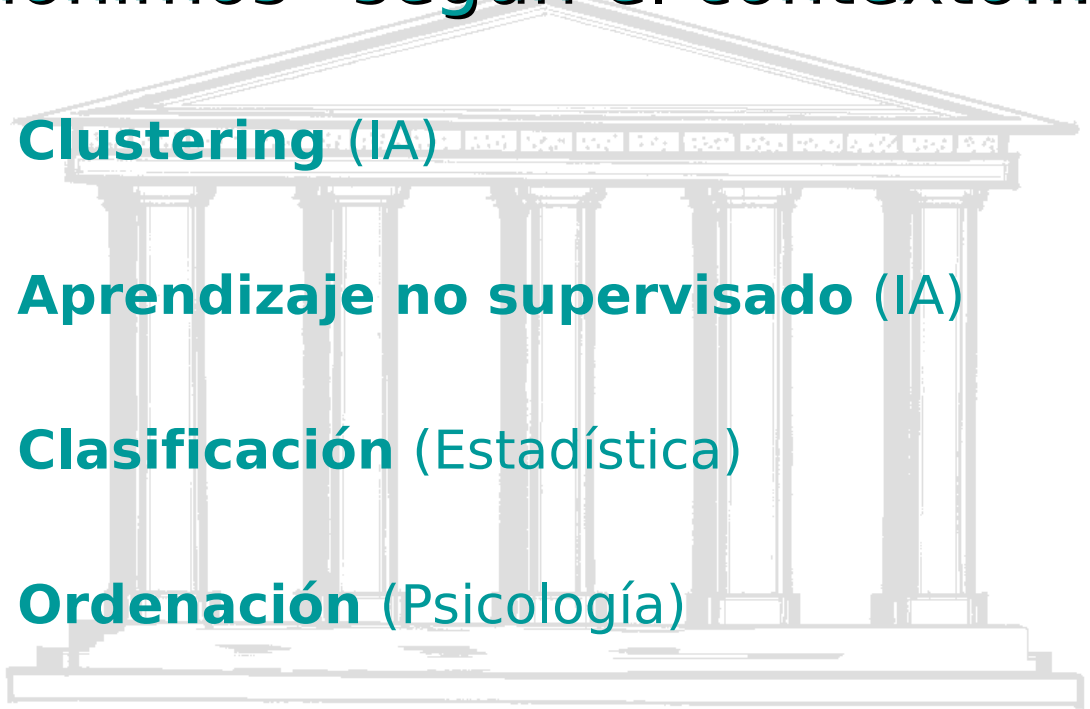
Clustering

Fernando Berzal
fberzal@decsai.ugr.es
http://elvex.ugr.es/idbis/dm/

Intelligent Databases and Information Systems research group
Department of Computer Science and Artificial Intelligence
E.T.S Ingeniería Informática – Universidad de Granada (Spain)

Clustering

"Sinónimos" según el contexto...



- **Clustering** (IA)
- **Aprendizaje no supervisado** (IA)
- **Clasificación** (Estadística)
- **Ordenación** (Psicología)
- **Segmentación** (Marketing)

Clustering

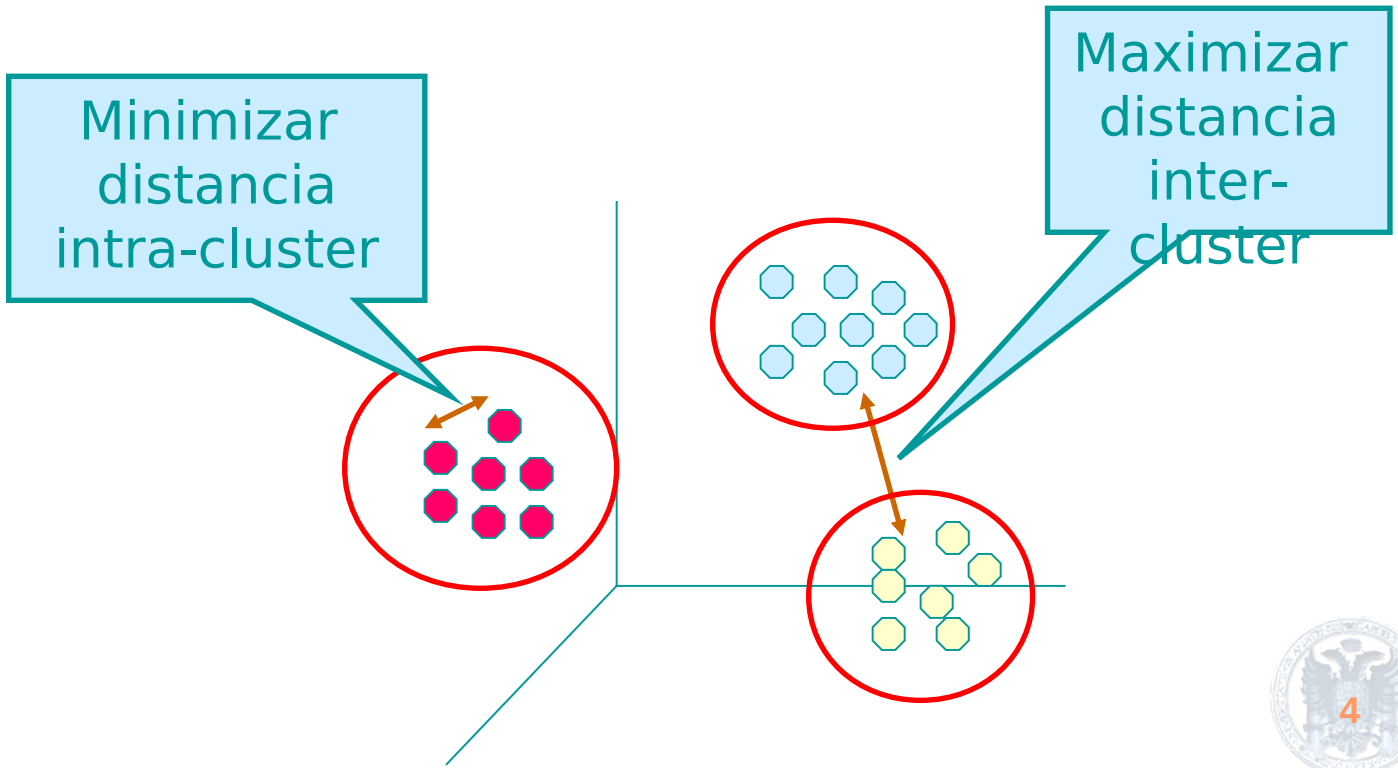
- **Objetivo**
  Agrupar objetos similares entre sí que sean distintos a los objetos de otros agrupamientos [clusters].

- **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

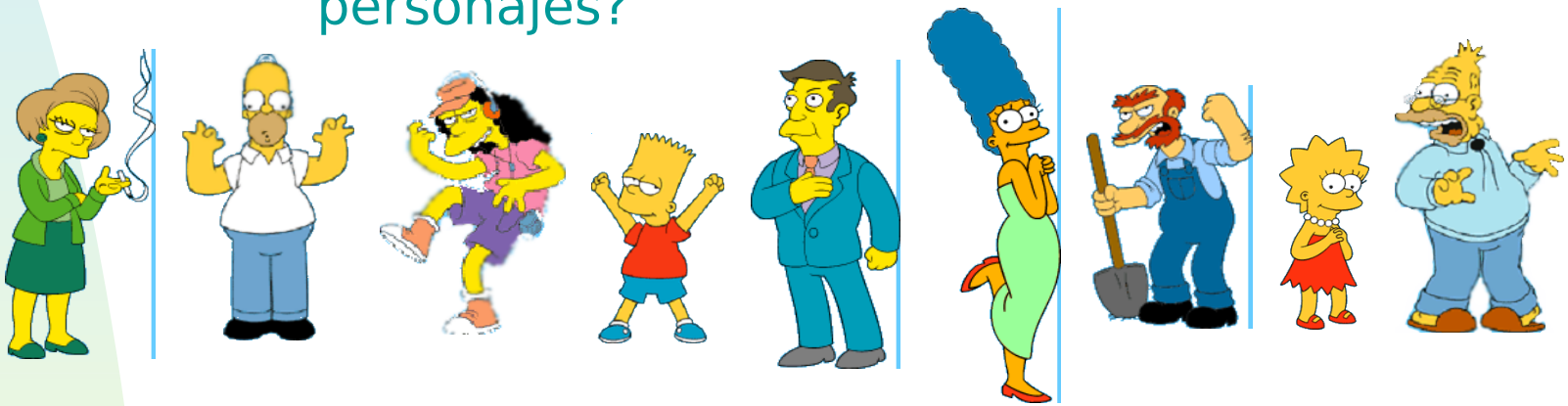
- Reconocimiento de formas.
- Mapas temáticos (GIS)
- Marketing: Segmentación de clientes
- Clasificación de documentos
- Análisis de web logs (patrones de acceso similares)
- ...

Aplicaciones típicas en *Data Mining*:

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

Clustering

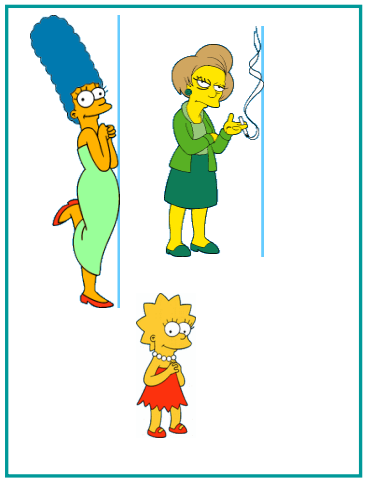
¿Cuál es la forma natural de agrupar los personajes?

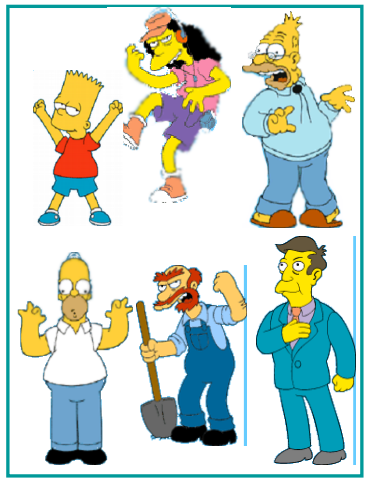


Hombres

vs.

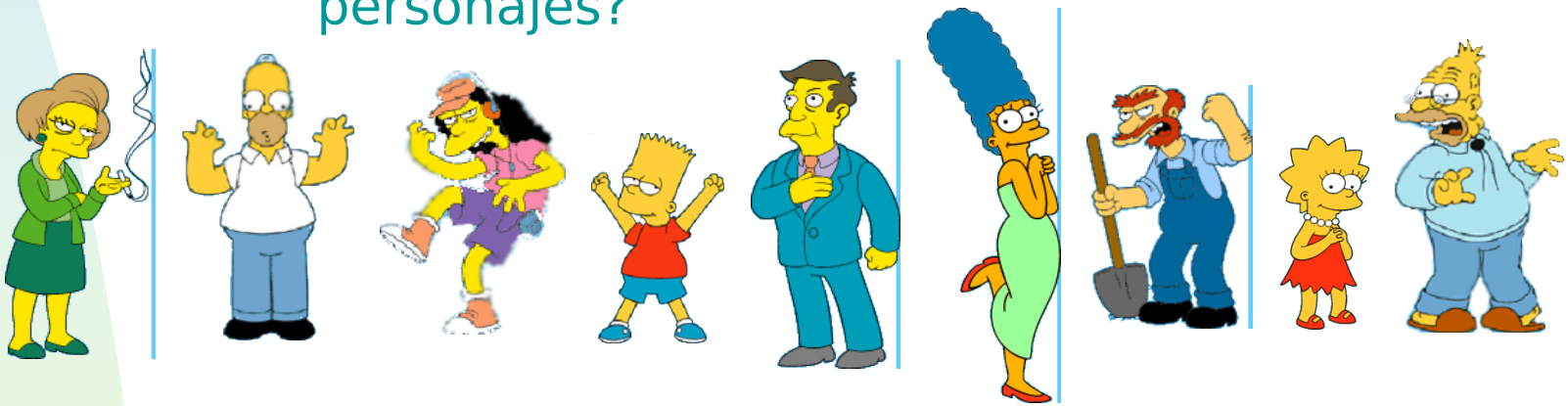
Mujeres



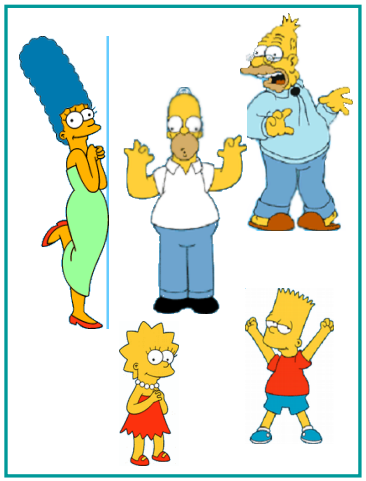


Clustering

¿Cuál es la forma natural de agrupar los personajes?

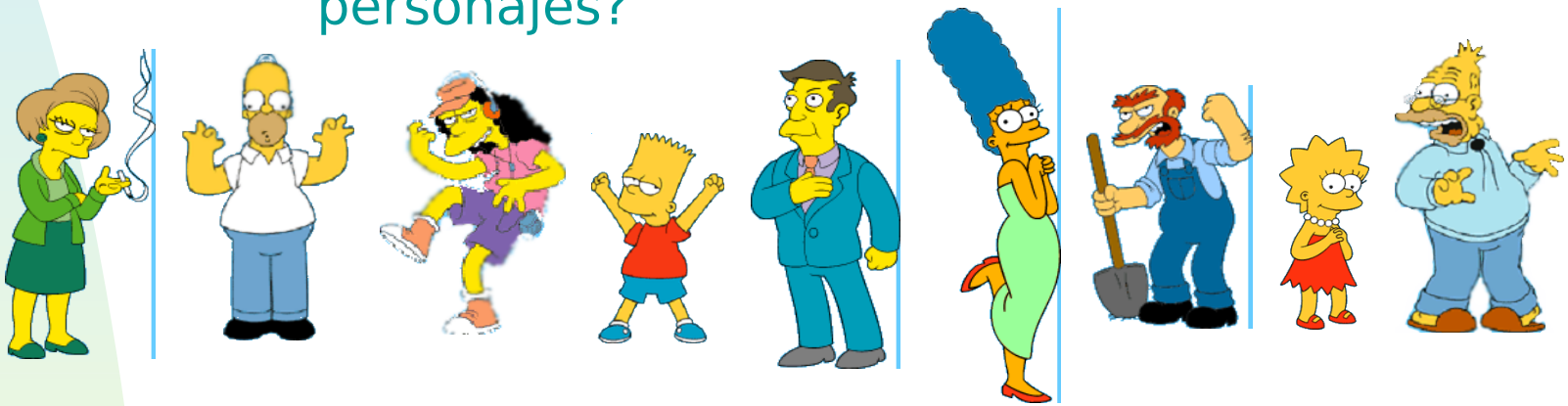


Simpsons
vs.
Empleados
de la escuela
de Springfield



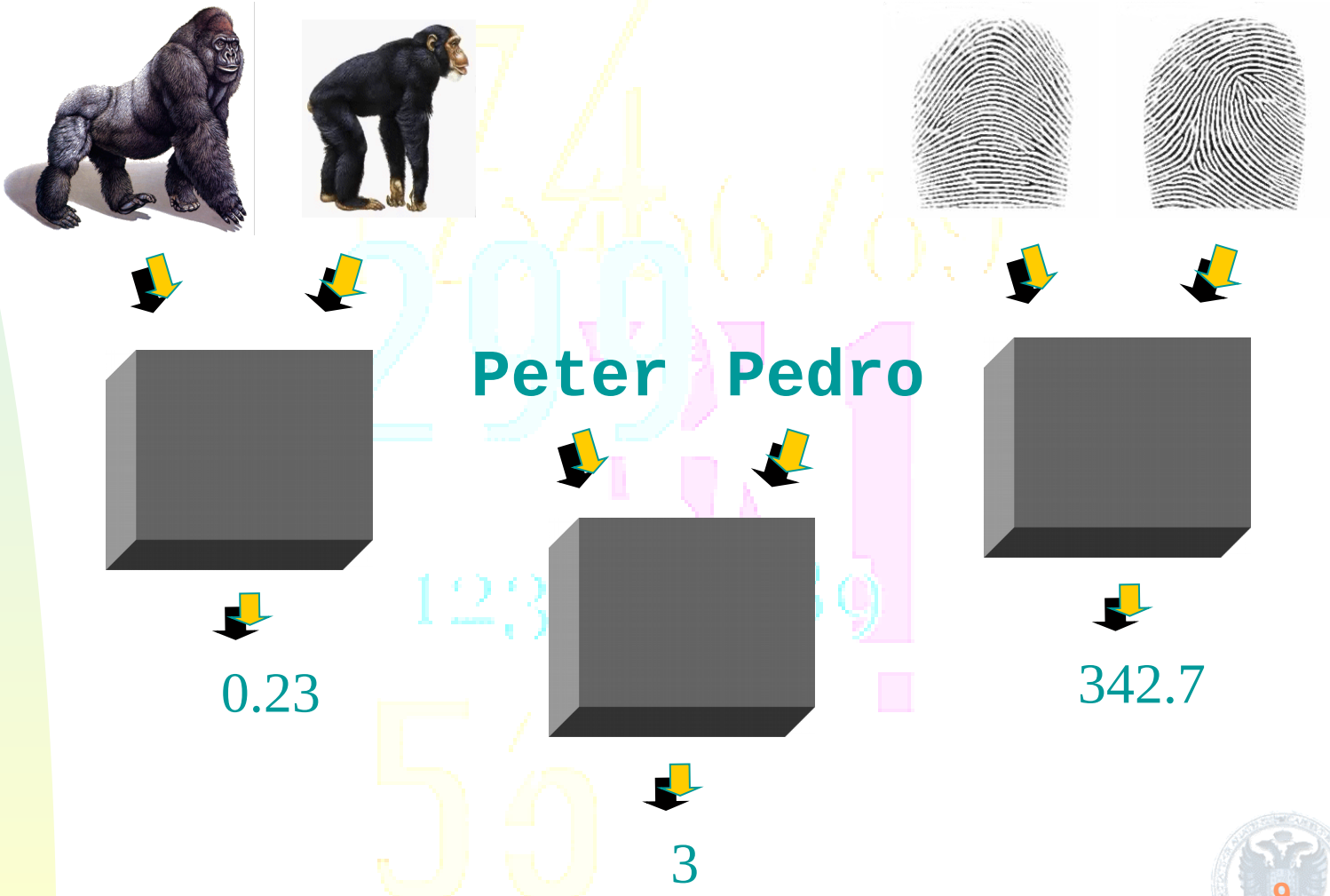
Clustering

¿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:

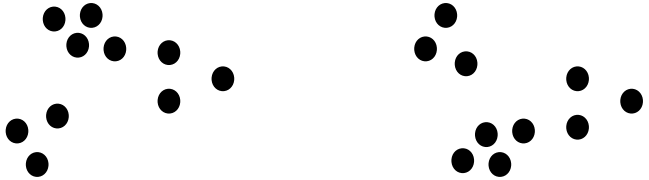
d(i,j) > d(i,k)

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**

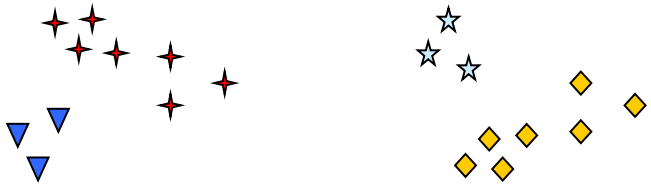
Medidas de similitud



¿Cuántos agrupamientos?

¿Dos?

¿Seis?



¿Cuatro?

