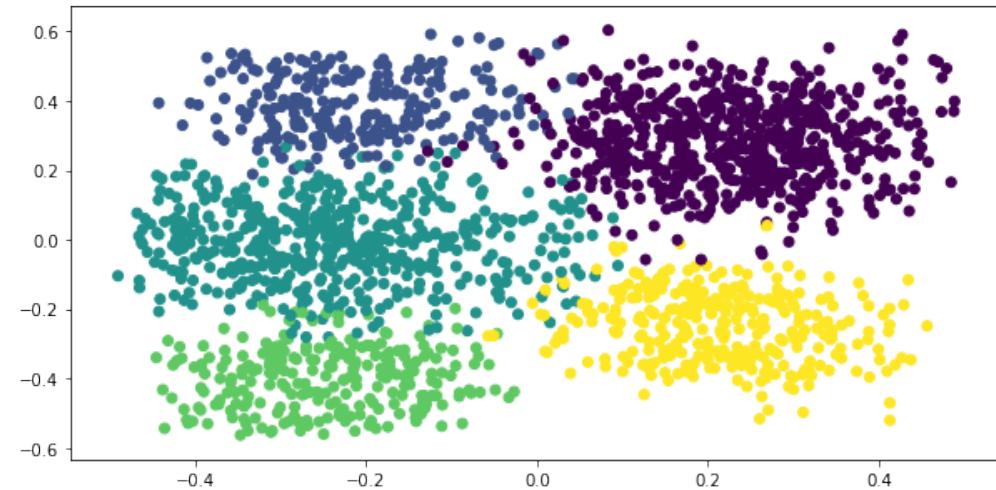
Agrupamiento basado en densidad

DBSCAN: Efecto de ϵ



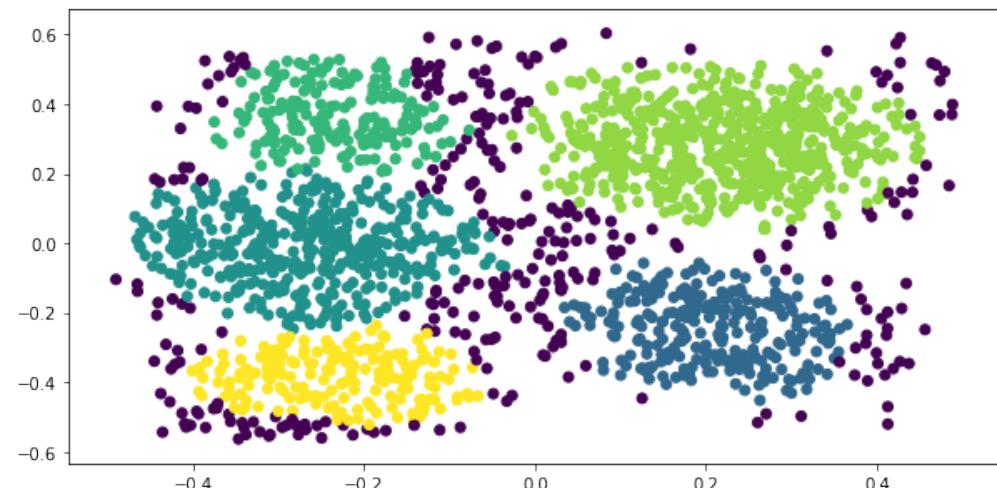
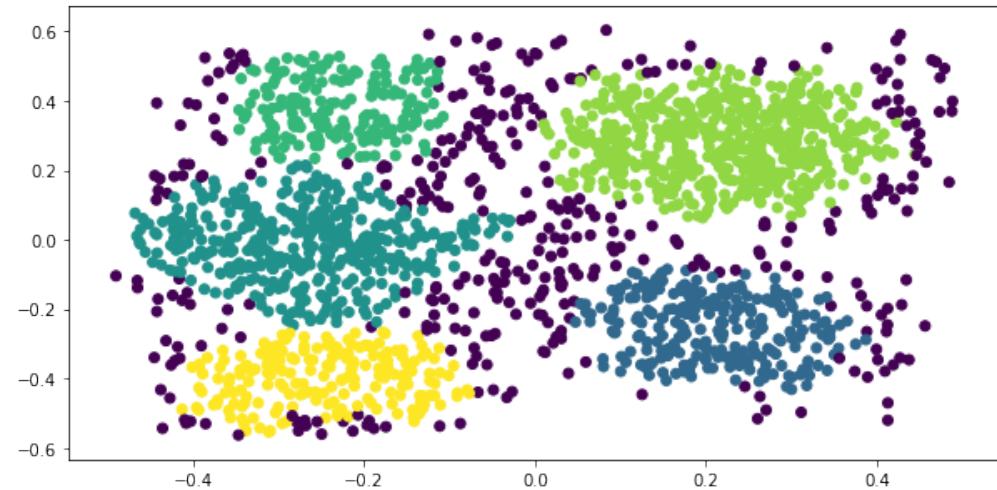
Agrupamiento basado en densidad

DBSCAN: Efecto de ϵ



Agrupamiento basado en densidad

DBSCAN: Efecto de ambos, M y ϵ



Agrupamiento basado en densidad

DBSCAN

Ventajas

- ▶ No es necesario especificar K
- ▶ Definición basada en densidad
- ▶ Funciona con clústeres de diferente tamaño y formas
- ▶ Generalizable a otros conceptos de densidad
- ▶ Puede funcionar con diferentes medidas de distancia

Agrupamiento basado en densidad

DBSCAN

Desventajas

- ▶ Definición compleja
- ▶ Problemas al lidiar con clústeres de diferente densidad
- ▶ Dos parámetros interdependientes a ajustar

Aprendizaje no supervisado

VC06: Agrupamiento basado en densidad – Mean Shift

Rocío del Amor del Amor

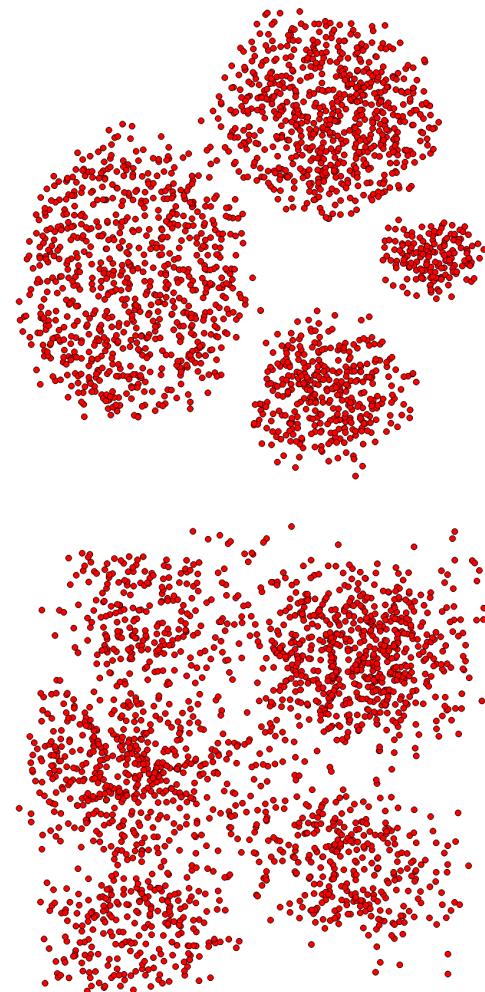
mrocio.delamor@campusviu.es

Universidad Internacional de Valencia

Agrupamiento

Tipos de algoritmos de agrupamiento

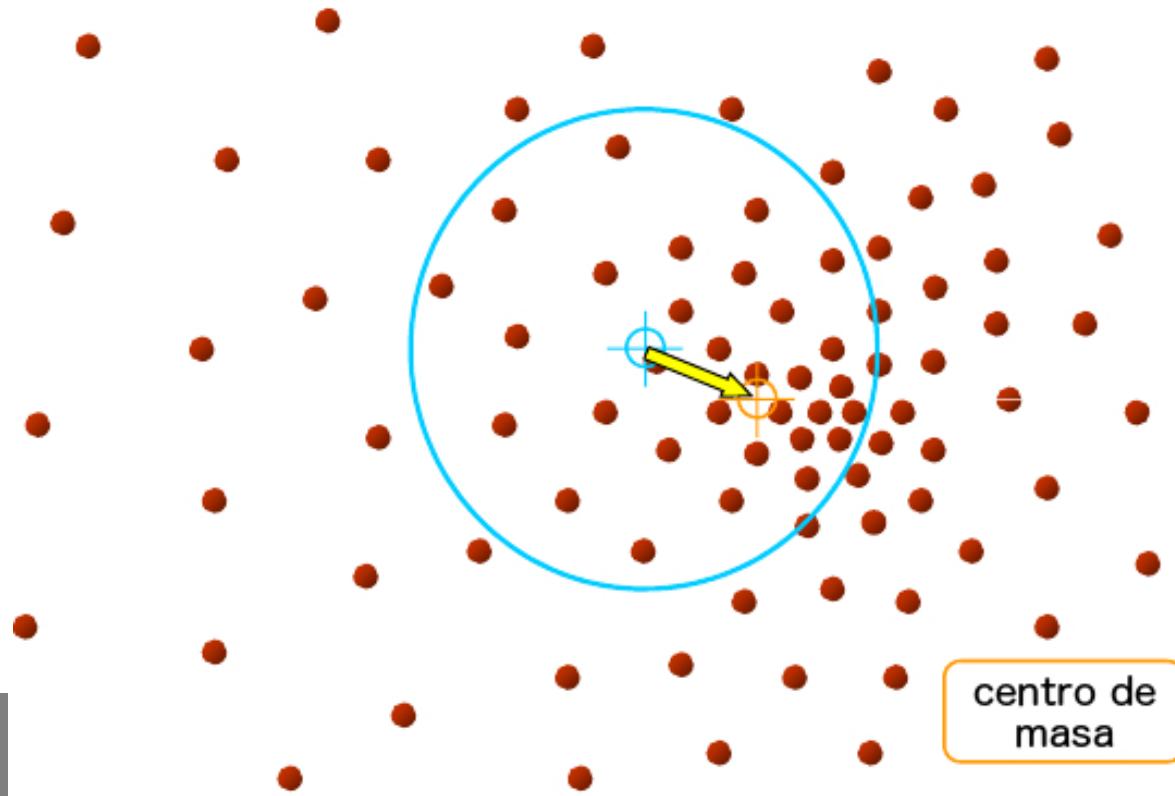
- ▶ Basados en particiones
- ▶ Jerárquicos
- ▶ Espectrales
- ▶ **Basados en densidad**
- ▶ Probabilísticos



Agrupamiento basado en densidad

Idea

Si desplazamos cada punto al centro de masa de su vecindario, los puntos se acaban agrupando de manera natural en grupos

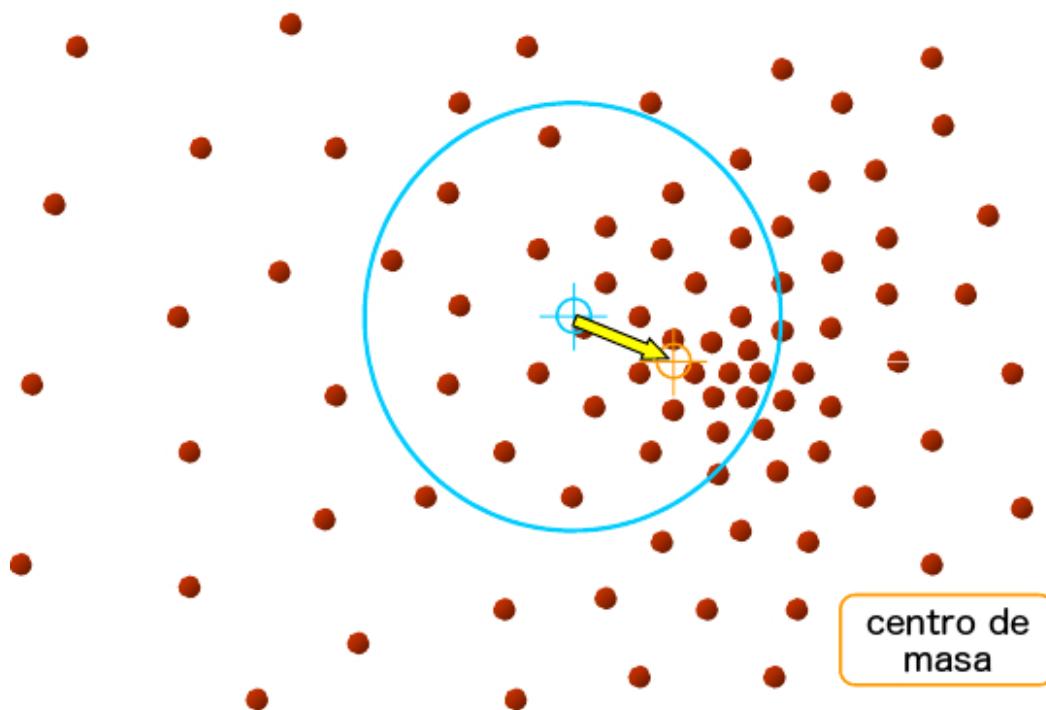


centro de
masa

Agrupamiento basado en densidad

Conceptos

- ▶ **Vecindario**, ¿qué casos se usan para calcular el centro?
- ▶ **Kernel**, ¿cómo se ponderan los casos usados para calcular el centro?



Agrupamiento basado en densidad

Media ponderada:

$$m(\mathbf{x}) = \frac{\sum_{i=1}^n k\left(\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{h^2}\right) \cdot \mathbf{x}_i}{\sum_{i=1}^n k\left(\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{h^2}\right)}$$

Kernel plano:

$$k(x) = \begin{cases} 1, & x \leq \lambda \\ 0, & x > \lambda \end{cases}$$

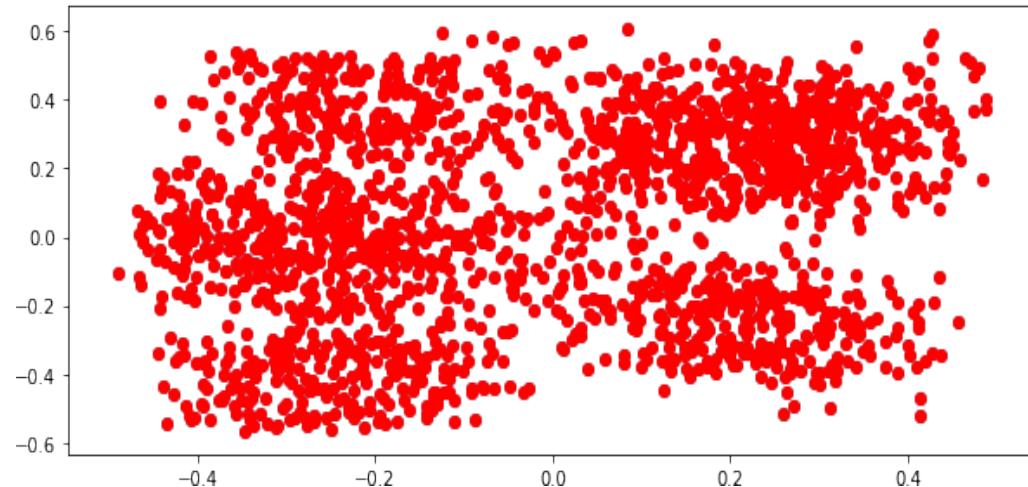
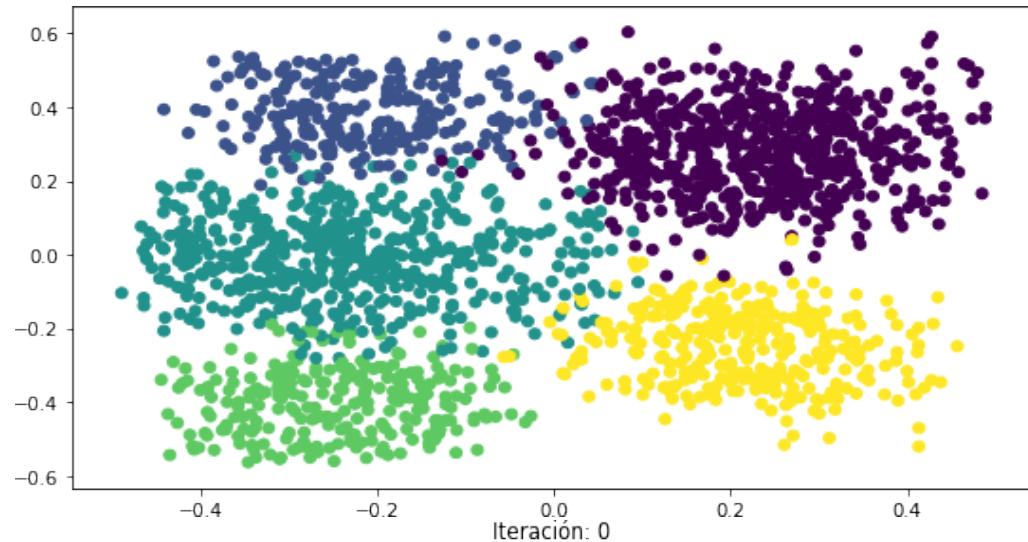
normalmente, $\lambda = 1$.

Kernel Gaussiano:

$$k(x) = e^{-x}$$

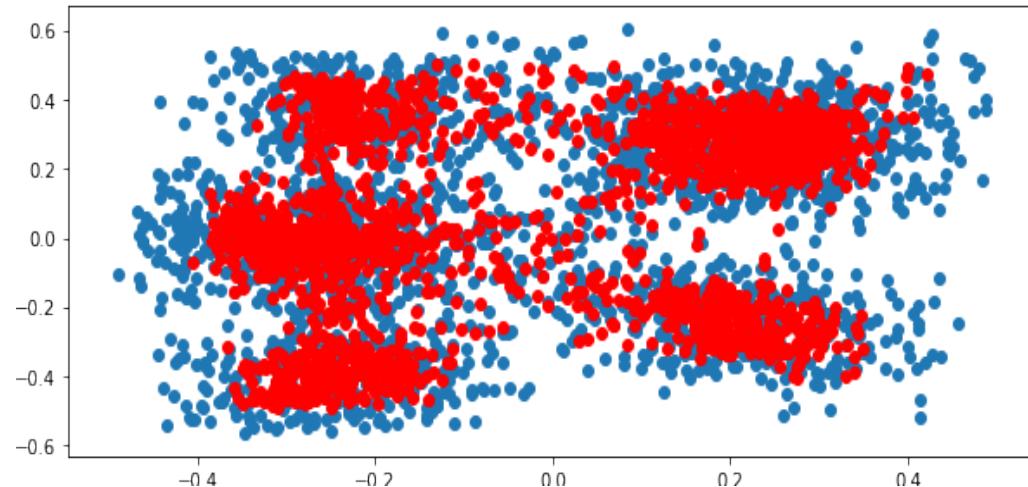
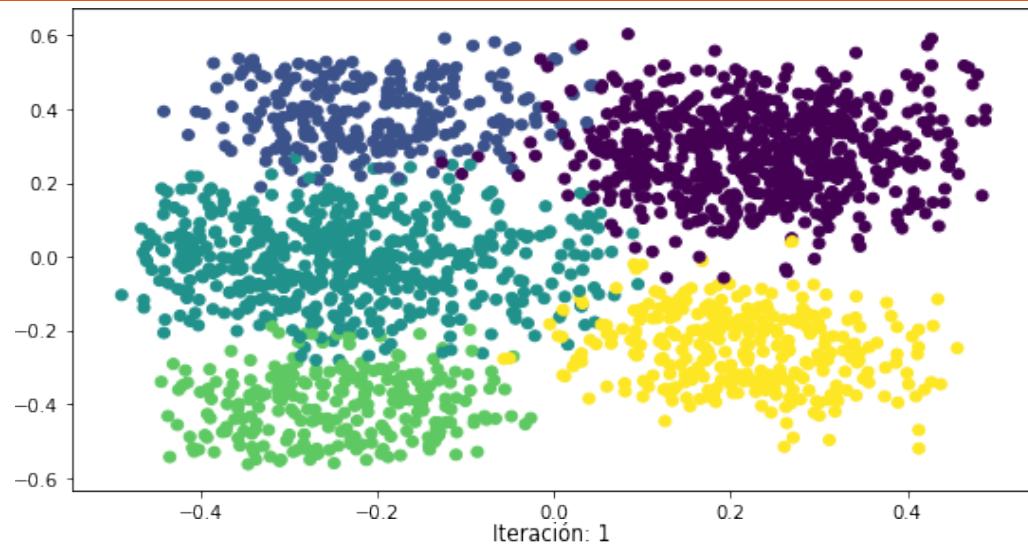
Agrupamiento basado en densidad

Mean-shift



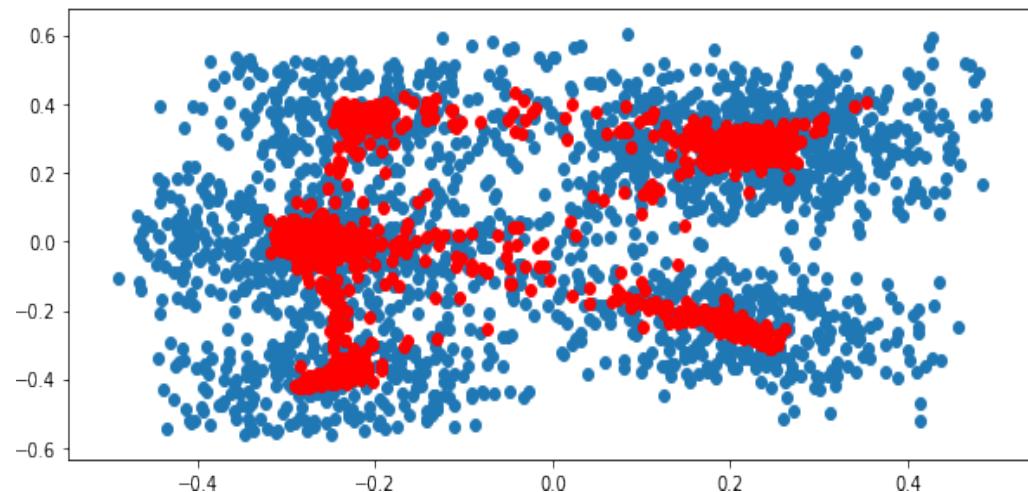
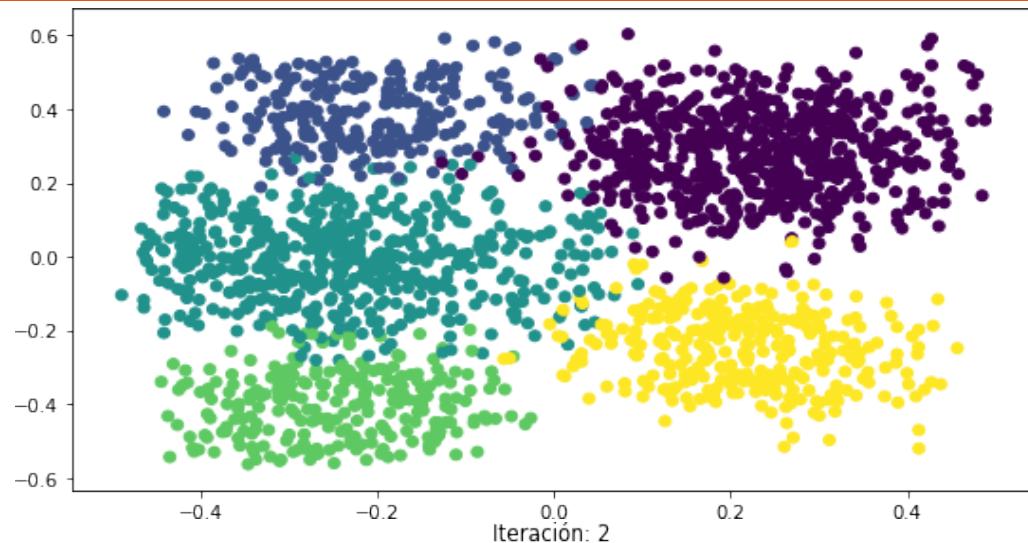
Agrupamiento basado en densidad

Mean-shift



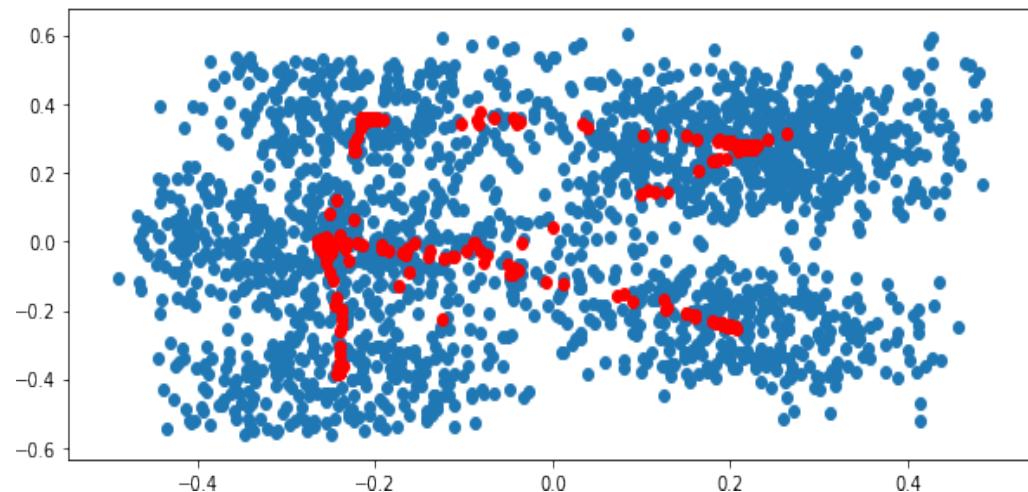
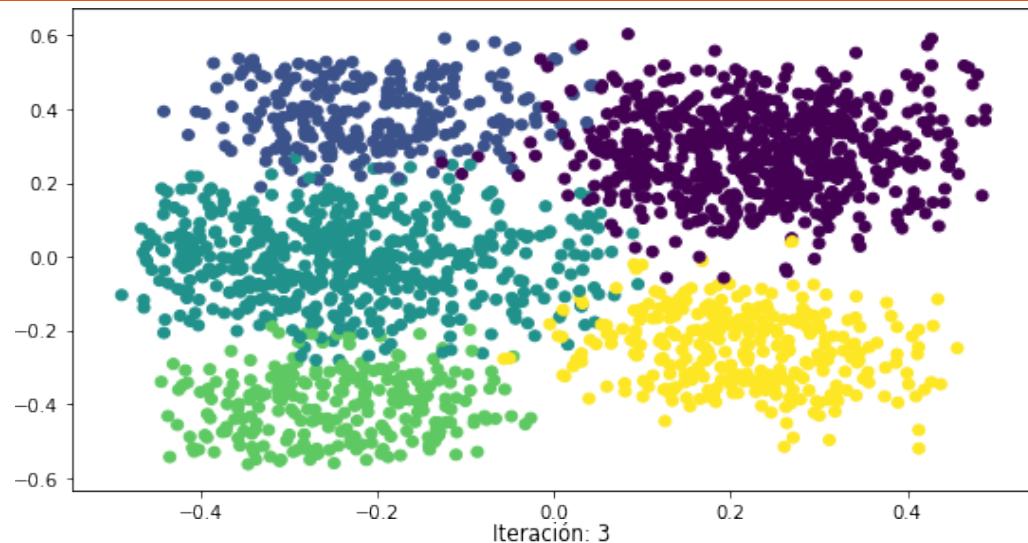
Agrupamiento basado en densidad

Mean-shift



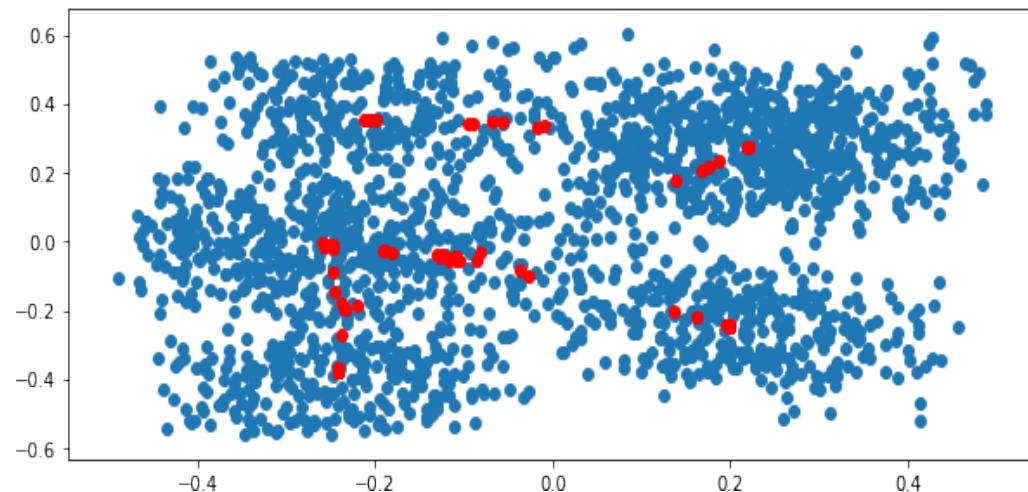
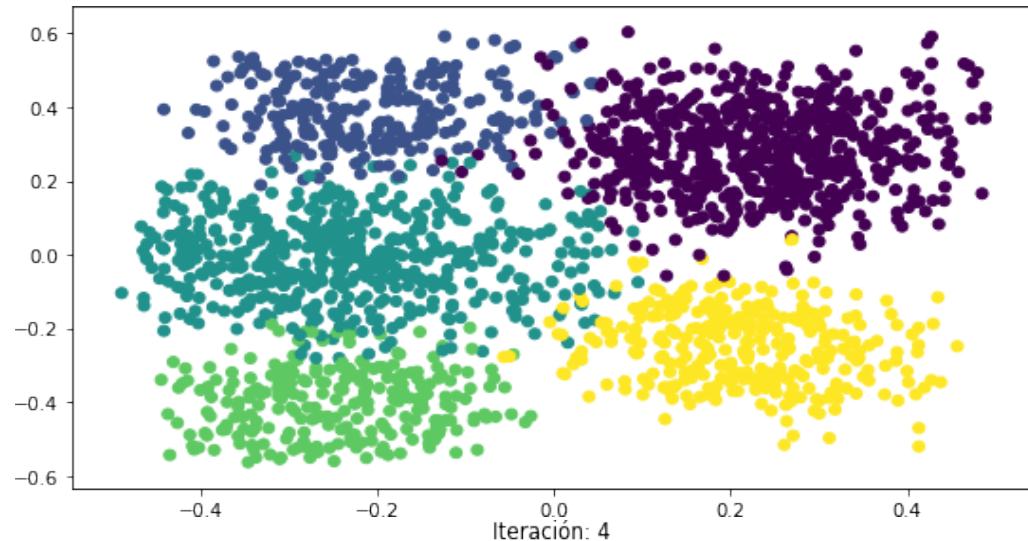
Agrupamiento basado en densidad

Mean-shift



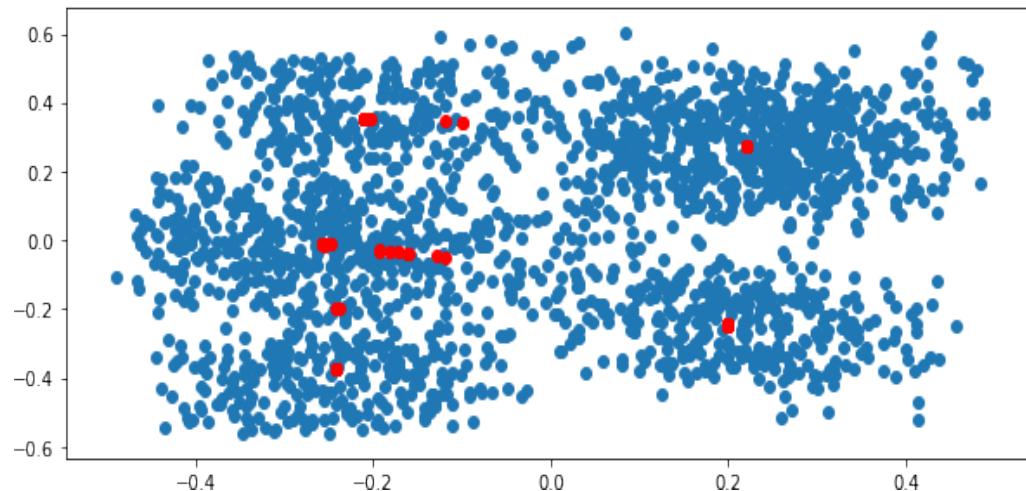
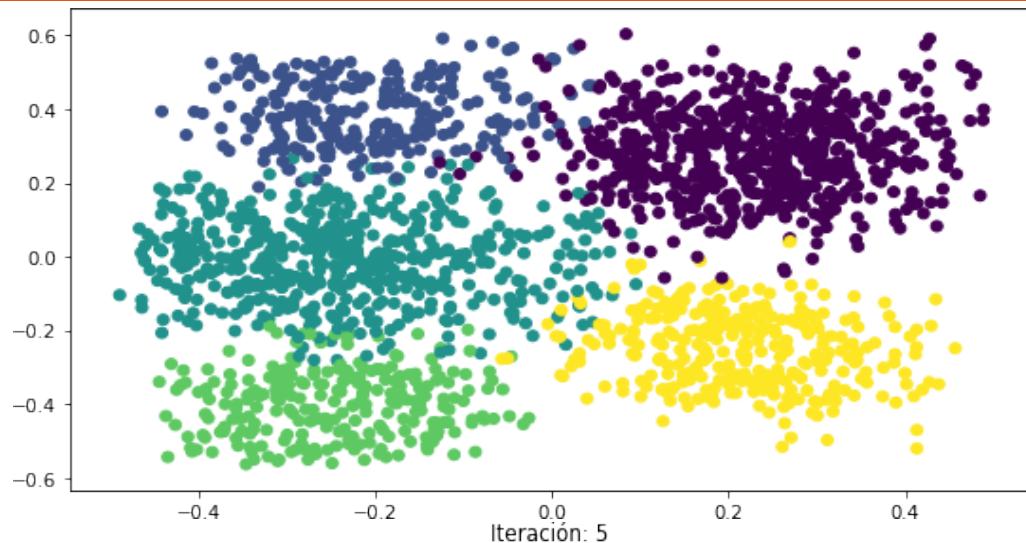
Agrupamiento basado en densidad

Mean-shift



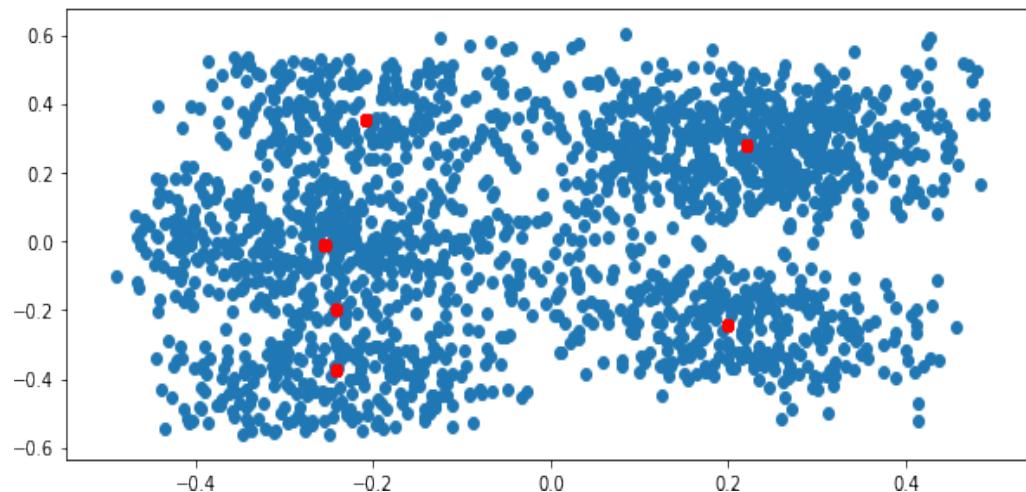
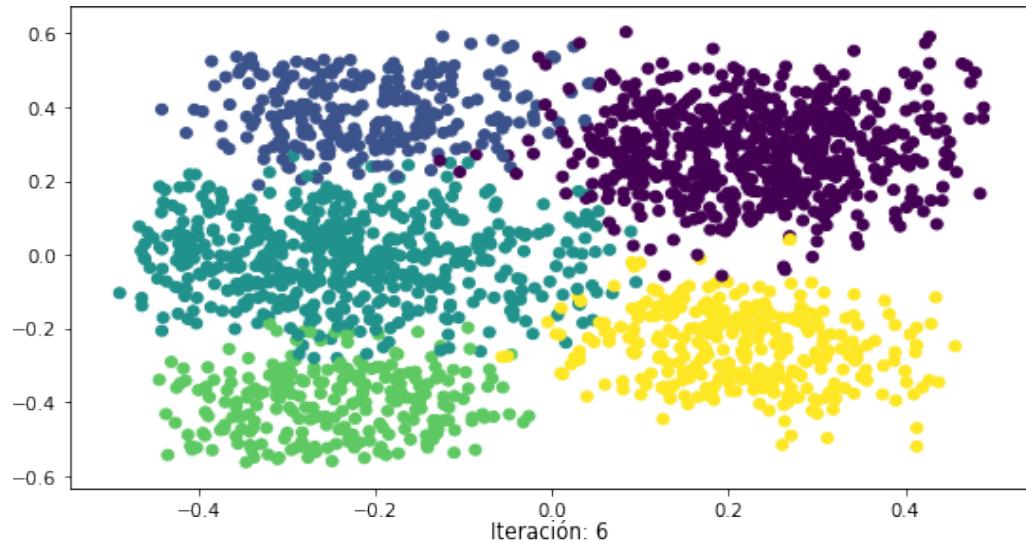
Agrupamiento basado en densidad

Mean-shift



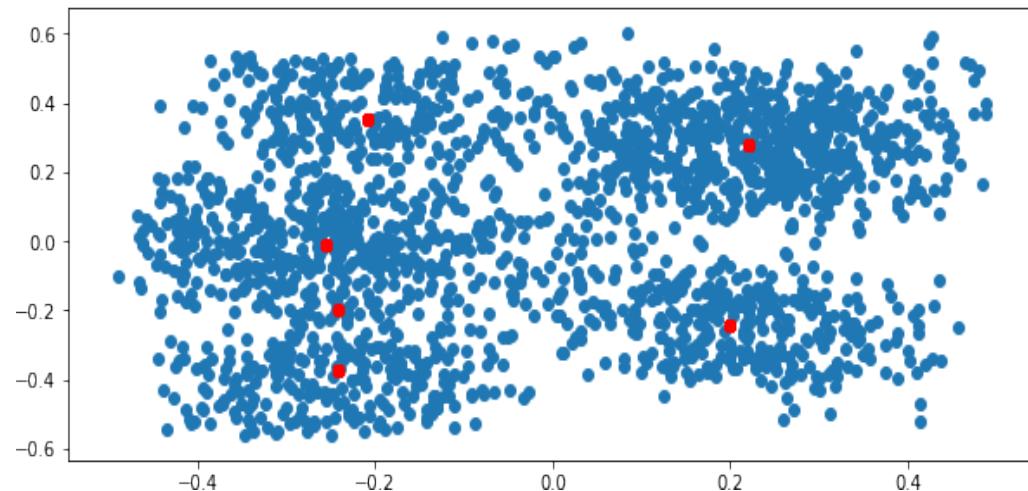
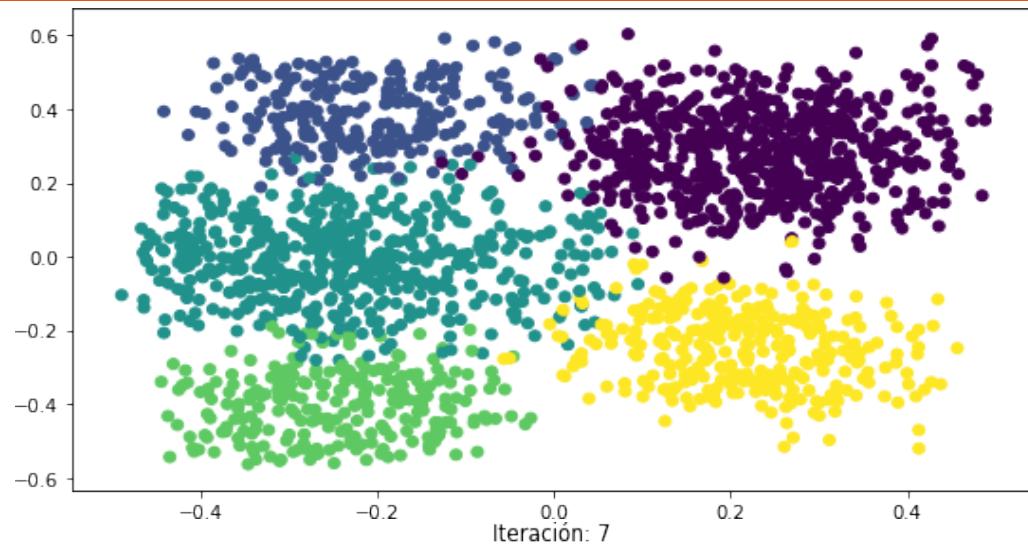
Agrupamiento basado en densidad

Mean-shift



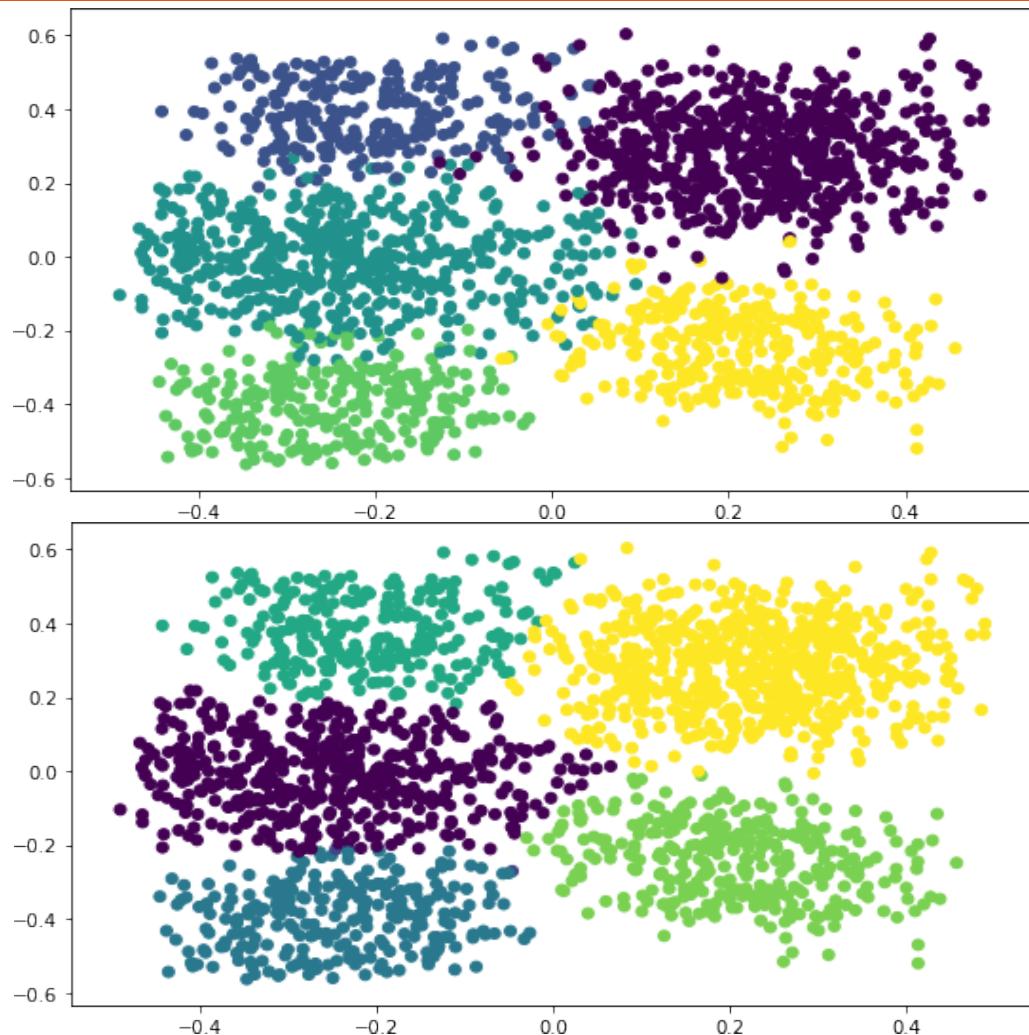
Agrupamiento basado en densidad

Mean-shift



Agrupamiento basado en densidad

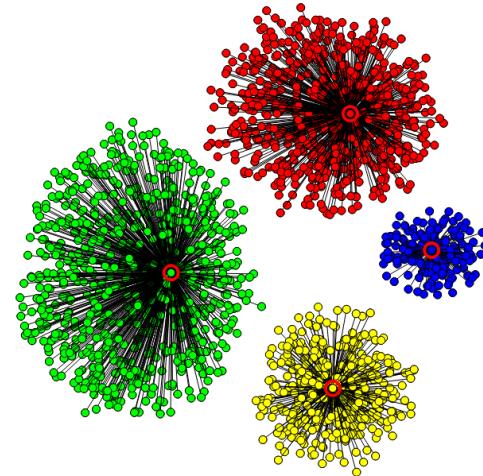
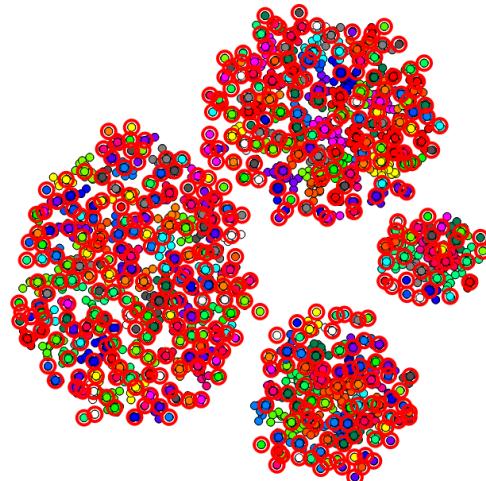
Mean-shift



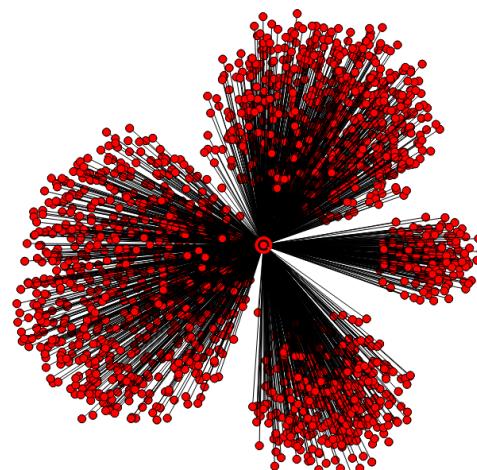
Agrupamiento basado en densidad

Mean-shift: efecto de h

$$h \equiv \varepsilon$$

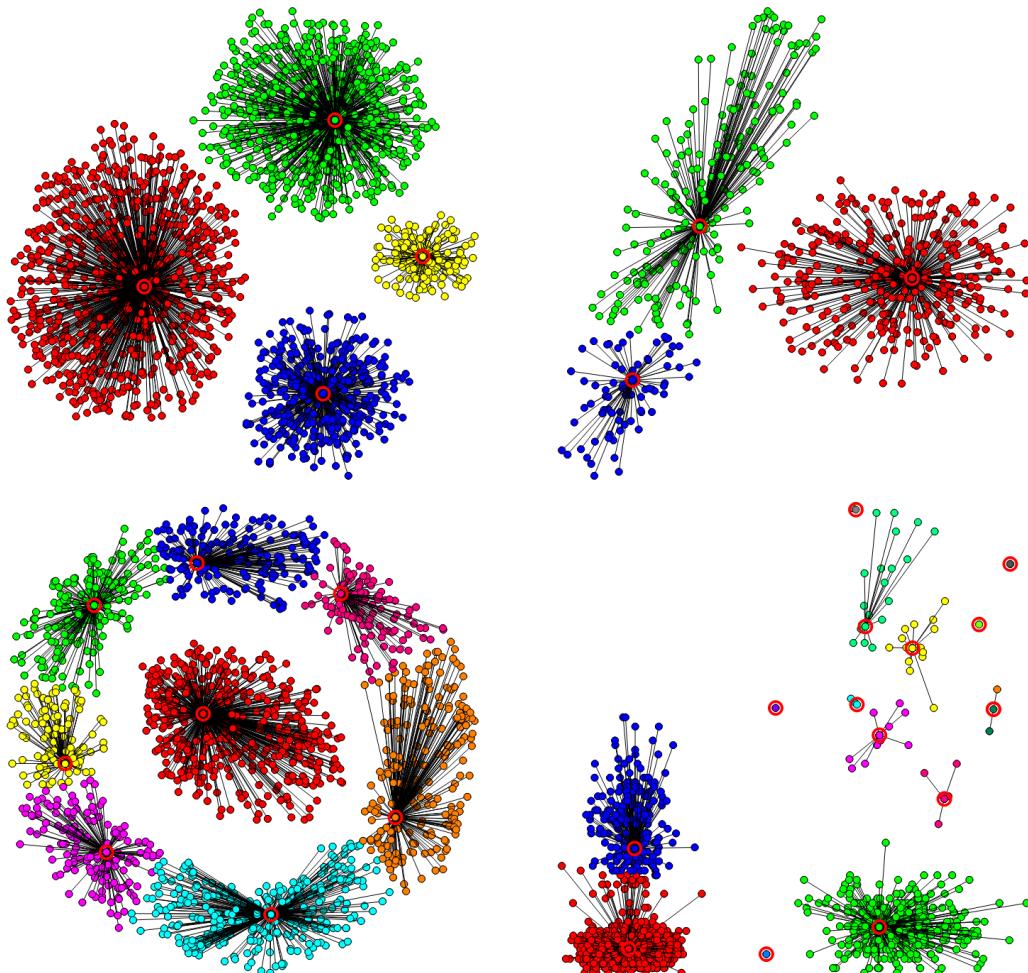


$$\downarrow h \Rightarrow \downarrow \text{nº clusters}$$



Agrupamiento basado en densidad

Mean-shift: efecto de h



Agrupamiento basado en densidad

Mean-shift



Agrupamiento basado en densidad

Mean-shift

Ventajas

- ▶ Conceptualmente sencilla
- ▶ No es necesario especificar K
- ▶ Definición basada en densidad
- ▶ Funciona con clústeres de diferente tamaño y formas
- ▶ Diferentes kernels

Agrupamiento basado en densidad

Mean-shift

Desventajas

- ▶ Clústeres no máximos
- ▶ Problemas al lidiar con clústeres de diferente densidad
- ▶ Sin demostración de convergencia en entornos reales

[NO EXAMEN]

Aprendizaje no supervisado

VC05: Agrupamiento basado en densidad – Propagación de afinidad

Rocío del Amor del Amor

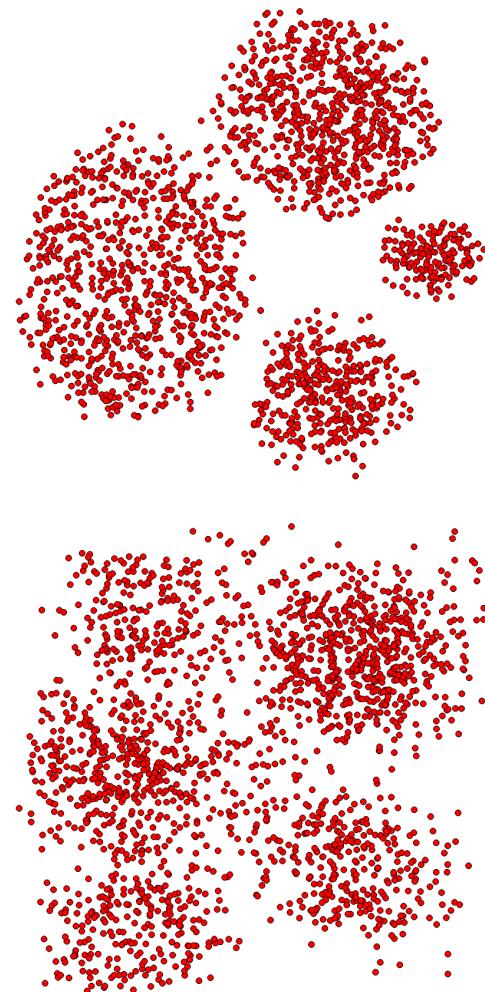
mrocio.delamor@campusviu.es

Universidad Internacional de Valencia

Agrupamiento

Tipos de algoritmos de agrupamiento

- ▶ Basados en particiones
- ▶ Jerárquicos
- ▶ Espectrales
- ▶ **Basados en densidad**
- ▶ Probabilísticos

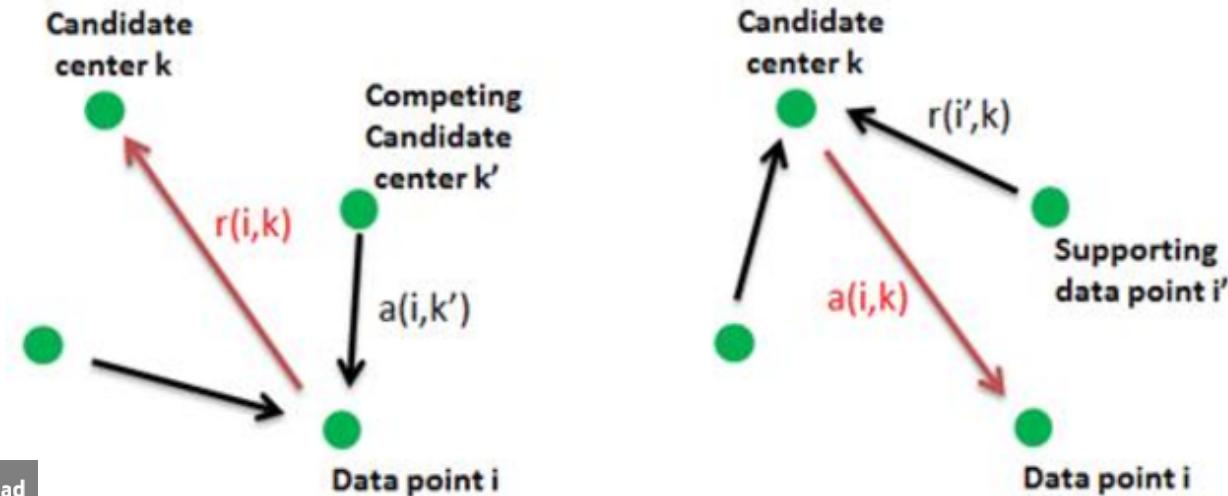


Agrupamiento basado en densidad

Idea

Pase de mensajes:

- ▶ Responsabilidad (R): evidencia de que x_k es el centroide más apropiado para x_i
- ▶ Disponibilidad (A): evidencia de lo apropiado que x_i escoja x_k como su centroide habiendo sido elegido como centroide por otros puntos



Agrupamiento basado en densidad

Inicialización

- ▶ Matriz de similitud (S): Similitud entre todos los pares de ejemplos del conjunto de datos
- ▶ Preferencias: creencia (autoconfianza) de que un ejemplo x_k puede ser centroide

Se modifica la **matriz de similitud S** para incorporar en su diagonal principal los valores de **preferencia** de cada ejemplo.

