Universidad Popular del Cesar

Ciencia de Datos

Clustering 2

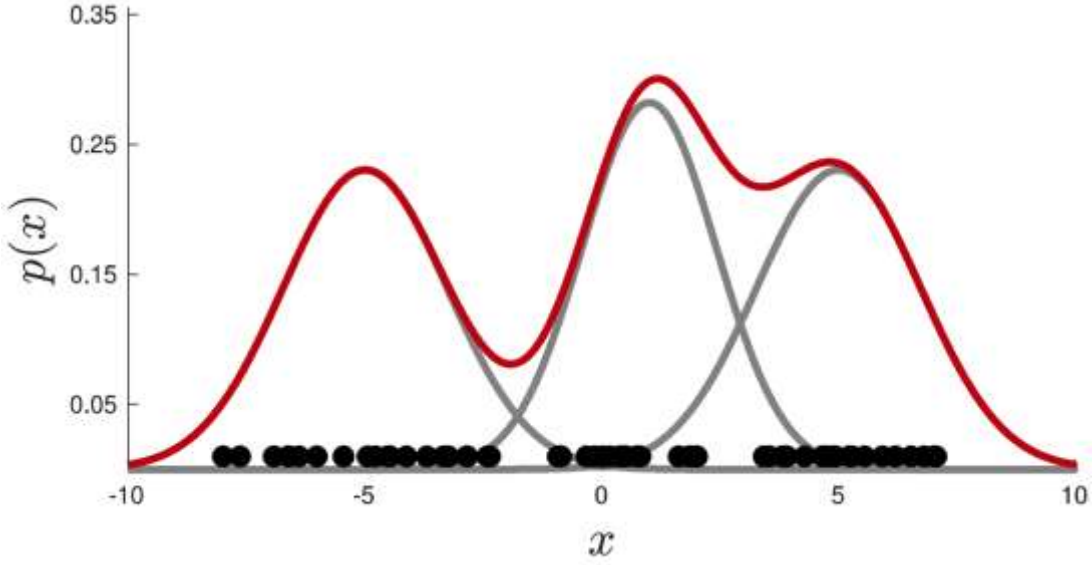
Dr. José Ramón Iglesias
DSP-ASIC BUILDER GROUP
Director Semillero TRIAC
Ingenieria Electronica
Universidad Popular del Cesar

Mapa de Ruta

1. Intuición general de clustering
2. Conocimiento de los Datos e Información Relevante al problema
3. Importancia del conocimiento de dominio
4. Similaridad y/o Distancia entre datos
5. Algoritmos de agrupamiento (GMM y Jerárquicos)
6. Evaluación de resultados: Visualización, Medidas y relevancia: utilidad o impacto

Mezcla de Gaussianas (GMM)

- Supongamos tener alguna información
  - Datos numéricos (reales),
  - producidos por una densidad mezcla de Gaussianas,



Cómo funciona el GMM?

- Se fija la cantidad de gaussianas,
  - se estiman los parámetros de cada gaussiana
  - se asigna c/dato como proveniente de una de las componentes de la mezcla.
  - La estimación se realiza mediante el algoritmo Expectation Maximization.

Cómo funciona GMM?

Expectation Maximization Algorithm

Input: X= $\{\mathbf{x}_i | i = 1, \ldots, n\}$ **datos,** m**=numero de componentes,** ϵ **tolerancia**

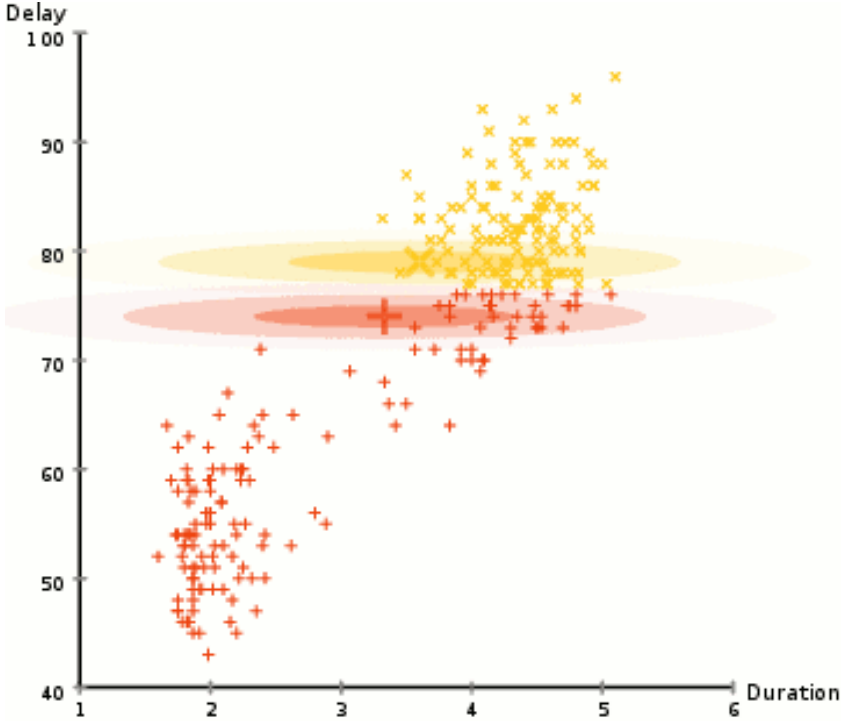
Output: $\hat{\boldsymbol{\Theta}}$

- ## $\hat{\boldsymbol{\Theta}}^0 \leftarrow \left[\left(\hat{\boldsymbol{\theta}}_1, \hat{p}_1\right)^0, \ldots, \left(\hat{\boldsymbol{\theta}}_m, \hat{p}_m\right)^0\right]$
- ## $t \leftarrow 0$

do
- ## $t \leftarrow t + 1$
- ## **Paso-E:** $Q\left(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}^t\right) \leftarrow \mathbb{E}\left[\sum_{i=1}^{\infty} \ln\left(p\left(\mathbf{x}_i | j; \hat{\boldsymbol{\Theta}}_j^t\right) p_j^t\right)\right]$
- ## **Paso-M:** $\hat{\boldsymbol{\Theta}}^{t+1} \leftarrow \arg\max_{\boldsymbol{\Theta}} Q\left(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}^t\right)$
- ## **until** $\left|Q\left(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}^t\right) - Q\left(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}^{t+1}\right)\right| < \epsilon$

Cómo funciona GMM?

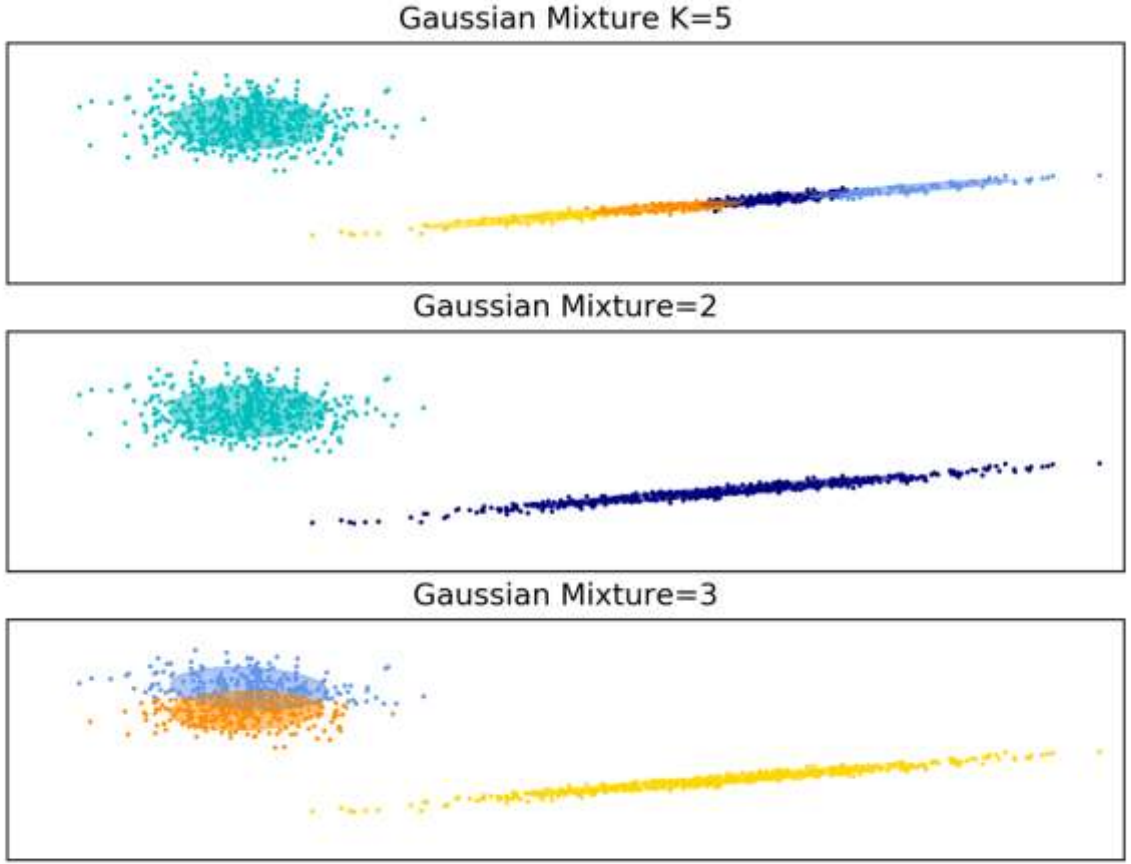
- Comenzamos con una partición aleatoria de la cual se sacan los parámetros de inicio y desde allí se itera