Medidas de similitud

Atributos continuos

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* / *taxica*

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

- Distar

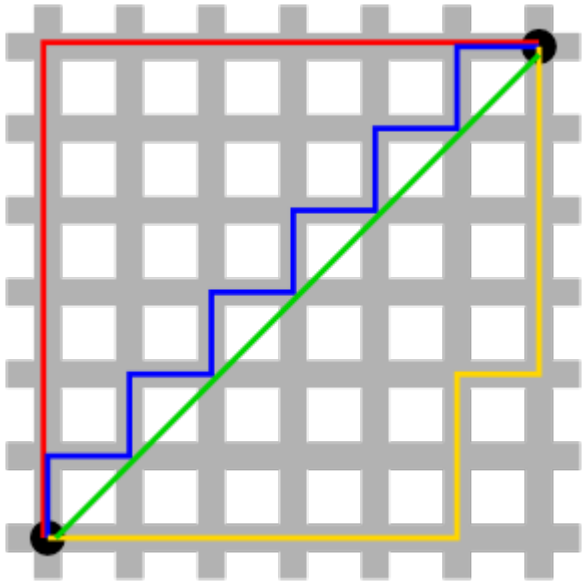
$$d_2(x,y) = \sqrt{\sum_{j=1}^{J} (x_j - y_j)^2}$$

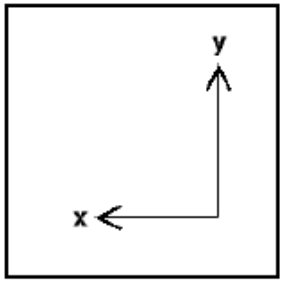
- Distar $d_\infty(x,y) = \max_{j=1..J} |x_j - y_j|$ *ominio* / *chessboard*

Medidas de similitud

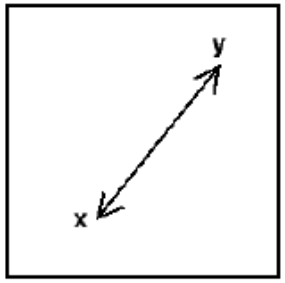
Métricas de distancia

Distancia de Minkowski





Manhattan



Euclidean

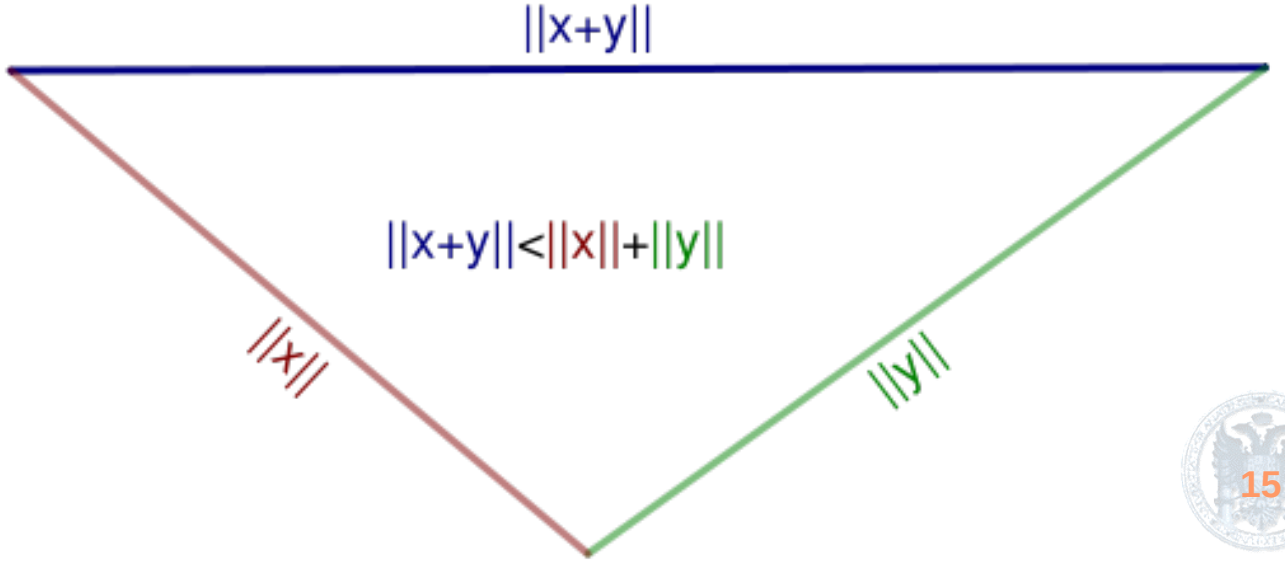
- Distancia de Manhattan = 12
- Distancia Euclídea ≅ 8.5
- Distancia de Chebyshev = 6

Medidas de similitud

Métricas de distancia

Distancia de Minkowski $d(i,j) \geq 0$

- Propiedad reflexiva $d(i,i) = 0$
- Propiedad simétrica $d(i,j) = d(j,i)$
- *Desigualdad triangular* $d(i,j) \leq d(i,k)+d(k,j)$



Medidas de similitud

Métricas de distancia

Distancia de Chebyshev

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

- También conocida como distancia de tablero de ajedrez (chessboard distance) Número de movimientos que el rey ha de hacer para llegar de una casilla a otra en un tablero de ajedrez.

Clustering

IIIBIS

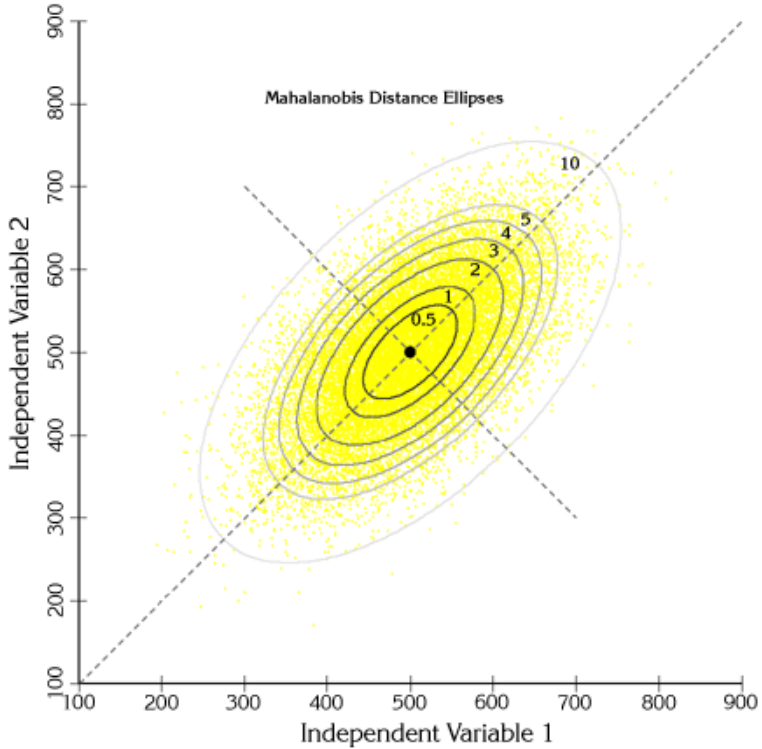
Medidas de similitud

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

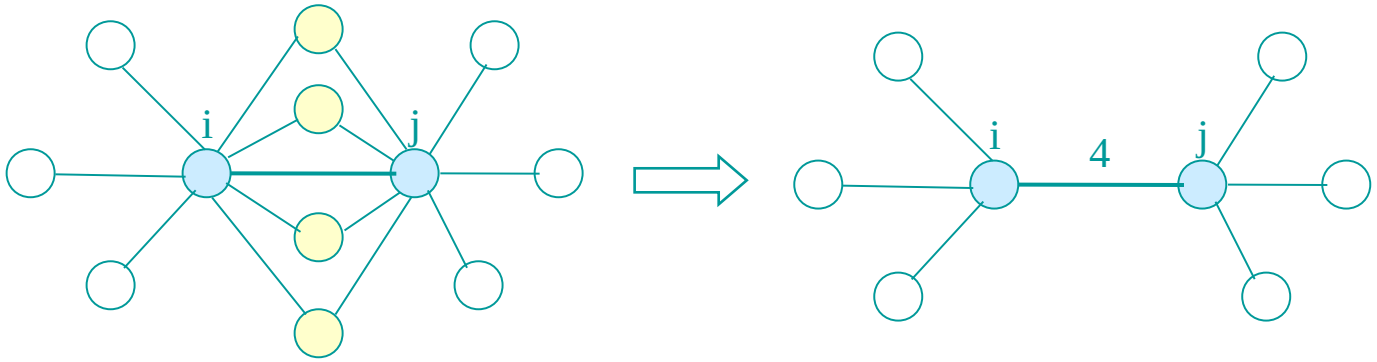
d(“night”,”natch”) = d(“natch”,”noche”) = 3

Aplicaciones: Correctores ortográficos, reconocimiento de voz, detección de plagios,

Medidas de similitud

Métricas de distancia

Vecinos compartidos



- "Mutual Neighbor Distance"

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

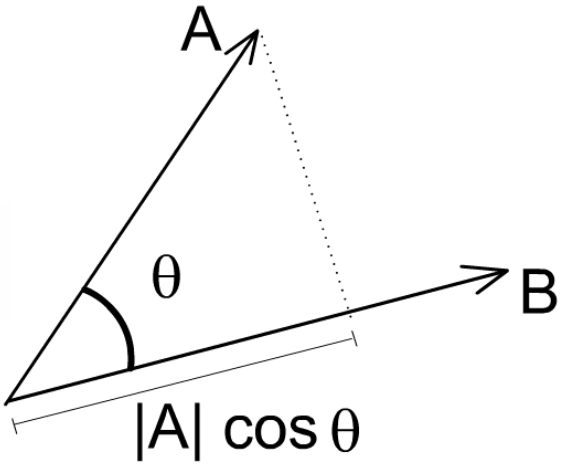
donde $NN(x_i,x_j)$ es el número de vecino de x_j con respecto a x_i

Medidas de similitud

Medidas de correlació

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

$$s(a,b) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A),$$
donde $\theta, \alpha, \beta \geq 0$

- Mode

$$-S_{Restle}(A,B) = |A \square B|$$
$$-S_{\square}(A,B) = \sup_x \mu_{A \square B}(x)$$

- Intersección

$$S_{MinSum}(A,B) = |A \cap B|$$
$$-S_{Enta}(A,B) = 1 - \sup_x \mu_{A \cap B}(x)$$

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

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

- D

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

Métodos de agrupamiento

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

Métodos de agrupamiento

Tipos de algoritmos de clustering

- Agrupamiento por particiones

  k-Means, CLARANS

- Clustering jerárquico

  BIRCH, ROCK, CHAMELEON

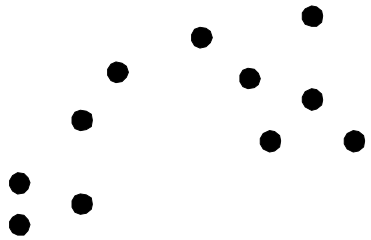
- Métodos basados en densidad

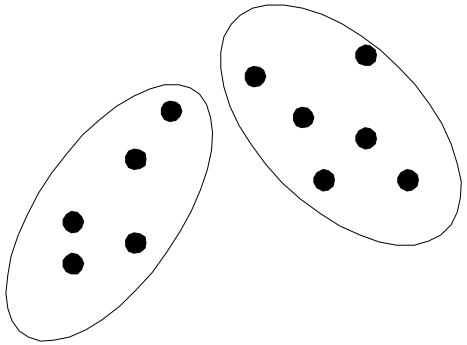
  DBSCAN

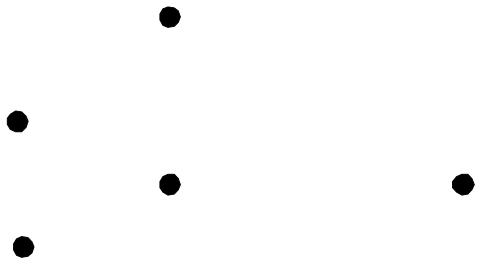
- …

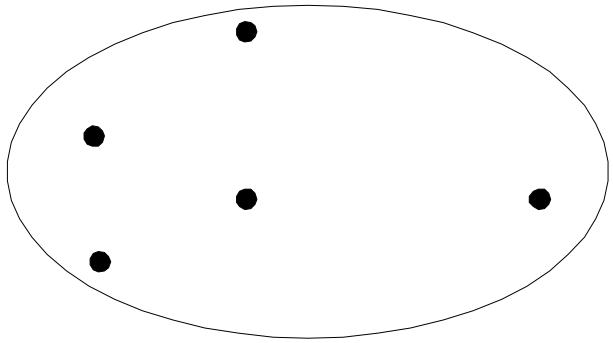
Métodos de agrupamiento

Clustering por particiones







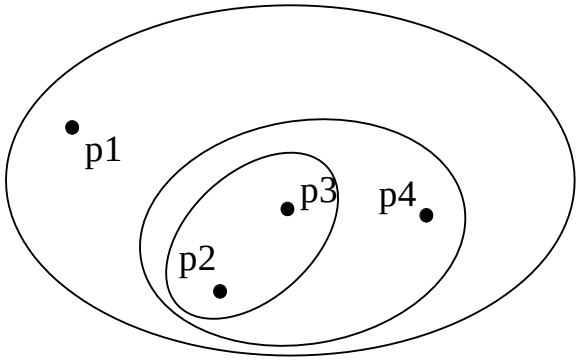


Datos originales

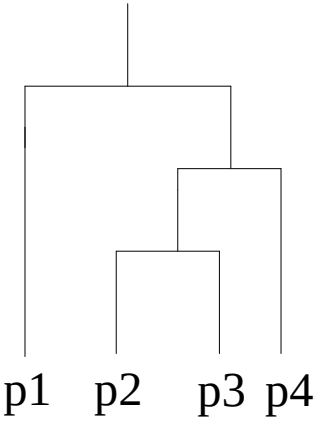
Datos agrupados

Métodos de agrupamiento

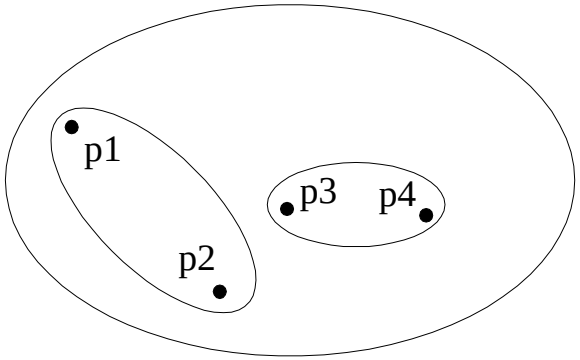
Clustering jerárquico



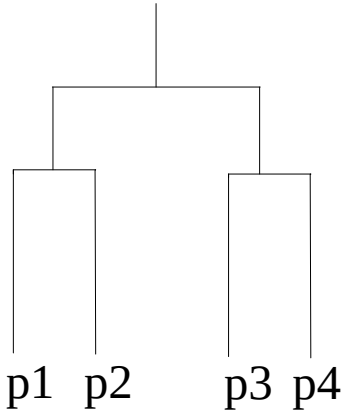
Tradicional



DENDOGRAMA



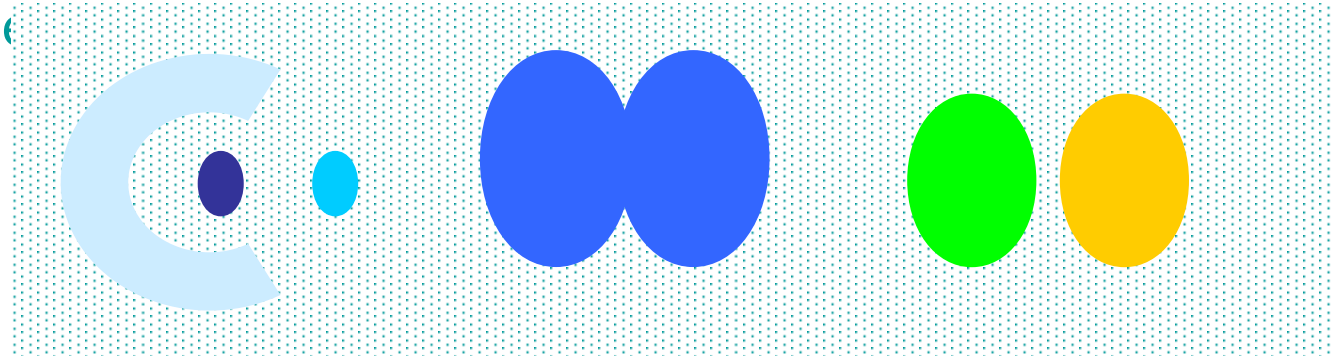
No tradicional



Métodos de agrupamiento

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

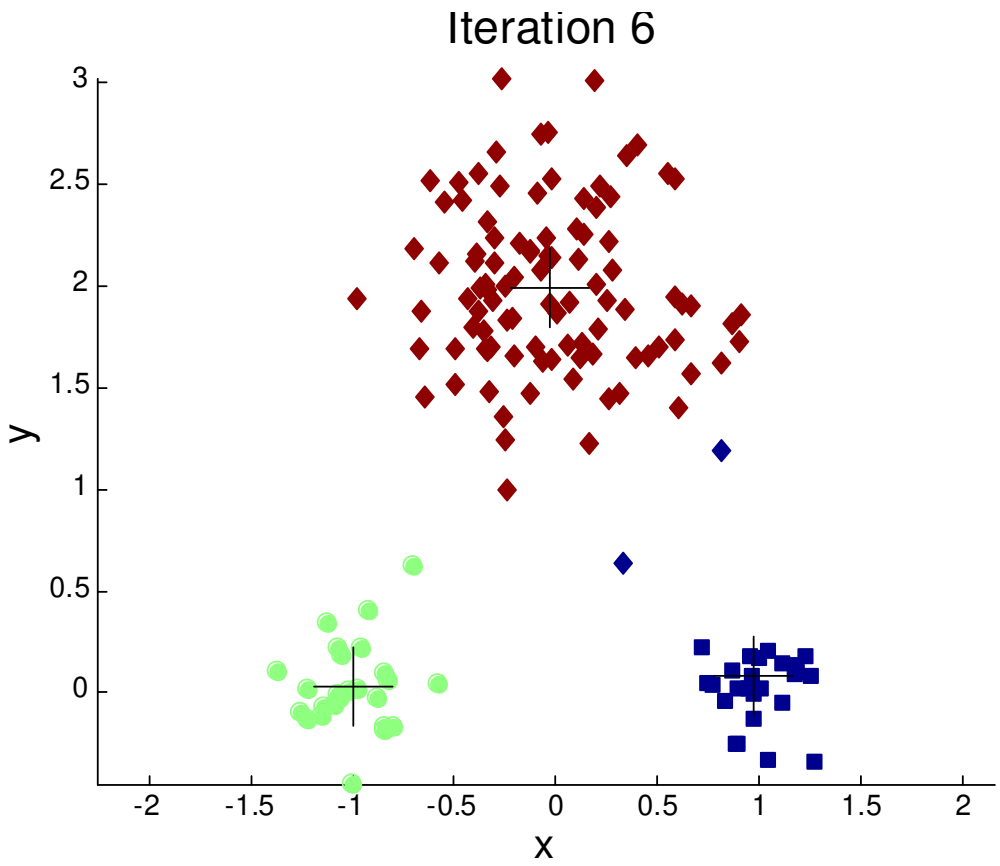


k-Means

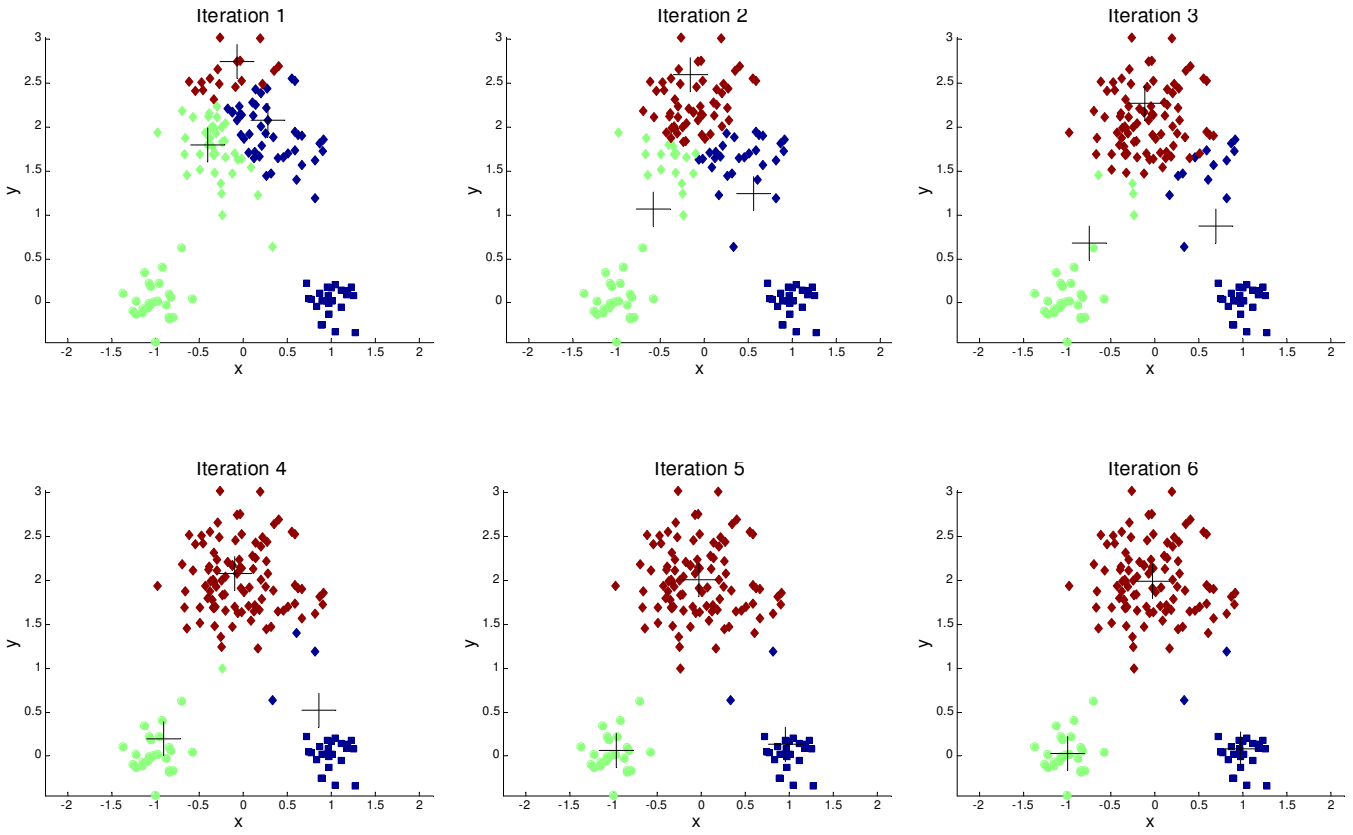
Algoritmo de agrupamiento por particiones (MacQueen, 1967)

