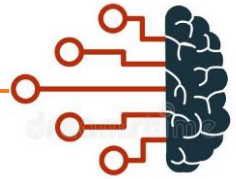
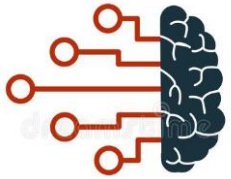


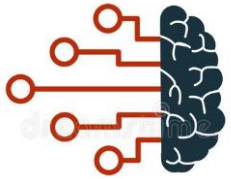
Aprendizaje No Supervisado

Contenido



1. Definiciones
2. Clustering
3. Tipos de Clustering

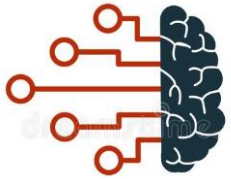
Aprendizaje no supervisado



Los datos de entrenamiento no están clasificados, ni organizados,

ni etiquetados

Clustering



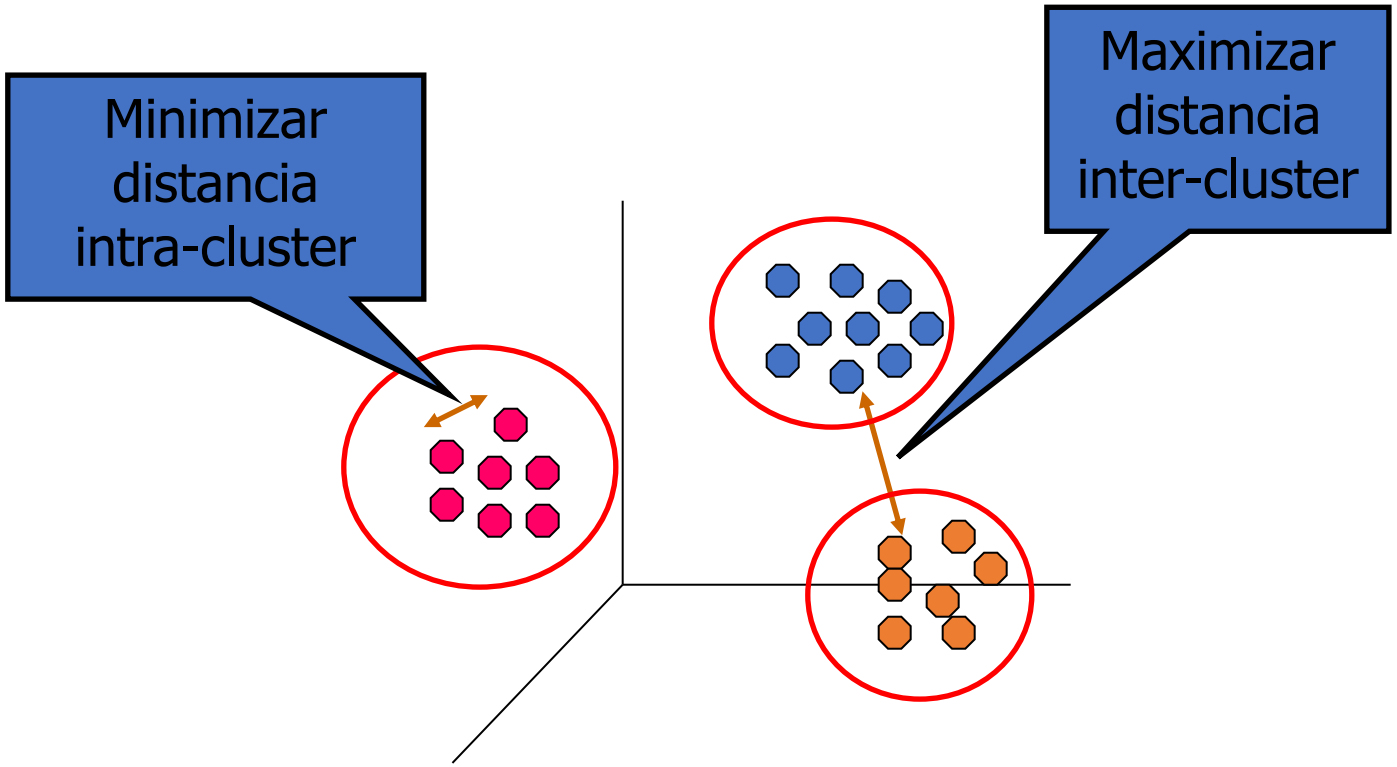
- **Objetivo,** Agrupar objetos similares entre sí que sean distintos a los objetos de otros agrupamientos [clusters].

- **Aprendizaje no supervisado,** No existen clases predefinidas

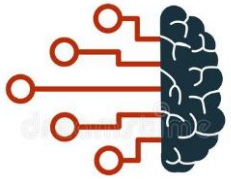
- Los resultados obtenidos dependerán de:
  - El algoritmo de agrupamiento seleccionado.
  - El conjunto de datos disponible
  - La medida de similitud utilizada para comparar objetos.

Clustering

Encontrar agrupamientos de tal forma que los objetos de un grupo sean similares entre sí y diferentes de los objetos de otros grupos:



Clustering - Aplicaciones



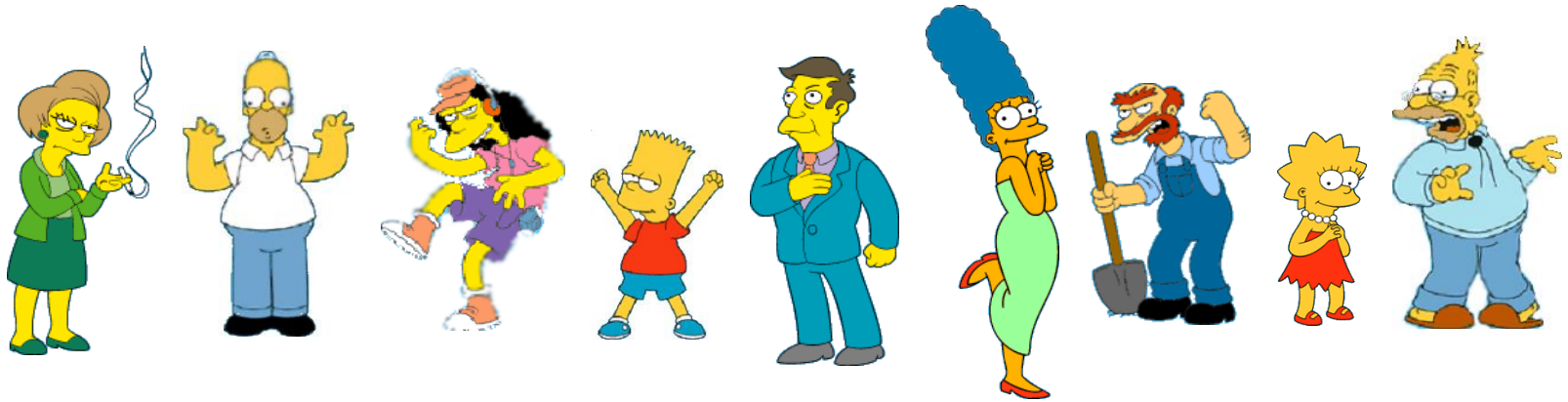
- Reconocimiento de formas.
- Mapas temáticos (GIS)
- Marketing: Segmentación de clientes
- Clasificación de documentos
- Análisis de web logs (patrones de acceso similares)
- ...

Aplicaciones típicas en Data Mining:

- Exploración de datos (segmentación & outliers)
- Preprocesamiento (p.ej. reducción de datos)

Clustering

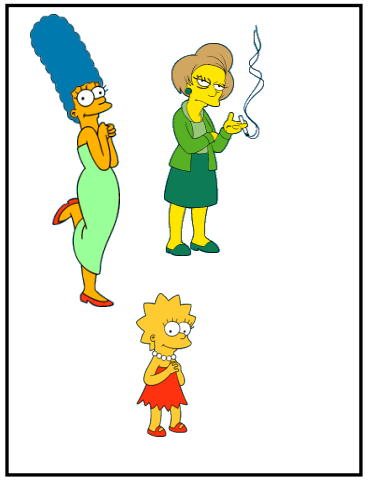
¿Cuál es la forma natural de agrupar los personajes?

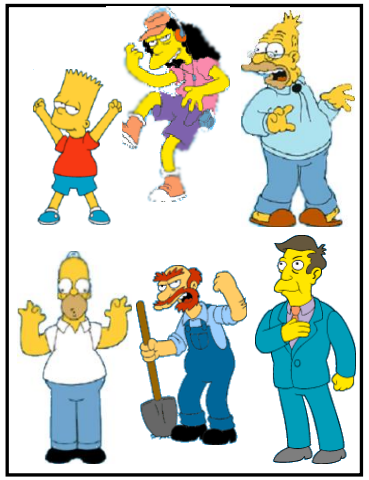


Hombres

vs.

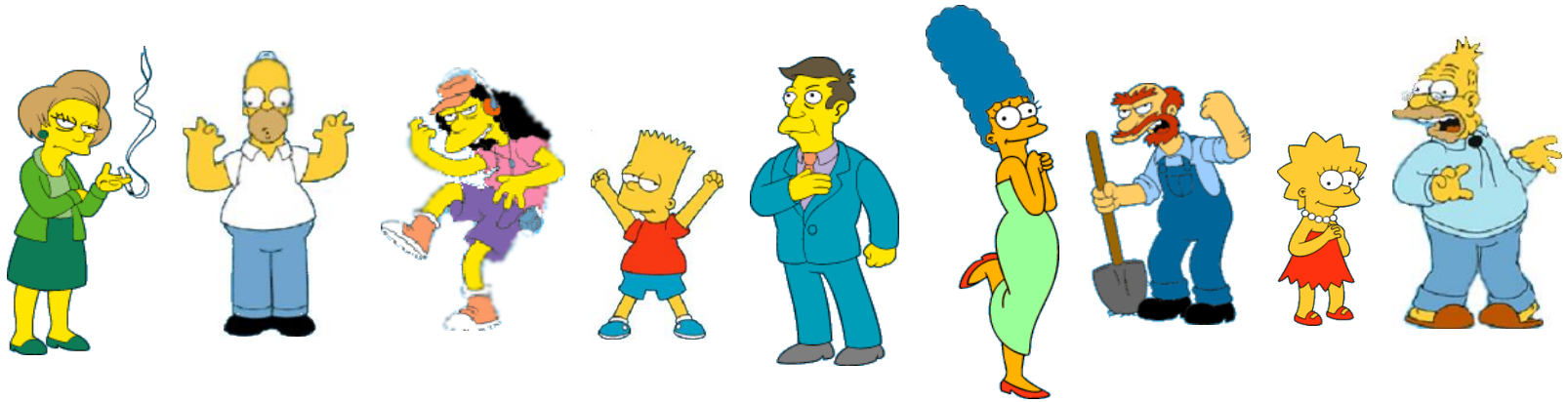
Mujeres



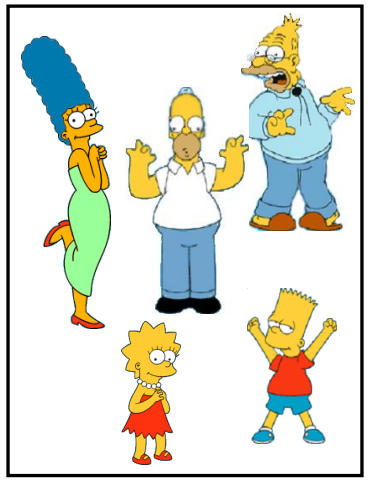


Clustering

¿Cuál es la forma natural de agrupar los personajes?