Parámetros

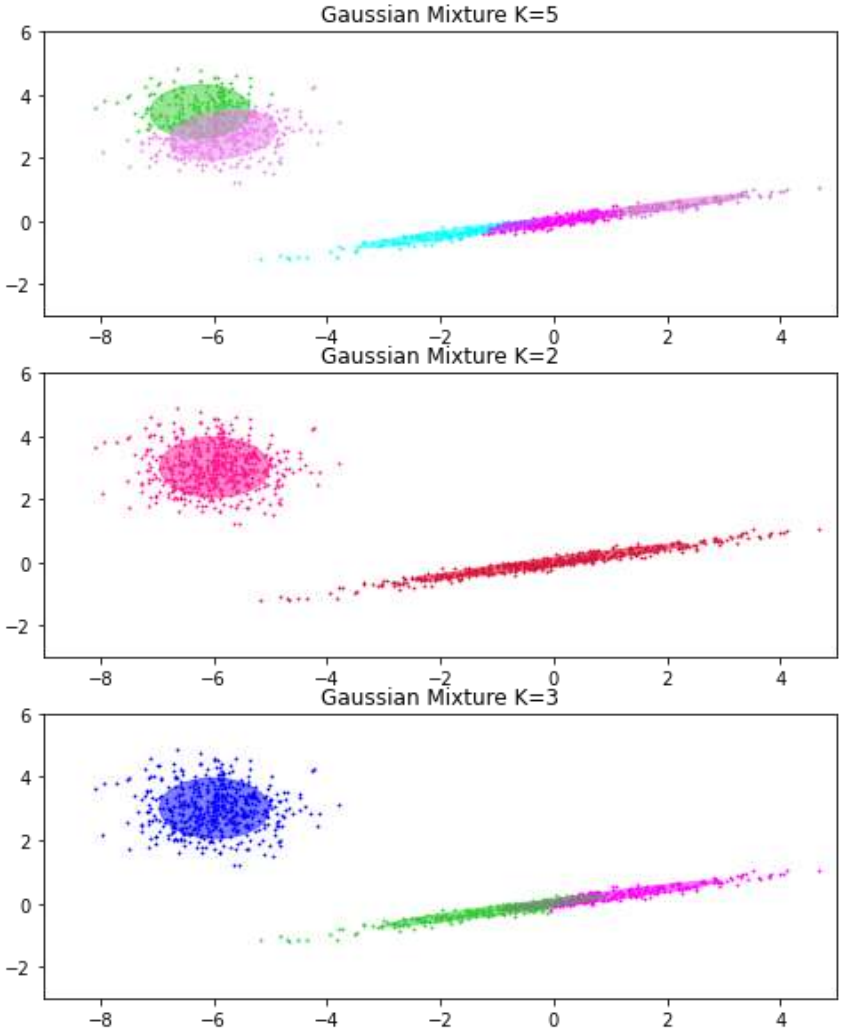
- Gran problema de GMM es la determinación del número de componentes de la mezcla
- Si no se elige un buen número, el modelo parte de forma aglutinada pero los clusters pueden no tener sentido.
- La otra característica que puede ser forzada de inicio es el tipo de matriz de varianza covarianza:
  - 'full' (each component has its own general covariance matrix),
  - 'tied' (all components share the same general covariance matrix),
  - 'diag' (each component has its own diagonal covariance matrix),
  - 'spherical' (each component has its own single variance).

Note_fig2.ipynb



Matriz de covarianza ‘full’ del módulo sklearn.

Note_fig2.ipynb



Matriz de covarianza ‘full’ del módulo sklearn.

Elección del modelo: k y tipo de matriz

- Bayesian Information Criterium (BIC) da un score al modelo con m parámetros.

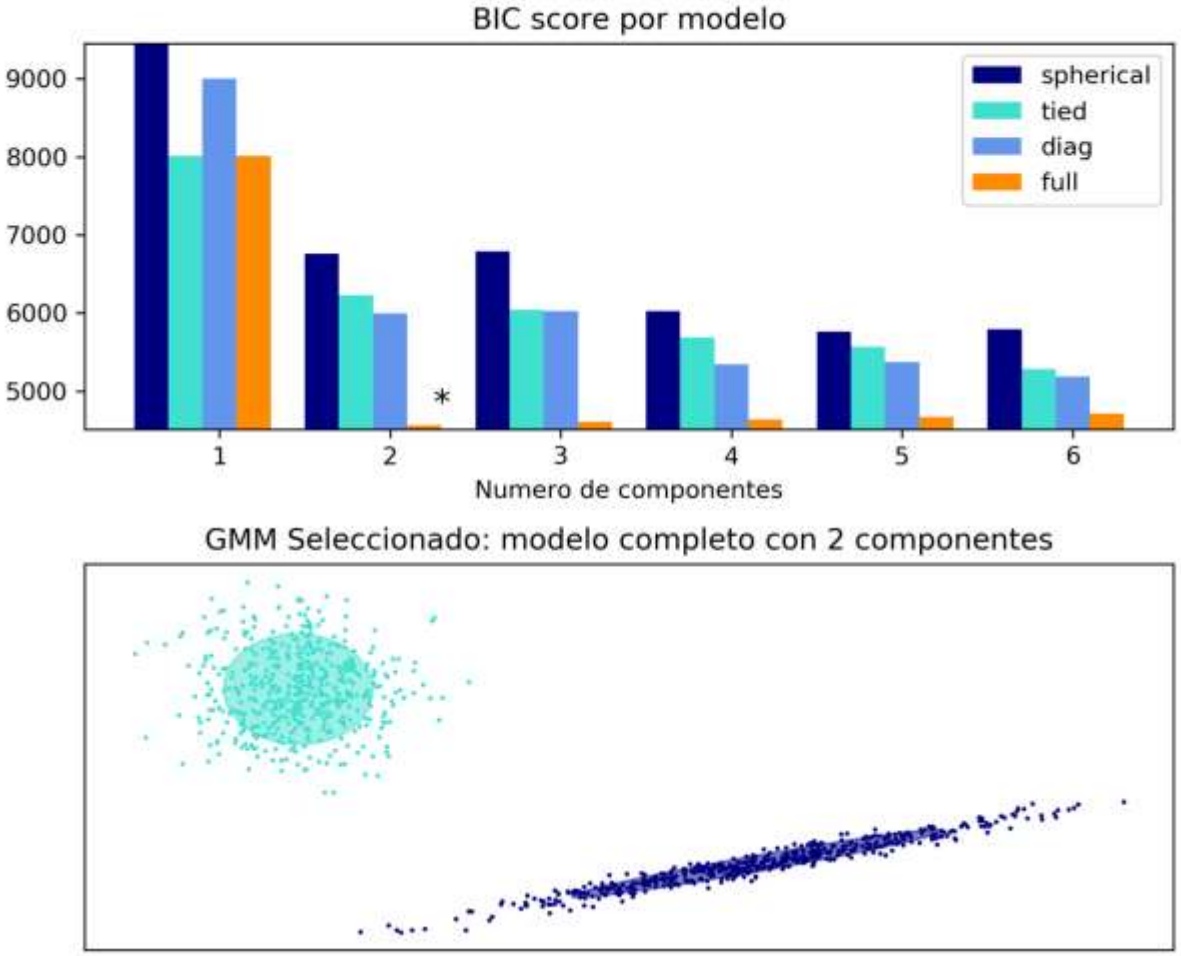
  $BIC = -2 * log(L(\Theta)) + log(n)m$

- Puede usarse otro índice llamado Akaike Information Criterium (AIC)
  $AIC = -2 * log(L(\Theta)) + 2m$

  donde L(Θ) es la verosimilitud, n el número de datos, y m el número de parámetros estimados, (k, el número de componentes, más las medias y entradas de la matriz de varianza covarianza.)

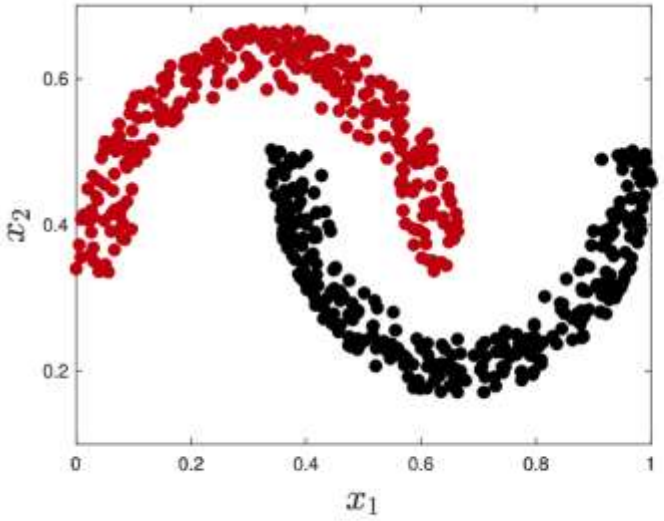
- La figura siguiente está generada por el script Note_fig3.ipynb

Note_fig3.ipynb

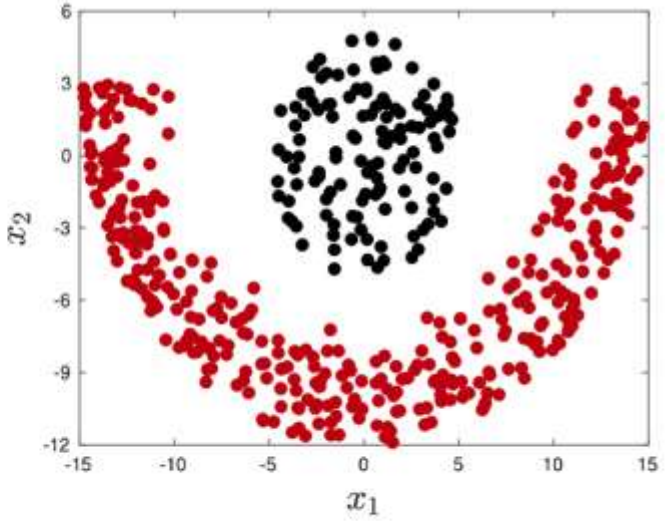


Mezcla de Gaussianas (GMM)
Gaussian Mixture Model

- Cualquier dato puede ser modelado con una mezcla de gaussianas?