- 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

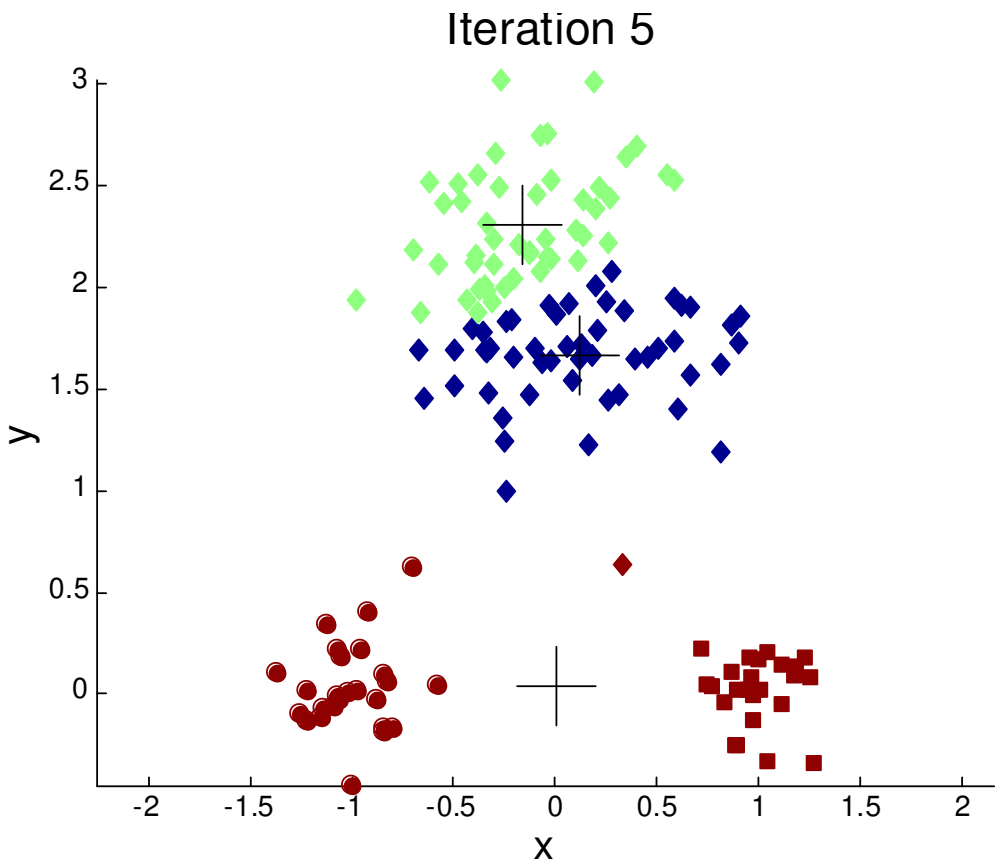
k-Means



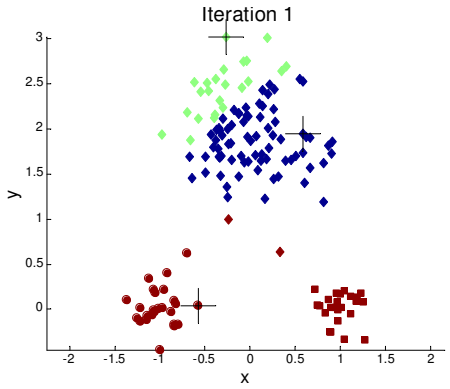
k-Means

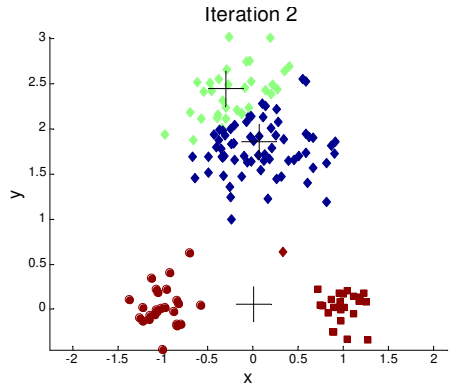

Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5
Iteration 6
x
y

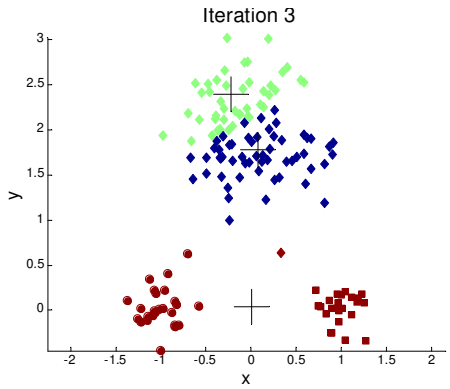
k-Means

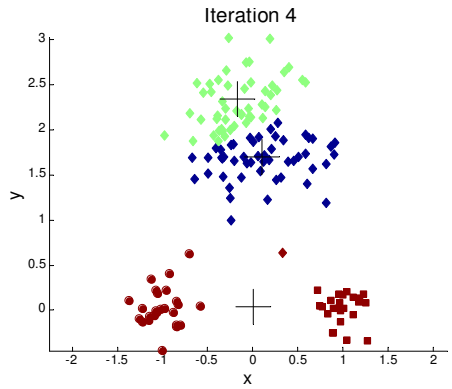


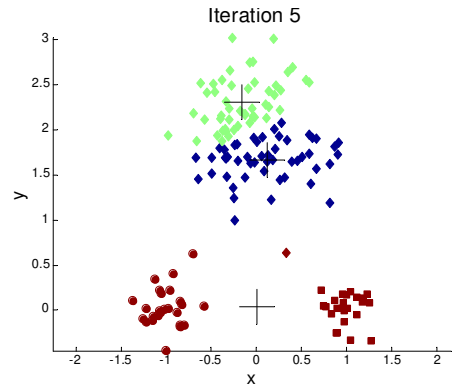
k-Means







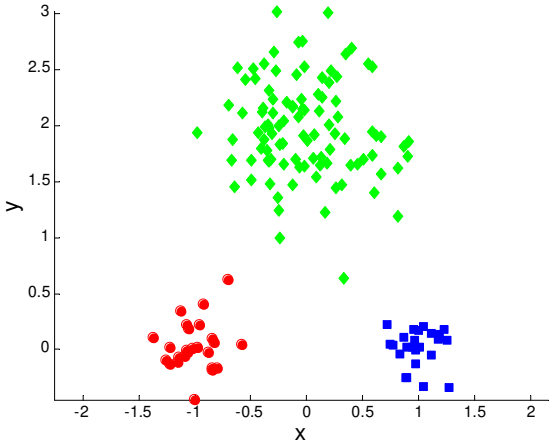




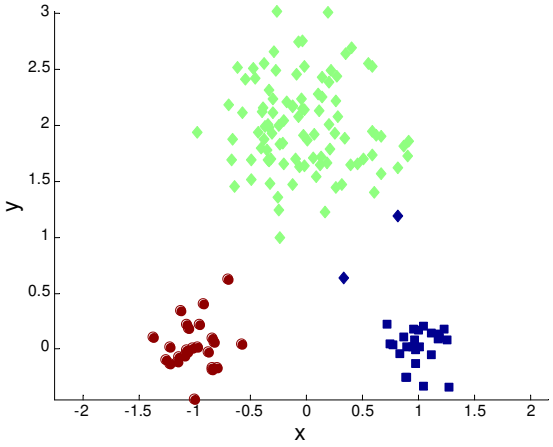
Clustering

IdBIS

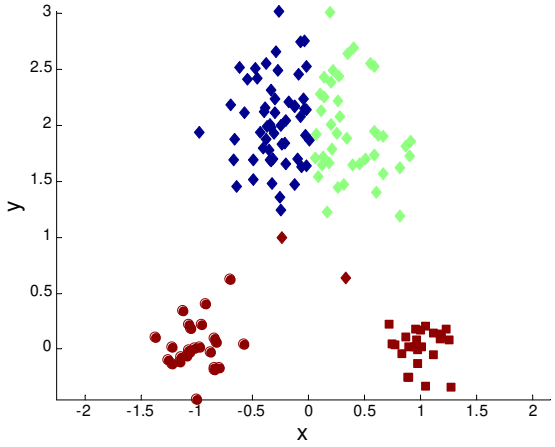
k-Means



Puntos originales



Solución óptima



Óptimo local

k-Means

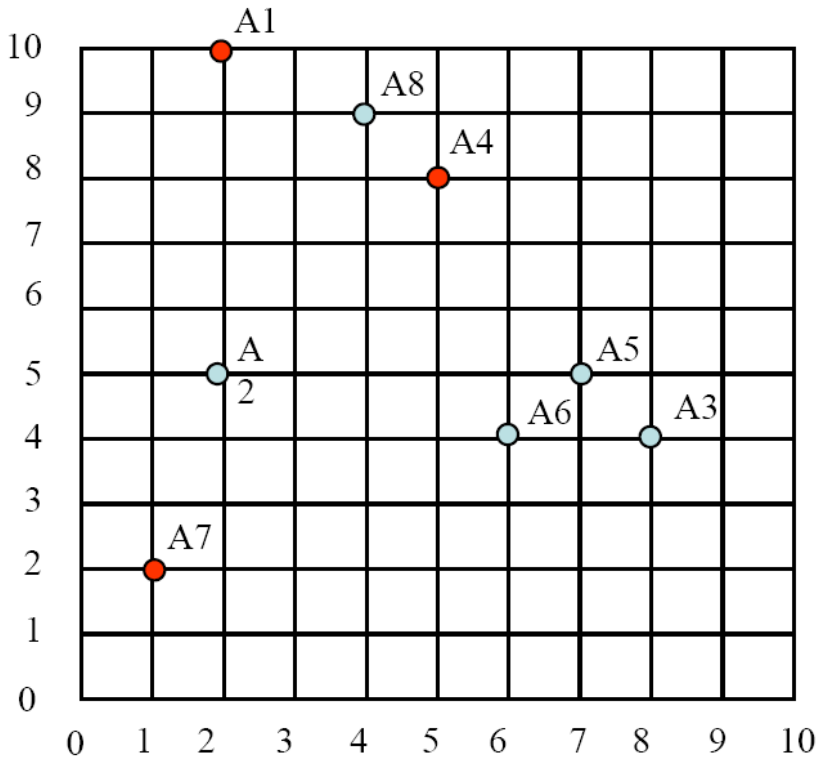
Ejercicio

Agrupar los 8 puntos de la figura en 3 clusters usando el algoritmo de las K medias

Centroides iniciales:
A1, A7 y A8

Métricas de distancia:

- Distancia euclídea
- Distancia de Manhattan
- Distancia de Chebyshev



k-Means

Ejercicio resuelto

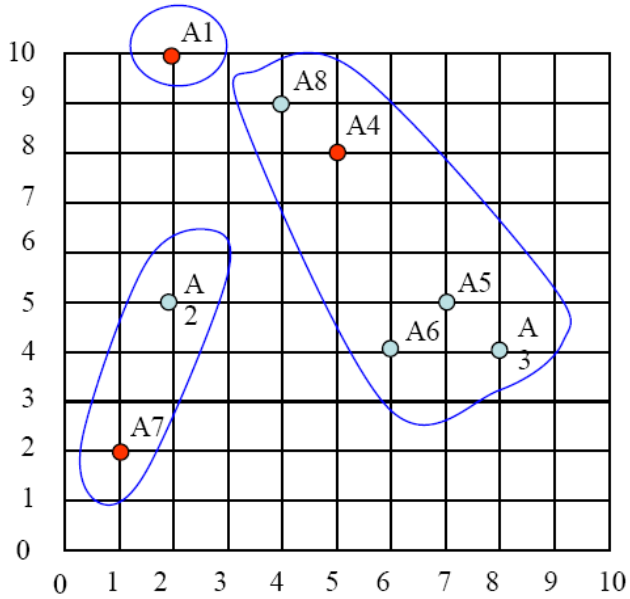
Distancia euclídea

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

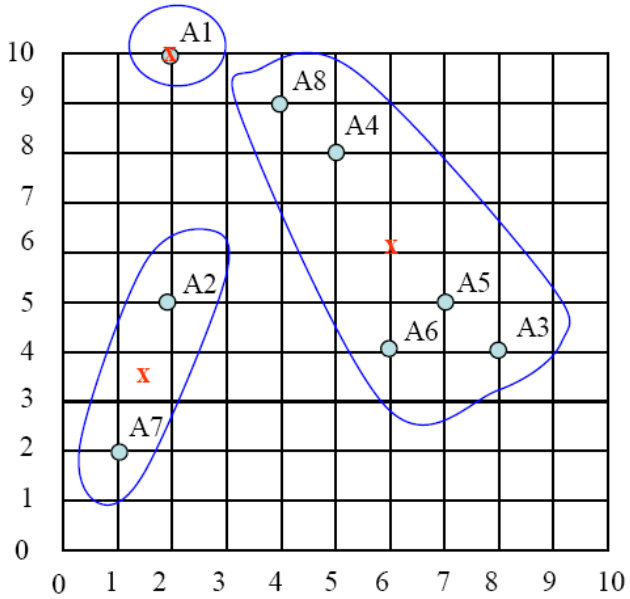
k-Means

Ejercicio resuelto

Distancia euclídea



Primera iteración

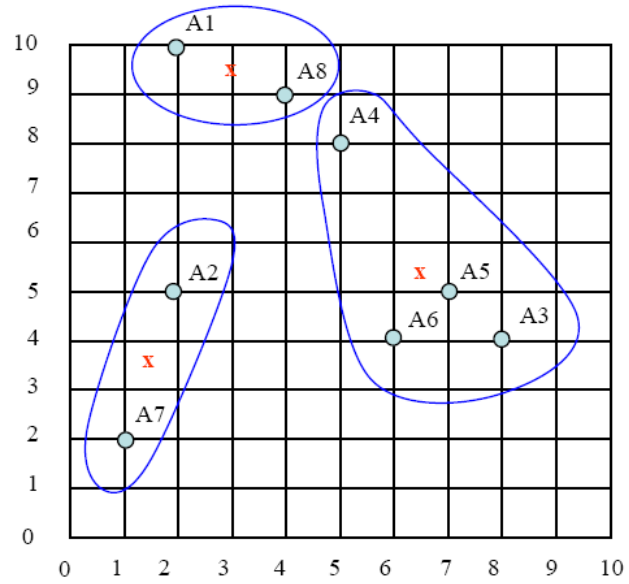


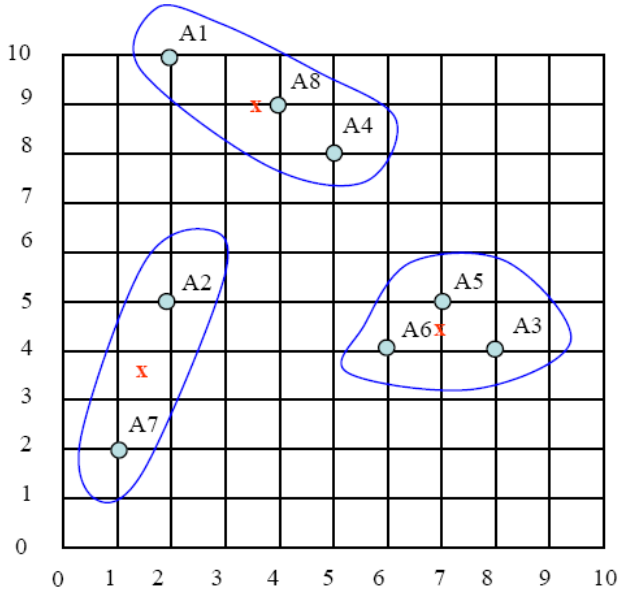
Segunda iteración

k-Means

Ejercicio resuelto

Distancia euclídea





Tercera iteración

Configuración final

k-Means

DEMO: K-Means

http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/AppletKM.html

k-Means

Ventaja

- Eficiencia **O(n·k·I·d)**

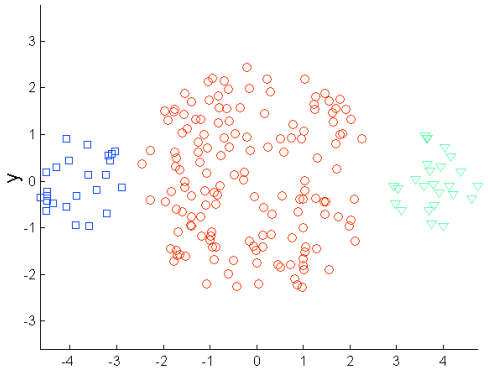
  vs. PAM $O(I \cdot k(n-k)^2)$

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

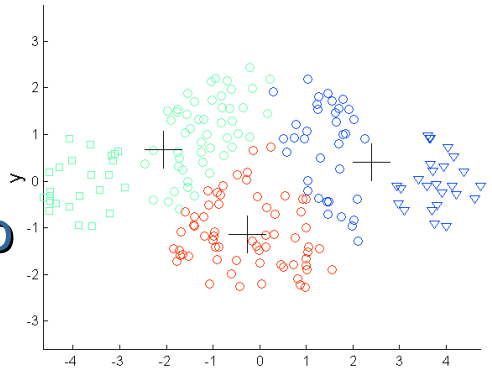
Desventajas

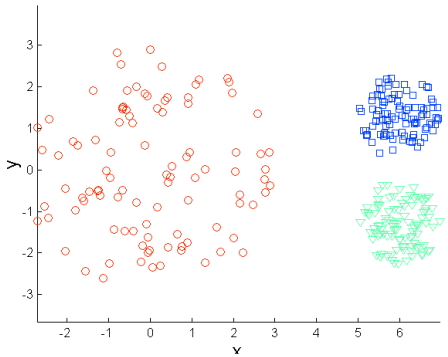
- Termina en un óptimo local:
  El resultado depende de la selección inicial de centroides.
- Necesidad de conocer el número de agrupamientos k
- Incapacidad para detectar ruido / identificar outliers.
- No resulta adecuado para detectar clusters no convexos
- Si tenemos datos de tipo categórico,