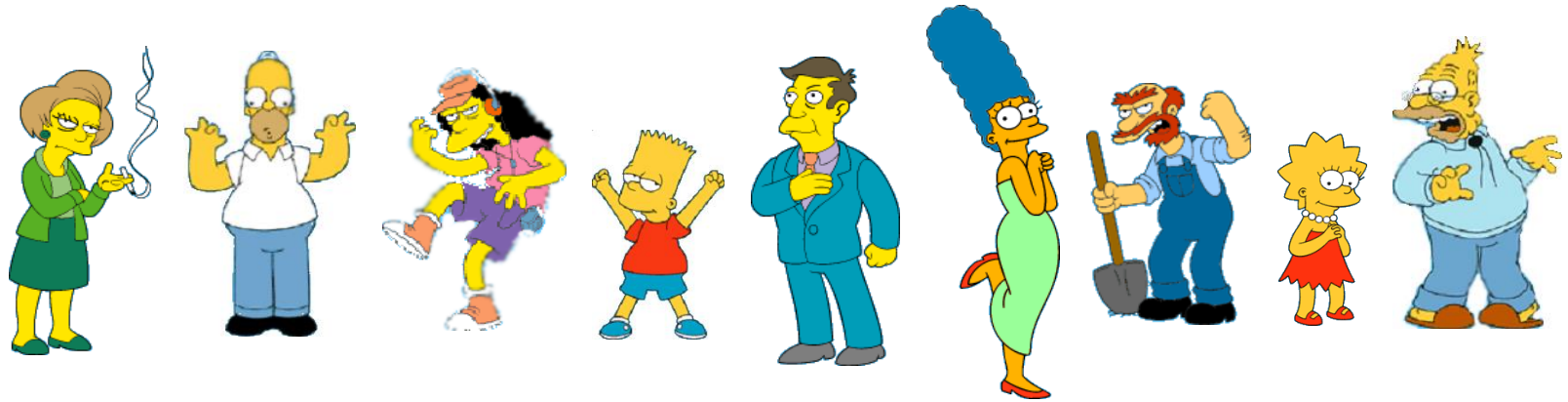


Simpsons
vs.
Empleados
de la escuela
de Springfield



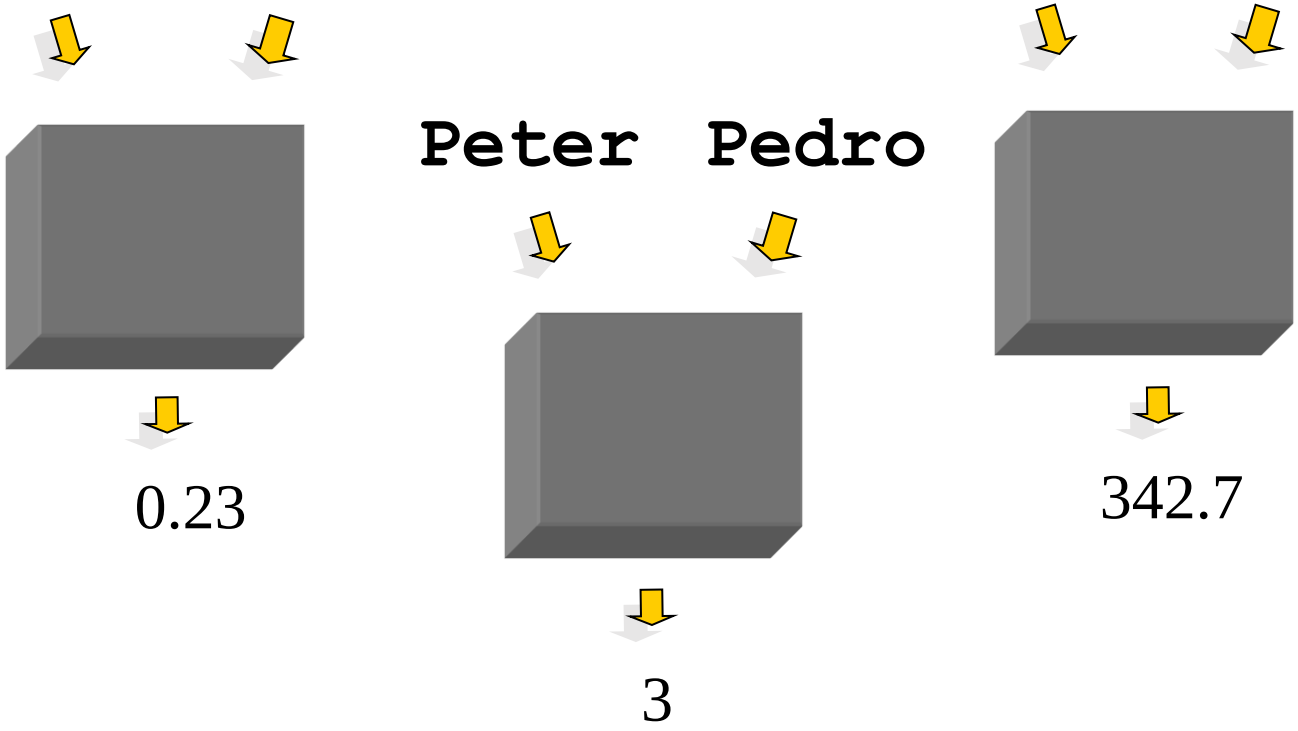
Clustering

¿Cuál es la forma natural de agrupar los personajes?

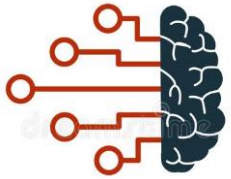


¡¡¡ El clustering es subjetivo !!!

Medidas de similitud



Medidas de similitud



Usualmente, se expresan en términos de distancias:

d(i,j) > d(i,k)

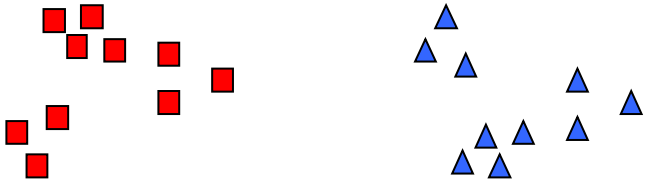
nos indica que el objeto i es más parecido a k que a j

La definición de la métrica de similitud/distancia será distinta en función del tipo de dato y de la interpretación semántica que nosotros hagamos.

En otras palabras, la similitud entre objetos es **subjetiva**.

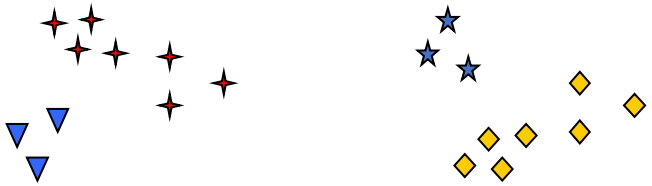
Medidas de similitud

¿Cuántos agrupamientos?



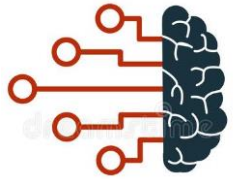
¿Dos?

¿Seis?



¿Cuatro?

Medidas de similitud - Atributos contínuos



Usualmente, se "estandarizan" a priori:

- Desviación absoluta media:

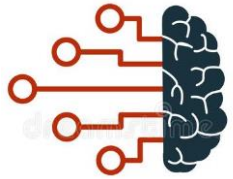
$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|)$$

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf}).$$

- z-score (medida estandarizada):

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

Medidas de similitud - Métricas de distancia



Distancia de Minkowski

$$d_r(x,y) = \left(\sum_{j=1}^{J} |x_j - y_j|^r\right)^{\frac{1}{r}}, \quad r \geq 1$$

Distancia de Manhattan (r=1) / city block / taxicab

$$d_1(x,y) = \sum_{j=1}^{J} |x_j - y_j|$$

Distancia euclídea (r=2):

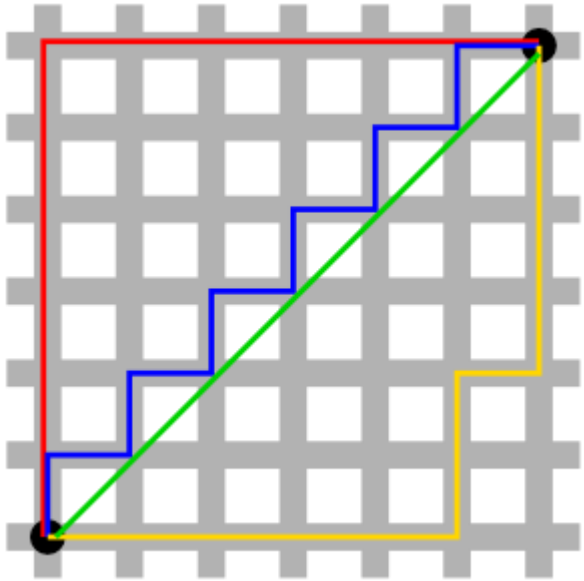
$$d_2(x,y) = \sqrt{\sum_{j=1}^{J} (x_j - y_j)^2}$$

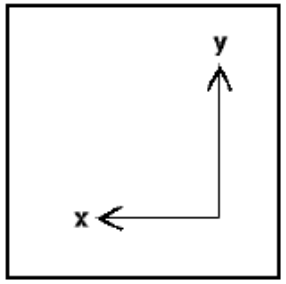
Distancia de Chebyshev (r→∞) / dominio / chessboard

$$d_\infty(x,y) = \text{máx}_{j=1..J} |x_j - y_j|$$

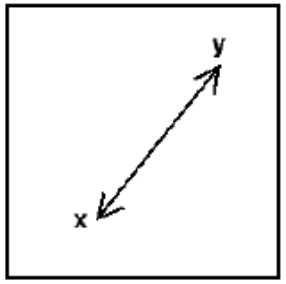
Medidas de similitud -Métricas de distancia

Distancia de Minkowski





Manhattan



Euclidean

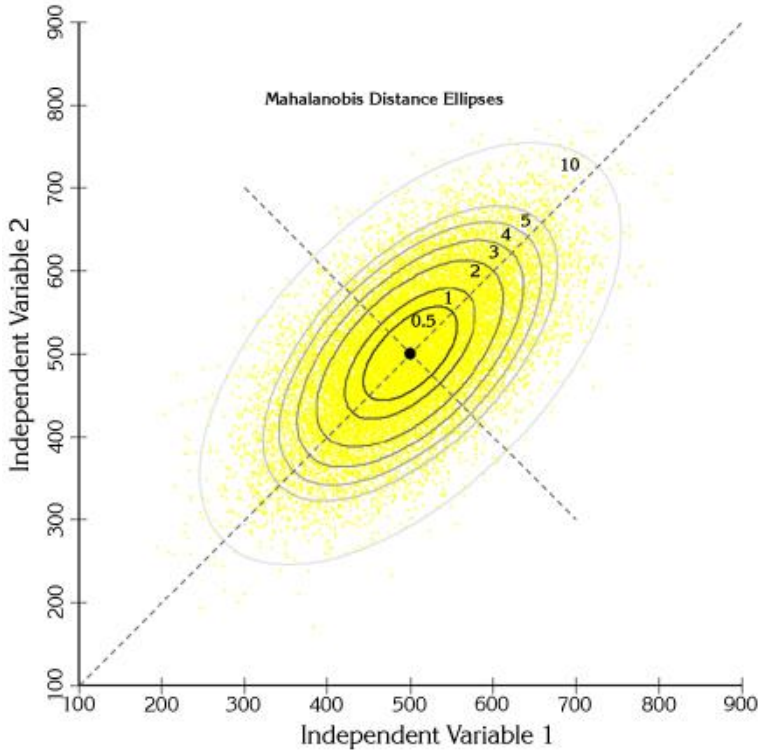
- Distancia de Manhattan = 12
- Distancia Euclídea ≅ 8.5
- Distancia de Chebyshev = 6

Medidas de similitud - Métricas de distancia

Distancia de Mahalanobis

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \Sigma^{-1} (\vec{x} - \vec{y})}.$$

- Considera las correlaciones entre variables.
- No depende de la escala de medida.