Agrupamiento basado en densidad

Responsabilidad

$$r(\mathbf{x}_i, \mathbf{x}_k) = s(\mathbf{x}_i, \mathbf{x}_k) - \max_{k' \neq k} (a(\mathbf{x}_i, \mathbf{x}_{k'}) + s(\mathbf{x}_i, \mathbf{x}_{k'}))$$

Agrupamiento basado en densidad

Disponibilidad

$$a(\mathbf{x}_i, \mathbf{x}_k) = \begin{cases} \min \left(0; r(\mathbf{x}_k, \mathbf{x}_k) + \sum_{i' \in \{1, \dots, n\}: i' \neq i \wedge i' \neq k} \max \left(0; r(\mathbf{x}_{i'}, \mathbf{x}_k) \right) \right), & i \neq k \\ \sum_{i' \in \{1, \dots, n\}: i' \neq k} \max \left(0; r(\mathbf{x}_{i'}, \mathbf{x}_k) \right), & i = k \end{cases}$$

Agrupamiento basado en densidad

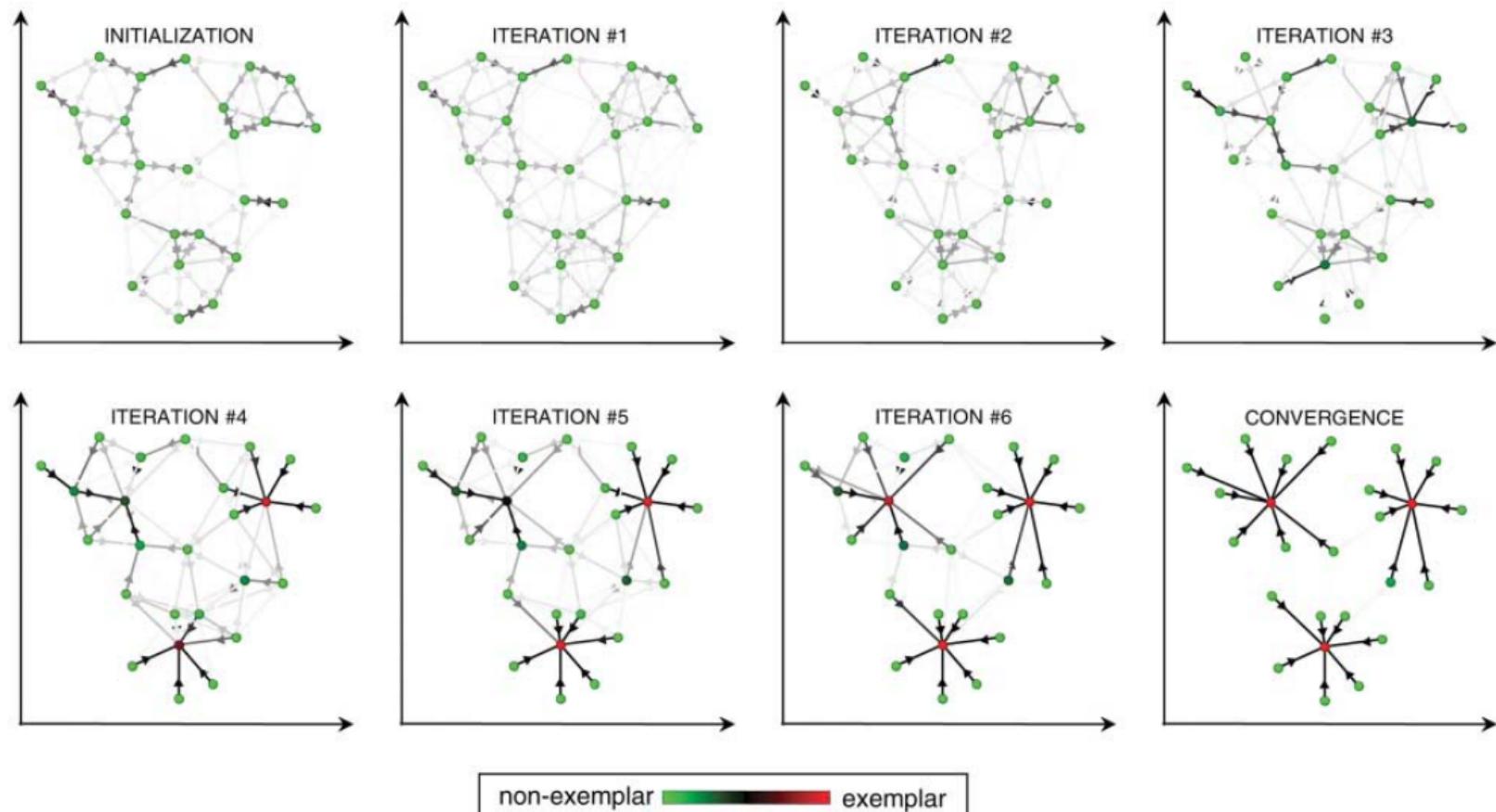
Affinity propagation

Algoritmo

1. Inicializar S con las preferencias
2. Inicializar A todo a cero
3. Repetir hasta convergencia:
 - 3.1 Actualizar R
 - 3.2 Actualizar A

Agrupamiento basado en densidad

Affinity propagation



Agrupamiento basado en densidad

Affinity propagation

Representantes o centroides

Aquellos puntos que tienen una responsabilidad y disponibilidad hacia ellos mismos positiva:

$$E \subset \{1, \dots, n\} : \forall k \in E, (r(\mathbf{x}_k, \mathbf{x}_k) + a(\mathbf{x}_k, \mathbf{x}_k)) > 0$$

Agrupamiento basado en densidad

Affinity propagation

Representantes o centroides

Aquellos puntos que tienen una responsabilidad y disponibilidad hacia ellos mismos positiva:

$$E \subset \{1, \dots, n\} : \forall k \in E, (r(\mathbf{x}_k, \mathbf{x}_k) + a(\mathbf{x}_k, \mathbf{x}_k)) > 0$$

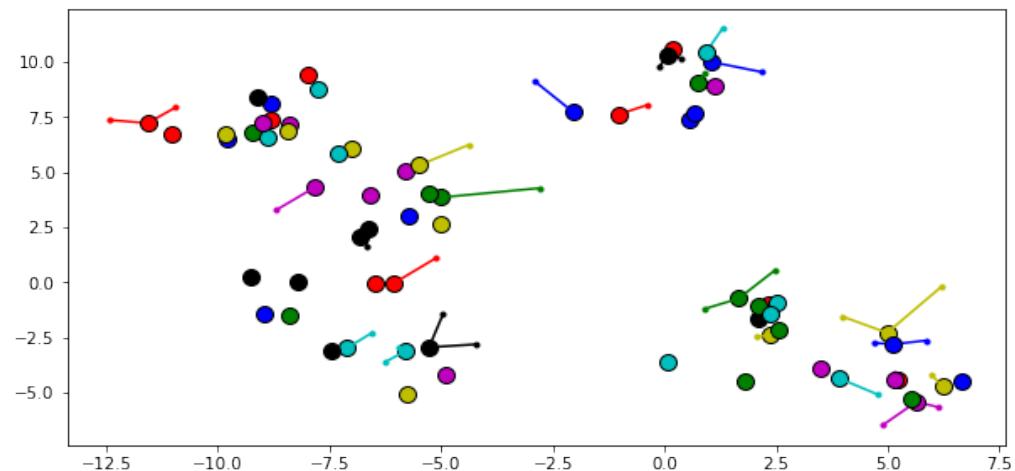
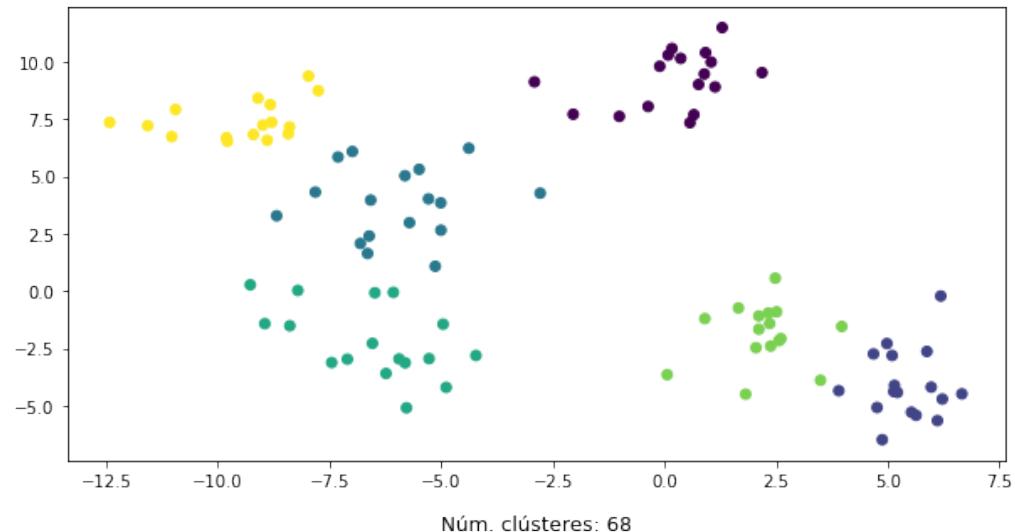
Asignación a clústeres (representantes)

Aquel representante que tiene una mayor responsabilidad y disponibilidad hacia el que tenemos asignado:

$$C(\mathbf{x}_i) = \arg \max_k (r(\mathbf{x}_i, \mathbf{x}_k) + a(\mathbf{x}_i, \mathbf{x}_k))$$

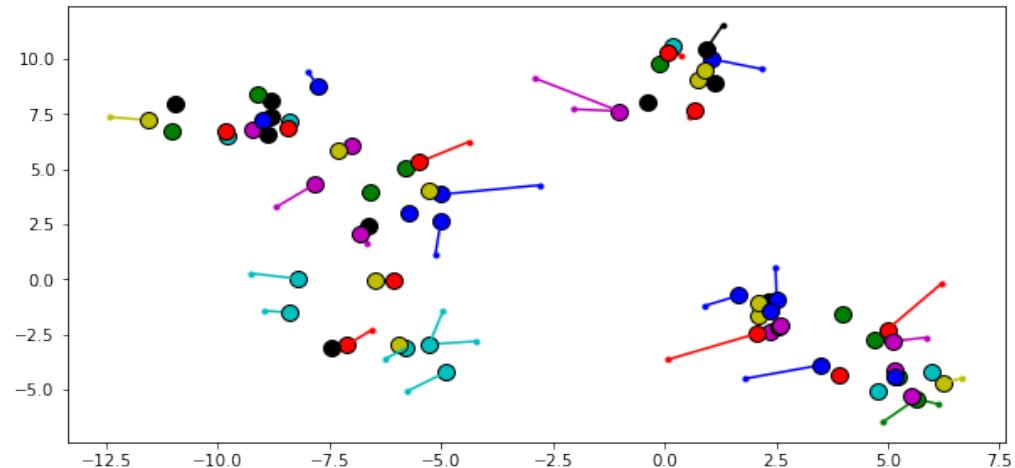
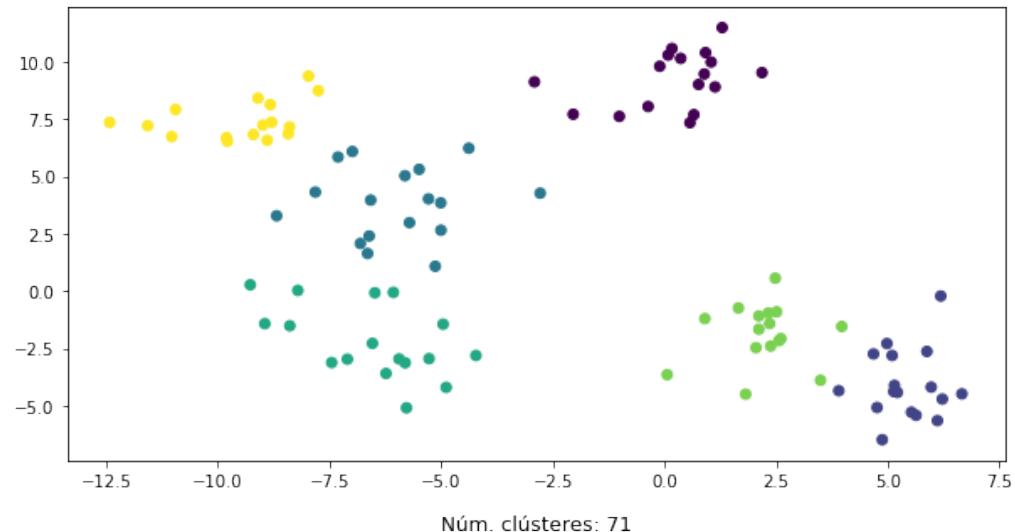
Agrupamiento basado en densidad

Mean-shift



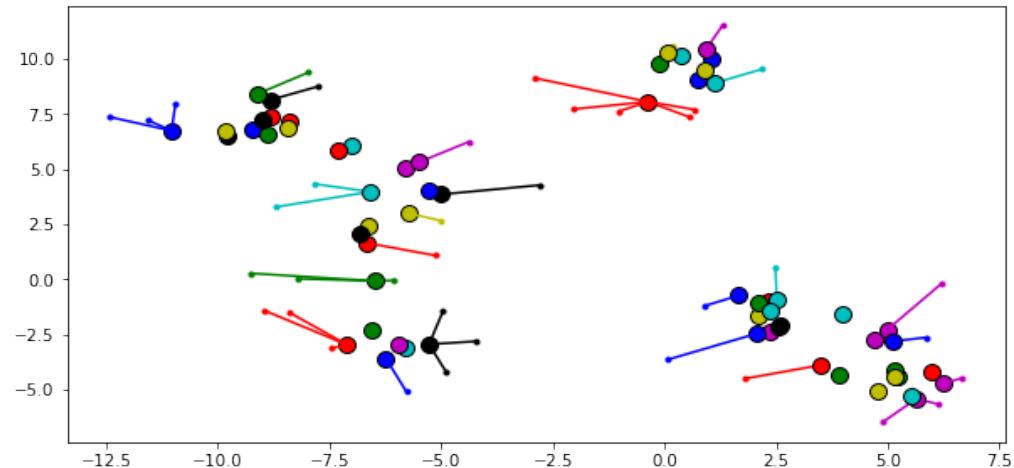
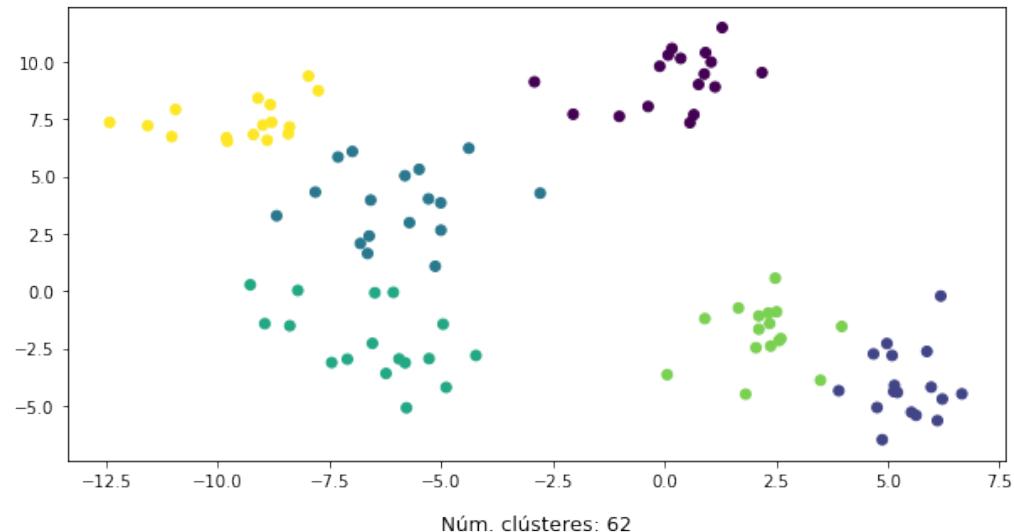
Agrupamiento basado en densidad

Mean-shift



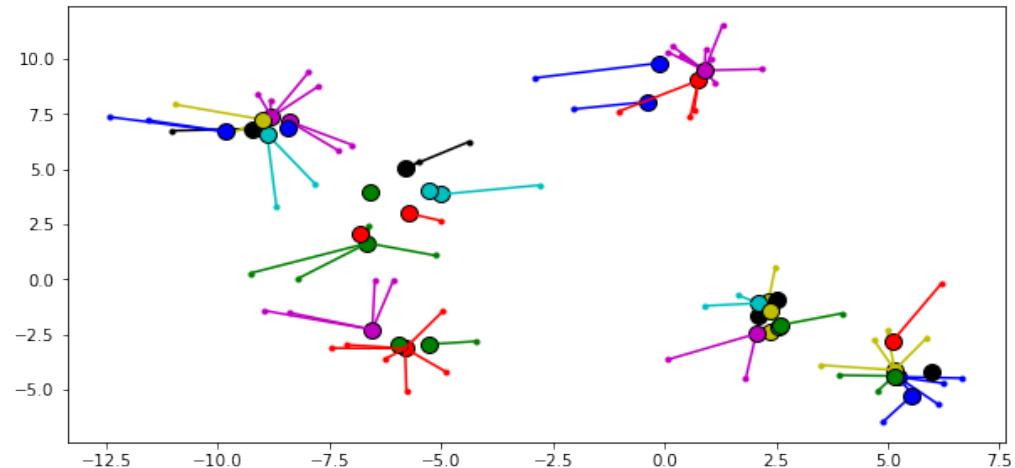
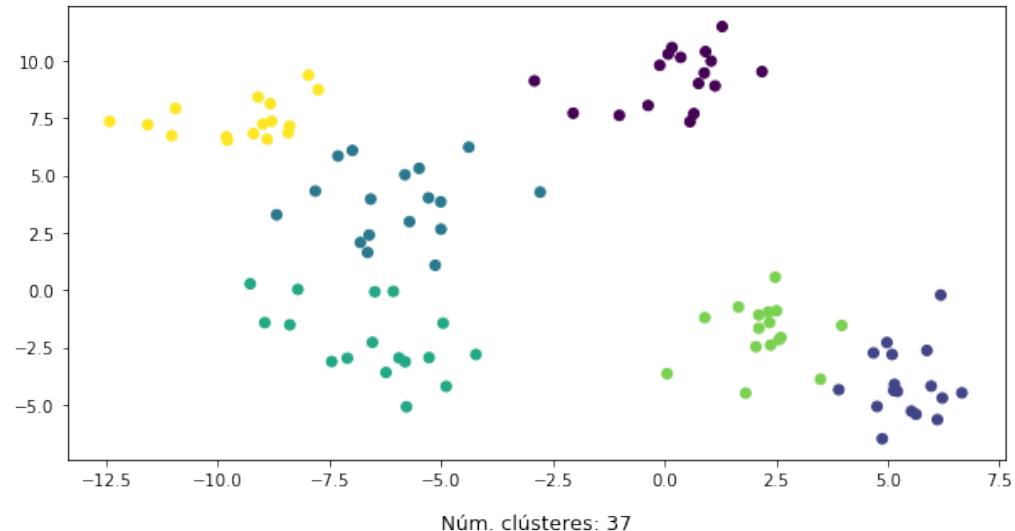
Agrupamiento basado en densidad

Mean-shift



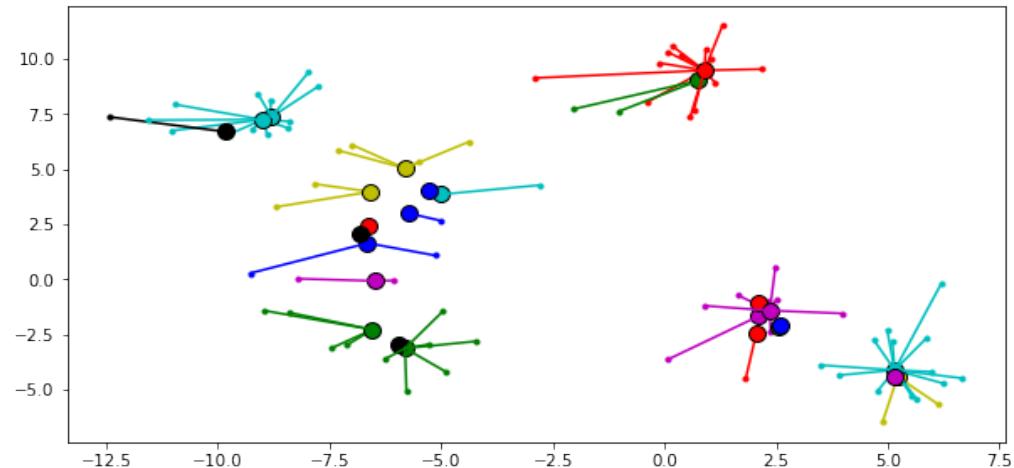
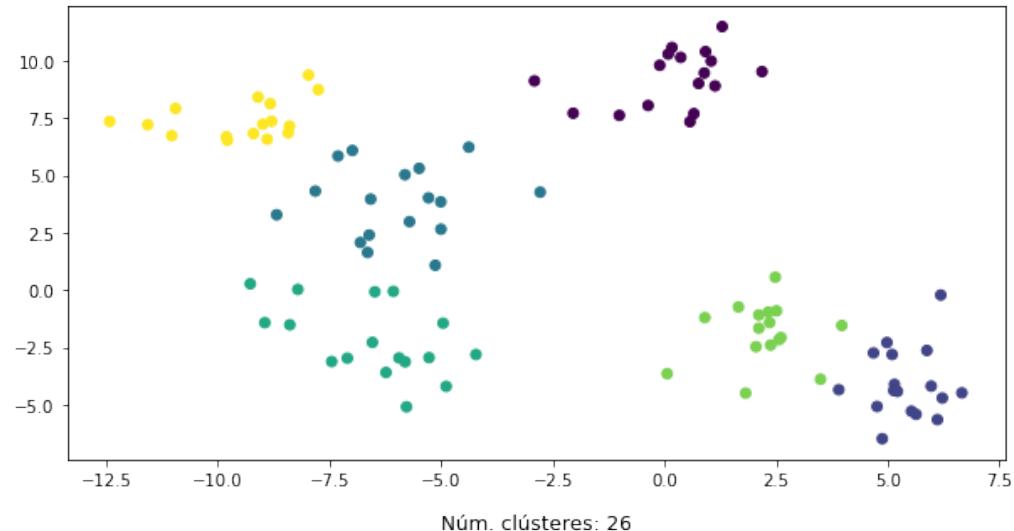
Agrupamiento basado en densidad

Mean-shift



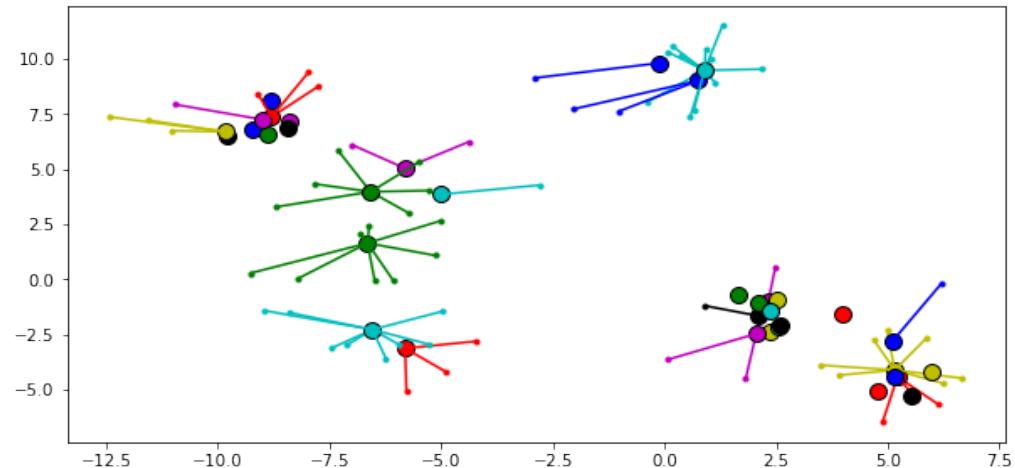
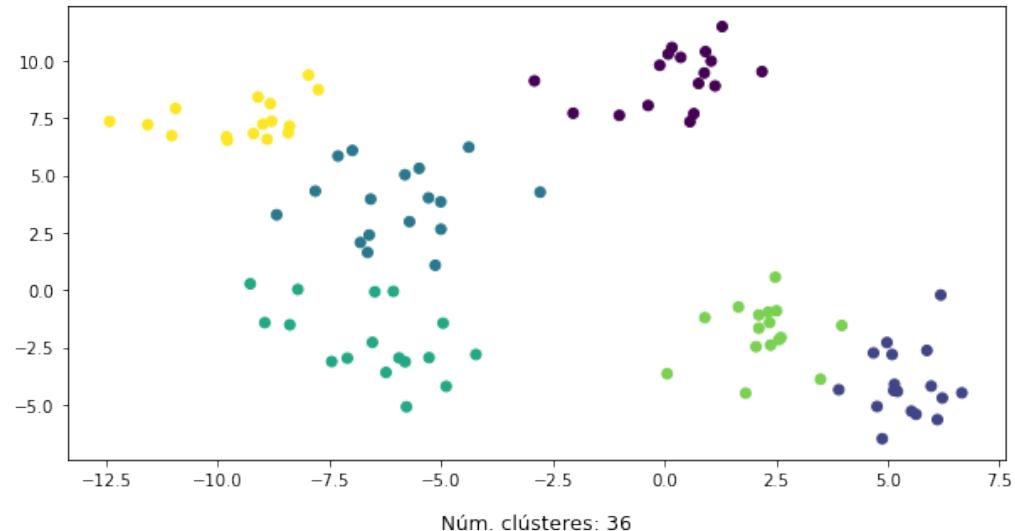
Agrupamiento basado en densidad

Mean-shift



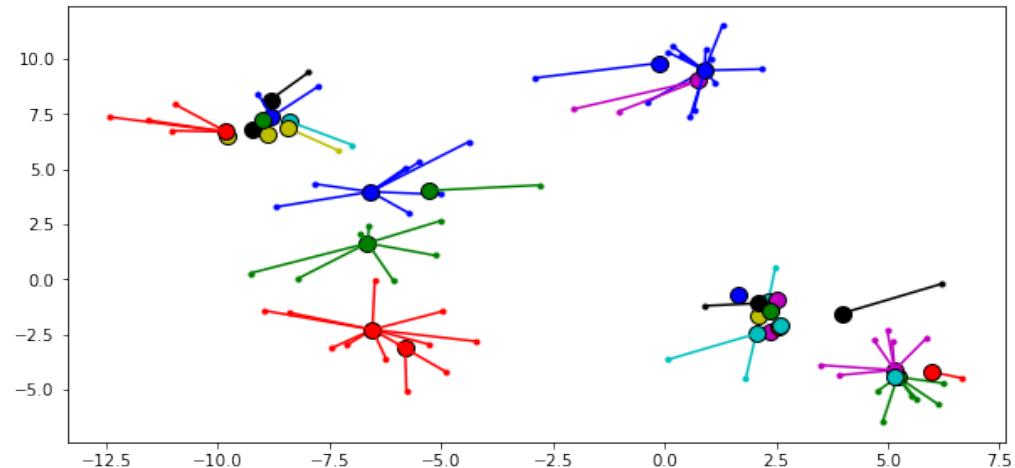
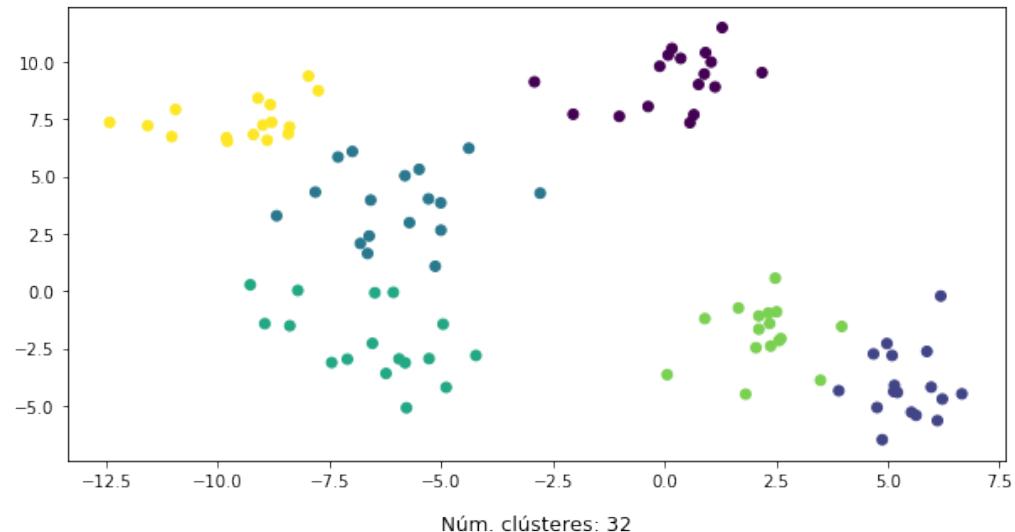
Agrupamiento basado en densidad

Mean-shift



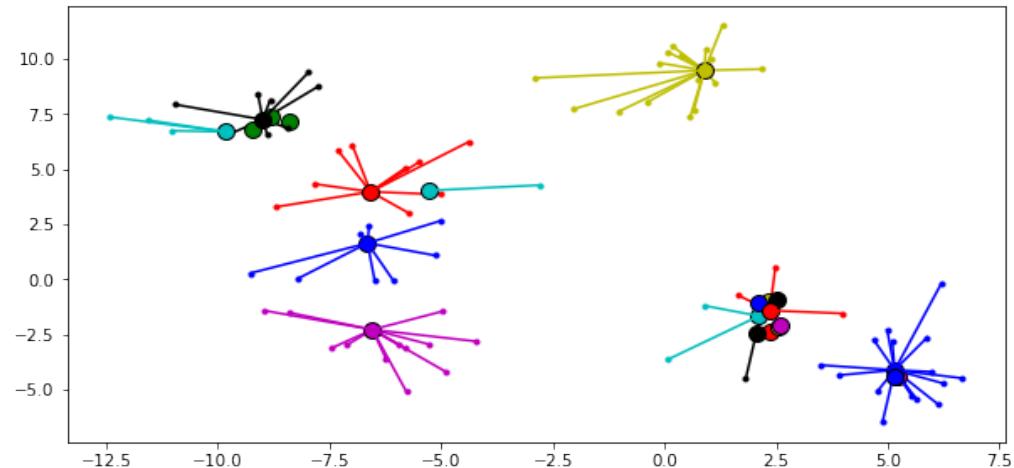
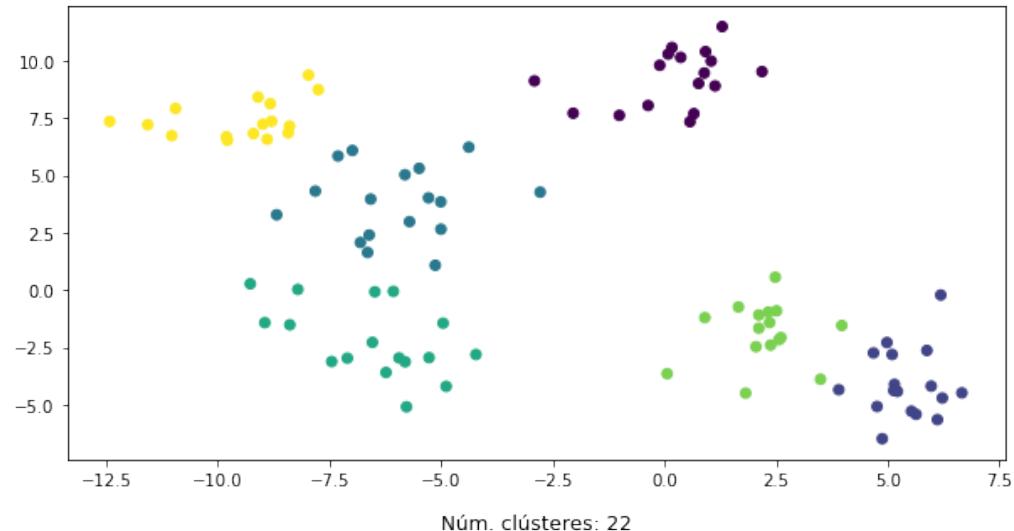
Agrupamiento basado en densidad

Mean-shift



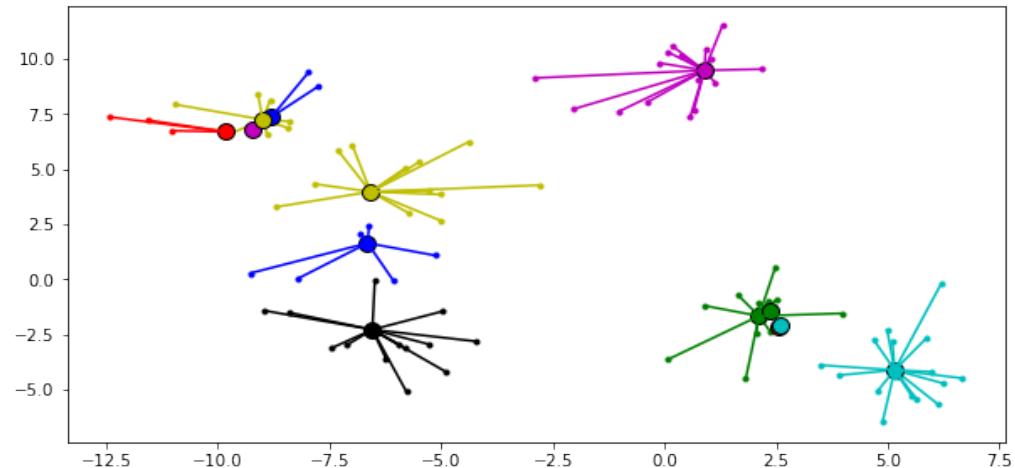
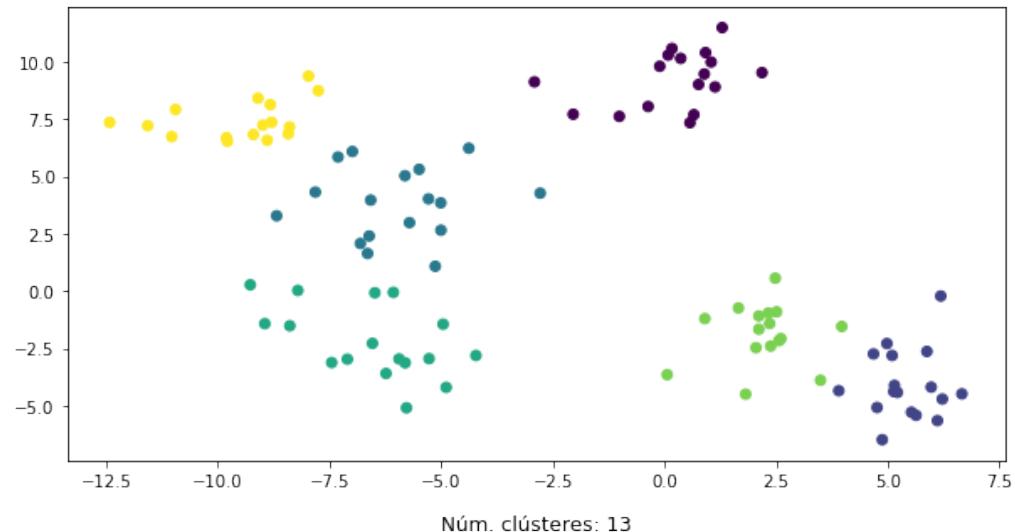
Agrupamiento basado en densidad

Mean-shift



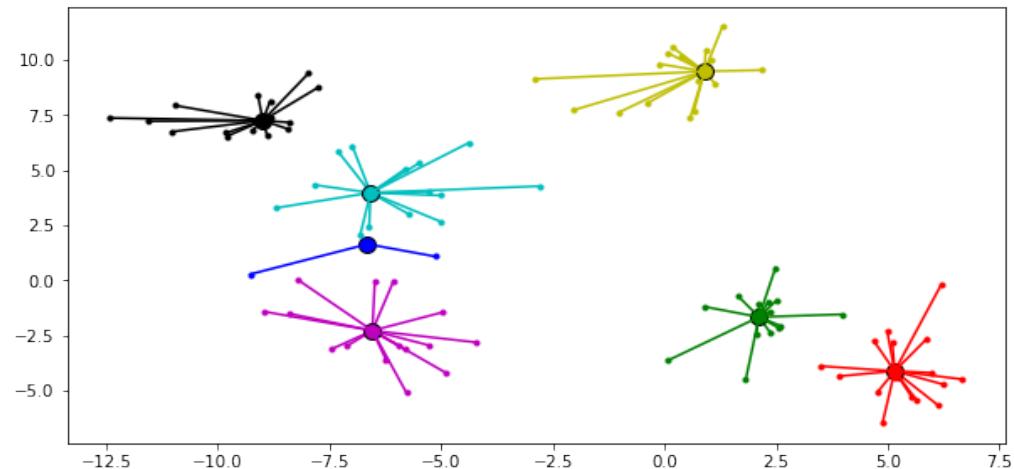
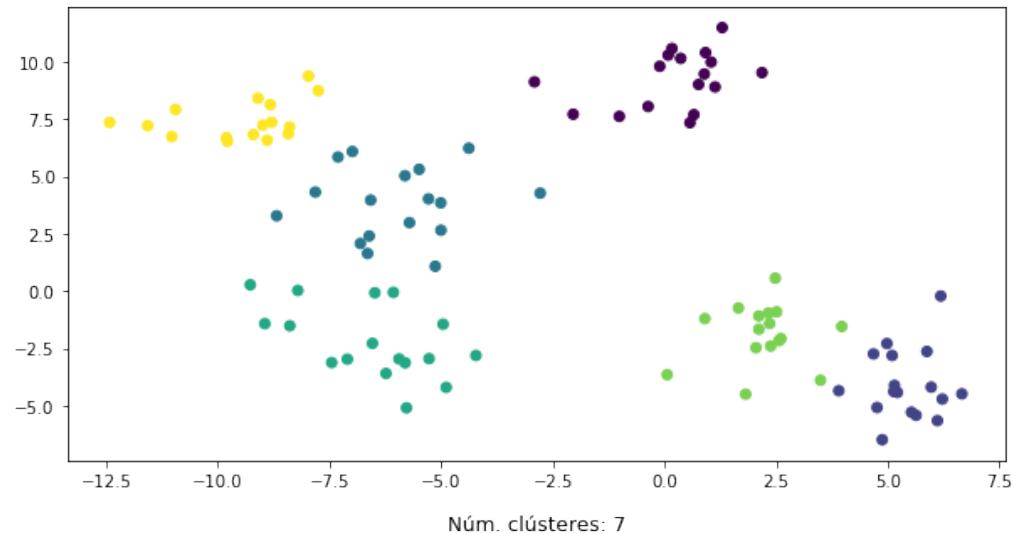
Agrupamiento basado en densidad

Mean-shift



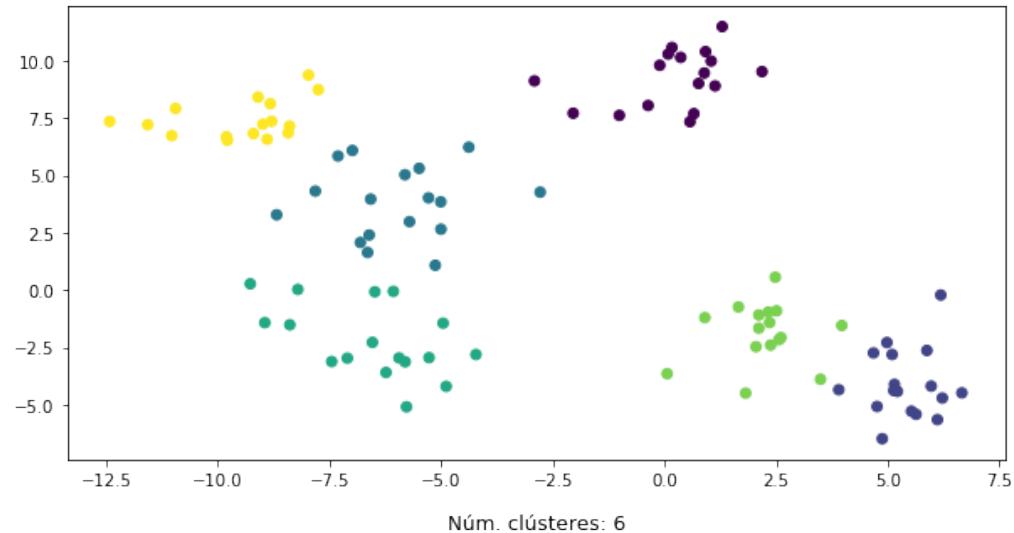
Agrupamiento basado en densidad

Mean-shift

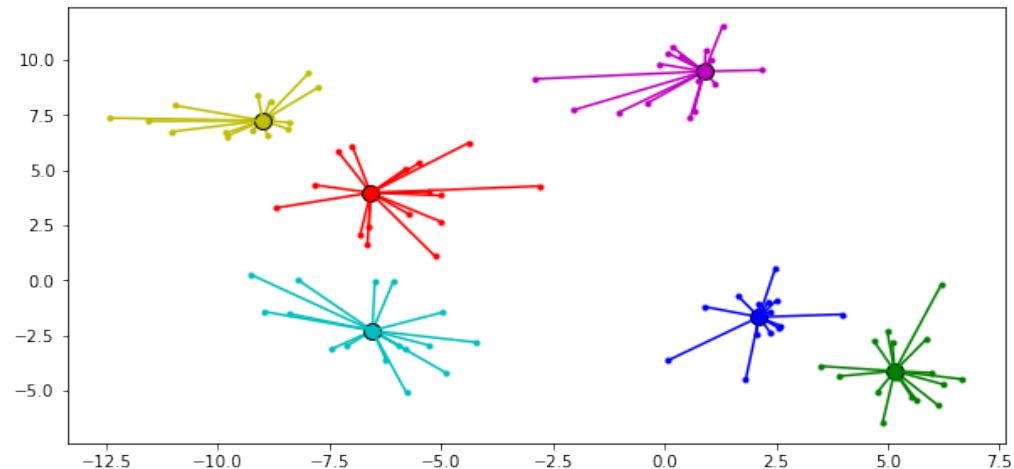


Agrupamiento basado en densidad

Mean-shift

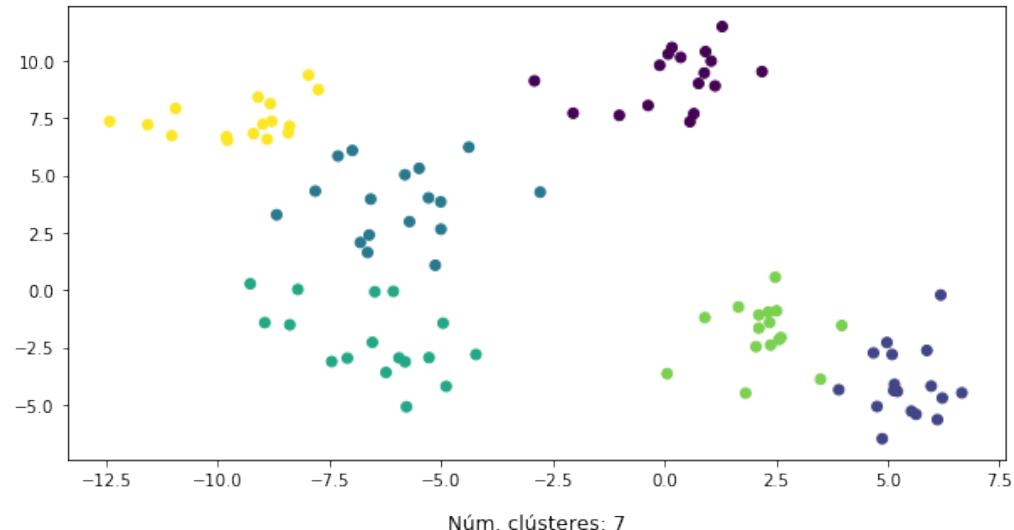


Núm. clústeres: 6

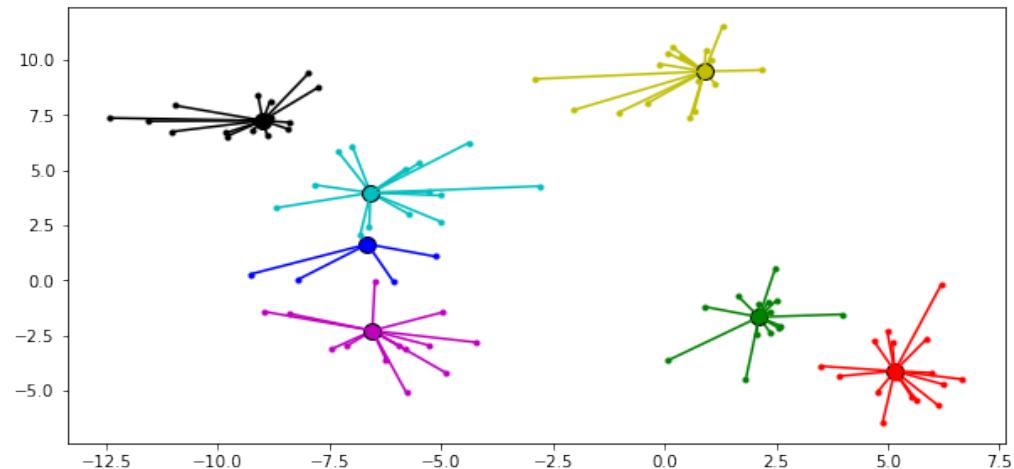


Agrupamiento basado en densidad

Mean-shift

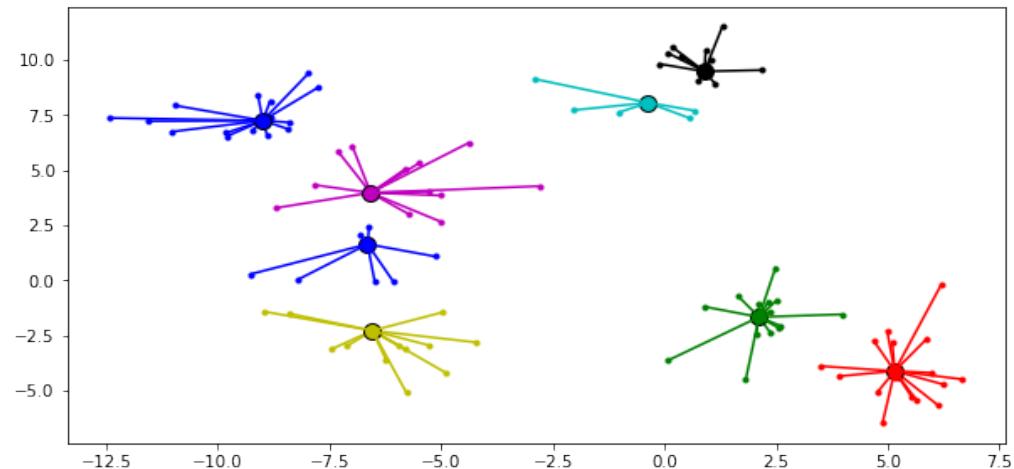
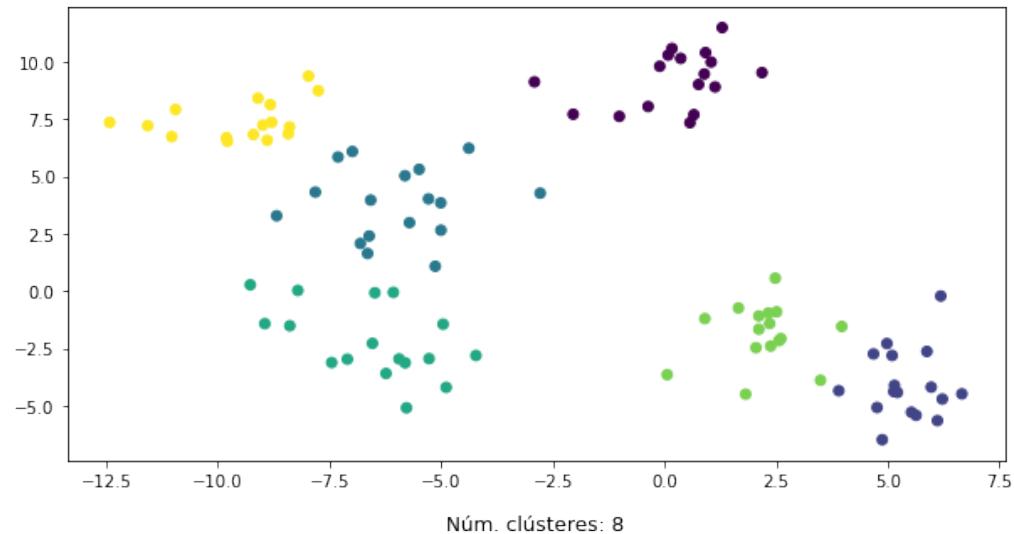


Núm. clústeres: 7



Agrupamiento basado en densidad

Mean-shift



Agrupamiento basado en densidad

Affinity propagation

Ventajas

- ▶ No es necesario especificar K
- ▶ Definición basada en similitud
- ▶ Podría incorporar diferentes medidas de similitud
- ▶ Funciona con clústeres de diferente tamaño

Agrupamiento basado en densidad

Affinity propagation

Desventajas

- ▶ Conceptualmente difícil de explicar
- ▶ Problemas para lidiar con clústeres de formas diversas
- ▶ Selección de las preferencias

Aprendizaje no supervisado

VC06: Agrupamiento basado en modelos probabilísticos – Mixtura de Gaussianas y Algoritmo EM

Rocío del Amor del Amor

mrocio.delamor@campusviu.es

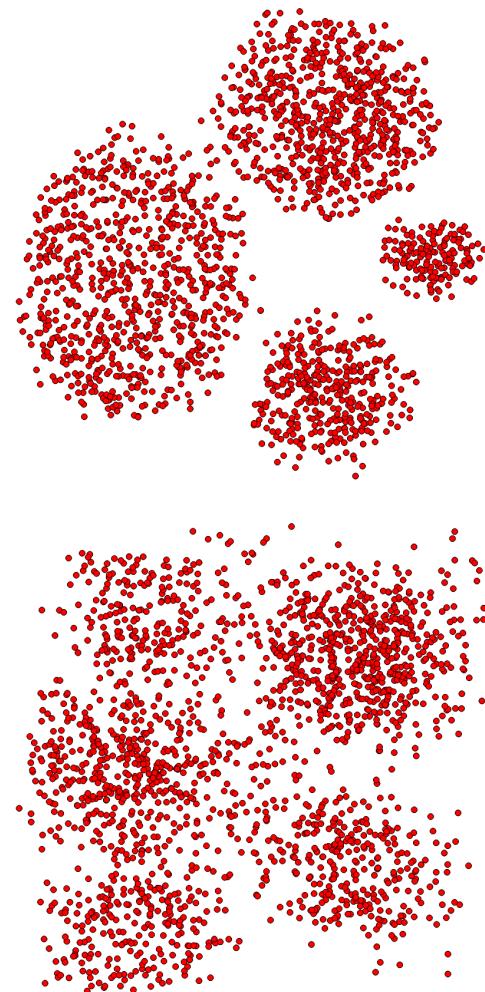
Universidad Internacional de Valencia

Uncertainty del modelo

Agrupamiento

Tipos de algoritmos de agrupamiento

- ▶ Basados en particiones
- ▶ Jerárquicos
- ▶ Espectrales
- ▶ Basados en densidad
- ▶ **Probabilísticos**



Agrupamiento

Agrupamiento probabilístico

- ▶ Se asume la existencia de un **modelo probabilístico** a partir del cual se han **generado** los **datos observados**.
- ▶ **Agrupamiento:** encontrar, a partir de los datos de entrenamiento, el mejor **ajuste del modelo generador** que se asume.
Procedimiento de estimación de los parámetros del modelo generador asumido
↳ media y varianza o media y matriz de covarianza.
- ▶ Se asume la existencia de una variable oculta (no observada) que **asigna** los **ejemplos a los clústeres**
Sin esta información, es necesario acudir a técnicas de estimación en presencia de datos incompletos (algoritmo EM)

↳ Expectation Maximization

Agrupamiento probabilístico

Conceptos básicos

► Variable aleatoria, X

Función que asigna un valor al resultado de un experimento aleatorio

No se conoce el valor que tomará al ser medida, pero se sabe cuál es la distribución de probabilidad asociada al conjunto de valores posibles

Discretas vs. continuas

Ej., cara obtenida al lanzar un dado (todos los valores, misma probabilidad 1/6)

Ej., peso de una persona tomada al azar en una población (escuela, pueblo, etc.)