k-Means

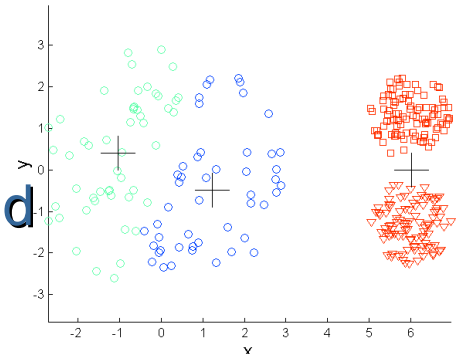


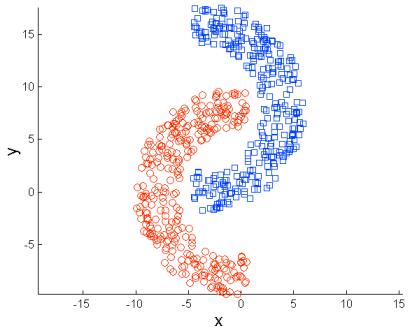
Clusters de distinto tamaño



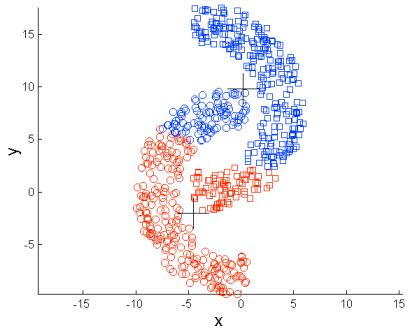


Clusters de distinta densidad





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

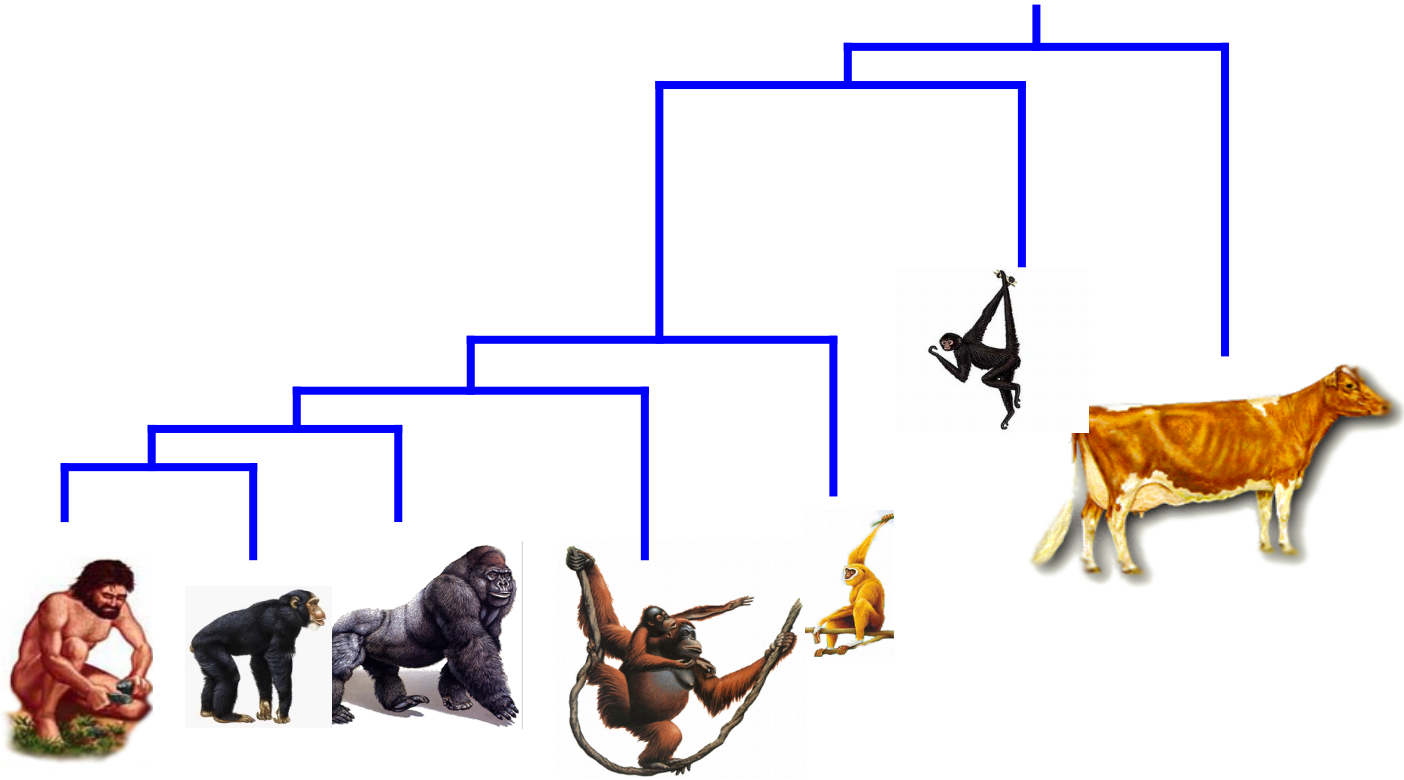
  vg. **PAM** (Partitioning Around Medoids, 1987)
  CLARA (Clustering LARge Applications, 1990)
  CLARANS (CLARA + Randomized Search,

k-Means

DEMO: Fuzzy C-Means

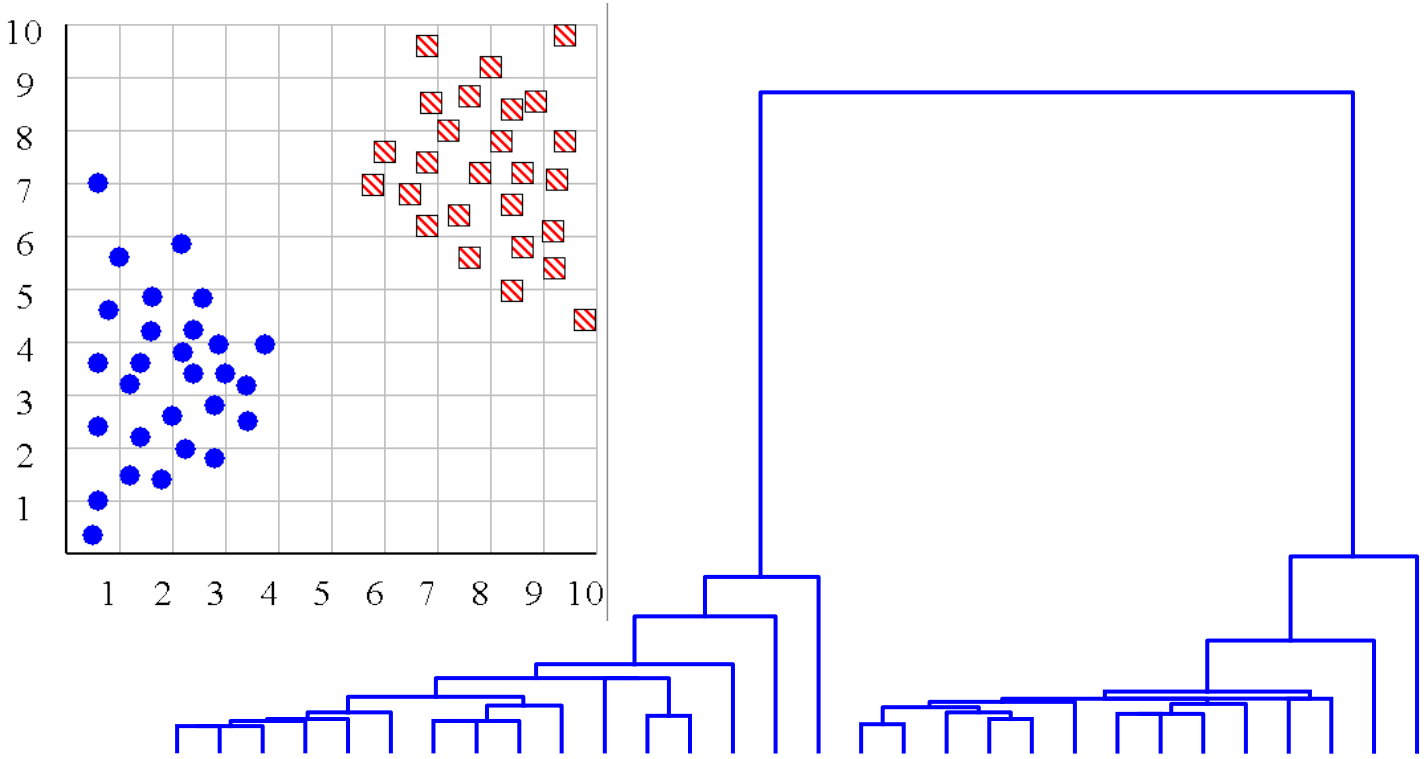
http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/AppletFCM.html

Clustering jerárquico



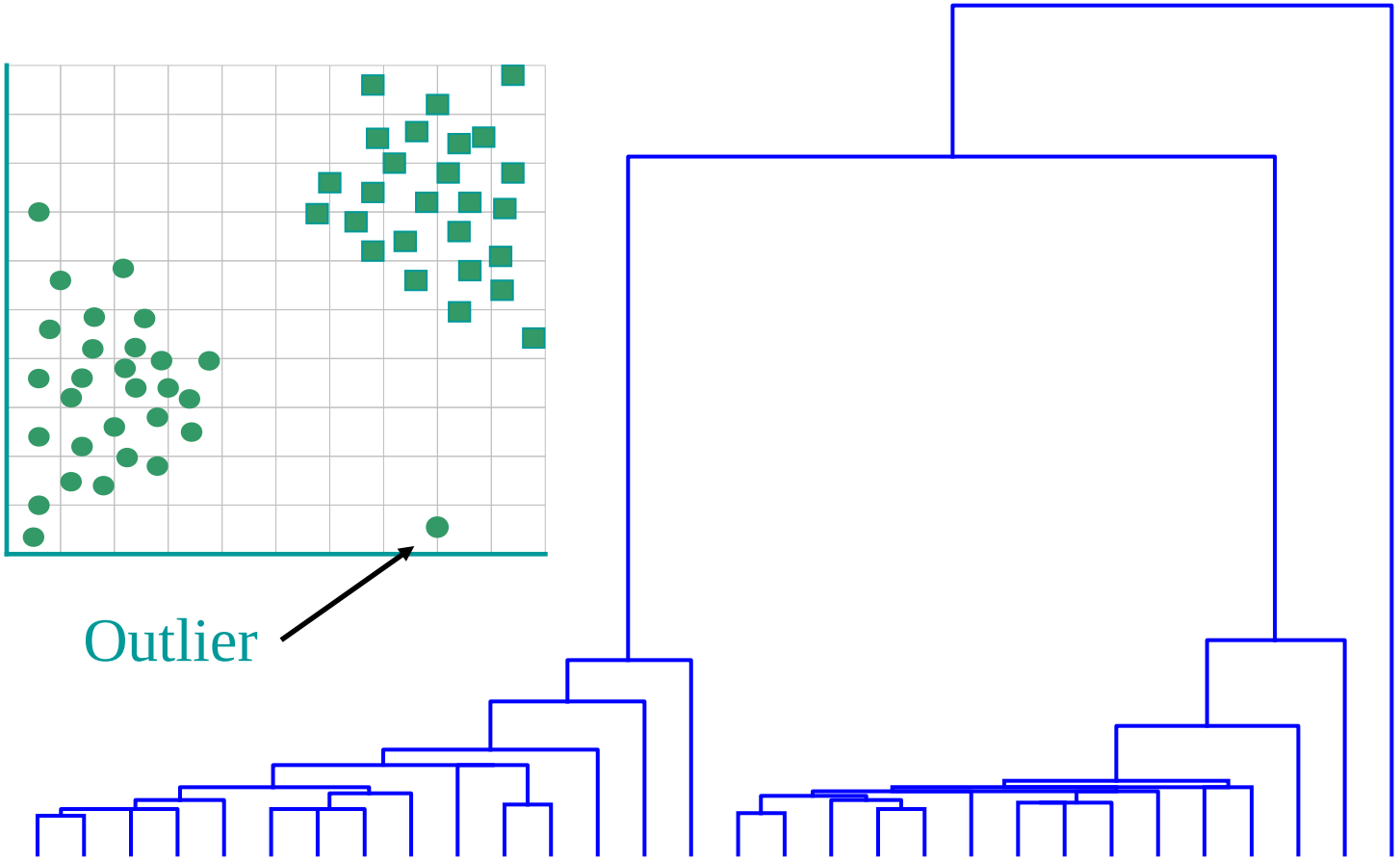
DENDROGRAMA: La similitud entre dos objetos viene dada por la “altura” del nodo común más cercano.

Clustering jerárquico



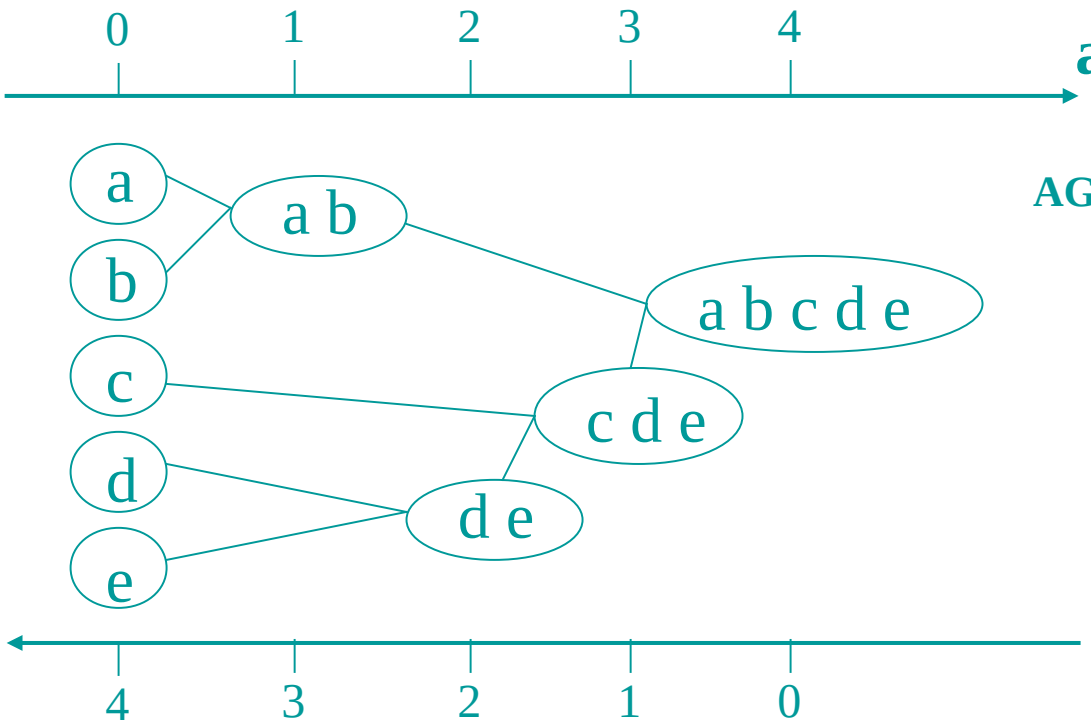
El DENDROGRAMA nos puede ayudar a determinar el número adecuado de agrupamientos (aunque normalmente no será tan fácil).

Clustering jerárquico



El DENDROGRAMA
también nos puede servir para detectar outliers.

Clustering jerárquico



aglomerativo (AGNES)
AGglomerative NESting

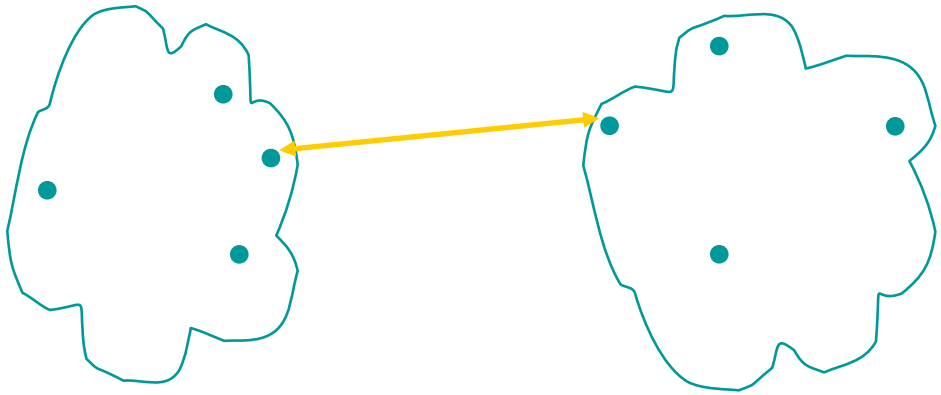
divisivo (DIANA)
Divisive ANAlysis

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

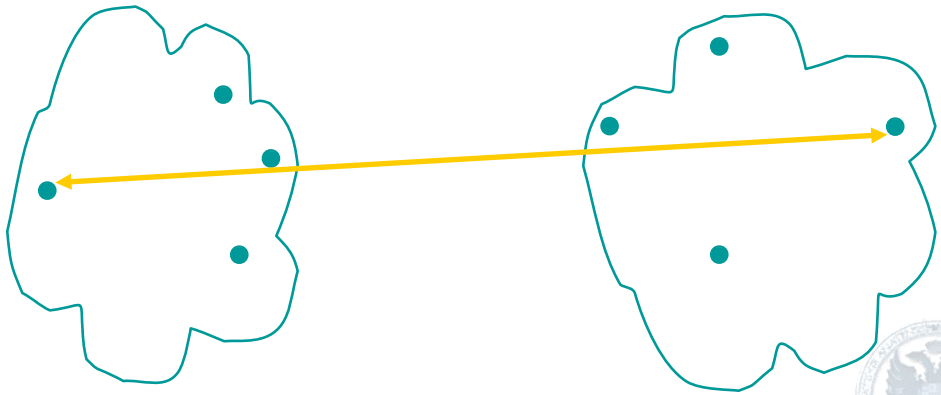
Clustering jerárquico

¿Cómo medir la distancia entre clusters?

- MIN
  single-link



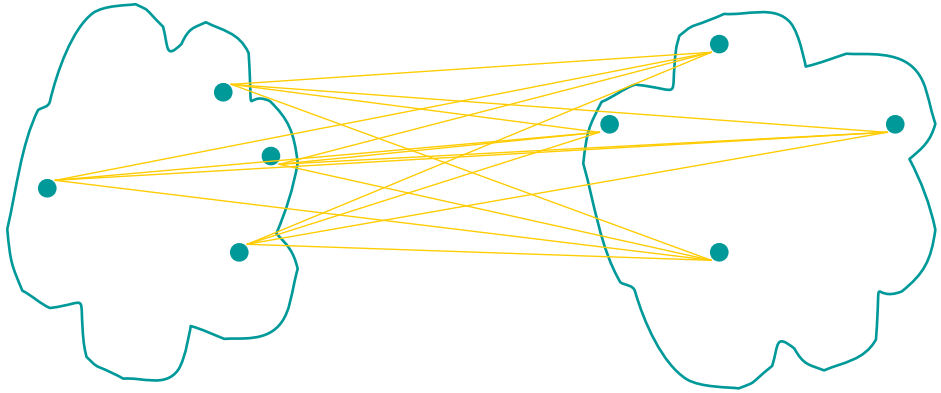
- MAX
  complete
  linkage
  (diameter)



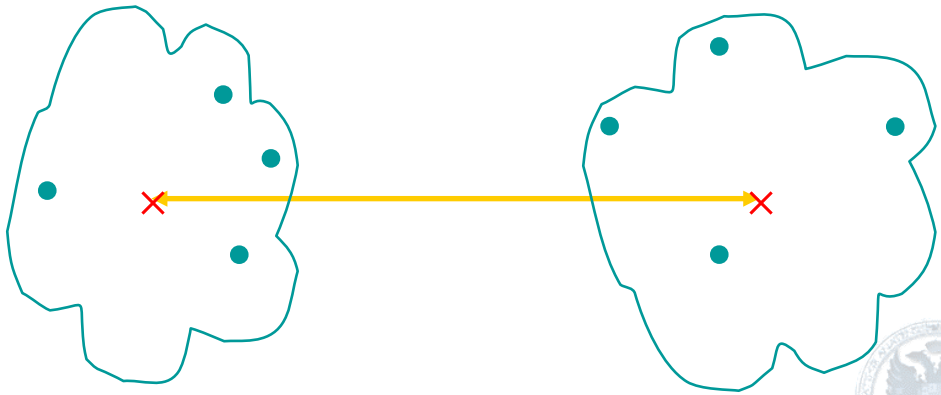
Clustering jerárquico

¿Cómo medir la distancia entre clusters?

