



DECSAI

Departamento de Ciencias de la Computación e I.A.

Universidad de Granada



Clustering

© Fernando Berzal, berzal@acm.org

Clustering



- Introducción
- Medidas de similitud
 - Para atributos continuos: Métricas de distancia.
 - Para atributos no continuos, p.ej. distancia de edición.
- Métodos de agrupamiento
 - K-Means
 - DBSCAN
 - Clustering jerárquico
- Evaluación de resultados
- Validación de resultados
- Apéndice: El problema de la dimensionalidad



Introducción



“Sinónimos” según el contexto...

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

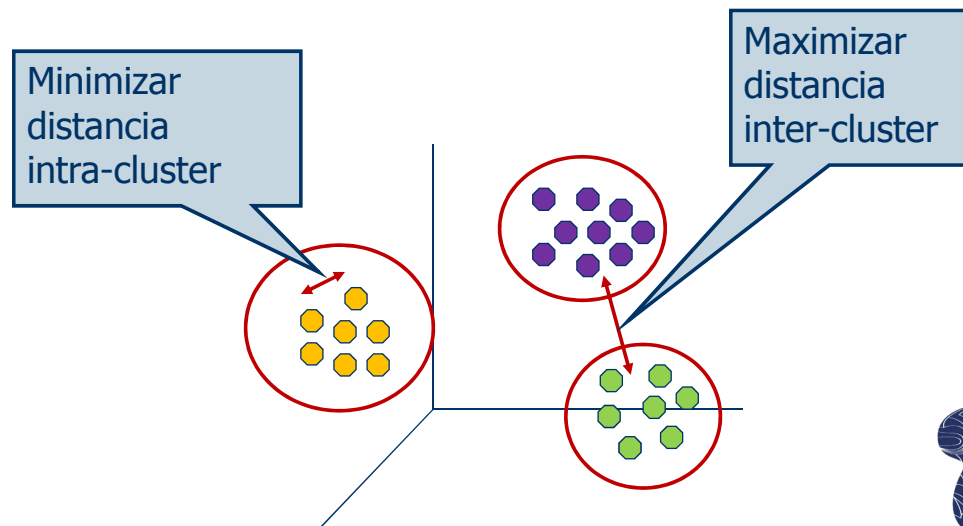


Introducción



Objetivo

Encontrar agrupamientos de tal forma que los objetos de un grupo sean similares entre sí y diferentes de los objetos de otros grupos [*clusters*].



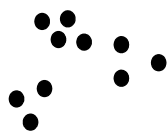
Introducción



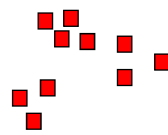
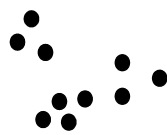
- **Aprendizaje no supervisado:**
No existen clases predefinidas.
- Los resultados obtenidos dependerán de:
 - ❖ El algoritmo de agrupamiento seleccionado.
 - ❖ El conjunto de datos disponible.
 - ❖ La medida de similitud utilizada para comparar objetos (usualmente, definida como medida de distancia).



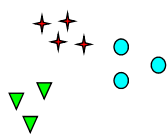
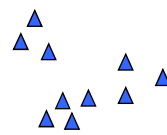
Introducción



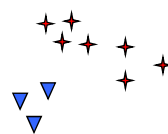
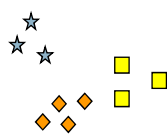
¿Cuántos
agrupamientos?



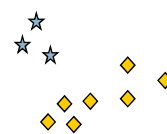
¿Dos?



¿Seis?



¿Cuatro?



Introducción



Aplicaciones

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

También se usa como paso previo a otras técnicas de Minería de Datos:

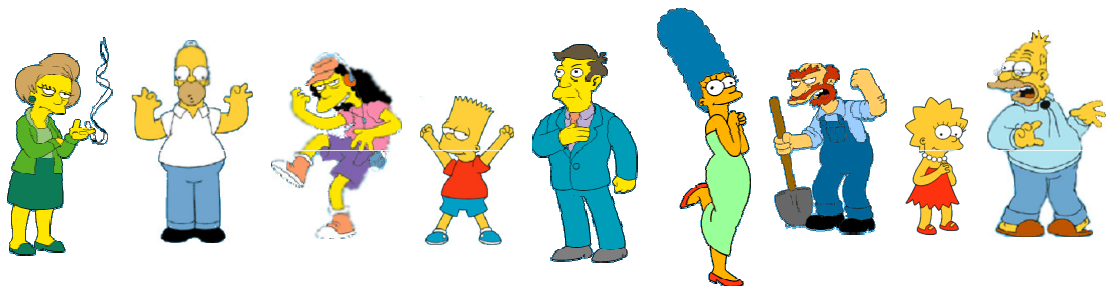
- Exploración de datos (segmentación & outliers)
- Preprocesamiento (p.ej. reducción de datos)



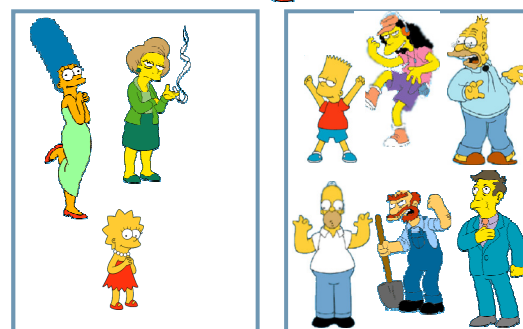
Medidas de similitud



¿Cuál es la forma natural de agrupar los personajes?



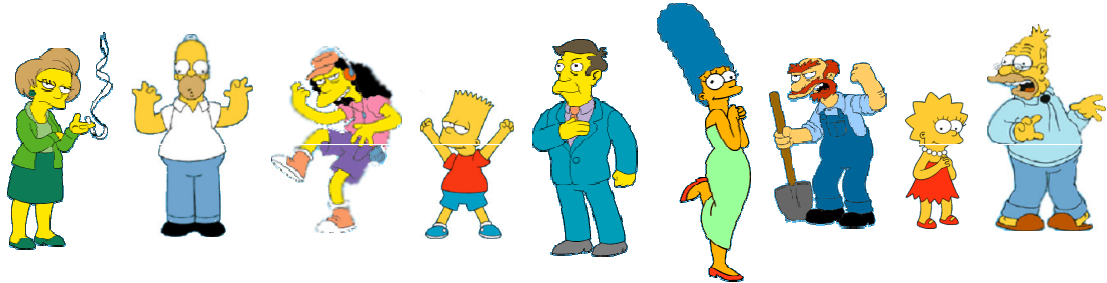
Mujeres
vs.
Hombres



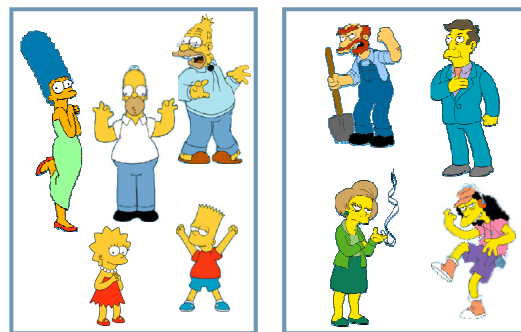
Medidas de similitud



¿Cuál es la forma natural de agrupar los personajes?



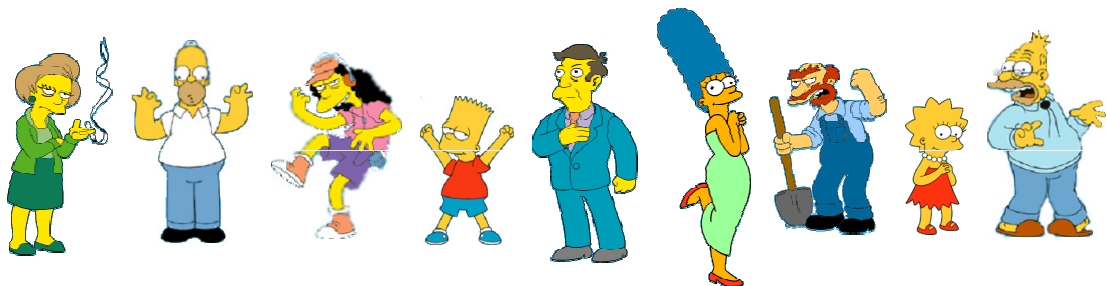
Simpsons
vs.
Empleados de
la escuela de
Springfield



Medidas de similitud



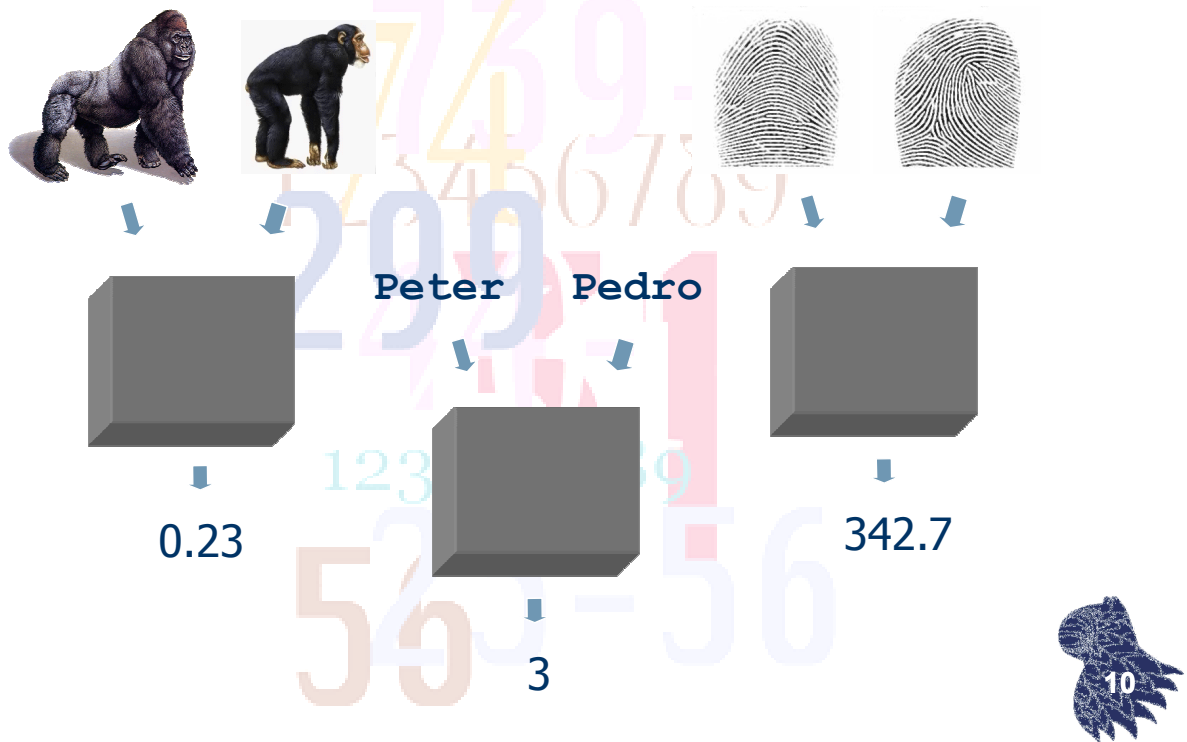
¿Cuál es la forma natural de agrupar los personajes?



iii El clustering es subjetivo !!!



Medidas de similitud



Medidas de similitud



	id	sexo	fechnac	educ	cattlab	salario	salini	T.emp	expprev	minoría
Grupo 1	121	Mujer	6-ago-1936	15	Administrativo	\$18.750	\$10.500	90	54	No
	122	Mujer	26-sep-1965	15	Administrativo	\$32.550	\$13.500	90	22	No
	123	Mujer	24-abr-1949	12	Administrativo	\$33.300	\$15.000	90	3	No
	124	Mujer	29-may-1963	16	Administrativo	\$38.550	\$16.500	90	Ausente	No
	125	Hombre	6-ago-1956	12	Administrativo	\$27.450	\$15.000	90	173	Sí
Grupo 2	126	Hombre	21-ene-1951	15	Seguridad	\$24.300	\$15.000	90	191	Sí
	127	Hombre	1-sep-1950	12	Seguridad	\$30.750	\$15.000	90	209	Sí
Grupo 3	128	Mujer	25-jul-1946	12	Administrativo	\$19.650	\$9.750	90	229	Sí
	129	Hombre	18-jul-1959	17	Directivo	\$68.750	\$27.510	89	38	No
	130	Hombre	6-sep-1958	20	Directivo	\$59.375	\$30.000	89	6	No
	131	Hombre	8-feb-1962	15	Administrativo	\$31.500	\$15.750	89	22	No
	132	Hombre	17-may-1953	12	Administrativo	\$27.300	\$17.250	89	175	No
	133	Hombre	12-sep-1959	15	Administrativo	\$27.000	\$15.750	89	87	No

NOTA: No será posible que todas las variables tengan valores similares en un mismo grupo, por lo que habrá que usar una medida global de semejanza entre los elementos de un mismo grupo.



Medidas de similitud



id	sexo	fechnac	educ	catlab	salario	salini	T.emp	expprev	minoría
121	Mujer	6-ago-1936	15	Administrativo	\$18.750	\$10.500	90	54	No
122	Mujer	26-sep-1965	15	Administrativo	\$32.550	\$13.500	90	22	No
123	Mujer	24-abr-1949	12	Administrativo	\$33.300	\$15.000	90	3	No

A la hora de calcular la similitud entre dos objetos:

- No tienen por qué utilizarse todos los atributos disponibles en nuestro conjunto de datos.
- Hay que tener cuidado con las magnitudes de cada variable.



Medidas de similitud



Atributos continuos

Para evitar que unas variables dominen sobre otras, los valores de los atributos se "normalizan" a priori:

- Desviación absoluta media:

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

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

- z-score (medida estandarizada):

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



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.