Agrupamiento probabilístico

Conceptos básicos

► Variable aleatoria, X

Función que asigna un valor al resultado de un experimento aleatorio

No se conoce el valor que tomará al ser medida, pero se sabe cuál es la distribución de probabilidad asociada al conjunto de valores posibles

Discretas vs. continuas

Ej., cara obtenida al lanzar un dado (todos los valores, misma probabilidad 1/6)

Ej., peso de una persona tomada al azar en una población (escuela, pueblo, etc.)

► Distribución de probabilidad, $p(X; \theta)$

Función que asigna, a cada valor posible x de la variable X , la probabilidad de que se obtenga dicho valor

Discretas vs. continuas

Ej., prob. de las caras de un dado (discreta, 6 valores)

Ej., prob. del peso las personas de una población (continua, infinitos valores)

Ej., prob. del número de quejas diarias de un negocio (discreta, infinitos valores)

Agrupamiento probabilístico

Conceptos básicos

► Variable aleatoria, X

Función que asigna un valor al resultado de un experimento aleatorio

No se conoce el valor que tomará al ser medida, pero se sabe cuál es la distribución de probabilidad asociada al conjunto de valores posibles

Discretas vs. continuas

Ej., cara obtenida al lanzar un dado (todos los valores, misma probabilidad 1/6)

Ej., peso de una persona tomada al azar en una población (escuela, pueblo, etc.)

► Distribución de probabilidad, $p(X; \theta)$

Función que asigna, a cada valor posible x de la variable X , la probabilidad de que se obtenga dicho valor

Discretas vs. continuas



Ej., prob. de las caras de un dado (discreta, 6 valores)

Ej., prob. del peso las personas de una población (continua, infinitos valores)

Ej., prob. del número de quejas diarias de un negocio (discreta, infinitos valores)

Agrupamiento probabilístico

Conceptos básicos

► Variable aleatoria, X

Función que asigna un valor al resultado de un experimento aleatorio

No se conoce el valor que tomará al ser medida, pero se sabe cuál es la distribución de probabilidad asociada al conjunto de valores posibles

Discretas vs. continuas

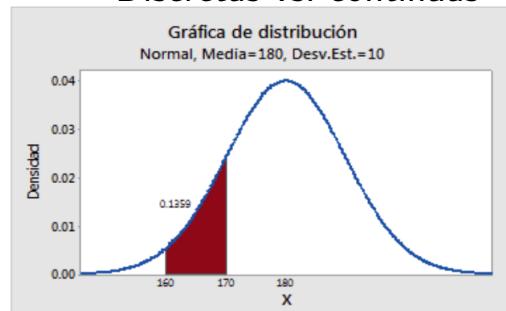
Ej., cara obtenida al lanzar un dado (todos los valores, misma probabilidad 1/6)

Ej., peso de una persona tomada al azar en una población (escuela, pueblo, etc.)

► Distribución de probabilidad, $p(X; \theta)$

Función que asigna, a cada valor posible x de la variable X , la probabilidad de que se obtenga dicho valor

Discretas vs. continuas



Ej., prob. de las caras de un dado (discreta, 6 valores)

Ej., prob. del peso las personas de una población (continua, infinitos valores)

Ej., prob. del número de quejas diarias de un negocio (discreta, infinitos valores)

Agrupamiento probabilístico

Conceptos básicos

► Variable aleatoria, X

Función que asigna un valor al resultado de un experimento aleatorio

No se conoce el valor que tomará al ser medida, pero se sabe cuál es la distribución de probabilidad asociada al conjunto de valores posibles

Discretas vs. continuas

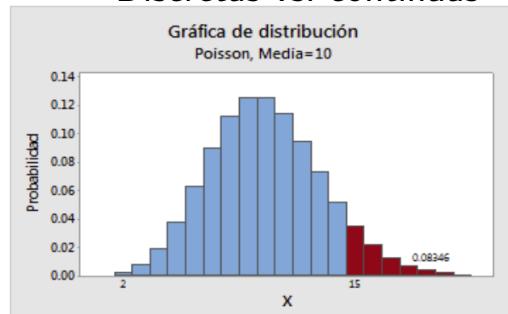
Ej., cara obtenida al lanzar un dado (todos los valores, misma probabilidad 1/6)

Ej., peso de una persona tomada al azar en una población (escuela, pueblo, etc.)

► Distribución de probabilidad, $p(X; \theta)$

Función que asigna, a cada valor posible x de la variable X , la probabilidad de que se obtenga dicho valor

Discretas vs. continuas



Ej., prob. de las caras de un dado (discreta, 6 valores)

Ej., prob. del peso las personas de una población (continua, infinitos valores)

Ej., prob. del número de quejas diarias de un negocio (discreta, infinitos valores)

Agrupamiento probabilístico

Conceptos

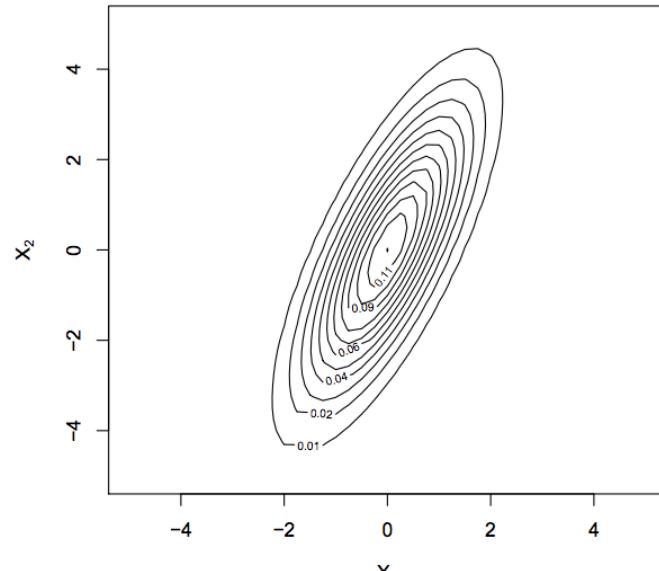
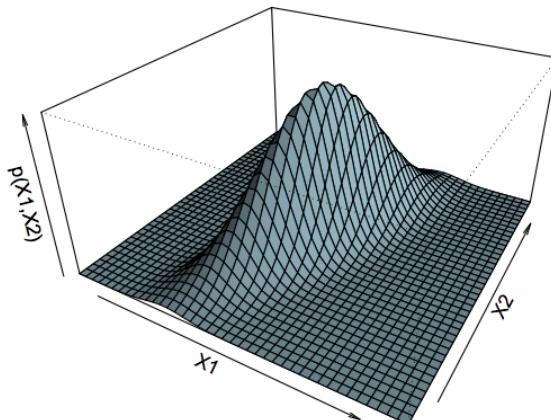
- ▶ Vector aleatorio (multi-variada), (X_1, \dots, X_v)

Función que asigna un vector de valores al resultado de un experimento aleatorio

La distribución de probabilidad asociada es multivariada, $p(X_1, \dots, X_v)$

Ej., coordenadas (x, y) en las que bota un balón al dejarlo caer desde un tercer piso

Ej., caras obtenidas al lanzar un dado y una moneda



Agrupamiento probabilístico

Conceptos

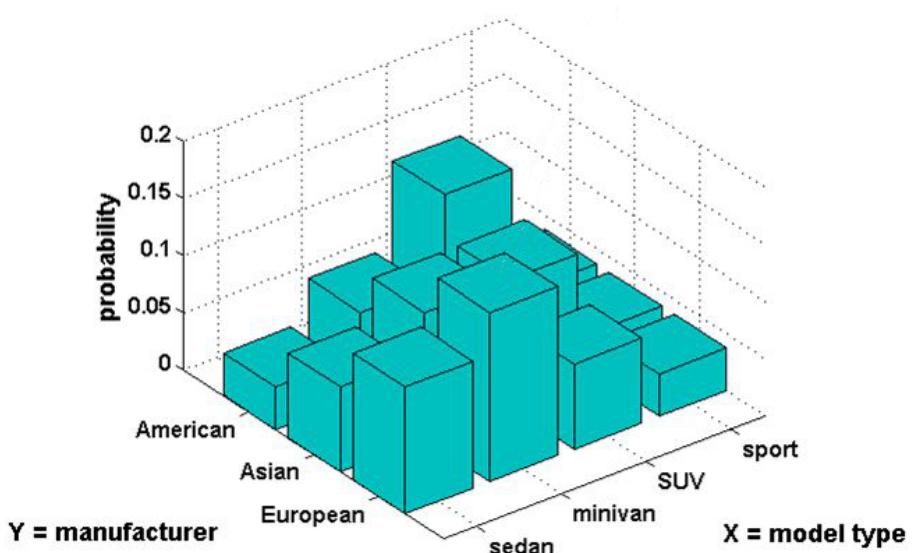
- ▶ Vector aleatorio (multi-variada), (X_1, \dots, X_v)

Función que asigna un vector de valores al resultado de un experimento aleatorio

La distribución de probabilidad asociada es multivariada, $p(X_1, \dots, X_v)$

Ej., coordenadas (x, y) en las que bota un balón al dejarlo caer desde un tercer piso

Ej., caras obtenidas al lanzar un dado y una moneda



Agrupamiento probabilístico

Conceptos

- ▶ Vector aleatorio (multi-variada), (X_1, \dots, X_v)

Función que asigna un vector de valores al resultado de un experimento aleatorio

La distribución de probabilidad asociada es multivariada, $p(X_1, \dots, X_v)$

Ej., coordenadas (x, y) en las que bota un balón al dejarlo caer desde un tercer piso

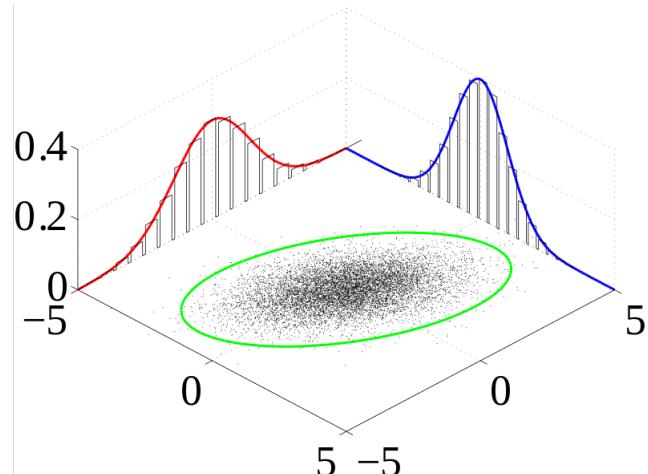
Ej., caras obtenidas al lanzar un dado y una moneda

- ▶ Modelo generativo

Es el modelo probabilístico que produjo un conjunto de datos dado

- ▶ Caso de una muestra

Es una instancia del modelo generativo. Un valor (vector) producto de muestrear la distribución de prob. generadora.



Agrupamiento probabilístico

Conceptos

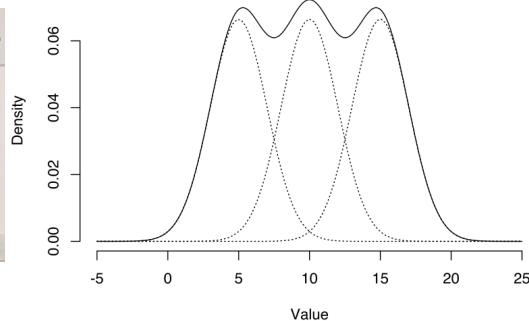
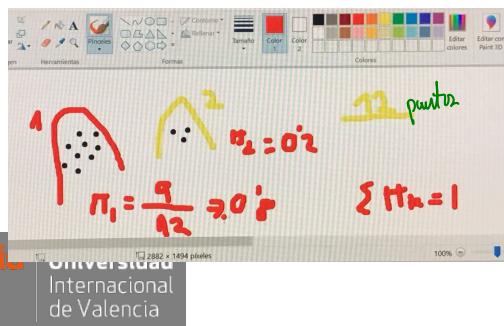
► Mixtura de distribuciones (Gaussianas) *conjunto de gaussianas.*

Función que asigna, a cada valor posible x de la variable X , la probabilidad de que se obtenga dicho valor

La probabilidad de cada valor se obtiene usando una serie de variables aleatorias (una de selección, y varias componentes):

- Una variable de selección que indica a qué componente pertenece (distribuida según $\{\pi_k\}_{k=1}^K$)
 - La variable de la componente seleccionada (distribuida según $p_k(x; \theta_k)$)
- número de gaussianas o "clusters"* *número de puntos que representa la gaussiana k*
- $$p(x) = \sum_{k=1}^K \pi_k \cdot p_k(x; \theta_k)$$
- probabilidad, dada la gaussiana k y dentro las parámetros de dicha gaussiana, de que el punto x haya sido generado por esa gaussiana*

Ej., edad de las personas de una escuela (alumnos/as, docentes)



Agrupamiento probabilístico

Conceptos

► Mixtura de distribuciones (Gaussianas)

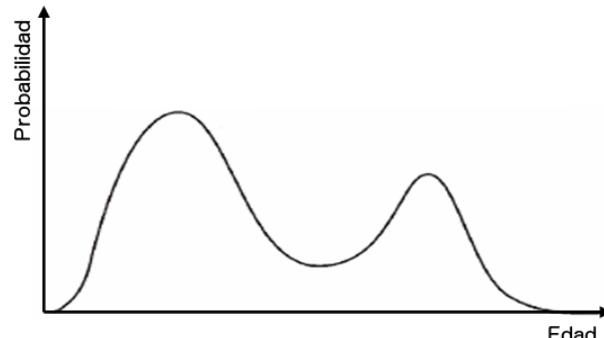
Función que asigna, a cada valor posible x de la variable X , la probabilidad de que se obtenga dicho valor

La probabilidad de cada valor se obtiene usando una serie de variables aleatorias (una de selección, y varias componentes):

- Una variable de selección que indica a qué componente pertenece (distribuida según $\{\pi_k\}_{k=1}^K$)
- La variable de la componente seleccionada (distribuida según $p_k(x; \theta_k)$)

$$p(x) = \sum_{k=1}^K \pi_k \cdot p_k(x; \theta_k)$$

Ej., edad de las personas de una escuela (alumnos/as, docentes)



Agrupamiento probabilístico

Mixtura de Gaussianas

Modelo generador de una Mixtura de Gaussianas sobre un vector aleatorio, (x_1, \dots, x_v) . La función de densidad de la mixtura es:

$$p(x) = \sum_{k=1}^K \pi_k \cdot p_k(x; \theta_k)$$

Coeficientes de mezcla, $\{\pi_1, \dots, \pi_K\}$: denotan la probabilidad de que un punto sea la instancia de cada una de las componentes. Cumplen que:

$$0 \leq \pi_k \leq 1, \forall k \in \{1, \dots, K\} \wedge \sum_{k=1}^K \pi_k = 1$$

La distribución de probabilidad de la k -ésima componente:

$$p_k(x; \theta_k) = \mathcal{N}(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^v |\Sigma|}} e^{-(x-\mu)^T \Sigma^{-1} (x-\mu)/2}$$

donde $\theta_k = \{\mu, \Sigma\}$ son los parámetros de la componente.

Agrupamiento probabilístico

Mixtura de Gaussianas

Modelo generador de una Mixtura de Gaussianas sobre un vector aleatorio, (x_1, \dots, x_v) . La función de densidad de la mixtura es:

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \cdot p_k(\mathbf{x}; \theta_k)$$

Coeficientes de mezcla, $\{\pi_1, \dots, \pi_K\}$: denotan la probabilidad de que un punto sea la instancia de cada una de las componentes. Cumplen que:

$$0 \leq \pi_k \leq 1, \forall k \in \{1, \dots, K\} \wedge \sum_{k=1}^K \pi_k = 1$$

La distribución de probabilidad de la k -ésima componente:

$$p_k(\mathbf{x}; \theta_k) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^v |\boldsymbol{\Sigma}|}} e^{-(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) / 2}$$

donde $\theta_k = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ son los parámetros de la componente.

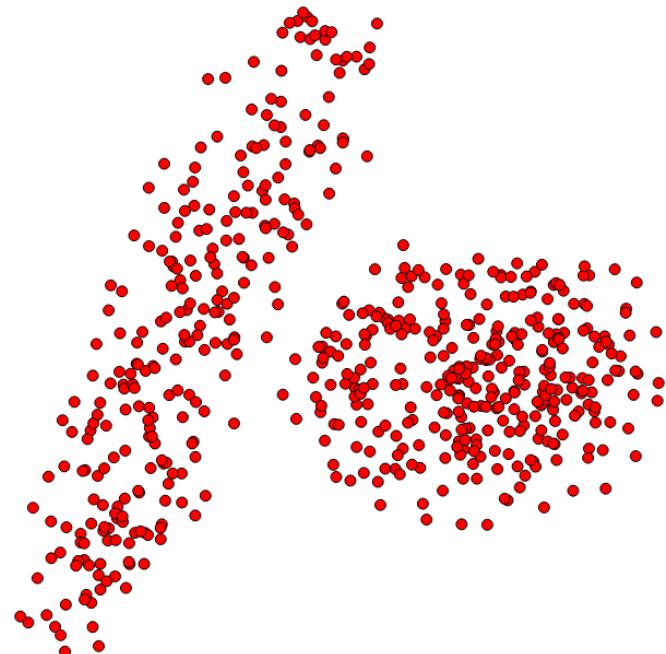
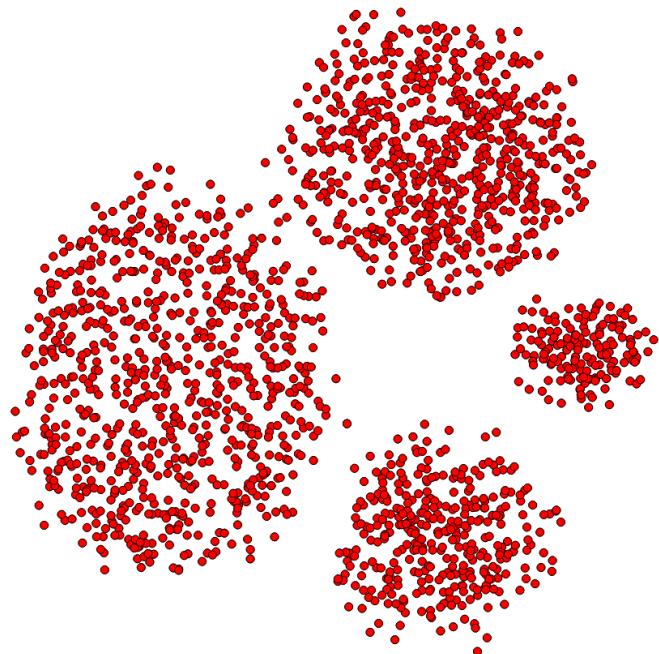
El proceso generador puede escalonarse en dos pasos:

- (i) seleccionar la componente k ; (ii) muestrear su distribución $p(x; \theta_k)$

Agrupamiento probabilístico

Mixtura de Gaussianas

Relación con el análisis de agrupamiento:



Agrupamiento probabilístico

Mixtura de Gaussianas

¿Qué queremos hacer?

- ▶ Plantear un modelo generador y entenderlo
- ▶ Estimar los parámetros del modelo generador dados los datos observados

Agrupamiento probabilístico

Mixtura de Gaussianas

La aproximación más habitual:

Asumir como modelo probabilístico generador de los datos una mixtura de distribuciones de probabilidad

- ▶ Mixtura de Gaussianas
- ▶ Cada componente (distribución) de la mixtura, un clúster
- ▶ **Incertidumbre**: asignación de los ejemplos a una u otra componente

- ▶ Un ejemplo pertenece a todos los clústeres (con distinta probabilidad)
- ▶ Al final, cada ejemplo pertenece al clúster cuya componente le otorga mayor probabilidad

Agrupamiento probabilístico

Mixtura de Gaussianas

La aproximación más habitual:

Asumir como modelo probabilístico generador de los datos una mixtura de distribuciones de probabilidad

- ▶ Mixtura de Gaussianas
- ▶ Cada componente (distribución) de la mixtura, un clúster
- ▶ **Incertidumbre**: asignación de los ejemplos a una u otra componente
 - *** Algoritmo Esperanza-Maximización (EM) ***
- ▶ Un ejemplo pertenece a todos los clústeres (con distinta probabilidad)
- ▶ Al final, cada ejemplo pertenece al clúster cuya componente le otorga mayor probabilidad

Agrupamiento probabilístico

Mixtura de Gaussianas

Proceso generador escalonado:

1. Muestrear $p(z)$ para seleccionar la componente k desde la que se generará la muestra
*** Existencia de una variable aleatoria latente que determina la selección de la componente a muestrear***
2. Muestrear la distribución de probabilidad seleccionada

$$p_k(\mathbf{x}; \theta_k) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \Sigma_k) \text{ with } \theta_k = \{\boldsymbol{\mu}_k, \Sigma_k\}$$

Agrupamiento probabilístico

Mixtura de Gaussianas

Asignación de casos a componentes

Variable aleatoria (multinomial con un único intento)

$$z \sim Mult(1, \{\pi_k\}_{k=1}^K)$$

Una muestra z es un vector K -dimensional con valor 1 para la componente seleccionada, y 0 para el resto

La distribución marginal de z es:

$$p(z) = \prod_{k=1}^K \pi_k^{z_k}$$

y la distribución condicional del caso x dado z :

$$p(x|z) = \prod_{k=1}^K p_k(x; \theta_k)^{z_k}$$

Agrupamiento probabilístico

Mixtura de Gaussianas

Modelo completo

Distribución marginal de \mathbf{x} :

$$p(\mathbf{x}) = \sum_z p(\mathbf{x}, z) = \sum_z p(\mathbf{x}|z) \cdot p(z) = \sum_{k=1}^K \pi_k \cdot p(\mathbf{x}; \theta_k)$$

La probabilidad condicionada de z dado \mathbf{x} (usando la regla de Bayes):

$$\begin{aligned} p(z_k = 1 | \mathbf{x}) &= \frac{p(\mathbf{x}|z_k = 1) \cdot p(z_k = 1)}{\sum_{k'=1}^K p(\mathbf{x}|z_{k'} = 1) \cdot p(z_{k'} = 1)} \\ &= \frac{\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \cdot \pi_k}{\sum_{k'=1}^K \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'}) \cdot \pi_{k'}} \end{aligned}$$

Agrupamiento probabilístico

Mixtura de Gaussianas

Aprendizaje

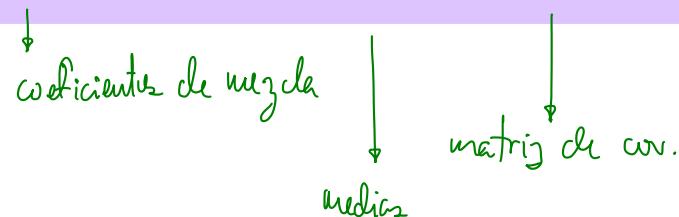
que los datos muestrados del modelo generador se parezcan lo máximo posible a los datos originales

Verosimilitud: Plausibilidad, dado un conjunto de datos, de un conjunto de parámetros Θ para un modelo.

Se calcula como la probabilidad asignada al conjunto de datos por el modelo parametrizado mediante Θ :

$$\mathcal{L}(\Theta | \{x_1, \dots, x_n\}) = \prod_{i=1}^n p(x_i; \Theta)$$

Parámetros del modelo: $\Theta = \{\pi_1, \dots, \pi_K, \mu_1, \dots, \mu_K, \Sigma_1, \dots, \Sigma_K\}$



Agrupamiento probabilístico

Mixtura de Gaussianas

Aprendizaje: estimación máximo verosímil

$$\begin{aligned}\Theta_{ML} &= \arg \max_{\Theta} \log \mathcal{L}(\Theta | \{\mathbf{x}_1, \dots, \mathbf{x}_n\}) \\ &= \arg \max_{\Theta} \sum_{i=1}^n \log p(\mathbf{x}_i; \Theta)\end{aligned}$$

Aprendizaje de parámetros máximo verosímiles: fórmula cerrada.

** Probabilidad de cada componente:

$$\hat{\pi}_k = \frac{\sum_{i=1}^n \hat{z}_{ik}}{n}$$

número de datos

Otra nota : *número de variables (o dimensiones) de un punto*
si: $\mathbf{X} = (\text{peso, edad, coordenadas})$

Agrupamiento probabilístico

Mixtura de Gaussianas

Aprendizaje: estimación máximo verosímil

$$\Theta_{ML} = \arg \max_{\Theta} \log \mathcal{L}(\Theta | \{x_1, \dots, x_n\})$$

$$= \arg \max_{\Theta} \sum_{i=1}^n \log p(x_i; \Theta)$$

Aprendizaje de parámetros máximo verosímiles: fórmula cerrada.

** Centro de la normal de la k -ésima componente:

$$\hat{\mu}_k = \frac{\sum_{i=1}^n \hat{z}_{ik} \cdot x_i}{\sum_{i=1}^n \hat{z}_{ik}}$$

↳ tiene la probabilidad de pertenencia del punto "i" a la gaussiana "k"

Ejemplo: $x_{-1} \rightarrow z_1 = \begin{pmatrix} k=1 \\ 0'8 \\ , \\ 0'2 \end{pmatrix}$

Agrupamiento probabilístico

Mixtura de Gaussianas

Aprendizaje: estimación máximo verosímil

$$\begin{aligned}\Theta_{ML} &= \arg \max_{\Theta} \log \mathcal{L}(\Theta | \{x_1, \dots, x_n\}) \\ &= \arg \max_{\Theta} \sum_{i=1}^n \log p(x_i; \Theta)\end{aligned}$$

Aprendizaje de parámetros máximo verosímiles: fórmula cerrada.

** Matriz de covarianza de la normal de la k -ésima componente:

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^n \hat{z}_{ik} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T}{\sum_{i=1}^n \hat{z}_{ik}}$$

Agrupamiento probabilístico

Mixtura de Gaussianas

Aprendizaje: estimación máximo verosímil

$$\Theta_{ML} = \arg \max_{\Theta} \log \mathcal{L}(\Theta | \{x_1, \dots, x_n\})$$

$$= \arg \max_{\Theta} \sum_{i=1}^n \log p(x_i; \Theta)$$

Aprendizaje de parámetros máximo verosímiles: fórmula cerrada.

Calcular la asignación prob. de cada ejemplo a componentes:

$$\hat{z}_{ik} = p(z_{ik} = 1 | x_i) = \frac{\hat{\pi}_k \cdot \mathcal{N}(x_i | \hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{k'=1}^K \hat{\pi}_{k'} \cdot \mathcal{N}(x_i | \hat{\mu}_{k'}, \hat{\Sigma}_{k'})}$$

[Examen]

¿Qué necesito calcular?

- Gaussianas (k) $\rightarrow \Theta = (\mu_k, \Sigma_k, \pi_k)$; $\sum_{i=1}^K \pi_k = 1$
- z_k = probabilidad de pertenencia a la gaussiana "k"

Agrupamiento probabilístico

Mixtura de Gaussianas

Aprendizaje: inter-dependencia

Calcular asignación prob. de cada ejemplo a componentes:

$$\hat{z}_{ik} = \hat{\pi}_k \cdot \mathcal{N}(\mathbf{x}_i | \hat{\mu}_k, \hat{\Sigma}_k) / \left(\sum_{k'=1}^K \hat{\pi}_{k'} \cdot \mathcal{N}(\mathbf{x}_i | \hat{\mu}_{k'}, \hat{\Sigma}_{k'}) \right)$$

Aprender los parámetros:

- Probabilidad de cada componente:

$$\hat{\pi}_k = \sum_{i=1}^n \hat{z}_{ik} / n$$

- Centro de la normal de la k -ésima componente:

$$\hat{\mu}_k = \left(\sum_{i=1}^n \hat{z}_{ik} \cdot \mathbf{x}_i \right) / \left(\sum_{i=1}^n \hat{z}_{ik} \right)$$

- Matriz de covarianza de la normal de la k -ésima componente:

$$\hat{\Sigma}_k = \left(\sum_{i=1}^n \hat{z}_{ik} (\mathbf{x}_i - \hat{\mu}_k)(\mathbf{x}_i - \hat{\mu}_k)^T \right) / \left(\sum_{i=1}^n \hat{z}_{ik} \right)$$

Agrupamiento probabilístico

Mixtura de Gaussianas

Aprendizaje: algoritmo Esperanza-Maximización

Paso E: Calcular asignación prob. de cada ejemplo a componentes:

$$\hat{z}_{ik} = \hat{\pi}_k \cdot \mathcal{N}(\mathbf{x}_i | \hat{\mu}_k, \hat{\Sigma}_k) / \left(\sum_{k'=1}^K \hat{\pi}_{k'} \cdot \mathcal{N}(\mathbf{x}_i | \hat{\mu}_{k'}, \hat{\Sigma}_{k'}) \right)$$

calcula μ y Σ aleatorio

Paso M: Aprender los parámetros

- Probabilidad de cada componente:

$$\hat{\pi}_k = \sum_{i=1}^n \hat{z}_{ik} / n$$

con ese calculo z_{ik}

- Centro de la normal de la k -ésima componente:

$$\hat{\mu}_k = \left(\sum_{i=1}^n \hat{z}_{ik} \cdot \mathbf{x}_i \right) / \left(\sum_{i=1}^n \hat{z}_{ik} \right)$$

- Matriz de covarianza de la normal de la k -ésima componente:

$$\hat{\Sigma}_k = \left(\sum_{i=1}^n \hat{z}_{ik} (\mathbf{x}_i - \hat{\mu}_k)(\mathbf{x}_i - \hat{\mu}_k)^T \right) / \left(\sum_{i=1}^n \hat{z}_{ik} \right)$$

Agrupamiento probabilístico

Mixtura de Gaussianas

Aprendizaje: algoritmo Esperanza-Maximización

Inicialización aleatoria

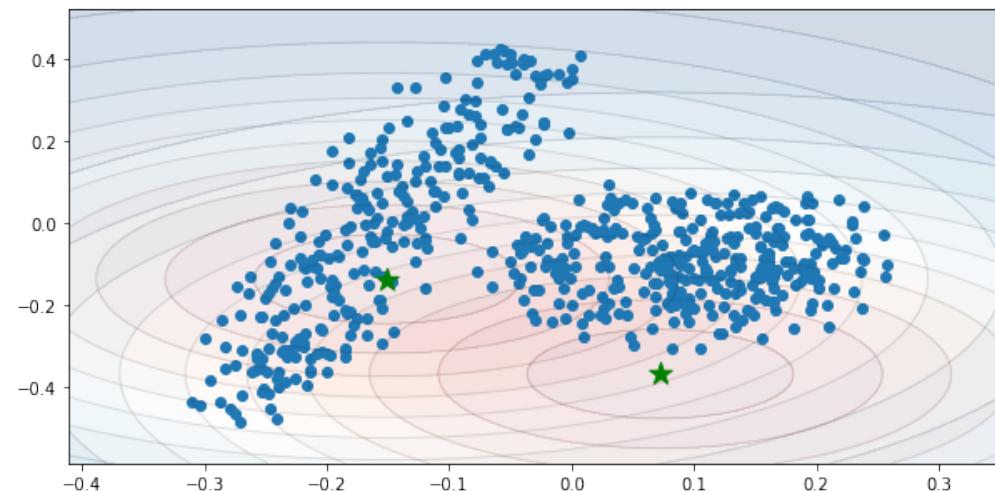
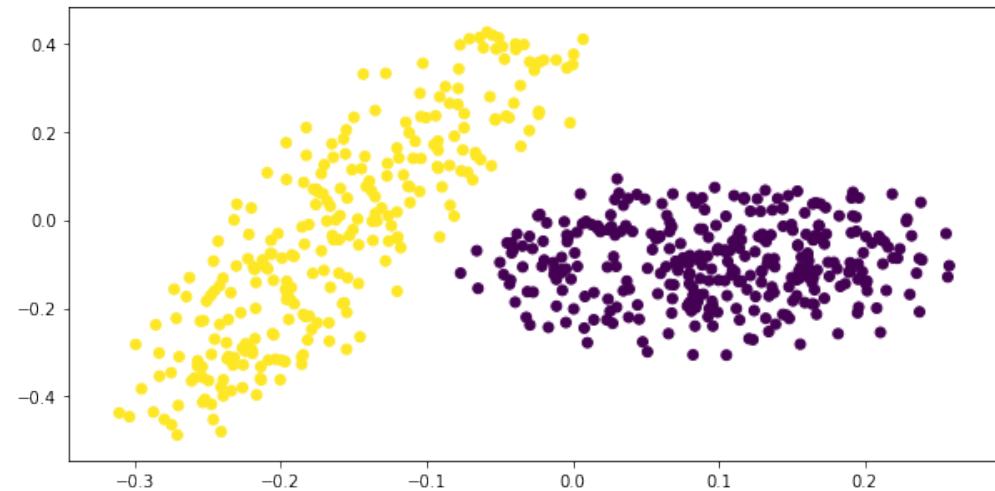
Paso E: Calcular asignación prob. de cada ejemplo a componentes

Paso M: Aprender los parámetros

Convergencia a un óptimo local de la función de verosimilitud

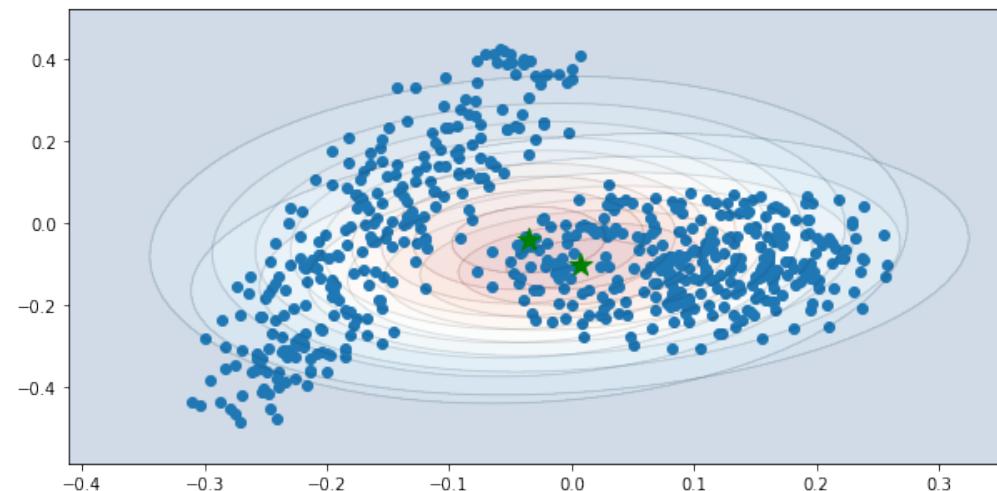
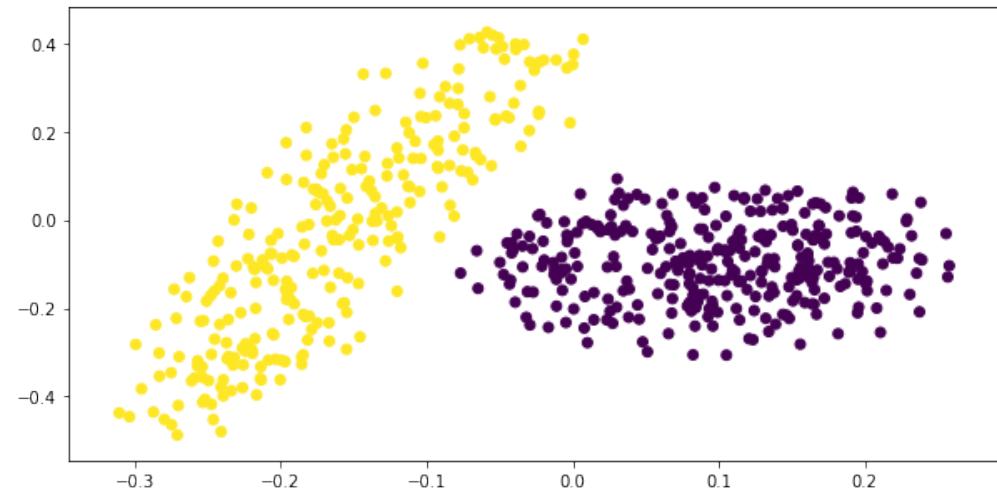
Agrupamiento probabilístico

Mixtura de Gaussianas



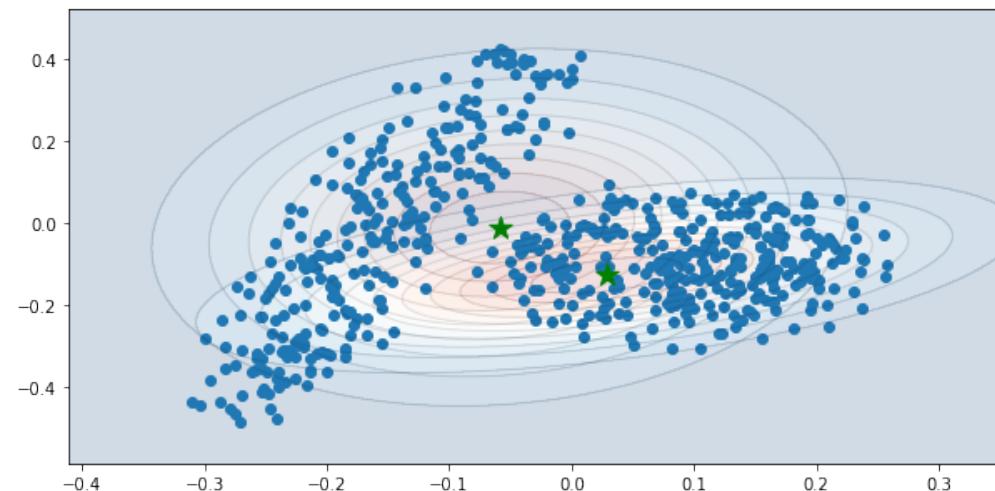
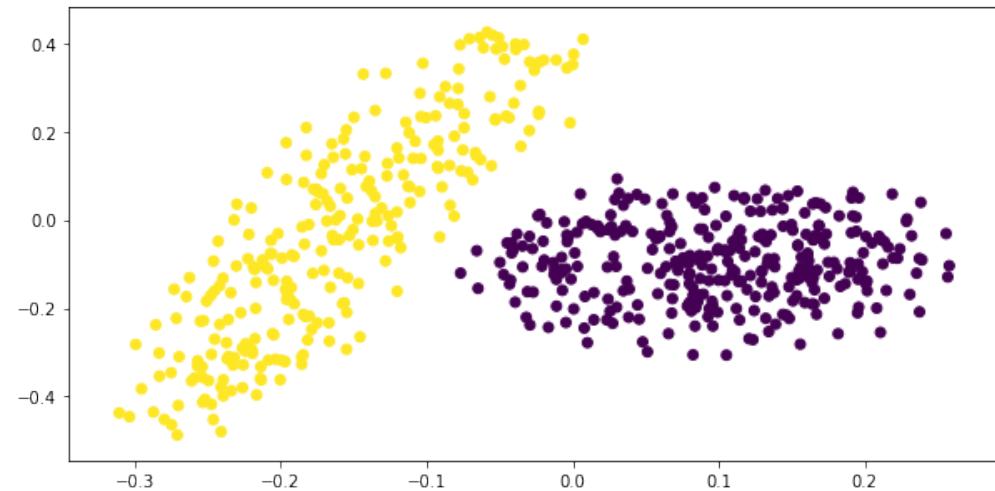
Agrupamiento probabilístico

Mixtura de Gaussianas



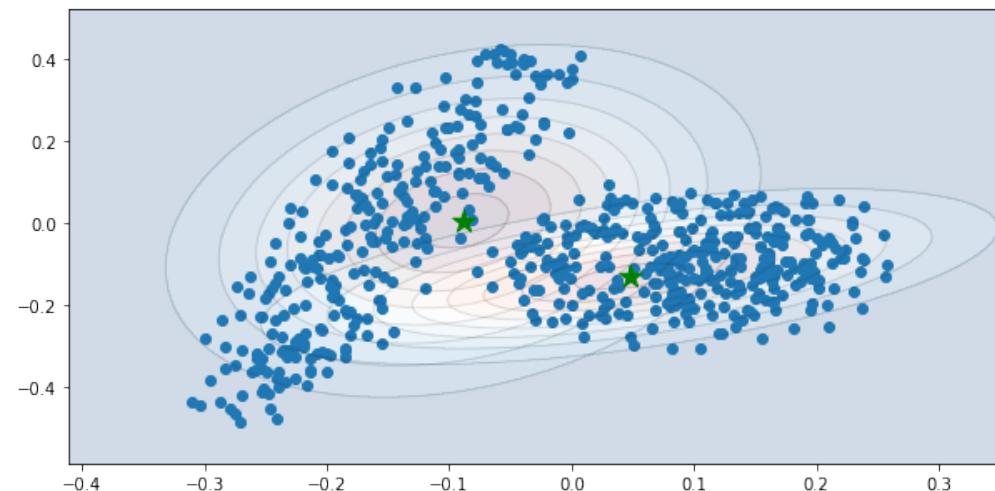
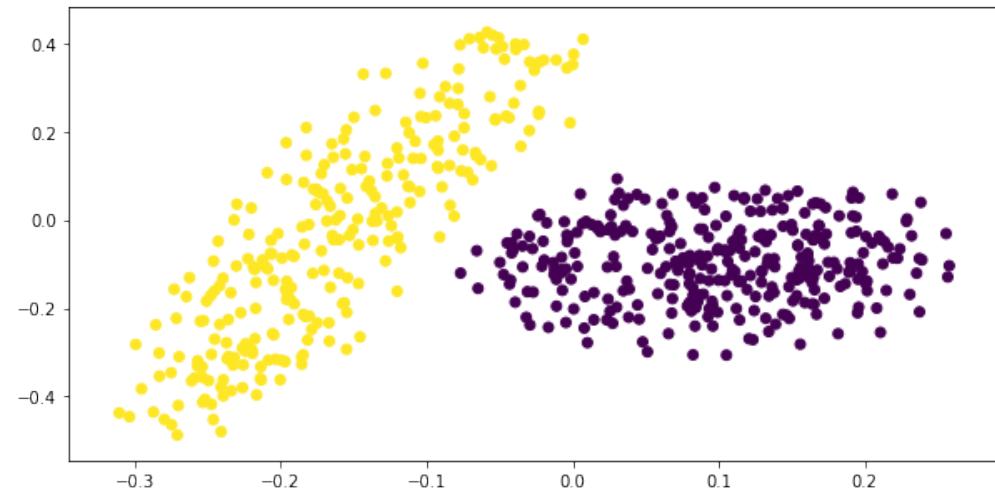
Agrupamiento probabilístico

Mixtura de Gaussianas



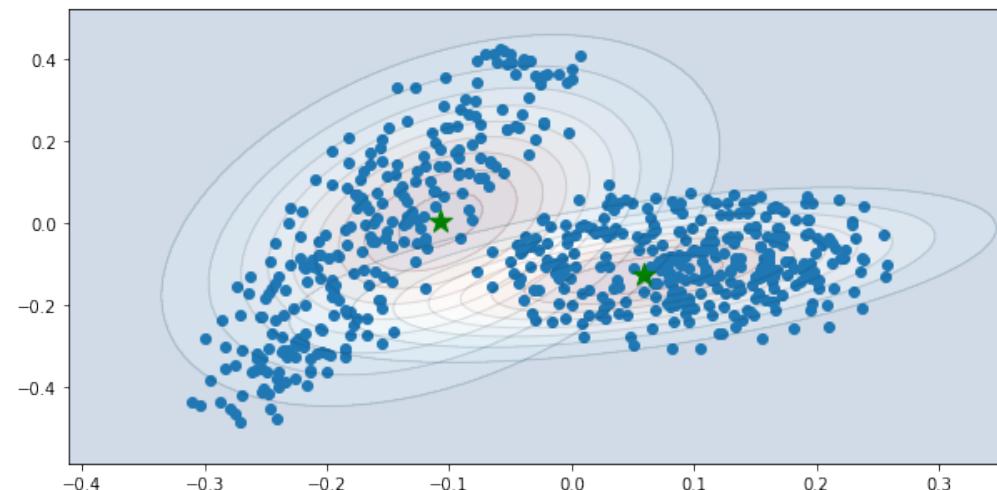
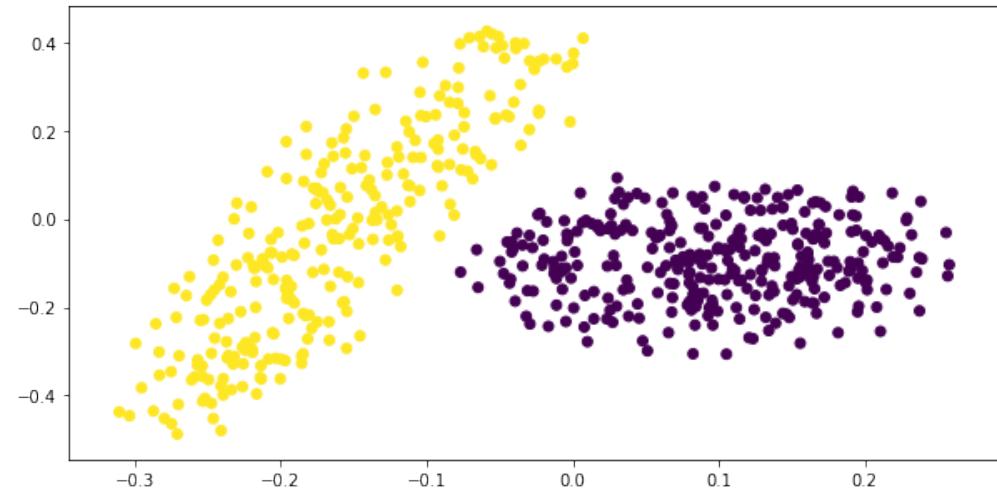
Agrupamiento probabilístico

Mixtura de Gaussianas



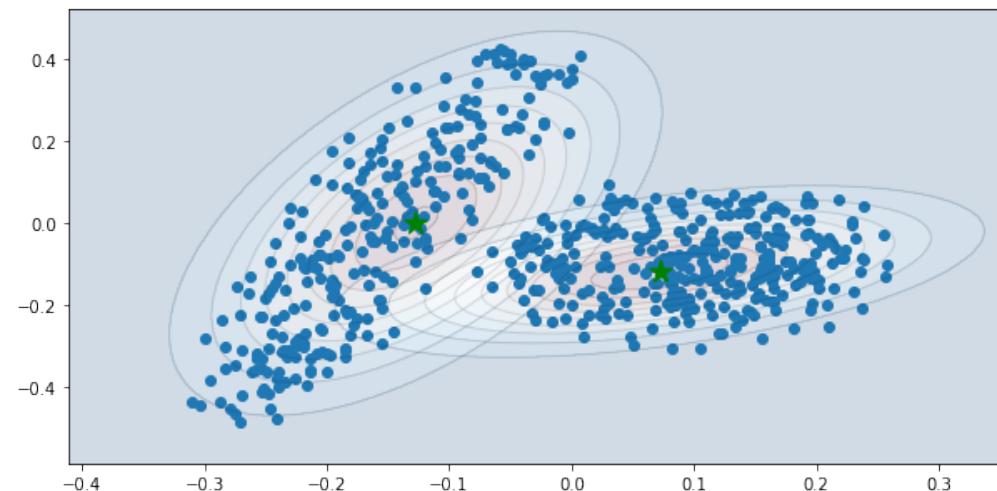
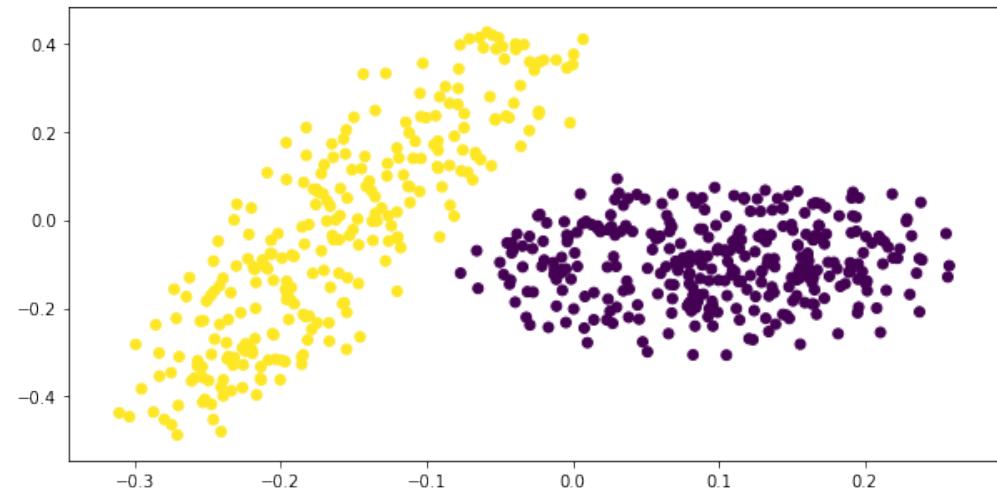
Agrupamiento probabilístico