- Promedio



- Centroides
  p.ej. BIRCH



Clustering jerárquico

Ejercicio

Utilizar un algoritmo aglomerativo de clustering jerárquico para agrupar los datos descritos por la siguiente matriz de distancias:

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

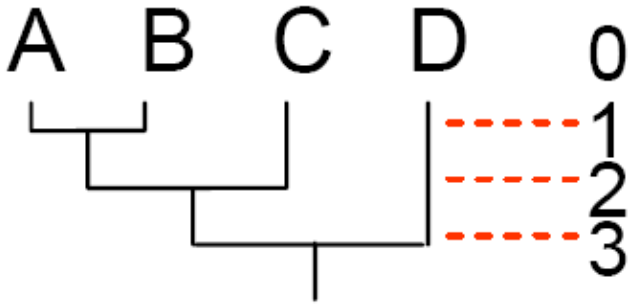
Variantes:

- Single-link (mínima distancia entre agrupamientos)
- Complete-link (máxima distancia entre agrupamientos)

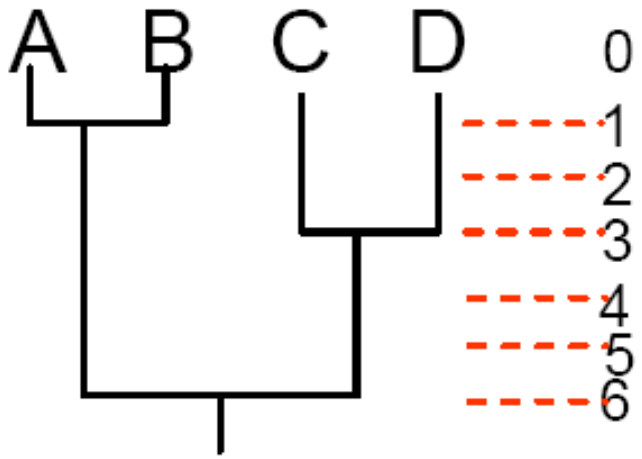
Clustering jerárquico

Ejercicio resuelto

Single-link



Complete-link



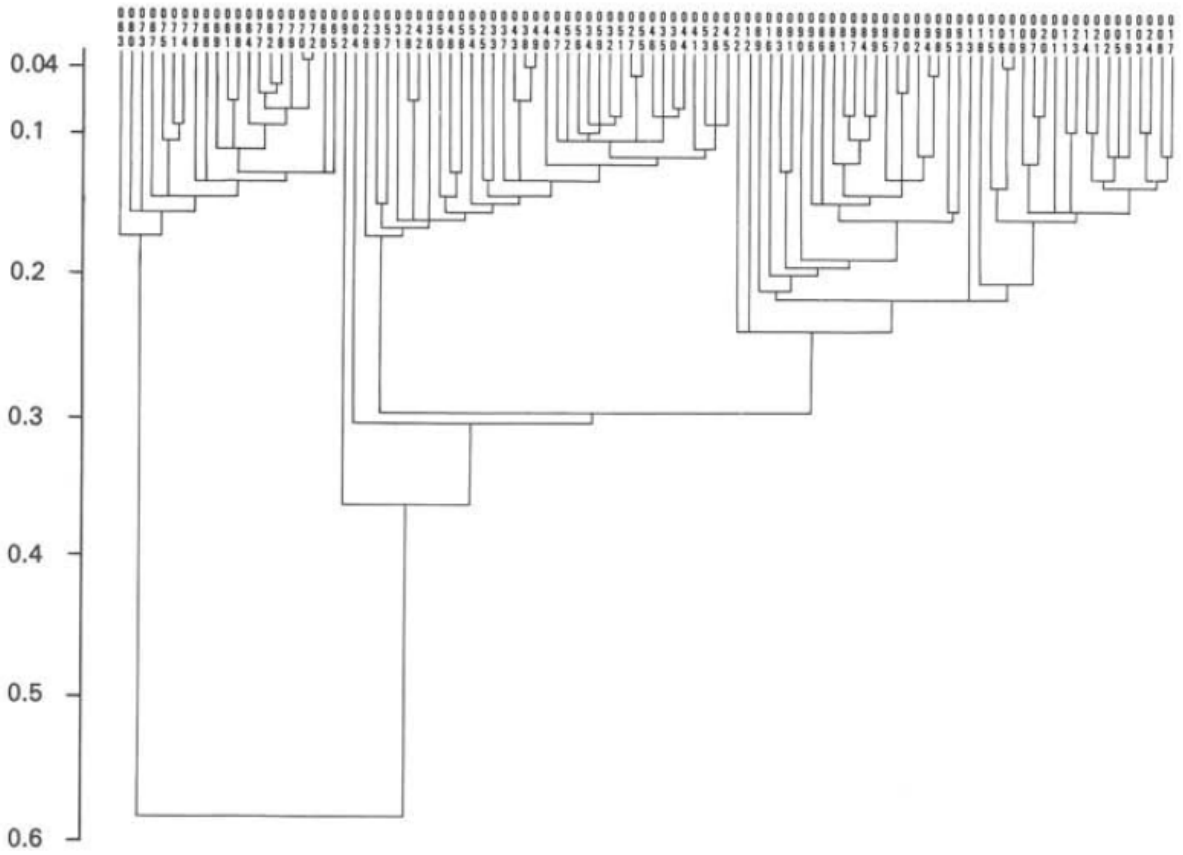
Clustering jerárquico

DEMO: Algoritmo aglomerativo

http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/AppletH.html

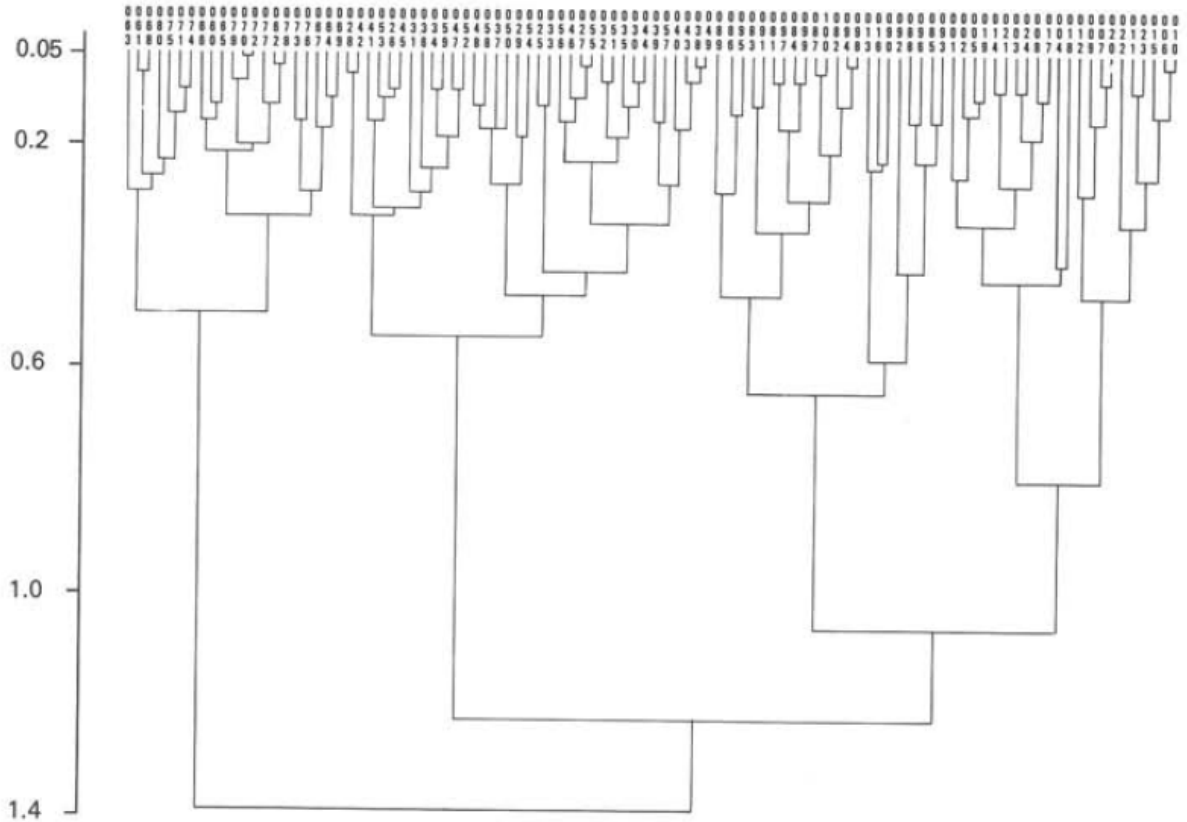
Clustering jerárquico

Datos sintéticos (4 clusters): Single-link



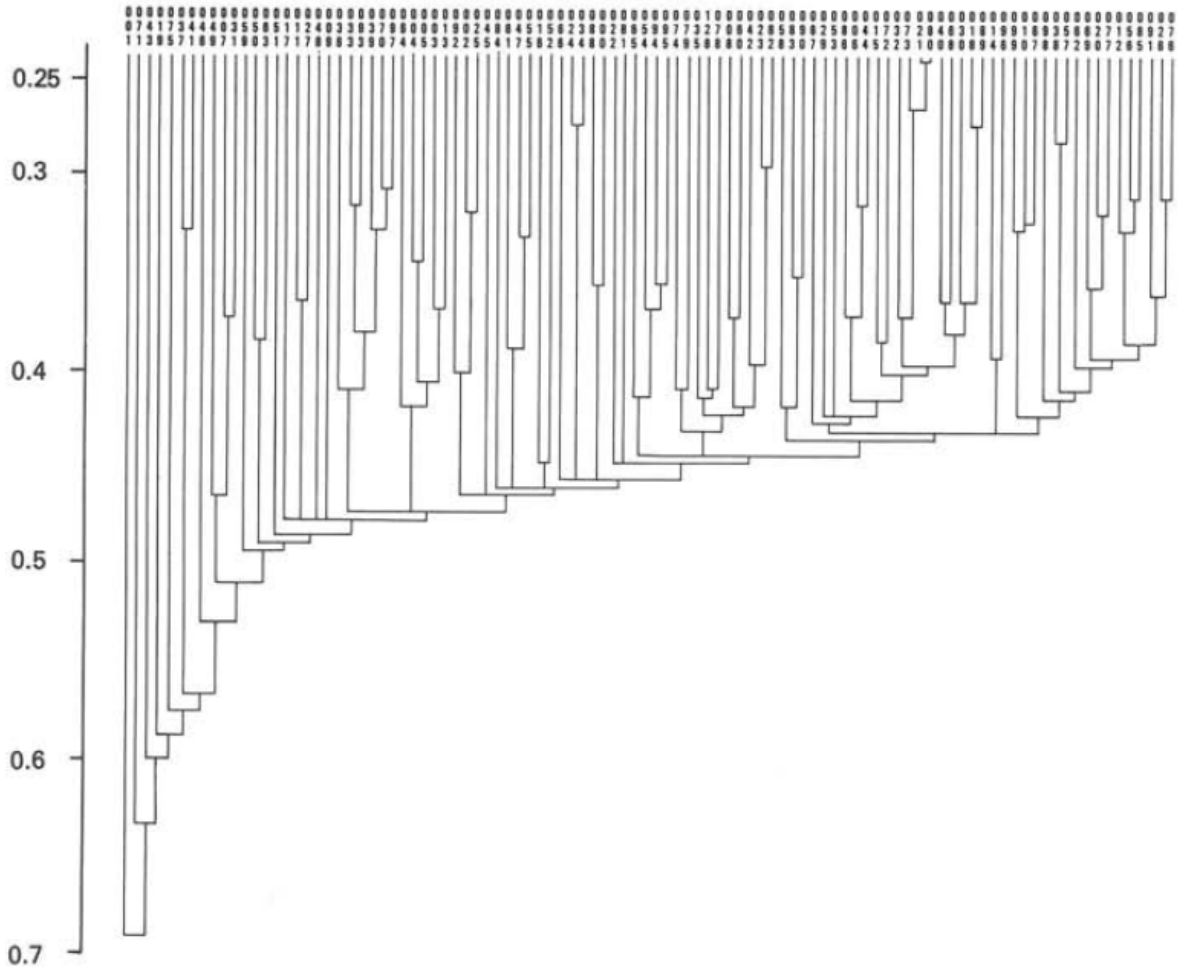
Clustering jerárquico

Datos sintéticos (4 clusters): Complete-link



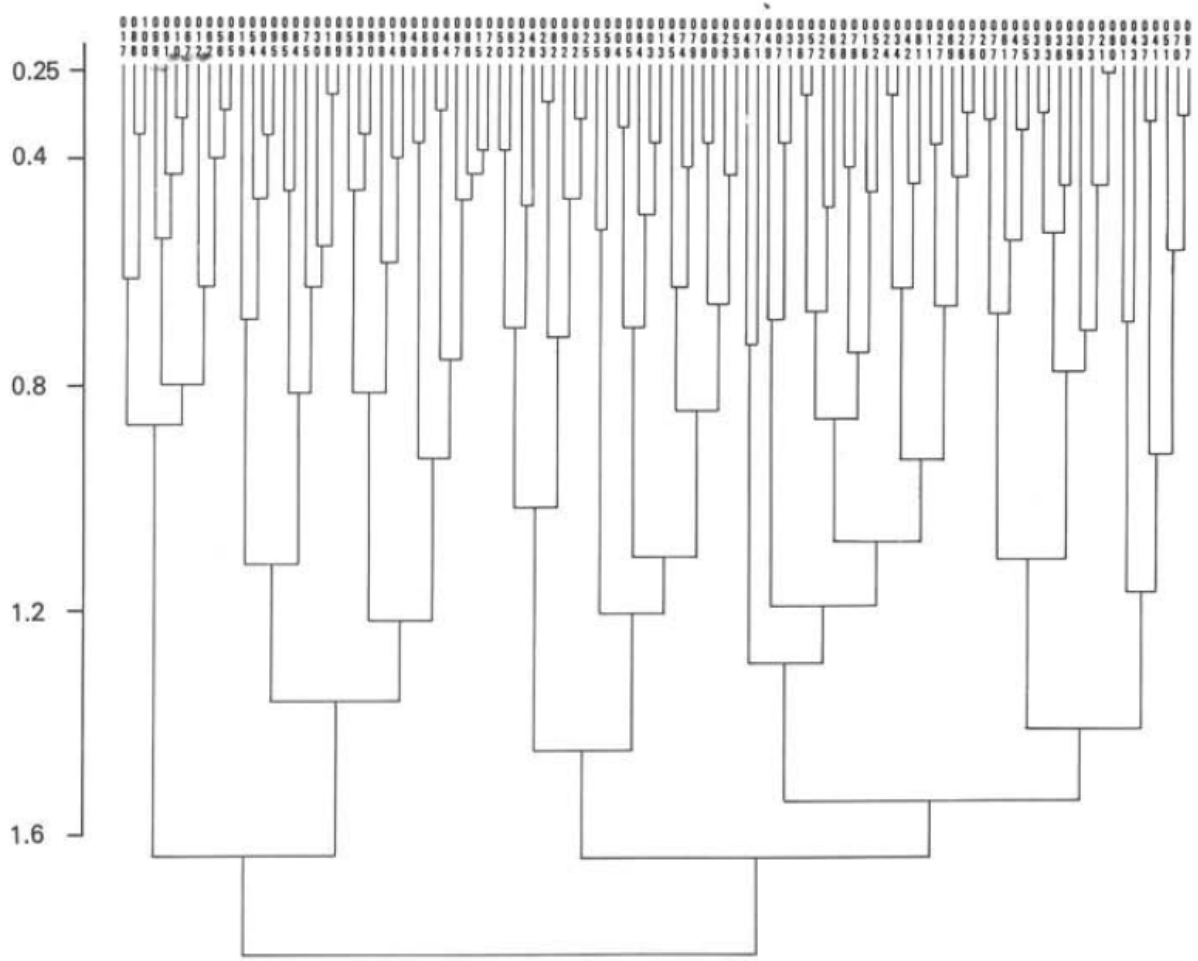
Clustering jerárquico

Datos sintéticos (aleatorios): Single-link



Clustering jerárquico

Datos sintéticos (aleatorios): Complete-link



Clustering jerárquico

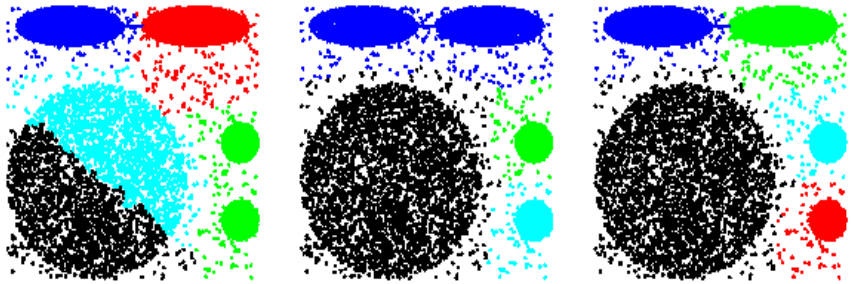
Principal inconveniente del clustering jerárquico:

Baja escalabilidad ≥ O(n²)

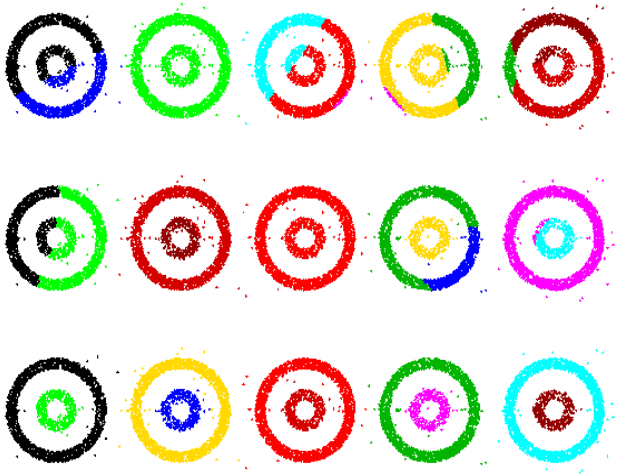
Algoritmos “escalables”:

- **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

Clustering jerárquico



a) BIRCH b) MST METHOD c) CURE



a) BIRCH

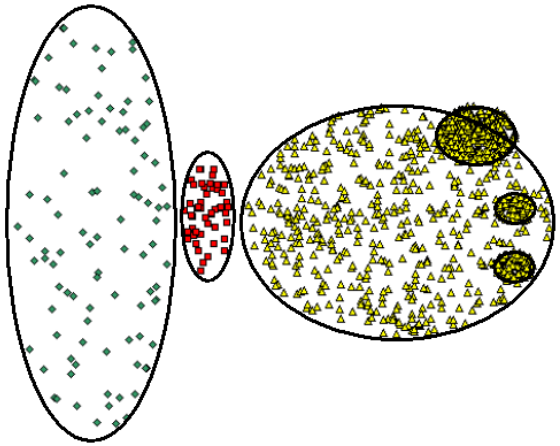
b) MST METHOD

c) CURE

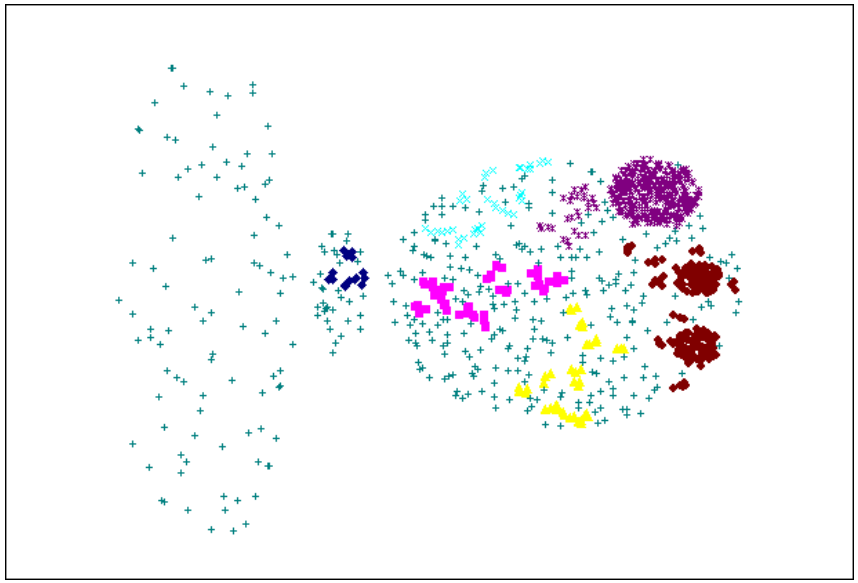
CURE

Clustering jerárquico

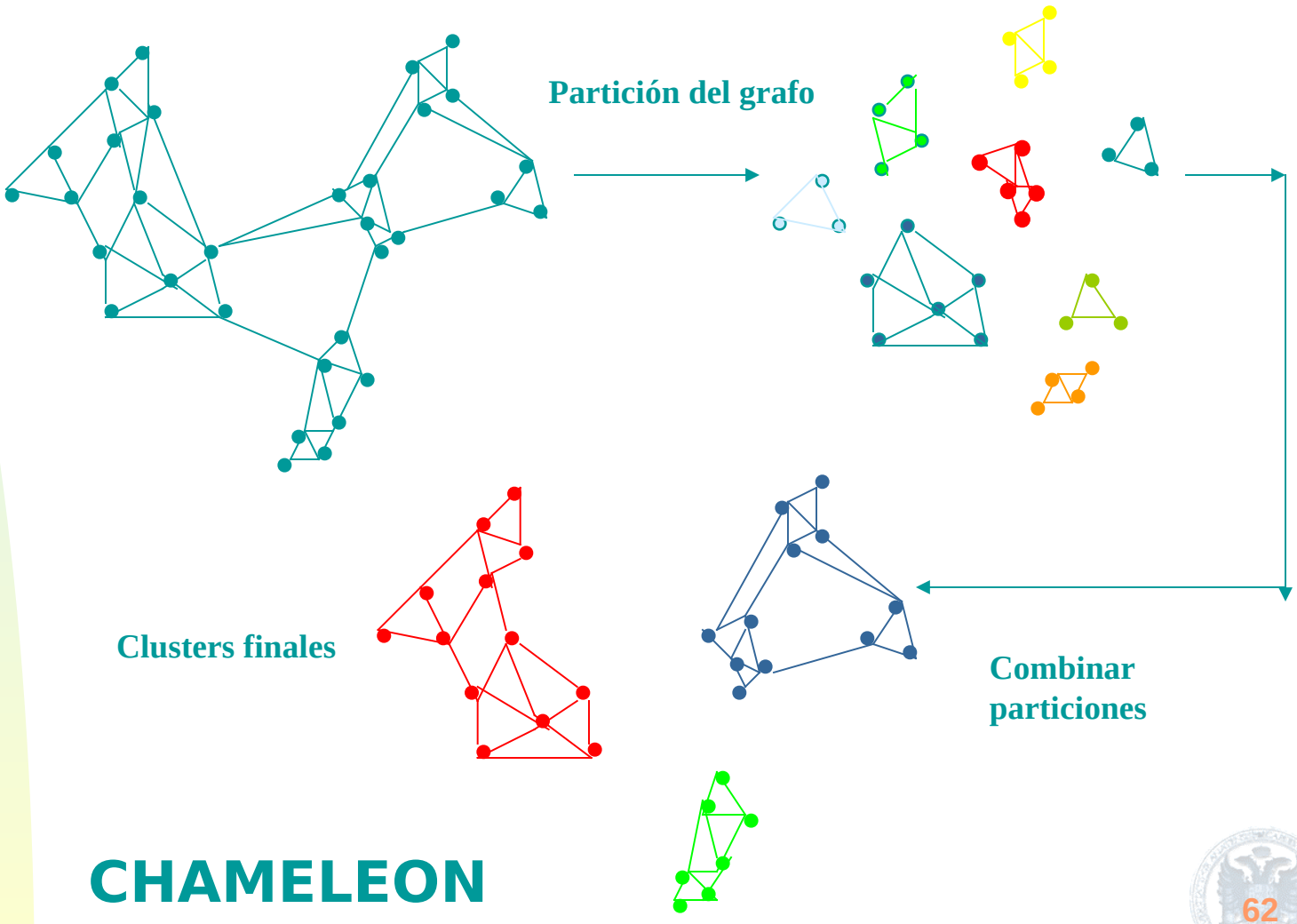
Agrupamientos con distintas densidades



CURE

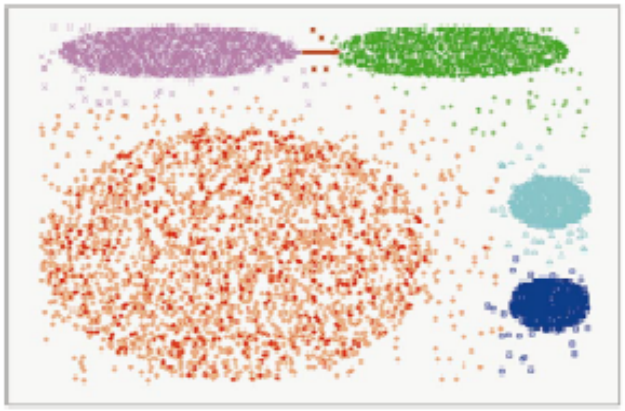


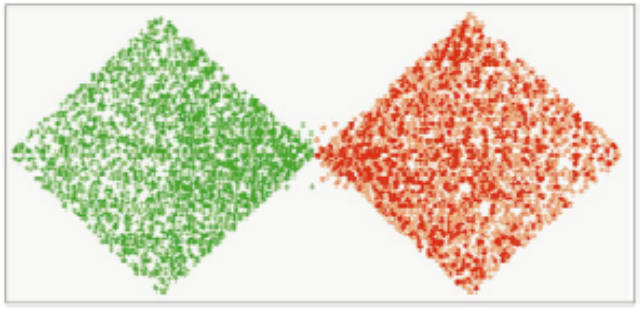
Clustering jerárquico

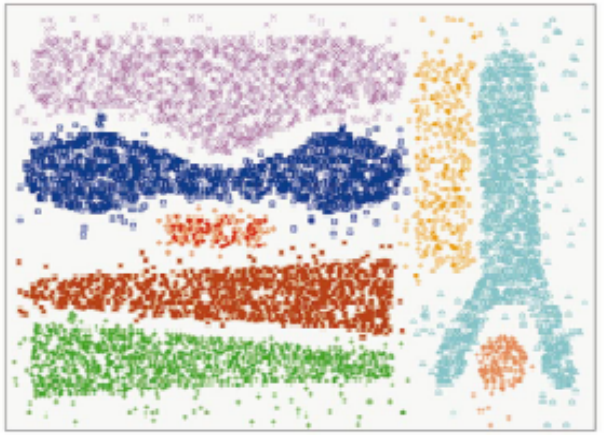


CHAMELEON

Clustering jerárquico







CHAMELEON

Density-based Clustering

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

Density-based Clustering

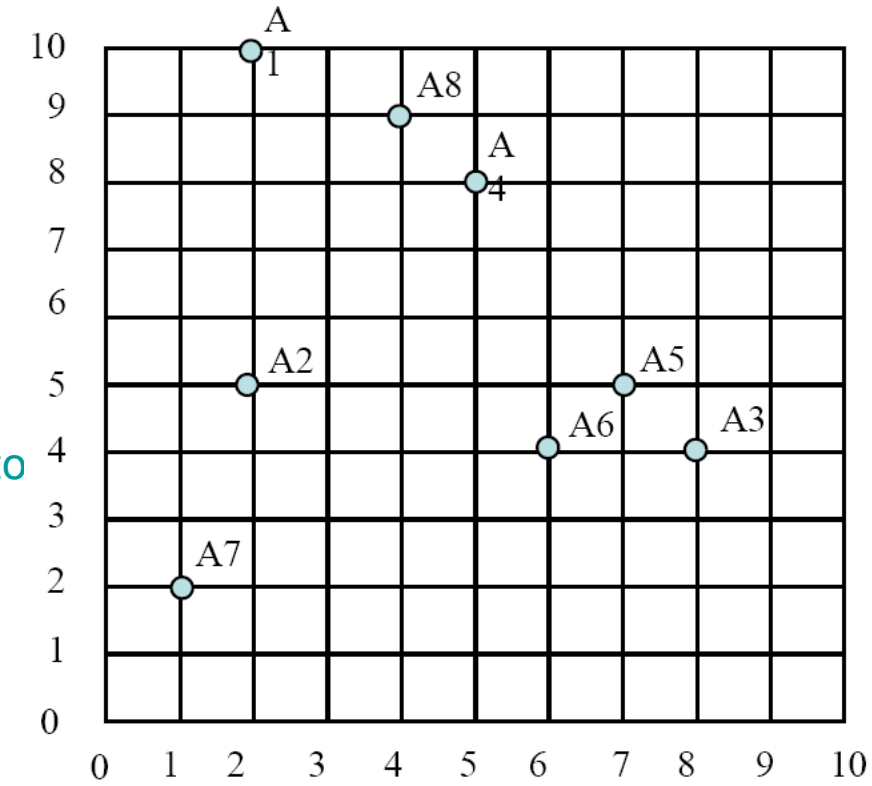
Algoritmos

- **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)

Density-based Clustering

Ejercicio

Agrupar los 8 puntos de la figura utilizando el algoritmo DBSCAN.



Número mínimo de punto en el “vecindario”:

MinPts = 2

Radio del “vecindario”:

Epsilon $\sqrt{2} \Rightarrow \sqrt{10}$

Density-based Clustering

Ejercicio resuelto

Distancia euclídea

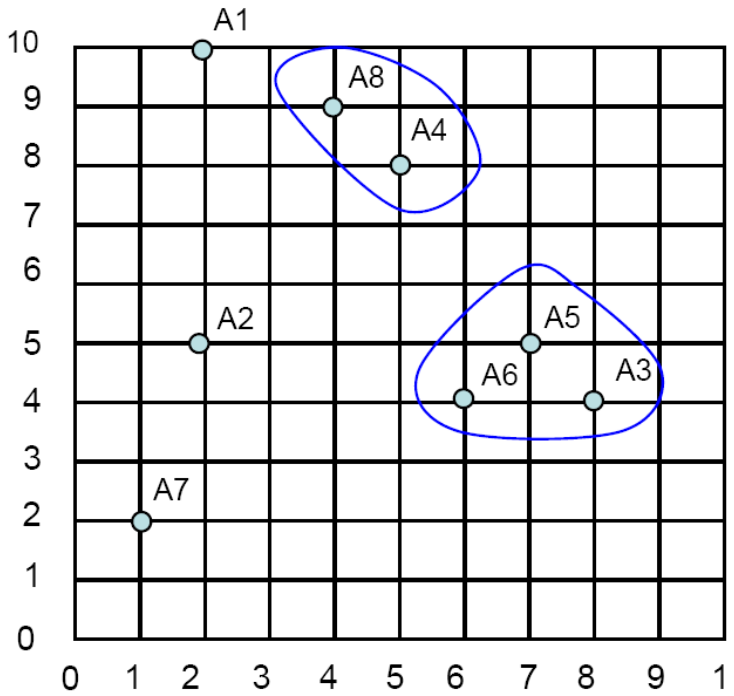
	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

Density-based Clustering

Ejercicio resuelto

Epsilon = $\sqrt{2}$

A1, A2 y A7 no tienen vecinos en su vecindario,
por lo que se consideran “outliers” (no están en zonas densas):



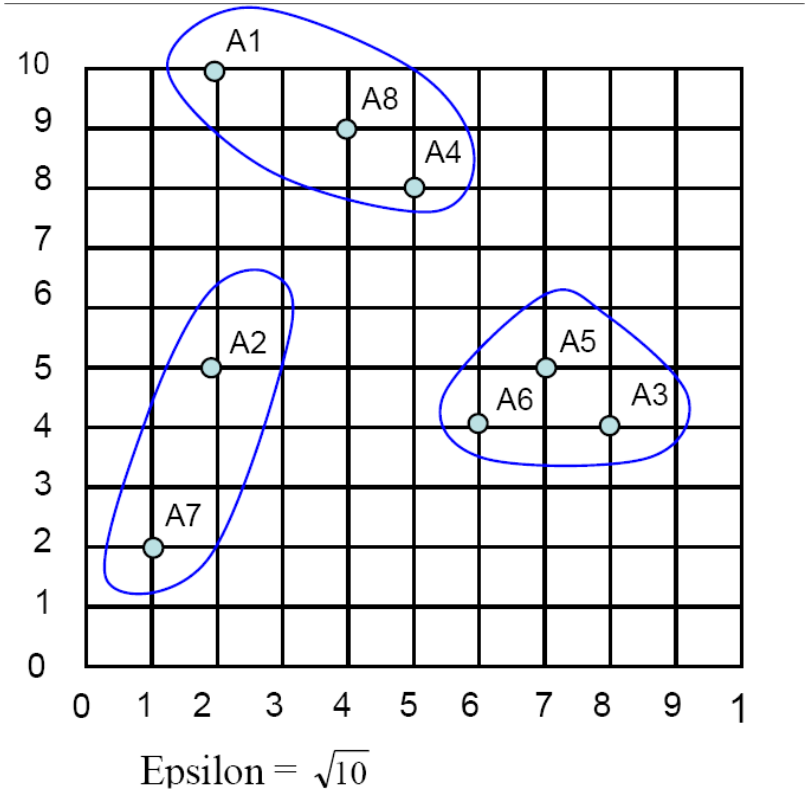
Epsilon = 2

Density-based Clustering

Ejercicio resuelto

Epsilon = $\sqrt{10}$

Al aumentar el valor del parámetro *Epsilon*,
el vecindario de los puntos aumenta y todos quedan agrupados:



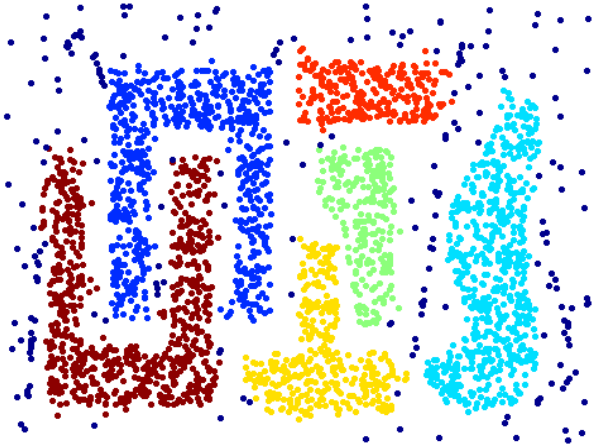
Epsilon = $\sqrt{10}$

Density-based Clustering

DEMO: DBSCAN et al.