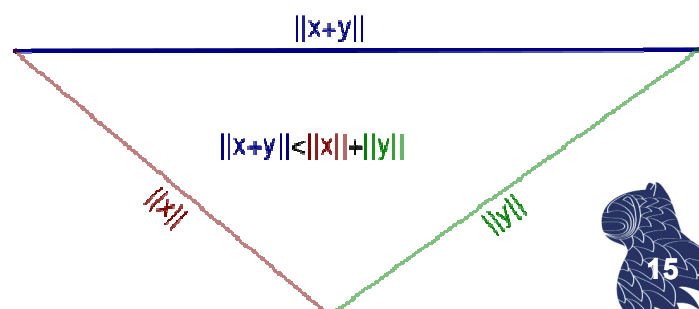


Medidas de similitud



Se suelen usar medidas de distancia porque verifican las siguientes propiedades:

- Propiedad reflexiva $d(i,j) = 0$ si y sólo si $i=j$
- 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 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 ($r \rightarrow \infty$) / dominio / chessboard

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

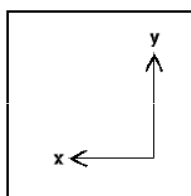


Medidas de similitud

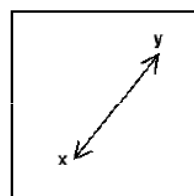


Métricas de distancia:

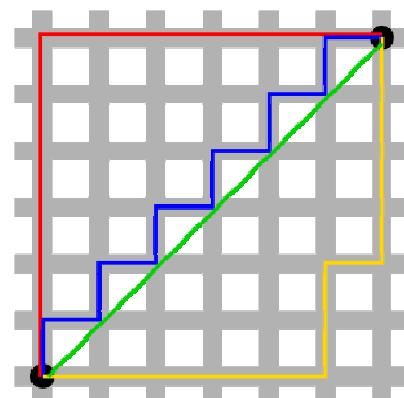
Distancia de Minkowski



Manhattan



Euclidean



- Distancia de Manhattan = 12 (roja, azul o amarilla)
- Distancia euclídea ≈ 8.5 (verde - continua)
- Distancia de Chebyshev = 6 (verde - discreta)



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.

	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1	♔	1	2	6
5	5	4	3	2	1	1	1	2	5
4	5	4	3	2	2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1
	a	b	c	d	e	f	g	h	



Medidas de similitud

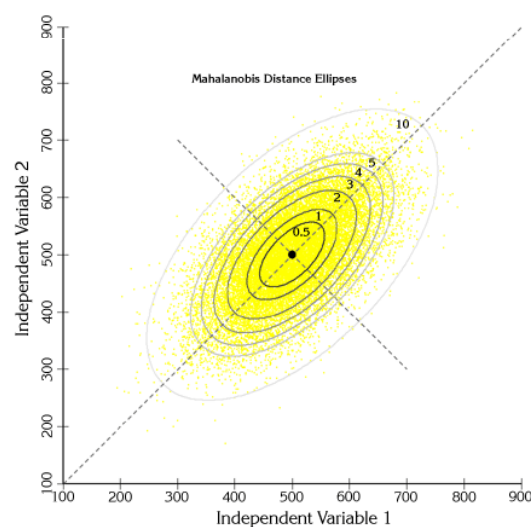


Métricas de distancia:

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 para atributos no continuos:

Distancia de edición = Distancia de Levenshtein

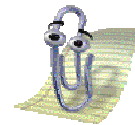
Número de operaciones necesario para transformar una cadena en otra.

$$d(\text{"data mining"}, \text{"data minino"}) = 1$$

$$d(\text{"efecto"}, \text{"defecto"}) = 1$$

$$d(\text{"poda"}, \text{"boda"}) = 1$$

$$d(\text{"night"}, \text{"natch"}) = d(\text{"natch"}, \text{"noche"}) = 3$$



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



Para datos binarios: Distancia de Hamming

Medidas de similitud



Métricas de similitud para atributos no continuos:

Modelos basados en Teoría de Conjuntos

Modelo de Tversky

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

donde $\theta, \alpha, \beta \geq 0$

■ Modelo de Restle

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

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

■ Intersección

$$S_{MinSum}(A, B) = |A \cap B|$$

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



Medidas de similitud



Métricas de similitud para atributos no continuos:

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

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

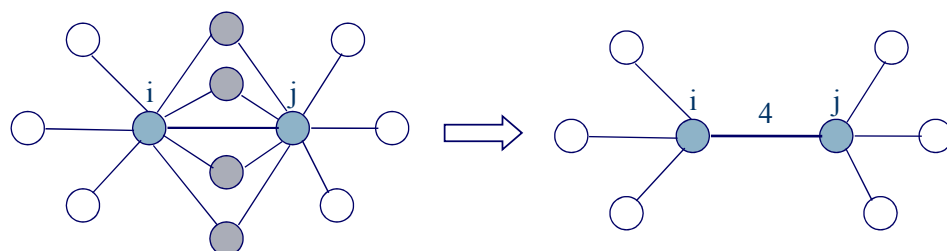


Medidas de similitud



Otras medidas de similitud:

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 $NN(\mathbf{x}_i, \mathbf{x}_j)$ es el número de vecino de \mathbf{x}_j con respecto a \mathbf{x}_i



Medidas de similitud

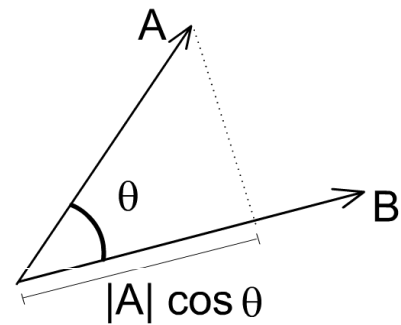


Otras medidas de similitud:

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

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



Métodos de agrupamiento



Tipos de algoritmos de clustering:

- **Agrupamiento por particiones**

k-Means, CLARANS

- **Métodos basados en densidad**

DBSCAN

- **Clustering jerárquico**

BIRCH, ROCK, CHAMELEON

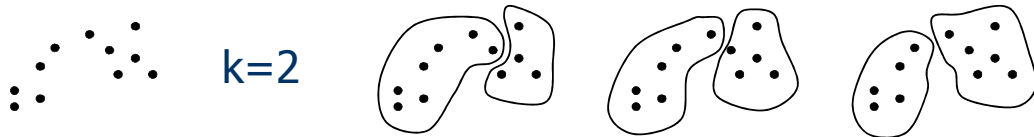
...



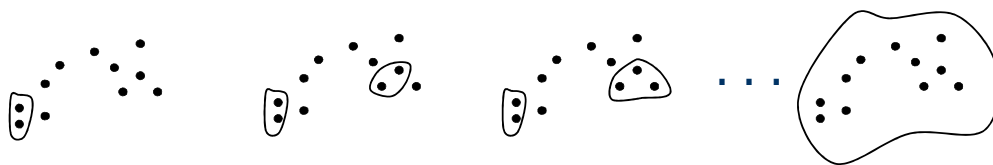
Métodos de agrupamiento



Clustering por particiones (suele fijarse k)



Clustering jerárquico (no se fija k)



Se obtiene como resultado final un conjunto de agrupamientos.



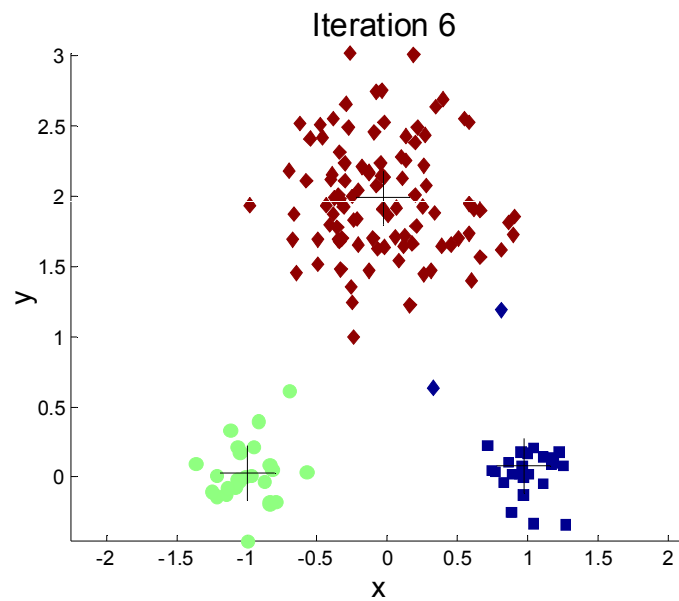
k-Means



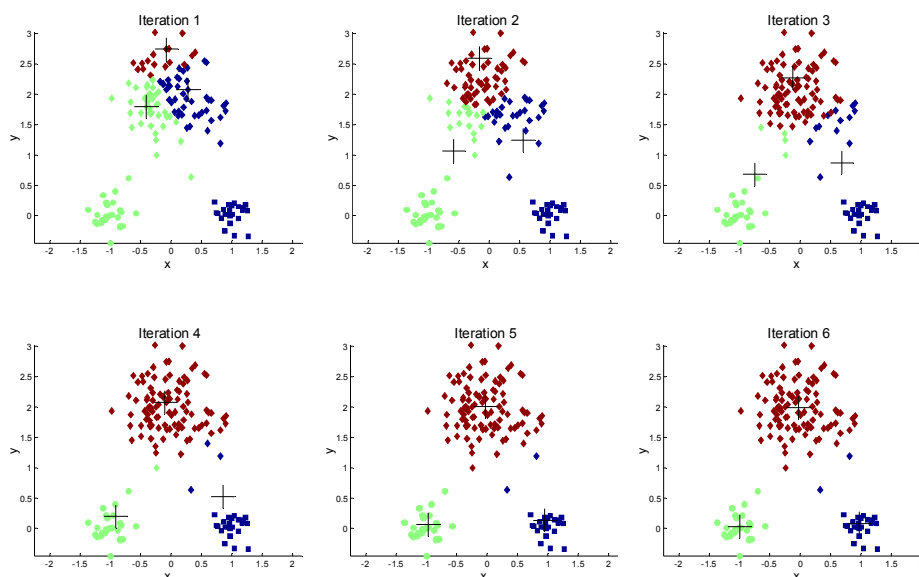
- Algoritmo de agrupamiento por particiones.
- Número de clusters conocido (k).
- Cada cluster tiene asociado un centroide (centro geométrico del cluster).
- Los puntos se asignan al cluster cuyo centroide esté más cerca (utilizando cualquier métrica de distancia).
- Iterativamente, se van actualizando los centroides en función de las asignaciones de puntos a clusters, hasta que los centroides dejen de cambiar.



k-Means



k-Means



k-Means



Inicialización

- Escoger k centroides aleatoriamente (hay métodos más sofisticados).
- Formar k grupos, asignando cada punto al centroide más cercano

Proceso iterativo

Mientras que los centroides cambien:

- Calcular las distancias de todos los puntos a los k centroides.
- Formar k grupos, asignando cada punto al centroide más cercano.
- Recalcular los nuevos centroides.



k-Means



Complejidad

n = número de puntos,
k = número de clusters,
I = número iteraciones,
d = número de atributos

Problema NP si k no se fija.

Ahora bien, si fijamos los parámetros n, k, d, I:

$$O(n * k * I * d)$$



k-Means



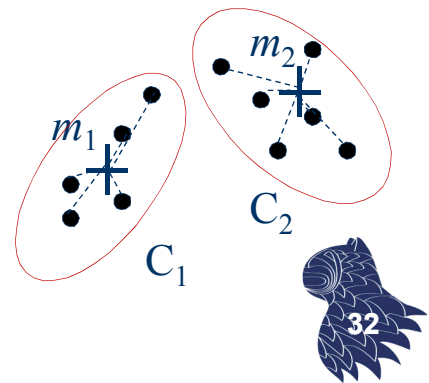
¿Cómo se recalculan los centroides?

Partiendo de k grupos de puntos, cada grupo determinará un nuevo centroide m_i .

- Se elige una medida global (función objetivo)

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} d^2(m_i, x)$$

- Se escogen los valores de m_i que minimizan dicha función.



k-Means



¿Cómo se recalculan los centroides?

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} d^2(m_i, x)$$

- Cuando se utiliza la distancia euclídea, SSE se minimiza usando la media aritmética (por cada atributo o variable)

$$x_a = (0.4, 0.6, 0.6, 0.7)$$

$$x_b = (0.3, 0.2, 0.1, 0.4)$$

$$x_c = (0.3, 0.2, 0.2, 0.4)$$

$$m_1 = (0.33, 0.33, 0.3, 0.5)$$

- Cuando se emplea la distancia de Manhattan, SSE se minimiza usando la mediana.
- La media funciona bien con muchas distancias (por ejemplo, cualquier divergencia de Bregman).



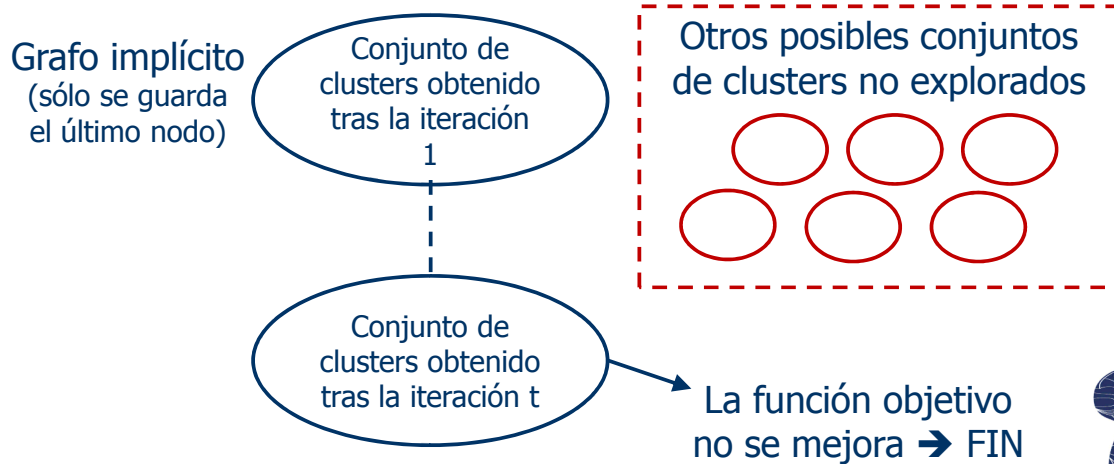
k-Means



Estrategia de control en la búsqueda:

Ascensión de colinas por la máxima pendiente

Después de cada iteración, no se recuerda el estado para volver atrás y probar otros posibles centroides.