Medidas de similitud- Métricas de distancia

Distancia de edición = Distancia de Levenshtein

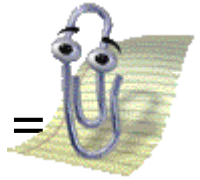
Número de operaciones necesario
para transformar una cadena en otra.

d("data mining", "data minino") = 1

d("efecto", "defecto") = 1

d("poda", "boda") = 1

d("night","natch") = d("natch","noche") =

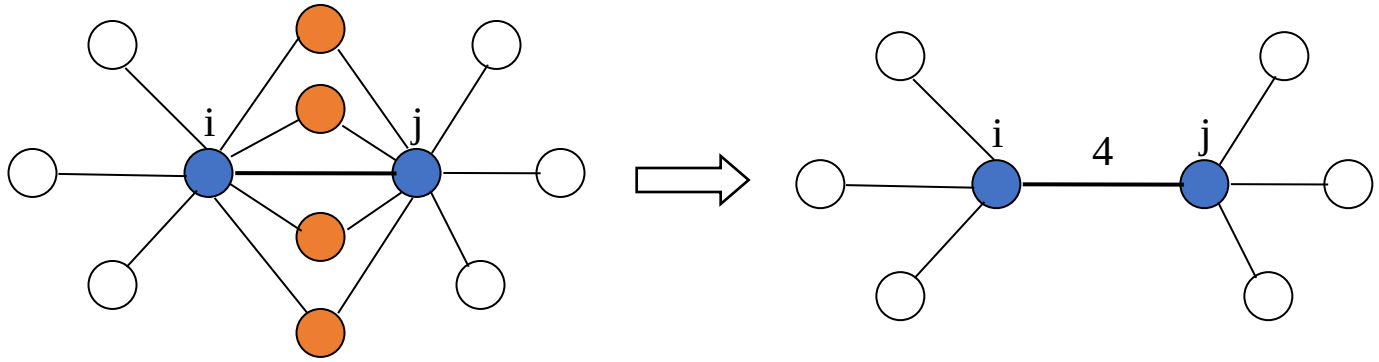


Aplicaciones: Correctores ortográficos, reconocimiento de voz, detección de plagios, análisis de ADN...

Para datos binarios: Distancia de Hamming

Medidas de similitud- Métricas de distancia

Vecinos compartidos



- "Mutual Neighbor Distance"

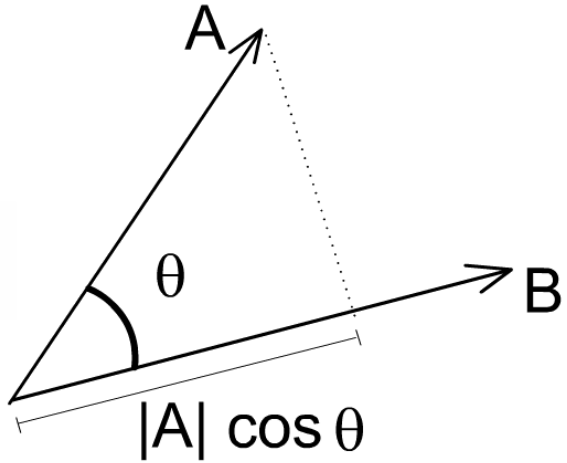
$$MND(\mathbf{x}_i, \mathbf{x}_j) = NN(\mathbf{x}_i, \mathbf{x}_j) + NN(\mathbf{x}_j, \mathbf{x}_i),$$

donde $NN(x_i,x_j)$ es el número de vecino de x_j con respecto a x_i

Medidas de similitud - Medidas de correlación

Producto escalar

$$S.(x, y) = x \cdot y = \sum_{j=1}^{J} x_j y_j$$



- "Cosine similarity"

$$\cos(\vec{x}, \vec{y}) = \sum_i \frac{x_i \cdot y_i}{\sqrt{\sum_i x_i^2} \cdot \sqrt{\sum_i y_i^2}}$$

- Coeficiente de Tanimoto

$$s(\vec{X}, \vec{Y}) = \frac{\vec{X}^t \cdot \vec{Y}}{\vec{X}^t \cdot \vec{X} + \vec{Y}^t \cdot \vec{Y} - \vec{X}^t \cdot \vec{Y}},$$

Medidas de similitud

Modelos basados en Teoría de Conjuntos

Modelo proporcional

$$s(a,b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha f(A - B) + \beta f(B - A)}$$

donde $\alpha, \beta \geq 0$

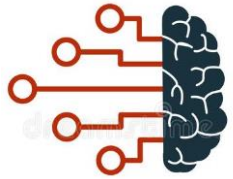
Modelo de Gregson = Coeficiente de Jaccard

$$S_{Gregson}(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Distancia de Tanimoto

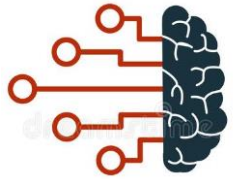
$$T(S_1,S_2) = \frac{|S_1| + |S_2| - 2|S_1 \cap S_2|}{|S_1| + |S_2| - |S_1 \cap S_2|}$$

Requisitos del algoritmo “perfecto”



- Escalabilidad
- Manejo de distintos tipos de datos
- Identificación de clusters con formas arbitrarias
- Número mínimo de parámetros
- Tolerancia frente a ruido y outliers
- Independencia con respecto al orden de presentación de los patrones de entrenamiento
- Posibilidad de trabajar en espacios con muchas dimensiones diferentes
- Capacidad de incorporar restricciones especificadas por el usuario (“domain knowledge”)
- Interpretabilidad / Usabilidad

Tipos de algoritmos de clustering



- Agrupamiento por particiones

  k-Means, CLARANS

- Clustering jerárquico

  BIRCH, ROCK, CHAMELEON

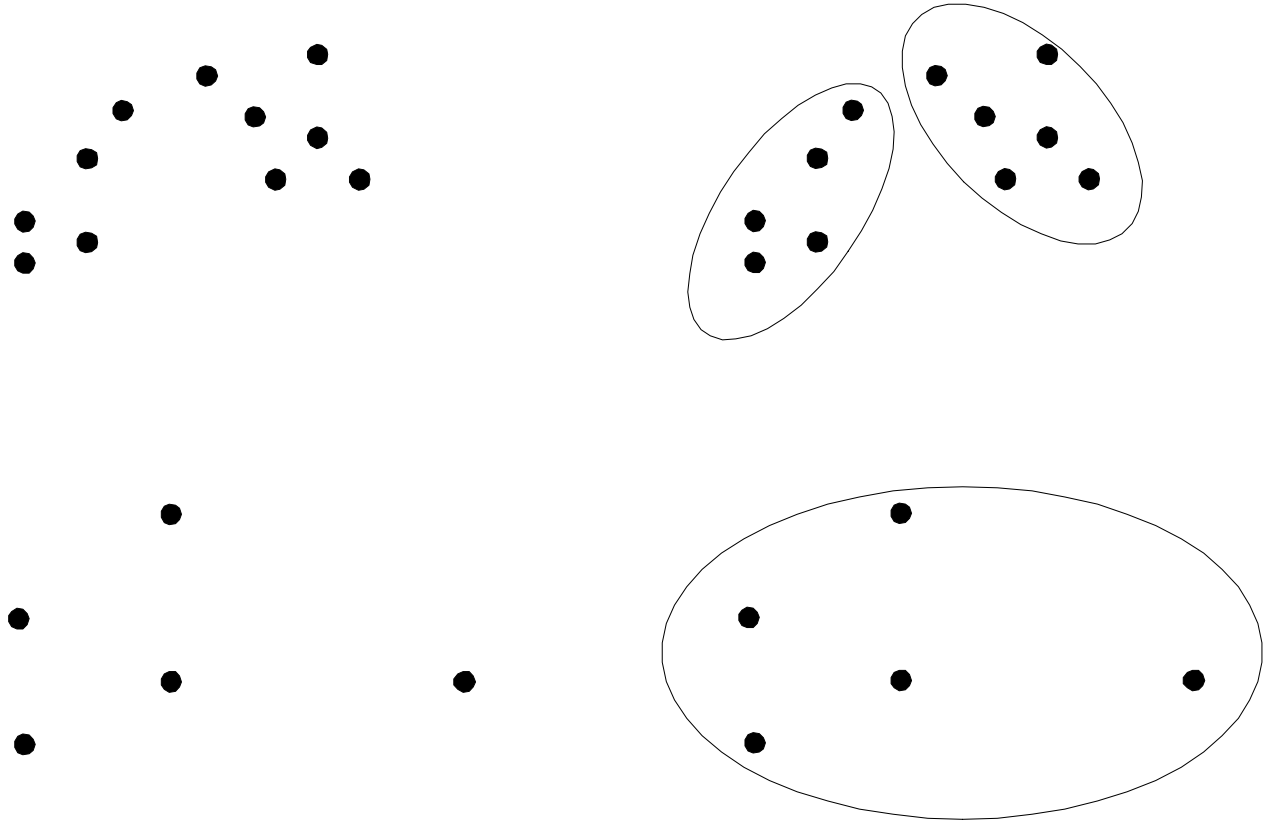
- Métodos basados en densidad

  DBSCAN

- ...