Mixtura de Gaussianas



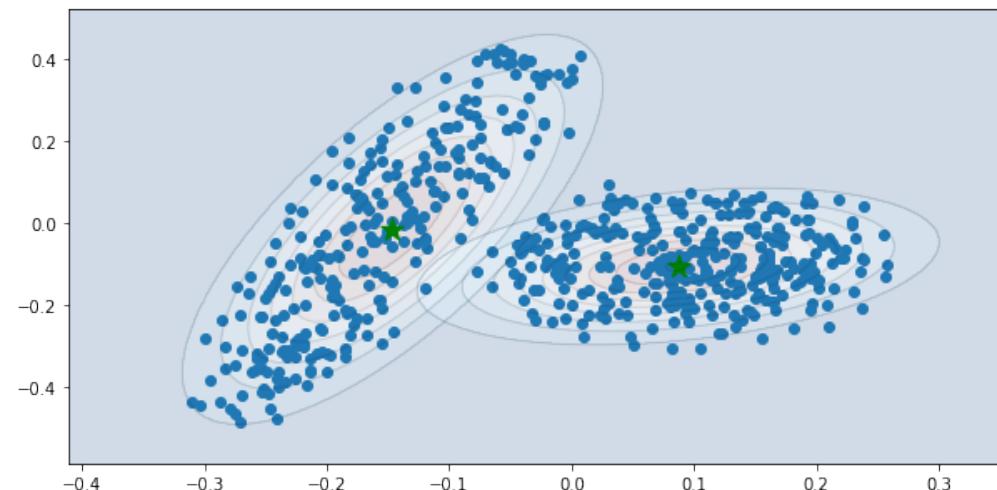
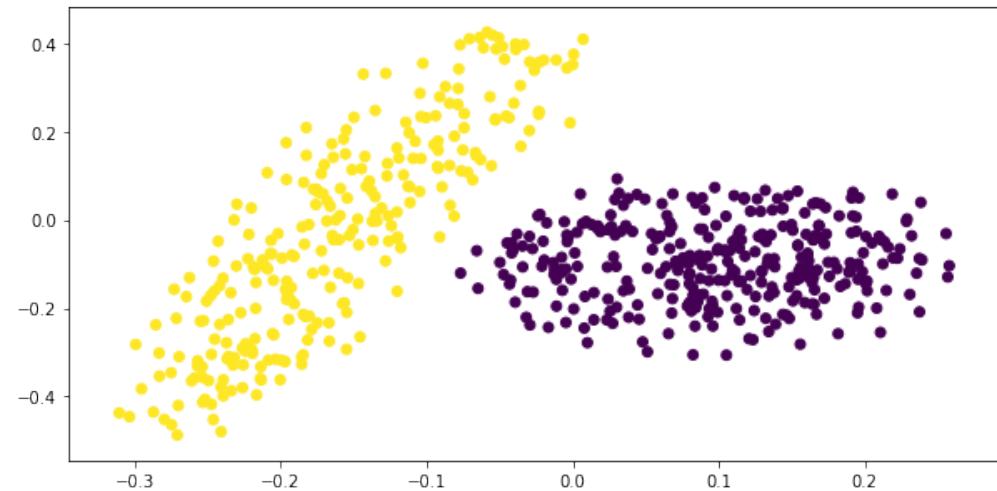
Agrupamiento probabilístico

Mixtura de Gaussianas



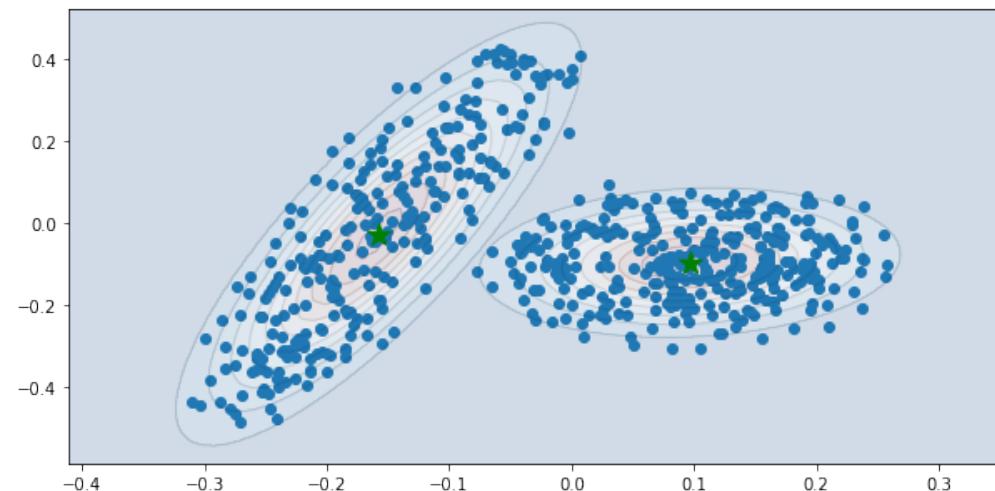
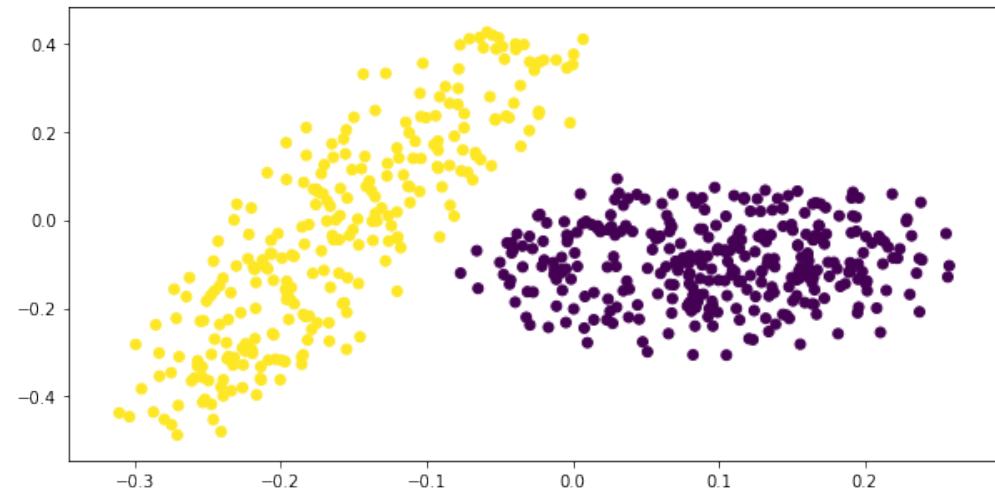
Agrupamiento probabilístico

Mixtura de Gaussianas



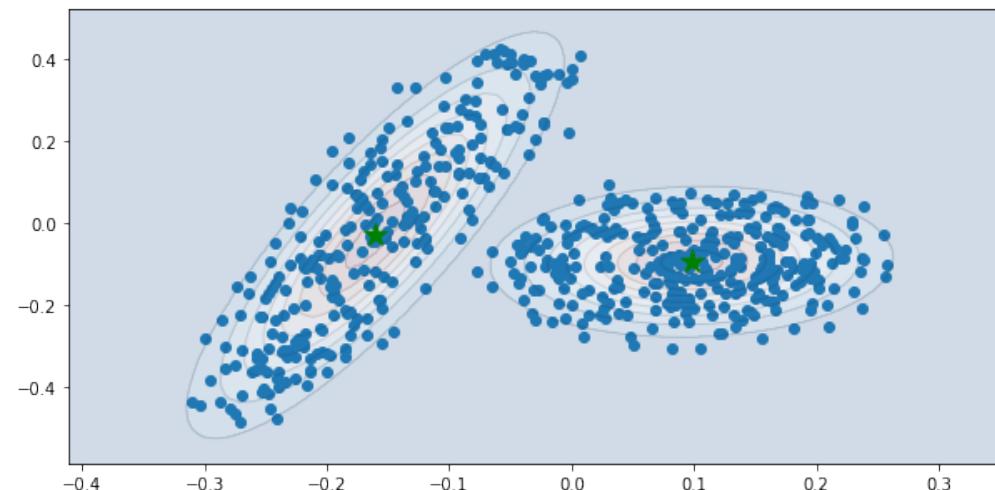
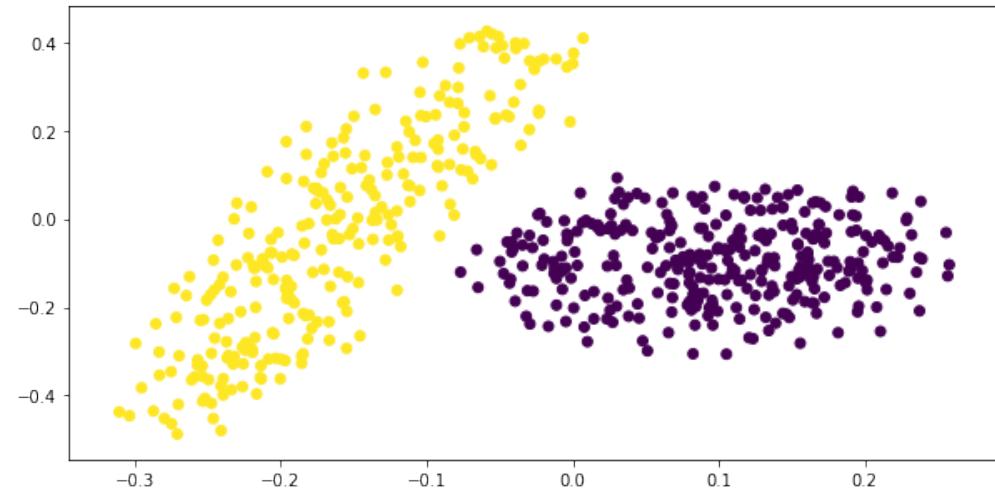
Agrupamiento probabilístico

Mixtura de Gaussianas



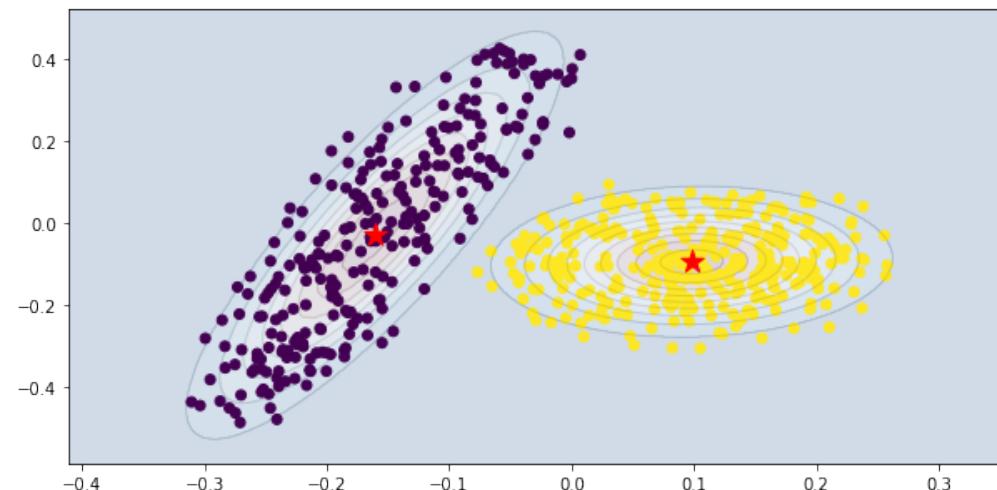
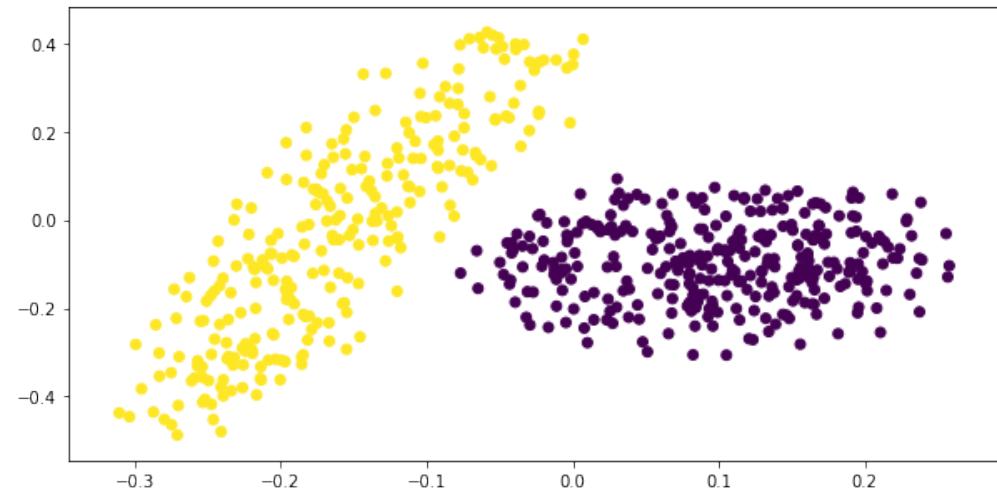
Agrupamiento probabilístico

Mixtura de Gaussianas



Agrupamiento probabilístico

Mixtura de Gaussianas



Agrupamiento probabilístico

Mixtura de Gaussianas

Ventajas

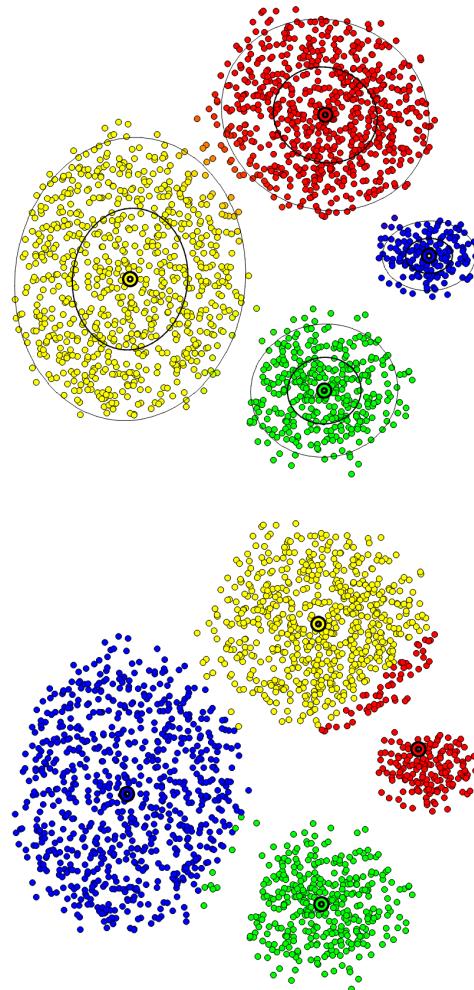
- ▶ Definición basada en probabilidad
- ▶ Definición sencilla
- ▶ Podría funcionar con diferentes distribuciones
- ▶ Funciona con clústeres cóncavos elípticos (vs. K -means)

Agrupamiento probabilístico

Mixtura de Gaussianas

Relación con *K*-means

- ▶ Asignación probabilística
vs. determinista
- ▶ Clústeres elípticos
vs. circulares



Agrupamiento probabilístico

Mixtura de Gaussianas

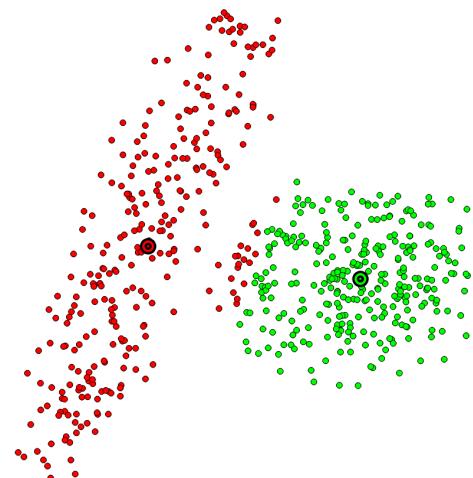
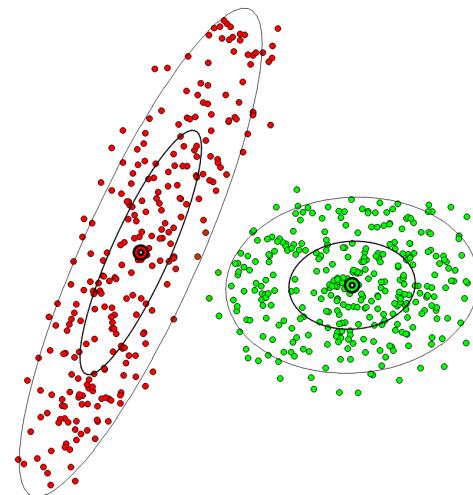
[Examen]

Para clústeres elípticos ¿qué es mejor?

- K-means
- Agrupamiento Probabilístico

Relación con K-means

- ▶ Asignación probabilística
vs. determinista
- ▶ Clústeres elípticos
vs. circulares



Agrupamiento probabilístico

Mixtura de Gaussianas

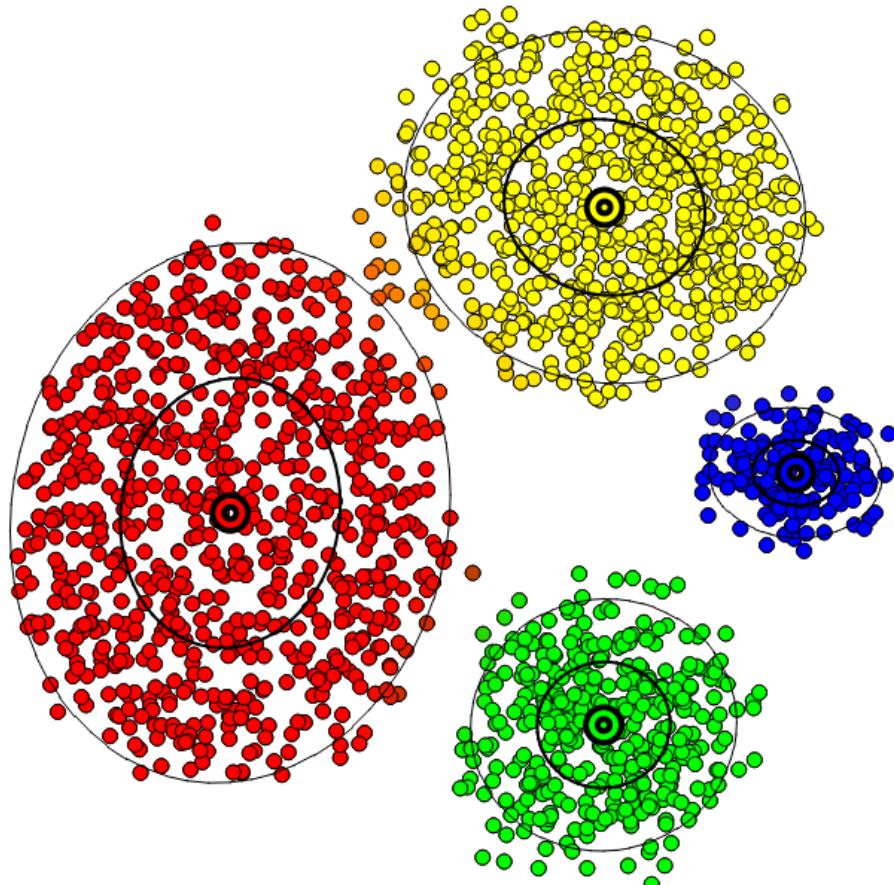
Desventajas

- ▶ Depende de K (parámetro)
- ▶ Asume clústeres cóncavos
- ▶ Problemas al lidiar con clústeres de diferente forma
- ▶ Depende de la inicialización

Agrupamiento probabilístico

Mixtura de Gaussianas

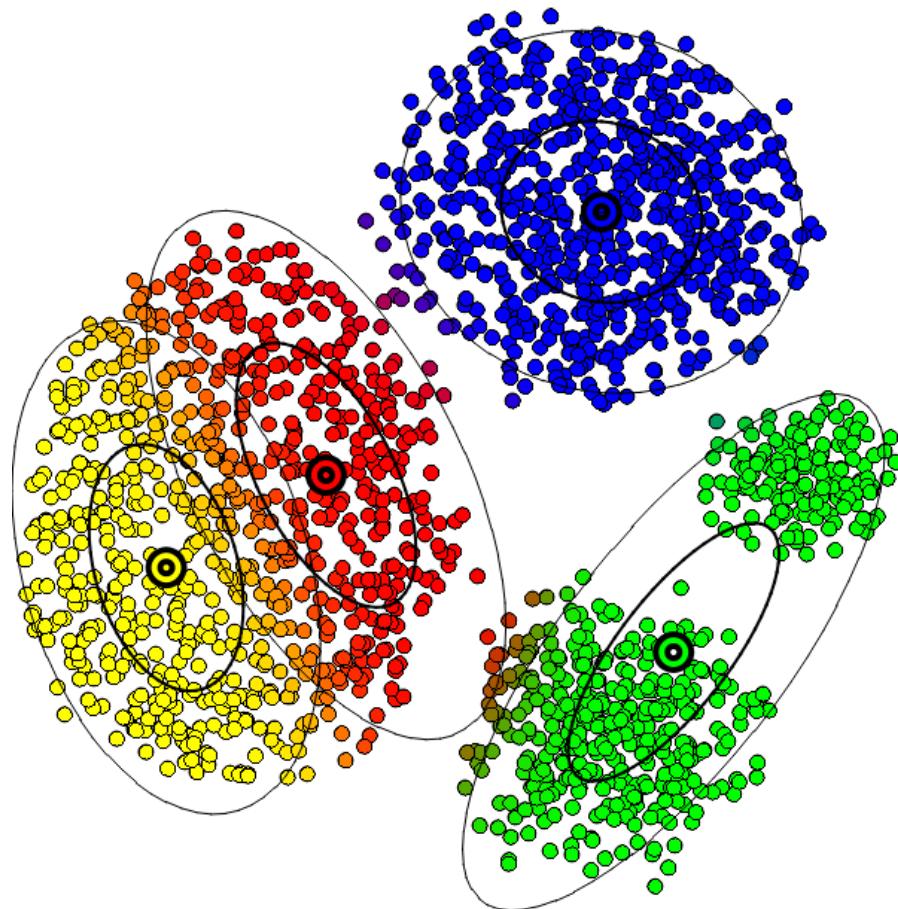
Dependencia de la inicialización



Agrupamiento probabilístico

Mixtura de Gaussianas

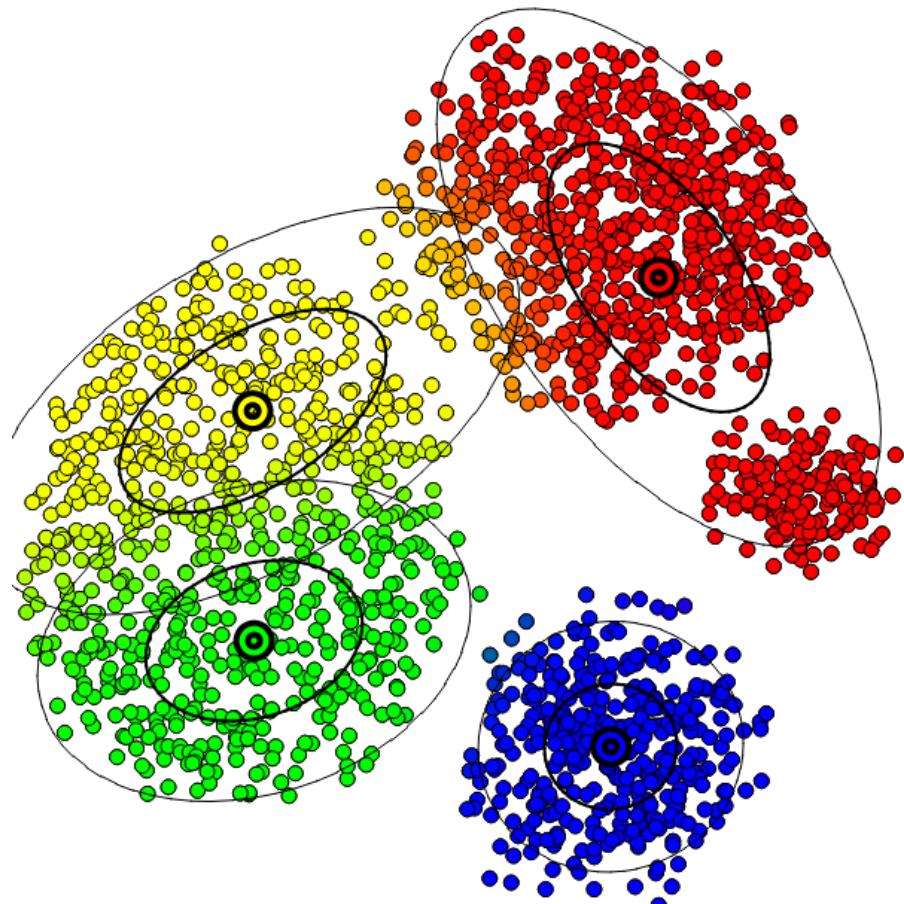
Dependencia de la inicialización



Agrupamiento probabilístico

Mixtura de Gaussianas

Dependencia de la inicialización



Aprendizaje no supervisado

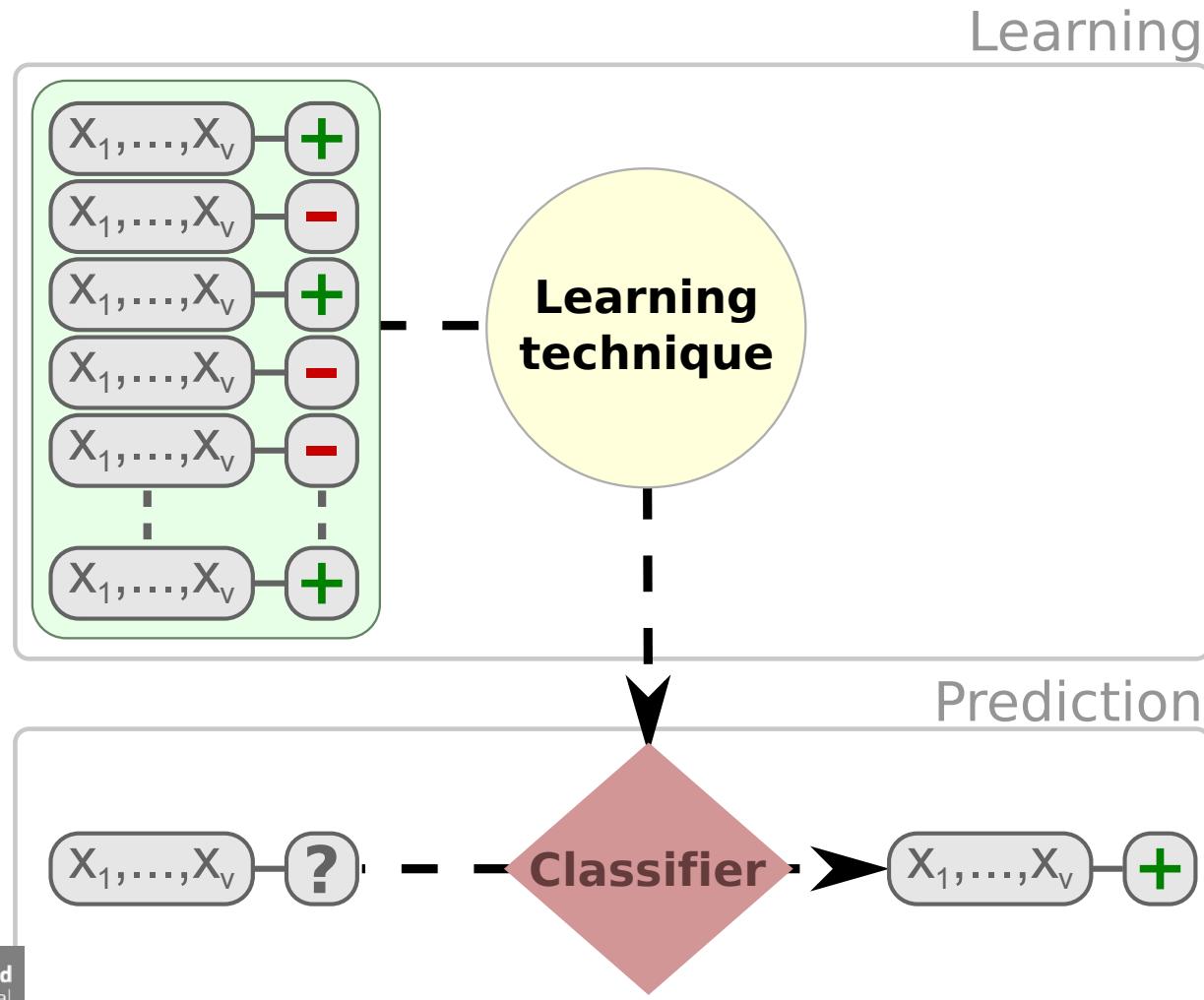
VC07: Naive Bayes

Rocío del Amor del Amor
mrocio.delamor@campusviu.es

Universidad Internacional de Valencia

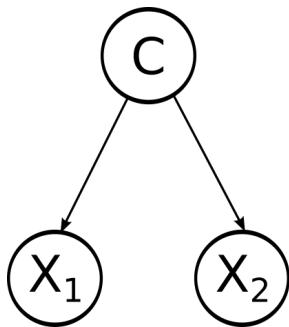
[Examen] Assume independencia entre las características (variables)
(independientes entre sí)

Clasificación Supervisada



Naive Bayes

$$p(X_1, X_2, C) = p(C|X_1, X_2) \times p(X_1, X_2)$$
$$p(C|X_1, X_2) = \frac{p(X_1, X_2, C)}{p(X_1, X_2)}$$



Regla de la cadena

$$p(X_1, X_2, C) = p(X_1|X_2, C) \times p(X_2, C)$$

$$p(X_1, X_2, C) = p(X_1|X_2, C) \times p(X_2|C) \times p(C)$$

Asunción de independencia X's—C del Naive Bayes

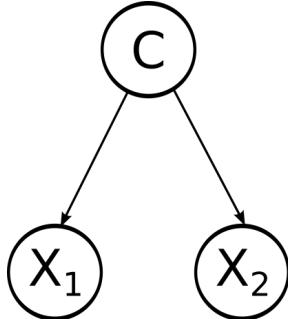
$$p(X_1, X_2, C) = p(X_1|C) \times p(X_2|C) \times p(C)$$
$$p(C|X_1, X_2) = \frac{p(X_1|C) \times p(X_2|C) \times p(C)}{p(X_1, X_2)}$$

$$\operatorname{argmáx}_c p(C|X_1, X_2)$$

$$p(C|X_1, X_2) \propto p(X_1|C) \times p(X_2|C) \times p(C)$$

$$\operatorname{argmáx}_c p(X_1|C) \times p(X_2|C) \times p(C)$$

Naive Bayes



$$\theta_C = p(C = 1)$$

$$\theta_{X_1|C=0} = p(X_1 = 1|C = 0)$$

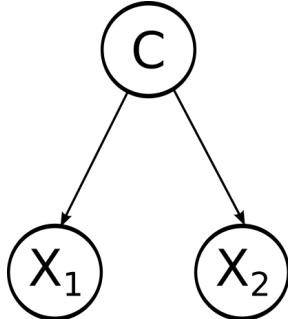
$$\theta_{X_1|C=1} = p(X_1 = 1|C = 1)$$

$$\theta_{X_2|C=0} = p(X_2 = 1|C = 0)$$

$$\theta_{X_2|C=1} = p(X_2 = 1|C = 1)$$

		X_1	C	$p(X_1 C)$			X_2	C	$p(X_2 C)$
		0	0	0,50			0	0	0,25
		1	0	0,50			1	0	0,75
C	$p(C)$	0	1	0,33			0	1	0,66
0	0,4	1	1	0,66			1	1	0,33
1	0,6								

Naive Bayes



$$\theta_C = p(C = 1)$$

$$\theta_{X_1|C=0} = p(X_1 = 1|C = 0)$$

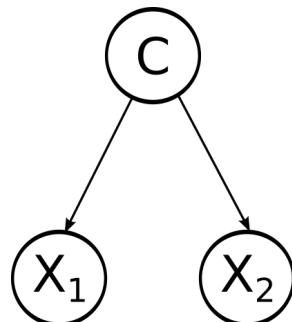
$$\theta_{X_1|C=1} = p(X_1 = 1|C = 1)$$

$$\theta_{X_2|C=0} = p(X_2 = 1|C = 0)$$

$$\theta_{X_2|C=1} = p(X_2 = 1|C = 1)$$

		X_1	C	$p(X_1 C)$			X_2	C	$p(X_2 C)$		
C	$p(C)$	0	0	0,50			0	0	0,25		
0	0,4	1	0	0,50	=	$\theta_{X_1 C=0}$	1	0	0,75	=	$\theta_{X_2 C=0}$
1	$0,6 = \theta_C$	0	1	0,33			0	1	0,66		
		1	1	0,66	=	$\theta_{X_1 C=1}$	1	1	0,33	=	$\theta_{X_2 C=1}$

Naive Bayes



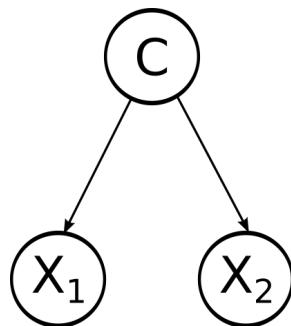
C	$p(C)$	X_1	C	$p(X_1 C)$	X_2	C	$p(X_2 C)$
0	0,4		0	0,50		0	0,25
1	0,6		1	0,50		1	0,75
		0	1	0,33		0	0,66
		1	1	0,66		1	0,33

$$\hat{c} = f(\mathbf{x}) = \operatorname{argmáx}_c p(c) \prod_{i=1}^2 p(x_i|c)$$

X_1	X_2	C
0	0	?
1	0	?

$$\begin{aligned} C=0 & \quad (1 - \theta_C) \times (1 - \theta_{X_1|C=0}) \times (1 - \theta_{X_2|C=0}) \\ C=1 & \quad \theta_C \times (1 - \theta_{X_1|C=1}) \times (1 - \theta_{X_2|C=1}) \end{aligned}$$

Naive Bayes



C	$p(C)$
0	0,4
1	0,6

X_1	C	$p(X_1 C)$	X_2	C	$p(X_2 C)$
0	0	0,50	0	0	0,25
1	0	0,50	1	0	0,75
0	1	0,33	0	1	0,66
1	1	0,66	1	1	0,33

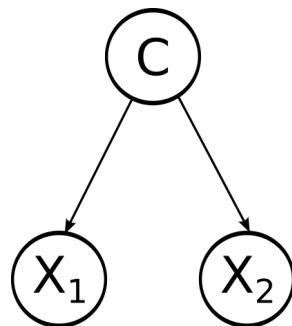
$$\hat{c} = f(\mathbf{x}) = \operatorname{argmáx}_c p(c) \prod_{i=1}^2 p(x_i|c)$$

X_1	X_2	C
0	0	?
1	0	?



$$\left\{ \begin{array}{ll} C=0 & 0,4 \times 0,5 \times 0,25 \\ C=1 & 0,6 \times 0,33 \times 0,66 \end{array} \right.$$

Naive Bayes



C	$p(C)$
0	0,4
1	0,6

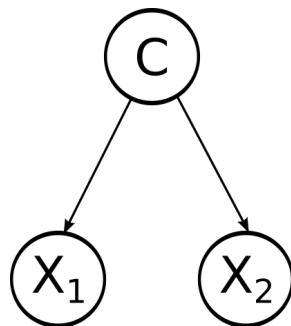
X_1	C	$p(X_1 C)$	X_2	C	$p(X_2 C)$
0	0	0,50	0	0	0,25
1	0	0,50	1	0	0,75
0	1	0,33	0	1	0,66
1	1	0,66	1	1	0,33

$$\hat{c} = f(\mathbf{x}) = \operatorname{argmáx}_c p(c) \prod_{i=1}^2 p(x_i|c)$$

X_1	X_2	C
0	0	?
1	0	?

$$\begin{aligned} C=0 & \quad (1 - \theta_C) \times \theta_{X_1|C=0} \times (1 - \theta_{X_2|C=0}) \\ C=1 & \quad \theta_C \times \theta_{X_1|C=1} \times (1 - \theta_{X_2|C=1}) \end{aligned}$$

Naive Bayes



C	$p(C)$
0	0,4
1	0,6

X_1	C	$p(X_1 C)$	X_2	C	$p(X_2 C)$
0	0	0,50	0	0	0,25
1	0	0,50	1	0	0,75
0	1	0,33	0	1	0,66
1	1	0,66	1	1	0,33

$$\hat{c} = f(\mathbf{x}) = \operatorname{argmáx}_c p(c) \prod_{i=1}^2 p(x_i|c)$$

X_1	X_2	C
0	0	?
1	0	?

$$C=0 \quad 0,4 \times 0,5 \times 0,25$$

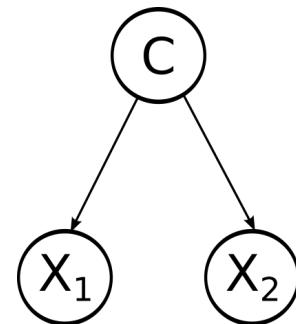
$$C=1 \quad 0,6 \times 0,66 \times 0,66$$

Naive Bayes

Aprendizaje del modelo

Estimando los parámetros (tablas de probabilidades):

X_1	X_2	C
1	0	0
0	0	1
0	1	0
0	1	0
1	0	1
0	0	1
1	1	0
1	1	1
0	1	0
1	1	0



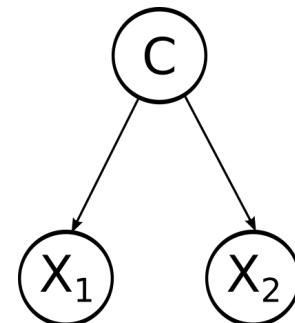
$$\begin{aligned} p(C) \\ p(X_1|C) \\ p(X_2|C) \end{aligned}$$

Naive Bayes

Aprendizaje del modelo

Estimando los parámetros (tablas de probabilidades):

X_1	X_2	C
1	0	0
0	0	1
0	1	0
0	1	0
1	0	1
0	0	1
1	1	0
1	1	1
0	1	0
1	1	0



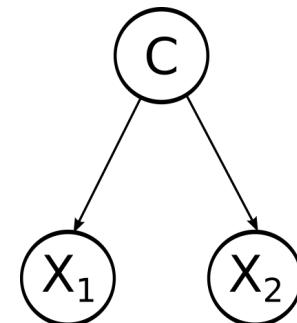
C	$p(C)$
0	
1	

Naive Bayes

Aprendizaje del modelo

Estimando los parámetros (tablas de probabilidades):

X_1	X_2	C
1	0	0
0	0	1
0	1	0
0	1	0
1	0	1
0	0	1
1	1	0
1	1	1
0	1	0
1	1	0



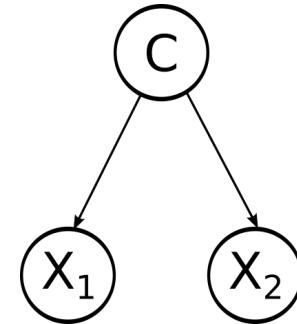
C	$p(C)$
0	$6/10 = 0,6$
1	$4/10 = 0,4$

Naive Bayes

Aprendizaje del modelo

Estimando los parámetros (tablas de probabilidades):

X_1	X_2	C
1	0	0
0	0	1
0	1	0
0	1	0
1	0	1
0	0	1
1	1	0
1	1	1
0	1	0
1	1	0



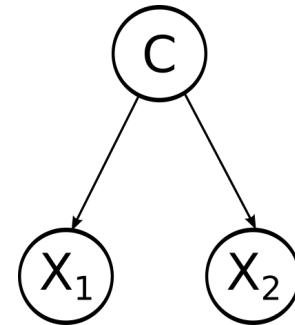
X_1	C	$p(X_1 C)$
0	0	3/6
1	0	
0	1	
1	1	

Naive Bayes

Aprendizaje del modelo

Estimando los parámetros (tablas de probabilidades):

X_1	X_2	C
1	0	0
0	0	1
0	1	0
0	1	0
1	0	1
0	0	1
1	1	0
1	1	1
0	1	0
1	1	0



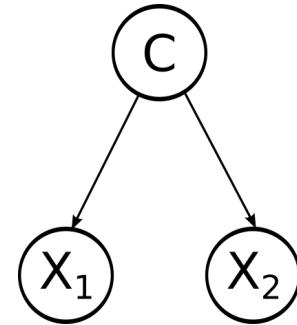
X_1	C	$p(X_1 C)$
0	0	$3/6 = 0,50$
1	0	$3/6 = 0,50$
0	1	$2/4 = 0,50$
1	1	$2/4 = 0,50$

Naive Bayes

Aprendizaje del modelo

Estimando los parámetros (tablas de probabilidades):

X_1	X_2	C
1	0	0
0	0	1
0	1	0
0	1	0
1	0	1
0	0	1
1	1	0
1	1	1
0	1	0
1	1	0



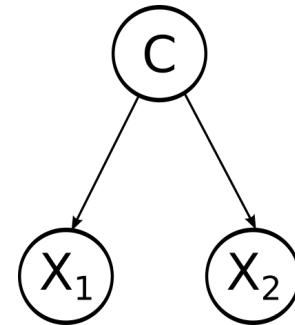
X_2	C	$p(X_2 C)$
0	0	
1	0	
0	1	
1	1	

Naive Bayes

Aprendizaje del modelo

Estimando los parámetros (tablas de probabilidades):

X_1	X_2	C
1	0	0
0	0	1
0	1	0
0	1	0
1	0	1
0	0	1
1	1	0
1	1	1
0	1	0
1	1	0



X_2	C	$p(X_2 C)$
0	0	$1/6 = 0,17$
1	0	$5/6 = 0,83$
0	1	$3/4 = 0,75$
1	1	$1/4 = 0,25$

Aprendizaje no supervisado

VC08: Aprendizaje semi-supervisado

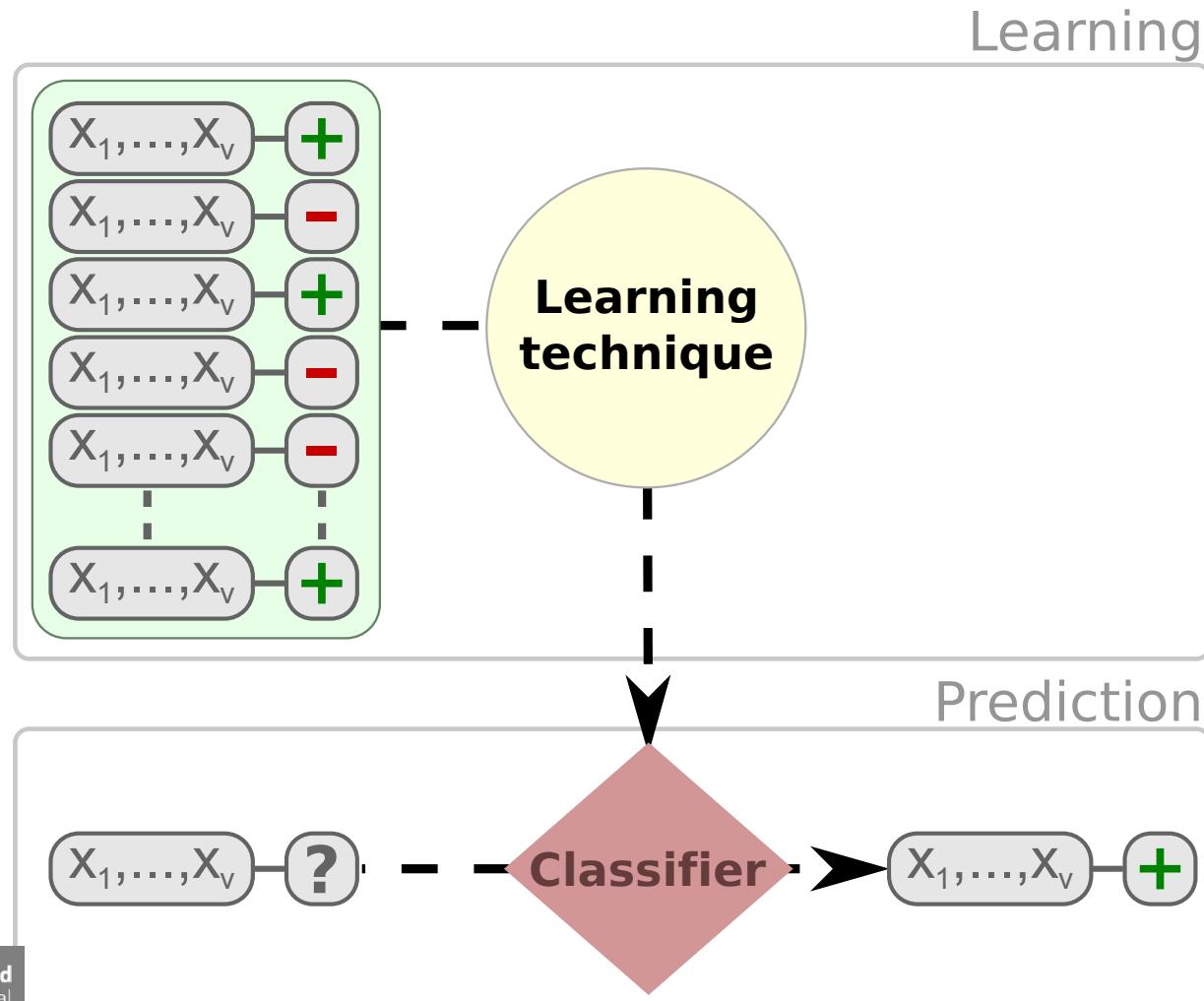
Rocío del Amor del Amor
mrocio.delamor@campusviu.es

Universidad Internacional de Valencia

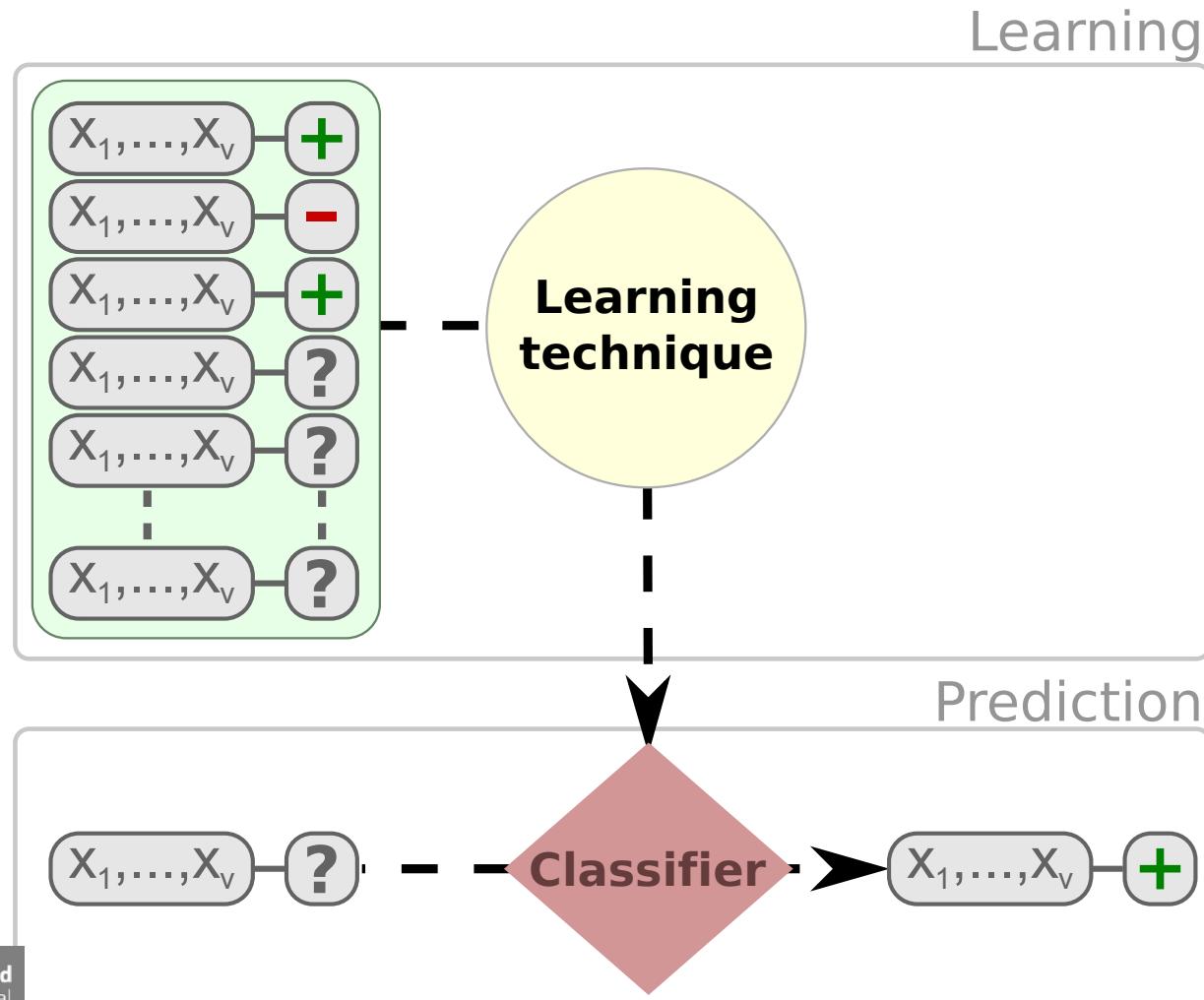
Algunos datos tienen etiquetas y otros no

- ↳ para reducir overfitting
- ↳ cuando hay pocos datos.

Aprendizaje semi-supervisado



Aprendizaje semi-supervisado



Aprendizaje semi-supervisado



Aprendizaje semi-supervisado

The image shows a screenshot of a semi-supervised learning process on a Gmail interface. On the left, the main Gmail inbox shows two forwarded messages from 'new user' with subject lines 'Foster Cold Stores & Truck - www.fostercoldstore.com - 65years of refrigeration' and 'Fwd: Password token - Foster Cold Stores & Truck'. The second message has a red box around its checkbox and the word 'Forwards (2)'. On the right, a search results page for 'in:spam' shows five spam emails: 'SoftMaker Software GmbH', 'no_reply', 'no_reply', 'Abelssoft', and '1-abc.net News'. A red arrow points from the 'Not spam' button in the search results back to the 'Forwards (2)' message in the inbox, illustrating the propagation of a learned classification rule.

Gmail - Inbox (1) - newuser7887@gmail....

+You Gmail Calendar Documents Photos Sites Search More

Your filters are forwarding some of your email to newuser87

Out of town

Gmail

Mail

Compose

Inbox (1)

Important

Chats

Sent Mail

Chat

Search people...

new user Set status here

Call phone

mynname.someone

Foster Cold Stores & Truck - www.fostercoldstore.com - 65years of refrigeration

Fwd: Password token - Foster Cold Stores & Truck

Forwards (2)

new user

new user

Hi

Google

in:spam

Gmail

Compose

Inbox

Important

Sent Mail

Drafts

All Mail

Spam (5)

SoftMaker Software GmbH

no_reply

no_reply

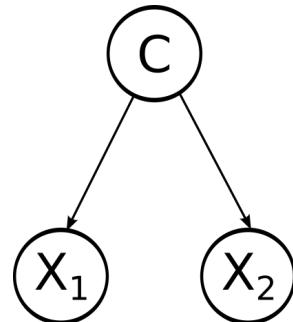
Abelssoft

1-abc.net News

Estrategia para estimar los parámetros de máxima verosimilitud (MLE) cuando hay datos incompletos.

¿Por qué no se pueden obtener directamente?

X_1	X_2	C
1	0	0
0	0	?
0	1	?
0	1	?
1	0	1
0	0	1
1	1	?
1	1	?
0	1	0
1	1	?



C	$p(C)$
0	?/10
1	?/10

Algoritmo *Expectation-Maximization*:

Procedimiento iterativo de dos pasos (E-M) que permite obtener los parámetros de máxima verosimilitud cuando hay datos perdidos (valores perdidos, variables latentes, etc.)

E-step: Se estima el valor de los datos perdidos usando la esperanza condicional de la verosimilitud

M-step: Se estiman unos nuevos parámetros dados los datos completados en el paso E.

Convergencia:

- ▶ Máximo (local)
- ▶ Casos raros: Punto de silla

Algoritmo *Expectation-Maximization*:

Procedimiento iterativo de dos pasos (E-M) que permite obtener los parámetros de máxima verosimilitud cuando hay datos perdidos (valores perdidos, variables latentes, etc.)

E-step:

$$Q(\theta; \theta^t) = E_{Z|X, \theta^t} [\log L(\theta; X, Z)]$$

M-step: Choose θ^{t+1} such that, for all $\theta \in \Theta$:

$$Q(\theta^{t+1}; \theta^t) \geq Q(\theta; \theta^t)$$

Donde Z son los datos perdidos, X los observados, y θ los parámetros del modelo. Se define verosimilitud como:

$$L(\theta; X, Z) = p(X, Z; \theta)$$

Algoritmo *Expectation-Maximization*:

Procedimiento iterativo de dos pasos (E-M) que permite obtener los parámetros de máxima verosimilitud cuando hay datos perdidos (valores perdidos, variables latentes, etc.)