k-Means



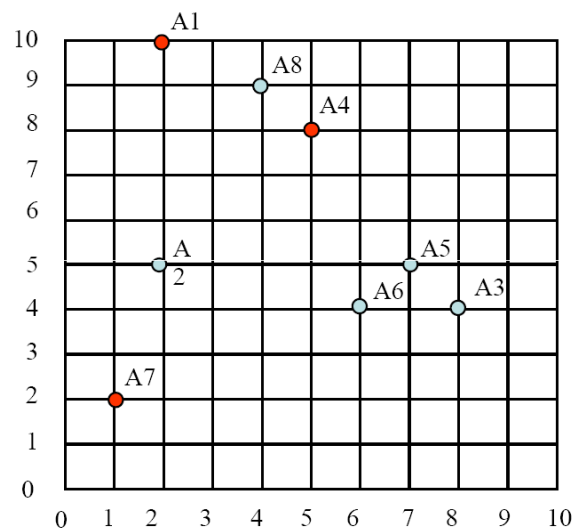
Ejercicio

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

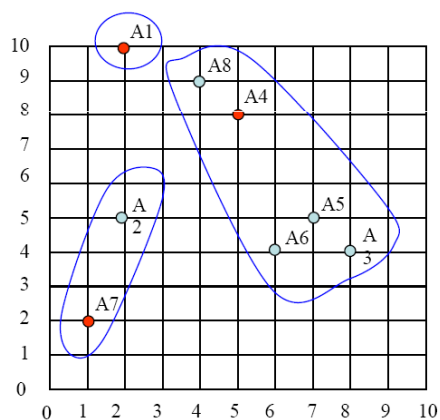


k-Means

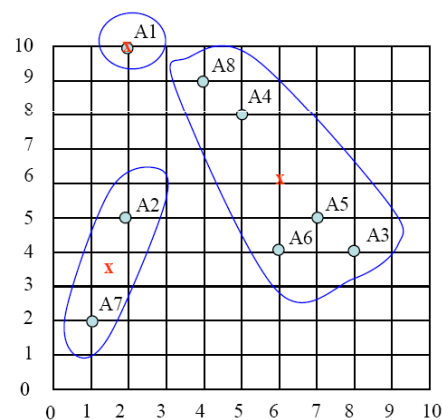


Ejercicio resuelto

Distancia euclídea



Primera iteración



Segunda iteración

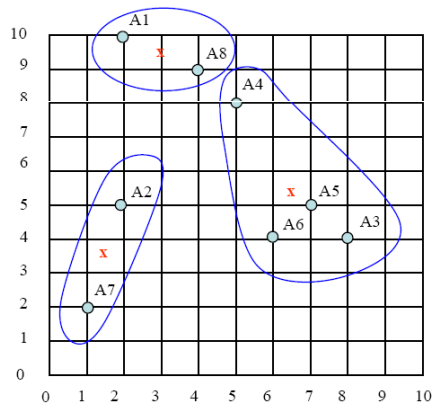


k-Means

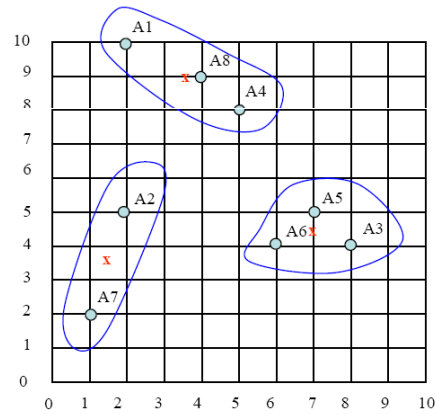


Ejercicio resuelto

Distancia euclídea



Tercera iteración



Configuración final



k-Means

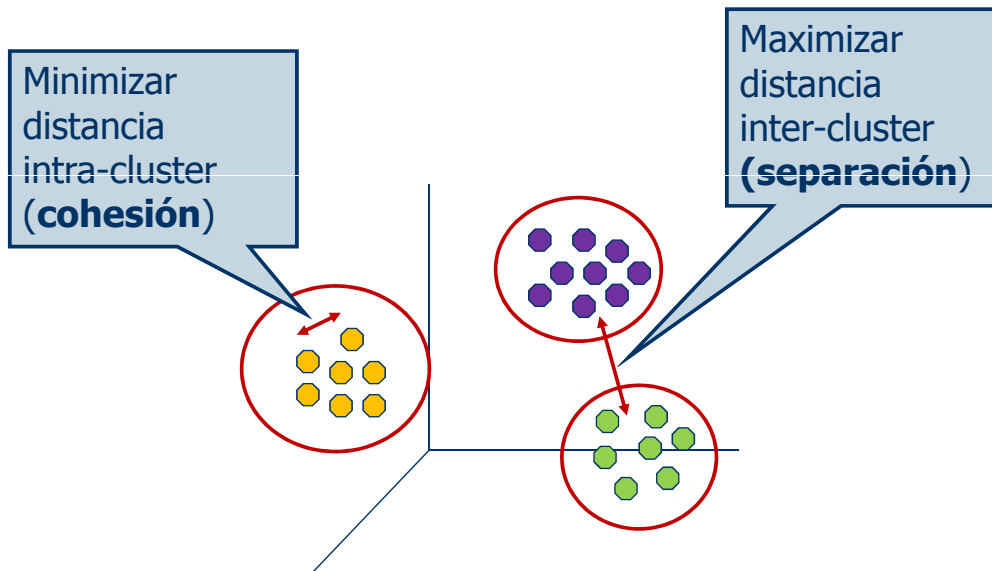


DEMO

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



Evaluación de resultados

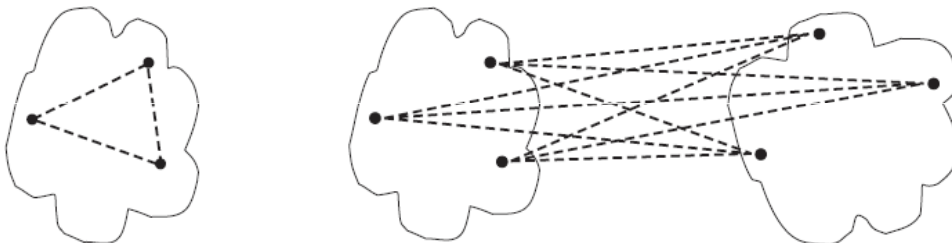


Evaluación de resultados

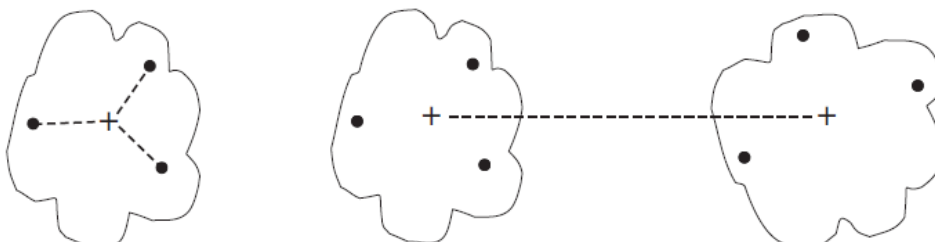


Formas de medir cohesión y separación:

- Sin usar centroides:



- Usando centroides:

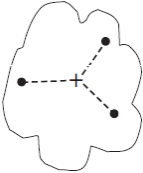
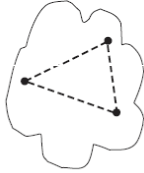


Evaluación de resultados



Cuando la distancia utilizada es la distancia euclídea:

1. El uso de centroides no influye al medir la cohesión:


$$\sum_{x \in C_i} \text{dist}^2(m_i, x) = \frac{1}{2(\#C_i)} \sum_{x \in C_i} \sum_{y \in C_i} \text{dist}^2(x, y)$$

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} \text{dist}^2(m_i, x) = \sum_{i=1}^K \frac{1}{2(\#C_i)} \sum_{x \in C_i} \sum_{y \in C_i} \text{dist}^2(x, y)$$

2. Minimizar cohesión y maximizar separación son equivalentes.



Evaluación de resultados



Así pues, cuando se usa la distancia euclídea, SSE es una buena medida del grado de ajuste (cohesión y separación) de los centroides hallados.

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} d^2(m_i, x)$$

Por otro lado, ya sabíamos que, en cada iteración del algoritmo de las k-medias, se maximizaba SSE al calcular los centroides usando la media aritmética.

¿Garantiza lo anterior que los centroides finales sean los que minimicen SSE globalmente? **NO**



Evaluación de resultados



Cuando usamos la distancia euclídea, el centroide determinado en cada iteración por el vector de medias garantiza la mejor solución con respecto a SSE, pero considerando:

- un valor de k fijo, y
- los centroides dados por la iteración anterior.

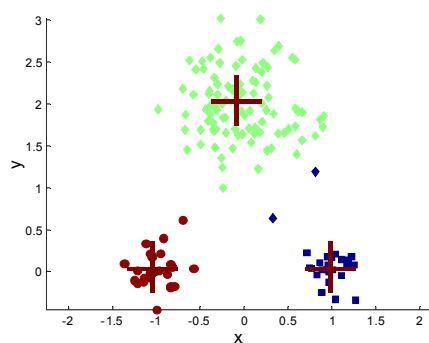
La solución final no será la óptima:

El algoritmo de las k medias **no garantiza** que los centroides finales obtenidos sean los que minimizan globalmente la función objetivo SSE.

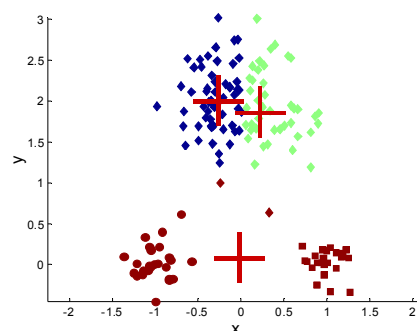


44

Evaluación de resultados



Solución óptima

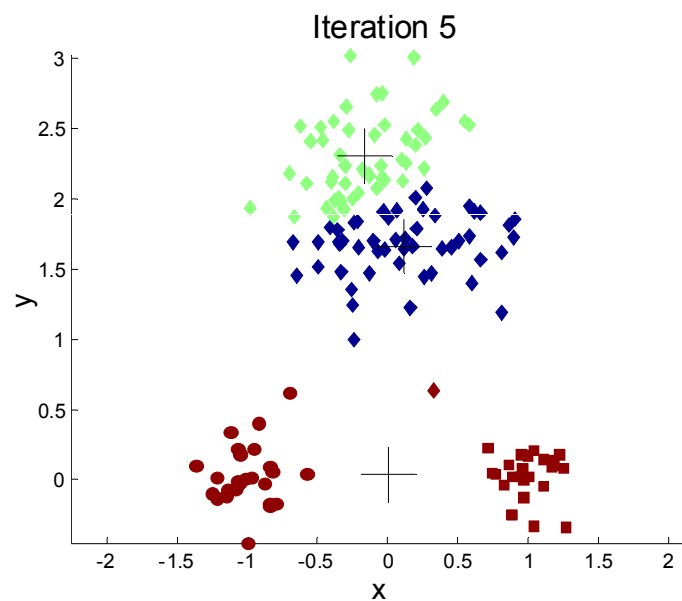


Posible resultado proporcionado por k-means
Óptimo local

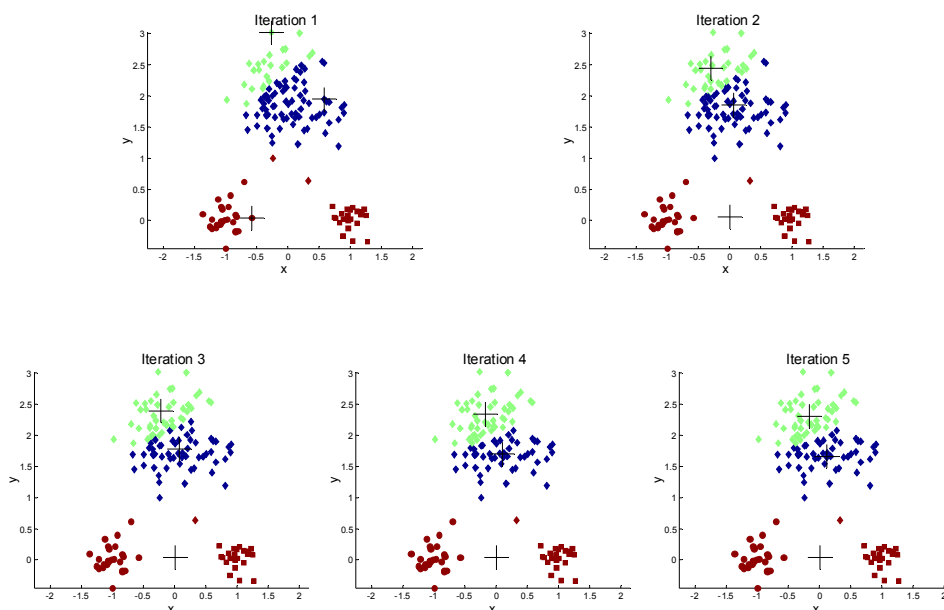


45

Evaluación de resultados



Evaluación de resultados

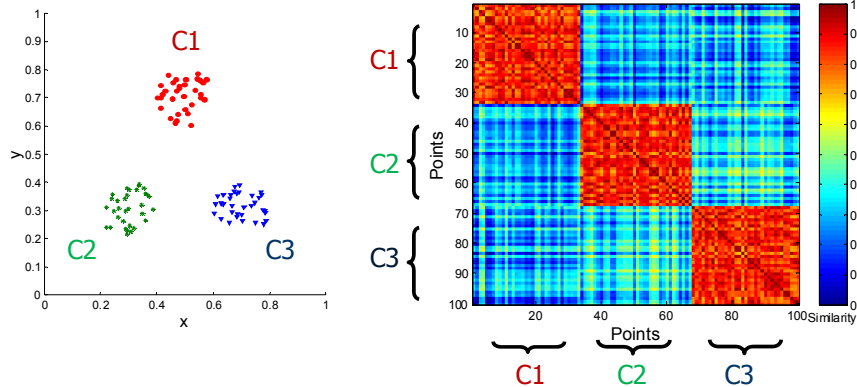


Evaluación de resultados



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...

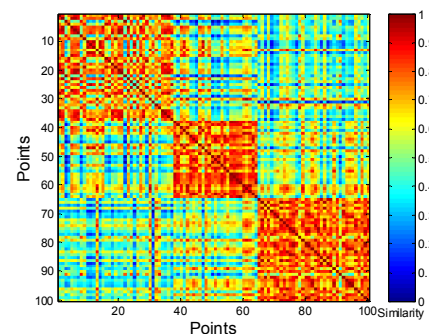
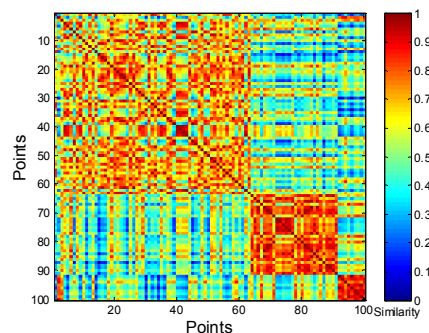
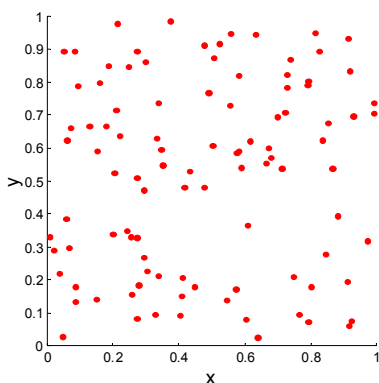


Evaluación de resultados



Problema

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

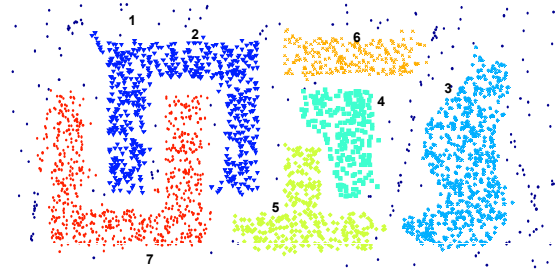
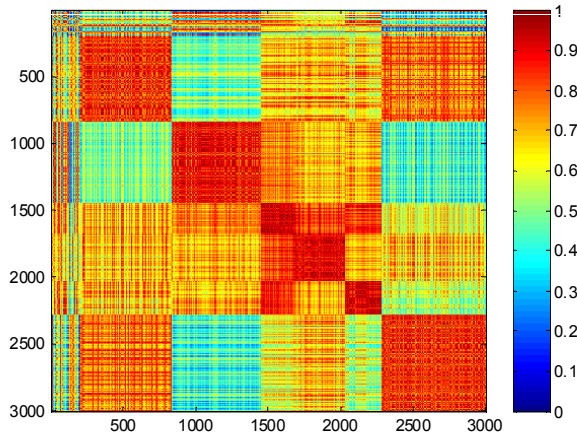


Evaluación de resultados



Matriz de similitud

DBSCAN



k-Means



Problema

Sensible a la elección inicial de los centroides.