Métodos de agrupamiento

Clustering por particiones

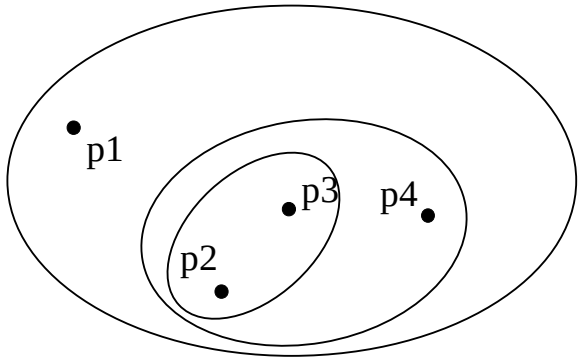


Datos originales

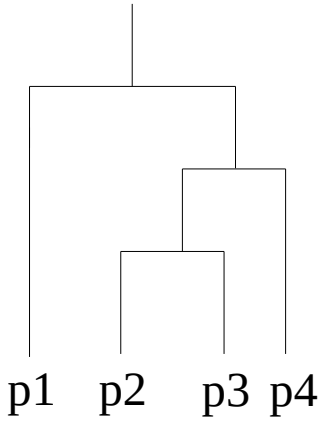
Datos agrupados

Métodos de agrupamiento

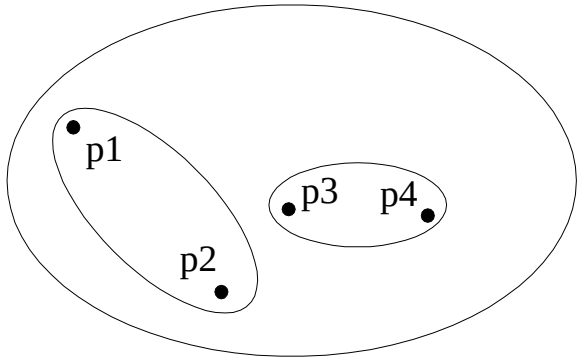
Clustering jerárquico



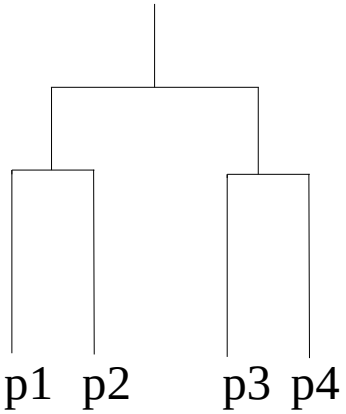
Tradicional



DENDOGRAMA

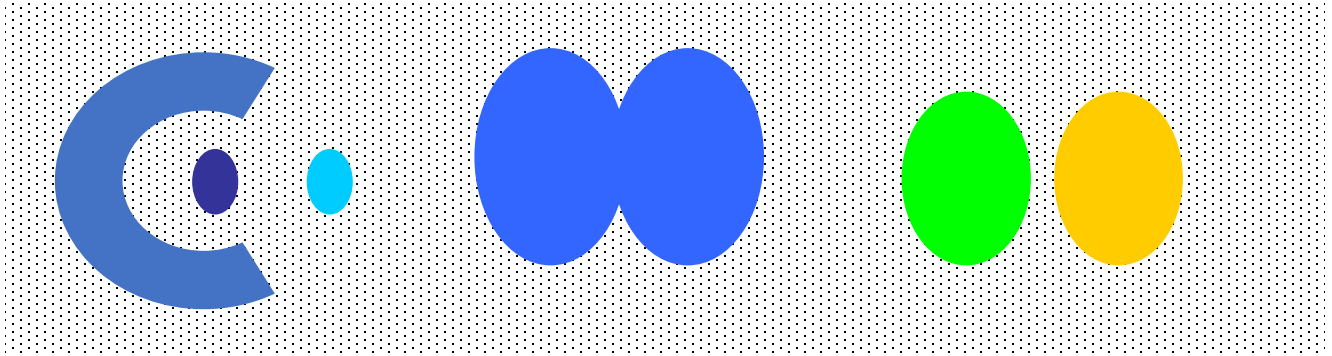


No tradicional

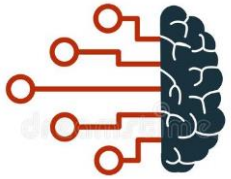


Métodos basados en densidad

- Un cluster en una región densa de puntos, separada por regiones poco densas de otras regiones densas.
- Útiles cuando los clusters tienen formas irregulares, están entrelazados o hay ruido/outliers en los datos.

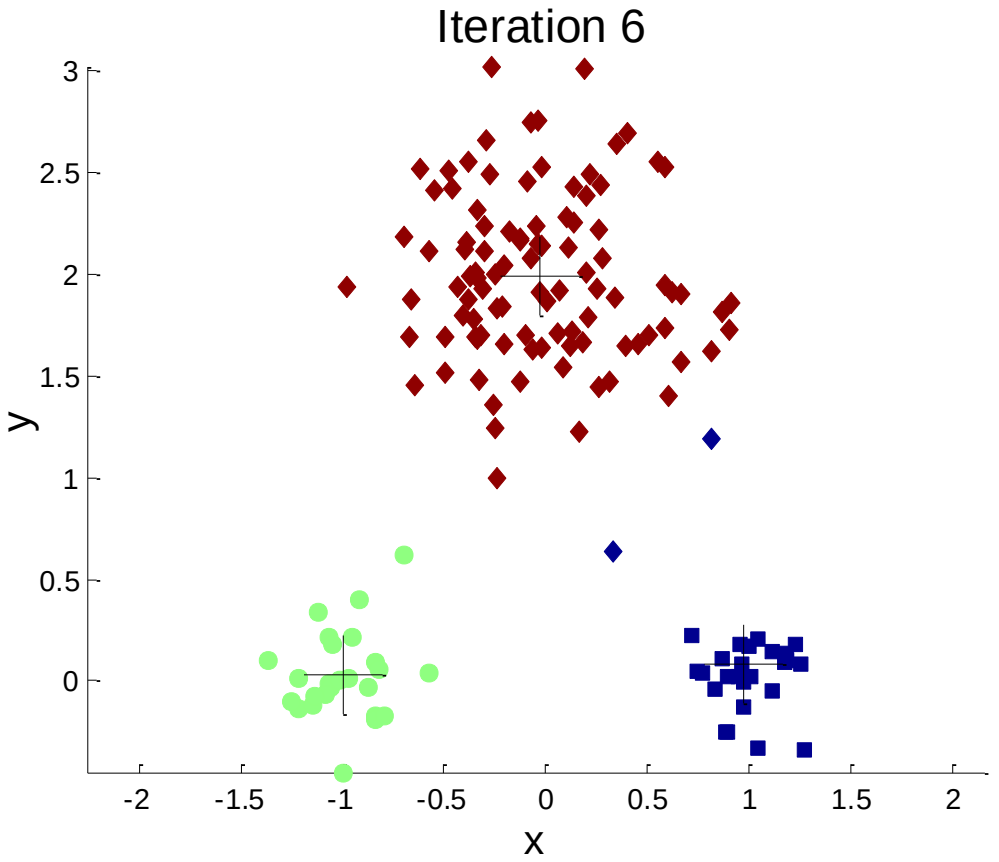


k-Means Algoritmo (MacQueen, 1967)

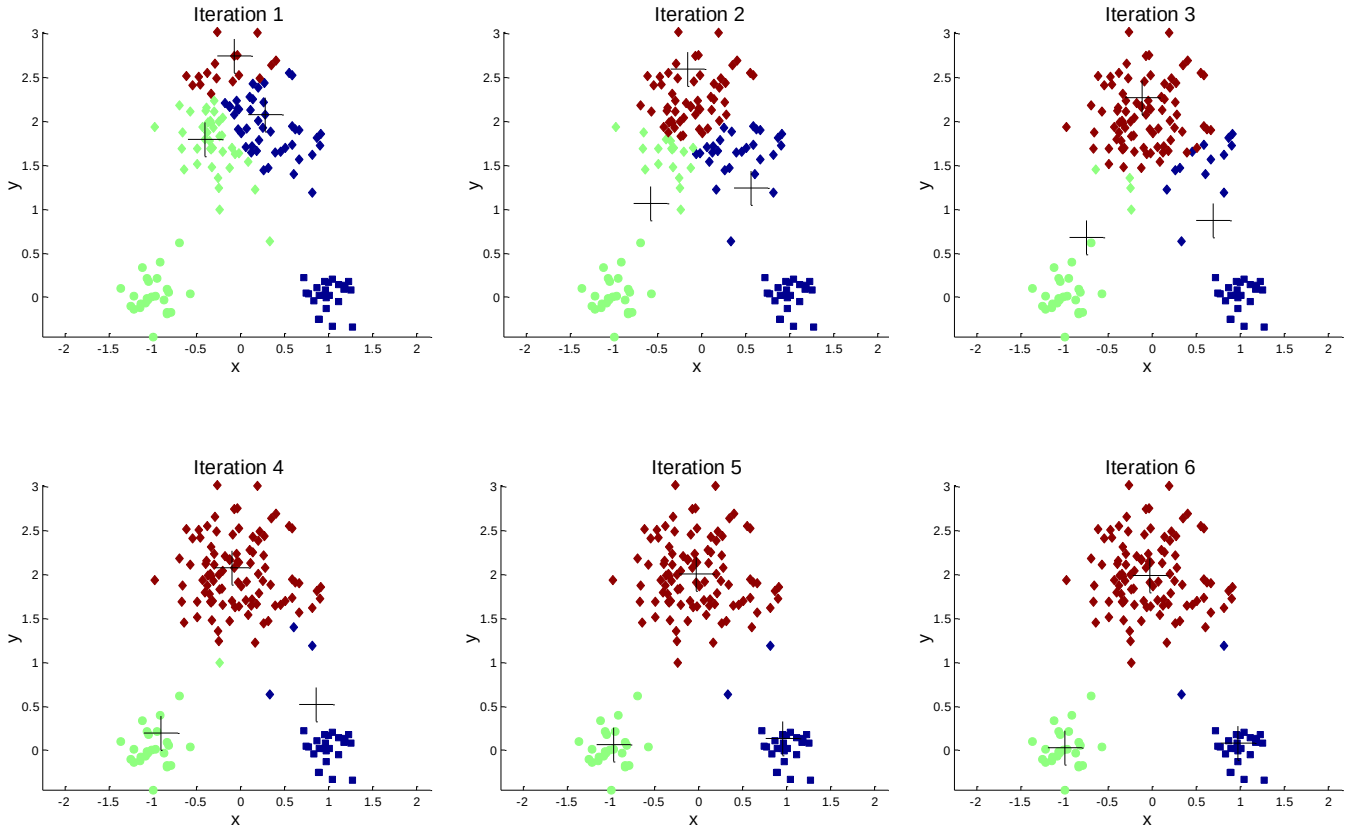


- Número de clusters conocido (**k**)
- Cada cluster tiene asociado un centroide (centro geométrico del cluster).
- Los puntos se asignan al cluster cuyo centroide esté más cerca (utilizando cualquier métrica de distancia).
- Iterativamente, se van actualizando los centroides en función de las asignaciones de puntos a clusters, hasta que los centroides dejen de cambiar.
- Complejidad **O(n*k*I*d)**
  donde n es el número de datos, k el número de clusters, I el número de iteraciones y d el número de atributos