E-step:

$$Q(\theta; \theta^t) = E_{Z|X, \theta^t} [\log L(\theta; X, Z)]$$

M-step:

$$\theta^{t+1} = \operatorname{argmáx}_{\theta} Q(\theta; \theta^t)$$

Donde Z son los datos perdidos, X los observados, y θ los parámetros del modelo. Se define verosimilitud como:

$$L(\theta; X, Z) = p(X, Z; \theta)$$

Verosimilitud:

$$L(\theta; X, Z) = p(X, Z; \theta)$$

$$L(\theta; X) = p(X; \theta) = \sum_Z p(X, Z; \theta)$$

$$p(X; \theta) = \frac{p(X, Z; \theta)}{p(Z|X; \theta)}$$

¿Maximizando Q maximizamos la verosimilitud?

$$\log p(X; \theta) = \log p(X, Z; \theta) - \log p(Z|X; \theta)$$

Verosimilitud:

$$L(\theta; X, Z) = p(X, Z; \theta)$$

$$L(\theta; X) = p(X; \theta) = \sum_Z p(X, Z; \theta)$$

$$p(X; \theta) = \frac{p(X, Z; \theta)}{p(Z|X; \theta)}$$

¿Maximizando Q maximizamos la verosimilitud?

$$\log p(X; \theta) = \sum_Z p(Z|X; \theta^t) \log p(X, Z; \theta) - \sum_Z p(Z|X; \theta^t) \log p(Z|X; \theta)$$

Verosimilitud:

$$L(\theta; X, Z) = p(X, Z; \theta)$$

$$L(\theta; X) = p(X; \theta) = \sum_Z p(X, Z; \theta)$$

$$p(X; \theta) = \frac{p(X, Z; \theta)}{p(Z|X; \theta)}$$

¿Maximizando Q maximizamos la verosimilitud?

$$\log p(X|\theta) = Q(\theta; \theta^t) + H(\theta; \theta^t)$$

Verosimilitud:

$$L(\theta; X, Z) = p(X, Z; \theta)$$

$$L(\theta; X) = p(X; \theta) = \sum_Z p(X, Z; \theta)$$

$$p(X; \theta) = \frac{p(X, Z; \theta)}{p(Z|X; \theta)}$$

¿Maximizando Q maximizamos la verosimilitud?

$$\log p(X|\theta) = Q(\theta; \theta^t) + H(\theta; \theta^t)$$

$$\log p(X|\theta) - \log p(X|\theta^t) = Q(\theta; \theta^t) - Q(\theta^t; \theta^t) + H(\theta; \theta^t) - H(\theta^t; \theta^t)$$

Verosimilitud:

$$L(\theta; X, Z) = p(X, Z; \theta)$$

$$L(\theta; X) = p(X; \theta) = \sum_Z p(X, Z; \theta)$$

$$p(X; \theta) = \frac{p(X, Z; \theta)}{p(Z|X; \theta)}$$

¿Maximizando Q maximizamos la verosimilitud?

$$\log p(X|\theta) = Q(\theta; \theta^t) + H(\theta; \theta^t)$$

$$\log p(X|\theta) - \log p(X|\theta^t) = Q(\theta; \theta^t) - Q(\theta^t; \theta^t) + C$$

con $C \geq 0$.

Verosimilitud:

$$L(\theta; X, Z) = p(X, Z; \theta)$$

$$L(\theta; X) = p(X; \theta) = \sum_Z p(X, Z; \theta)$$

$$p(X; \theta) = \frac{p(X, Z; \theta)}{p(Z|X; \theta)}$$

¿Maximizando Q maximizamos la verosimilitud?

$$\log p(X|\theta) = Q(\theta; \theta^t) + H(\theta; \theta^t)$$

$$\log p(X|\theta) - \log p(X|\theta^t) \geq Q(\theta; \theta^t) - Q(\theta^t; \theta^t)$$

EM en la práctica

Aprendizaje del modelo (NB)

Paso E:

Determinista

X_1	X_2	C
1	0	0
0	0	?
0	1	?
0	1	?
1	0	1
0	0	1
1	1	?
1	0	?
0	1	0
1	1	?

Probabilista

X_1	X_2	C
1	0	0
0	0	?
0	1	?
0	1	?
1	0	1
0	0	1
1	1	?
1	0	?
0	1	0
1	1	?

EM en la práctica

Aprendizaje del modelo (NB)

Paso E:

Determinista

X_1	X_2	C
1	0	0
0	0	1
0	1	0
0	1	1
1	0	1
0	0	1
1	1	0
1	0	1
0	1	0
1	1	0

Probabilista

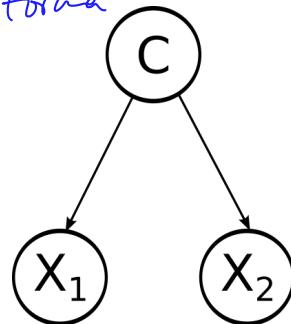
X_1	X_2	C	
		0	1
1	0	1.0	0.0
0	0	0.4	0.6
0	1	0.7	0.3
0	1	0.5	0.5
1	0	0.0	1.0
0	0	0.0	1.0
1	1	0.8	0.2
1	0	0.3	0.7
0	1	1.0	0.0
1	1	0.6	0.4

EM

Aprendizaje del modelo

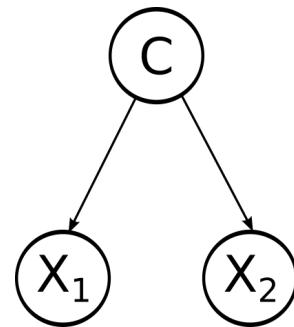
X_1	X_2	C
1	0	0
0	0	?
0	1	?
0	1	?
1	0	1
0	0	1
1	1	?
1	0	?
0	1	0
1	1	?

¿Cómo lo hacemos?
Inputamos de forma
probabilística



C	$p(C)$
0	?/10
1	?/10

X_1	X_2	C	
		0	1
1	0	1.0	0.0
0	0	0.5	0.5
0	1	0.5	0.5
0	1	0.5	0.5
1	0	0.0	1.0
0	0	0.0	1.0
1	1	0.5	0.5
1	0	0.5	0.5
0	1	1.0	0.0
1	1	0.5	0.5



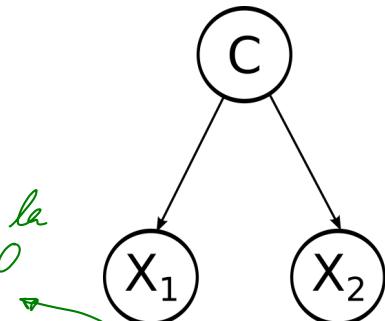
C	$p(C)$
0	
1	

EM

Aprendizaje del modelo

X_1	X_2	C	
		0	1
1	0	1.0	0.0
0	0	0.5	0.5
0	1	0.5	0.5
0	1	0.5	0.5
1	0	0.0	1.0
0	0	0.0	1.0
1	1	0.5	0.5
1	0	0.5	0.5
0	1	1.0	0.0
1	1	0.5	0.5

suma de la
columna 0



C	$p(C)$
0	$5/10 = 0,5$
1	$5/10 = 0,5$

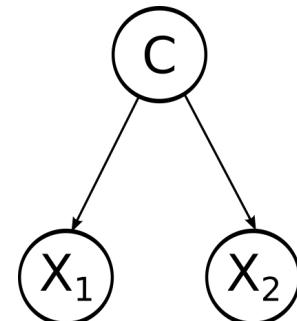
EM

Aprendizaje del modelo

[Examen]

Clase 19/10/2023 hora 20:25-20:27 , pregunta.

		C		
		0	1	
		X ₁	X ₂	
1	0	✓1.0	0.0	1
0.5	0 0	•0.5	0.5	0.5
0.5	0 1	•0.5	0.5	0.5
0.5	0 1	•0.5	0.5	0.5
1	0	0.0	1.0	
0	0	0.0	1.0	
1	1	•0.5	0.5	0.5
1	0	•0.5	0.5	0.5
1	0 1	✓1.0	0.0	1
1	1	•0.5	0.5	0.5

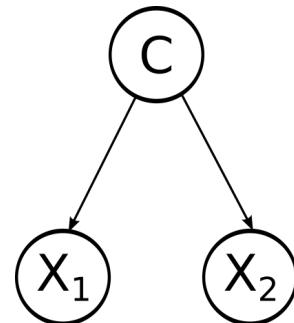


X ₁	C	p(X ₁ C)
0	0	1+3·0.5 /2+6·0.5
1	0	
0	1	
1	1	

EM

Aprendizaje del modelo

X_1	X_2	C	
		0	1
1	0	1.0	0.0
0	0	0.5	0.5
0	1	0.5	0.5
0	1	0.5	0.5
1	0	0.0	1.0
0	0	0.0	1.0
1	1	0.5	0.5
1	0	0.5	0.5
0	1	1.0	0.0
1	1	0.5	0.5

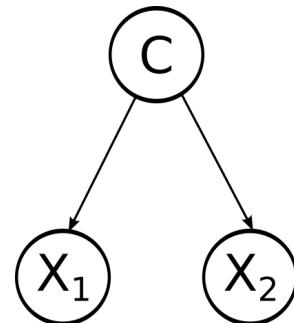


X_1	C	$p(X_1 C)$
0	0	$2,5/5 = 0,50$
1	0	$2,5/5 = 0,50$
0	1	$2,5/5 = 0,50$
1	1	$2,5/5 = 0,50$

EM

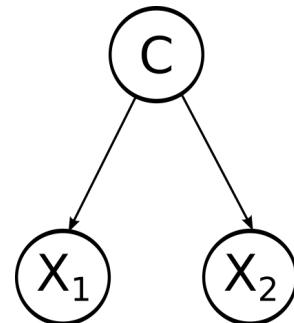
Aprendizaje del modelo

X_1	X_2	C	
		0	1
1	0	1.0	0.0
0	0	0.5	0.5
0	1	0.5	0.5
0	1	0.5	0.5
1	0	0.0	1.0
0	0	0.0	1.0
1	1	0.5	0.5
1	0	0.5	0.5
0	1	1.0	0.0
1	1	0.5	0.5



X_2	C	$p(X_2 C)$
0	0	
1	0	
0	1	
1	1	

X_1	X_2	C	
		0	1
1	0	1.0	0.0
0	0	0.5	0.5
0	1	0.5	0.5
0	1	0.5	0.5
1	0	0.0	1.0
0	0	0.0	1.0
1	1	0.5	0.5
1	0	0.5	0.5
0	1	1.0	0.0
1	1	0.5	0.5

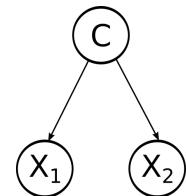


X_2	C	$p(X_2 C)$
0	0	2/5 = 0,40
1	0	3/5 = 0,60
0	1	3/5 = 0,60
1	1	2/5 = 0,40

EM

Re-estimación de pesos

X_1	X_2	C	
		0	1
1	0	1.0	0.0
0	0	0.5	0.5
0	1	0.5	0.5
0	1	0.5	0.5
1	0	0.0	1.0
0	0	0.0	1.0
1	1	0.5	0.5
1	0	0.5	0.5
0	1	1.0	0.0
1	1	0.5	0.5



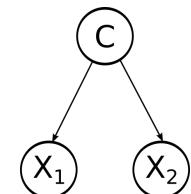
C	$p(C)$
0	0.5
1	0.5

X_i	C	$p(X_1 C)$	$p(X_2 C)$
0	0	0.5	0.4
1	0	0.5	0.6
0	1	0.5	0.6
1	1	0.5	0.4

$$\hat{c} = \operatorname{argmáx}_c p(c) \prod_{i=1}^2 p(x_i|c)$$

Re-estimación de pesos

		C	
		0	1
X_1	X_2		
1	0	1.0	0.0
0	0	0.5	0.5
0	1	0.5	0.5
0	1	0.5	0.5
1	0	0.0	1.0
0	0	0.0	1.0
1	1	0.5	0.5
1	0	0.5	0.5
0	1	1.0	0.0
1	1	0.5	0.5



C	$p(C)$
0	0.5
1	0.5

X_i	C	$p(X_1 C)$	$p(X_2 C)$
0	0	0.5	0.4
1	0	0.5	0.6
0	1	0.5	0.6
1	1	0.5	0.4

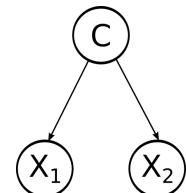
$$p(c = 0|\mathbf{x}) = \frac{1}{\theta} p(0) \prod_{i=1}^2 p(x_i|0)$$

$$p(c = 1|\mathbf{x}) = \frac{1}{\theta} p(1) \prod_{i=1}^2 p(x_i|1)$$

EM

Re-estimación de pesos

X_1	X_2	C	
		0	1
1	0	1.0	0.0
0	0	0.10	0.15
0	1	0.5	0.5
0	1	0.5	0.5
1	0	0.0	1.0
0	0	0.0	1.0
1	1	0.5	0.5
1	0	0.5	0.5
0	1	1.0	0.0
1	1	0.5	0.5



C	$p(C)$
0	0.5
1	0.5

X_i	C	$p(X_1 C)$	$p(X_2 C)$
0	0	0.5	0.4
1	0	0.5	0.6
0	1	0.5	0.6
1	1	0.5	0.4

$$p(c = 0|\mathbf{x}) = \frac{1}{\theta} p(0) \prod_{i=1}^2 p(x_i|0)$$

$$p(c = 1|\mathbf{x}) = \frac{1}{\theta} p(1) \prod_{i=1}^2 p(x_i|1)$$

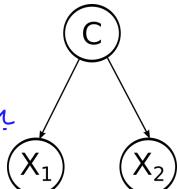
EM

Re-estimación de pesos

[Examen] (lunes 19/10/2023 hora 20:28)

		C	
		0	1
X_1	X_2	0	1
1	0	1.0	0.0
0	0	0.4	0.6
0	1	0.5	0.5
0	1	0.5	0.5
1	0	0.0	1.0
0	0	0.0	1.0
1	1	0.5	0.5
1	0	0.5	0.5
0	1	1.0	0.0
1	1	0.5	0.5

se normalizan



C	$p(C)$
0	0.5
1	0.5

X_i	C	$p(X_1 C)$	$p(X_2 C)$
0	0	0.5	0.4
1	0	0.5	0.6
0	1	0.5	0.6
1	1	0.5	0.4

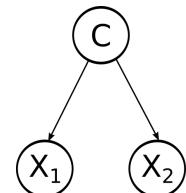
$$p(c=0|\mathbf{x}) = \frac{1}{\theta} p(0) \prod_{i=1}^2 p(x_i|0)$$

$$p(c=1|\mathbf{x}) = \frac{1}{\theta} p(1) \prod_{i=1}^2 p(x_i|1)$$

EM

Re-estimación de pesos

X_1	X_2	C	
		0	1
1	0	1.0	0.0
0	0	0.4	0.6
0	1	0.15	0.10
0	1	0.15	0.10
1	0	0.0	1.0
0	0	0.0	1.0
1	1	0.10	0.15
1	0	0.15	0.10
0	1	1.0	0.0
1	1	0.15	0.10



C	$p(C)$
0	0.5
1	0.5

X_i	C	$p(X_1 C)$	$p(X_2 C)$
0	0	0.5	0.4
1	0	0.5	0.6
0	1	0.5	0.6
1	1	0.5	0.4

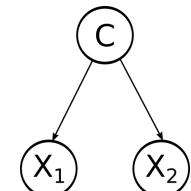
$$p(c = 0|\mathbf{x}) = \frac{1}{\theta} p(0) \prod_{i=1}^2 p(x_i|0)$$

$$p(c = 1|\mathbf{x}) = \frac{1}{\theta} p(1) \prod_{i=1}^2 p(x_i|1)$$

EM

Re-estimación de pesos

		C	
		0	1
X_1	X_2		
1	0	1.0	0.0
0	0	0.4	0.6
0	1	0.6	0.4
0	1	0.6	0.4
1	0	0.0	1.0
0	0	0.0	1.0
1	1	0.4	0.6
1	0	0.6	0.4
0	1	1.0	0.0
1	1	0.6	0.4



C	$p(C)$
0	0.5
1	0.5

X_i	C	$p(X_1 C)$	$p(X_2 C)$
0	0	0.5	0.4
1	0	0.5	0.6
0	1	0.5	0.6
1	1	0.5	0.4

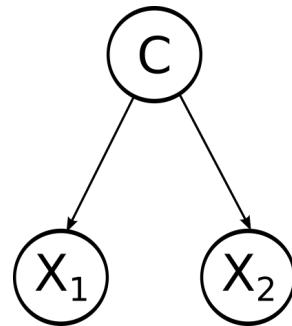
$$p(c = 0|\mathbf{x}) = \frac{1}{\theta} p(0) \prod_{i=1}^2 p(x_i|0)$$

$$p(c = 1|\mathbf{x}) = \frac{1}{\theta} p(1) \prod_{i=1}^2 p(x_i|1)$$

EM

Re-aprender el modelo

X_1	X_2	C	
		0	1
1	0	1.0	0.0
0	0	0.4	0.6
0	1	0.6	0.4
0	1	0.6	0.4
1	0	0.0	1.0
0	0	0.0	1.0
1	1	0.4	0.6
1	0	0.6	0.4
0	1	1.0	0.0
1	1	0.6	0.4

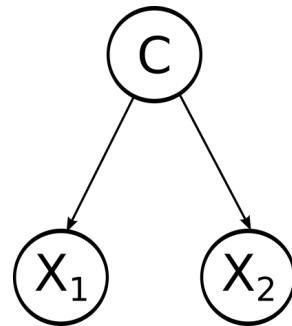


C	$p(C)$
0	
1	

EM

Re-aprender el modelo

X_1	X_2	C	
		0	1
1	0	1.0	0.0
0	0	0.4	0.6
0	1	0.6	0.4
0	1	0.6	0.4
1	0	0.0	1.0
0	0	0.0	1.0
1	1	0.4	0.6
1	0	0.6	0.4
0	1	1.0	0.0
1	1	0.6	0.4

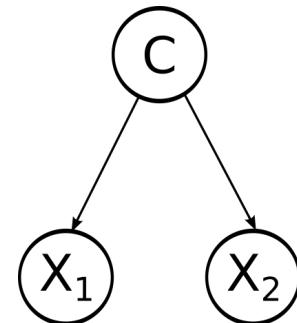


C	$p(C)$
0	0,52
1	0,48

EM

Re-aprender el modelo

X_1	X_2	C	
		0	1
1	0	1.0	0.0
0	0	0.4	0.6
0	1	0.6	0.4
0	1	0.6	0.4
1	0	0.0	1.0
0	0	0.0	1.0
1	1	0.4	0.6
1	0	0.6	0.4
0	1	1.0	0.0
1	1	0.6	0.4

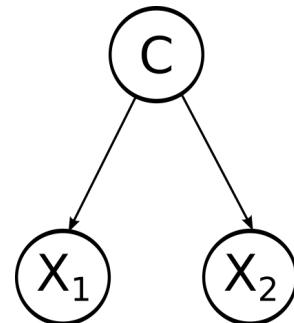


X_1	C	$p(X_1 C)$
0	0	
1	0	
0	1	
1	1	

EM

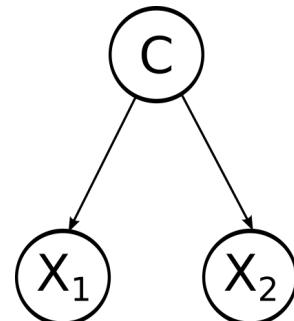
Re-aprender el modelo

X_1	X_2	C	
		0	1
1	0	1.0	0.0
0	0	0.4	0.6
0	1	0.6	0.4
0	1	0.6	0.4
1	0	0.0	1.0
0	0	0.0	1.0
1	1	0.4	0.6
1	0	0.6	0.4
0	1	1.0	0.0
1	1	0.6	0.4



X_1	C	$p(X_1 C)$
0	0	2,6/5,2 = 0,5
1	0	2,6/5,2 = 0,5
0	1	2,4/4,8 = 0,5
1	1	2,4/4,8 = 0,5

X_1	X_2	C	
		0	1
1	0	1.0	0.0
0	0	0.4	0.6
0	1	0.6	0.4
0	1	0.6	0.4
1	0	0.0	1.0
0	0	0.0	1.0
1	1	0.4	0.6
1	0	0.6	0.4
0	1	1.0	0.0
1	1	0.6	0.4

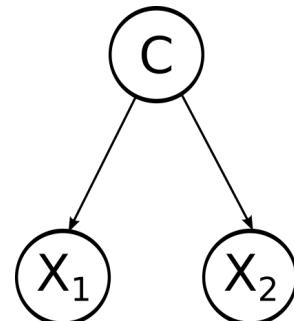


X_2	C	$p(X_2 C)$
0	0	
1	0	
0	1	
1	1	

EM

Re-aprender el modelo

X_1	X_2	C	
		0	1
1	0	1.0	0.0
0	0	0.4	0.6
0	1	0.6	0.4
0	1	0.6	0.4
1	0	0.0	1.0
0	0	0.0	1.0
1	1	0.4	0.6
1	0	0.6	0.4
0	1	1.0	0.0
1	1	0.6	0.4

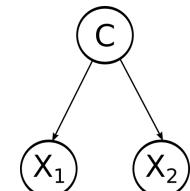


X_2	C	$p(X_2 C)$
0	0	2,0/5,2 = 0,385
1	0	3,2/5,2 = 0,615
0	1	3,0/4,8 = 0,625
1	1	1,8/4,8 = 0,375

EM

Re-estimación de pesos

X_1	X_2	C	
		0	1
1	0	1.0	0.0
0	0	0.4	0.6
0	1	0.6	0.4
0	1	0.6	0.4
1	0	0.0	1.0
0	0	0.0	1.0
1	1	0.4	0.6
1	0	0.6	0.4
0	1	1.0	0.0
1	1	0.6	0.4



C	$p(C)$
0	0.52
1	0.48

X_i	C	$p(X_1 C)$	$p(X_2 C)$
0	0	0.5	0.385
1	0	0.5	0.615
0	1	0.5	0.625
1	1	0.5	0.375

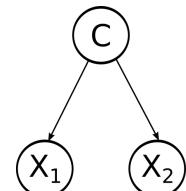
$$p(c = 0|\mathbf{x}) = \frac{1}{\theta} p(0) \prod_{i=1}^2 p(x_i|0)$$

$$p(c = 1|\mathbf{x}) = \frac{1}{\theta} p(1) \prod_{i=1}^2 p(x_i|1)$$

EM

Re-estimación de pesos

X_1	X_2	C	
		0	1
1	0	1.0	0.0
0	0	0.41	0.59
0	1	0.6	0.4
0	1	0.6	0.4
1	0	0.0	1.0
0	0	0.0	1.0
1	1	0.4	0.6
1	0	0.6	0.4
0	1	1.0	0.0
1	1	0.6	0.4



C	$p(C)$
0	0.52
1	0.48

X_i	C	$p(X_1 C)$	$p(X_2 C)$
0	0	0.5	0.385
1	0	0.5	0.615
0	1	0.5	0.625
1	1	0.5	0.375

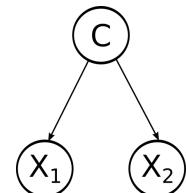
$$p(c = 0|\mathbf{x}) = \frac{1}{\theta} p(0) \prod_{i=1}^2 p(x_i|0)$$

$$p(c = 1|\mathbf{x}) = \frac{1}{\theta} p(1) \prod_{i=1}^2 p(x_i|1)$$

EM

Re-estimación de pesos

X_1	X_2	C	
		0	1
1	0	1.0	0.0
0	0	0.41	0.59
0	1	0.64	0.36
0	1	0.64	0.36
1	0	0.0	1.0
0	0	0.0	1.0
1	1	0.64	0.36
1	0	0.41	0.59
0	1	1.0	0.0
1	1	0.64	0.36



C	$p(C)$
0	0.52
1	0.48

X_i	C	$p(X_1 C)$	$p(X_2 C)$
0	0	0.5	0.385
1	0	0.5	0.615
0	1	0.5	0.625
1	1	0.5	0.375

$$p(c = 0|\mathbf{x}) = \frac{1}{\theta} p(0) \prod_{i=1}^2 p(x_i|0)$$

$$p(c = 1|\mathbf{x}) = \frac{1}{\theta} p(1) \prod_{i=1}^2 p(x_i|1)$$

En la práctica...

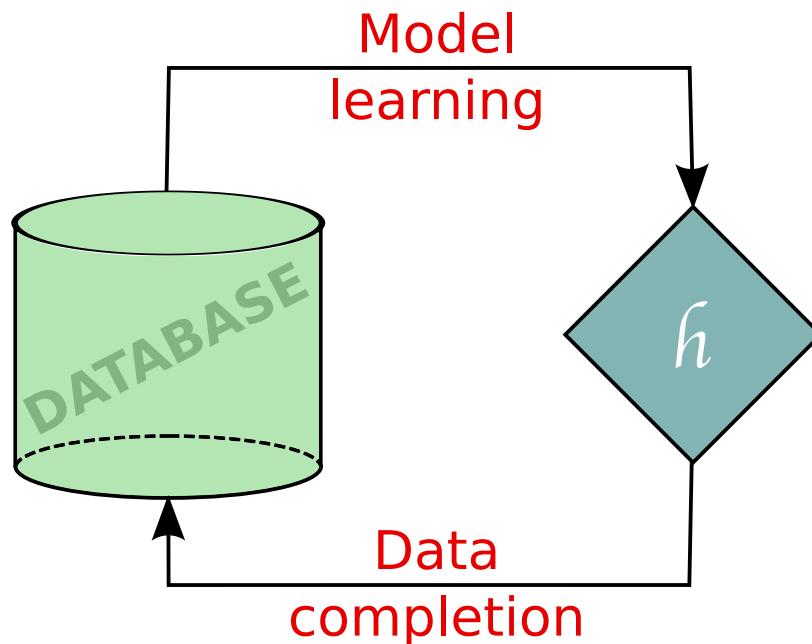
- ▶ Estimación de parámetros: Laplace Smoothing

$$\theta_i = \frac{N_i + 1}{N + |L|}$$

nº muestras clase "i"
nº muestras total
cardinalidad de la clase
(número de clases que tenemos)

- ▶ Cálculo de probabilidades: cálculo logarítmico

$$\hat{c} = \operatorname{argmáx}_c \exp \left[\log p(c) + \sum_{i=1}^2 \log p(x_i|c) \right]$$



* Paradigma de teacher - student.

Aprendizaje no supervisado

VC09: Análisis de Componentes Principales

Rocío del Amor del Amor
mrocio.delamor@campusviu.es

Universidad Internacional de Valencia

Cuestiones previas

- ▶ Datos originales vs. datos transformados
 - con PCA ▶↓ Interpretabilidad vs. ↑utilidad
 - ▶ Datos completos vs. pérdida de información
 - ▶ Gran cantidad de datos vs. Cantidad manejable de datos

Cuestiones previas

Valor medio y valor esperado

Dada una variable aleatoria X , el **valor esperado** de X es:

$$E[X] = \sum_{x \in \mathcal{X}} x \cdot P(X = x)$$

Dada una muestra S de X , el **valor medio** de S es:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n S_i$$

Cuestiones previas

Variance

Dada una variable aleatoria X , la **varianza** de X es:

$$\text{Var}(X) = E[(X - E[X])^2]$$

donde $E[X] = \sum_{x \in \mathcal{X}} x \cdot P(X = x)$

Dada una muestra S de X , la **varianza** de S es:

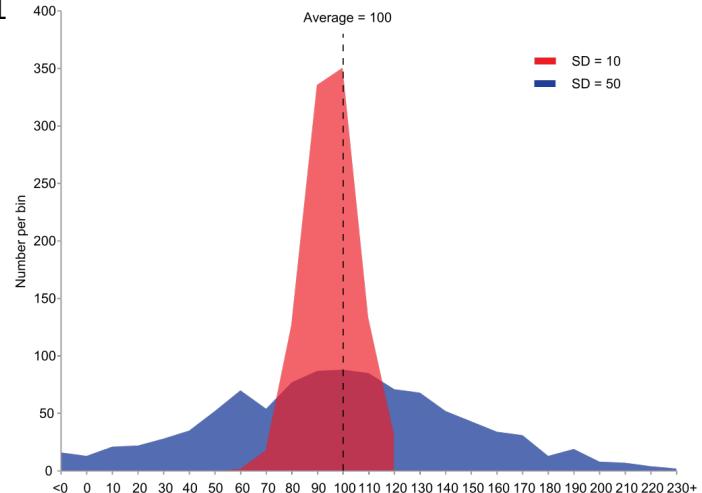
$$\sigma_n^2 = \frac{1}{n} \sum_{i=1}^n (S_i - \bar{X})^2$$

Ejemplos:

X : Altura de estudiantes

X : Edad

X : Horas de estudio



Justificación, necesidad de la transformación

Probablemente, el principal uso del análisis de componentes es la reducción de dimensionalidad

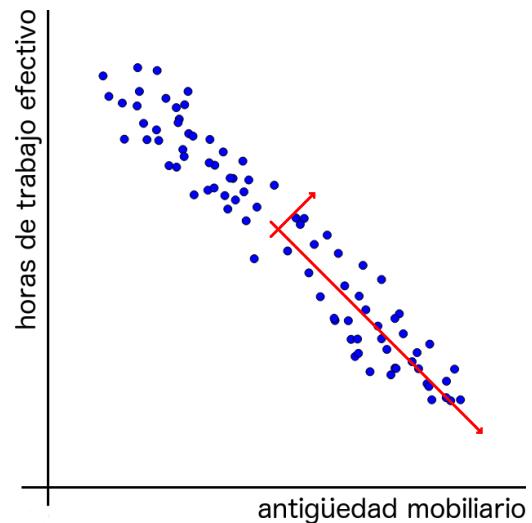
Expresar los mismos datos, con la menor pérdida de información posible, a través de un menor número de variables.

Otra info: descubrimiento de relaciones ocultas entre variables, espacio más apropiado para la aplicación de ciertas técnicas de análisis, etc.

Ejemplo

Estudio del rendimiento de trabajadores/as

- ▶ Variables: No. horas trabajadas, Antigüedad del material, Comodidad, Movilidad en el puesto de trabajo, Rendimiento, etc.
- ▶ Si el número de horas de trabajo real está directamente relacionado con la antigüedad del material, la relación puede quedar escondida a simple vista
- ▶ Mediante análisis de componentes, se descubriría la relación entre ellas y la presencia de información redundante



Análisis de Componentes Principales (PCA)

[Examen] Diferencia entre SE y PCA

→ NO lineal

Cierre 19/10/2023

lineal.

hora 21:18 - 21:19

Idea

- ▶ La idea es crear un conjunto de variables nuevo (reducido) que representen la misma información (pero no la va a representar toda)
- ▶ Serie de componentes (variables) ortogonales que explican, cada vez en menor medida, una porción de la información
Podríamos decir: PCA obtiene representaciones comprimidas de los datos
- ▶ Las componentes que explican en menor medida los datos se eliminan para conseguir la reducción de dimensionalidad
- ▶ Efectivo contra el ruido y los valores extraños
- ▶ La representación en espacios alternativos puede ser útil para ciertos tipos de técnicas de análisis

* Las nuevas variables son combinaciones lineales de las variables antiguas

PCA

→ componentes principales

- CPs: serie de proyecciones de los datos **mutuamente no correlacionadas, ordenadas** según la cantidad de varianza de los datos originales que explican
- Cada CP es el eje que mejor explica la mayor porción de varianza no explicada

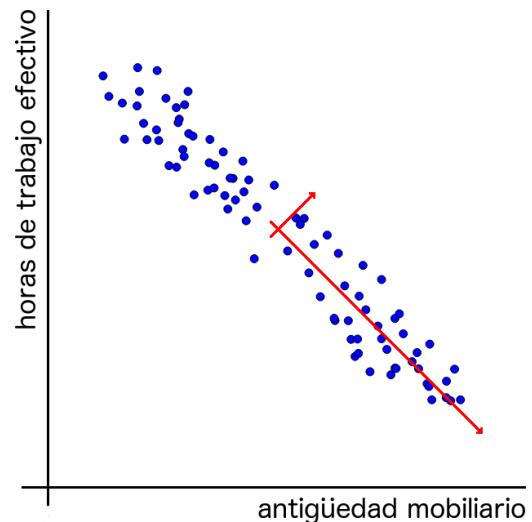
CP.1: Explica la mayor cantidad de varianza

CP.2: Ortogonal a CP.1, es el eje que explica la mayor cantidad de varianza no explicada por CP.1

CP.3: Ortogonal a CP.1 y CP.2, es el eje que explica la mayor cantidad de varianza no explicada por CP.1 ni por CP.2

...

CPs = # Variables originales



PCA

Paso previo: Estandarización de los datos

Objetivo

Conseguir que todas las variables originales tengan el mismo rango.

1. Centrar las variables (media = 0):

estandarización

$$\mathbf{x}_i \leftarrow \mathbf{x}_i - \frac{1}{n} \sum_{i'=1}^n \mathbf{x}_{i'} , \forall i \in \{1, \dots, n\}$$

2. Re-escalar las variables (varianza = 1):

$$x_{ij} \leftarrow x_{ij} / \sqrt{\frac{1}{n} \sum_{i'=1}^n (x_{i'j})^2} , \forall i \in \{1, \dots, n\} \wedge j \in \{1, \dots, v\}$$

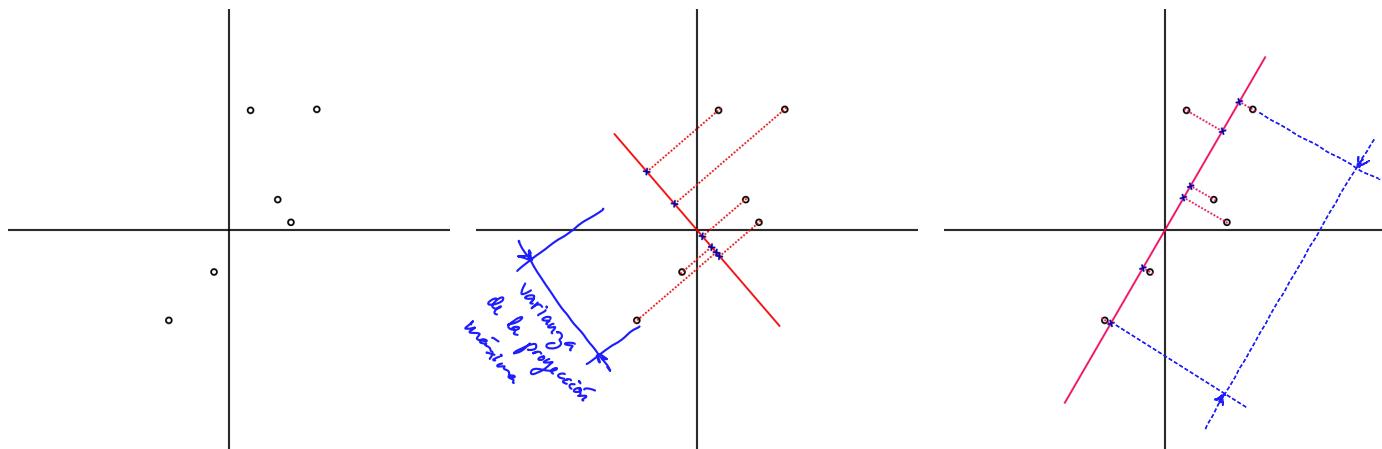
** Evitar que las variables de mayor rango dominen las de menor rango **

Objetivo

Encontrar la dirección sobre la que mejor se expresan los datos

[Examen]

Buscar un vector u tal que si los datos se proyectan en esa dirección, la varianza de la proyección es máxima



Varianza de una proyección

Buscar el vector \mathbf{u} que maximiza la varianza sobre todo el conjunto de datos, $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$:

$$\frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i^t \mathbf{u})^2 = \frac{1}{n} \sum_{i=1}^n \mathbf{u}^t \mathbf{x}_i \mathbf{x}_i^t \mathbf{u} = \mathbf{u}^t \left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^t \right) \mathbf{u} = \mathbf{u}^t \Sigma \mathbf{u}$$

donde $\Sigma = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^t$ es la matriz de covarianza.

Tarea

El problema se define como:

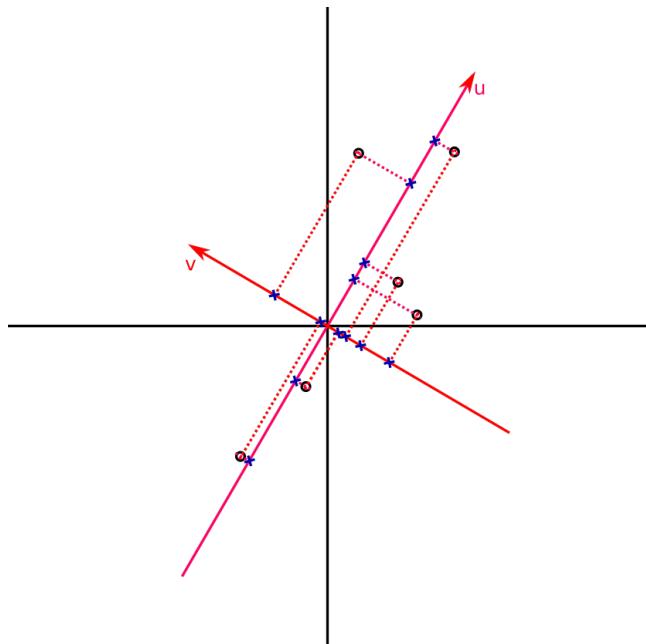
$$\arg \max_{\mathbf{u}} \mathbf{u}^t \Sigma \mathbf{u}$$

Respuesta: El vector propio principal de Σ

¡Los **vectores propios** de Σ son los vectores ortogonales que buscamos!

Procedimiento

1. Calcular la matriz de covarianzas, Σ
2. Descomponer Σ en vectores propios
(descomposición en valores singulares)
3. Seleccionar los q vectores propios principales como CPs
(los q vectores propios con mayor valor propio asociado)



- ▶ Cada componente principal \mathbf{u}_j es una combinación lineal de las variables originales
- ▶ El nuevo conjunto de datos Z en el espacio transformado es:

$$z_i = \begin{bmatrix} \mathbf{u}_1^t \mathbf{x}_i \\ \mathbf{u}_2^t \mathbf{x}_i \\ \vdots \\ \mathbf{u}_q^t \mathbf{x}_i \end{bmatrix}, \quad \forall i \in \{1, \dots, n\}$$

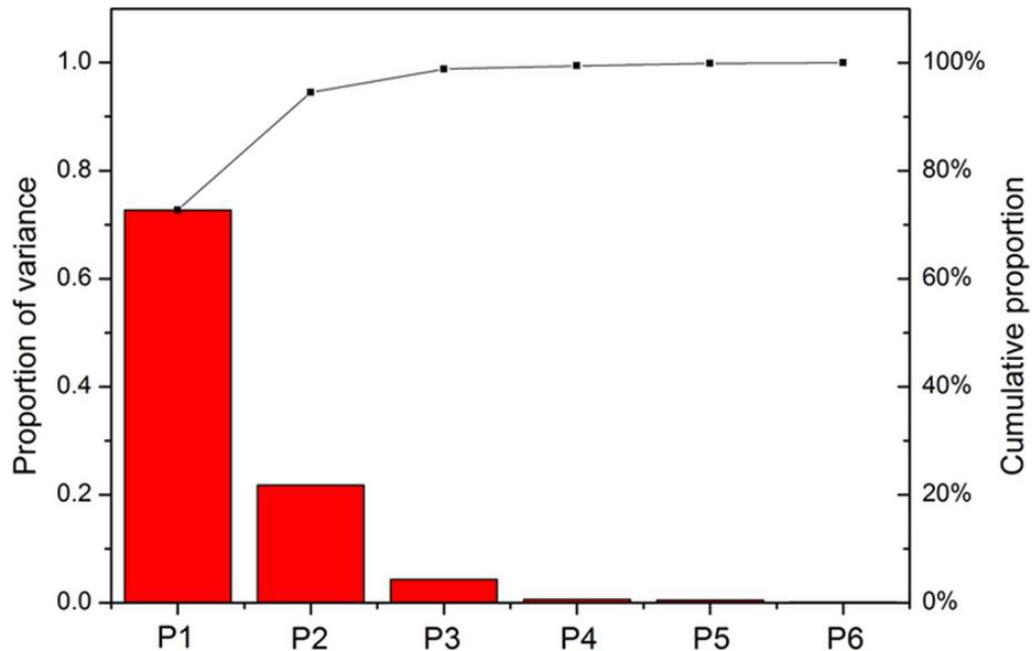
- ▶ La reducción de la dimensionalidad depende del número de componentes principales (q).

La reducción de la dimensionalidad depende de q
(número de componentes principales)

- ▶ Si $q = v$, no hay pérdida de información ni reducción de dimensionalidad
- ▶ A menor q , mayor reducción de dimensionalidad y pérdida de información
- ▶ Ritmo de pérdida de información: depende de la redundancia de las variables y las relaciones ocultas en los datos

PCA

Seleccionando q



1. Fijar un umbral s (ej., 95) en el acumulado de varianza explicada
2. Seleccionar las q CPs que expliquen al menos el $s\%$ de la varianza total de los datos

PCA

Pros y contras

Ventajas

- ▶ Técnica no paramétrica
- ▶ Único parámetro ajustable (posterior): número de componentes q
- ▶ En el espacio de optimización no existen máximos locales donde el método pudiese quedar atrapado

Desventajas

- ▶ El nuevo espacio puede no ser intuitivo
- ▶ La interpretabilidad de las variables se pierde (oscurece)
- ▶ Se limita a momentos muestrales de orden 2 (varianza) y a proyecciones lineales

Aprendizaje no supervisado

VC09: Análisis de Componentes Independientes

Rocío del Amor del Amor
mrocio.delamor@campusviu.es

Universidad Internacional de Valencia

ICA vs. PCA

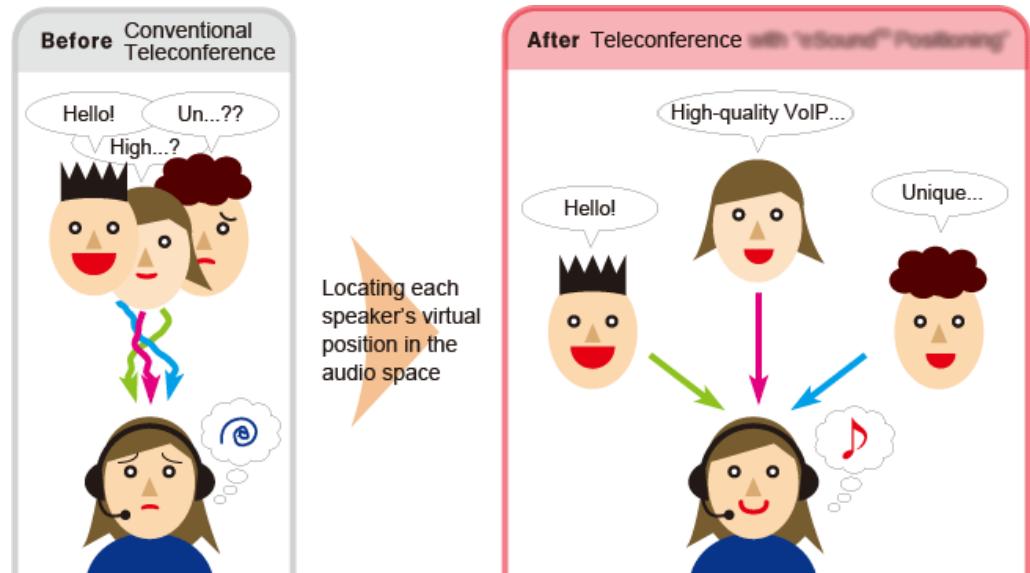
- ▶ Mismo resultado (datos transformados)
- ▶ Distinta hipótesis
- ▶ Reducción de dimensionalidad (dimensiones que mejor explican los datos)
- ▶ Separación de las fuentes/variables originales

Ejemplo clásico: Fiesta-cóctel

Personas que hablan simultáneamente

Una persona escuchando (ej., micrófonos) en diferentes puntos de la sala captaría una mezcla diferente de mensajes

¿se pueden separar las diferentes voces (fuentes) para descifrar los distintos mensajes?



Basic:

<https://www.youtube.com/watch?v=wIlrrddNbXDo>

Intermediate:

<https://www.youtube.com/watch?v=pSwR05d266I>

Final:

<https://www.youtube.com/watch?v=e4woe8GRjEI>

Definición del problema

- ▶ Existe un grupo de u **fuentes independientes**
- ▶ Un vector original es $\mathbf{s} \in \mathbb{R}^u$, un punto de cada fuente en un instante concreto
- ▶ Existe un grupo de v **receptores**
- ▶ El vector observado es $\mathbf{x} \in \mathbb{R}^v$
- ▶ Se asume la existencia de una **matriz de mezcla** A tal que:

$$\mathbf{x} = A\mathbf{s}$$

o una **matriz de separación**, W ($W = A^{-1}$):

$$\mathbf{s} = W\mathbf{x}$$

Problema de optimización

Encontrar el vector w_j que permite recomponer la j -ésima fuente:

$$s_j = \mathbf{w}_j^t \mathbf{x}$$

para todo $j \in \{1, \dots, u\}$

Problema de optimización

Encontrar el vector w_j que permite recomponer la j -ésima fuente:

$$s_j = \mathbf{w}_j^t \mathbf{x}$$

para todo $j \in \{1, \dots, u\}$

Solución

Buscaremos el estimador máximo verosímil de W

Propiedad de la suposición de independencia

La probabilidad conjunta es igual al producto de las distribuciones marginales de las fuentes:

$$p(\mathbf{s}) = \prod_{j=1}^u p_s(s_j)$$

Expresado según lo observado, \mathbf{x} :

$$p_x(\mathbf{x}) = \prod_{j=1}^u p_s(\mathbf{w}_j^t \mathbf{x}) \cdot |W|$$

el determinante de W hace que la distribución de probabilidad integre a 1

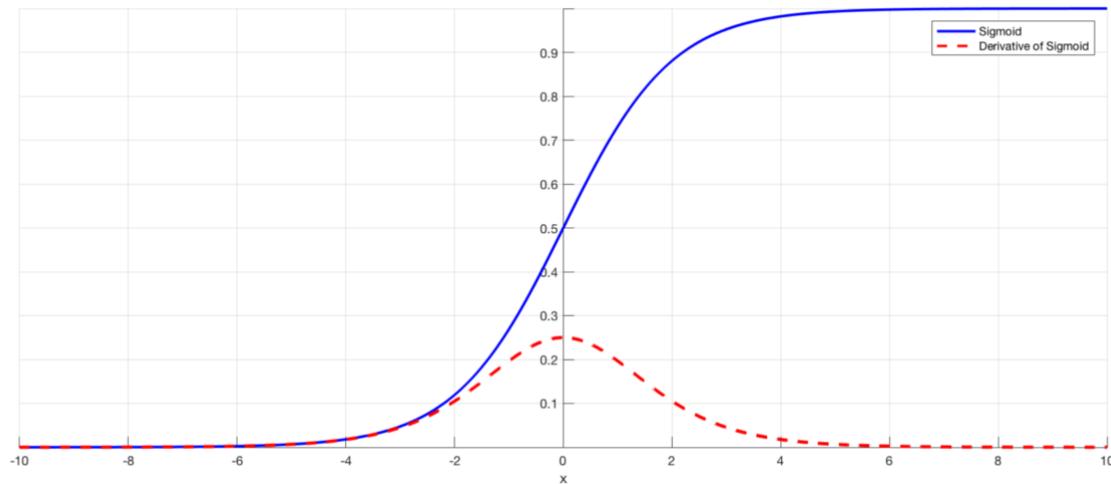
Usaremos la función sigmoide como distribución cumulativa:

$$g(s_j) = 1/(1 + e^{-s_j})$$

** Una distr. cumulativa es una función monótona que crece suavemente de 0 a 1

Su derivada es la función de densidad:

$$p_s(s_j) = g'(s_j)$$



Verosimilitud:

$$L(W; \{\mathbf{x}_1, \dots, \mathbf{x}_n\}) = \prod_{i=1}^n \left(\prod_{j=1}^u g'(\mathbf{w}_j^t \mathbf{x}_i) \cdot |W| \right)$$

Logaritmo de la verosimilitud:

$$\log L(W; \{\mathbf{x}_1, \dots, \mathbf{x}_n\}) = \sum_{i=1}^n \left(\sum_{j=1}^u \log g'(\mathbf{w}_j^t \mathbf{x}_i) + \log |W| \right)$$

Estimador máximo-verosímil de W

Algoritmo de ascenso de gradiente estocástico

- ▶ Obtener una matriz W inicial
- ▶ Iterativamente y para cada caso observado \mathbf{x}_i , tomar un paso en la dirección de máximo ascenso dada por el gradiente*:

$$W \leftarrow W + \alpha \begin{pmatrix} \begin{bmatrix} 1 - 2g(\mathbf{w}_1^t \mathbf{x}_i) \\ 1 - 2g(\mathbf{w}_2^t \mathbf{x}_i) \\ \vdots \\ 1 - 2g(\mathbf{w}_n^t \mathbf{x}_i) \end{bmatrix} & \mathbf{x}_i^t + (W^t)^{-1} \end{pmatrix}$$

donde α (ratio de aprendizaje) determina el paso de actualización de W

- ▶ Converge a un máximo local
(W no cambia sustancialmente entre iterat. consecutivas)

* generalización de la derivada a múltiples dimensiones

Dada una estimación (máximo verosímil) de la matriz W , los valores de las fuentes se obtienen a partir de una observación \mathbf{x}_i :

$$\mathbf{s}_i = W\mathbf{x}_i$$

- ▶ Suposición de independencia entre observaciones no realista
- ▶ Aun así, si el conjunto de datos suficientemente grande, el rendimiento del algoritmo no se ve comprometido
Si la correlación entre muestras es evidente, se realiza un recorrido aleatorio por las observaciones x_i en el ascenso del gradiente estocástico (acelerar la convergencia)
- ▶ El algoritmo converge a un óptimo local
- ▶ Dependiendo del ratio de aprendizaje, α , el algoritmo puede escapar de los óptimos locales con cierta facilidad
Hacer una exploración aleatoria del conjunto de datos en cada iteración ayuda a prevenir este problema