Posibles soluciones

- Realizar varias ejecuciones con varios conjuntos de centroides iniciales y comparar resultados (GRASP).
- Estimar a priori unos buenos centroides:
 - Métodos específicos: k-means++
 - Escoger una muestra y aplicar un método jerárquico



k-Means



Problema

Hay que elegir a priori el valor de k
(a priori, no sabemos cuántos grupos puede haber).

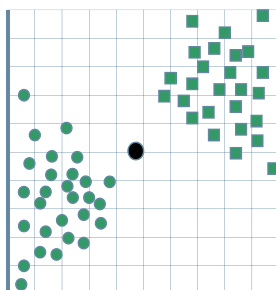
Posibles soluciones

- Usar un método jerárquico sobre una muestra de los datos (por eficiencia) para estimar el valor de k .
- Usar un valor de k alto, ver los resultados y ajustar. Siempre que se aumente el valor de k , disminuirá el valor SSE. Lo normal será ir probando con varios valores de k y comprobar cuándo no hay una mejora significativa en SSE.

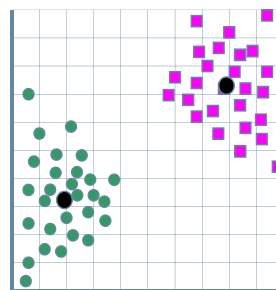


52

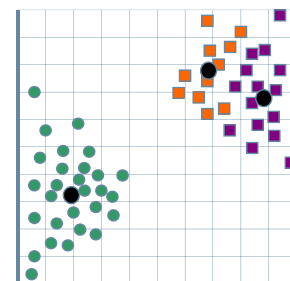
k-Means



$k = 1$
SSE = 873.0



$k = 2$
SSE = 173.1

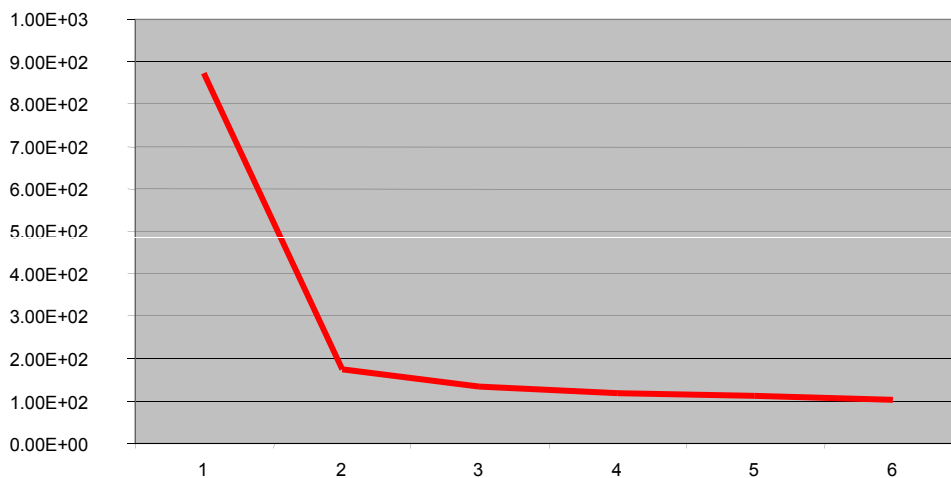


$k = 3$
SSE = 133.6



53

k-Means



El codo en $k=2$ sugiere que éste es el valor más adecuado para el número de agrupamientos.



k-Means



Problema

Cuando se usan la media para calcular los centroides, el método es sensible a outliers (valores anómalos).

Posibles soluciones

- Usar medianas en vez de medias (aun con la distancia euclídea).
- Eliminar previamente los outliers.
¡Ojo! Los outliers pueden ser valores interesantes
- Usar k-medoids: En vez de usar el vector de medias como centroide, se usa el vector correspondiente a un dato real (un representante).



k-Means



Problema

Manejo de atributos no numéricos.

Posibles soluciones

- Extender la medida de similitud para que incluya atributos nominales, p.ej.

$$d(a,b) = 1 \text{ si } a=b, 0 \text{ en otro caso}$$

Elegir como representante en el centroide la moda de los datos asignados a dicho cluster (método k-mode).



k-Means



Problema

K-Means no funciona bien cuando los clusters son:

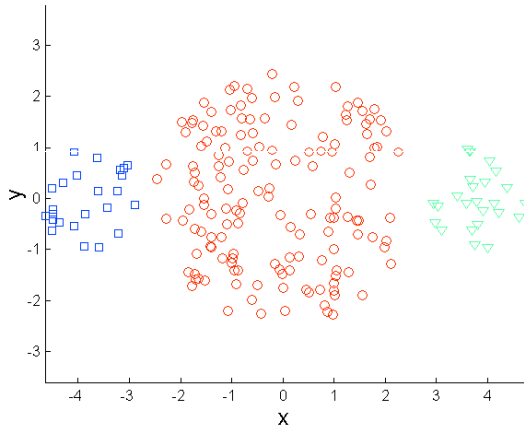
- de distinto tamaño
- de diferente densidad
- no convexos



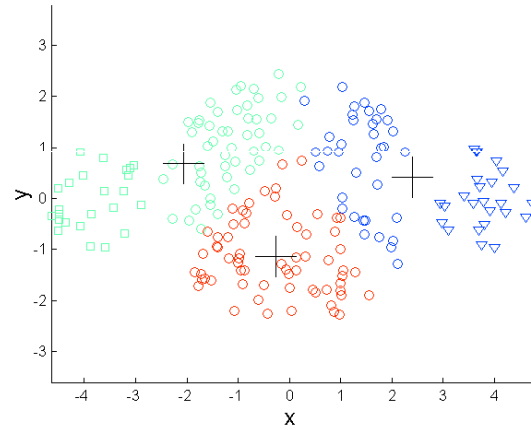
k-Means



Clusters de distinto tamaño



Puntos originales



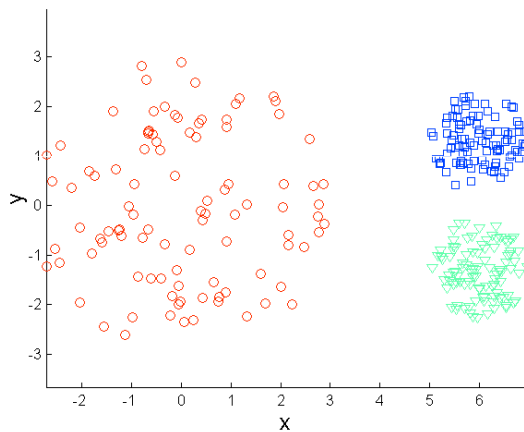
k-Means (3 clusters)



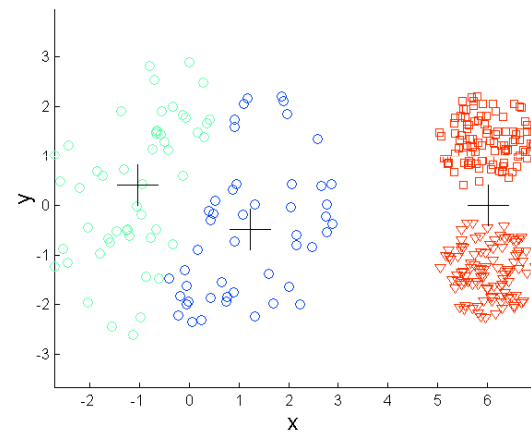
k-Means



Clusters de distinta densidad



Puntos originales



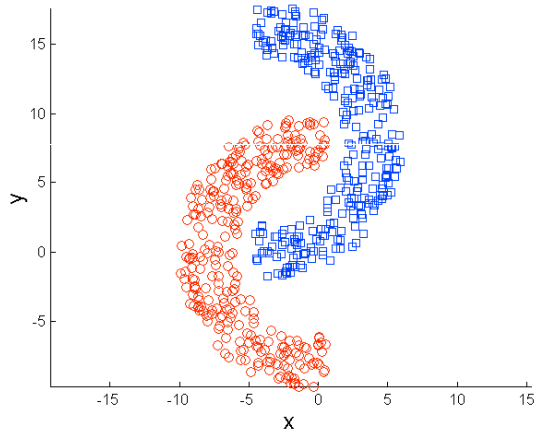
k-Means (3 clusters)



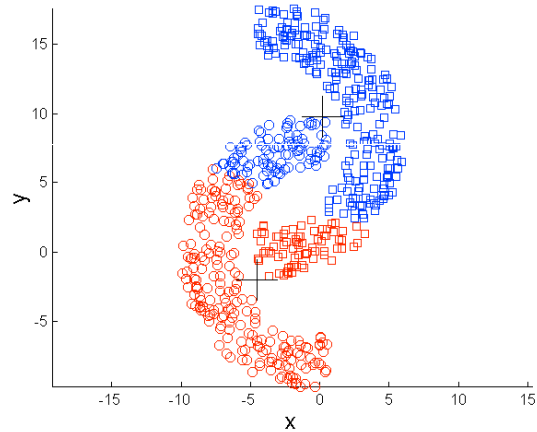
k-Means



Clusters no convexos



Puntos originales



k-Means (2 clusters)



k-Means



Problema

K-Means no funciona bien cuando los clusters son:

- de distinto tamaño
- de diferente densidad
- no convexos

Posibles soluciones

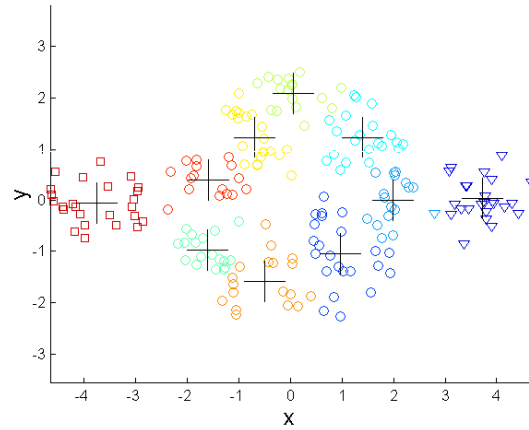
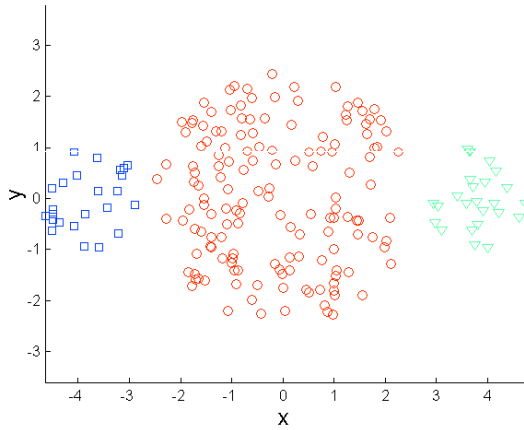
- Métodos ad-hoc.
- Usar un valor de k alto y revisar los resultados.



k-Means



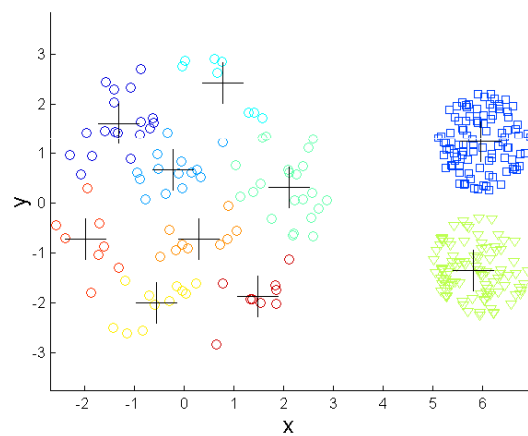
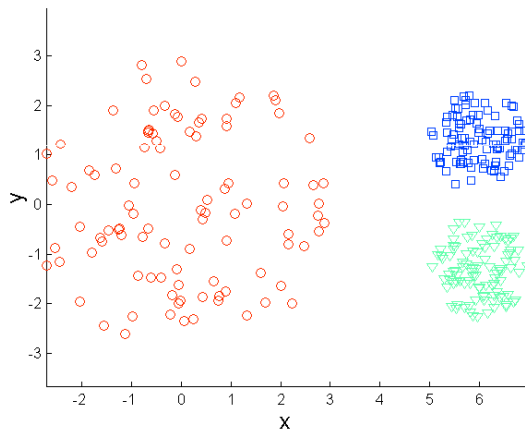
Clusters de distinto tamaño



k-Means



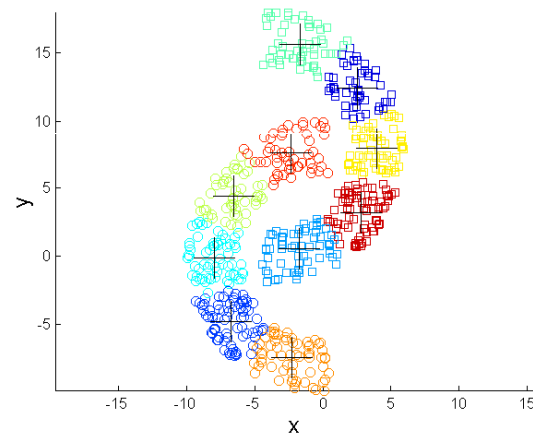
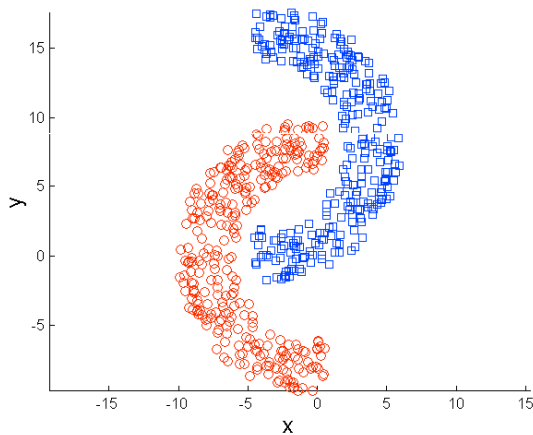
Clusters de distinta densidad



k-Means



Clusters no convexos



k-Means



Pre-procesamiento

- Normalizar los datos.
- Detectar outliers (eliminarlos, en su caso).