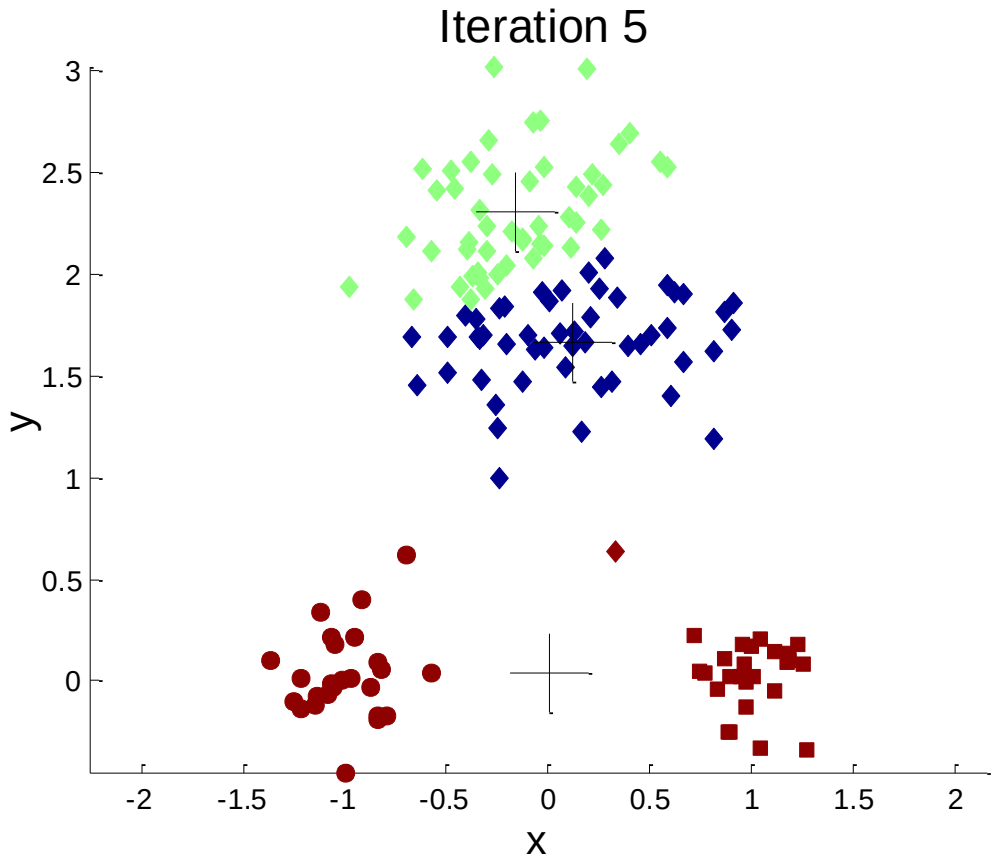
k-Means



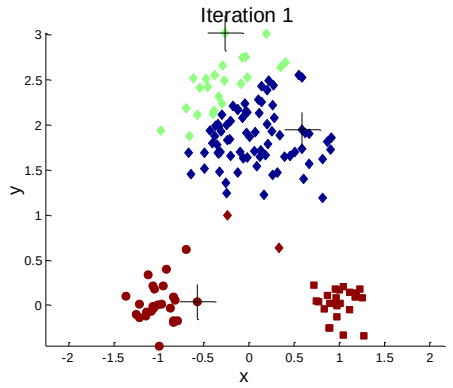
k-Means

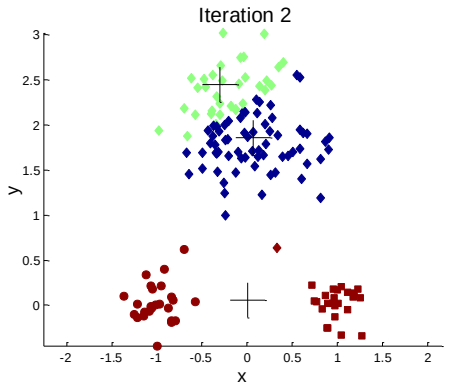

Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5
Iteration 6
x
y

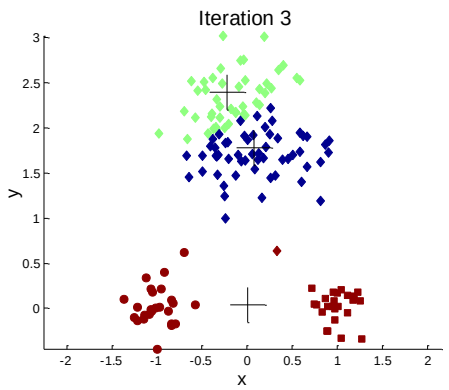
k-Means

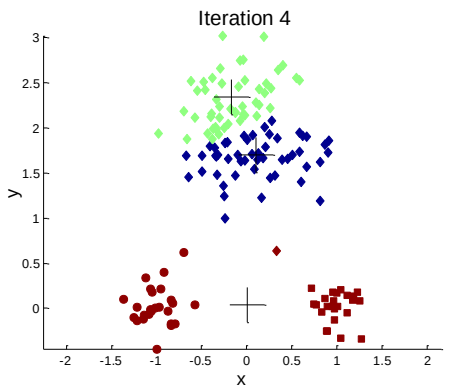


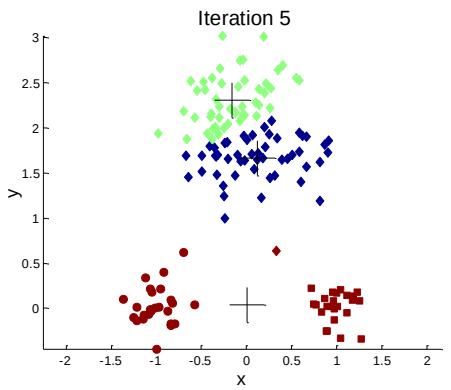
k-Means


Iteration 1
x
y

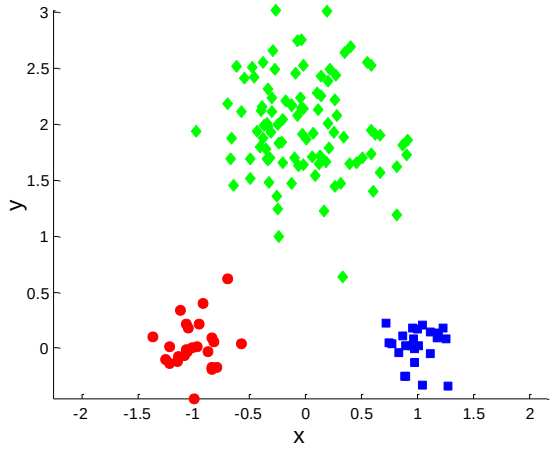

Iteration 2
x
y


Iteration 3
x
y

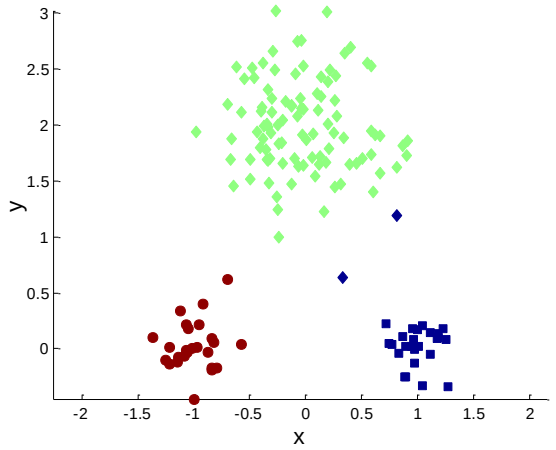

Iteration 4
x
y


Iteration 5
x
y

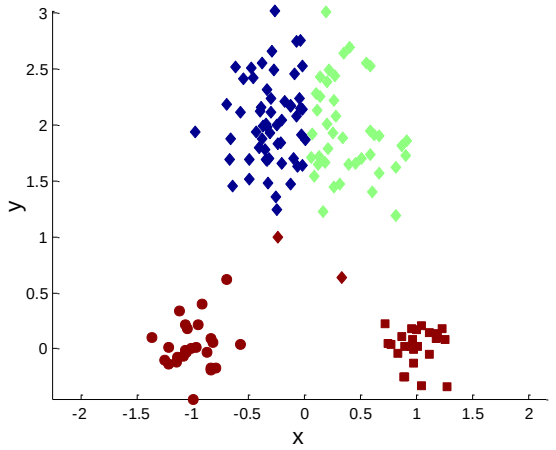
k-Means



Puntos originales

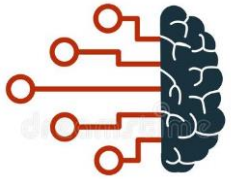


Solución óptima



Óptimo local

k-Means



Ventaja

- Eficiencia **O(n·k·I·d)**

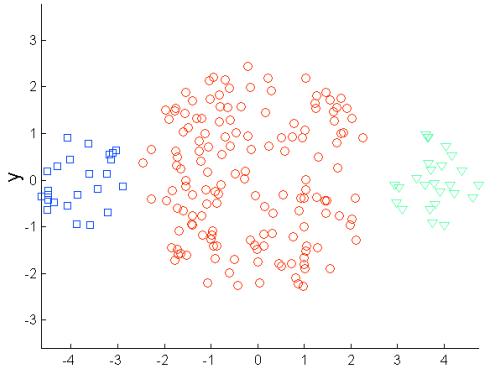
vs. PAM $O(I \cdot k(n-k)^2)$

CLARA $O(ks^2+k(n-k))$

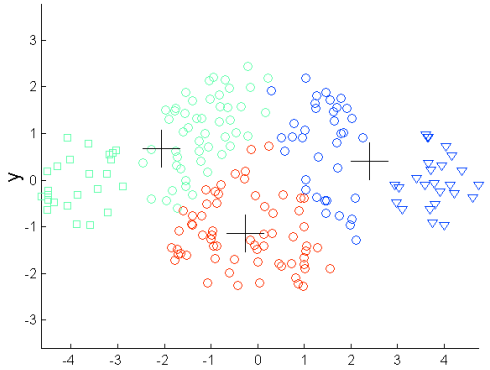
Desventajas

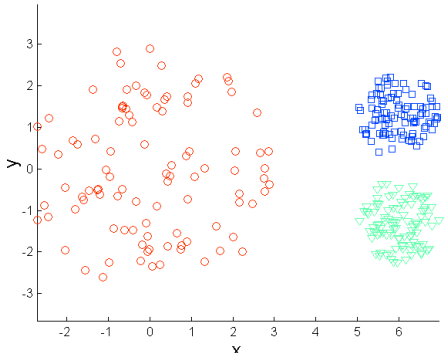
- Termina en un óptimo local:
  El resultado depende de la selección inicial de centroides.
- Necesidad de conocer el número de agrupamientos k
- Incapacidad para detectar ruido / identificar outliers.
- No resulta adecuado para detectar clusters no convexos
- Si tenemos datos de tipo categórico,
  ¿cómo calculamos la media?