http://www.cs.ualberta.ca/~yaling/Cluster/Applet/Code/Cluster.html

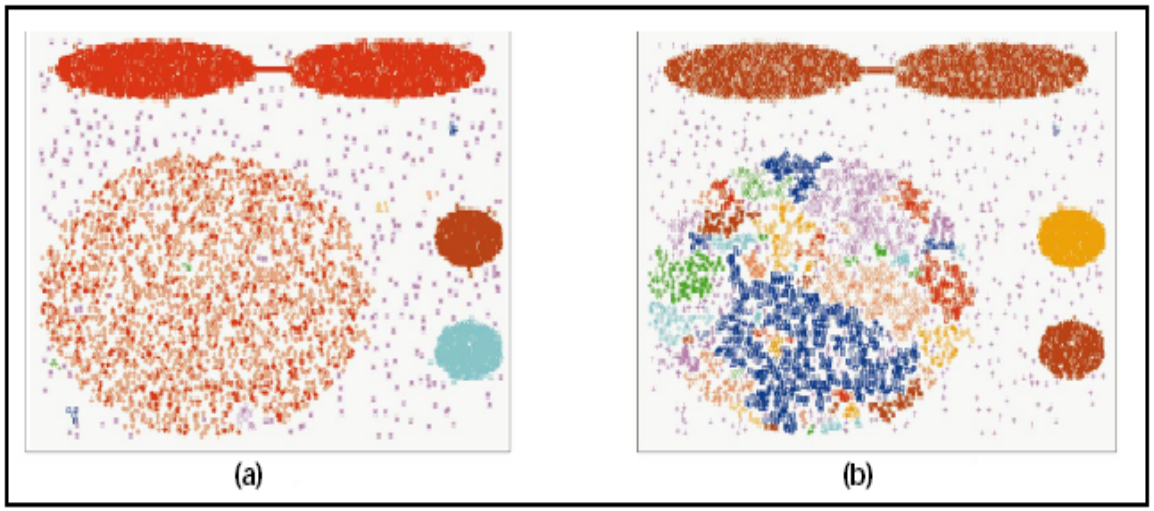
Density-based Clustering

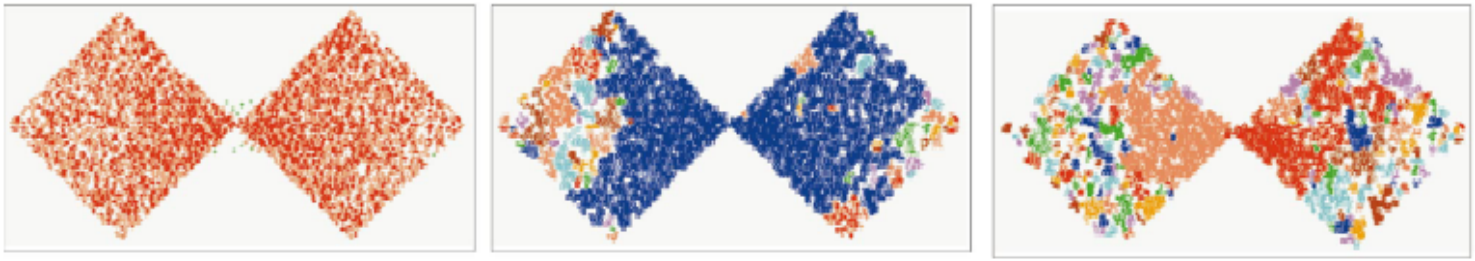


Clusters

DBSCAN ... cuando funciona bien

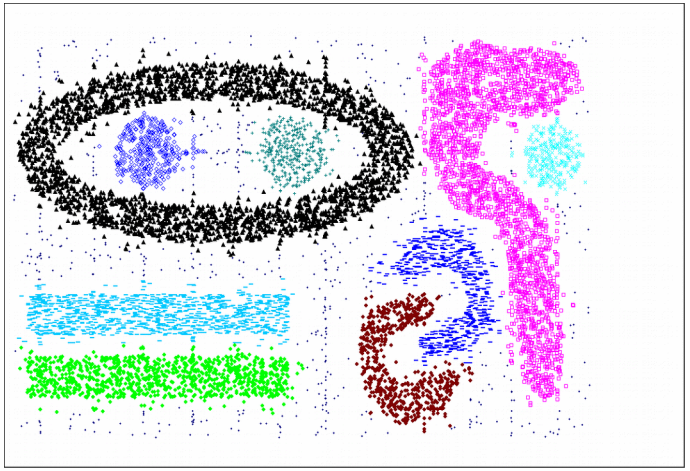
Density-based Clustering

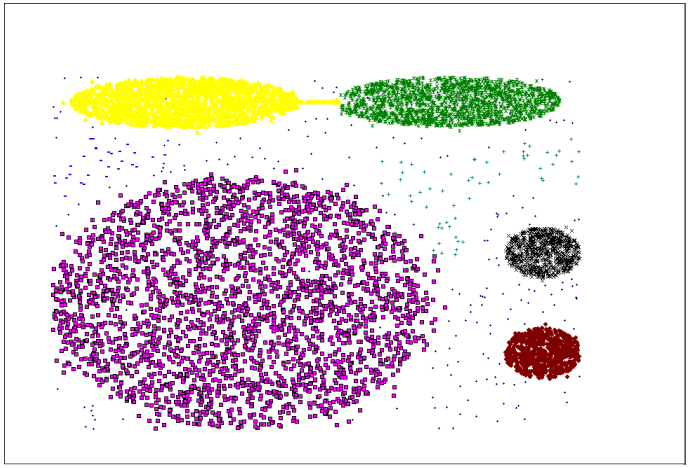


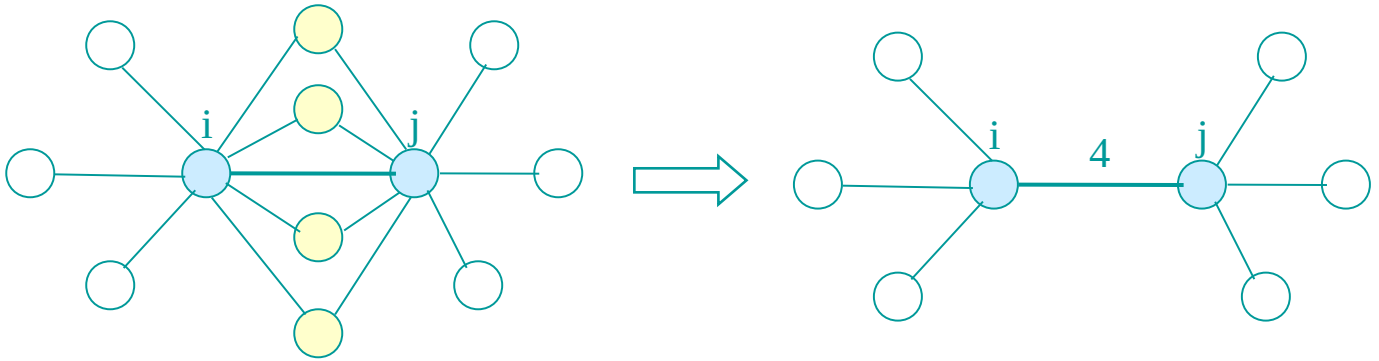


DBSCAN sensible al valor inicial de sus parámetros

Density-based Clustering



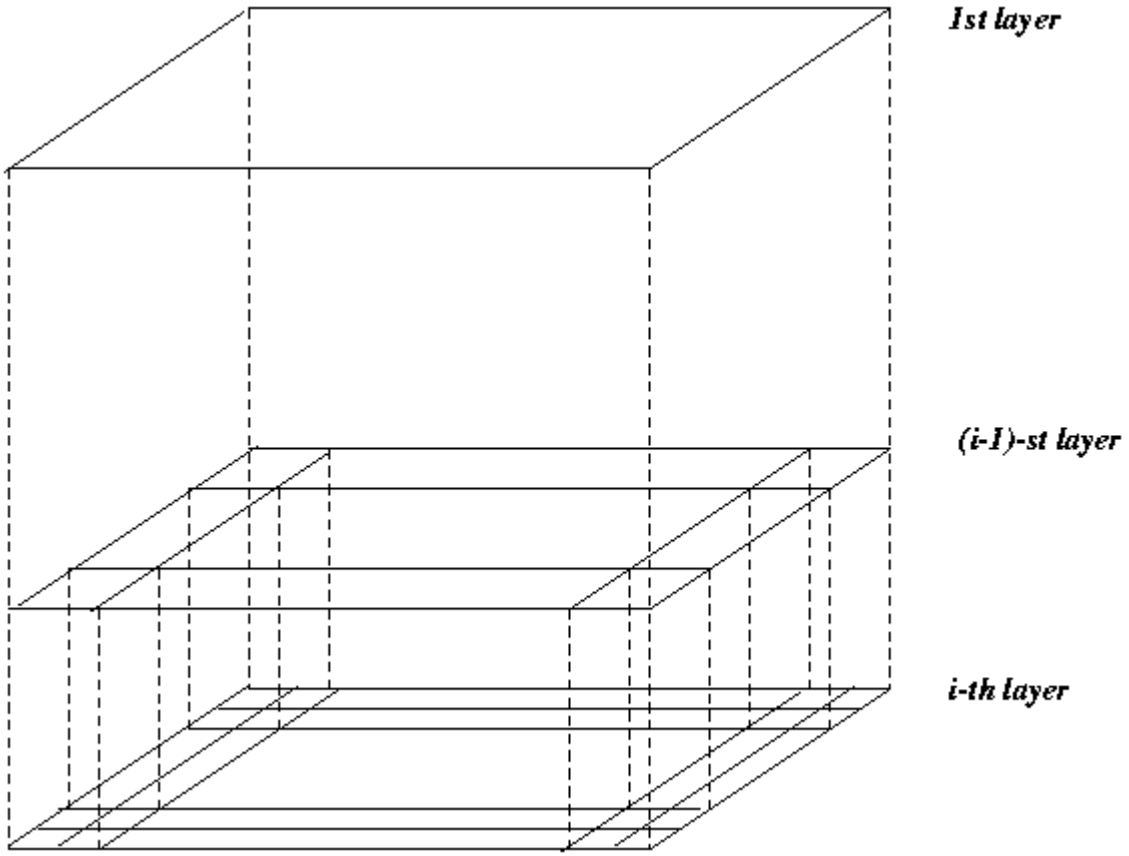




SNN density-based clustering… $O(n^2)$

Otros métodos

Grids multiresolución



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)

Otros métodos

Clustering basado en modelos

Ajustar los datos a un modelo matemático

Se supone que los datos provienen de la superposición de varias distribuciones de probabilidad.

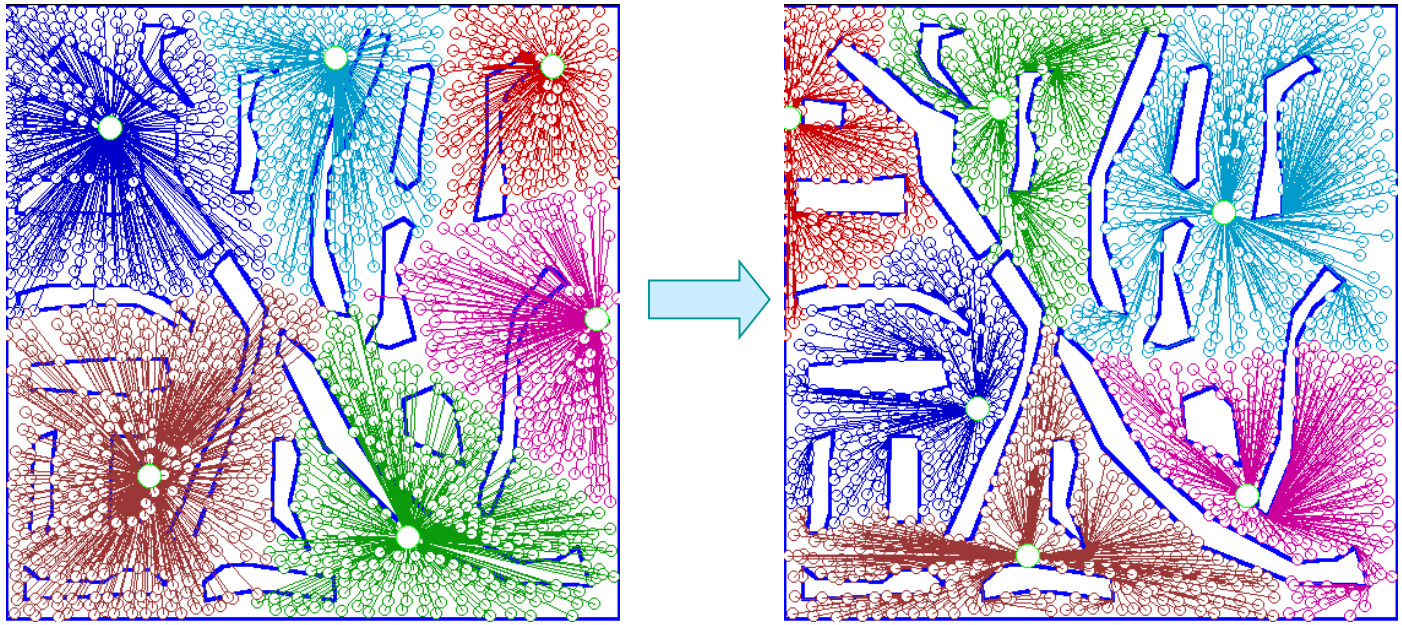
Algoritmos

- Estadística:
  EM [Expectation Maximization]**, AutoClass**
- Clustering conceptual (Machine Learning):
  COBWEB, CLASSIT
- Redes neuronales:
  SOM [Self-Organizing Maps]

Otros métodos

Clustering con restricciones

p.ej. Clustering con obstáculos



Posibles aplicaciones:

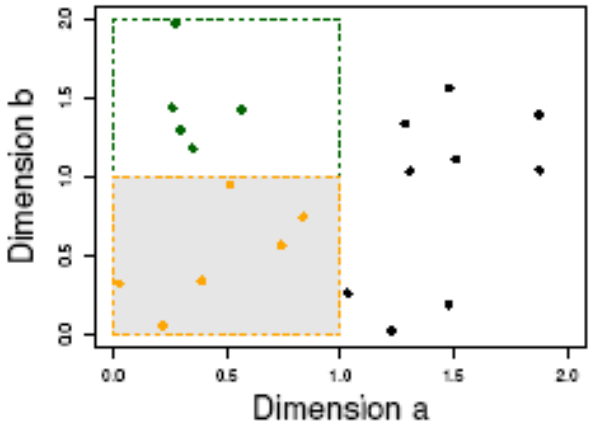
Distribución de cajeros automáticos/supermercados...

Subspace clustering

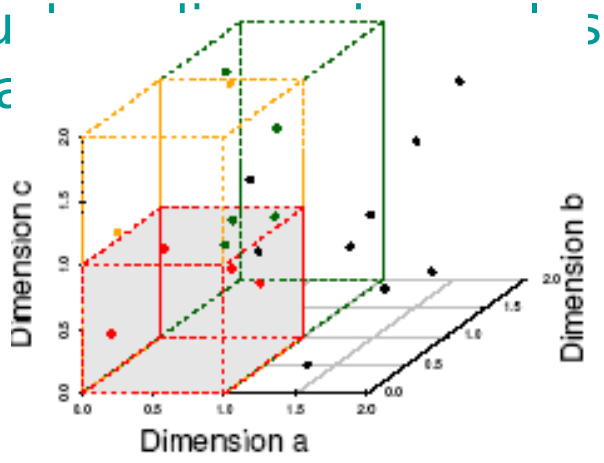
La dimensionalidad de los datos

¿Por qué es un problema?

- Los datos en una dimensión están relativamente cerca
- Al añadir una nueva dimensión, los datos se alejan.
- Cuando tenemos mu[illegible]cia[illegible]s



(b) 6 Objects in One Unit Bin



(c) 4 Objects in One Unit Bin

Clustering

IdBIS

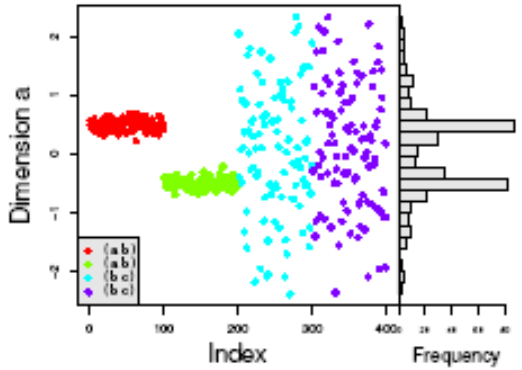
Subspace clustering

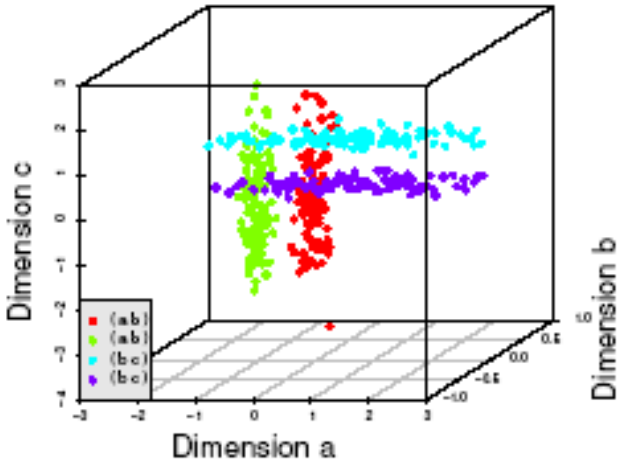
La dimensionalidad de los datos

Soluciones

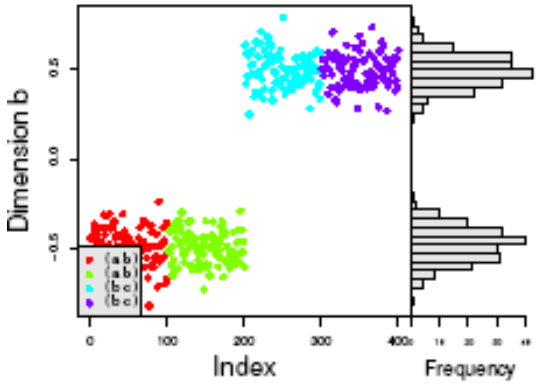
- **Transformación de características** (PCA, SVD)
  útil sólo si existe correlación/redundancia
- **Selección de características** (wrapper/filter)
  útil si se pueden encontrar clusters en subespacios
- **"Subspace clustering"**
  Buscar clusters en todos los subespacios posibles.

Subspace clustering

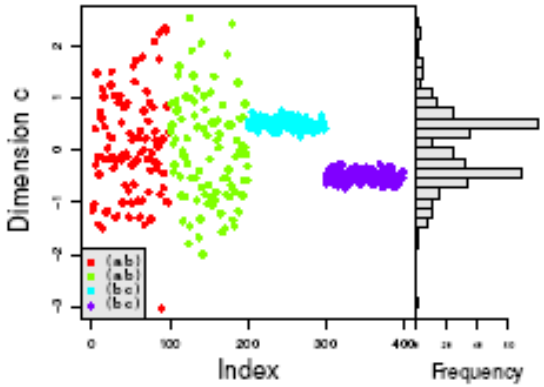




(a) Dimension a



(b) Dimension b

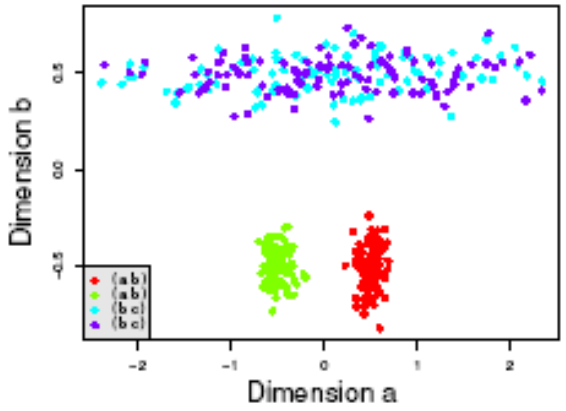


(c) Dimension c

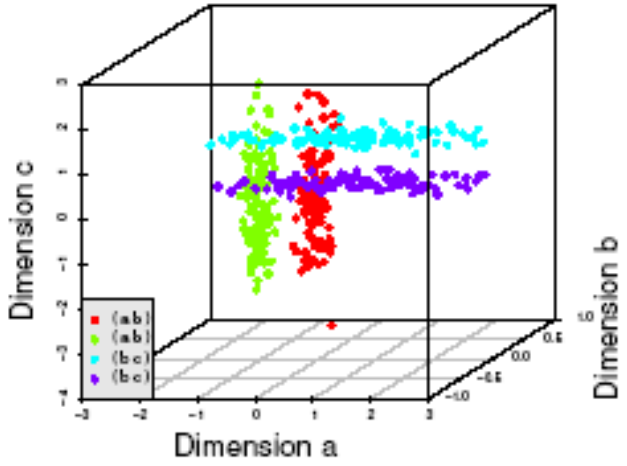
Clustering

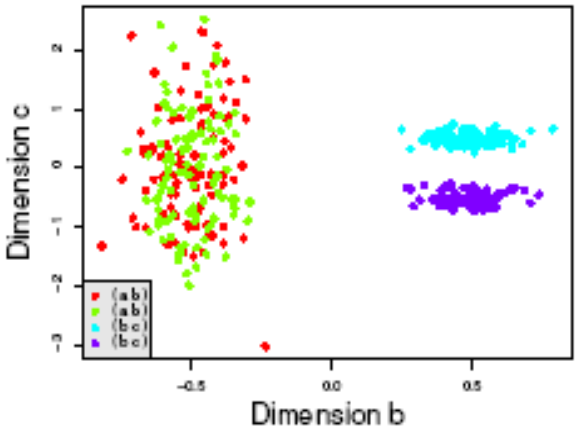
IdBIS

Subspace clustering

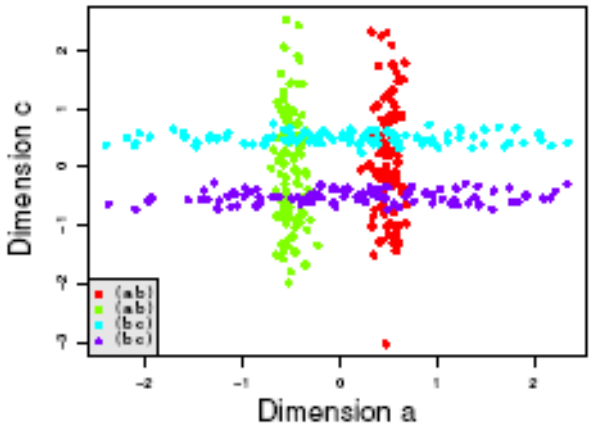


(a) Dims a & b





(b) Dims b & c



(c) Dims a & c

Subspace clustering

DEMO: CLIQUE et al.

http://www.cs.ualberta.ca/~yaling/Cluster/Applet/Code/Cluster.html

Validación

¿Cómo se puede evaluar
la calidad de los clusters obtenidos?

Depende de lo que estemos buscando...

Hay situaciones en las que nos interesa:

- Evitar descubrir clusters donde sólo hay ruido.
- Comparar dos conjuntos de clusters alternativos.
- Comparar dos técnicas de agrupamiento

Validación

- **Criterios externos**
  (aportando información adicional)

  p.ej. entropía/pureza (como en clasificación)

- **Criterios internos**
  (a partir de los propios datos),

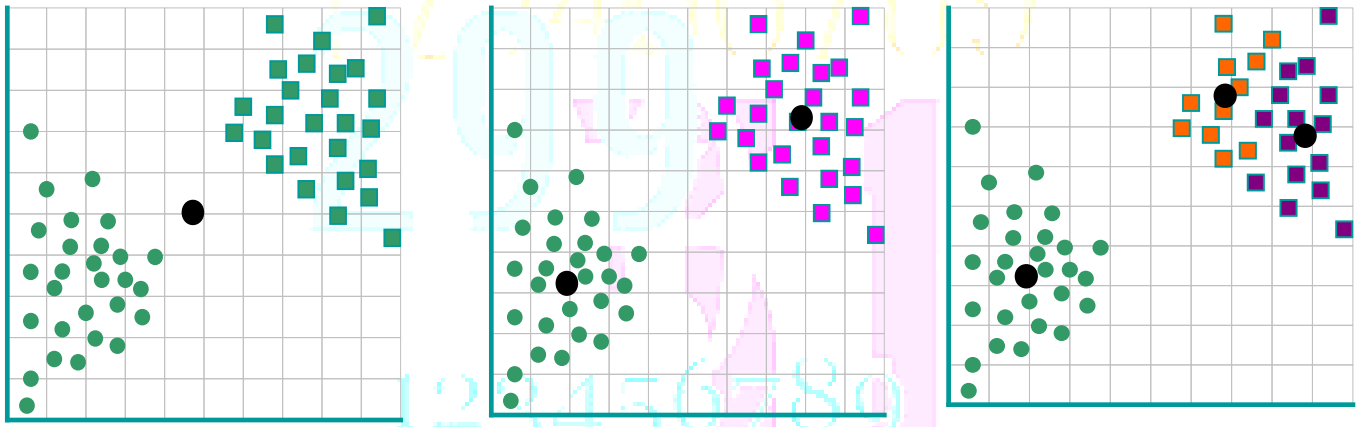
  p.ej. SSE (“Sum of Squared Error”)
  - para comparar clusters
  - para estimar el número de clusters

  Otras medidas:
  cohesión, separación, coeficientes de silueta…

Validación

¿Cuál es el número adecuado de agrupamientos?

p.ej. SSE ("Sum of Squared Error")



k = 1
J = 873.0

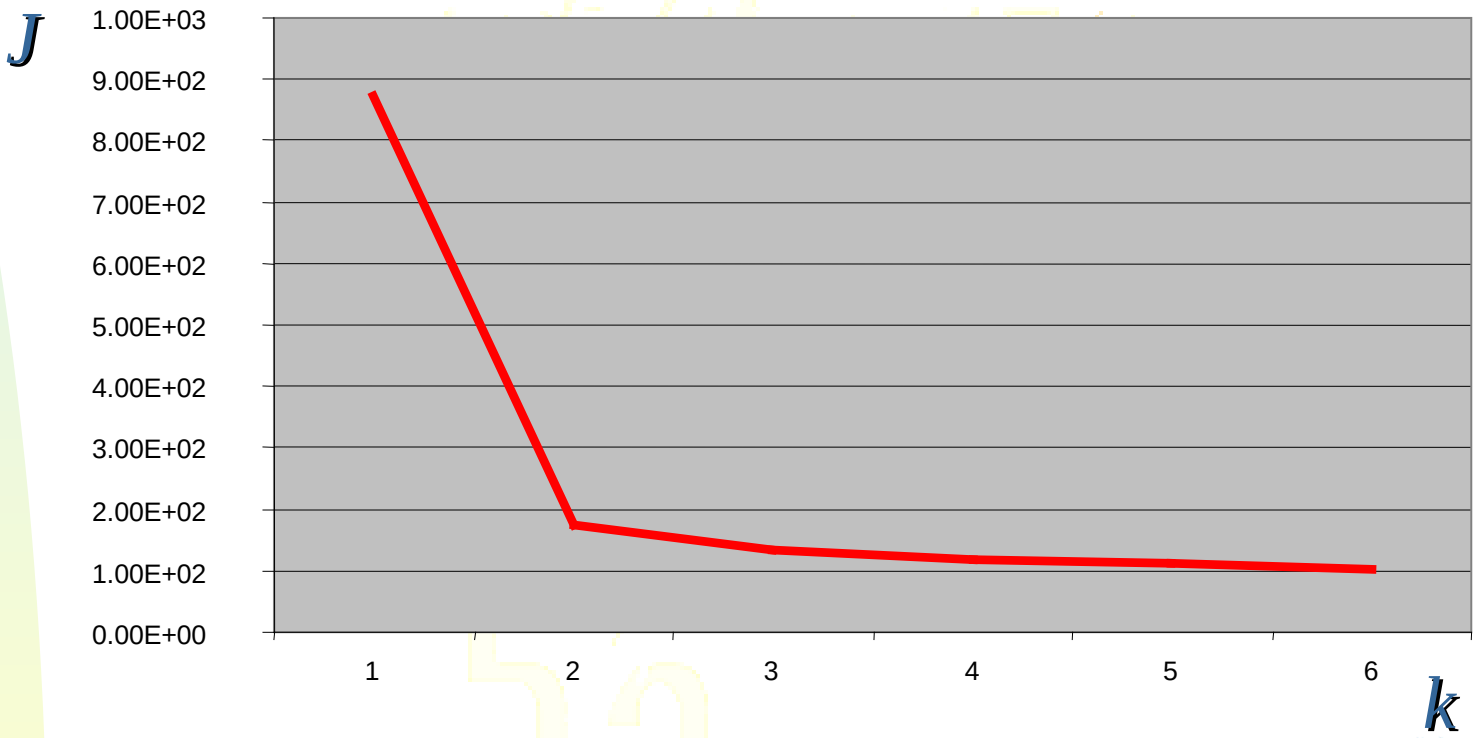
k = 2
J = 173.1

k = 3
J = 133.6

Validación

¿Cuál es el número adecuado de agrupamientos?