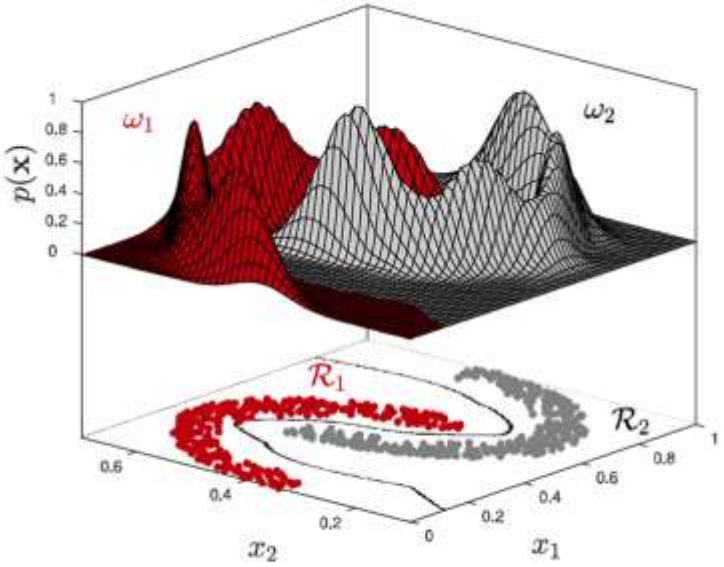
(a)



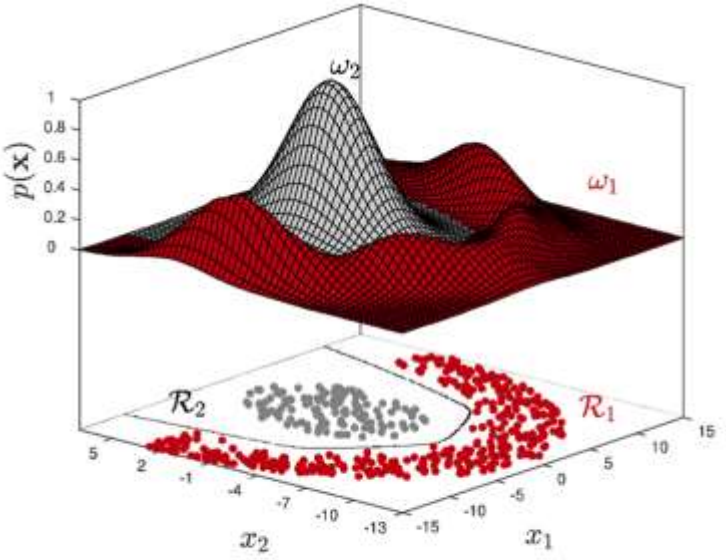
(b)

Mezcla de Gaussianas (GMM)

- No todos, pero muchos si se pueden modelar, si uno conoce la cantidad de gaussianas que forman la mezcla
-



(a)



(b)

Clustering jerárquico aglomerativo

Bottom-up

- Al comenzar, cada objeto es su propio cluster

- Se define una noción de semejanza **entre clusters** usando una medida de similaridad o distancia entre elementos

- Se unen en un solo cluster el par de clusters más semejantes

- La historia de uniones forma un árbol (dendograma)

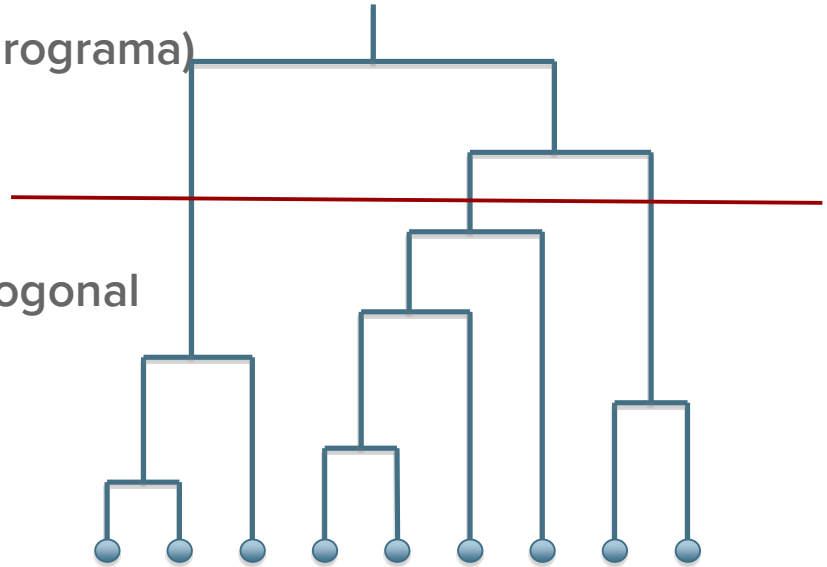
Clustering jerárquico

Si no queremos especificar k...

Algoritmos jerárquicos que generan una

taxonomía jerárquica de clusters (dendrograma)

- Interpretación más rica
- Más difícil de interpretar
- El corte del árbol tiene que ser ortogonal



Semejanza entre clusters

Single-linkage

1. Para cada par de clusters A y B, el par de objetos a, b más cercanos tal que a pertenece a A y b pertenece a B
2. Se unen los clusters con el par de objetos más semejante

Complete-linkage

1. Para cada par de clusters A y B, el par de objetos a, b más distantes tal que a pertenece a A y b pertenece a B
2. Se unen los clusters con el par de objetos más semejante

Semejanza entre clusters

Average-link

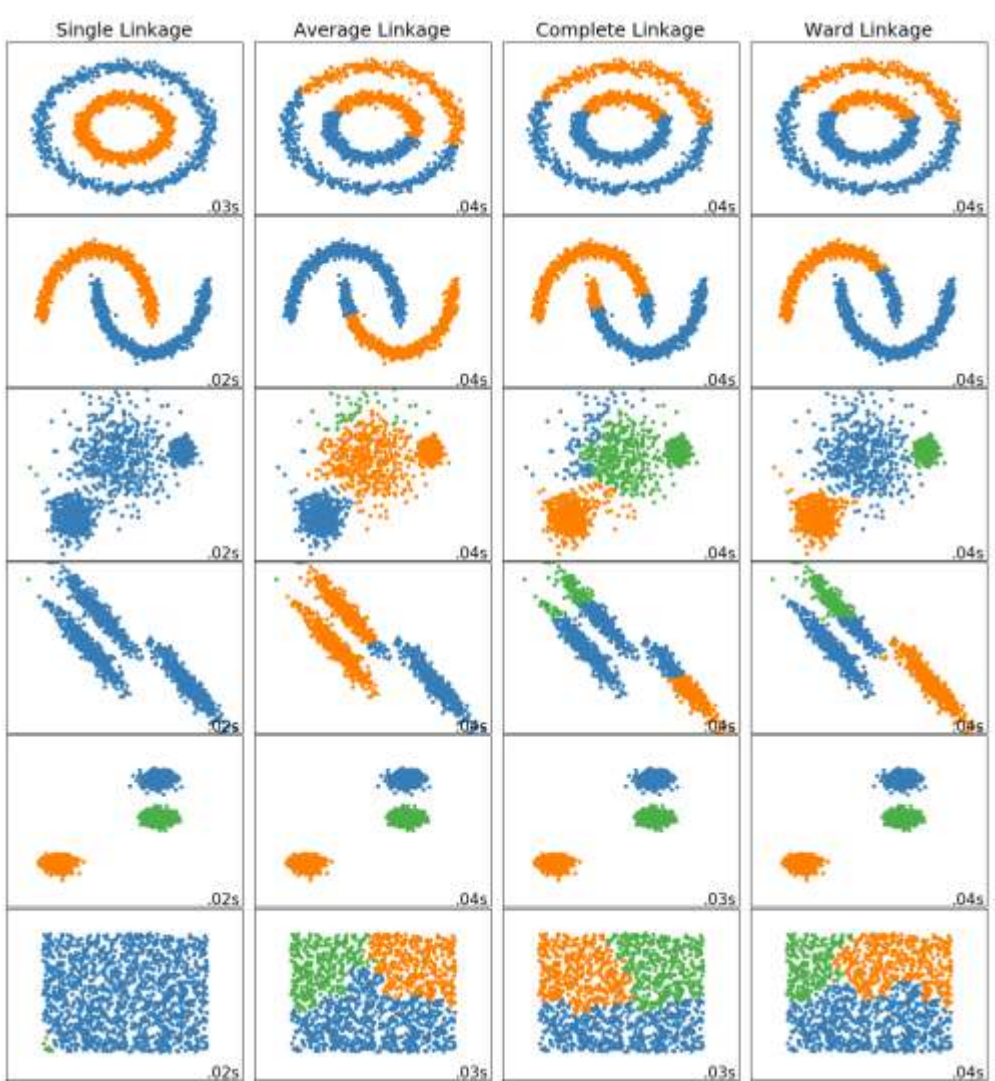
1. Para cada par de clusters A y B, se calcula la distancia entre todo par de objetos a, b tal que a pertenece a A y b pertenece a B
2. Se unen los clusters con el promedio de distancia más bajo