k-Means

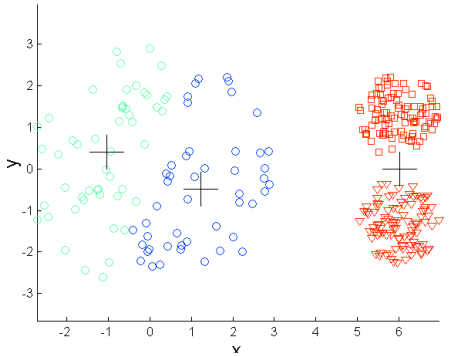


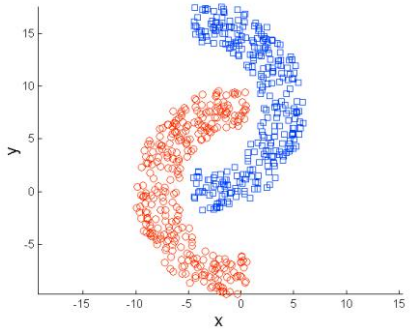
Clusters de distinto tamaño



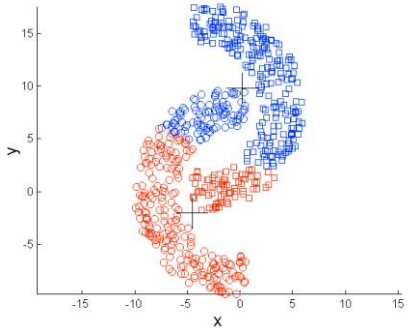


Clusters de distinta densidad

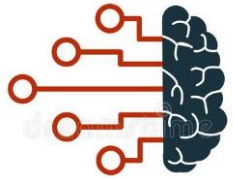




Clusters no convexos



k-Means -Variantes



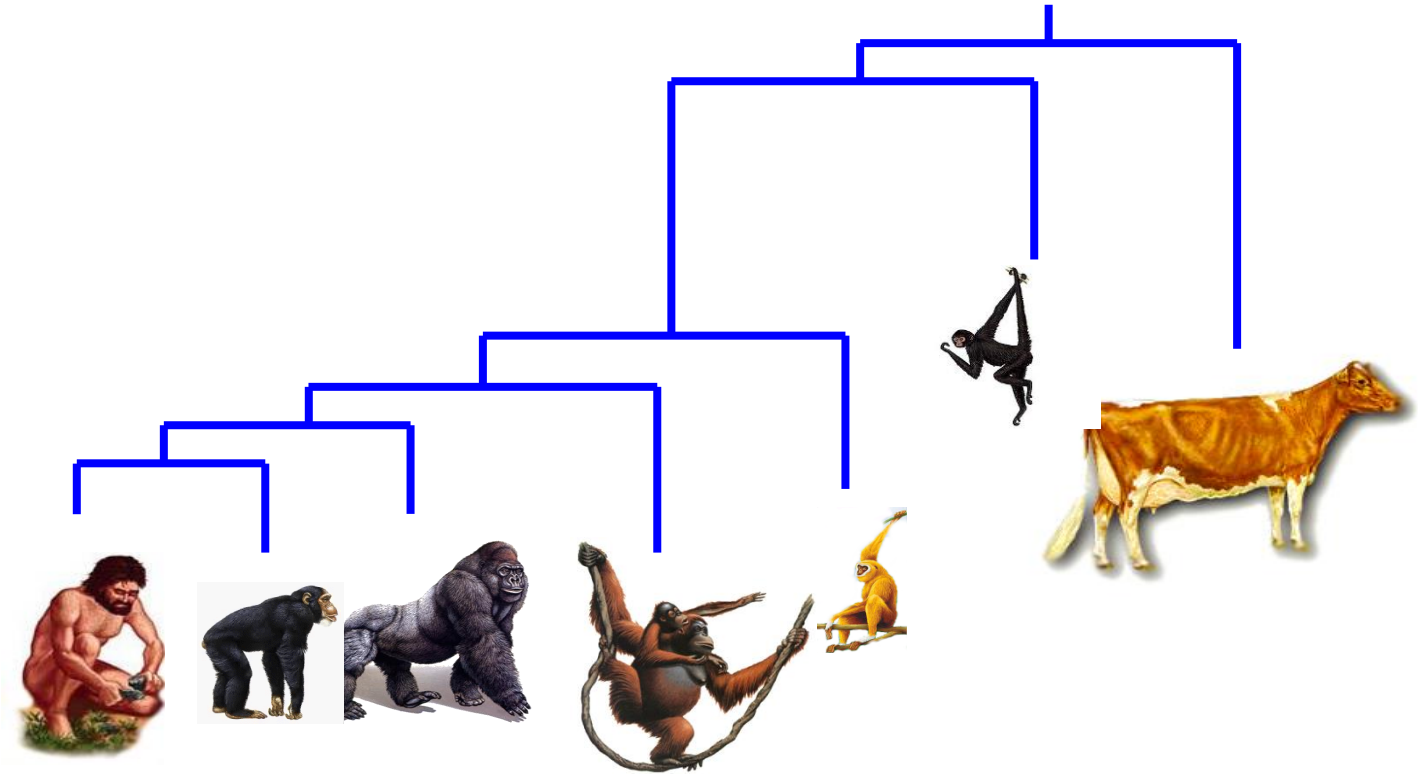
- **GRASP** [Greedy Randomized Adaptive Search Procedure] para evitar óptimos locales.

- **k-Modes** (Huang'1998) utiliza modas en vez de medias (para poder trabajar con atributos de tipo categórico).

- **k-Medoids** utiliza medianas en vez de medias para limitar la influencia de los outliers

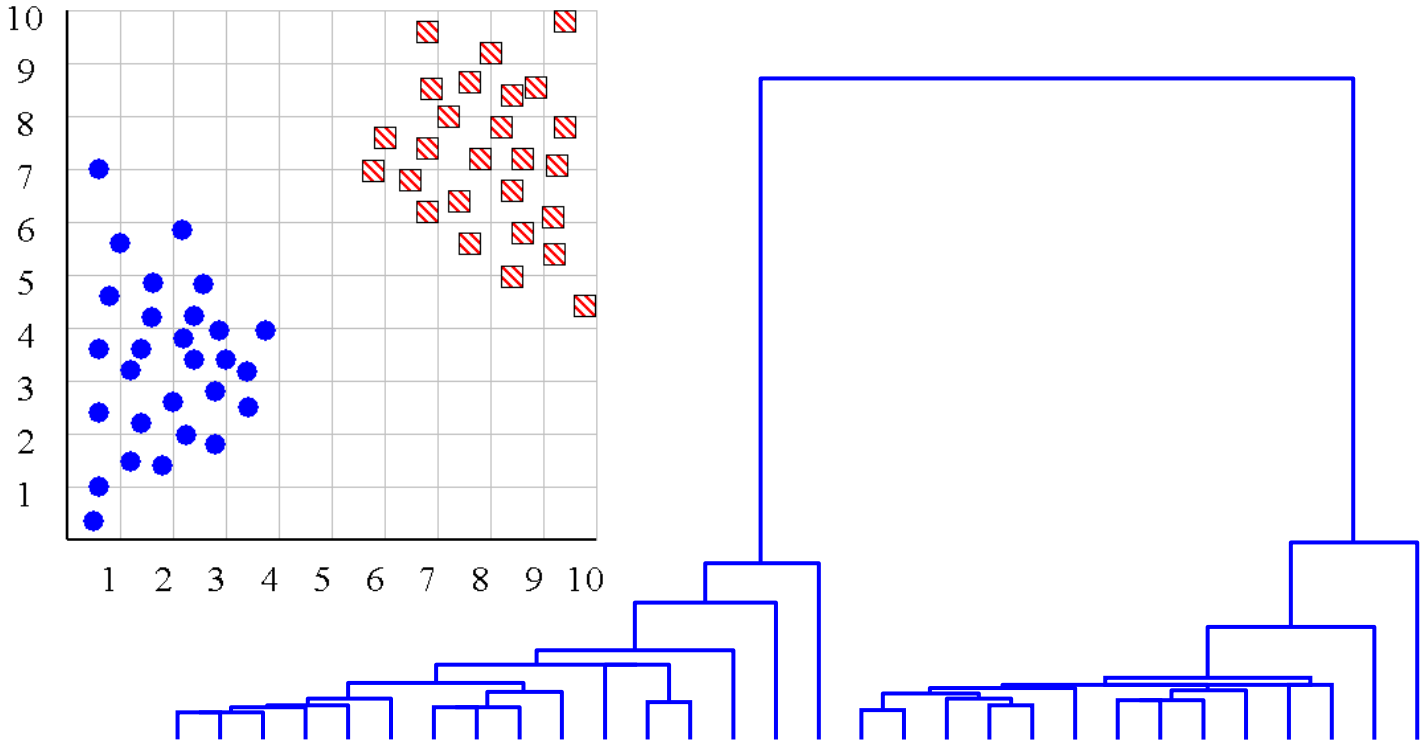
  vg. **PAM** (Partitioning Around Medoids, 1987)
  CLARA (Clustering LARge Applications, 1990)
  CLARANS (CLARA + Randomized Search, 1994)

Clustering Jerárquico



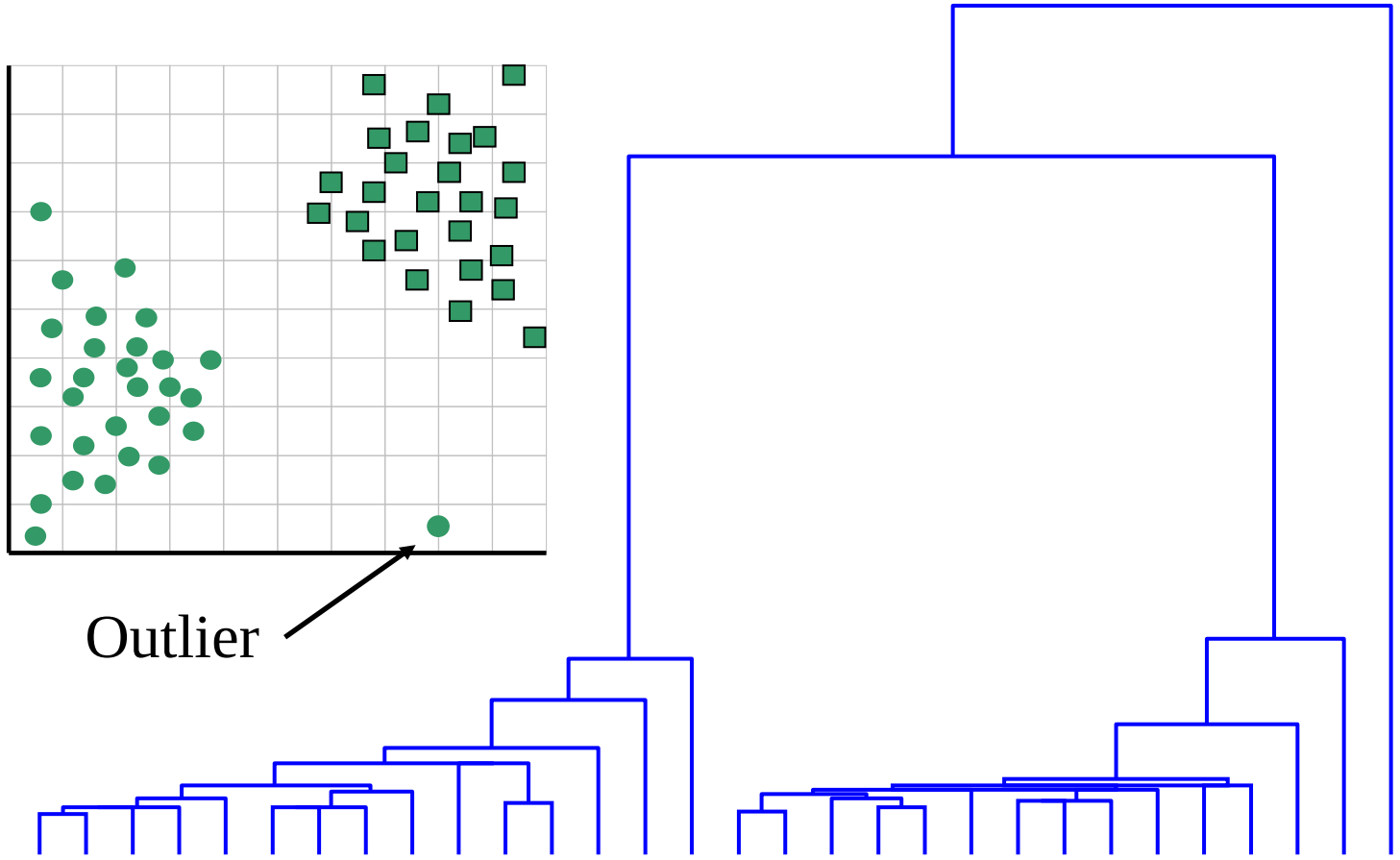
DENDROGRAMA: La similitud entre dos objetos viene dada por la "altura" del nodo común más cercano.