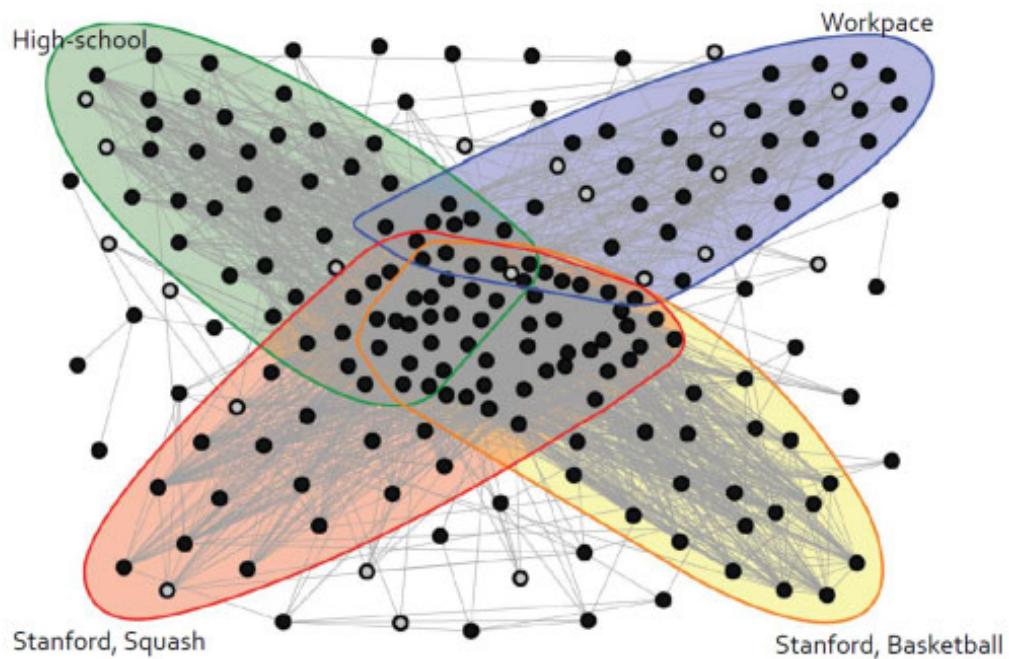
Aprendizaje no supervisado

VC10: Análisis de grafos, PageRank y otros

Rocío del Amor del Amor
mrocio.delamor@campusviu.es

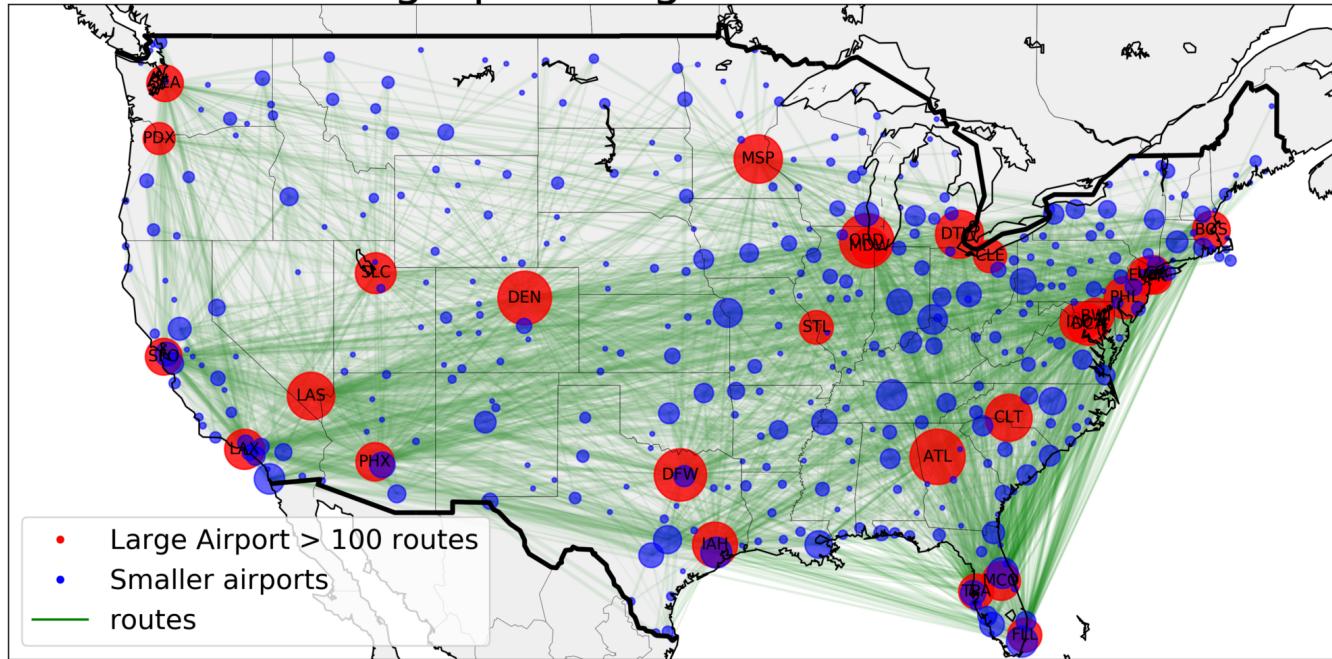
Universidad Internacional de Valencia

La Web como grafo

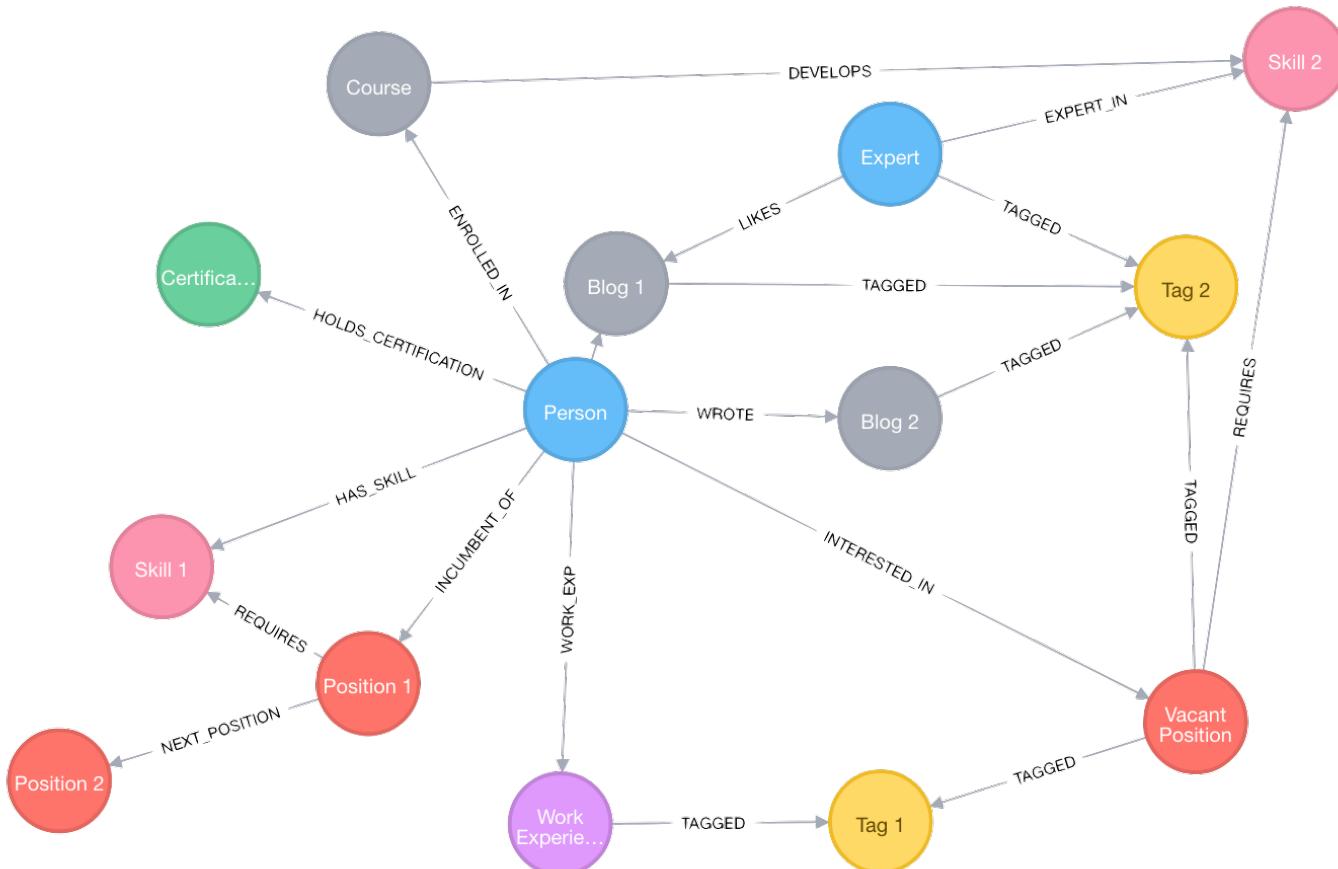


La Web como grafo

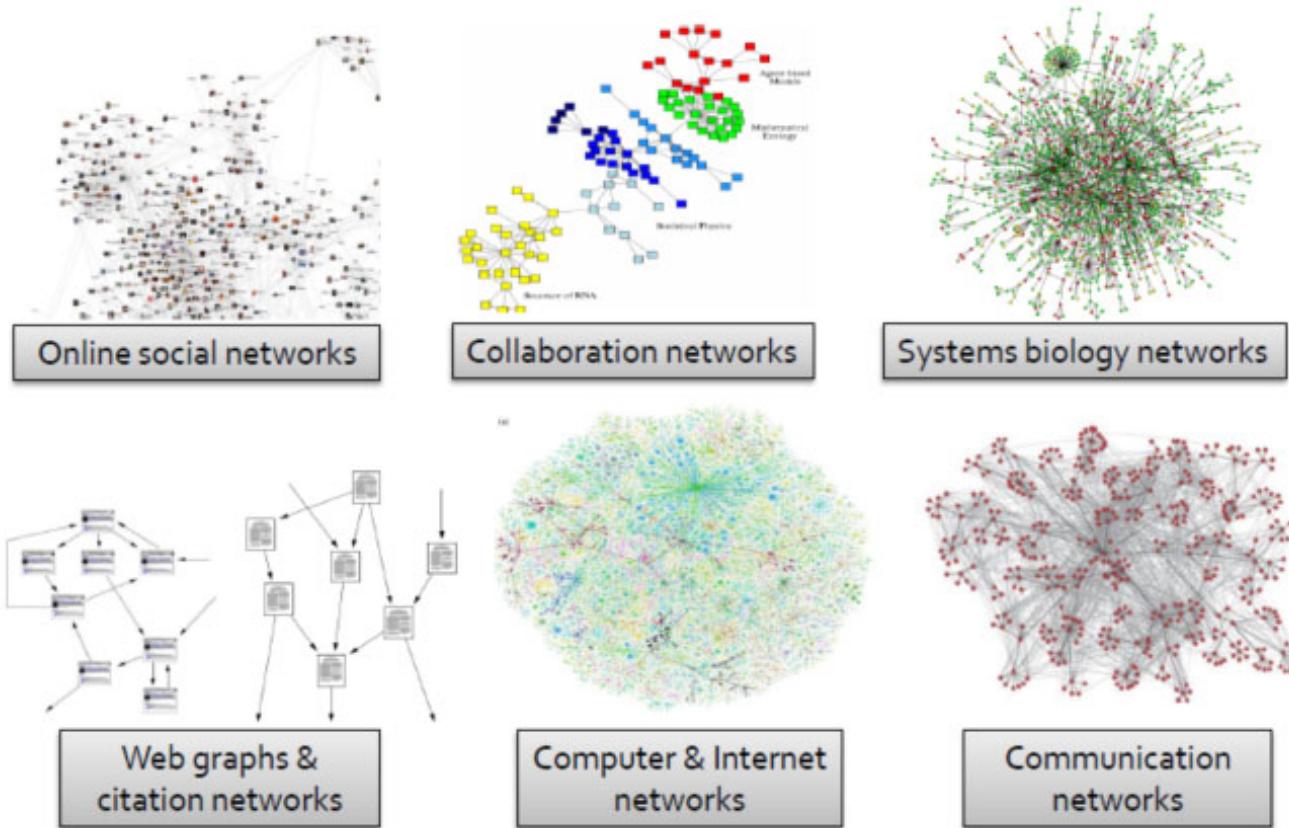
Network graph of flight routes in the USA



La Web como grafo



La Web como grafo



La Web como grafo



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item

Donald Trump

Donald John Trump (born June 14, 1946) is an American businessman, television producer, and politician who is the **Republican Party** nominee for President of the United States in the 2016 election. He is the chairman and president of **The Trump Organization**, which is the principal holding company for his real estate ventures and other business interests. During his career, Trump has built office towers, hotels, casinos, golf courses, an urban development project in Manhattan, and other branded facilities worldwide.

Trump was born and raised in **New York City** and received a **bachelor's degree** in economics from the **Wharton School** of the **University of Pennsylvania** in 1968. In 1971 he was given control of his father **Fred Trump**'s real estate and construction firm and later renamed it **The Trump Organization**, rising to public prominence shortly thereafter. Trump has appeared at the **Miss USA** pageants, which he owned from 1996 to 2015, and has made cameo appearances in films and television series. He sought the **Reform Party** presidential nomination in 2000, but withdrew before voting began. He hosted and co-produced **The Apprentice**, a reality television series on NBC, from 2004 to 2015. As of 2016, he was listed by **Forbes** as the 324th wealthiest person in the world, and 156th in the United States.

In June 2015, Trump announced his candidacy for president as a Republican and quickly emerged as the front-runner for his party's nomination. In May 2016, his remaining Republican rivals suspended their campaigns, and in July he was formally nominated for president at the **2016 Republican National Convention**. Trump's campaign has received unprecedented media coverage and international attention. Many of his statements in interviews, on Twitter, and at campaign rallies have been controversial or false. Several rallies during the primaries were accompanied by protests or riots. On October 7, a 2005 audio recording surfaced in which Trump bragged about forcibly kissing and groping women; at least fifteen women accused him of similar conduct shortly thereafter.^{[5][6]} He apologized for the 2005 comments and denied the allegations.

Trump's platform includes renegotiation of U.S.–China trade deals, opposition to particular trade agreements such as **NAFTA** and the **Trans-Pacific Partnership**, stronger enforcement of immigration laws together with "deportation of the bad guys", border wall, review of executive orders, repeal and replacement of the **Affordable Care Act**.

339. ^ Linski, Jack (July 7, 2015). "More People Are Running for Presidential Nomination Than Ever". *Time*. Retrieved February 14, 2016.

340. ^ Howell, Kellan (March 9, 2016). "Donald Trump helps GOP presidential debates break TV ratings records". *Time*. Retrieved October 8, 2016.

341. ^ "Donald Trump, Ted Cruz Angling For One-On-One Republican Race". *Fortune*. March 6, 2016.

342. ^ Bump, Philip (March 23, 2016). "Why Donald Trump is poised to win the nomination and lose the general election, in one poll". *The Washington Post*.

343. ^ Nussbaum, Matthew (May 3, 2016). "RNC Chairman: Trump is our nominee". *Politico.com*. Retrieved May 4, 2016.

344. ^ Bump, Philip. "Trump got the most GOP votes ever — both for and against him — and other fun facts". *The Washington Post*. Retrieved July 12, 2016.

345. ^ Berenson, Tessa (May 5, 2016). "Donald Trump Tells West Virginia Primary Voters to Stay Home". *Time*.

346. ^ "Fuller picture emerges of man arrested at Trump rally". *Associated Press*.

Donald Trump



Trump in August 2015

Born Donald John Trump
June 14, 1946 (age 70)
Queens, New York City, U.S.

Residence Manhattan, New York City

Alma mater Fordham University
University of Pennsylvania

Occupation Businessman · television
producer · politician

Net worth US\$3.7 billion (2016)^{[1][2]}

party 2011, 2012–present

705. ^ "Algemeiner Honors Joan Rivers, Donald Trump, Yuli Edelstein at Second Annual 'Jewish 100' Gala". *Algemeiner Journal*. Brooklyn, NY. February 5, 2015.

707. ^ Hascup, Henry (March 27, 2015). "2015 New Jersey Boxing Hall of Fame Inductees". *New Jersey Boxing Hall of Fame*.

708. ^ "MC-LEF Events". Marine Corps-Law Enforcement Foundation. 2015. Archived from the original on August 19, 2015. "Donald Trump received our Commandant's Leadership Award."

709. ^ Madan, Monique (March 4, 2015). "Donald Trump gets his key to Doral". *The Miami Herald*. Miami. Archived from the original on July 8, 2015. Retrieved August 19, 2015.

710. ^ Hidalgo, Daniel (August 5, 2015). "Doral lets Donald Trump keep key to city; also gives initial OK to four new developments". *The Miami Herald*. Miami. Archived from the original on August 19, 2015. Retrieved August 19, 2015.

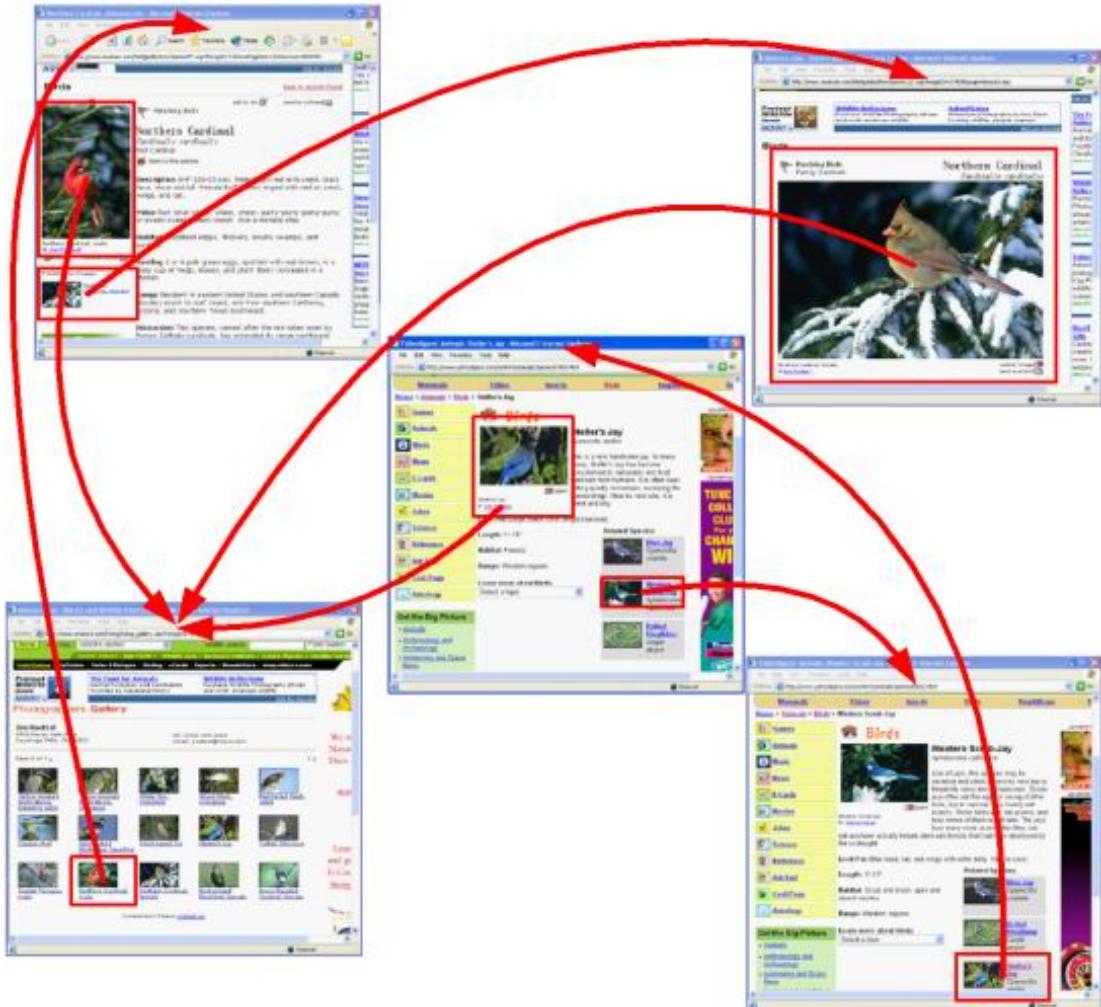
711. ^ Ellmers, Renee (April 21, 2016). "Donald Trump: 'The rule breaker'". *Time*.

External links

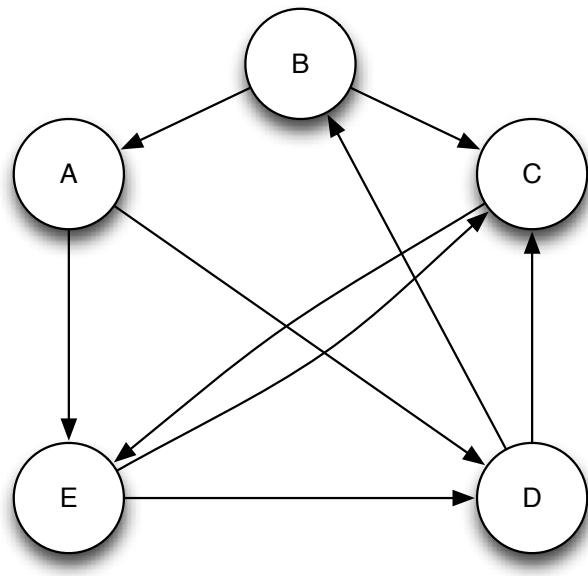
- Official website
- Donald Trump at the Internet Movie Database

Library resources about
Donald Trump

La Web como grafo



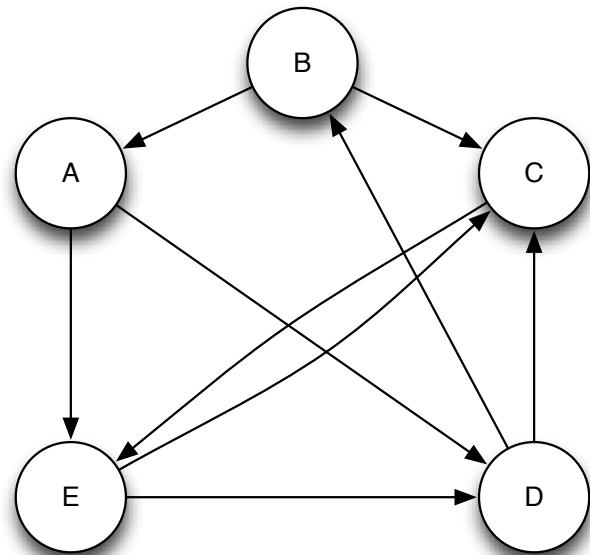
Grafo



Matriz de adyacencia

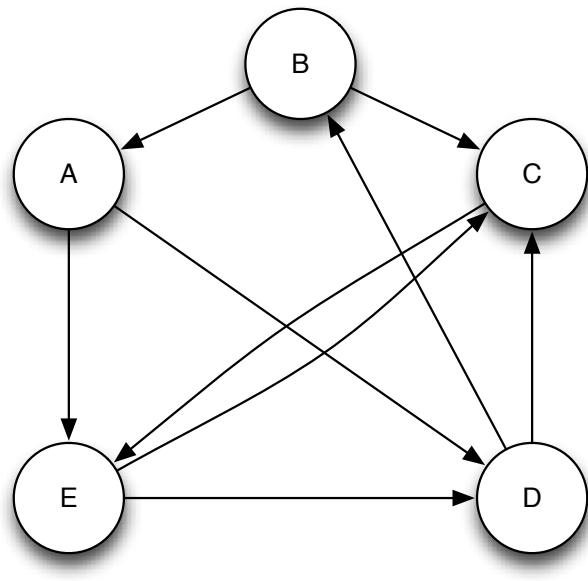
$$D = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Grafo



In-degree
Out-degree

Grafo



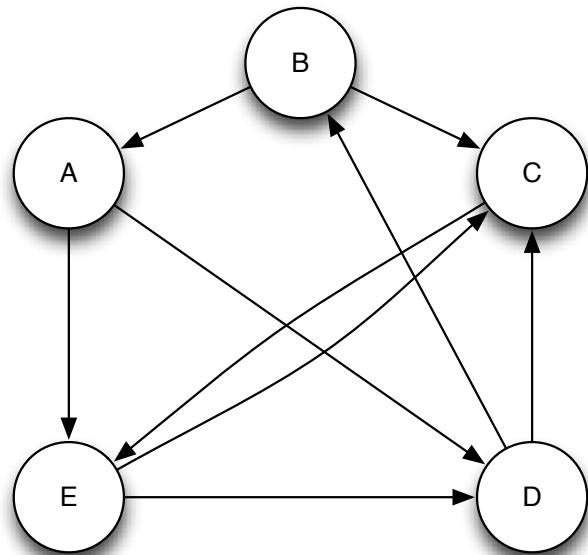
Matriz de transición

suman 1 por columnas.
↑
in-degree → (popularidad
de la página)
out-degree

$$A = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix}$$

probabilidades de ir de una a otra

Grafo



Descripción / clustering /
clasificación de grafos

Subgrafo

A priori algorithm

PageRank

Definición probabilista

El PageRank de una página web es la probabilidad de que un internauta acabe en dicha web tras navegar aleatoriamente por Internet desde un punto de inicio al azar.

Random Walk

Otra definiciones

- ▶ Álgebra
- ▶ Sistema dinámico

PageRank

Álgebra

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix} \cdot \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix} = v \cdot \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

$$a_{11} \cdot e_1 + a_{12} \cdot e_2 + a_{13} \cdot e_3 + a_{14} \cdot e_4 + a_{15} \cdot e_5 = v \cdot e_1$$

$$a_{21} \cdot e_1 + a_{22} \cdot e_2 + a_{23} \cdot e_3 + a_{24} \cdot e_4 + a_{25} \cdot e_5 = v \cdot e_2$$

$$a_{31} \cdot e_1 + a_{32} \cdot e_2 + a_{33} \cdot e_3 + a_{34} \cdot e_4 + a_{35} \cdot e_5 = v \cdot e_3$$

$$a_{41} \cdot e_1 + a_{42} \cdot e_2 + a_{43} \cdot e_3 + a_{44} \cdot e_4 + a_{45} \cdot e_5 = v \cdot e_4$$

$$a_{51} \cdot e_1 + a_{52} \cdot e_2 + a_{53} \cdot e_3 + a_{54} \cdot e_4 + a_{55} \cdot e_5 = v \cdot e_5$$

Cualquier matriz estocástica izquierda (columna) tiene 1 como valor propio

PageRank

Álgebra

$$\begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

$$e_1 = a_{11} \cdot e_1 + \frac{1}{2} \cdot e_2 + a_{13} \cdot e_3 + a_{14} \cdot e_4 + a_{15} \cdot e_5$$

$$e_2 = a_{21} \cdot e_1 + a_{22} \cdot e_2 + a_{23} \cdot e_3 + \frac{1}{2} \cdot e_4 + a_{25} \cdot e_5$$

$$e_3 = a_{31} \cdot e_1 + \frac{1}{2} \cdot e_2 + a_{33} \cdot e_3 + \frac{1}{2} \cdot e_4 + \frac{1}{2} \cdot e_5$$

$$e_4 = \frac{1}{2} \cdot e_1 + a_{42} \cdot e_2 + a_{43} \cdot e_3 + a_{44} \cdot e_4 + \frac{1}{2} \cdot e_5$$

$$e_5 = \frac{1}{2} \cdot e_1 + a_{52} \cdot e_2 + 1 \cdot e_3 + a_{54} \cdot e_4 + a_{55} \cdot e_5$$

PageRank

Álgebra

$$\begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

$$e_1 = \frac{1}{4} \cdot e_4$$

$$e_2 = \frac{1}{2} \cdot e_4$$

$$e_3 = \frac{1}{4} \cdot e_4 + \frac{1}{2} \cdot e_4 + \frac{1}{2} \cdot e_5$$

$$e_4 = \frac{1}{2} \cdot e_1 + \frac{1}{2} \cdot e_5$$

$$e_5 = \frac{1}{2} \cdot e_1 + e_3$$

PageRank

Álgebra

$$\begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

$$e_1 = \frac{1}{4} \cdot e_4$$

$$e_2 = \frac{1}{2} \cdot e_4$$

$$e_3 = \frac{3}{4} \cdot e_4 + \frac{1}{2} \cdot e_5$$

$$e_4 = \frac{1}{2} \cdot e_1 + \frac{1}{2} \cdot e_5$$

$$e_5 = \frac{1}{2} \cdot e_1 + e_3$$

PageRank

Álgebra

$$\begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

$$e_1 = \frac{1}{4} \cdot e_4$$

$$e_2 = \frac{1}{2} \cdot e_4$$

$$e_3 = \frac{3}{4} \cdot e_4 + \frac{1}{2} \cdot e_5$$

$$e_4 = \frac{1}{8} \cdot e_4 + \frac{1}{2} \cdot e_5$$

$$e_5 = \frac{1}{8} \cdot e_4 + e_3$$

PageRank

Álgebra

$$\begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

$$e_1 = \frac{1}{4} \cdot e_4$$

$$e_2 = \frac{1}{2} \cdot e_4$$

$$e_3 = \frac{3}{4} \cdot e_4 + \frac{1}{2} \cdot e_5$$

$$\frac{7}{8} \cdot e_4 = \frac{1}{2} \cdot e_5$$

$$e_5 = \frac{1}{8} \cdot e_4 + e_3$$

PageRank

Álgebra

$$\begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

$$e_1 = \frac{1}{4} \cdot e_4$$

$$e_2 = \frac{1}{2} \cdot e_4$$

$$e_3 = \frac{3}{4} \cdot e_4 + \frac{7}{8} \cdot e_4$$

$$e_4 = e_4$$

$$e_5 = \frac{7}{4} \cdot e_4$$

PageRank

Álgebra

$$\begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

$$e_1 = \frac{1}{4} \cdot e_4$$

$$e_2 = \frac{1}{2} \cdot e_4$$

$$e_3 = \frac{13}{8} \cdot e_4$$

$$e_4 = e_4$$

$$e_5 = \frac{7}{4} \cdot e_4$$

PageRank

Álgebra

$$\begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix} = \begin{bmatrix} \frac{1}{4} \cdot e_4 \\ \frac{1}{2} \cdot e_4 \\ \frac{13}{8} \cdot e_4 \\ e_4 \\ \frac{7}{4} \cdot e_4 \end{bmatrix}$$

$$e_1 = \frac{1}{4} \cdot e_4$$

$$e_2 = \frac{1}{2} \cdot e_4$$

$$e_3 = \frac{13}{8} \cdot e_4$$

$$e_4 = e_4$$

$$e_5 = \frac{7}{4} \cdot e_4$$

PageRank

Álgebra

$$\begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix} = \frac{e_4}{8} \cdot \begin{bmatrix} 2 \\ 4 \\ 13 \\ 8 \\ 14 \end{bmatrix}$$

$$e_1 = \frac{1}{4} \cdot e_4$$

$$e_2 = \frac{1}{2} \cdot e_4$$

$$e_3 = \frac{13}{8} \cdot e_4$$

$$e_4 = e_4$$

$$e_5 = \frac{7}{4} \cdot e_4$$

PageRank

Álgebra

$$\begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 4 \\ 13 \\ 8 \\ 14 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \\ 13 \\ 8 \\ 14 \end{bmatrix}$$

$$e_1 = \frac{1}{4} \cdot e_4$$

$$e_2 = \frac{1}{2} \cdot e_4$$

$$e_3 = \frac{13}{8} \cdot e_4$$

$$e_4 = 8$$

$$e_5 = \frac{7}{4} \cdot e_4$$

PageRank

Álgebra

$$\begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0,049 \\ 0,098 \\ 0,317 \\ 0,195 \\ 0,341 \end{bmatrix} = \begin{bmatrix} 0,049 \\ 0,098 \\ 0,317 \\ 0,195 \\ 0,341 \end{bmatrix}$$

$$e_1 = \frac{1}{4} \cdot e_4$$

$$e_2 = \frac{1}{2} \cdot e_4$$

$$e_3 = \frac{13}{8} \cdot e_4$$

$$e_4 = \frac{8}{41}$$

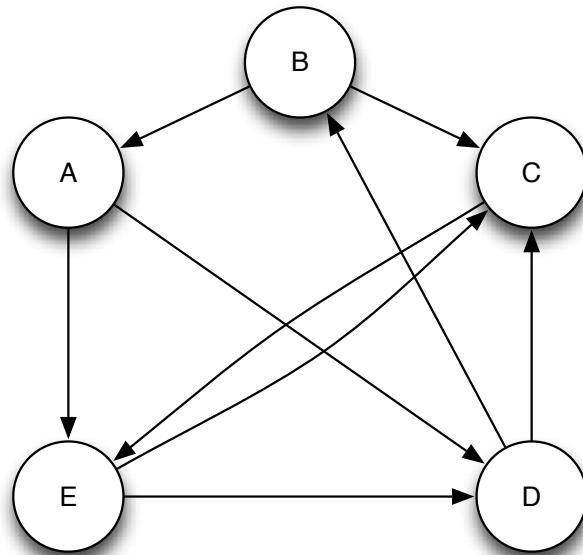
$$e_5 = \frac{7}{4} \cdot e_4$$

PageRank

Sistema dinámico

$$A = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$A^2 = \begin{bmatrix} 0 & 0 & 0 & \frac{1}{4} & 0 \\ \frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 & 0 \\ 0 & \frac{3}{4} & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$



PageRank

Sistema dinámico

$$A = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix}$$

Inicio aleatorio

$$F = \begin{bmatrix} \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \end{bmatrix}$$
$$A^2 = \begin{bmatrix} 0 & 0 & 0 & \frac{1}{4} & 0 \\ \frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 & 0 \\ 0 & \frac{3}{4} & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

PageRank

Sistema dinámico

$$A = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$F = \begin{bmatrix} \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \end{bmatrix}$$

$$A \cdot F = \begin{bmatrix} \frac{1}{10} \\ \frac{1}{10} \\ \frac{3}{10} \\ \frac{2}{10} \\ \frac{3}{10} \end{bmatrix}$$

PageRank

Sistema dinámico

$$A = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix} \quad A^2 = \begin{bmatrix} 0 & 0 & 0 & \frac{1}{4} & 0 \\ \frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 & 0 \\ 0 & \frac{3}{4} & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \quad F = \begin{bmatrix} \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \end{bmatrix}$$

$$A^2 \cdot F = \begin{bmatrix} \frac{1}{20} \\ \frac{1}{10} \\ \frac{3}{10} \\ \frac{2}{10} \\ \frac{7}{20} \end{bmatrix}$$

PageRank

Sistema dinámico

$$A = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix} \quad A^3 = \begin{bmatrix} \frac{1}{8} & 0 & 0 & 0 & \frac{1}{8} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{4} & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{3}{8} \\ 0 & \frac{3}{8} & 0 & \frac{3}{8} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} & \frac{3}{8} & \frac{1}{4} \end{bmatrix} \quad F = \begin{bmatrix} \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \end{bmatrix}$$

$$A^3 \cdot F = \begin{bmatrix} \frac{1}{20} \\ \frac{1}{10} \\ \frac{13}{40} \\ \frac{1}{5} \\ \frac{13}{40} \end{bmatrix}$$

PageRank

Sistema dinámico

$$A = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix} \quad A^* = \begin{bmatrix} & & & & \\ & & & & \\ & & \dots & & \\ & & & & \\ & & & & \end{bmatrix} \quad F = \begin{bmatrix} \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \end{bmatrix}$$

$$A^* \cdot F = \begin{bmatrix} & & & & \\ & & & & \\ & & \dots & & \\ & & & & \end{bmatrix}$$

PageRank

Sistema dinámico

$$A = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix} \quad A^6 = \begin{bmatrix} 0,08 & 0 & 0,06 & 0,05 & 0,05 \\ 0,09 & 0,14 & 0,09 & 0,06 & 0,11 \\ 0,34 & 0,28 & 0,38 & 0,30 & 0,28 \\ 0,17 & 0,19 & 0,22 & 0,23 & 0,16 \\ 0,31 & 0,39 & 0,25 & 0,36 & 0,41 \end{bmatrix} \quad F = \begin{bmatrix} \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \end{bmatrix}$$

$$A^6 \cdot F = \begin{bmatrix} 0,047 \\ 0,100 \\ 0,316 \\ 0,194 \\ 0,344 \end{bmatrix} \quad E = \begin{bmatrix} 0,049 \\ 0,098 \\ 0,317 \\ 0,195 \\ 0,341 \end{bmatrix}$$

PageRank

Sistema dinámico

$$A = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix} \quad A^* = \begin{bmatrix} & & & & \\ & & & & \\ & & \dots & & \\ & & & & \\ & & & & \end{bmatrix} \quad F = \begin{bmatrix} \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \end{bmatrix}$$

$$A^* \cdot F = \begin{bmatrix} & & & & \\ & & & & \\ & & \dots & & \\ & & & & \end{bmatrix} \quad E = \begin{bmatrix} 0,049 \\ 0,098 \\ 0,317 \\ 0,195 \\ 0,341 \end{bmatrix}$$

PageRank

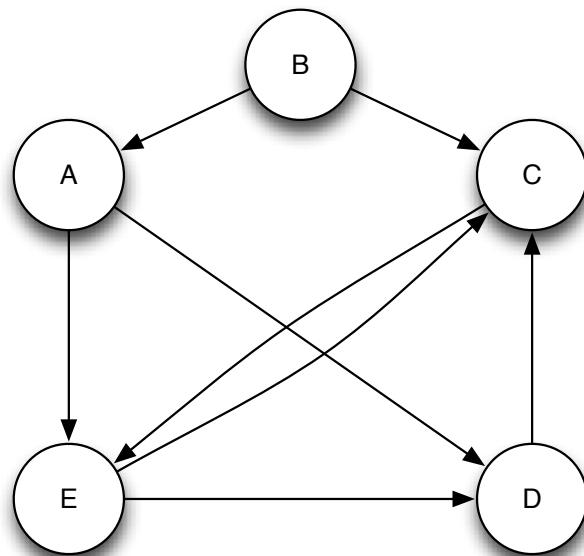
Sistema dinámico

$$A = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix} \quad A^{10} = \begin{bmatrix} 0,05 & 0,04 & 0,06 & 0,04 & 0,04 \\ 0,09 & 0,11 & 0,09 & 0,10 & 0,10 \\ 0,32 & 0,31 & 0,33 & 0,32 & 0,31 \\ 0,20 & 0,18 & 0,20 & 0,20 & 0,19 \\ 0,34 & 0,36 & 0,32 & 0,36 & 0,36 \end{bmatrix} \quad F = \begin{bmatrix} \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \\ \frac{1}{5} \end{bmatrix}$$

$$A^{10} \cdot F = \begin{bmatrix} 0,049 \\ 0,098 \\ 0,317 \\ 0,194 \\ 0,342 \end{bmatrix} \quad E = \begin{bmatrix} 0,049 \\ 0,098 \\ 0,317 \\ 0,195 \\ 0,341 \end{bmatrix}$$

PageRank

Corrección



Matriz de adyacencia

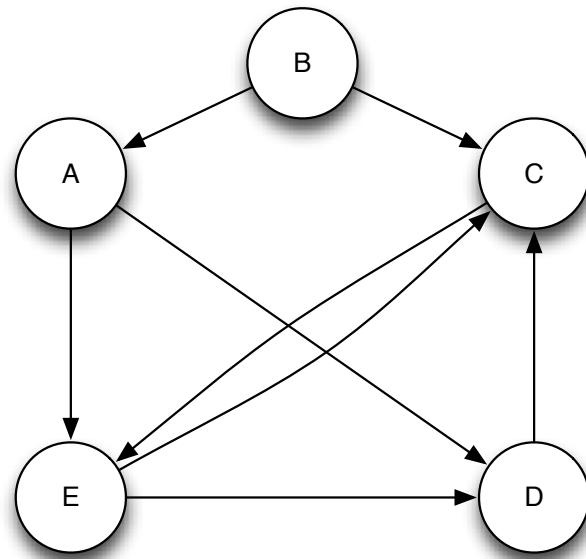
$$D = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Matriz de transición

$$A = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 1 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix}$$

PageRank

Corrección



PageRank

$$\begin{bmatrix} 0 \\ 0 \\ 0,4 \\ 0,2 \\ 0,4 \end{bmatrix}$$

corrección son prácticamente 0

PageRank

Corrección

$$\text{PageRank} = A^k \cdot F$$

$$\text{PageRank} = A \cdot \text{PageRank}$$

PageRank

Corrección

$$\text{PageRank} = M^k \cdot F$$

$$M = (1 - p) \cdot A + p \cdot B$$

$$p \in [0, 1]$$

$$B = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix}$$

factor de corrección, puedo acabar en una página de forma aleatoria saliendo del subgrafo en el que estoy trabajando.

PageRank

Corrección

$$\text{PageRank} = ((1 - p) \cdot A + p \cdot B) \cdot \text{PageRank}$$

$$p \in [0, 1]$$

$$\frac{1}{n} \cdot e_1 + \frac{1}{n} \cdot e_2 + \cdots + \frac{1}{n} \cdot e_i + \cdots + \frac{1}{n} \cdot e_n = \frac{1}{n}, \sum_{i=1}^n e_i = 1$$

$$B = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix}$$

PageRank

Corrección

$$PageRank = (1 - p) \cdot A \cdot PageRank + p \cdot b$$

$$p \in [0, 1]$$

$$b = \begin{bmatrix} \frac{1}{n} \\ \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{bmatrix}$$

PageRank

Corrección

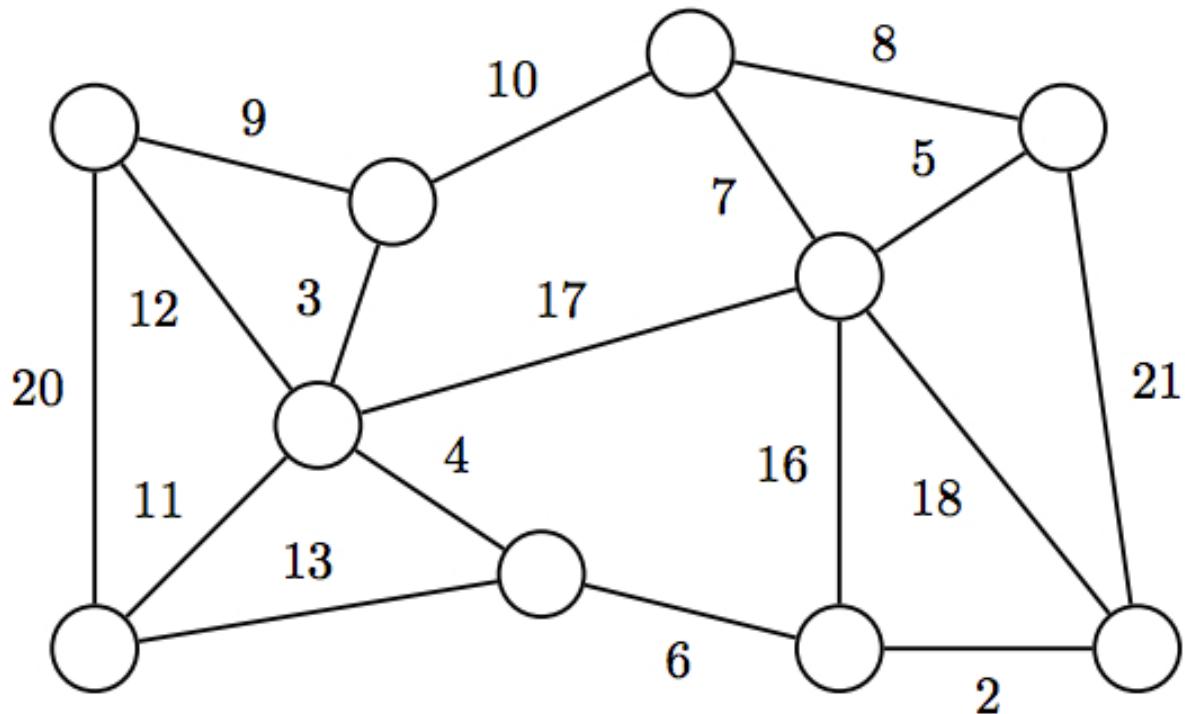
$$A = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 1 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$p = 0,15$$

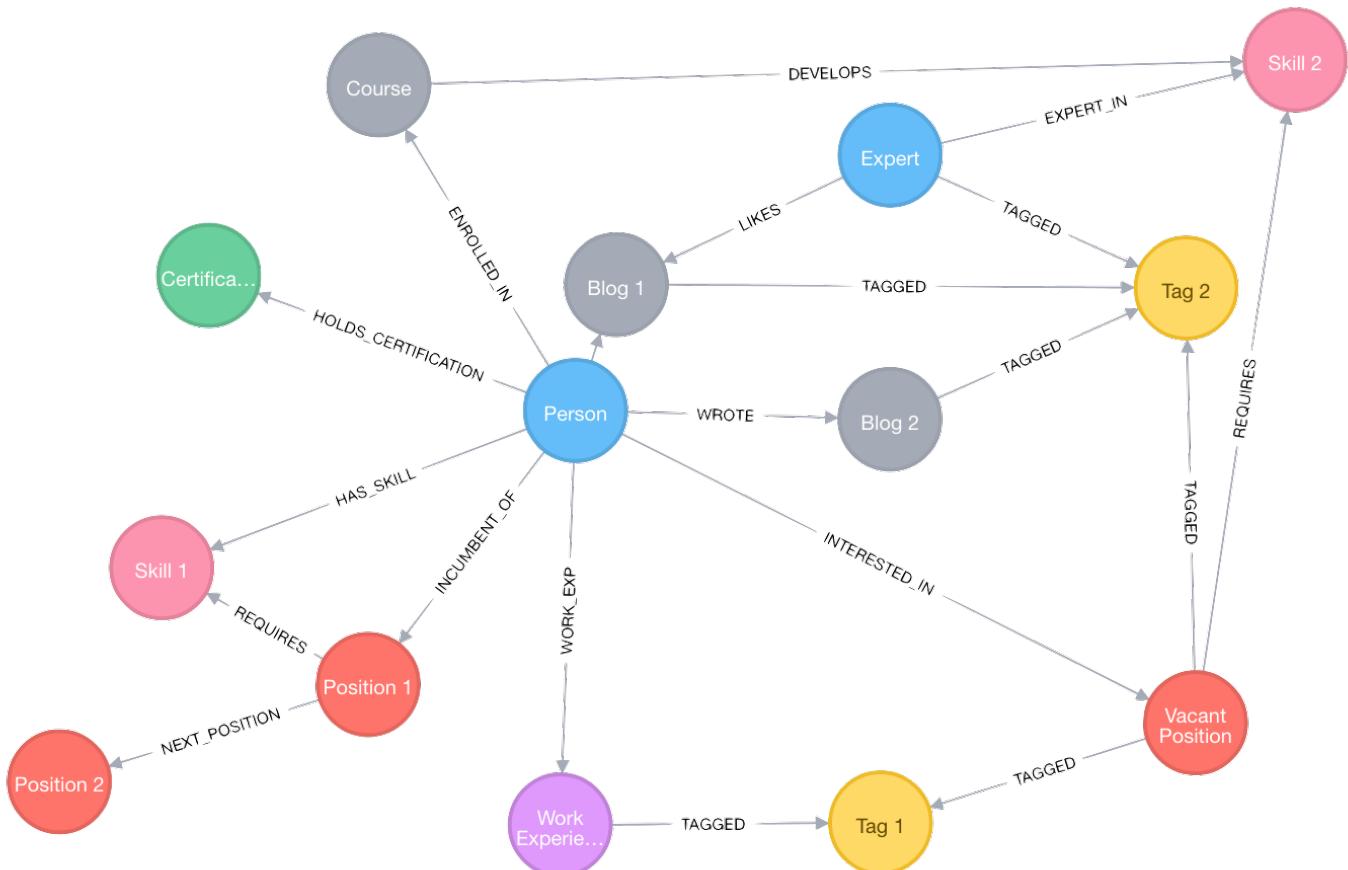
$$M = \begin{bmatrix} 0,030 & 0,455 & 0,03 & 0,03 & 0,030 \\ 0,030 & 0,030 & 0,03 & 0,03 & 0,030 \\ 0,030 & 0,455 & 0,03 & 0,88 & 0,455 \\ 0,455 & 0,030 & 0,03 & 0,03 & 0,455 \\ 0,455 & 0,030 & 0,88 & 0,03 & 0,030 \end{bmatrix}$$

$$E = \begin{bmatrix} 0,043 \\ 0,030 \\ 0,366 \\ 0,201 \\ 0,360 \end{bmatrix}$$

Grafos



Grafos



Grafos

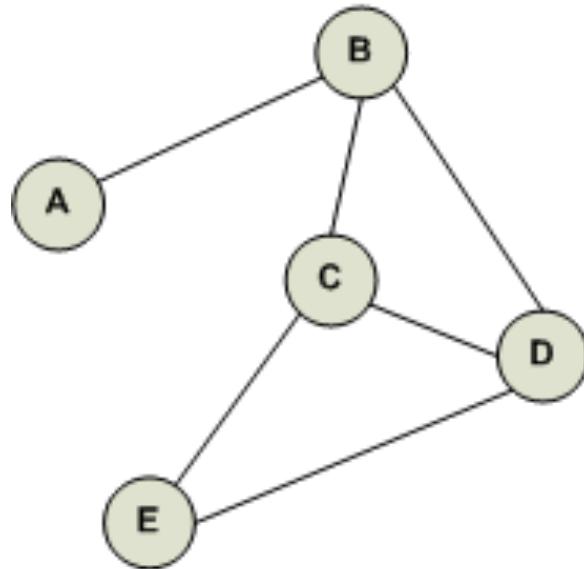


Fig 1. Undirected Graph

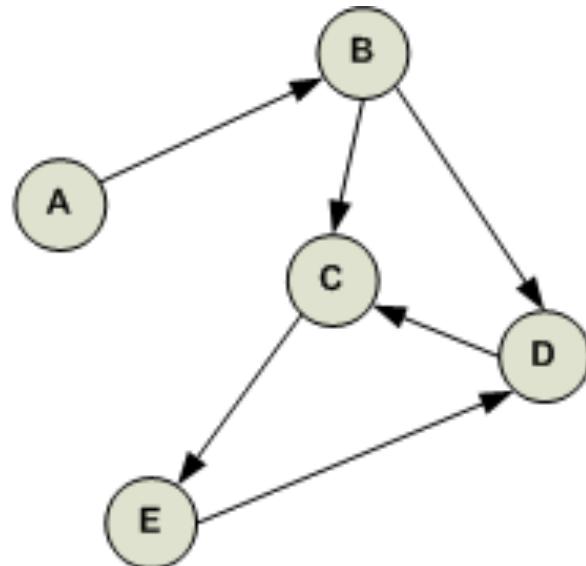
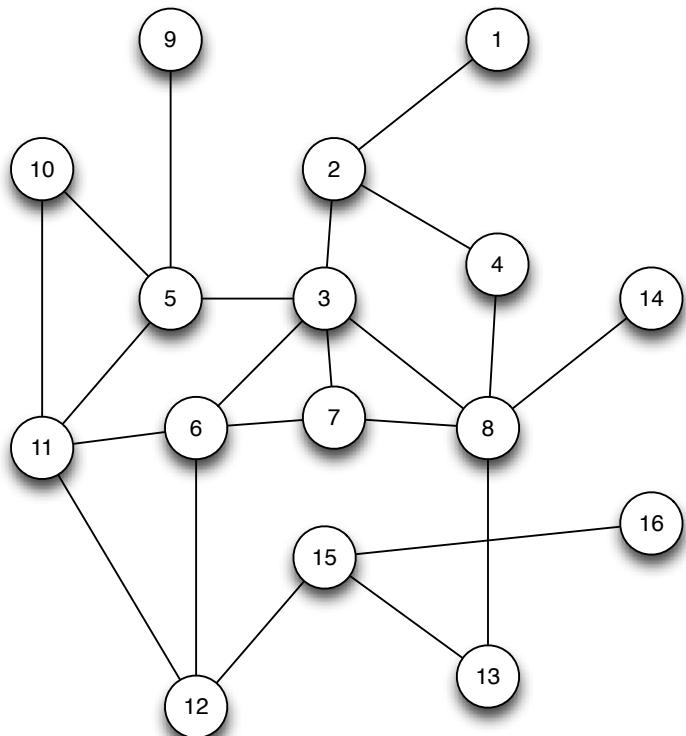


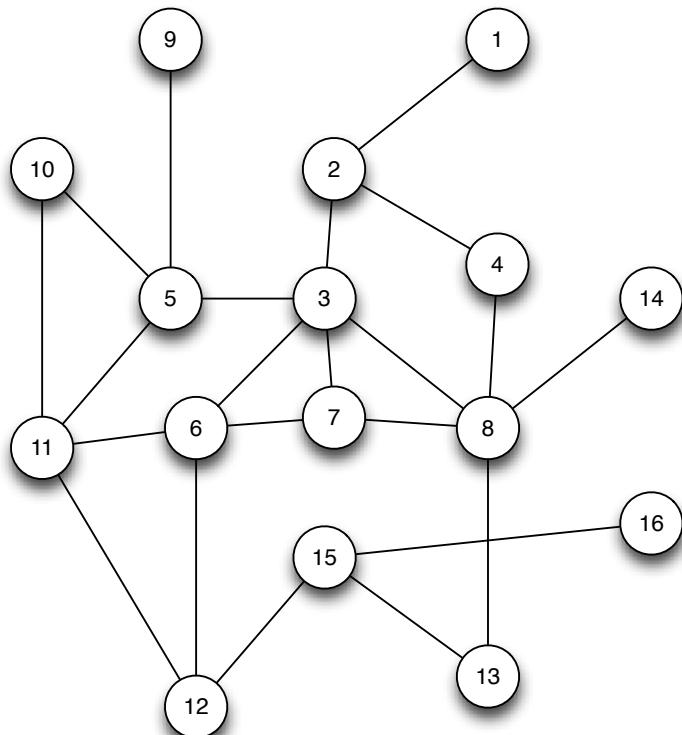
Fig 2. Directed Graph

Personalized PageRank (PPR)



$$\text{rank} = M' \cdot \text{rank}$$

Personalized PageRank (PPR)