Centroid: Se unen los clusters con los centroides más cercanos

Ward-link

1. Para cada par de clusters A y B, se calcula la varianza al unirlos
2. Se unen los clusters con la mínima varianza

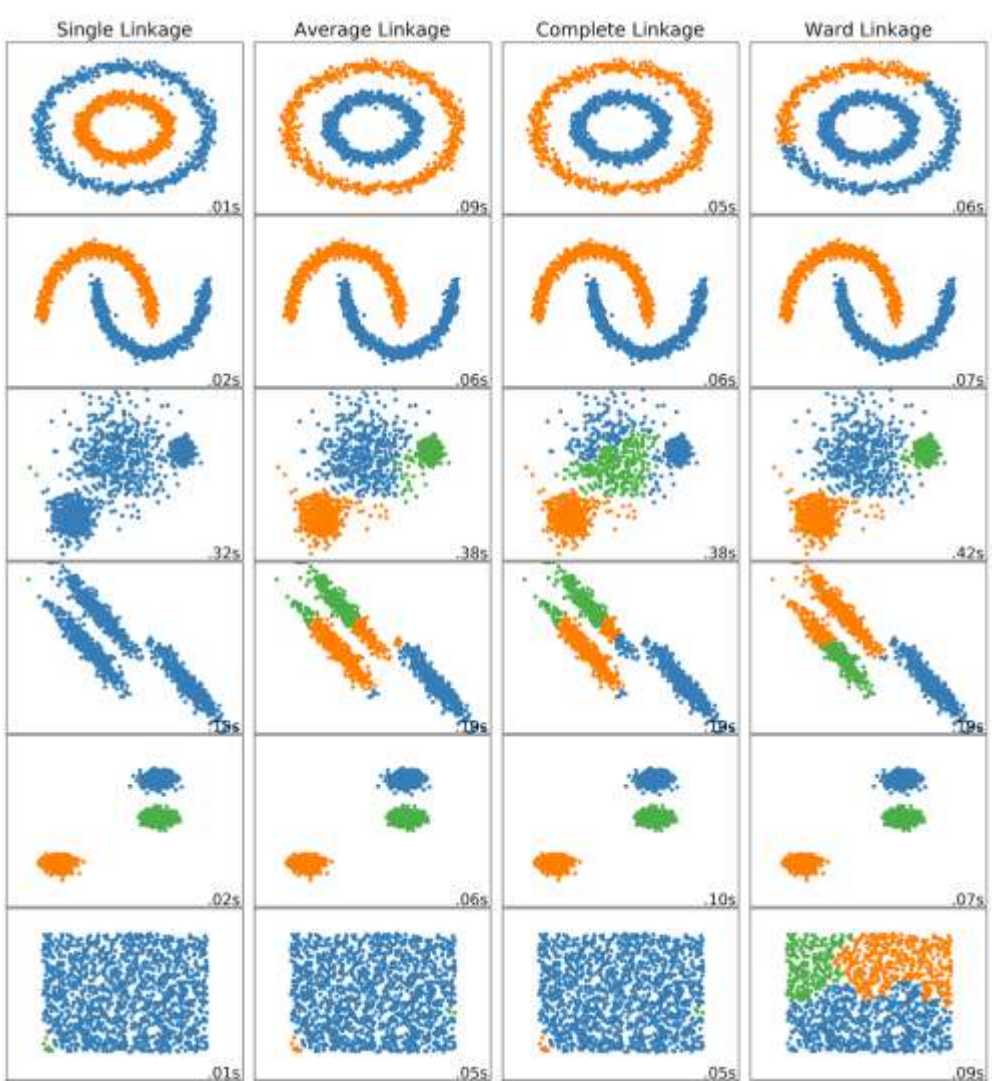
Clustering jerárquico note_fig9.ipynb



Agregando restricción de conectividad

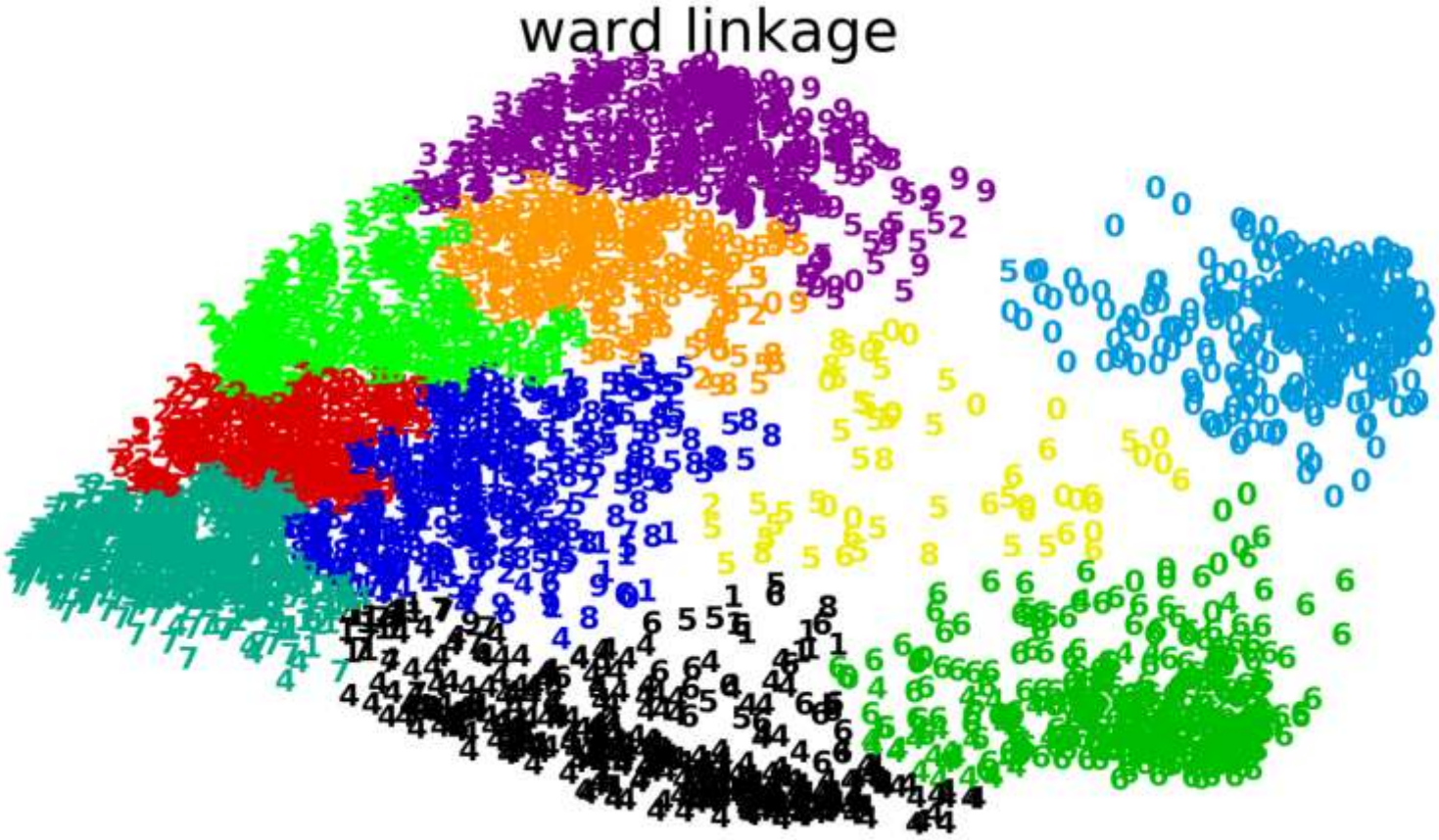
1. Las restricciones de conectividad implican que solo clusters adyacentes pueden unirse
2. Para ello se usa una matriz de conectividad que define los vecinos de cada muestra dada una estructura de los datos
3. Puede incitar a armar clusters desbalanceados, algunos muy grandes y otros muy pequeños.