Post-procesamiento

- Eliminar pequeños clusters que puedan representar outliers.
- Dividir clusters dispersos ('loose' clusters); esto es, clusters con un SSE relativamente alto.
- Combinar clusters cercanos que tengan un SSE relativamente bajo.

NOTA: Estos criterios se pueden incluir en el propio algoritmo de clustering (p.ej. algoritmo ISODATA).



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)

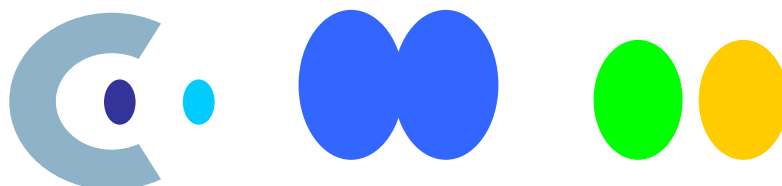


DBSCAN



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.





Métodos basados en densidad

Criterio de agrupamiento local:

Densidad de puntos

Regiones densas de puntos separadas de otras regiones densas por regiones poco densas.

Características

- Identifican clusters de formas arbitrarias.
- Robustos ante la presencia de ruido.
- Escalables: Un único recorrido del conjunto de datos



Algoritmos basados en densidad

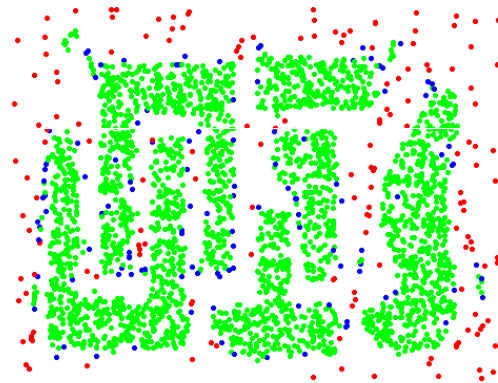
- **DBSCAN**: Density Based Spatial Clustering of Applications with Noise (Ester et al., KDD'1996)
- **OPTICS**: Ordering Points To Identify the Clustering Structure (Ankerst et al. SIGMOD'1999)
- **DENCLUE**: DENSity-based CLUstEring (Hinneburg & Keim, KDD'1998)
- **CLIQUE**: Clustering in QUEst (Agrawal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



DBSCAN



Detecta regiones densas de puntos separadas de otras regiones densas por regiones poco densas:



Parámetros: Epsilon = 10, MinPts = 4

Puntos: **core** (cluster), **border** (frontera) y **noise** (ruido)

Eficiencia: **$O(n \log n)$**



DBSCAN



Ejercicio

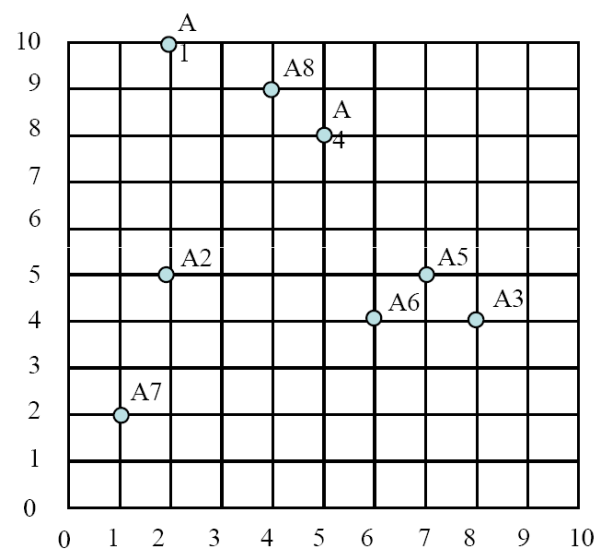
Agrupar los 8 puntos de la figura utilizando el algoritmo DBSCAN.

Número mínimo de puntos en el "vecindario":

$$\text{MinPts} = 2$$

Radio del "vecindario":

$$\text{Epsilon} \sqrt{2} > \sqrt{10}$$





Ejercicio resuelto

Distancia euclídea

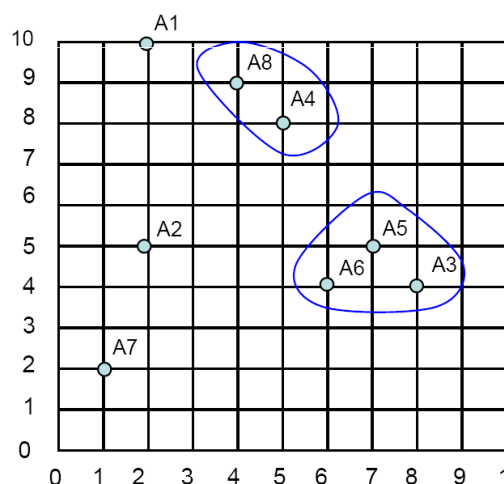
	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$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



Ejercicio resuelto

Epsilon = $\sqrt{2}$

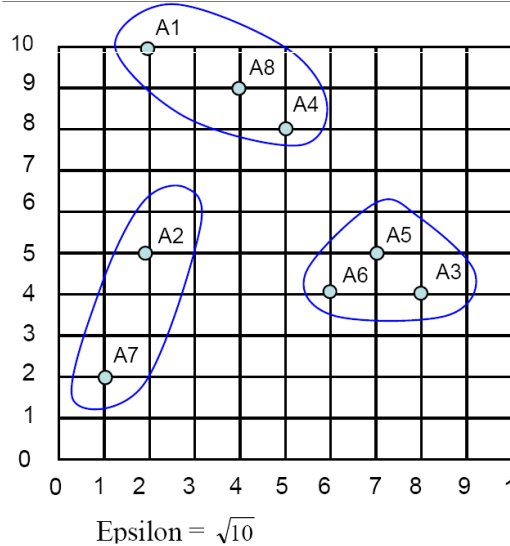
A1, A2 y A7 no tienen vecinos en su vecindario, por lo que se consideran "outliers" (no están en zonas densas):



Ejercicio resuelto

$$\text{Epsilon} = \sqrt{10}$$

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

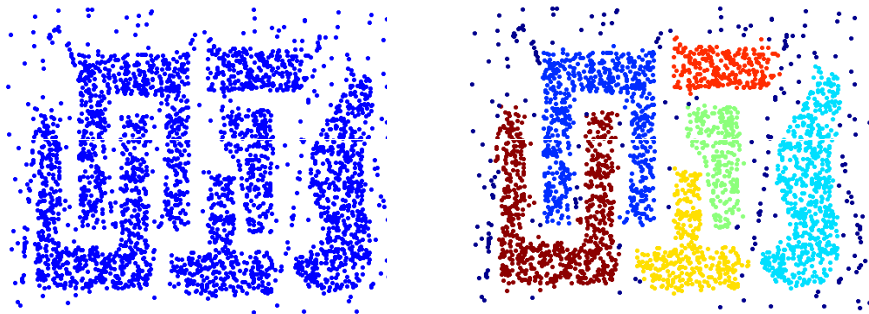


DEMO

<http://webdocs.cs.ualberta.ca/~yaling/Cluster/Applet/Code/Cluster.html>



DBSCAN

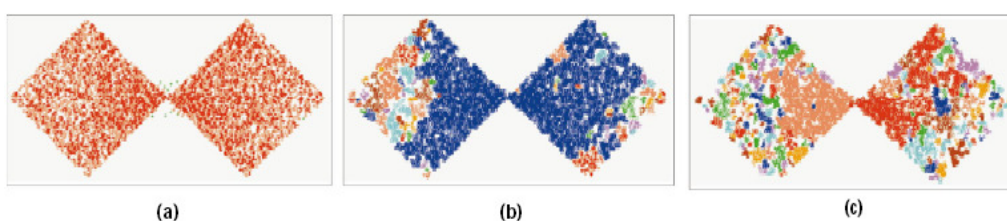
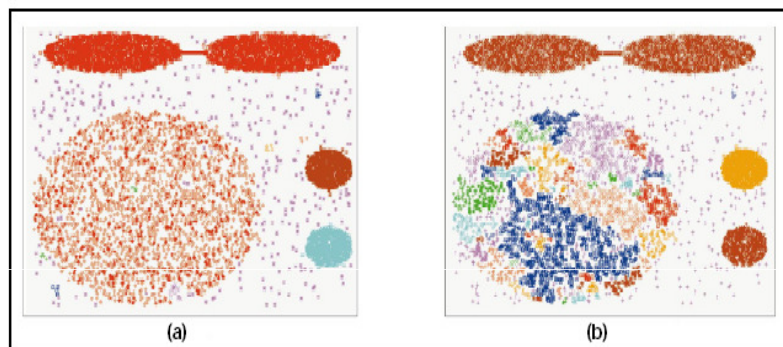


Clusters

DBSCAN... cuando funciona bien :-)



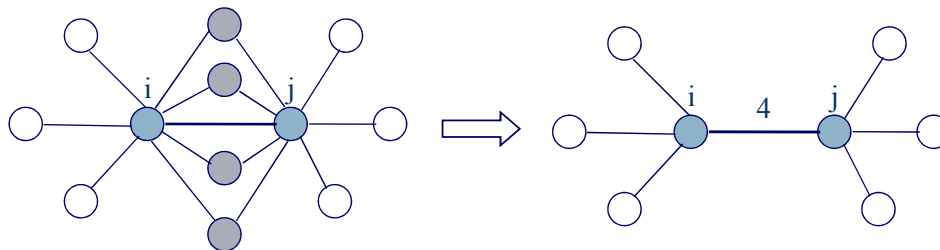
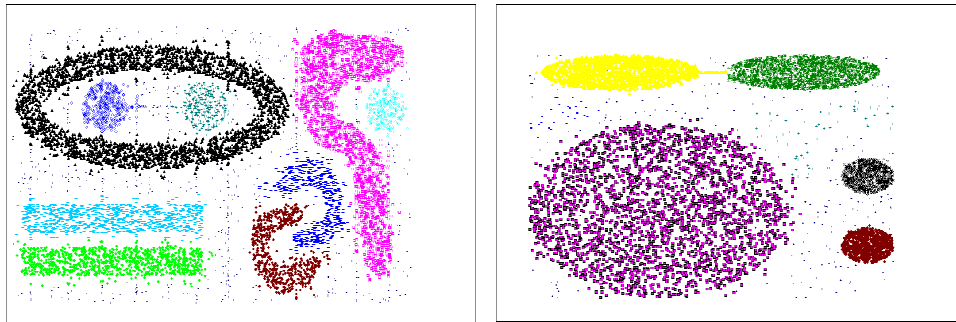
DBSCAN



DBSCAN... cuando no funciona :-)



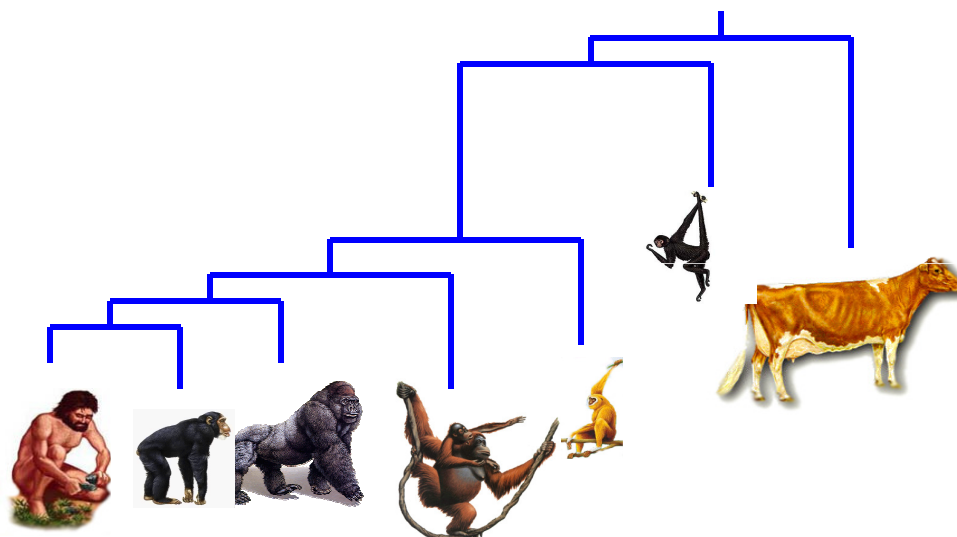
DBSCAN++



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



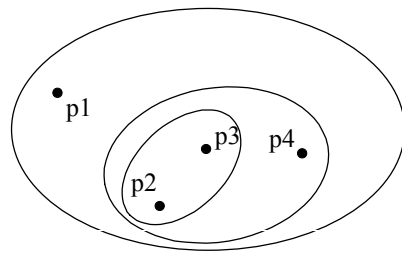
Clustering jerárquico



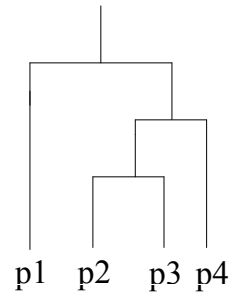
- **DENDROGRAMA:** La similitud entre dos objetos viene dada por la "altura" del nodo común más cercano.