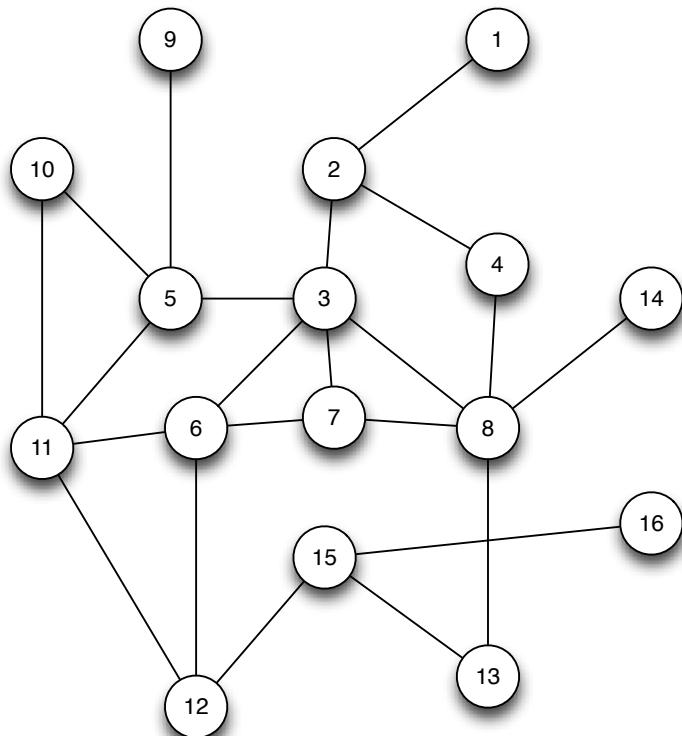
Nodo	Rank ⁰
1	100
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0
13	0
14	0
15	0
16	0

$$\text{rank} = (1 - p) \cdot M \cdot \text{rank} + p v$$

$$v_i = \frac{1}{|\mathbf{T}|} \mathbf{1}_{[i \in \mathbf{T}]}$$

{ match de mi búsqueda,
coincidencias de palabras
que busco con la web }

Personalized PageRank (PPR)

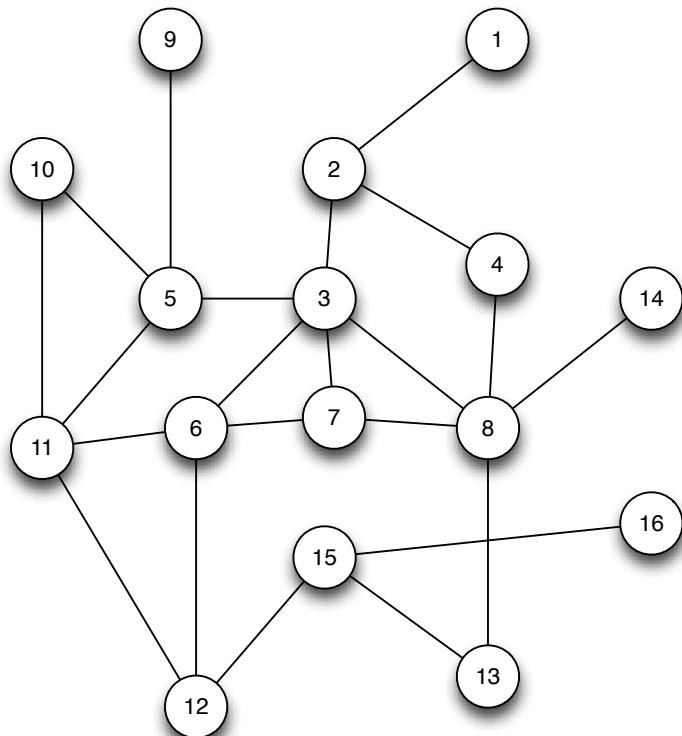


Nodo	Rank ¹
1	15
2	85
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0
13	0
14	0
15	0
16	0

$$rank = (1 - p) \cdot M \cdot rank + p v$$

$$v_i = \frac{1}{|\mathbf{T}|} \mathbf{1}_{[i \in \mathbf{T}]}$$

Personalized PageRank (PPR)

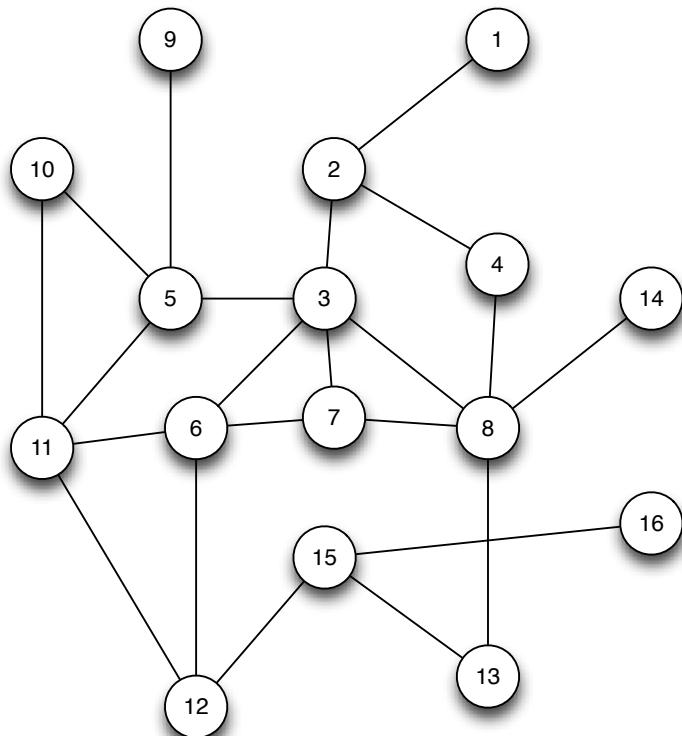


Nodo	Rank ²
1	39
2	13
3	24
4	24
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0
13	0
14	0
15	0
16	0

$$rank = (1 - p) \cdot M \cdot rank + p v$$

$$v_i = \frac{1}{|\mathbf{T}|} \mathbf{1}_{[i \in \mathbf{T}]}$$

Personalized PageRank (PPR)

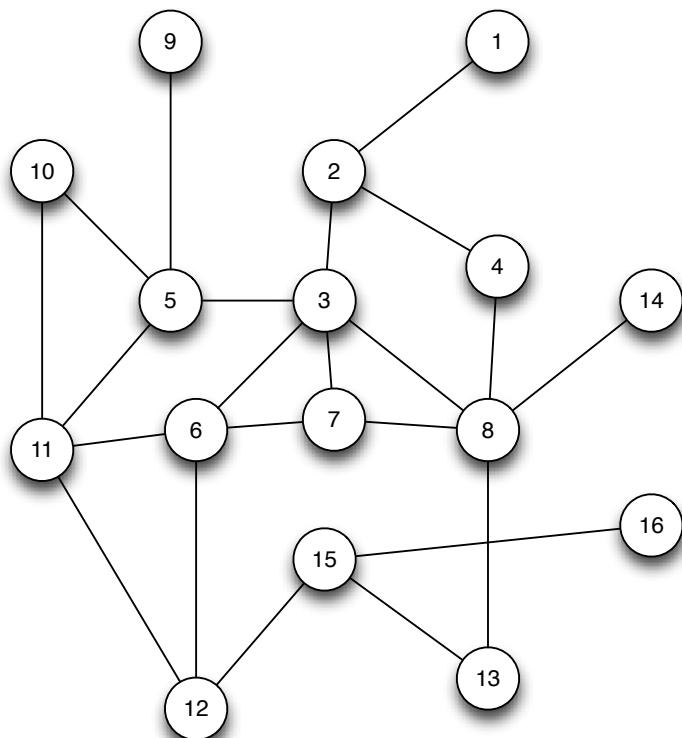


Nodo	Rank ³
1	19
2	48
3	4
4	4
5	4
6	4
7	4
8	14
9	0
10	0
11	0
12	0
13	0
14	0
15	0
16	0

$$rank = (1 - p) \cdot M \cdot rank + p v$$

$$v_i = \frac{1}{|\mathbf{T}|} \mathbf{1}_{[i \in \mathbf{T}]}$$

Personalized PageRank (PPR)

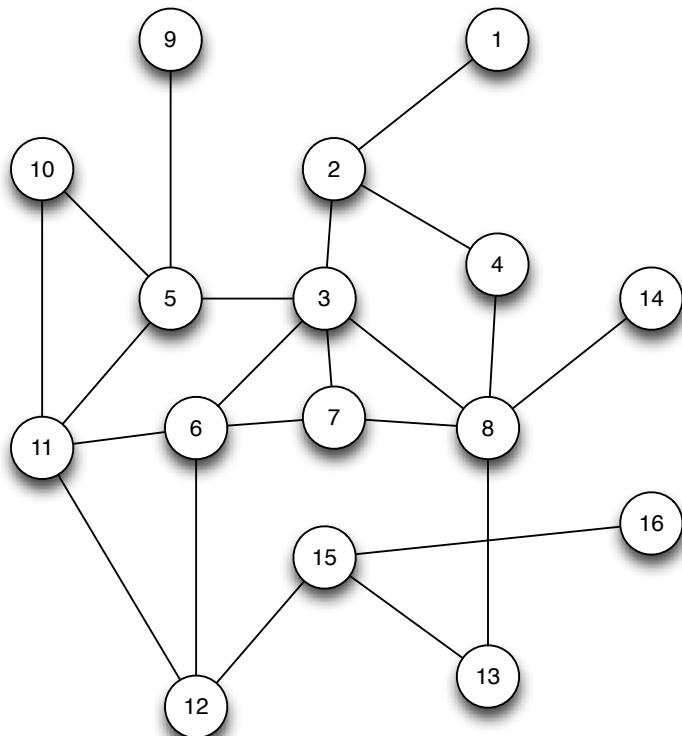


Nodo	Rank ⁴
1	28
2	18
3	19
4	16
5	1
6	2
7	4
8	3
9	1
10	1
11	2
12	1
13	2
14	2
15	0
16	0

$$rank = (1 - p) \cdot M \cdot rank + p v$$

$$v_i = \frac{1}{|\mathbf{T}|} \mathbf{1}_{[i \in \mathbf{T}]}$$

Personalized PageRank (PPR)

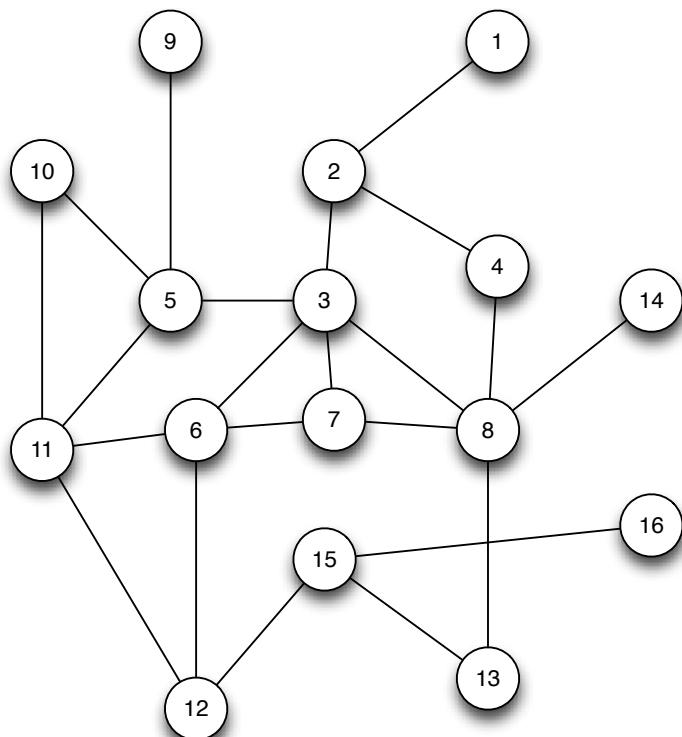


Nodo	Rank ⁵
1	20
2	34
3	7
4	6
5	5
6	5
7	4
8	14
9	0
10	1
11	1
12	1
13	1
14	1
15	1
16	0

$$rank = (1 - p) \cdot M \cdot rank + p v$$

$$v_i = \frac{1}{|\mathbf{T}|} \mathbf{1}_{[i \in \mathbf{T}]}$$

Personalized PageRank (PPR)

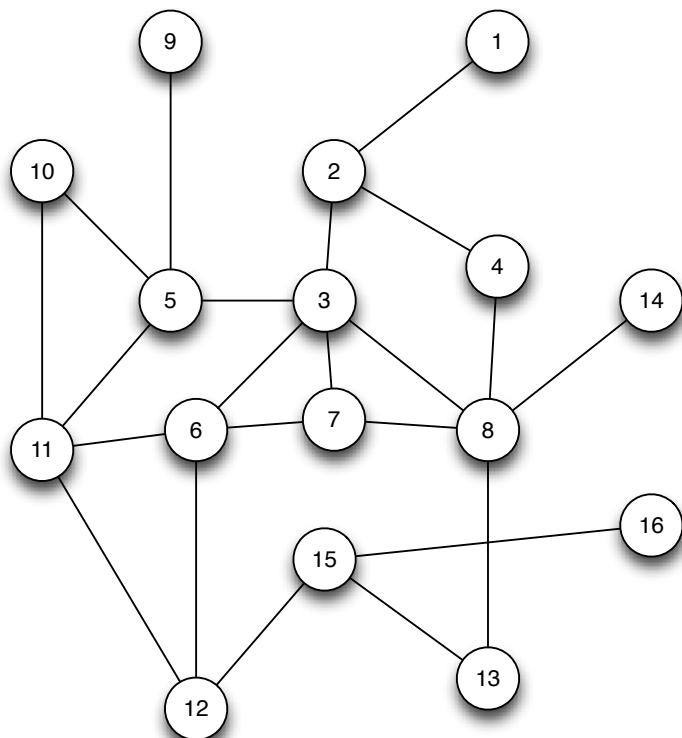


Nodo	Rank ⁶
1	25
2	21
3	15
4	12
5	2
6	3
7	5
8	6
9	1
10	1
11	2
12	2
13	3
14	2
15	0
16	0

$$rank = (1 - p) \cdot M \cdot rank + p v$$

$$v_i = \frac{1}{|\mathbf{T}|} \mathbf{1}_{[i \in \mathbf{T}]}$$

Personalized PageRank (PPR)

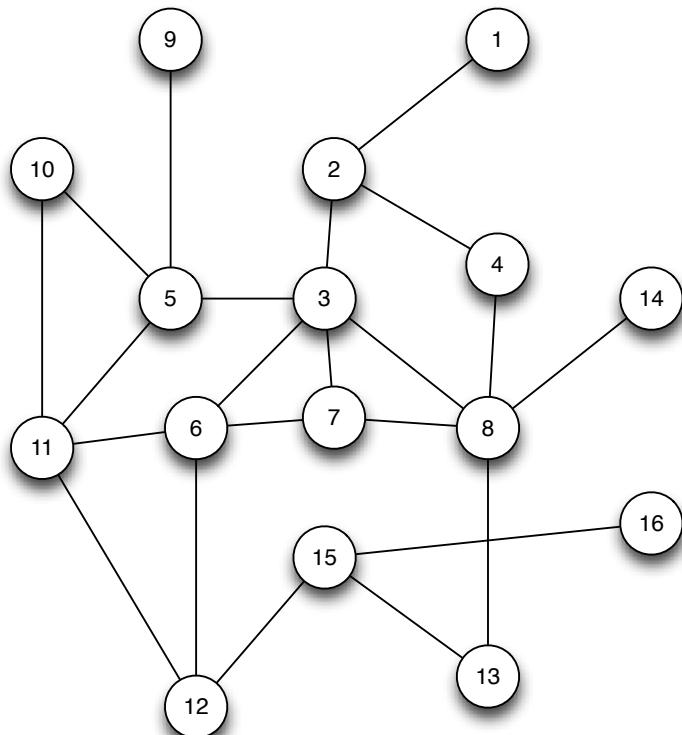


Nodo	Rank ⁷
1	21
2	29
3	9
4	7
5	4
6	5
7	4
8	12
9	0
10	1
11	2
12	1
13	1
14	1
15	2
16	0

$$rank = (1 - p) \cdot M \cdot rank + p v$$

$$v_i = \frac{1}{|\mathbf{T}|} \mathbf{1}_{[i \in \mathbf{T}]}$$

Personalized PageRank (PPR)

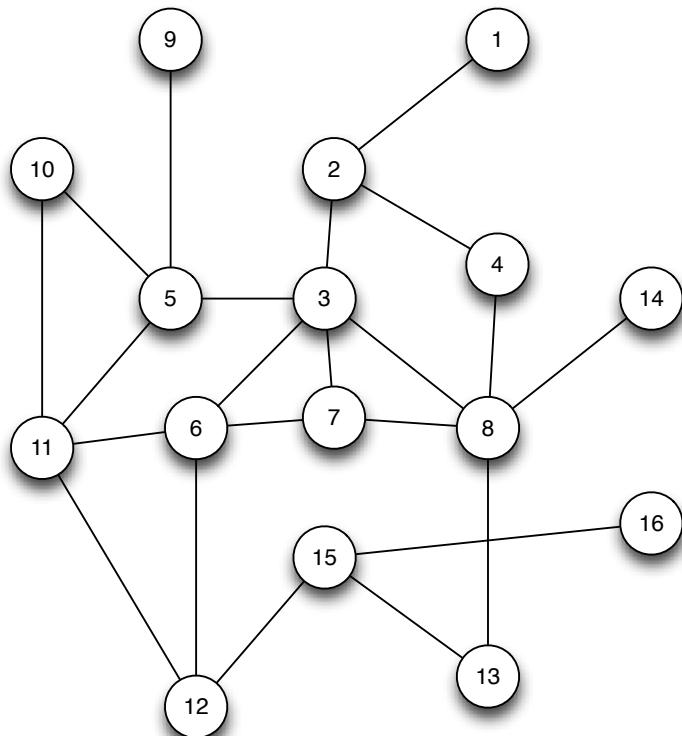


Nodo	Rank ⁸
1	23
2	22
3	13
4	10
5	3
6	4
7	5
8	7
9	1
10	1
11	3
12	2
13	3
14	2
15	1
16	1

$$rank = (1 - p) \cdot M \cdot rank + p v$$

$$v_i = \frac{1}{|\mathbf{T}|} \mathbf{1}_{[i \in \mathbf{T}]}$$

Personalized PageRank (PPR)

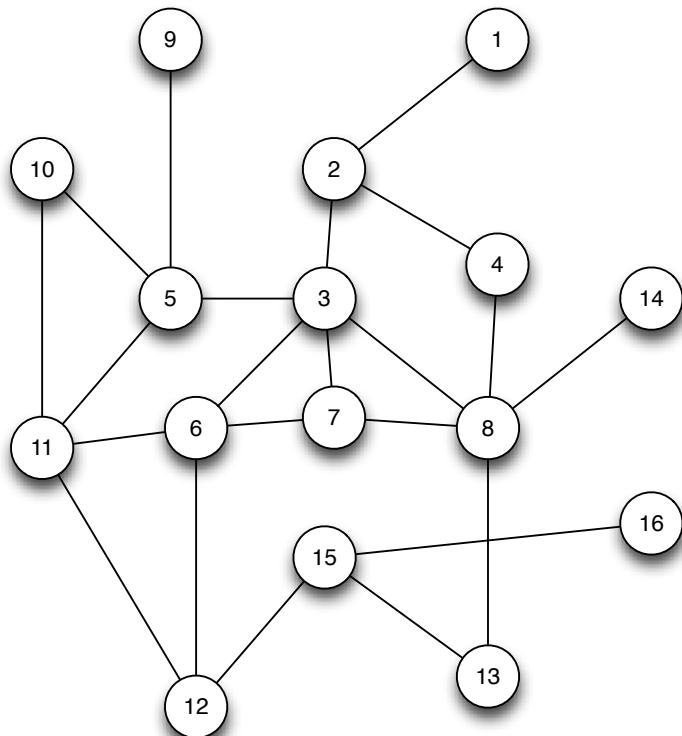


Nodo	Rank ⁹
1	21
2	26
3	10
4	7
5	4
6	5
7	4
8	11
9	1
10	1
11	2
12	2
13	1
14	1
15	2
16	0

$$rank = (1 - p) \cdot M \cdot rank + p v$$

$$v_i = \frac{1}{|\mathbf{T}|} \mathbf{1}_{[i \in \mathbf{T}]}$$

Personalized PageRank (PPR)



Nodo	Rank ¹⁰
1	22
2	23
3	12
4	9
5	3
6	4
7	5
8	8
9	1
10	1
11	3
12	2
13	2
14	2
15	1
16	1

$$rank = (1 - p) \cdot M \cdot rank + p v$$

$$v_i = \frac{1}{|\mathbf{T}|} \mathbf{1}_{[i \in \mathbf{T}]}$$

Búsqueda sesgada por el tema

HITS

Hubs and authorities

Authorities: Webs con el mejor contenido relevante

Hubs: Webs que apuntan eficientemente al contenido relevante

Búsqueda sesgada por el tema

HITS

Procedimiento

Dada una consulta Q :

- ▶ Coger el subgrafo R_Q formado por las N ($= 200$) webs que más referencias a Q contengan
- ▶ Obtener S_Q , el subgrafo que amplia R_Q con webs apuntadas o que apuntan a las webs de R_Q
- ▶ Calcular un peso a_i y h_i para cada web (w_i) de S_Q :

$$a_i = \sum_{j \in IN(w_i)} h_j$$

$$h_i = \sum_{j \in OUT(w_i)} a_j$$

donde $IN(w_i)$ agrupa todas las webs que apuntan a w_i y $OUT(w_i)$ agrupa todas las webs apuntadas por w_i .

Búsqueda sesgada por el tema

HITS

Procedimiento

Dada una consulta Q :

- ▶ Inicializar el vector de pesos \mathbf{h} todo a 1.
- ▶ Actualizar los valores iterativamente:

$$\mathbf{h} = \mathbf{D} \cdot \mathbf{a}$$

$$\mathbf{a} = \mathbf{D}^t \cdot \mathbf{h}$$

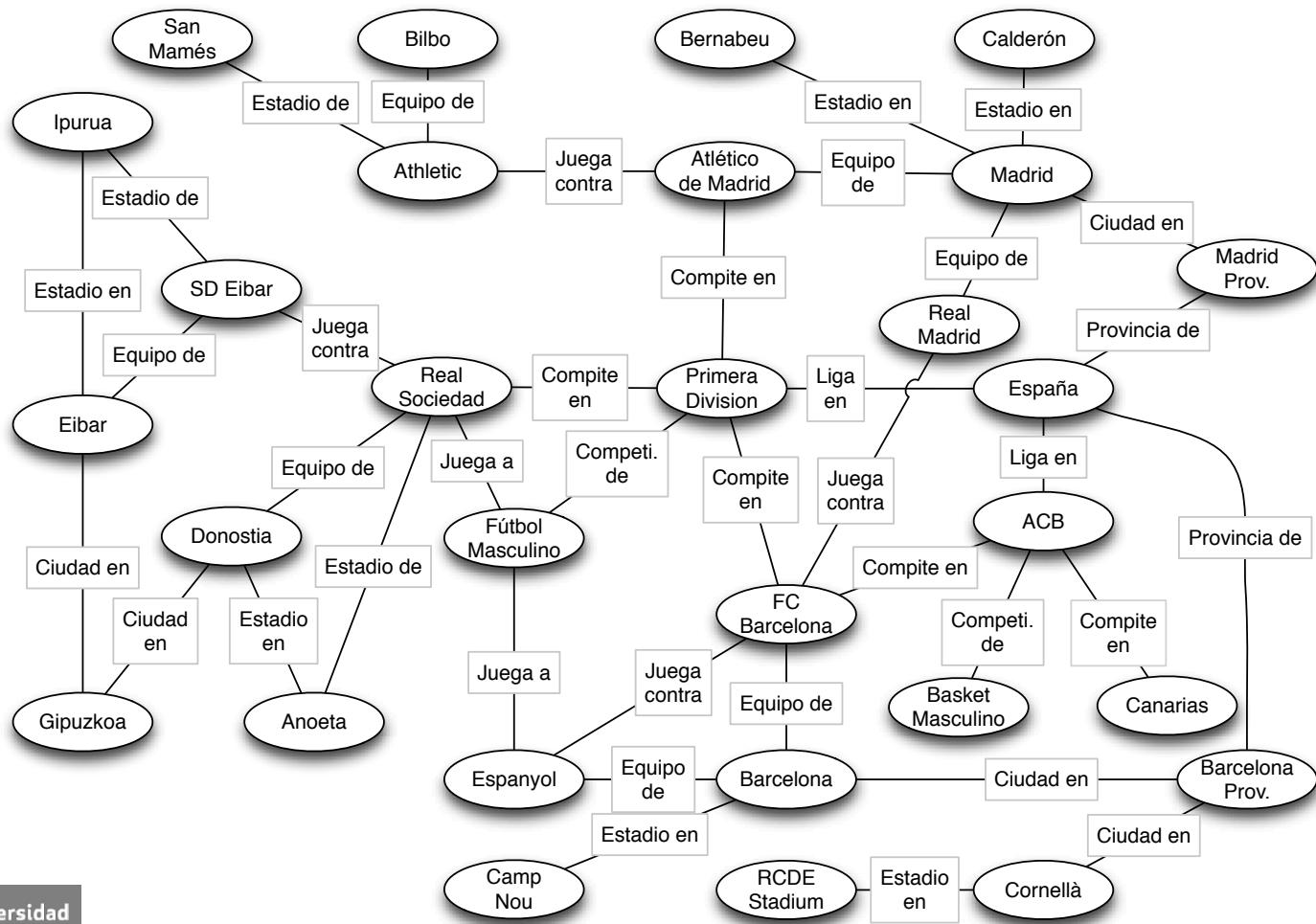
Aunque también se pueden ver como vectores autocomputados:

$$\mathbf{h}^k = (\mathbf{D} \cdot \mathbf{D}^t) \cdot \mathbf{h}^{k-1}$$

$$\mathbf{a}^k = (\mathbf{D}^t \cdot \mathbf{D}) \cdot \mathbf{a}^{k-1}$$

Inferir nuevas relaciones de *cierto* tipo

PRA



Inferir nuevas relaciones de *cierto* tipo

PRA

Algoritmo en tres fases

- ▶ Buscar caminos (y seleccionar)
- ▶ Rellenar matriz con probabilidades
- ▶ Aprender clasificador

Consideraciones

- ▶ Datos iniciales: Pares reales del tipo de relación en cuestión
- ▶ ¿Pares negativos?
- ▶ Random walk with restart (RWR ó PPR)
- ▶ Con la matriz rellena, se puede aprender cualquier clasificador

Inferir nuevas relaciones de *cierto* tipo

PRA

Primera fase

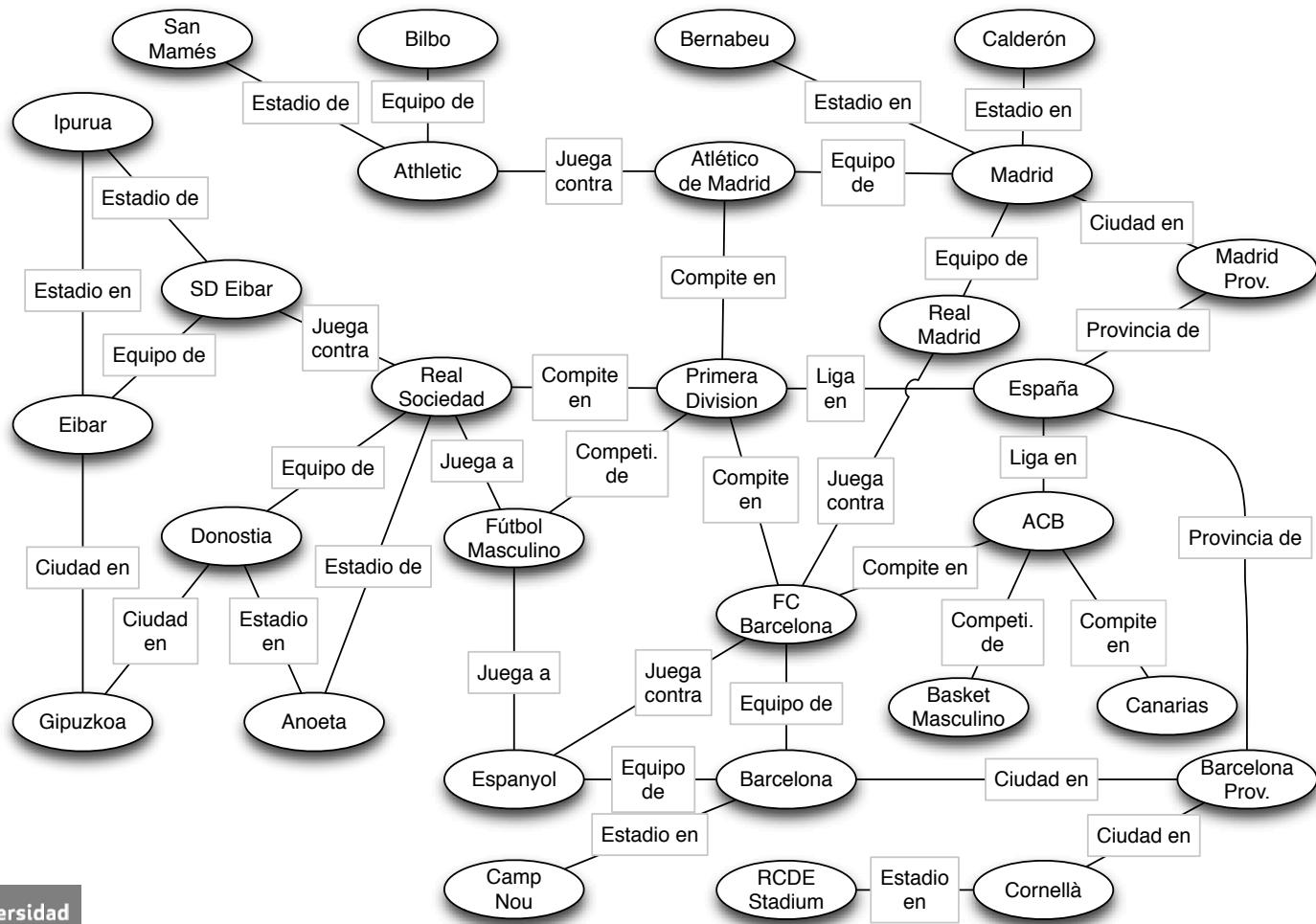
- ▶ Lanzar múltiples random walks desde un componente de uno de los pares de nodos de aprendizaje
- ▶ Guardar todos los caminos que haya llevado (en algún momento del proceso) a pasar por el otro nodo del par

Ejemplo: Relación “juega_a”, con pares reales tales como (Real Sociedad, Fútbol masculino) ó (Espanyol, Fútbol masculino)
Algunos caminos encontrados:

compite_en + competición_de
compite_en + \neg compite_en + compete_en + competición_de
 juega_contra + compete_en + competición_de
 equipo_de + \neg equipo_de + compete_en + competición_de

Inferir nuevas relaciones de *cierto* tipo

PRA



Inferir nuevas relaciones de *cierto* tipo

PRA

Segunda fase

- ▶ Para cada camino detectado –columna–, llenar cada celda con la probabilidad de llegar al *target* (t_i) desde el *origin* (o_i), dado un par – $r_i = (o_i, t_i)$, fila– usando el camino en cuestión
- ▶ Generar los negativos y completar igualmente el dataset

	camino_1	camino_2	...	camino_m	C
(r_1)	$p_{c_1}(o_1, t_1)$	$p_{c_2}(o_1, t_1)$...	$p_{c_m}(o_1, t_1)$	+
(r_2)	$p_{c_1}(o_2, t_2)$	$p_{c_2}(o_2, t_2)$...	$p_{c_m}(o_2, t_2)$	+
...
(r_{n+1})	$p_{c_1}(o_{n+1}, t_{n+1})$	$p_{c_2}(o_{n+1}, t_{n+1})$...	$p_{c_m}(o_{n+1}, t_{n+1})$	-
(r_{n+2})	$p_{c_1}(o_{n+2}, t_{n+2})$	$p_{c_2}(o_{n+2}, t_{n+2})$...	$p_{c_m}(o_{n+2}, t_{n+2})$	-
...

Inferir nuevas relaciones de *cierto* tipo

PRA

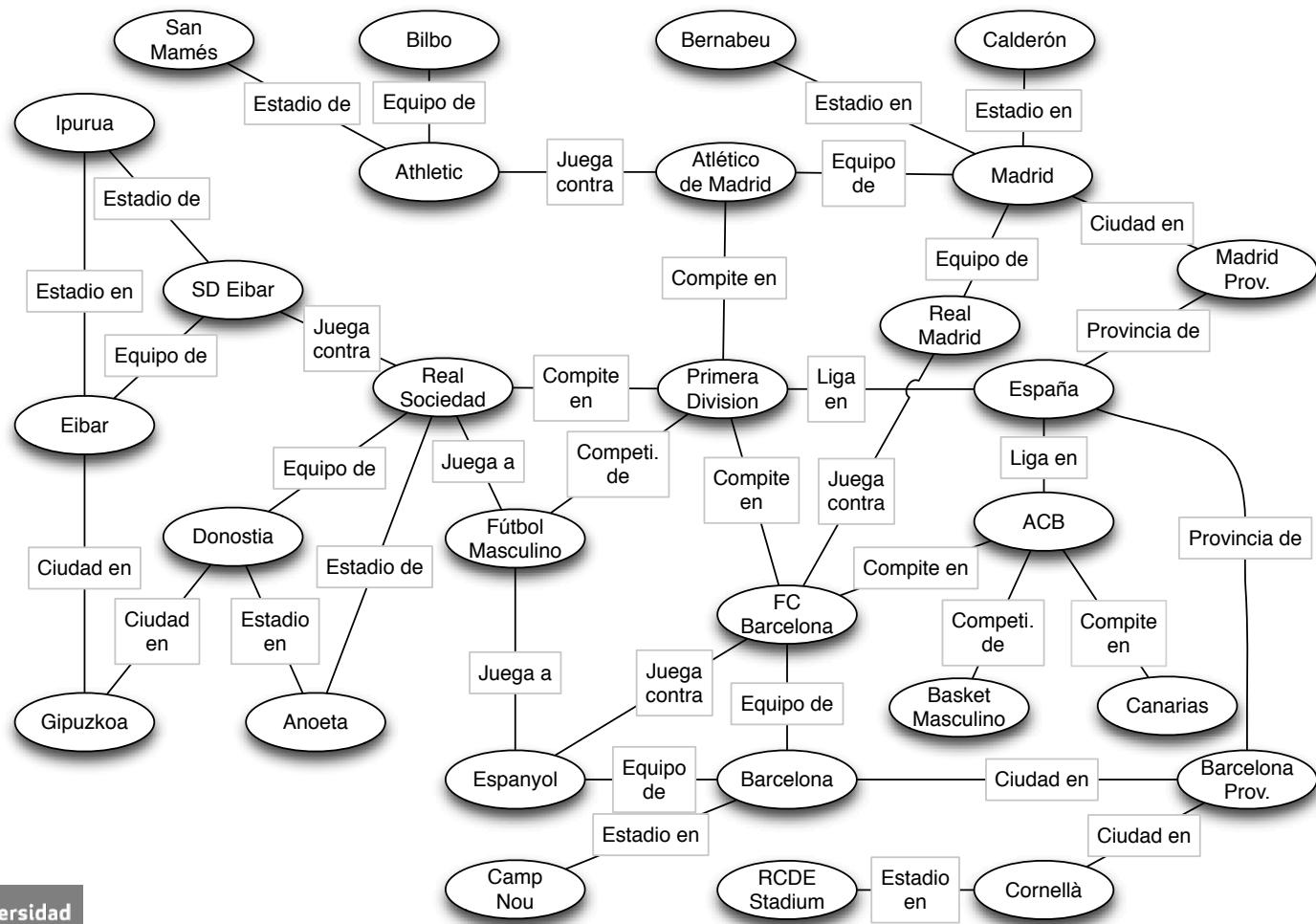
Tercera fase

- ▶ Aprender un clasificador (regresión logística)
- ▶ Usar para predecir (inferir) nuevos pares de la relación

	camino_1	camino_2	...	camino_m	C
(r_1)	$p_{c_1}(o_1, t_1)$	$p_{c_2}(o_1, t_1)$...	$p_{c_m}(o_1, t_1)$	+
(r_2)	$p_{c_1}(o_2, t_2)$	$p_{c_2}(o_2, t_2)$...	$p_{c_m}(o_2, t_2)$	+
...
(r_{n+1})	$p_{c_1}(o_{n+1}, t_{n+1})$	$p_{c_2}(o_{n+1}, t_{n+1})$...	$p_{c_m}(o_{n+1}, t_{n+1})$	-
(r_{n+2})	$p_{c_1}(o_{n+2}, t_{n+2})$	$p_{c_2}(o_{n+2}, t_{n+2})$...	$p_{c_m}(o_{n+2}, t_{n+2})$	-
...

Inferir nuevas relaciones de *cierto* tipo

PRA



Bibliografía

- R. Agrawal, R. Srikant. Fast algorithms for mining association rules. Proc. 20th International Conference on Very Large Data Bases (VLDB), 487-499, 1994.
- L. Page, S. Brin, R. Motwani, T. Winograd. The pagerank citation ranking: Bringing order to the web. Tech. Report 1999-66, Stanford InfoLab, 1999.
- T. H. Haveliwala. Topic-sensitive pagerank. Proc. 11th International World Wide Web Conference, 517–526, 2002.
- J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal for the Association of Computing Machinery, 46(19):604-632, 1999.
- N. Lao, W. W. Cohen. Relational retrieval using a combination of path-constrained random walks. Machine Learning, 81(1):53–67,

[Examen] Para generación de imágenes se usan los GAN y los VAEs. AE más para reducción de la dimensionalidad

Aprendizaje no supervisado

VC10: Análisis de transacciones y reglas de asociación

Rocío del Amor del Amor
mrocio.delamor@campusviu.es

Universidad Internacional de Valencia

Reglas de asociación



Reglas de asociación

leche → galletas

leche ∧ cereales → galletas

Métricas

Soporte (support) de un (sub)conjunto de ítems $X \subseteq I$

Proporción de transacciones del conjunto S en las que aparece el subconjunto de ítems X :

$$\text{soporte}(X; S) = \frac{|\{T \in S : X \subseteq T\}|}{|S|}$$

Se puede entender como la probabilidad conjunta (marginal) de X .

El **soporte de una regla** de asociación $X \rightarrow Y$: soporte del conjunto resultante de la unión de antecedente y consecuente,

$$\text{soporte}(X \rightarrow Y; S) = \text{soporte}(\underbrace{X \cup Y}_{\substack{\text{proba. de} \\ \text{que compre}}} ; S)$$

de todas las transacciones

$X \cup Y$

Confianza (confidence) de una regla de asociación $X \rightarrow Y$

Frecuencia con la que se cumple la regla de asociación:

$$\text{confianza}(X \rightarrow Y; S) = \frac{\text{soporte}(X \wedge Y; S)}{\text{soporte}(X; S)}$$

La confianza tiende a 1 a medida que se observa con mayor frecuencia Y cada vez que aparece X .

Se puede entender como la precisión de la regla o la probabilidad condicionada de Y dado X .

Métricas

Mejora (lift) de una regla de asociación $X \rightarrow Y$

Ratio del soporte como indicador de la creencia de que la regla pueda ser producto del azar:

$$\text{mejora}(X \rightarrow Y; S) = \frac{\text{soporte}(X \cap Y; S)}{\text{soporte}(X; S) \cdot \text{soporte}(Y; S)}$$

donde según el valor de $\text{mejora}(X \rightarrow Y; S) =$

- 1 indica que ambos subconjuntos de ítems se relacionan al azar
- > 1 indica cierto grado de co-ocurrencia
- < 1 indica complementariedad (cuando se observa un subconjunto, no se observa el otro)

Búsqueda de subconjuntos frecuentes

Algoritmo APRIORI

Método de búsqueda en anchura de subconjuntos de elementos frecuentes

[Examen]

Intuición: un conjunto sólo puede ser frecuente si todos sus subconjuntos también lo son

Se reduce el espacio de búsqueda y se aborda el problema de manera iterativa-aglomerativa

Definiciones:

- **Itemset**: conjunto de elementos
- **Itemset frecuente**: conjunto de elementos que aparece en al menos ϵ transacciones del conjunto de referencia S
- **k -itemset**: conjunto de k elementos

Búsqueda de subconjuntos frecuentes

Algoritmo APRIORI

Dado un umbral de soporte ϵ , se seleccionan los 1-itemsets (items individuales) que superan el ϵ

Se buscan iterativamente los k -itemsets con $k = \{2, 3, \dots\}$

- Para que un k -itemset I sea frecuente, los k diferentes $(k - 1)$ -itemsets subconjuntos de I deben ser frecuentes

El algoritmo Apriori devuelve un conjuntos de itemsets frecuentes. A partir de estos, se pueden construir reglas de asociación.

Algoritmo APRIORI

Algoritmo Apriori

Recibe: Conjunto de transacciones, $S = \{T_1, T_2, \dots, T_n\}$; Umbral de soporte, ϵ

1. $L_1 \leftarrow$ Todos los 1-itemsets (dado ϵ)

2. Para $k = 2, 3, \dots$

 2.1. Se crea un conjunto de k -itemsets candidatos a partir de los $(k - 1)$ -itemsets fuertes obtenidos en el paso anterior:

$$C_k \leftarrow \{T = T^{k-1} \cup \{i\}: (T^{k-1} \in L_{k-1}) \wedge (i \notin T^{k-1}) \wedge (\forall j \in T, (T \setminus \{j\}) \in L_{k-1})\}$$

 2.2. Contar las apariciones de los k -itemsets candidatos de C_k en el conjunto de transacciones S

 2.3. Filtrar los k -itemsets de C_k que son realmente fuertes o frecuentes:

$$L_k = \{T \in C_k: \text{soporte}(T) > \epsilon\}$$

Parar si $L_k = \emptyset$

Devuelve: k -itemsets frecuentes, para todo k

Problemas

- ▶ El listado de todos los posibles candidatos en cada paso, es un procedimiento exhaustivo (coste computacional).

Es habitual buscar los itemsets candidatos solamente como una combinación de los itemsets frecuentes de la iteración anterior ($k - 1$)

- ▶ Recorre múltiples veces el conjunto de transacciones de referencia S

Se puede ir reduciendo el conjunto de transacciones S (no se encontrarán itemsets de tamaño k en una transacción que no de tamaño $k - 1$).

Aprendizaje no supervisado

Deep clustering

Rocío del Amor del Amor
mrocio.delamor@campusviu.es

Universidad Internacional de Valencia

Agrupamiento

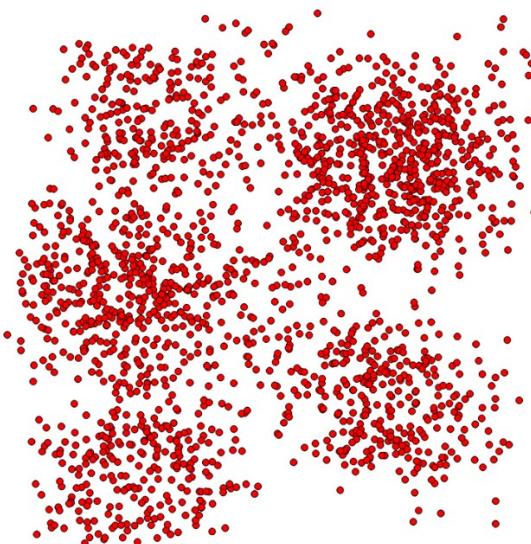
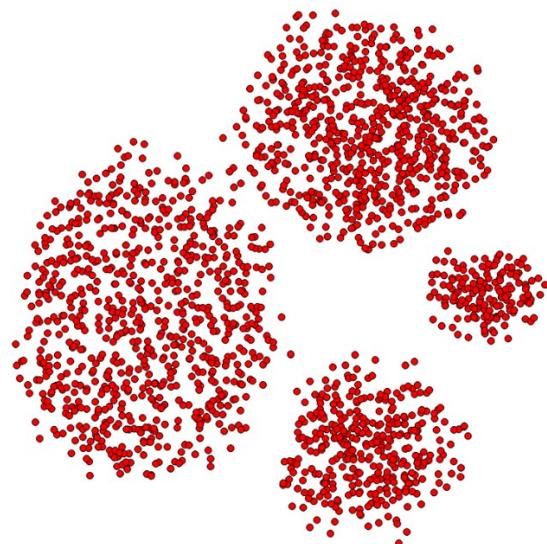
Definición

Dado un conjunto de datos, el agrupamiento trata de identificar subgrupos homogéneos de ejemplos que manifiestan diferencias relevantes con los otros subgrupos que se formen.

Agrupamiento

Definición

Dado un conjunto de datos, el agrupamiento trata de identificar subgrupos homogéneos de ejemplos que manifiestan diferencias relevantes con los otros subgrupos que se formen.

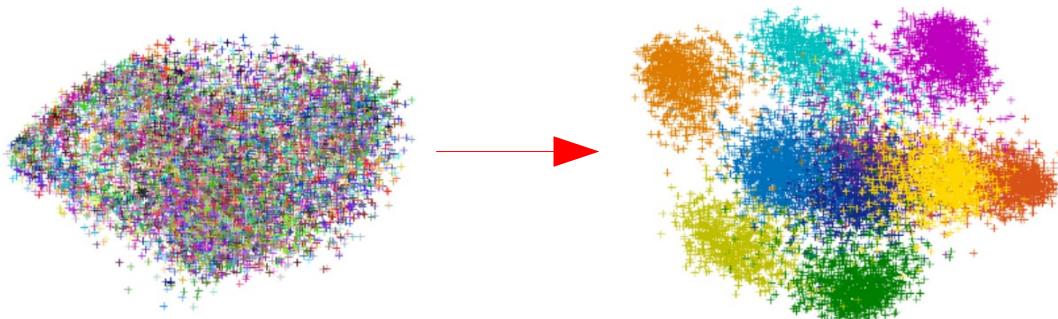


Agrupamiento profundo

Definición

Dado un conjunto de datos, el agrupamiento trata de identificar subgrupos homogéneos de ejemplos que manifiestan diferencias relevantes con los otros subgrupos que se formen.

Para ello, primero realiza una transformación de los datos a un espacio de representación diferente, casi siempre de **menor dimensionalidad**.

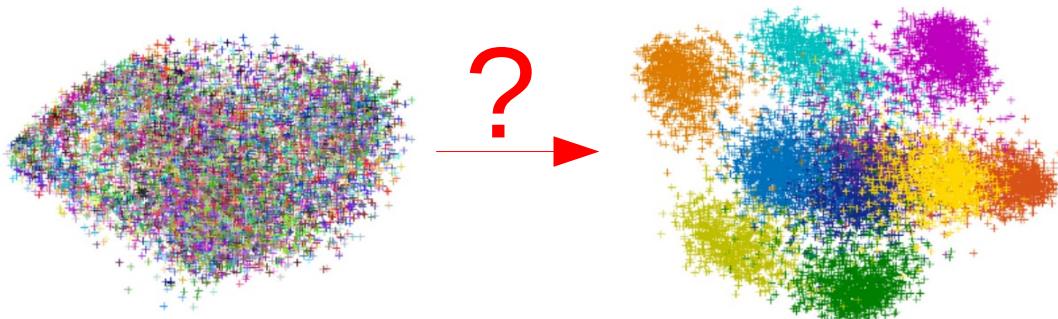


Agrupamiento profundo

Definición

Dado un conjunto de datos, el agrupamiento trata de identificar subgrupos homogéneos de ejemplos que manifiestan diferencias relevantes con los otros subgrupos que se formen.

Para ello, primero realiza una transformación de los datos a un espacio de representación diferente, casi siempre de **menor dimensionalidad**.



Agrupamiento profundo

Definición

Dado un conjunto de datos, el agrupamiento trata de identificar subgrupos homogéneos de ejemplos que manifiestan diferencias relevantes con los otros subgrupos que se formen.

Para ello, primero realiza una transformación de los datos a un espacio de representación diferente, casi siempre de **menor dimensionalidad**.

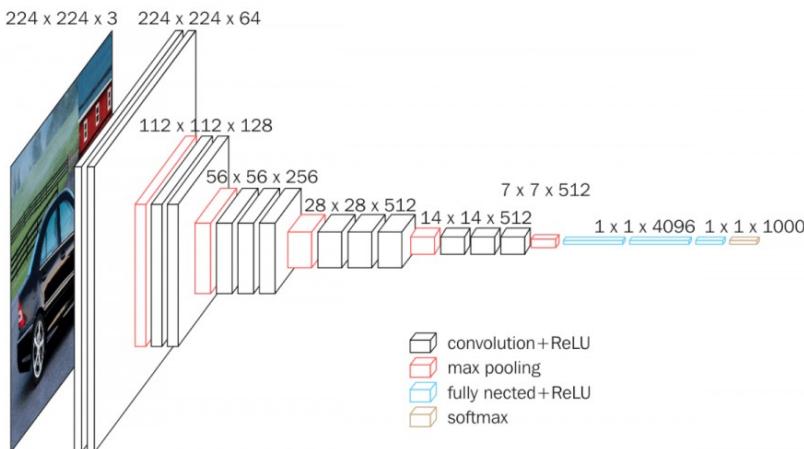
El primer paso es:
reducir la dimensionalidad de los datos

Agrupamiento profundo

¿Cómo encontramos las transformaciones adecuadas?

Definición

Las redes neuronales **profundas** permiten transformar los datos de un espacio de representación a otro de menor dimensionalidad y mayor densidad de información, el cual utilizan para llevar a cabo la tarea que se les exige.



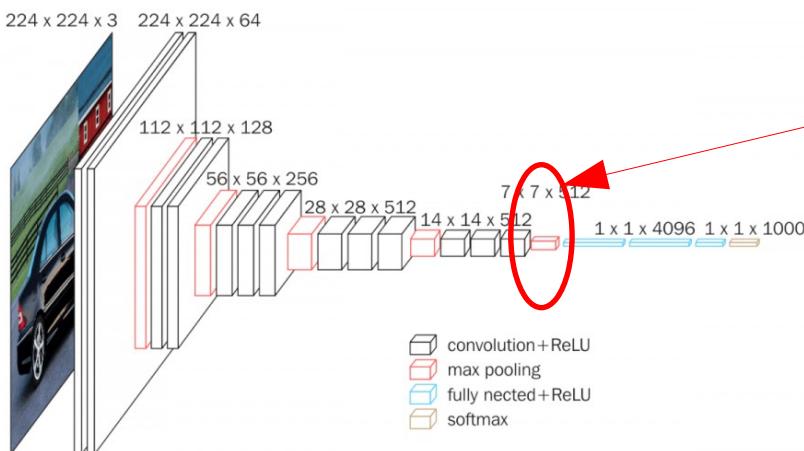
Red Neuronal Convolucional (CNN) - VGG16

Agrupamiento profundo

¿Cómo encontramos las transformaciones adecuadas?

Definición

Las redes neuronales **profundas** permiten transformar los datos de un espacio de representación a otro de menor dimensionalidad y mayor densidad de información, el cual utilizan para llevar a cabo la tarea que se les exige.



En este punto la red ha comprimido la información de la entrada con la menor pérdida de información posible

Red Neuronal Convolucional (CNN) - VGG16

Agrupamiento profundo

Tipos de algoritmos de agrupamiento profundo

- Basados en auto-encoders
- Basados en modelos generativos
- Basados en optimización directa de cluster

Agrupamiento profundo: AEs

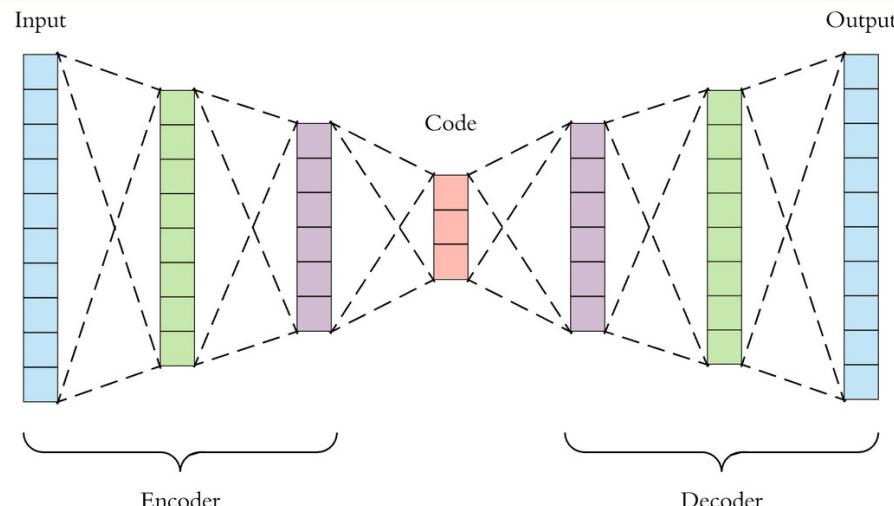
Tipos de algoritmos de agrupamiento profundo

- **Basados en auto-encoders**
- Basados en modelos generativos
- Basados en optimización directa de cluster

Agrupamiento profundo: AEs

Definición

El propósito de los **auto-encoders** es reconstruir los datos que tienen a la entrada con el menor error posible tras haberlos **comprimido** a un espacio de menor dimensionalidad (espacio latente o *bottleneck*).