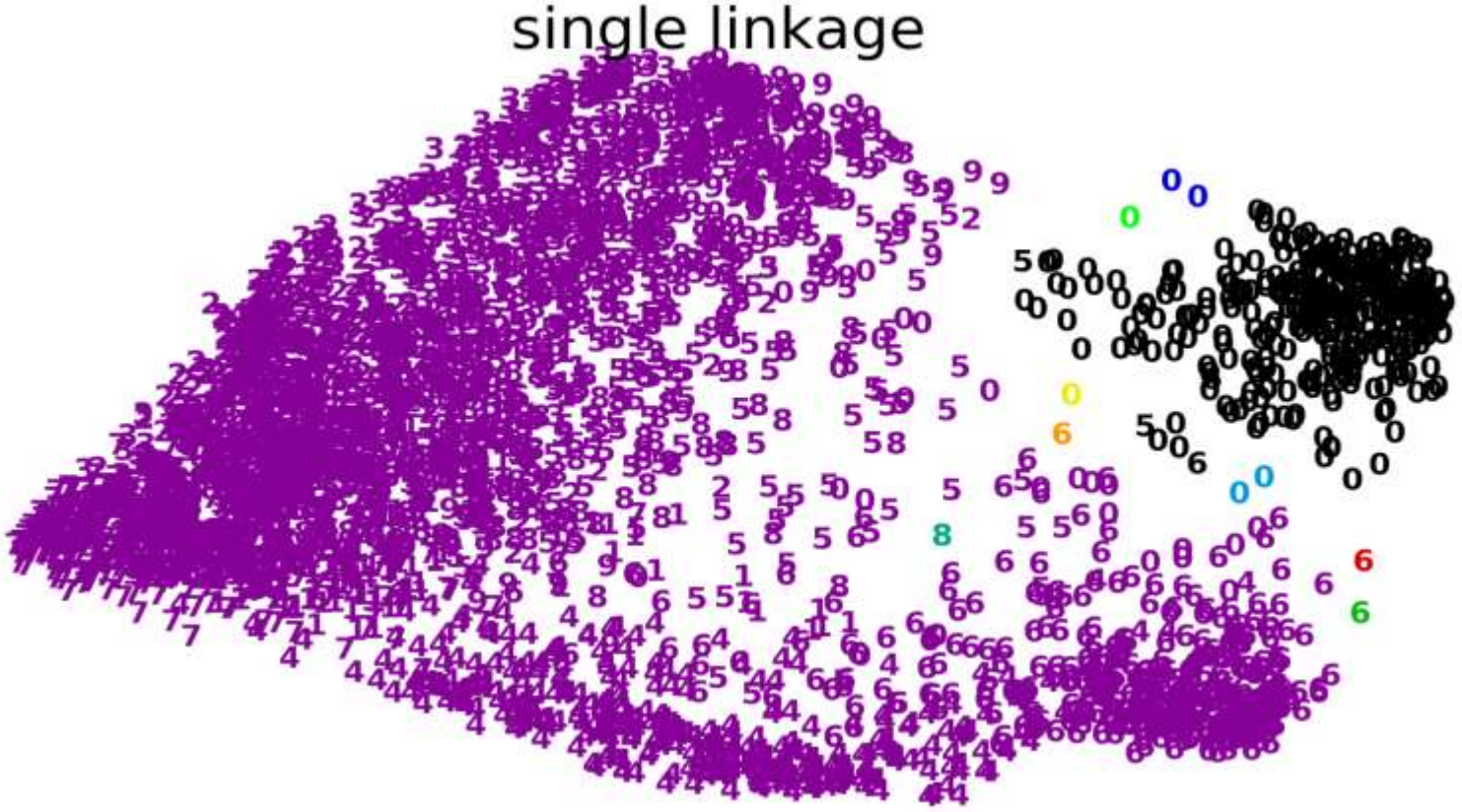
Clustering jerárquico con restriccion de conectividad

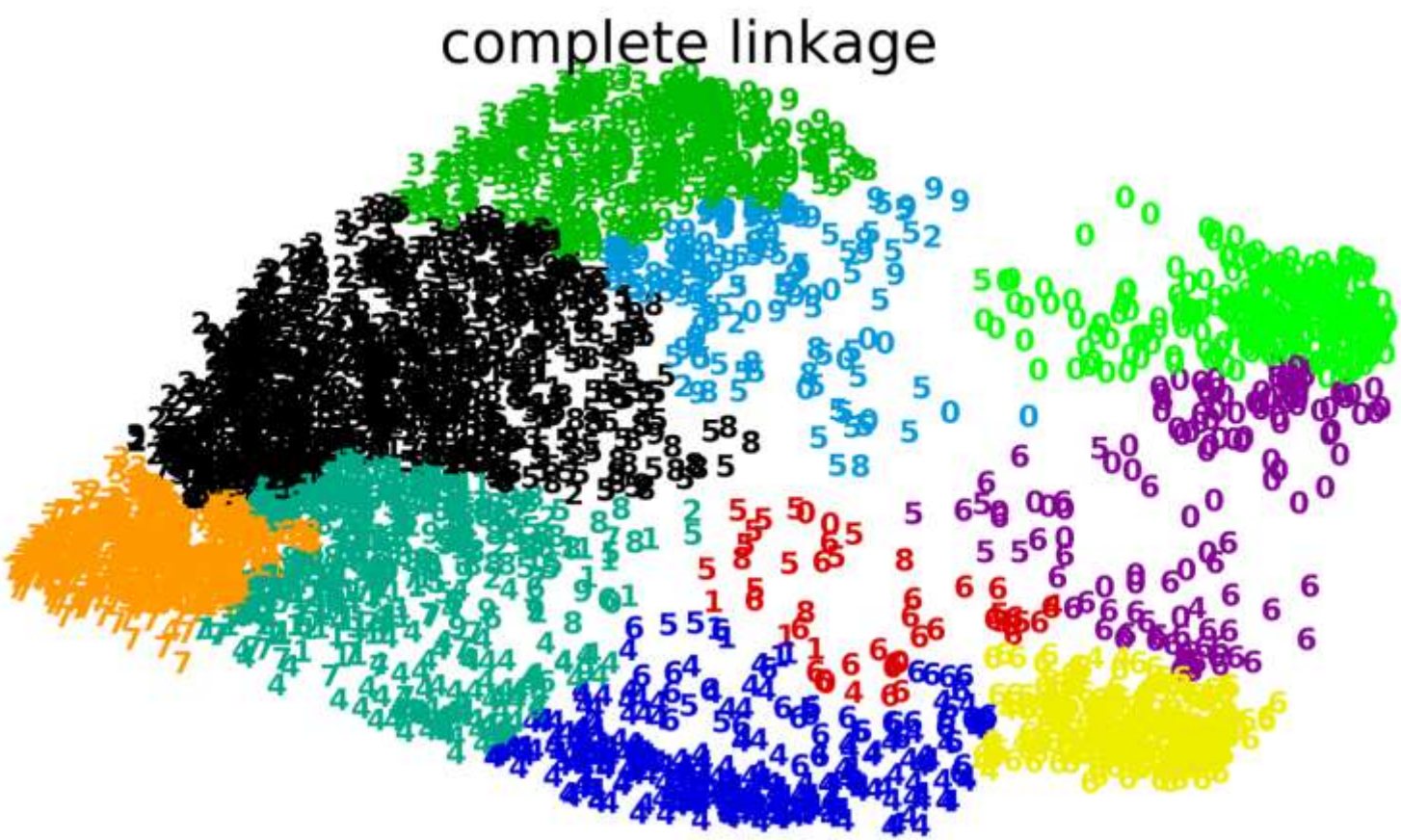


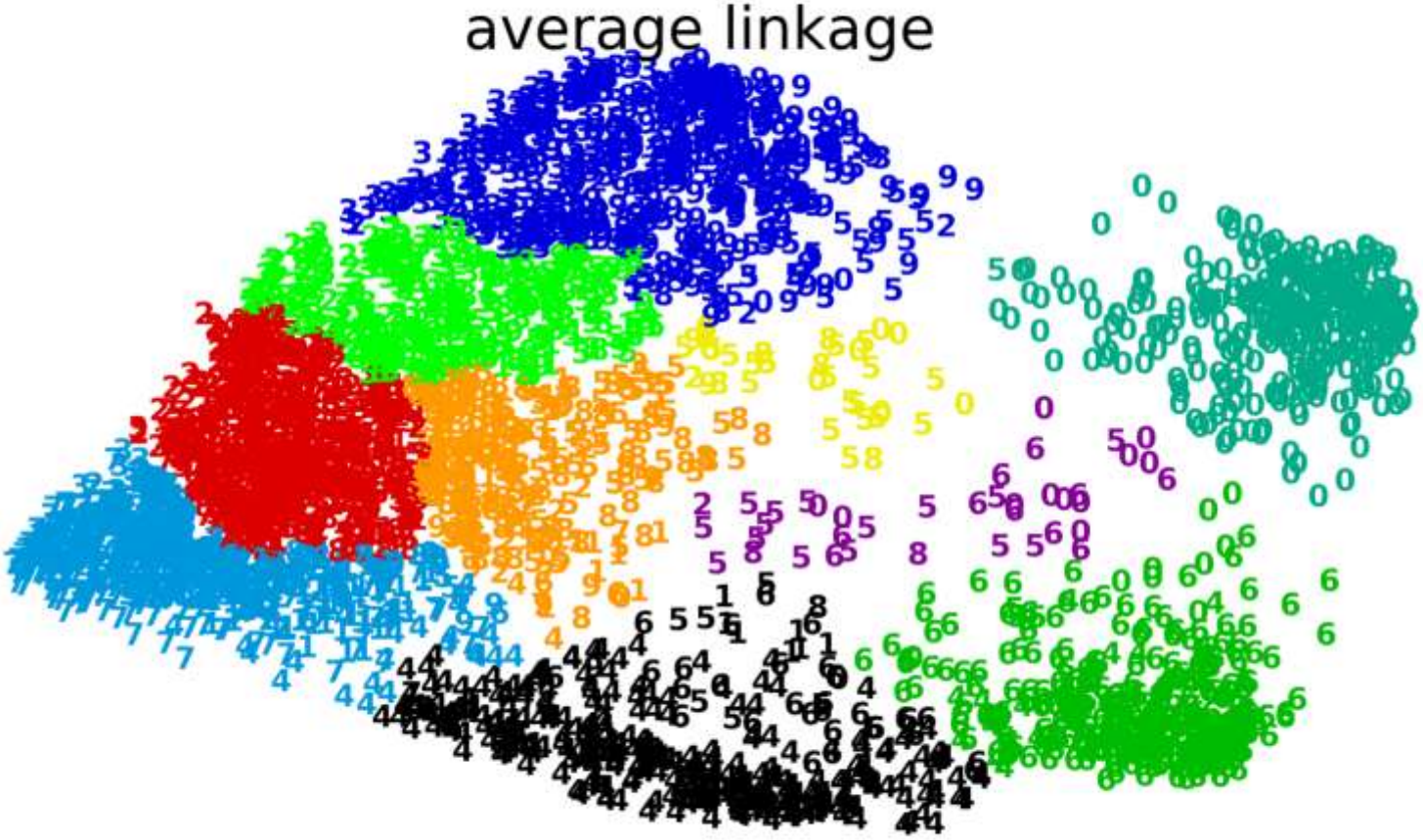
Proyección usando spectral embedding

1. Databases grandes pueden ser visualizadas con embeddings.
2. En este caso se usa spectral embedding para observar agrupamiento de imágenes de números manuscritos.
3. Las clases son balanceadas, sin embargo se observa la tendencia a armar clases desbalanceadas.
4. Single linkage tiene tendencia a hacer crecer desproporcionadamente un cluster.