Clustering Jerárquico



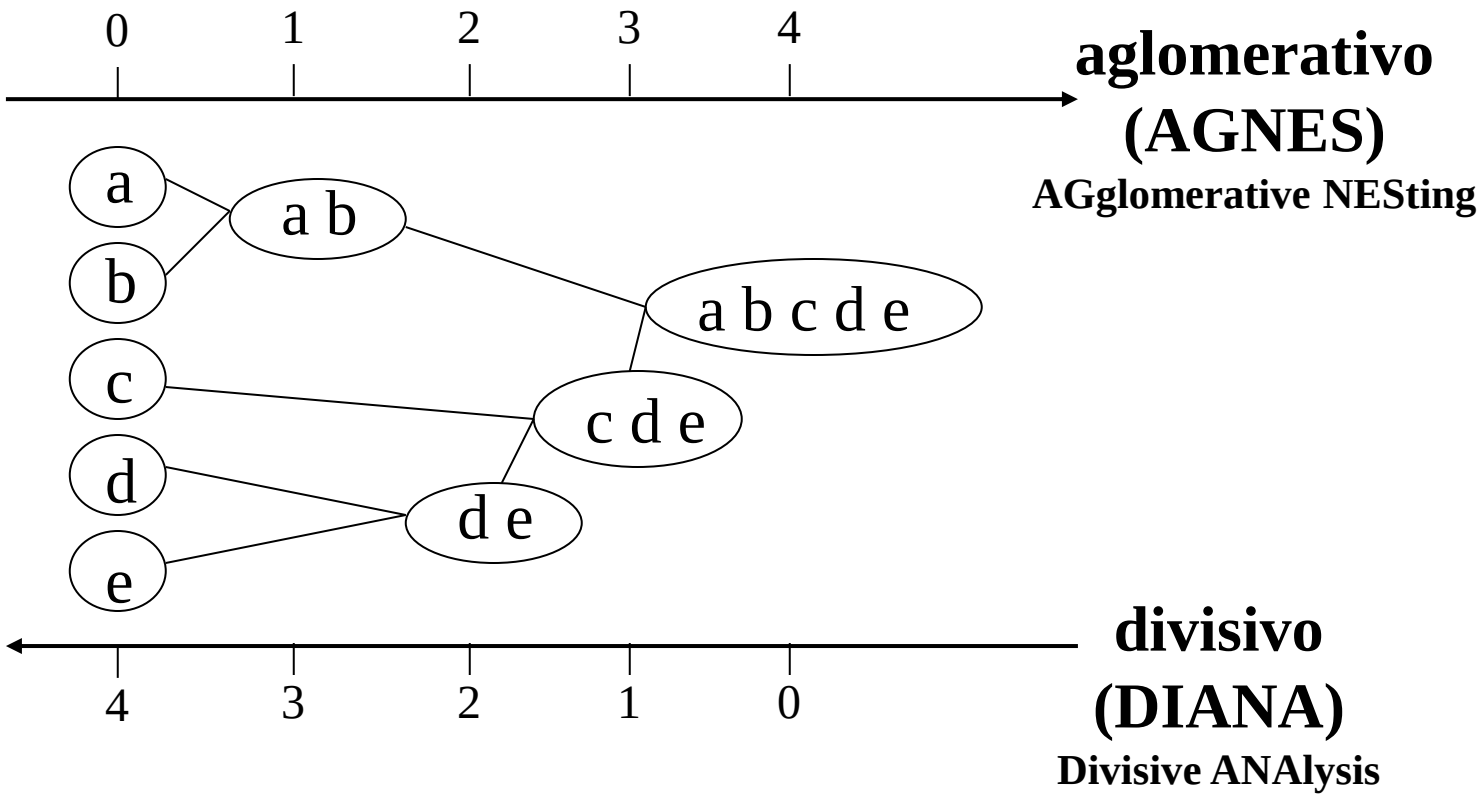
El DENDROGRAMA nos puede ayudar a determinar el número adecuado de agrupamientos (aunque normalmente no será tan fácil).

Clustering Jerárquico



El DENDROGRAMA
también nos puede servir para detectar outliers.

Clustering Jerárquico

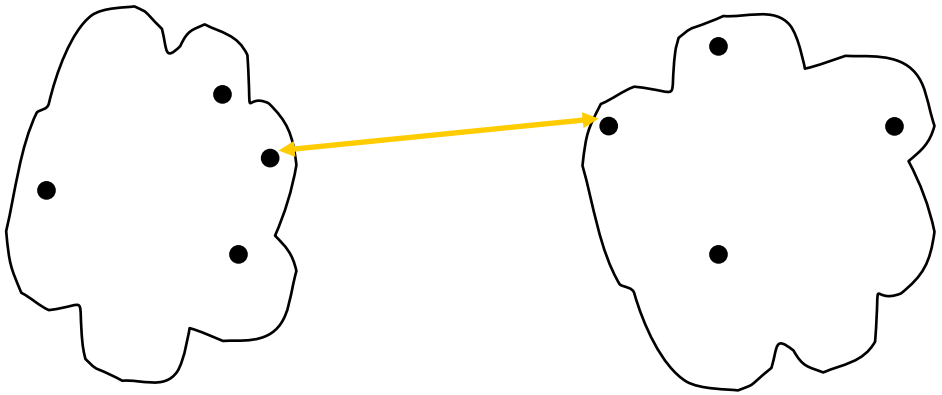


En lugar de establecer de antemano el número de clusters, tenemos que definir un criterio de parada

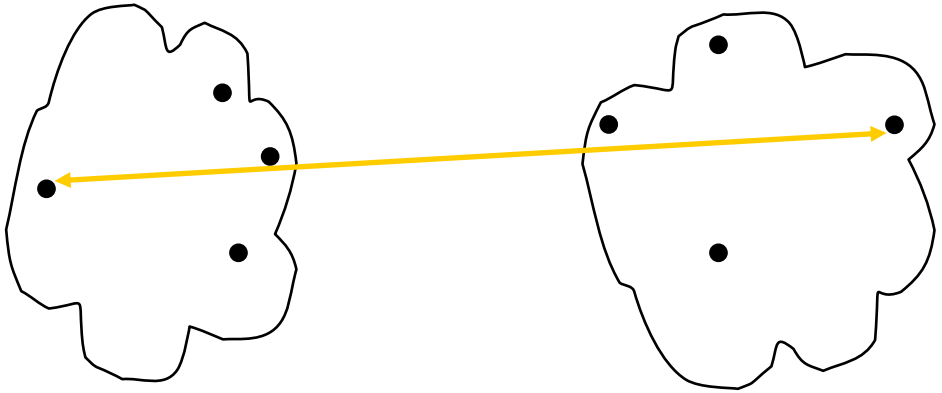
Clustering Jerárquico

¿Cómo medir la distancia entre clusters?

- MIN
  single-link



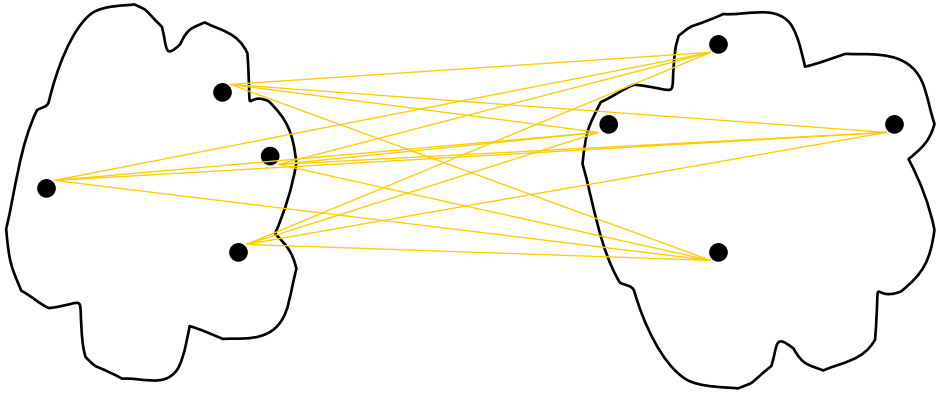
- MAX
  complete
  linkage
  (diameter)



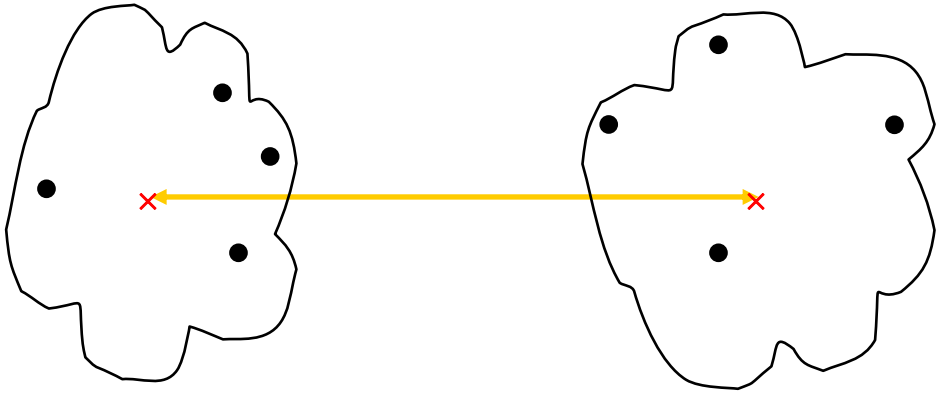
Clustering Jerárquico

¿Cómo medir la distancia entre clusters?