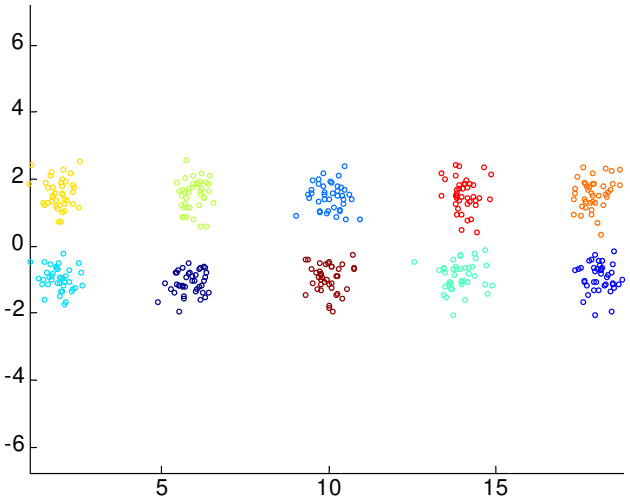
p.ej. SSE (“Sum of Squared Error”)

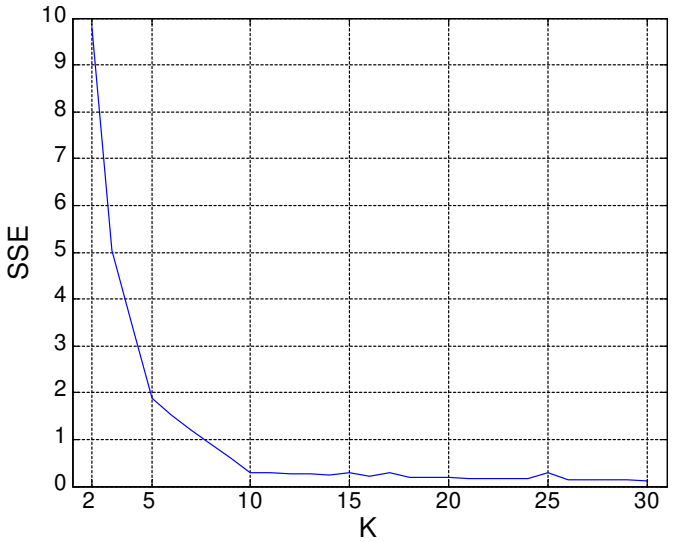


El codo en k=2 sugiere que éste es el valor

adecuado para el número de

Validación

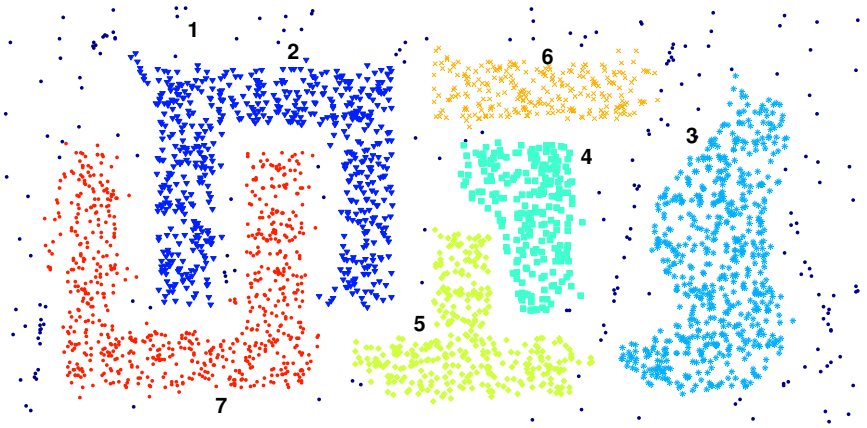


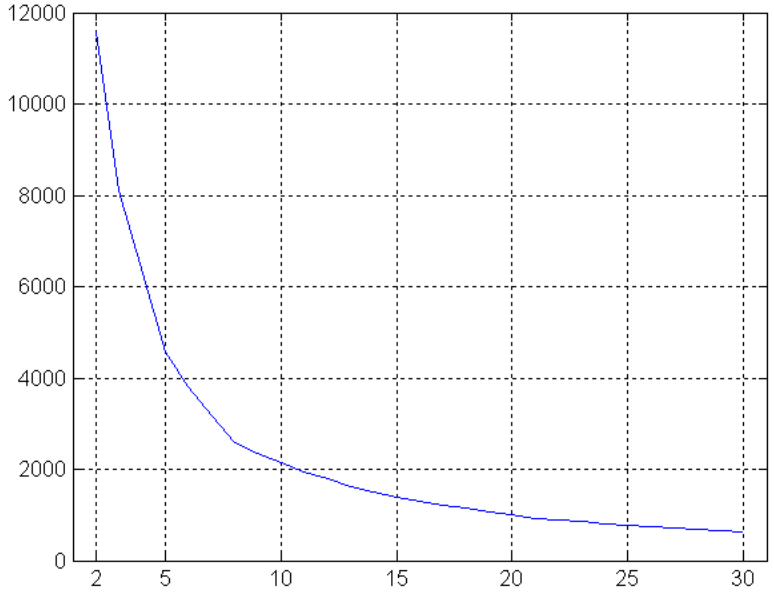


Clustering

IdBIS

Validación

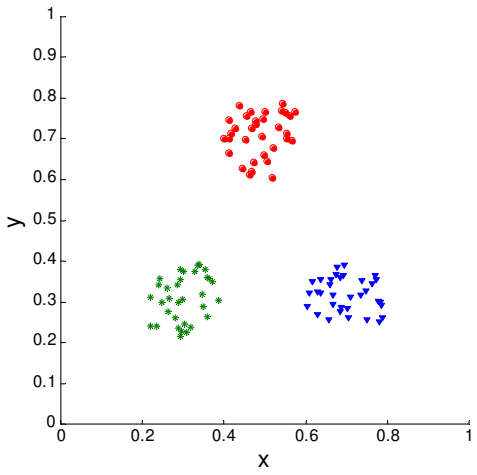


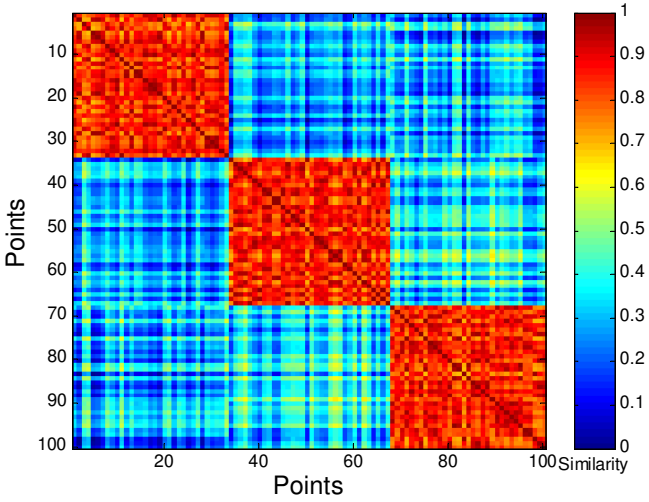


Validación

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...





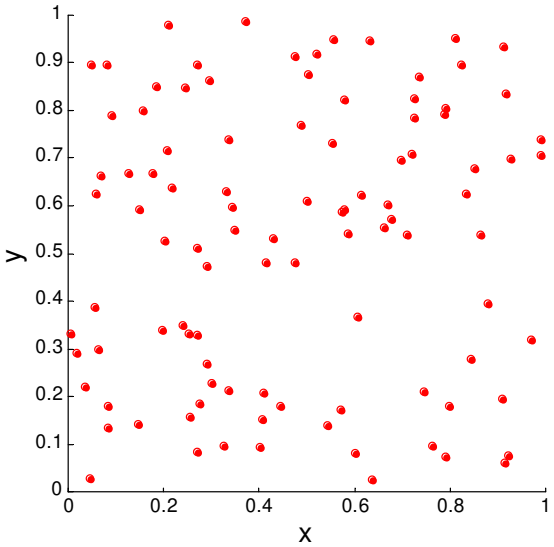
Clustering

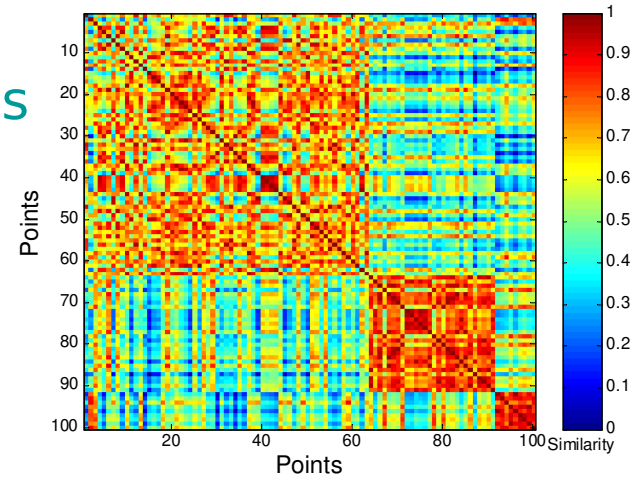
IdBIS

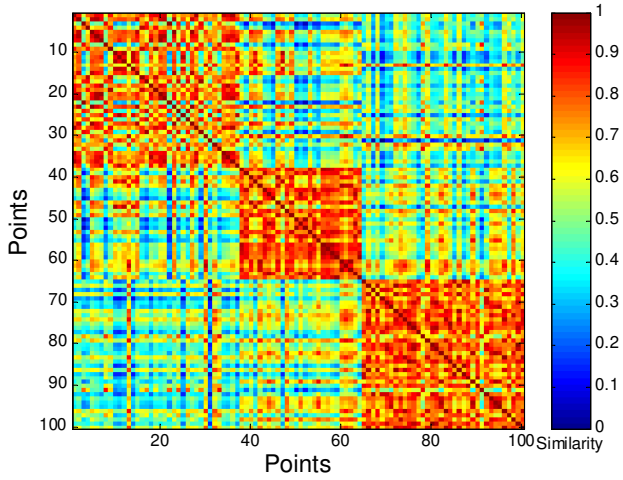
Validación

Matriz de similitud

Clusters en datos aleatorios (DBSCAN y k-Means)







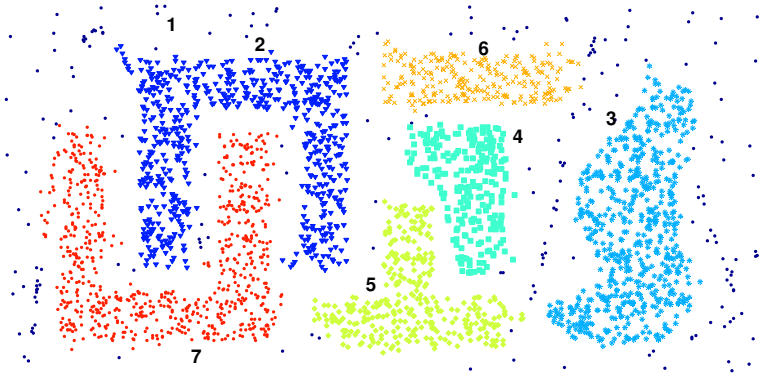
Clustering

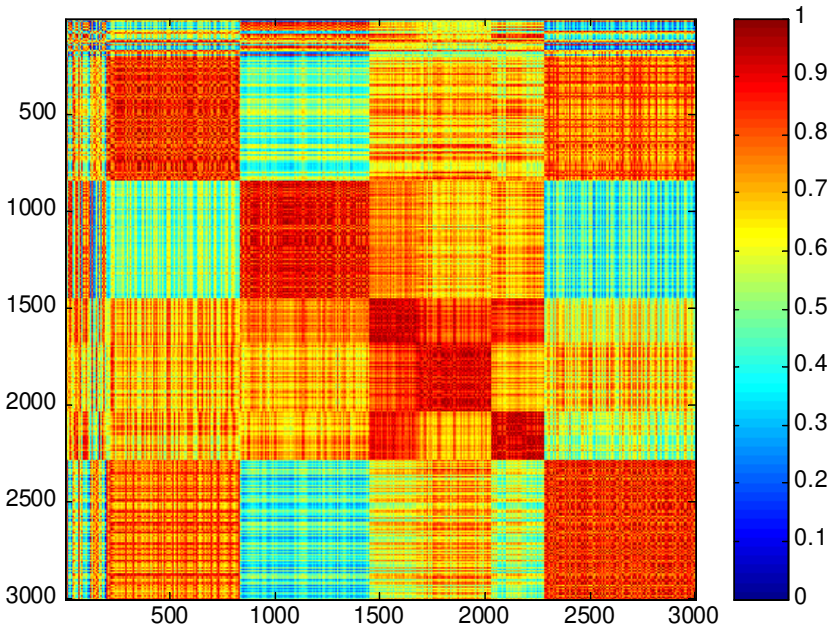
IVBIS

Validación

Matriz de similit

DBSCAN





Clustering

IdBIS

Bibliografía

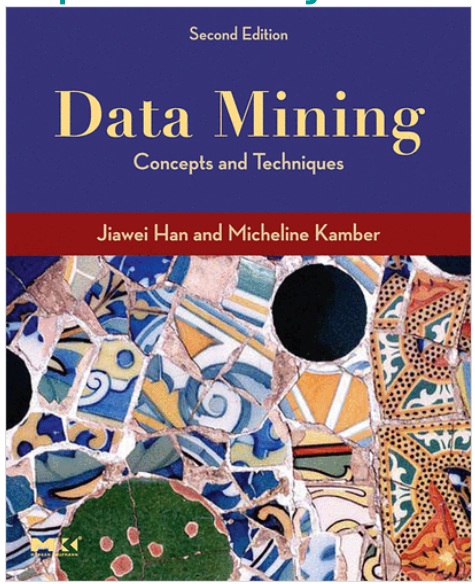
- R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. **Automatic subspace clustering of high dimensional data for data mining applications**. SIGMOD'98
- M. Ankerst, M. Breunig, H.-P. Kriegel, and J. Sander. **Optics: Ordering points to identify the clustering structure**, SIGMOD'99.
- L. Ertöz, M. Steinbach, and V. Kumar. **Finding clusters of different sizes, shapes, and densities in noisy, high-dimensional data**, SDM'2003
- M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. **A density-based algorithm for discovering clusters in large spatial databases**. KDD'96.
- D. Fisher. **Knowledge acquisition via incremental conceptual clustering**. Machine Learning, 2:139-172, 1987.
- D. Gibson, J. Kleinberg, and P. Raghavan. **Clustering categorical data: An approach based on dynamic systems**. VLDB'98
- S. Guha, R. Rastogi, and K. Shim. **Cure: An efficient clustering algorithm for large databases**. SIGMOD'98.
- S. Guha, R. Rastogi, and K. Shim. **ROCK: A robust clustering algorithm for categorical attributes**. In *ICDE'99*, Sydney, Australia, March 1999.

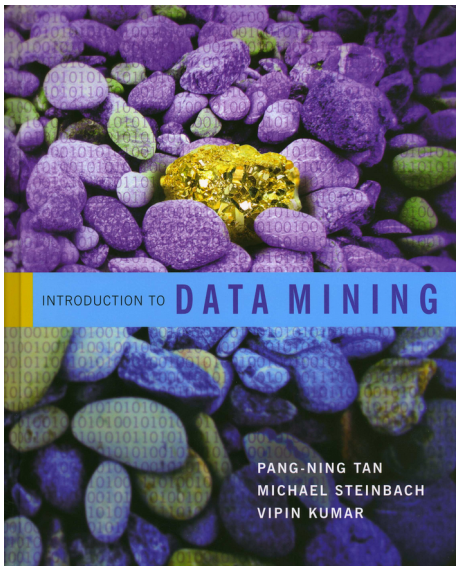
Bibliografía

- A. Hinneburg, D.l A. Keim: **An Efficient Approach to Clustering in Large Multimedia Databases with Noise**. KDD'98.
- G. Karypis, E.-H. Han, and V. Kumar. **CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling**. *COMPUTER, 32(8): 68-75, 1999.*
- L. Parsons, E. Haque and H. Liu, **Subspace Clustering for High Dimensional Data: A Review** , SIGKDD Explorations, 6(1), June 2004
- G. Sheikholeslami, S. Chatterjee, and A. Zhang. **WaveCluster: A multi-resolution clustering approach for very large spatial databases**. VLDB'98.
- A. K. H. Tung, J. Hou, and J. Han. **Spatial Clustering in the Presence of Obstacles** , *ICDE'01*
- H. Wang, W. Wang, J. Yang, and P.S. Yu. **Clustering by pattern similarity in large data sets**, *SIGMOD'* 02.
- W. Wang, Yang, R. Muntz, **STING: A Statistical Information grid Approach to Spatial Data Mining**, VLDB'97.
- T. Zhang, R. Ramakrishnan, and M. Livny. **BIRCH : an efficient**

Créditos

- Jiawei Han (University of Illinois at Urbana-Champaign): “Data Mining: Concepts and Techniques”, capítulo 7, 2006
- Pang-Ning Tan (Michigan State University), Michael Steinbach & Vipin Kumar (University of Minnesota): “Introduction to Data Mining”, capítulos 8 y 9, 2006





Apéndice: Notación O

El impacto de la eficiencia de un algoritmo...

n	10	100	1000	10000	100000
O(n)	10ms	0.1s	1s	10s	100s
O(n·log$_2$ n)	33ms	0.7s	10s	2 min	28 min
O(n²)	100ms	10s	17 min	28 horas	115 días
O(n³)	1s	17min	12 días	31 años	32 milenios