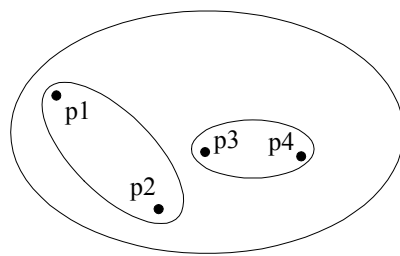
Clustering jerárquico



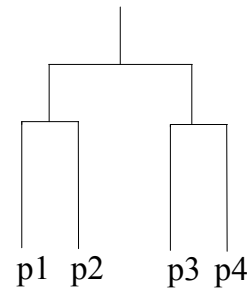
Tradicional



DENDOGRAMA



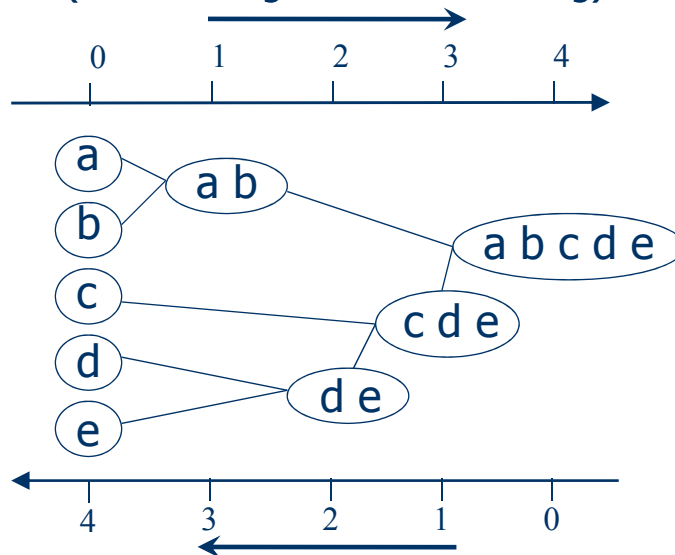
No tradicional



Clustering jerárquico



Aglomerativo
(AGNES: AGglomerative NESTing)



Divisivo
(DIANA: Divisive ANALysis)



Clustering jerárquico



Dos tipos de técnicas de clustering jerárquico

■ Técnicas aglomerativas

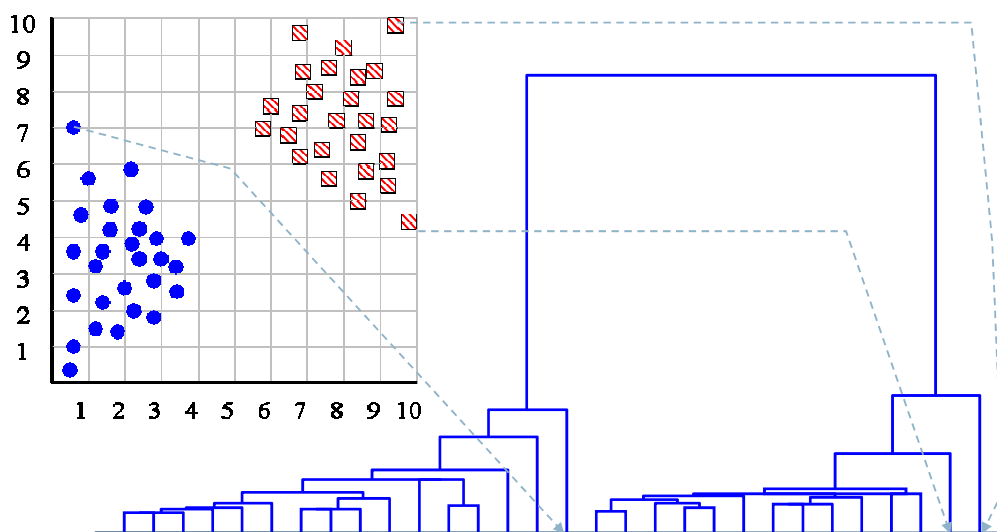
- Comenzar con cada caso como cluster individual.
- En cada paso, combinar el par de clusters más cercano hasta que sólo quede uno (o k).

■ Técnicas divisivas

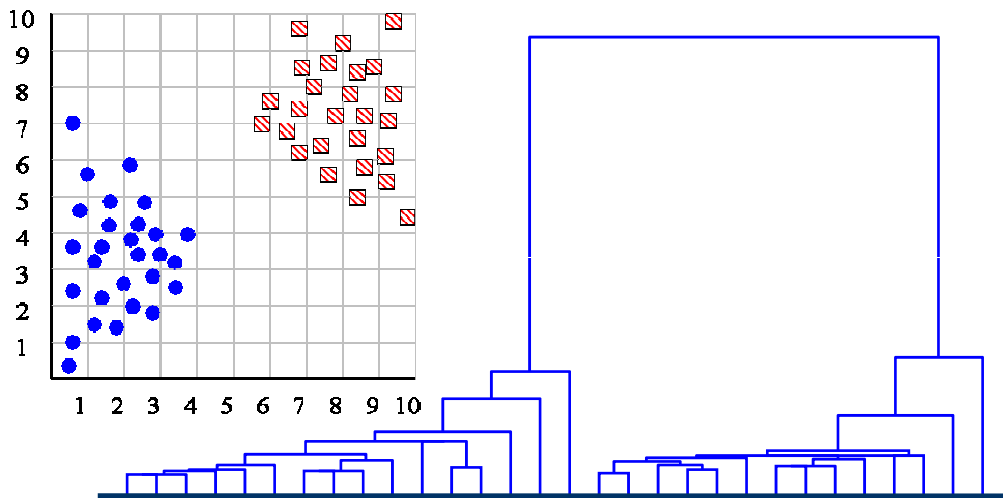
- Comenzar con un único cluster que englobe todos los casos de nuestro conjunto de datos.
- En cada paso, partir un cluster hasta que cada cluster contenga un único caso (o queden k clusters).



Clustering jerárquico



Clustering jerárquico



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

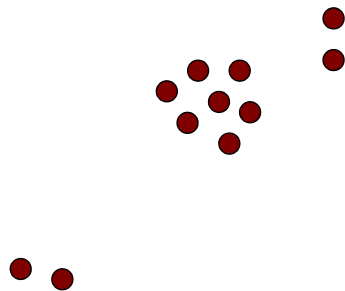


Clustering jerárquico



Ejemplo

Construir el correspondiente dendograma.
¿Cuál es el número ideal de clusters?



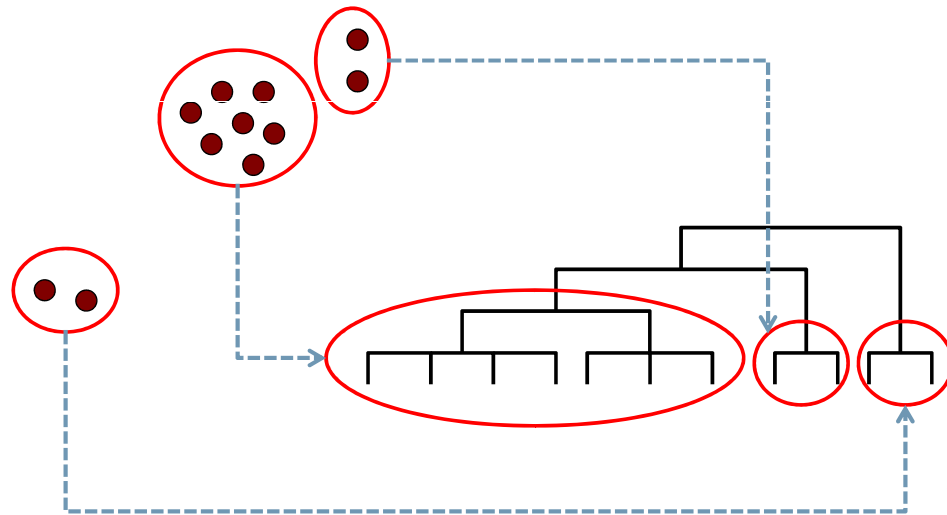
Clustering jerárquico



Ejemplo

Construir el correspondiente dendograma.

¿Cuál es el número ideal de clusters?



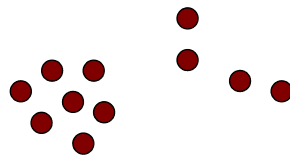
Clustering jerárquico



Ejemplo

Construir el correspondiente dendograma.

¿Cuál es el número ideal de clusters?



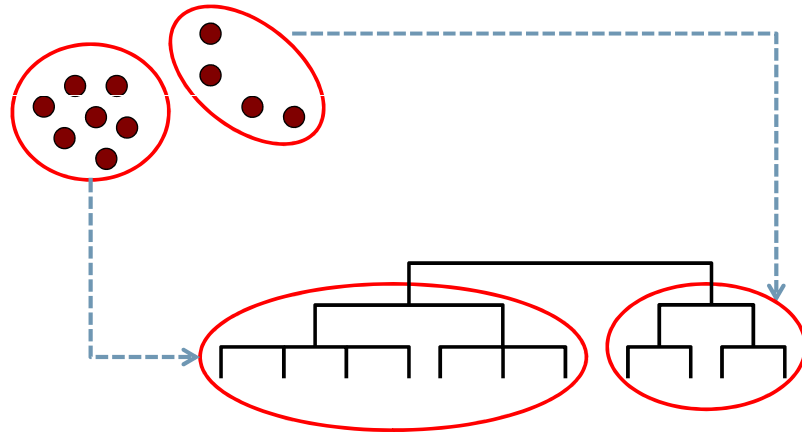
Clustering jerárquico



Ejemplo

Construir el correspondiente dendograma.

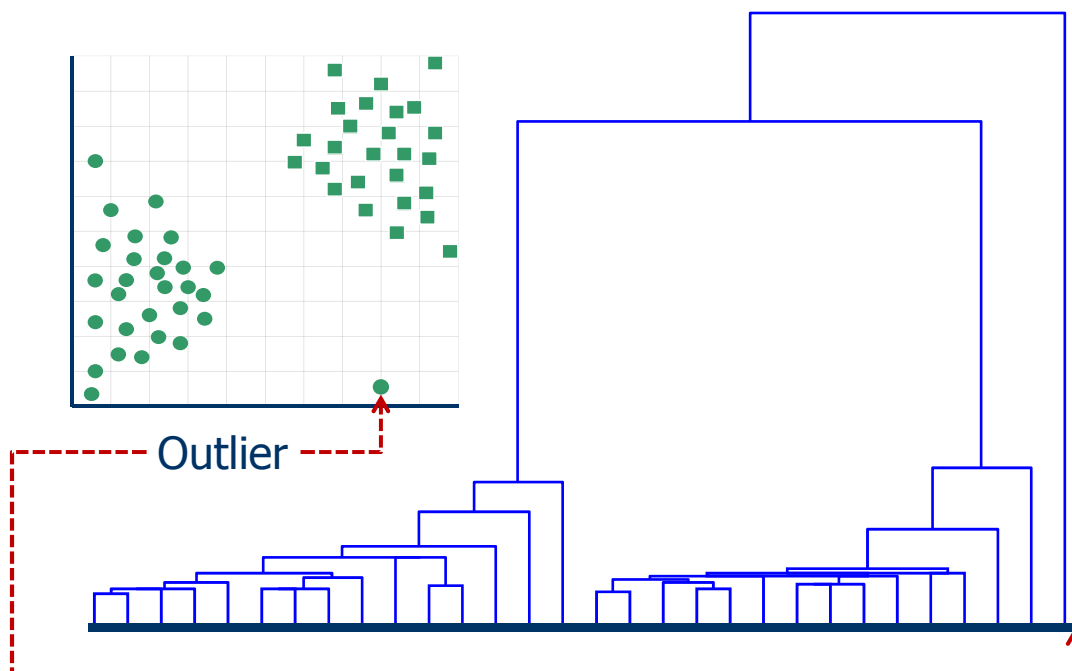
¿Cuál es el número ideal de clusters?



Clustering jerárquico



Detección de la presencia de outliers:



Clustering jerárquico



Algoritmo básico (aglomerativo)

Calcular la matriz de similitud/distancias

Inicialización: Cada caso, un cluster

Repetir

Combinar los dos clusters más cercanos

Actualizar la matriz de similitud/distancias hasta que sólo quede un cluster

- Estrategia de control irrevocable (greedy):
Cada vez que se unen dos clusters,
no se reconsidera otra posible unión.

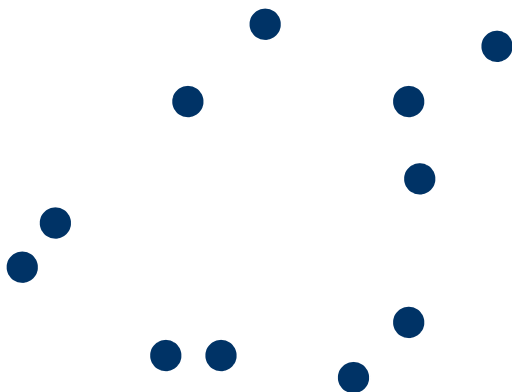


Clustering jerárquico



Inicialización:

Clusters de casos individuales
y matriz de distancias



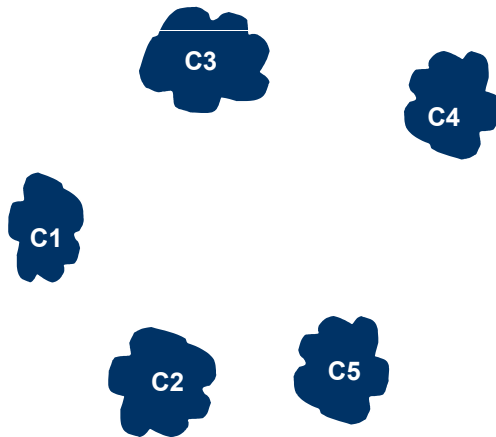
	p1	p2	p3	p4	p5
p1					
p2					
p3					
p4					
p5					



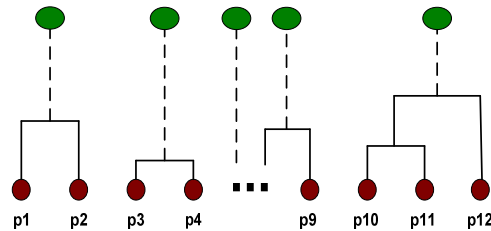
Clustering jerárquico



Tras varias iteraciones:
Varios clusters...



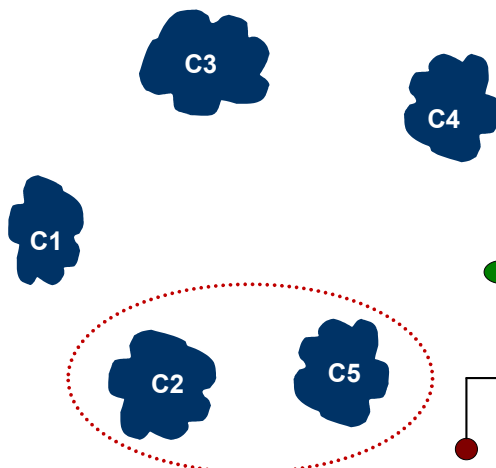
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					



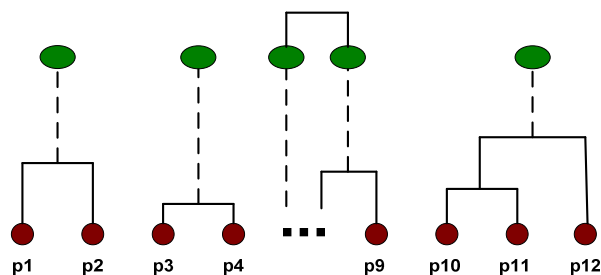
Clustering jerárquico



Combinamos los clusters 2 y 5
y actualizamos la matriz de distancias
¿cómo?