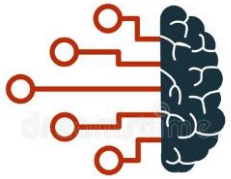
- Promedio



- Centroides
  p.ej. BIRCH



Clustering Jerárquico



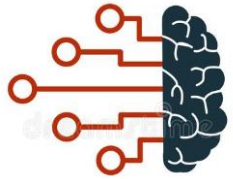
Principal inconveniente del clustering jerárquico:

Baja escalabilidad **≥ O(n^2)**

Algoritmos "escalables":

- **BIRCH:** Balanced Iterative Reducing and Clustering using Hierarchies (Zhang, Ramakrishnan & Livny, SIGMOD'1996)
- **ROCK:** RObust Clustering using linKs (Guha, Rastogi & Shim, ICDE'1999)
- **CURE**: Clustering Using REpresentatives (Guha, Rastogi & Shim, SIGMOD'1998)
- **CHAMELEON:** Hierarchical Clustering Using Dynamic Modeling (Karypis, Han & Kumar, 1999)

Clustering basado en densidad



Criterio de agrupamiento local:

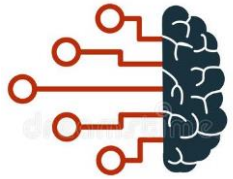
Densidad de puntos

Región densas de puntos separadas
de otras regiones densas por regiones poco densas

Características

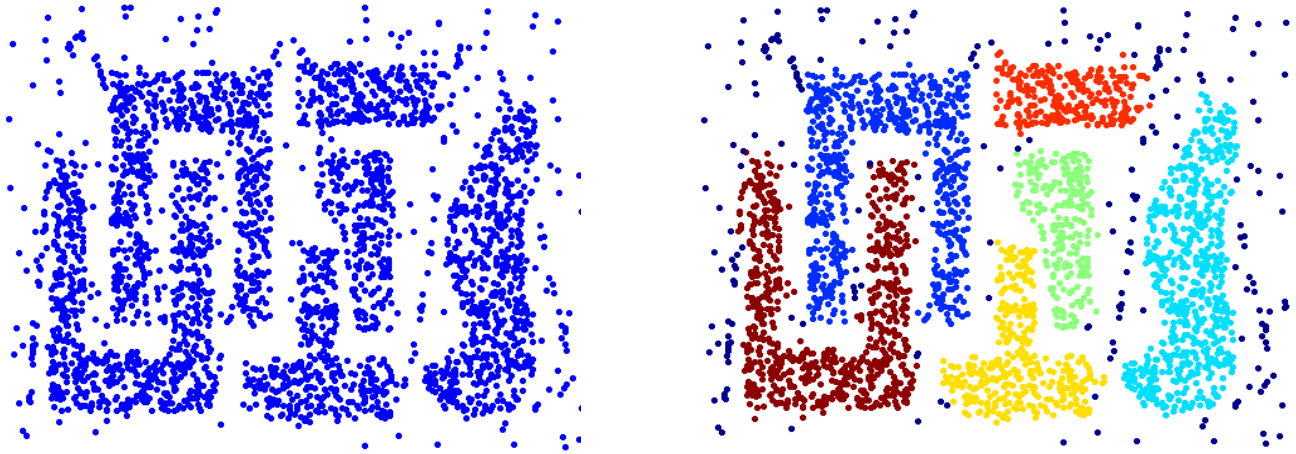
- Identifica clusters de formas arbitrarias.
- Robusto ante la presencia de ruido
- Escalable: Un único recorrido del conjunto de datos

Clustering basado en densidad - Algoritmos



- **DBSCAN**: Density Based Spatial Clustering of Applications with Noise (Ester et al., KDD'1996)
- **OPTICS**: Ordering Points To Identify the Clustering Structure (Ankerst et al. SIGMOD'1999)
- **DENCLUE**: DENsity-based CLUstEring (Hinneburg & Keim, KDD'1998)
- **CLIQUE**: Clustering in QUEst (Agrawal et al., SIGMOD'1998)
- **SNN** (Shared Nearest Neighbor) density-based clustering (Ertöz, Steinbach & Kumar, SDM'2003)

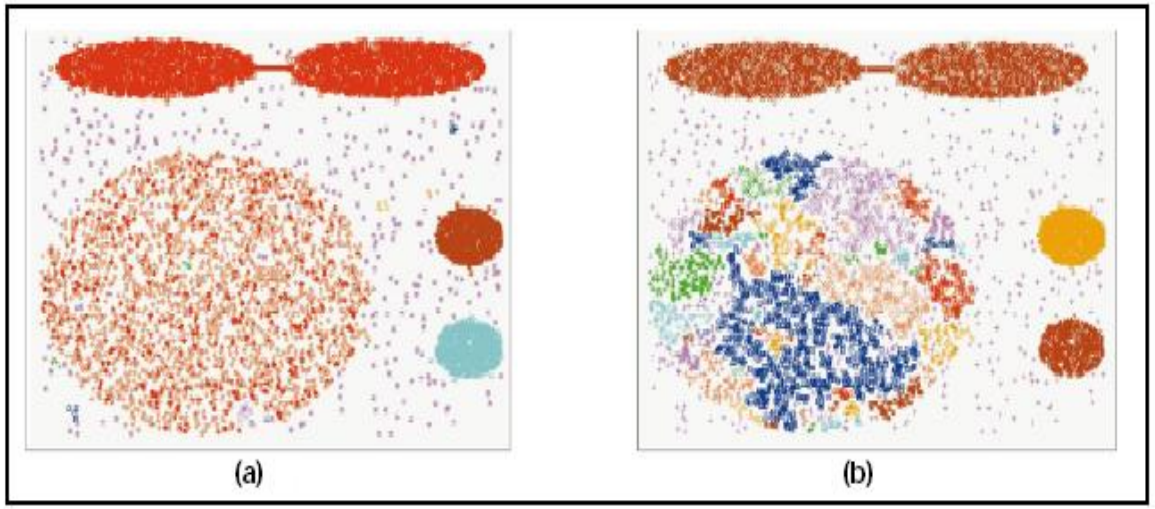
Clustering basado en densidad

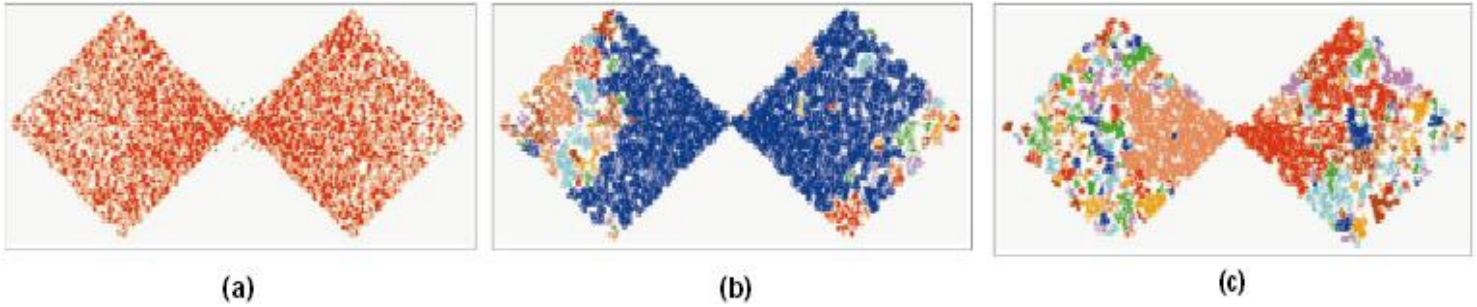


Clusters

DBSCAN … cuando funciona bien

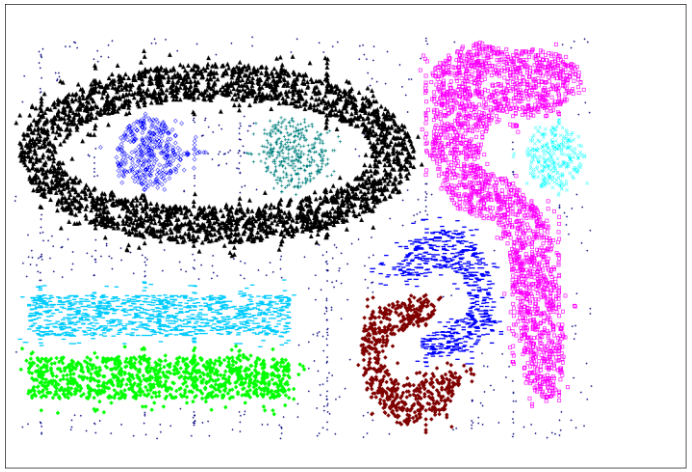
Clustering basado en densidad

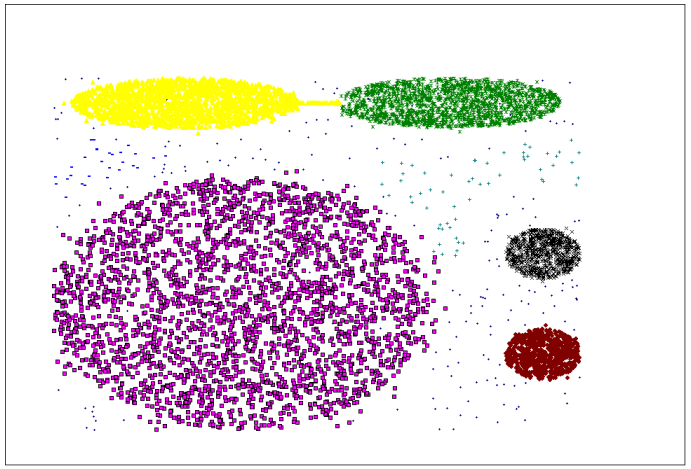


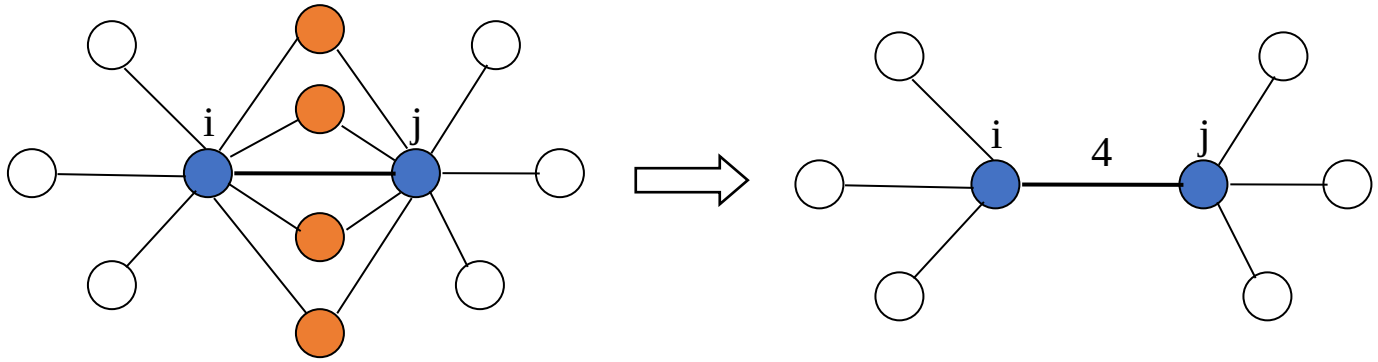


DBSCAN sensible al valor inicial de sus parámetros

Clustering basado en densidad

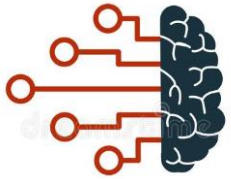






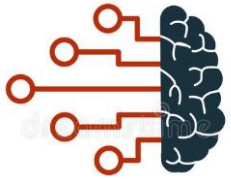
SNN density-based clustering… O(n^2)

Otros métodos



Grids multiresolución

- **STING**, a STatistical INformation Grid approach (Wang, Yang & Muntz, VLDB'1997)
- **WaveCluster**, basado en wavelets (Sheikholeslami, Chatterjee & Zhang, VLDB'1998)
- **CLIQUE:** CLustering In QUEst (Agrawal et al., SIGMOD'1998)



Gracias……