ward linkage


single linkage


complete linkage


average linkage

Evaluación

Un experto de dominio **interpreta** los clusters y encuentra información valiosa

¿Cómo mostrar el contenido de los clusters?

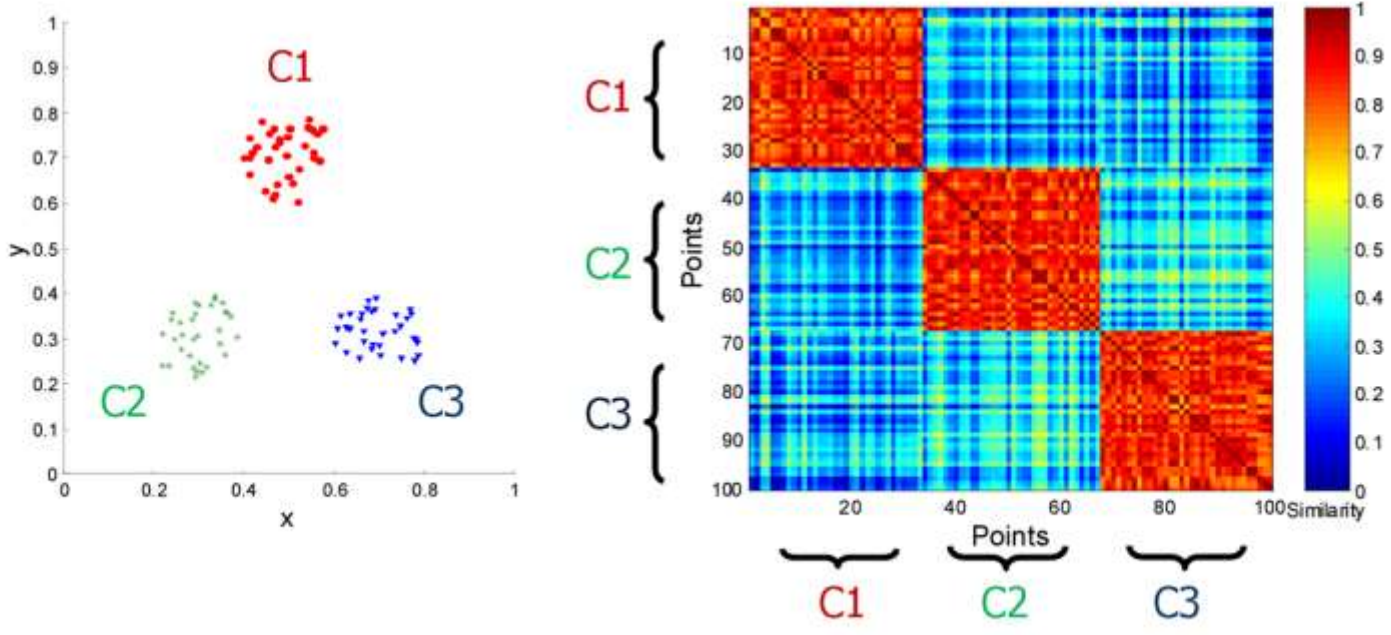
- Centroides (o medoides)
- Resumen de características
- Características más distintivas de cada cluster
- Aplicar un algoritmo de aprendizaje automático interpretable (Decision Tree)

Evaluación intrínseca

- Coeficiente Silhouette
  - Mide la semejanza de cada objeto al cluster al que se asigna (cohesión), comparada con otros clusters (separación).
  - Valores entre -1 y 1
  - Si el valor es bajo o negativo, el agrupamiento no es bueno compacto.
  - https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html
- Gráfico de codo (Elbow method)
  - Se observa la inercia en función del número de clusters.

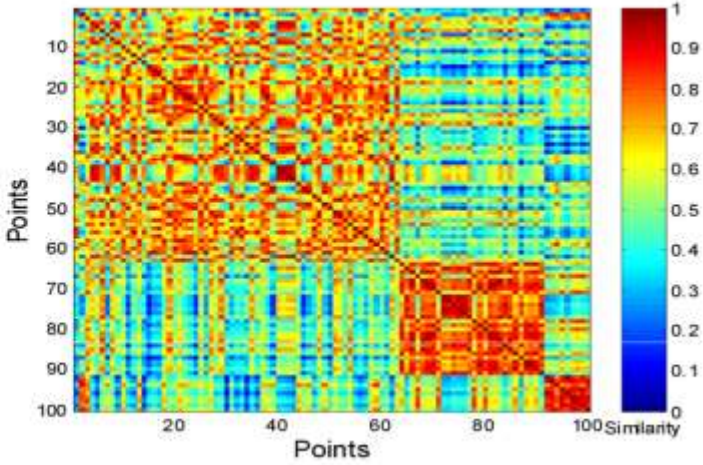
Matriz de similitud

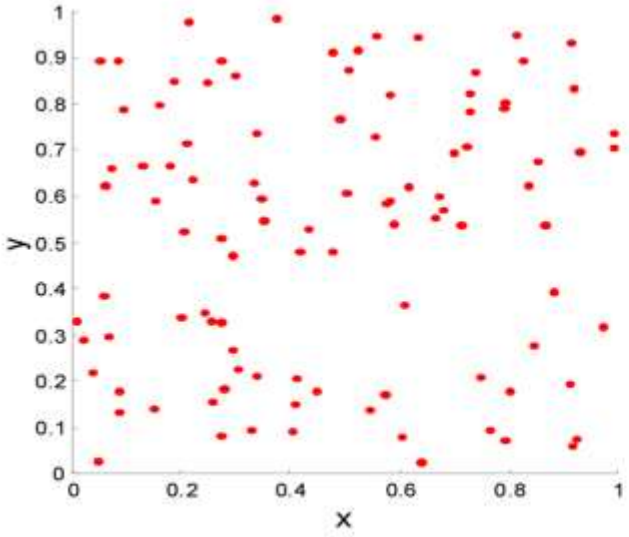
Ordenamos los datos en la matriz de similitud con respecto a los clusters en los que quedan los datos e inspeccionamos visualmente...

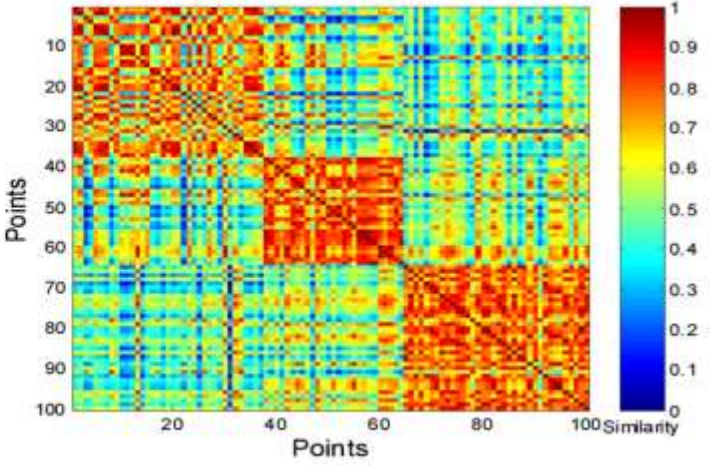


Problema

Incluso en datos aleatorios, si nos empeñamos, encontramos clusters: DBSCAN (arriba) y k-Means (abajo)