	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

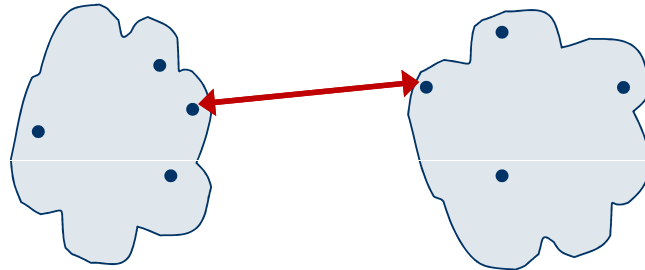


Clustering jerárquico

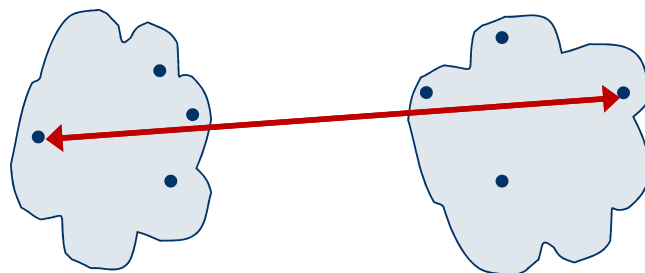


¿Cómo se mide la distancia entre clusters?

- MIN
single-link



- MAX
complete
linkage
(diameter)



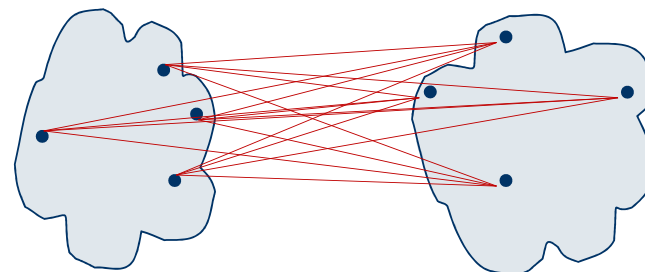
94

Clustering jerárquico

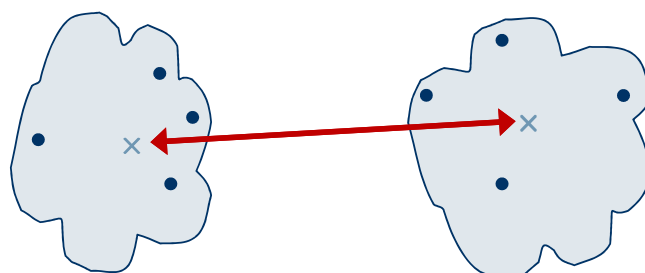


¿Cómo se mide la distancia entre clusters?

- Promedio



- Centroides
p.ej. BIRCH

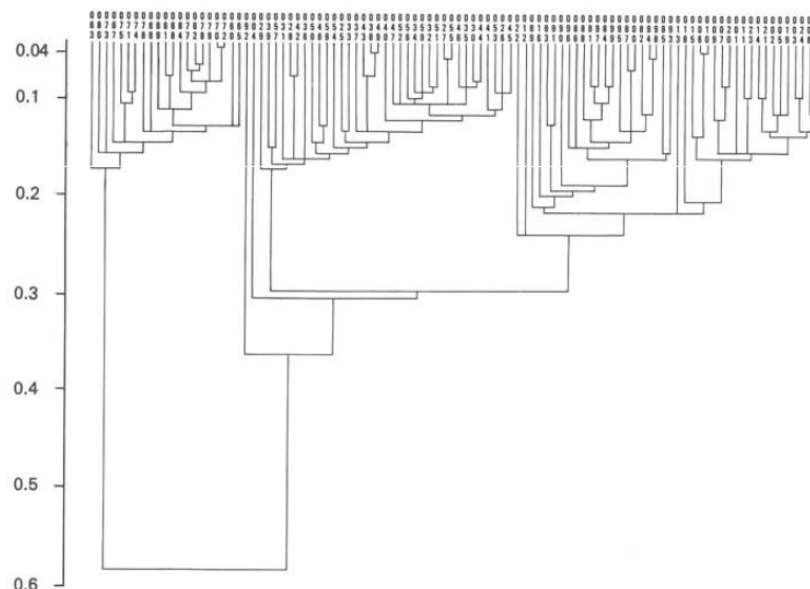


95

Clustering jerárquico



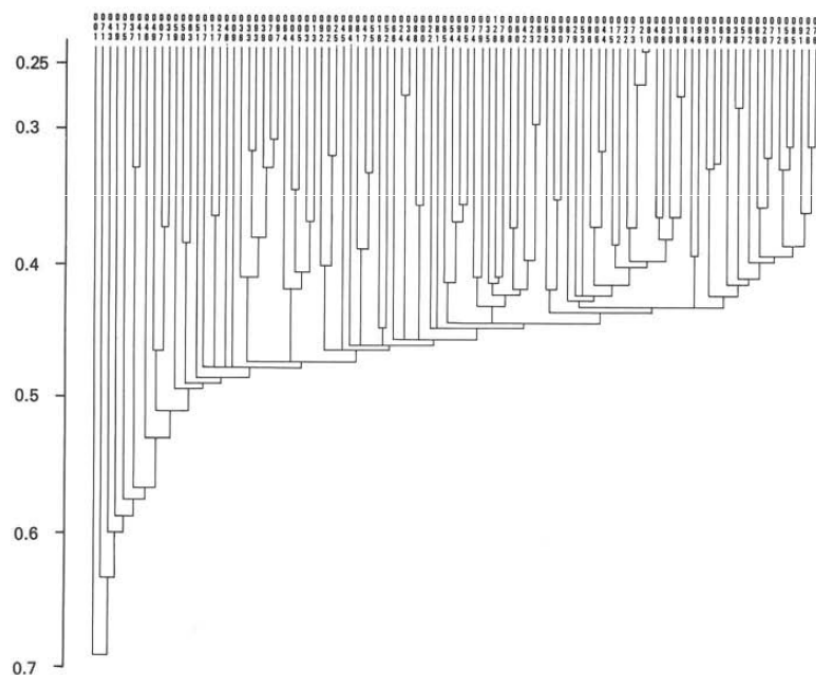
Datos sintéticos (4 clusters): Single-link



Clustering jerárquico



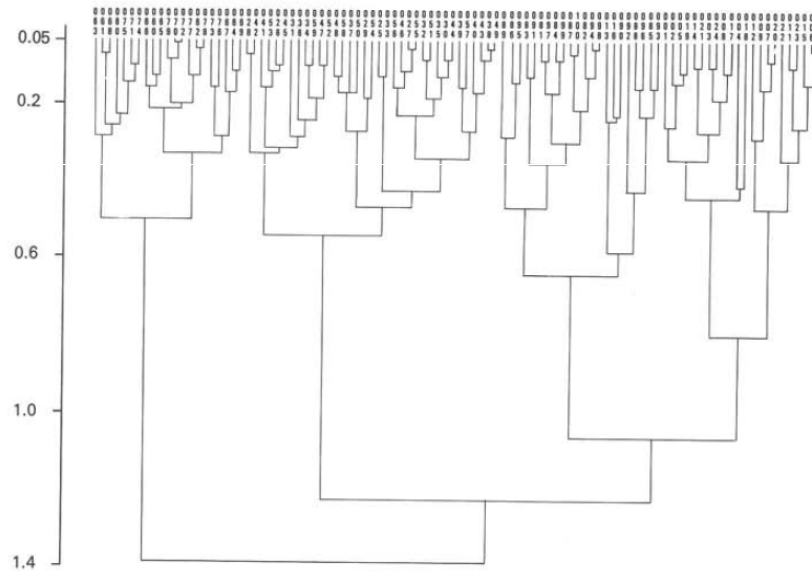
Datos sintéticos (aleatorios): Single-link



Clustering jerárquico



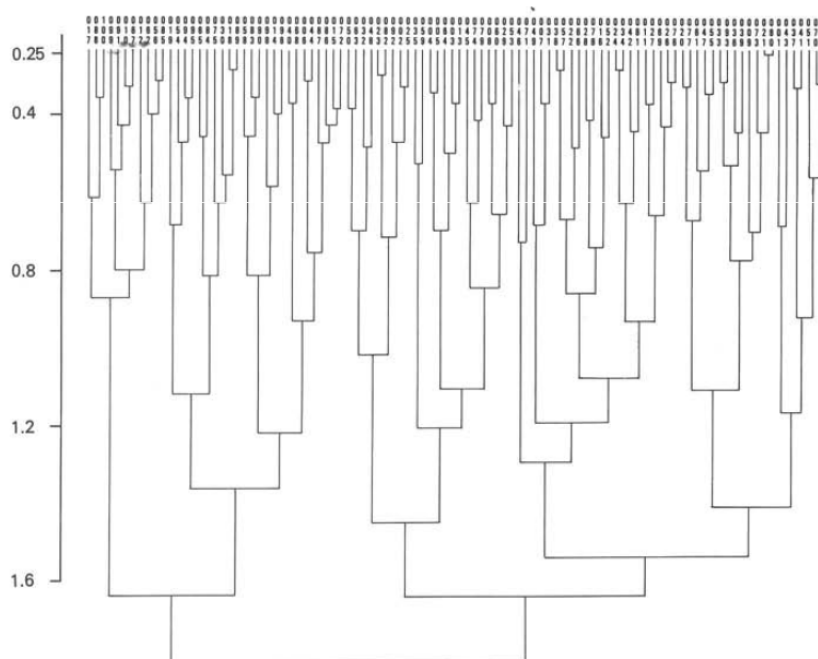
Datos sintéticos (4 clusters): Complete-link



Clustering jerárquico



Datos sintéticos (aleatorios): Complete-link



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).

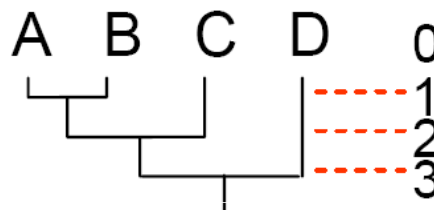


Clustering jerárquico

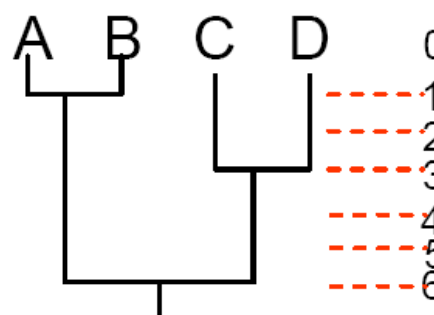


Ejercicio resuelto

- Single-link



- Complete-link



Clustering jerárquico



DEMO

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



Clustering jerárquico



Principal inconveniente del clustering jerárquico

Baja escalabilidad $\geq O(n^2)$

Por este motivo, si se usa un método jerárquico para estimar el número de grupos k (para un k -means), se suele emplear una muestra de los datos y no el conjunto de datos completo.



Clustering jerárquico



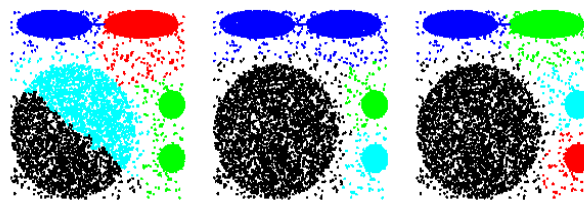
Algoritmos de clustering jerárquico

- **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)

✓ SPSS: Two-Step Clustering (variante de BIRCH)



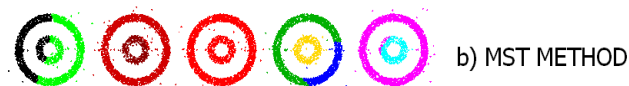
Clustering jerárquico



a) BIRCH b) MST METHOD c) CURE



a) BIRCH



b) MST METHOD



c) CURE



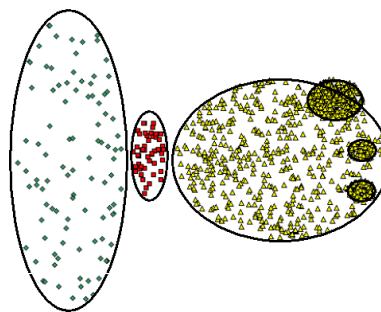
CURE



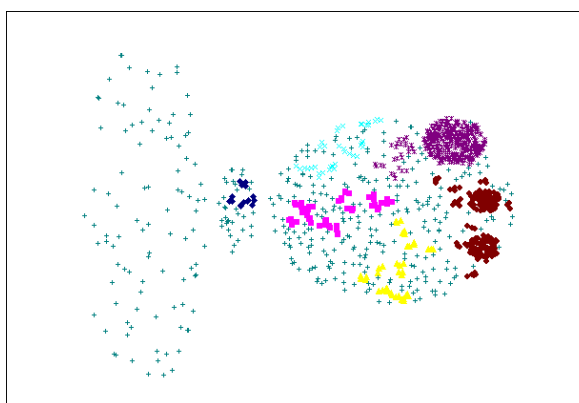
Clustering jerárquico



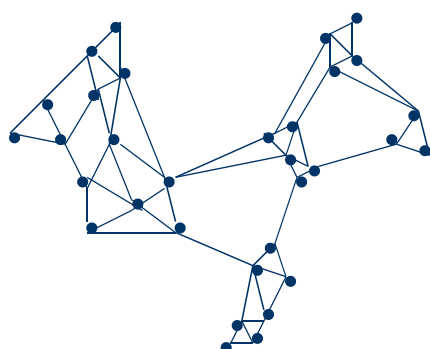
Agrupamientos con
varias densidades



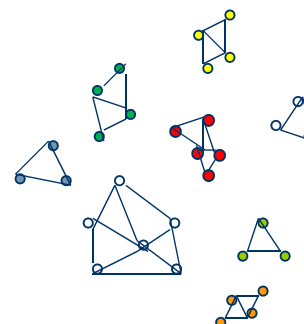
CURE



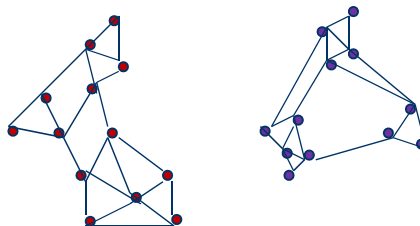
Clustering jerárquico



**Partición
del grafo**



**Combinar
particiones**



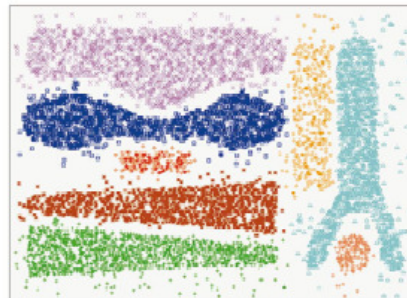
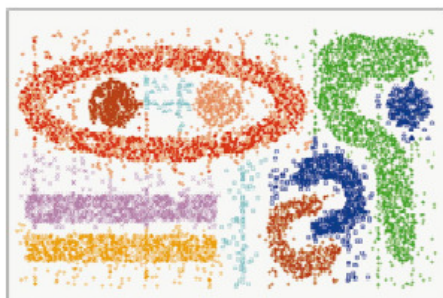
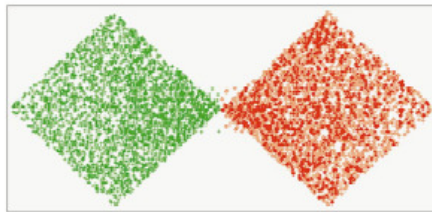
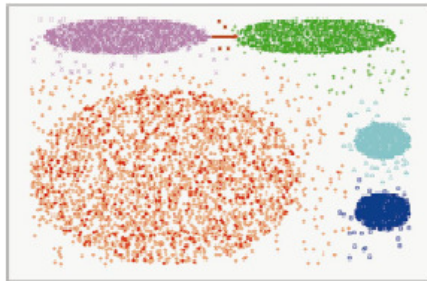
CHAMELEON