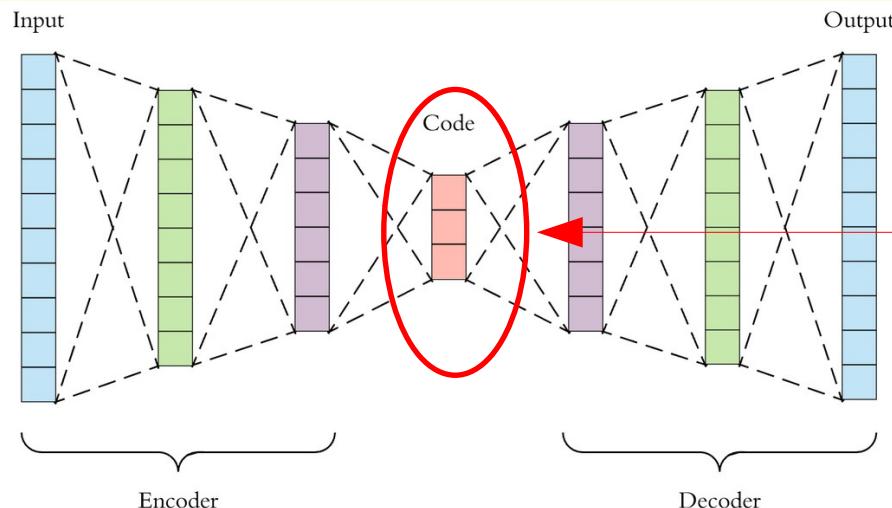


Arquitectura típica de un auto-encoder

Agrupamiento profundo: AEs

Definición

El propósito de los **auto-encoders** es reconstruir los datos que tienen a la entrada con el menor error posible tras haberlos **comprimido** a un espacio de menor dimensionalidad (espacio latente o *bottleneck*).

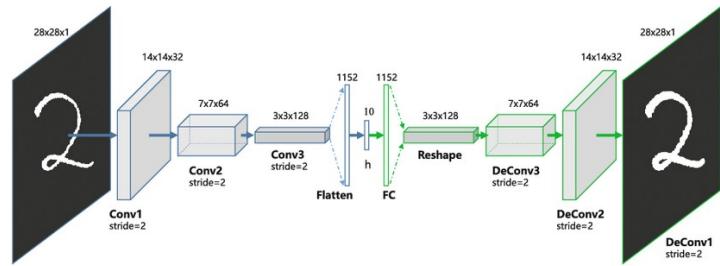
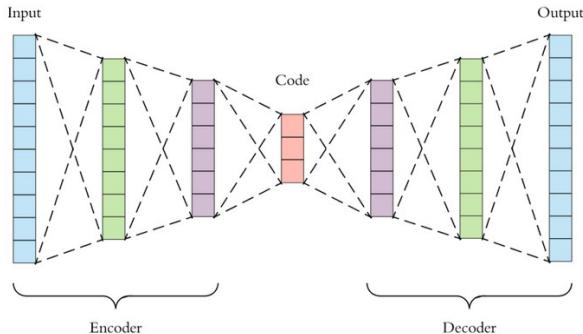


Arquitectura típica de un auto-encoder

En este punto la red ha comprimido la información de la entrada con la menor pérdida de información posible

Agrupamiento profundo: AEs

¿Cómo funcionan los auto-encoders?



- El **encoder** tiene como entrada los datos originales (x) y devuelve el **código** (z)

$$z = \sigma(\mathbf{W}x + \mathbf{b})$$

- El **decoder** toma como entrada el **código** y trata de reconstruir los datos originales (x')

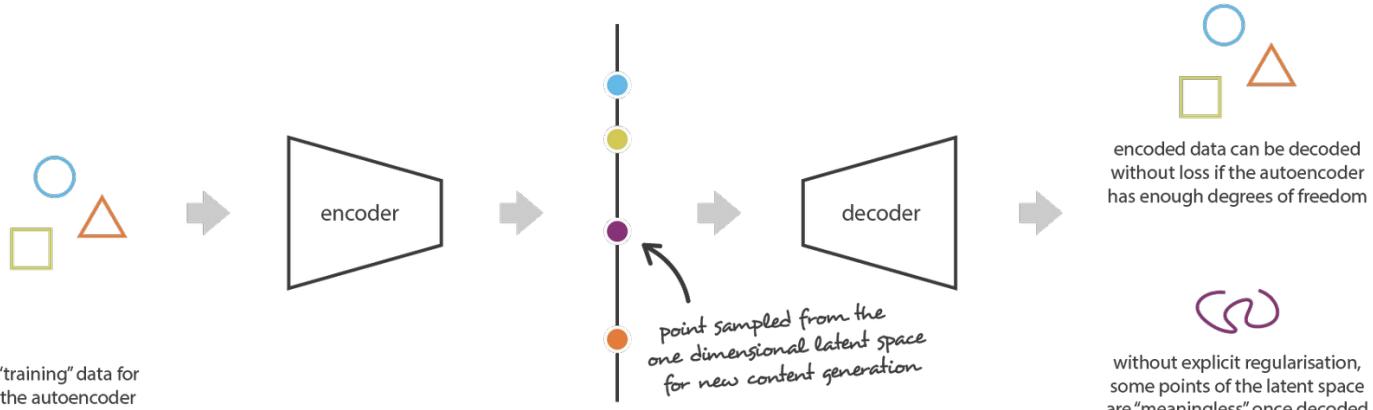
$$x' = \sigma'(\mathbf{W}'z + \mathbf{b}')$$

- La función de pérdidas se encarga de que x y x' sean lo más parecidas posible calculando el **error de reconstrucción**

$$\mathcal{L}(x, x') = \|x - x'\|^2 = \|x - \sigma'(\mathbf{W}'(\sigma(\mathbf{W}x + \mathbf{b})) + \mathbf{b}')\|^2$$

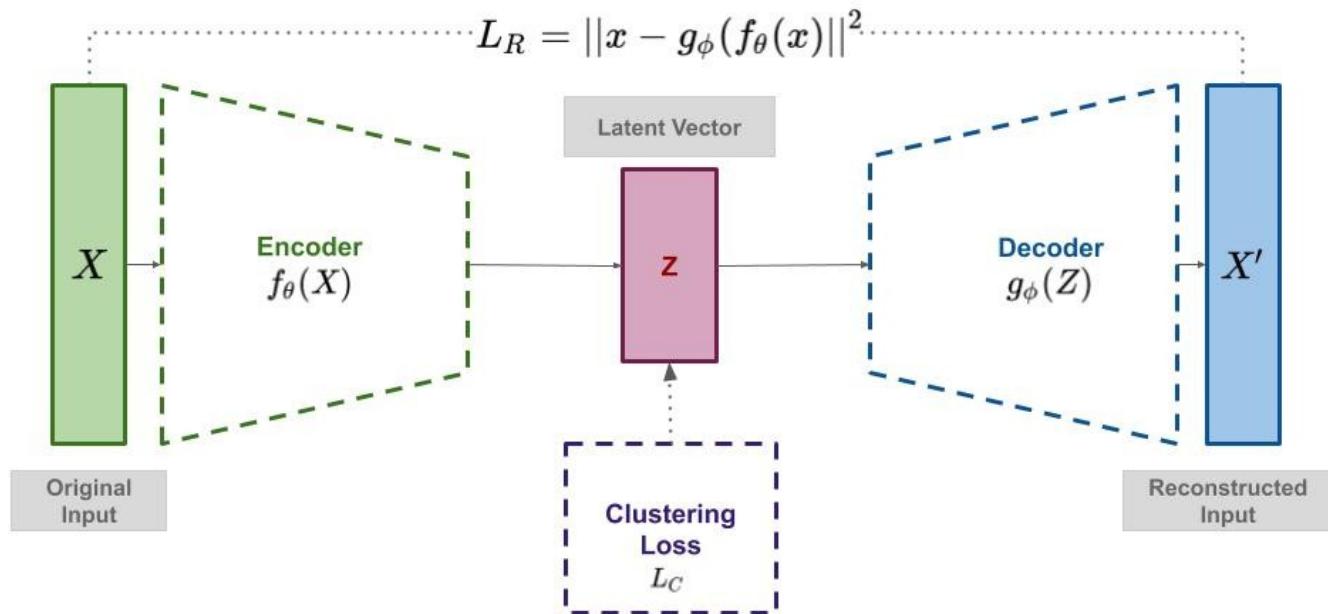
Agrupamiento profundo: AEs

¿Cómo funcionan los auto-encoders?



- El **auto-encoder** se preocupa únicamente de que la reconstrucción sea lo más parecida a la entrada, no de crear un espacio latente *organizado*.
- La muestra de entrada se codifica como un **punto** del espacio latente.

Agrupamiento profundo: AEs



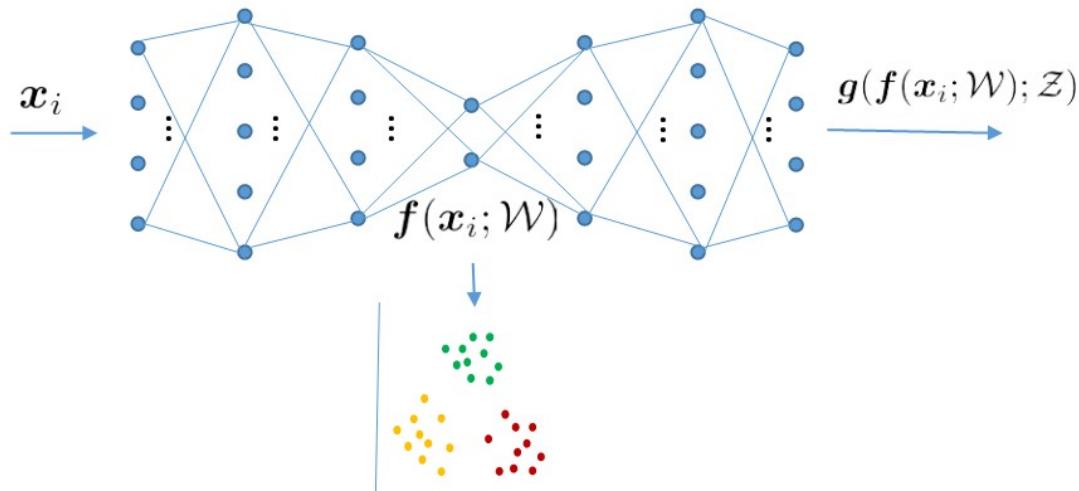
- Para que el espacio latente sea adecuado para realizar un clustering de buena calidad, es necesario añadir una **condición extra** al entrenamiento.
- La mayoría de métodos introducen un **término de pérdidas** que optimice el **clustering** obtenido.

Agrupamiento profundo: AEs

Método: Deep Clustering Network (DCN)

Utiliza un auto-encoder para aprender representaciones de los datos que puedan ser útiles a un K-means.

Primero, pre-entrena el AE, después, optimiza las **pérdidas de reconstrucción** y las **pérdidas del K-means** iterativamente.

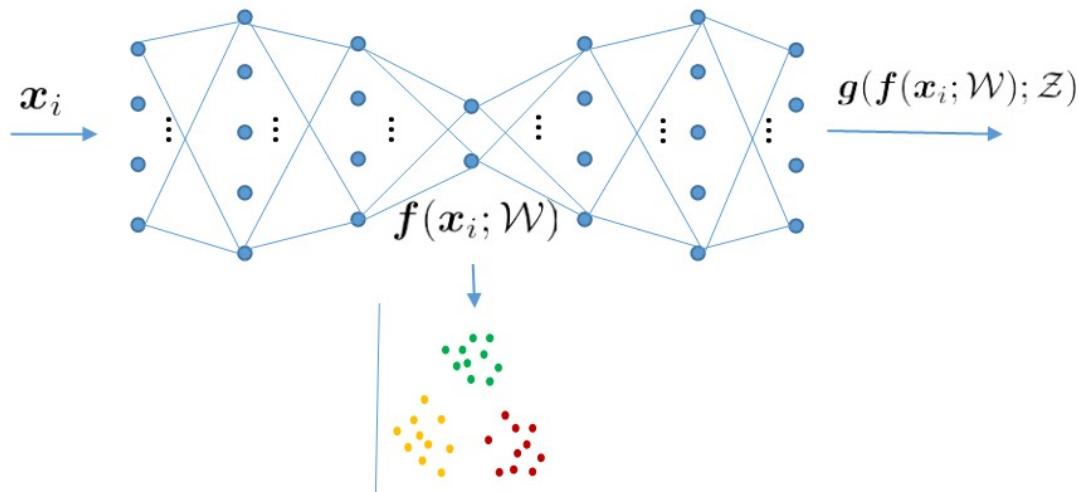


Agrupamiento profundo: AEs

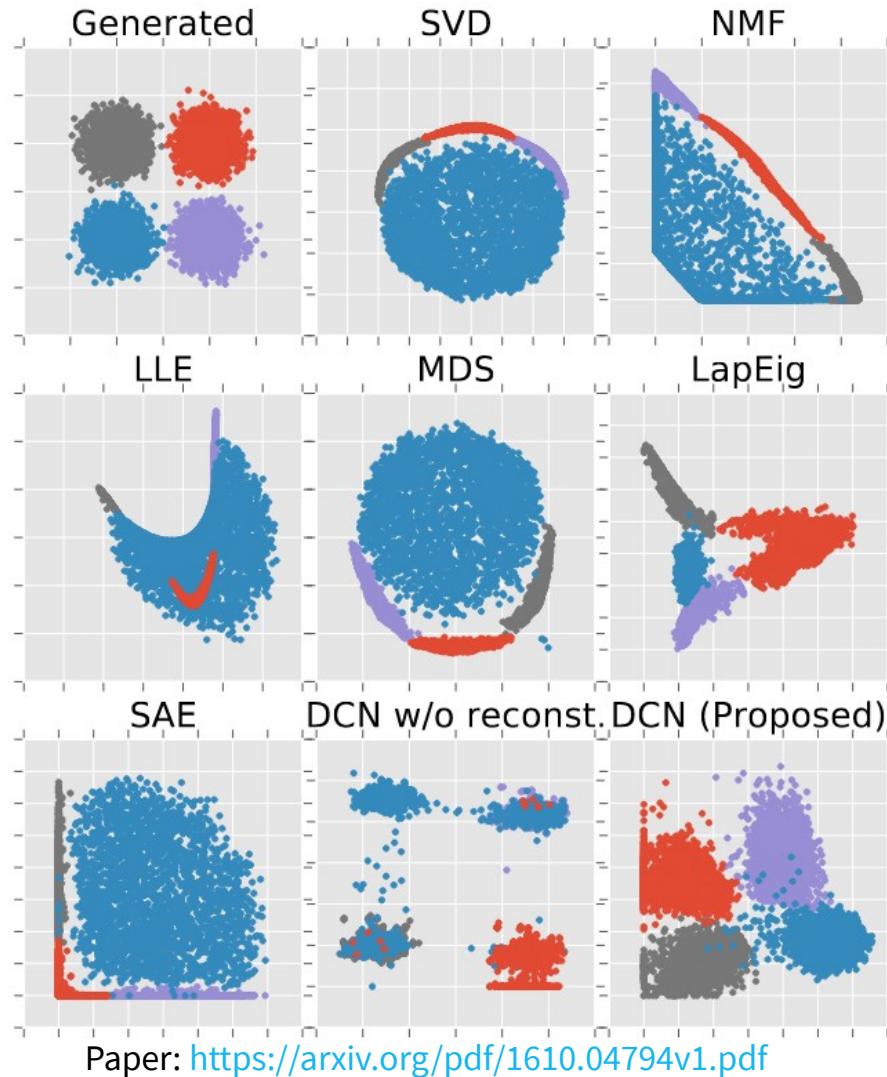
La optimización **minimiza**:

- El error de **reconstrucción**
- Las pérdidas introducidas por la función de pérdidas correspondiente al **K-means**

$$\min \sum_{i=1}^N \left(\ell(g(f(x_i)), x_i) + \frac{\lambda}{2} \|f(x_i) - M s_i\|_2^2 \right)$$



Agrupamiento profundo: AEs



Paper: <https://arxiv.org/pdf/1610.04794v1.pdf>

Agrupamiento profundo: AEs

¡Implementación disponible!

<https://github.com/xuyxu/Deep-Clustering-Network>

Agrupamiento profundo

Tipos de algoritmos de agrupamiento profundo

- Basados en auto-encoders
- **Basados en modelos generativos**
- Basados en optimización directa de cluster

Agrupamiento profundo: modelos generativos

Definición

El propósito de los **modelos generativos** es aprender la distribución de los datos originales para ser capaces de reproducirlos con la mayor fidelidad posible.

Dos métodos:

- Variational auto-encoder (VAE)
- Generative Adversarial Network (GAN)

Agrupamiento profundo: modelos generativos

Definición

El propósito de los **modelos generativos** es aprender la distribución de los datos originales para ser capaces de reproducirlos con la mayor fidelidad posible.

Dos métodos:

- **Variational auto-encoder (VAE)**
- Generative Adversarial Network (GAN)

Agrupamiento profundo: VAEs

Definición

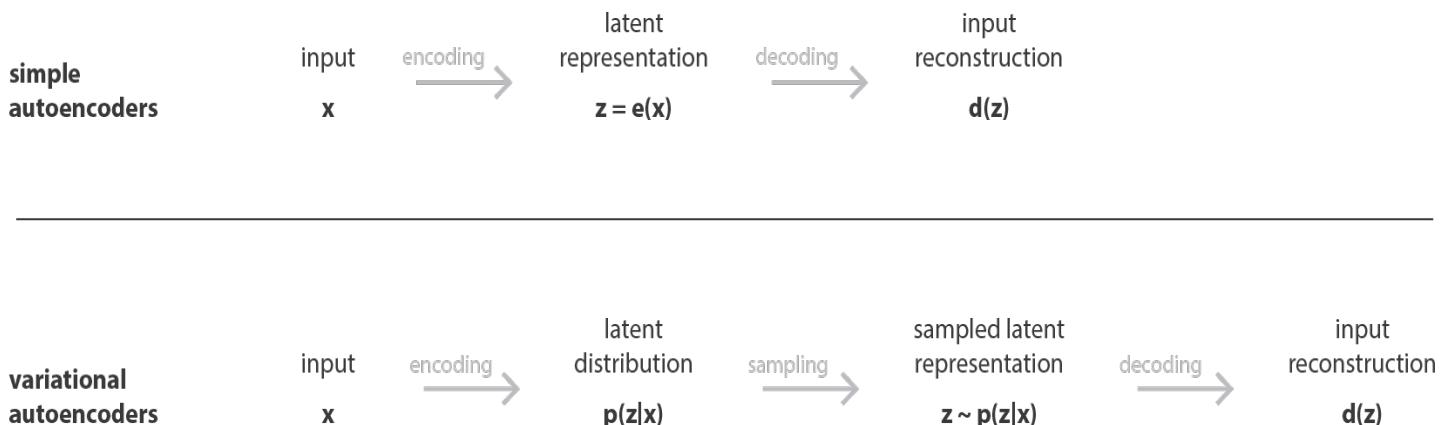
El propósito de los **auto-encoders variacionales** es reconstruir los datos que tienen a la entrada con el menor error posible tras haberlos **comprimido** a un espacio de menor dimensionalidad (espacio latente o *bottleneck*).

A diferencia del AE tradicional, el VAE codifica las muestras originales como una **distribución de probabilidad**, en vez de como un único punto del espacio latente.

Agrupamiento profundo: VAEs

Definición

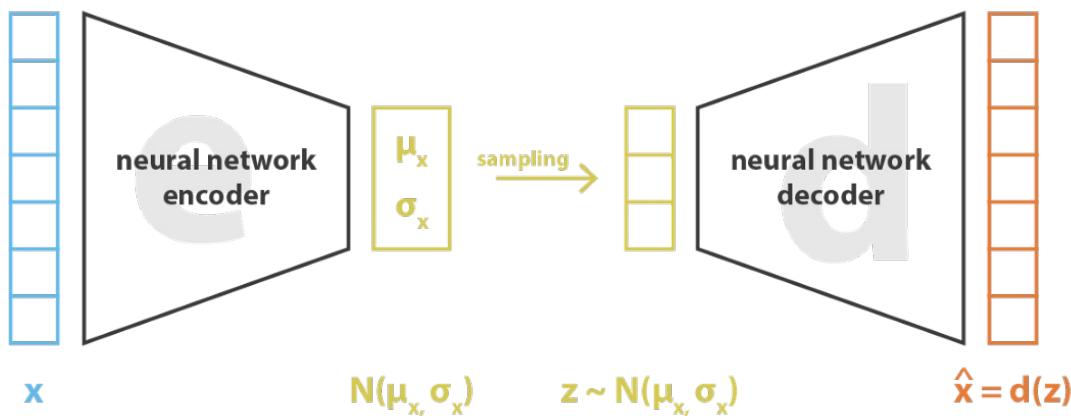
A diferencia del AE tradicional, el **VAE** codifica las muestras originales como una **distribución de probabilidad**, en vez de como un único punto del espacio latente.



Agrupamiento profundo: VAEs

Definición

A diferencia del AE tradicional, el **VAE** codifica las muestras originales como una **distribución de probabilidad**, en vez de como un único punto del espacio latente.

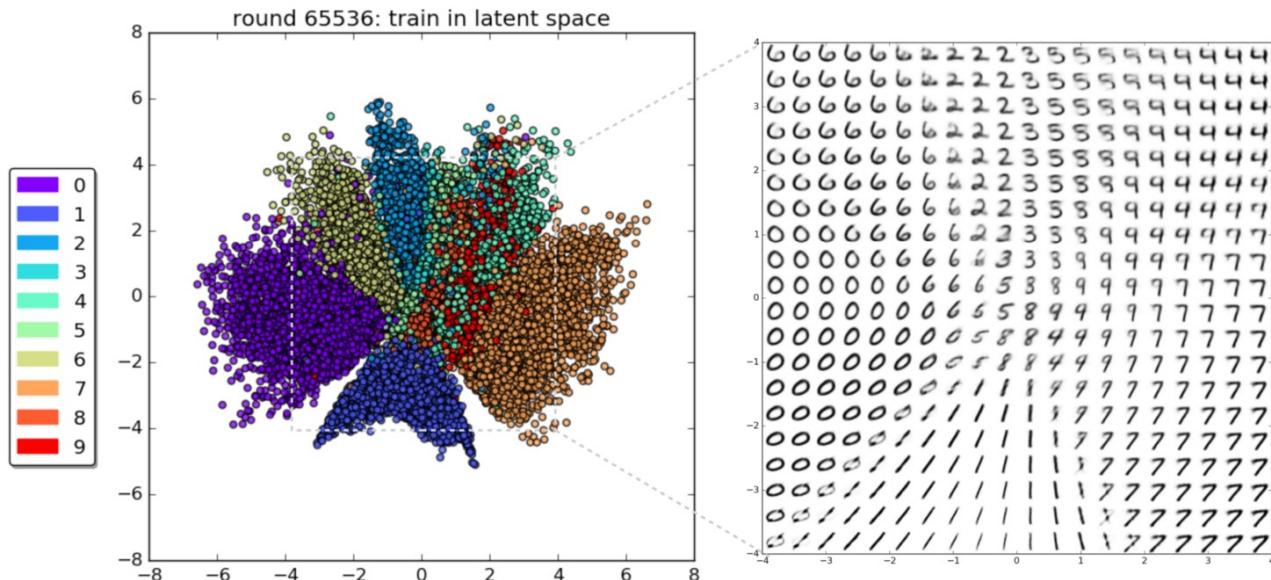


$$\text{loss} = \|x - \hat{x}\|^2 + \text{KL}[N(\mu_x, \sigma_x), N(0, I)] = \|x - d(z)\|^2 + \text{KL}[N(\mu_x, \sigma_x), N(0, I)]$$

Agrupamiento profundo: VAEs

Objetivo

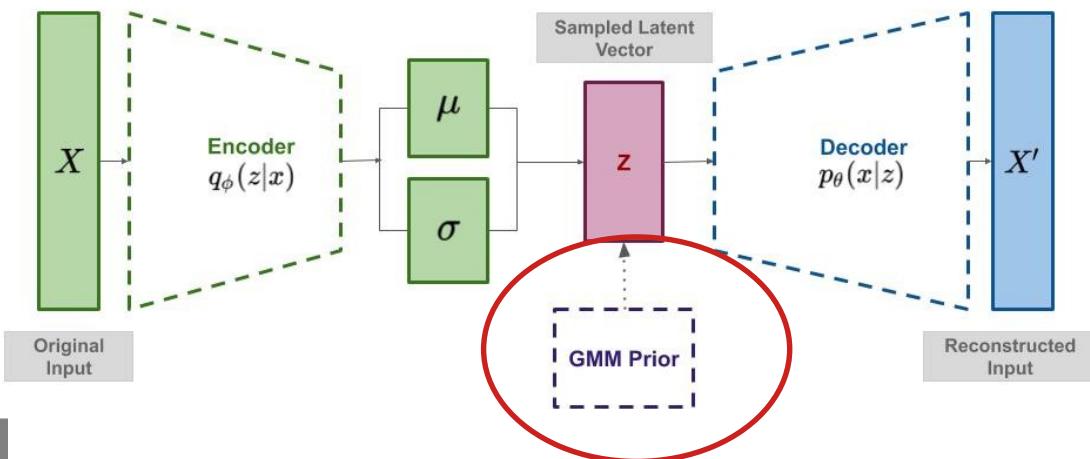
El objetivo del *Auto-Encoder Variacional* (VAE) es obtener un espacio latente “ordenado” del cual podamos muestrear y obtener nuevos datos similares (pero no iguales) a los originales.



Agrupamiento profundo: VAEs

Método: Variational Deep Embedding (VaDE)

Se impone una distribución de probabilidad **a priori** (*prior*) consistente en una **mixtura de Gaussianas** (GMM) y se optimizan los parámetros del modelo de forma que las **probabilidades** de que los datos puedan haber sido generados por dicha mixtura sea **máxima**.

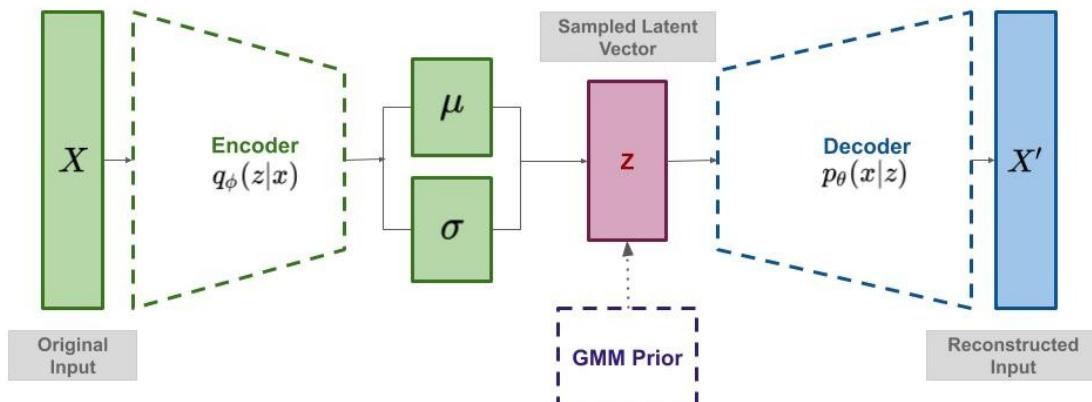


Agrupamiento profundo: VAEs

La optimización minimiza:

- El error de **reconstrucción**
- La **divergencia KL** entre la **mixtura de Gaussianas** y la **distribución aprendida por el VAE**

$$\mathcal{L}_{\text{ELBO}}(\mathbf{x}) = E_{q(\mathbf{z}, c|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] - D_{KL}(q(\mathbf{z}, c|\mathbf{x})\|p(\mathbf{z}, c))$$



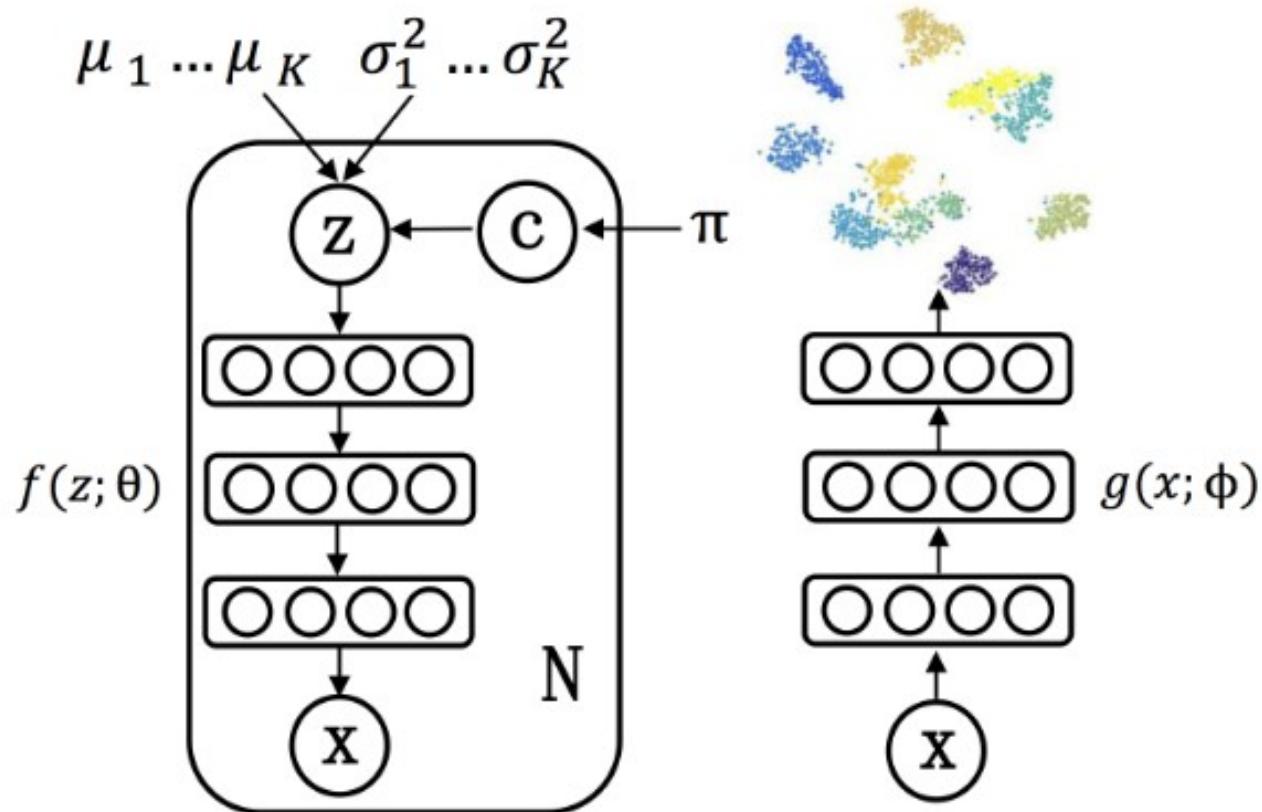
Agrupamiento profundo: VAEs

Método: Variational Deep Embedding (VaDE)

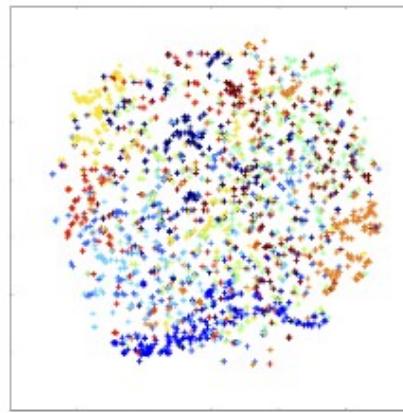
El modo de funcionamiento es el siguiente:

- 1) Se selecciona una componente de la mixtura de Gaussianas
- 2) Se muestrea dicha componente, obteniendo un código latente
- 3) Se introduce el código latente al decoder y se obtiene una posible reconstrucción de x
- 4) Se calculan las pérdidas y se actualizan los pesos

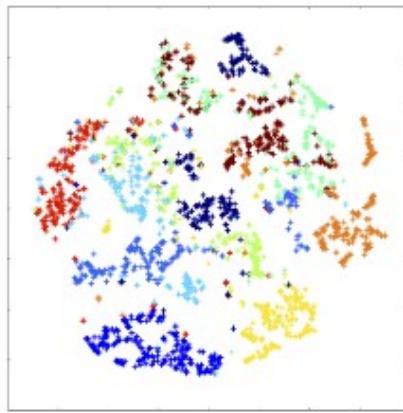
Agrupamiento profundo: VAEs



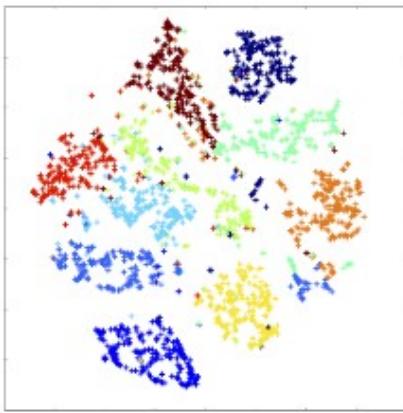
Agrupamiento profundo: VAEs



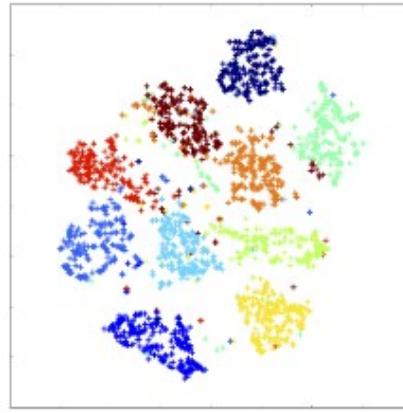
(a) Epoch 0 (11.35%)



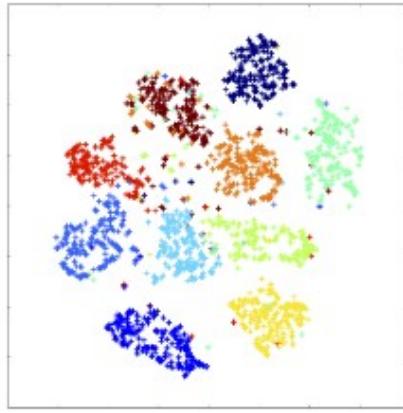
(b) Epoch 1 (55.63%)



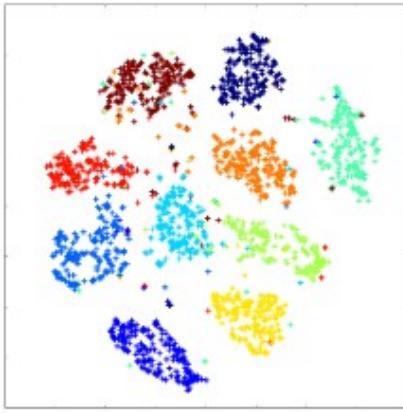
(c) Epoch 5 (72.40%)



(d) Epoch 50 (84.59%)



(e) Epoch 120 (90.76%)



(f) Epoch End (94.46%)

Agrupamiento profundo: VAEs

[Examen] ¿Qué se tiene que optimizar en un algoritmo de deep clustering basado en VAE?

- VAE
GMM
- { - los pesos de la VAE
 - { - " factores " mezcla
 - { - las medias de las gaussianas
 - " matrices de covarianza de las gaussianas.

¡Implementación disponible!

<https://github.com/slim1017/VaDE>

Agrupamiento profundo: modelos generativos

Definición

El propósito de los **modelos generativos** es aprender la distribución de los datos originales para ser capaces de reproducirlos con la mayor fidelidad posible.

Dos métodos:

- Variational auto-encoder (VAE)
- **Generative Adversarial Network (GAN)**

Agrupamiento profundo: GANs

Definición

El propósito de las **redes generativas adversarias** es generar nuevo contenido capaz de hacerse pasar por el original partiendo de **ruido** a la entrada (en su versión más básica).

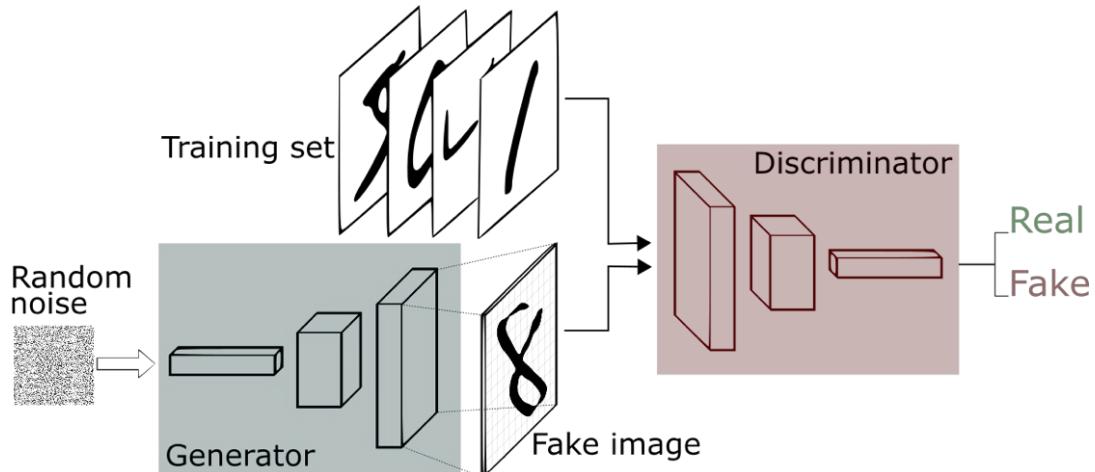
Durante el entrenamiento, el modelo aprende a generar contenido **realista** partiendo del ruido original, es decir, encuentra un **“mapeo”** adecuado entre vectores (o matrices) de elementos aleatorios a contenido realista.

Se usan sobretodo, pero no únicamente, en el ámbito de la **imagen**.

Agrupamiento profundo: GANs

Definición

El propósito de las **redes generativas adversarias** es generar nuevo contenido capaz de hacerse pasar por el original partiendo de **ruido** a la entrada (en su versión más básica).

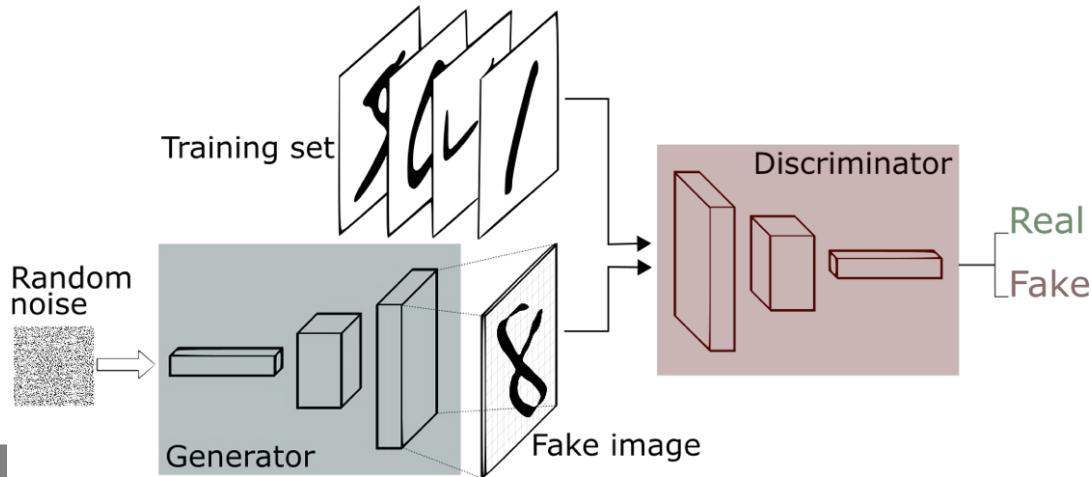


Arquitectura típica de una GAN

Agrupamiento profundo: GANs

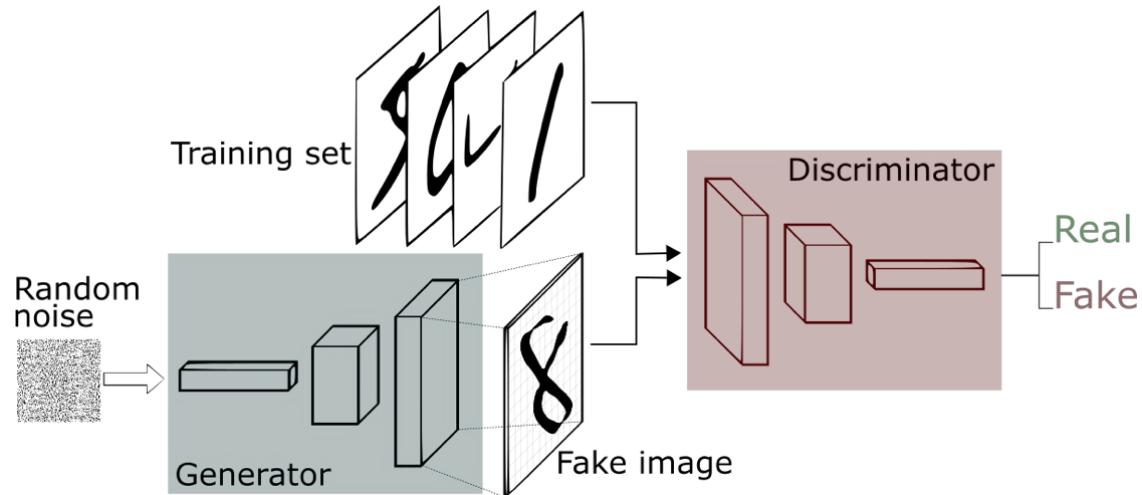
¿Cómo funcionan las GANs?

- El **generador** (“falsificador”) trata de generar contenido lo suficientemente realista como para **engañar** al discriminador (“policía”)
- El discriminador trata de **distinguir** el contenido real del falso
¡Competen entre ellos! → inestabilidad :(



Agrupamiento profundo: GANs

¿Cómo funcionan las GANs?



$$\mathcal{L}_{adv} = \min_G \max_D V(D, G) = \underbrace{\mathbb{E}_{x \sim p_{data}(x)} [\log D(x)]}_{\text{probabilidad de que } D(x) \text{ prediga que los datos reales son verdaderos}} + \underbrace{\mathbb{E}_{z \sim p_z(z)} [1 - \log D(G(z))]}_{\text{probabilidad de que } D(x) \text{ prediga que los datos generados, } x=G(z), \text{ no son verdaderos}}$$

probabilidad de que
D(x) prediga que los
datos reales **son**
verdaderos

probabilidad de que D(x)
prediga que los datos
generados, x=G(z), **no** son
verdaderos

Agrupamiento profundo: GANs

¿Cómo se entrena las GANs?

El entrenamiento ocurre en **2 fases**:

- El discriminador recibe imágenes **reales y falsas**, con las **etiquetas correspondientes** (0 para las falsas y 1 para las reales), y debe aprender a predecir correctamente. Cuando se **equivoca**, sus **pesos** se **actualizan**.
- Con los pesos del **discriminador congelados**, el **generador** introduce **imágenes generadas** con **etiqueta “real”** (1) al **discriminador**. **Cuando** el discriminador predice que la imagen es falsa, al no coincidir la etiqueta con la predicción, se genera un **error** que permite **actualizar los pesos del generador** para que aprenda a sintetizar imágenes más realistas.

Agrupamiento profundo: GANs

Método: Information Maximizing GAN (InfoGAN)

El propósito principal de InfoGAN es el de “desenredar” las distintas componentes presentes en los datos (“*disentangled representations*”).

Para ello, InfoGAN descompone la entrada en dos partes:

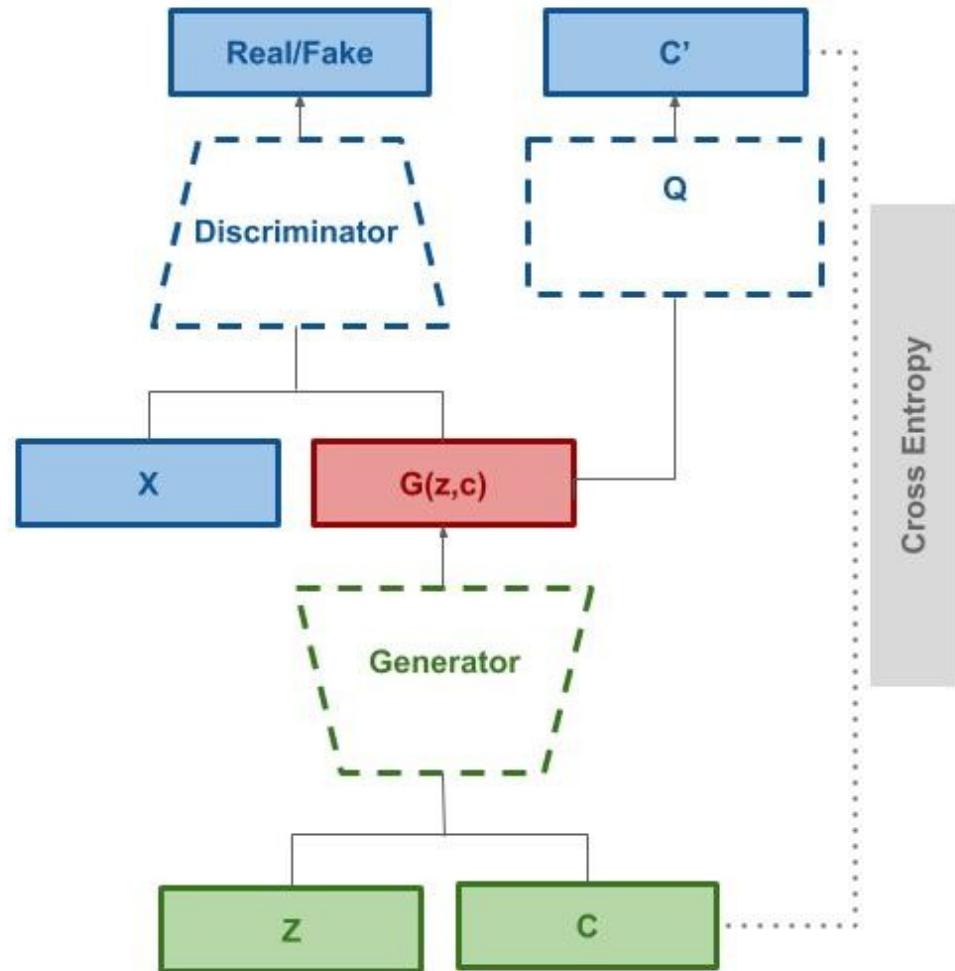
- Ruido incomprensible: z
- El código latente: c

El generador pasa a tener 2 entradas: $G(z, c)$.

Se combina el objetivo GAN con un término de regularización basado en la información mútua: $I(c; G(z, c))$.

El objetivo principal **no es el clustering**, pero se aprovecha para ello.

Agrupamiento profundo: GANs



Paper: <https://arxiv.org/pdf/1606.03657.pdf>

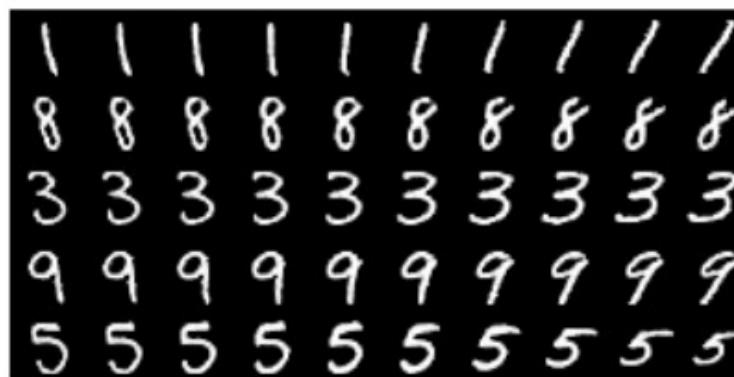
Agrupamiento profundo: GANs



(a) Varying c_1 on InfoGAN (Digit type)



(b) Varying c_1 on regular GAN (No clear meaning)



(c) Varying c_2 from -2 to 2 on InfoGAN (Rotation)



(d) Varying c_3 from -2 to 2 on InfoGAN (Width)

Agrupamiento profundo: GANs

¡Implementación disponible!

<https://github.com/openai/InfoGAN>

Agrupamiento profundo

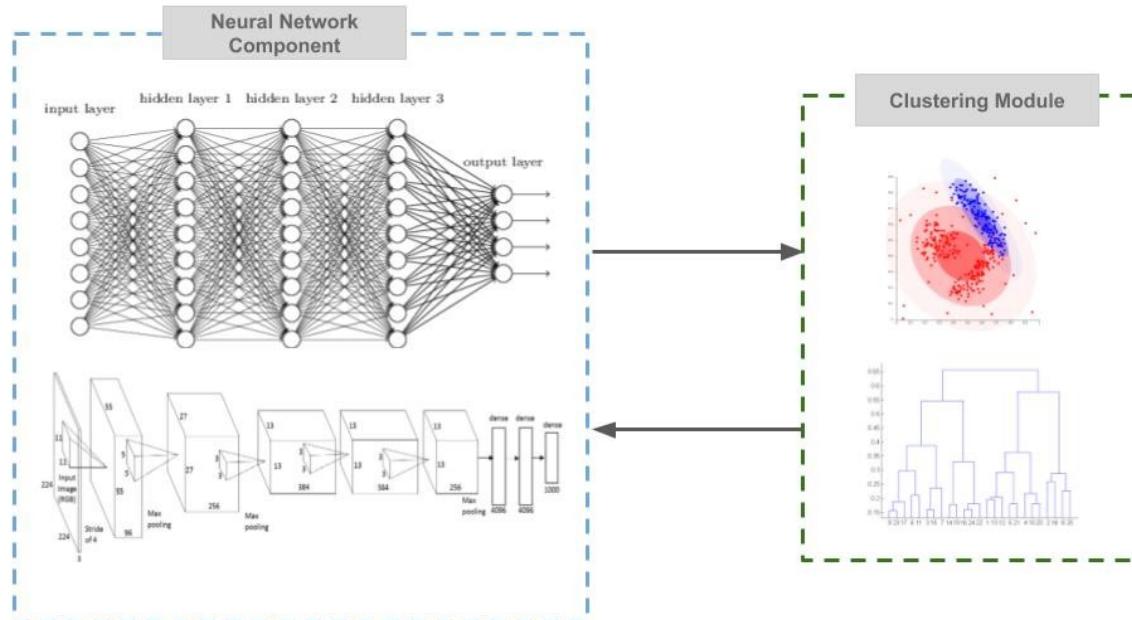
Tipos de algoritmos de agrupamiento profundo

- Basados en auto-encoders
- Basados en modelos generativos
- Basados en optimización directa de cluster**

Agrupamiento profundo: optimización directa

Definición

En esta clase de métodos se prescinde de cualquier posible término de pérdidas de reconstrucción y se utiliza **únicamente** el término de **pérdidas de agrupamiento** para optimizar la red neuronal.



Agrupamiento profundo: optimización directa

Método: Joint Unsupervised Learning (JULE)

Este método emplea una **red neuronal convolucional** junto a una función de **pérdidas** de **agrupamiento aglomerativo**.

En cada **predicción** (*forward pass*) se realiza **clustering jerárquico** usando una medida de **afinidad** y las representaciones se **optimizan** en el *backward pass*.

Tiene el **inconveniente** de que requiere la construcción de una matriz de afinidad no dirigida que requiere de mucha **memoria** y capacidad de **cómputo**.

Agrupamiento profundo: GANs

¡Implementación disponible!

<https://github.com/FJR-Nancy/joint-cluster-cnn>

Agrupamiento profundo: GANs

Si queréis profundizar:

<https://deepnotes.io/deep-clustering>

[Unsupervised Clustering for Deep Learning: A tutorial survey](#)

[Deep clustering: methods and implements](#)

Agrupamiento profundo

Ventajas

- Las redes neuronales profundas permiten obtener **representaciones reducidas** de los datos que son más adecuadas para realizar **buenas agrupaciones**.
- El entrenamiento de estos modelos es **end-to-end**: combina *feature extraction, dimensionality reduction* y *clustering*.
- **Escalabilidad**: gracias a las redes neuronales profundas podemos procesar enormes datasets altamente dimensionales.

Agrupamiento profundo

Desventajas

- **Hiper-parámetros:** las redes neuronales y las pérdidas utilizadas en los algoritmos de agrupamiento profundo dependen de hiper-parámetros que no son sencillos de escoger.
- **Falta de interpretabilidad:** las redes neuronales profundas dificultan bastante la interpretabilidad de los modelos.
- **Falta de base teórica:** aunque en la práctica funcionen, los fundamentos teóricos no son sólidos.

Aprendizaje no supervisado

Deep clustering

Rocío del Amor del Amor
mrocio.delamor@campusviu.es

Universidad Internacional de Valencia