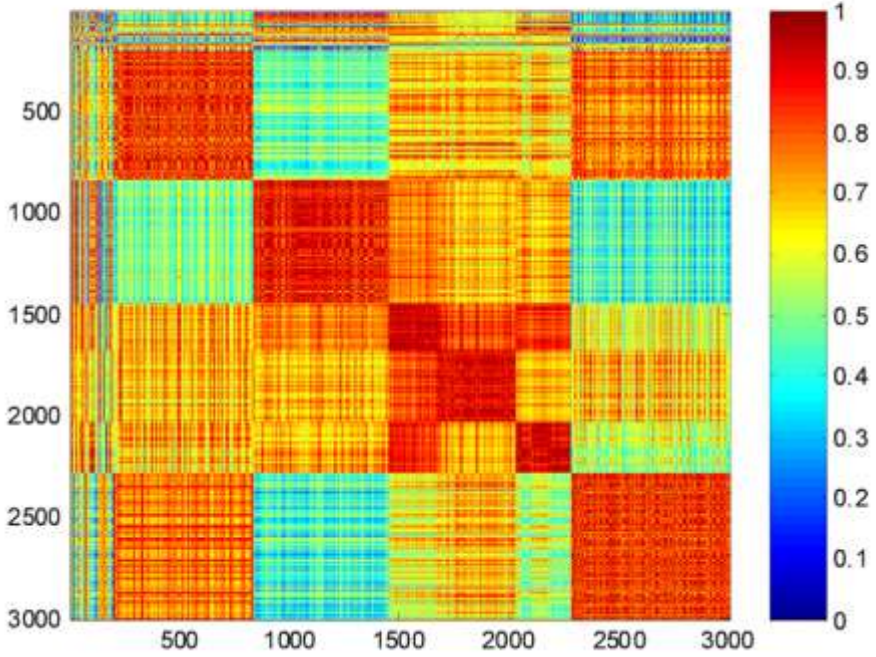


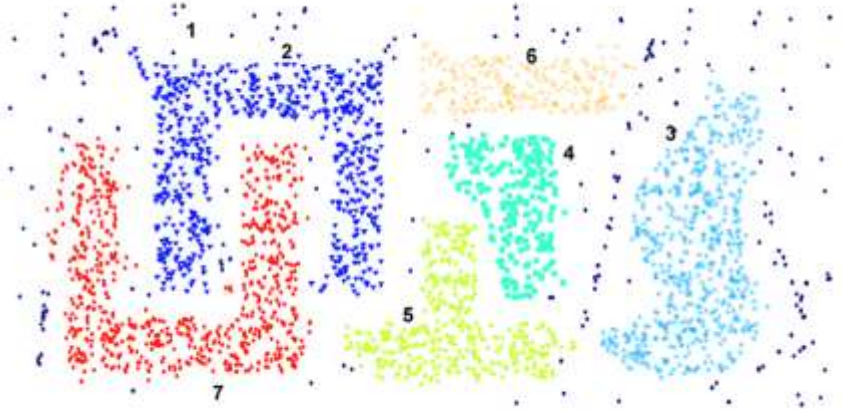




Matriz de similitud

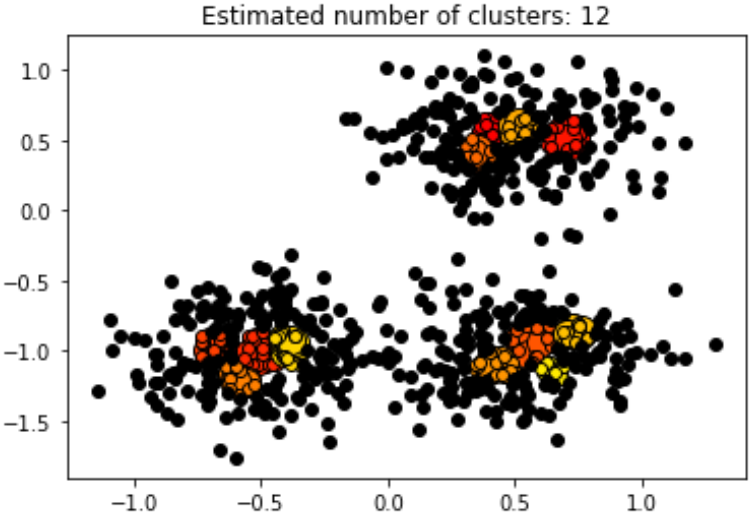
DBSCAN





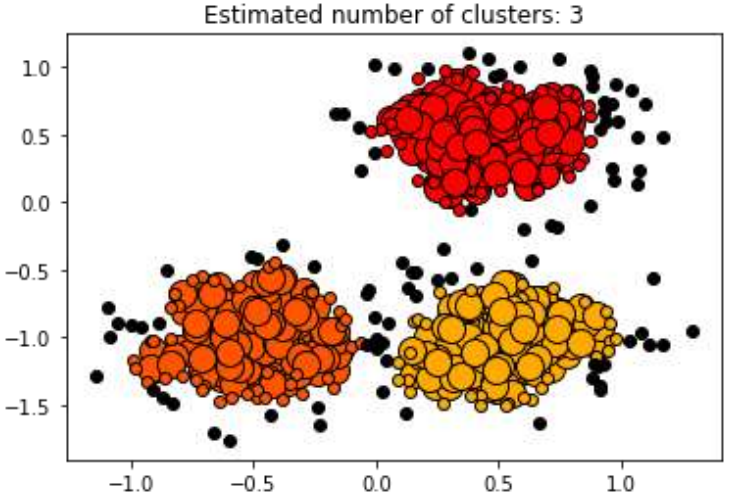
DBSCAN

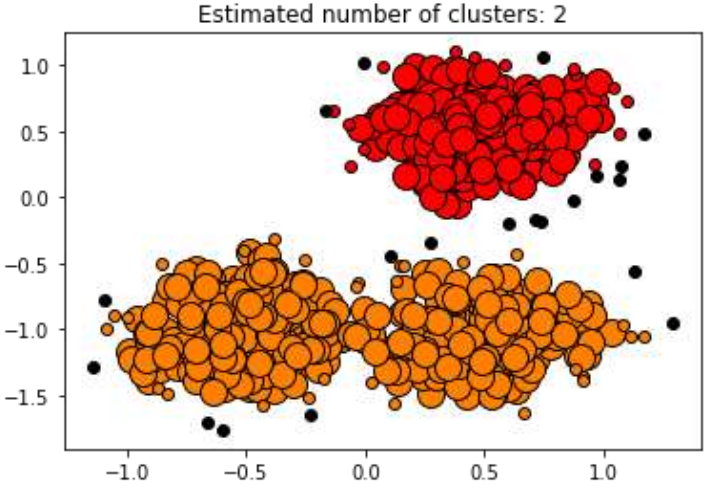
min_samples=10



eps=0.1

eps=0.2





eps=0.3

note_fig10.ipynb

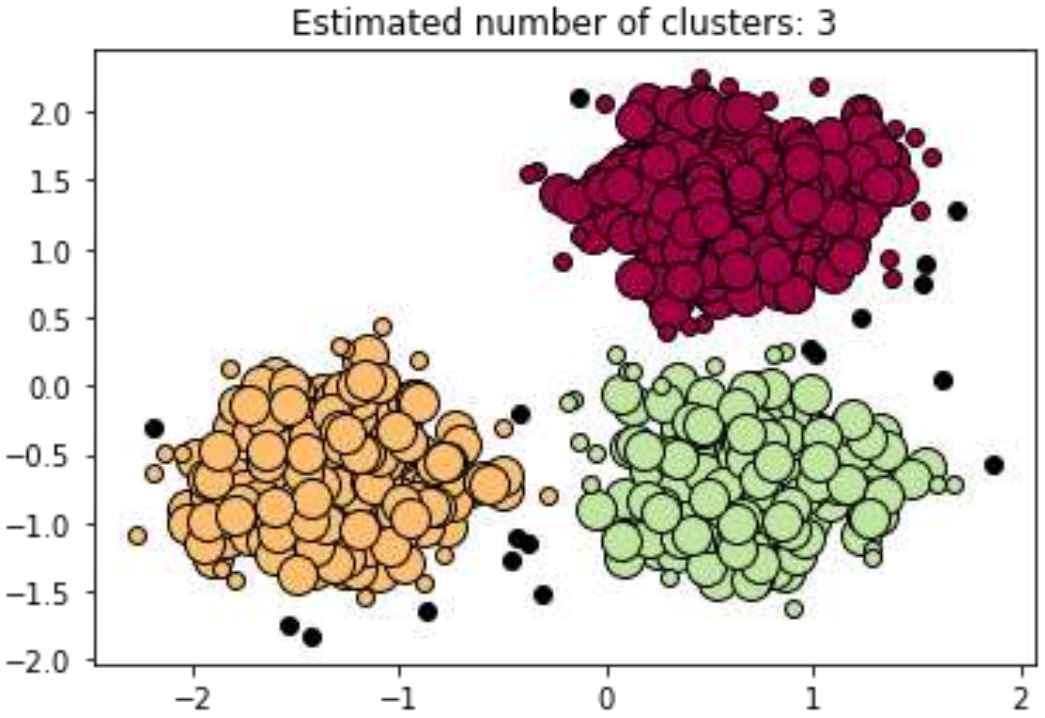
Evaluación con clases

Si los objetos tienen alguna etiqueta, observamos su distribución en los clusters, estamos cerca de el problema semi supervisado!!