**Clusters
finales**



Clustering jerárquico



CHAMELEON



Validación de resultados



¿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 de resultados



- **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

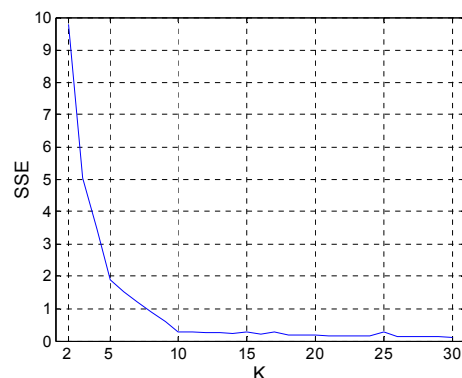
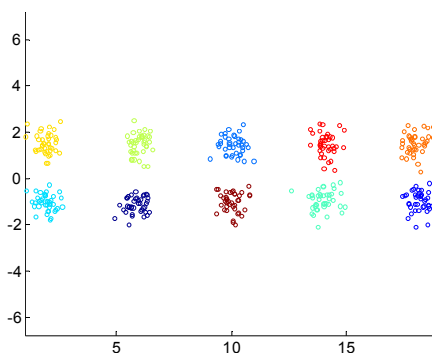
Otras medidas:
cohesión, separación, coeficientes de silueta...



Validación de resultados



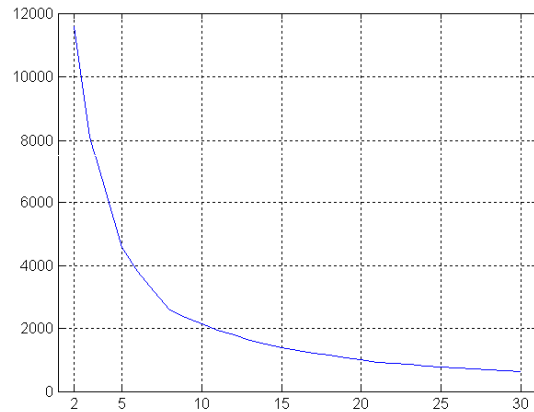
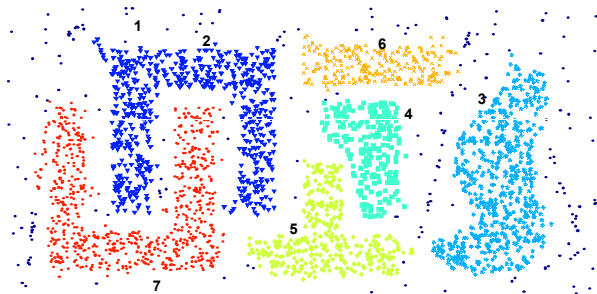
SSE ("Sum of Squared Error")



Validación de resultados



SSE ("Sum of Squared Error")



Algoritmos de clustering



Requisitos del algoritmo "perfecto"

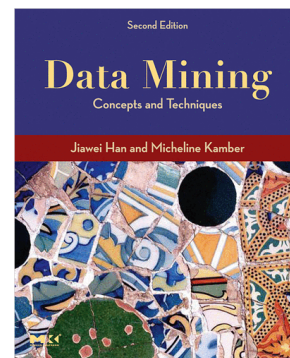
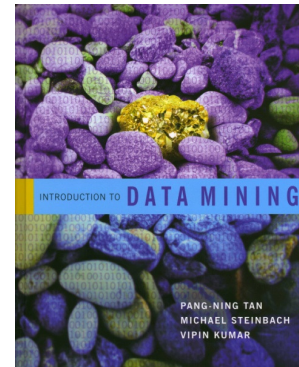
- Escalabilidad.
- Manejo de distintos tipos de datos.
- Identificación de clusters con formas arbitrarias.
- Número mínimo de parámetros.
- Tolerancia frente a ruido y outliers.
- Independencia con respecto al orden de presentación de los patrones de entrenamiento.
- Posibilidad de trabajar en espacios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



Bibliografía



- Pang-Ning Tan, Michael Steinbach & Vipin Kumar: **Introduction to Data Mining** Addison-Wesley, 2006. ISBN 0321321367 [capítulos 8&9]
- Jiawei Han & Micheline Kamber: **Data Mining: Concepts and Techniques** Morgan Kaufmann, 2006. ISBN 1558609016 [capítulo 7]



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.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 data clustering method for very large databases**. SIGMOD'96.



Apéndice: Notación O



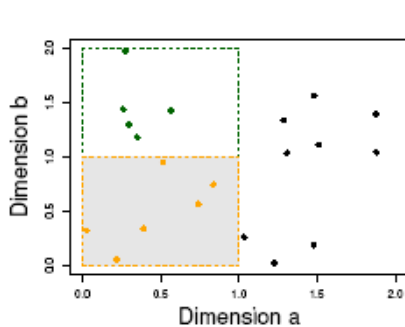
El impacto de la eficiencia de un algoritmo...

n	10	100	1000	10000	100000
O(n)	10ms	0.1s	1s	10s	100s
O(n·log₂ 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

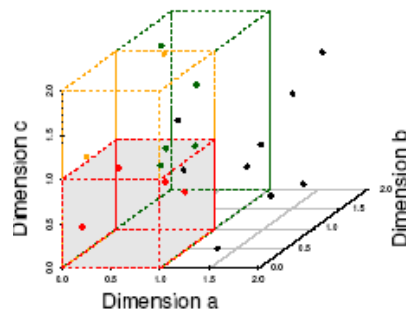


¿Por qué es un problema?

- Los datos en una dimensión están relativamente cerca
- Al añadir una nueva dimensión, los datos se alejan.
- Cuando tenemos muchas dimensiones, las medidas de distancia no son útiles ("equidistancia").



(b) 6 Objects in One Unit Bin



(c) 4 Objects in One Unit Bin



Posibles soluciones

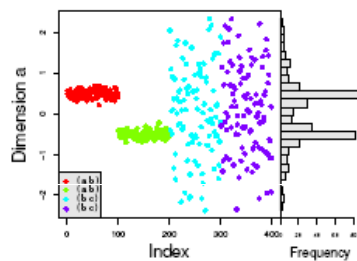
- **Transformación de características** (PCA, SVD) para reducir la dimensionalidad de los datos, útil sólo si existe correlación/redundancia.
- **Selección de características** (wrapper/filter) útil si se pueden encontrar clusters en subespacios.
- **"Subspace clustering"**
 Buscar clusters usando distintas combinaciones de atributos, vg. CLIQUE o PROCLUS.



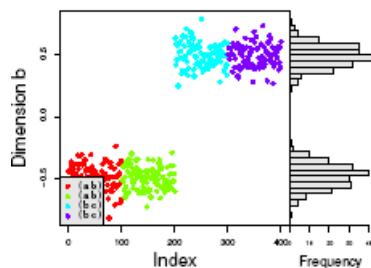
Apéndice

La dimensionalidad de los datos

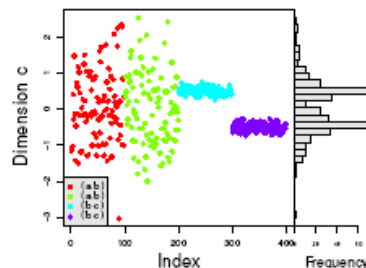
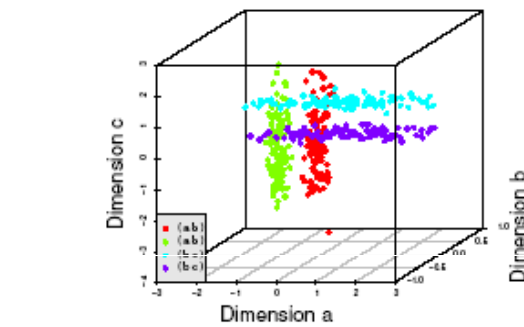
Subspace clustering



(a) Dimension *a*



(b) Dimension *b*



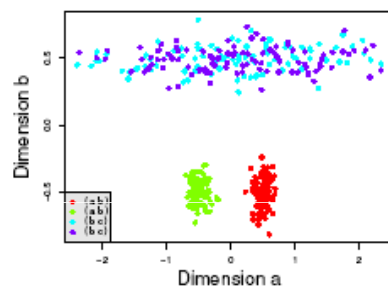
(c) Dimension *c*



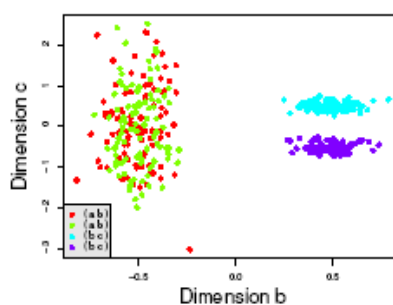
Apéndice

La dimensionalidad de los datos

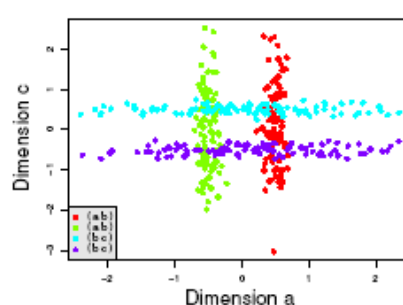
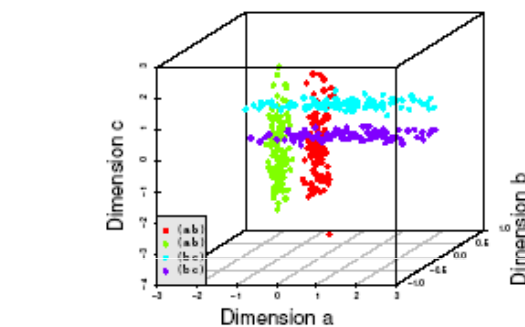
Subspace clustering



(a) Dims *a* & *b*



(b) Dims *b* & *c*



(c) Dims *a* & *c*





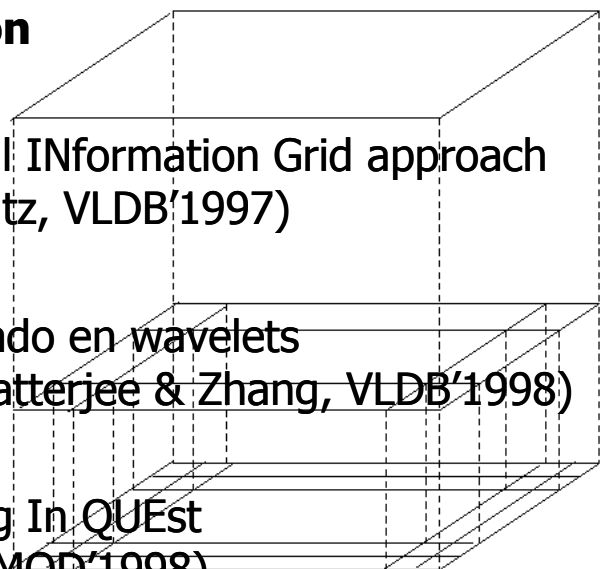
DEMO

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



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)



1st layer

(i-1)-st layer

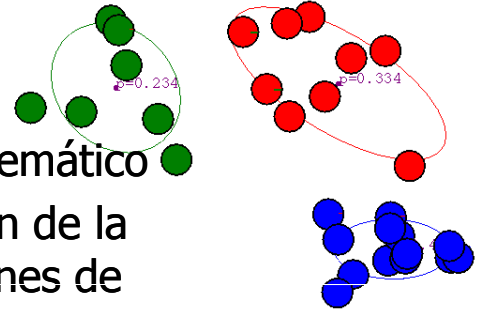
i-th layer





Clustering basado en modelos

Ajustar los datos a un modelo matemático
(se supone que los datos provienen de la
superposición de varias distribuciones de
probabilidades)

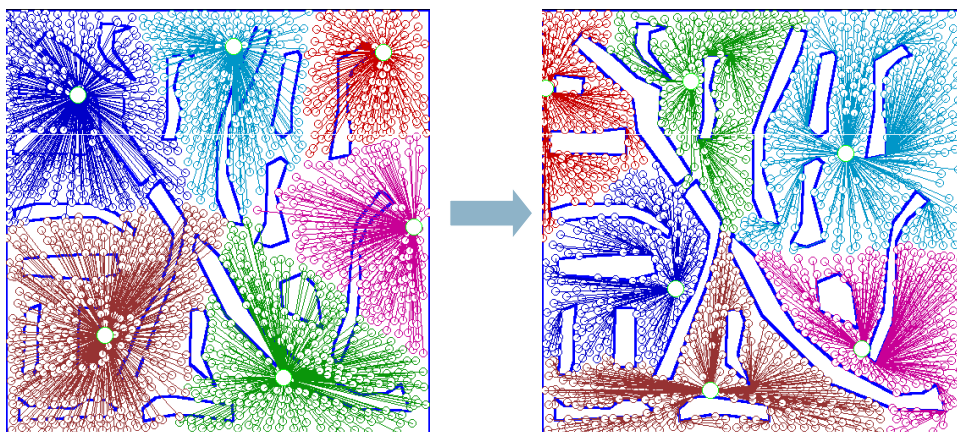


- Estadística:
EM [Expectation Maximization], **AutoClass**
- Clustering conceptual (Machine Learning):
COBWEB, **CLASSIT**
- Redes neuronales:
SOM [Self-Organizing Maps]



Clustering con restricciones

p.ej. Clustering con obstáculos



Posibles aplicaciones:
Distribución de cajeros automáticos/supermercados...