- Homogeneidad: cada cluster contiene sólo miembros de una clase
- Completitud: todos los miembros de una clase están en el mismo cluster
- V-measure: media armónica de los anteriores
- Adjusted Rand index: semejanza entre las etiquetas originales y las asignadas
- Información Mutua entre etiquetas originales y asignadas

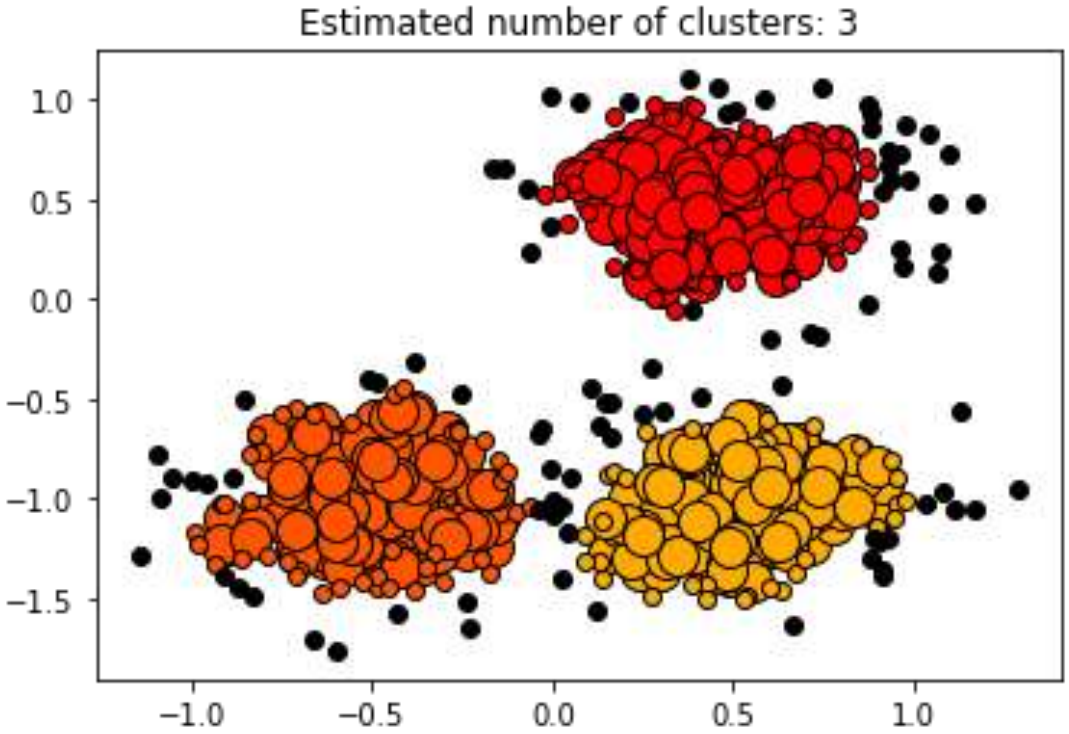
Evaluación con clases: note_fig10.ipynb

Estimated number of clusters: 3
Estimated number of noise points: 18
Homogeneity: 0.953
Completeness: 0.883
V-measure: 0.917
Adjusted Rand Index: 0.952
Adjusted Mutual Information: 0.916
Silhouette Coefficient: 0.626



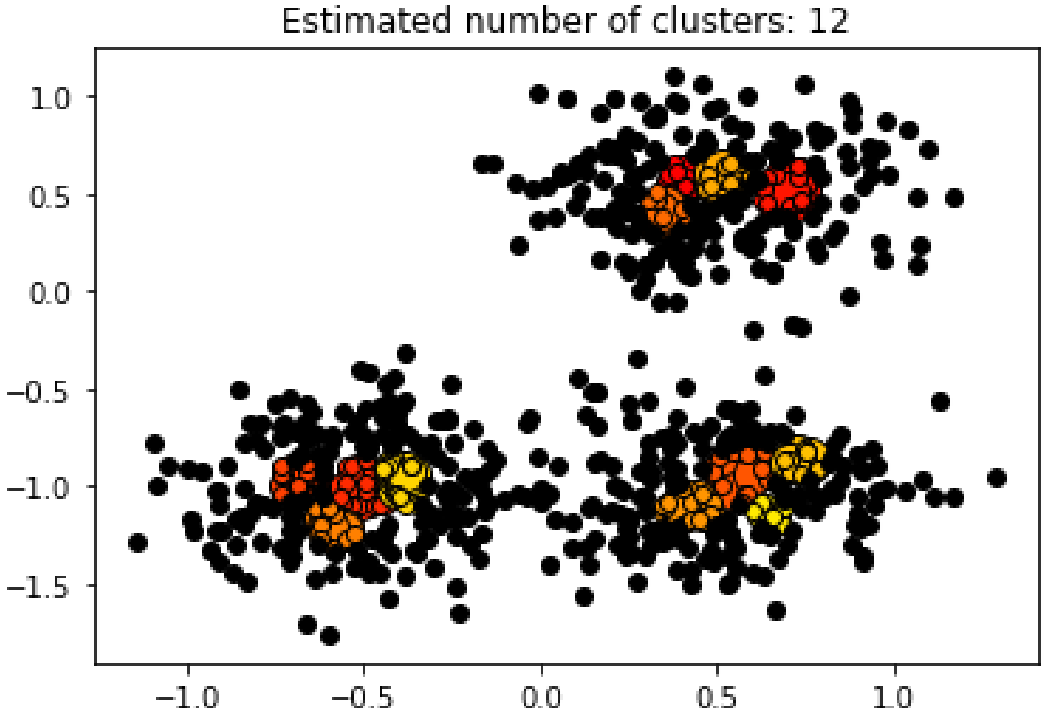
Evaluación con clases: note_fig10.ipynb

Estimated number of clusters: 3
Estimated number of noise points: 91
Homogeneity: 0.847
Completeness: 0.697
V-measure: 0.765
Adjusted Rand Index: 0.791
Adjusted Mutual Information: 0.764
Silhouette Coefficient: 0.532



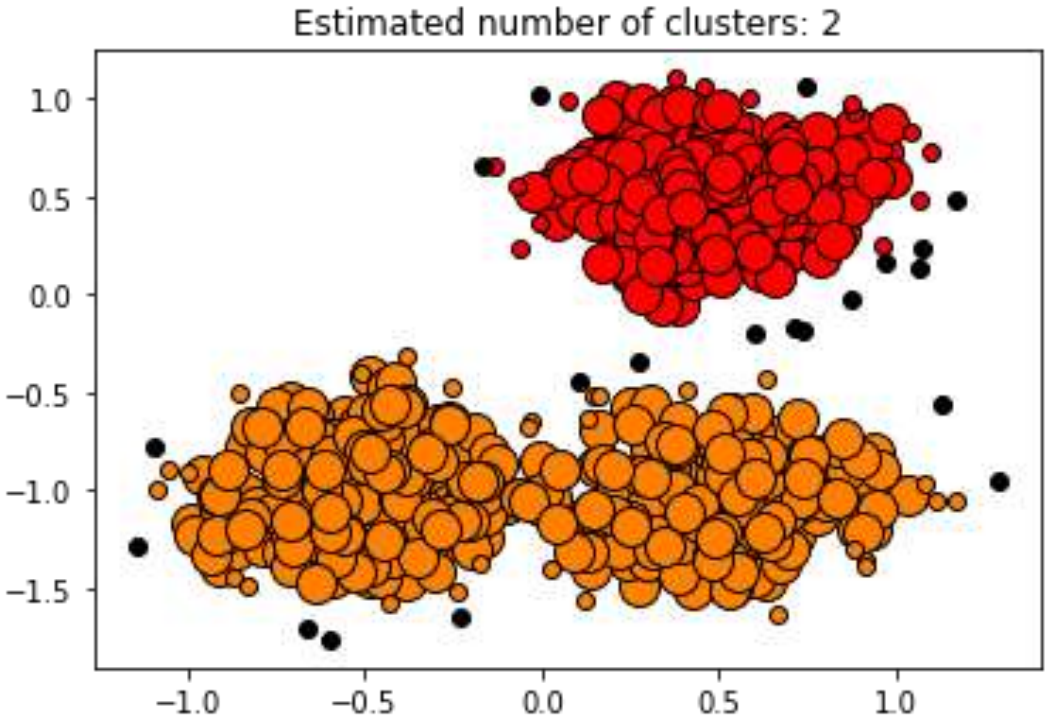
Evaluación con clases: note_fig10.ipynb

Estimated number of clusters: 12
Estimated number of noise points: 518
Homogeneity: 0.310
Completeness: 0.247
V-measure: 0.275
Adjusted Rand Index: 0.022
Adjusted Mutual Information: 0.265
Silhouette Coefficient: -0.365



Evaluación con clases: note_fig10.ipynb

Estimated number of clusters: 2
Estimated number of noise points: 20
Homogeneity: 0.563
Completeness: 0.837
V-measure: 0.673
Adjusted Rand Index: 0.545
Adjusted Mutual Information: 0.672
Silhouette Coefficient: 0.448



Evaluación con testigos

1. Se seleccionan aleatoriamente pares de objetos del dataset
2. Un experto del dominio decide si tienen que estar en el mismo cluster o en diferentes clusters
3. Observamos el grado de acuerdo entre cada solución y los testigos

1. Se seleccionan aleatoriamente objetos del dataset
2. Se los etiqueta
3. Se observa cómo se distribuyen en el dataset

Indicadores de malas soluciones

- Una clase muy grande y el resto mucho más chicas
- Clases con uno o pocos elementos
- Clusters con las mismas características, poco distinguibles
- Soluciones muy diferentes con diferentes inicializaciones

Malas soluciones pueden deberse a: “malas” características elegidas, número no adecuado de clases, no hay grupos para encontrar, ...

Clustering no es clasificación

No vamos a obtener clases bien diferenciadas, sino más bien mucho ruido

Es fuertemente sensible a las características de los objetos, a los parámetros, a los outliers

La mayor parte de aproximaciones son muy inestables

La primera aproximación suele ser inservible, hay que refinar características e iterar

Aplicaciones

- Segmentación de clientes, usuarios... para marketing personalizado
- Encontrar temas ➜ topic detection
- Segmentación de Imágenes
- Agrupamiento de productos
- Detección de anomalías
- Taxonomías de plantas y otros organismos
- Detección de clases con significados semejantes