

Clase 5

Aprendizaje no supervisado: clustering

Leticia C. Cagnina^{1,2}

¹Universidad Nacional de San Luis, Argentina

²Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
correo: lcagnina@unsl.edu.ar



Curso: Minería de Datos
Universidad Nacional de San Luis - Año 2018

Resumen

1 Medidas de Similitud y Disimilitud

- Background
- Medidas
- Similitud
- Distancias

2 Clustering: Conceptos básicos

- ¿qué es el Análisis de Clusters?
- Tipos de Clustering y de Clusters

3 Algoritmos de agrupamiento

- Algoritmos iterativos: *K-means*
- Algoritmo basado en densidad: DBSCAN

4 Validación de los agrupamientos

- Medidas de Validez Internas (MVI)
- Medidas de Validez Externas (MVE)

Introducción

En esta clase veremos dos conceptos importantes:

- Medidas de **proximidad** de objetos
- **Agrupamiento** de objetos

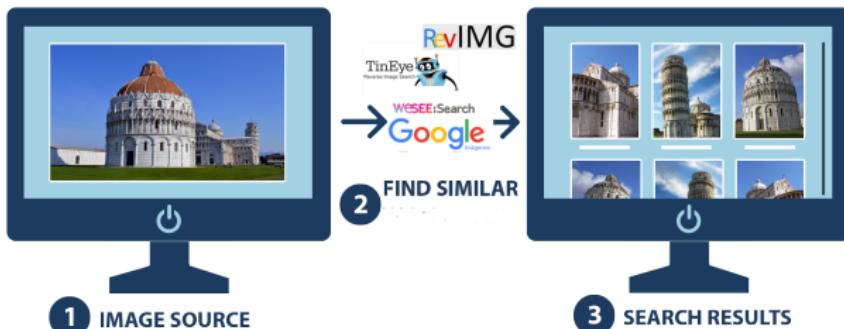
¿Qué es una medida de proximidad?

- Es un estimación numérica del grado de **semejanza o parecido** que tienen dos objetos.
 - Puede darse como medida de:
 - **Similitud**
 - **Disimilitud o distancia**

¿Por qué son importantes las medidas de proximidad?

Por su uso intensivo en muchas tareas que involucran la **comparación** de objetos de la Web:

- ### ● Recuperación de información



Background

¿Por qué son importantes las medidas de proximidad?

Por su uso intensivo en muchas tareas que involucran la **comparación** de objetos de la Web:

- Clustering



[Web](#) [Wiki](#) [Images](#) [News](#) [Yahoo](#)

salsa

[Search](#) [More options](#)

[Tree](#) [Visualization](#)

[All Topics \(100\)](#)
[Dance Salsa \(18\)](#)
[Salsa Classes \(10\)](#)
[Salsa Recipes \(7\)](#)
[Salsa Music \(6\)](#)
[Mambo \(5\)](#)
[Salsa Tanzschule \(4\)](#)

Cluster **Salsa Recipes** with 7 documents

1 [© Salsa Cycles 2008](#)   
Product index, **salsa** recipes, catalog
<http://www.salsacycles.com/> [Entire]

4 [Salsa Recipes - Dips and Spreads - All](#)
Looking for **salsa** recipes? Allrecipes ha

Background

¿Por qué son importantes las medidas de proximidad?

Por su uso intensivo en muchas tareas que involucran la **comparación** de objetos de la Web:

- ### • Sistemas recomendadores

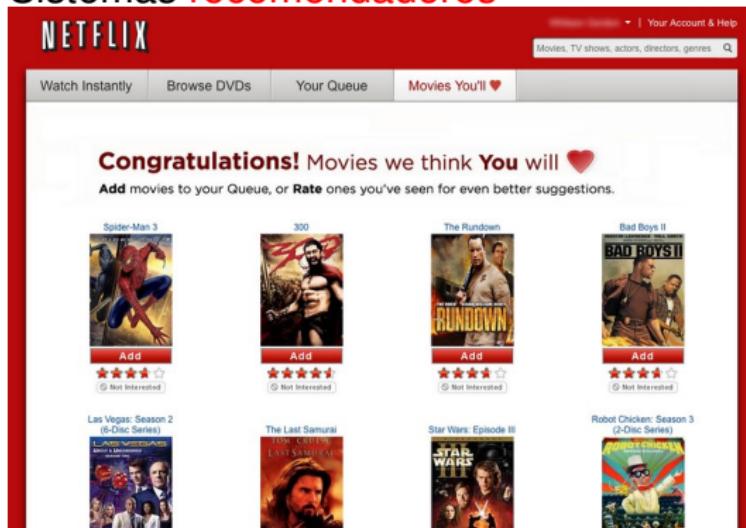


Background

¿Por qué son importantes las medidas de proximidad?

Por su uso intensivo en muchas tareas que involucran la **comparación** de objetos de la Web:

- Sistemas **recomendadores**



¿Por qué son importantes las medidas de proximidad?

Por su uso intensivo en muchas tareas que involucran la **comparación** de objetos de la Web:

- Detección de **plagio**:
 - Patentes
 - Código de software
 - Textos (reportes/artículos/tesis)
 - Música
- **Clasificación** (ej. k -vecinos más cercanos)

Medidas de Similitud y Disimilitud de Objetos

- Componente fundamental de cualquier algoritmo de clustering.
- Intentan estimar cuan **semejantes** o **diferentes** son dos objetos.
- Si $o_1, o_2 \in \mathcal{O}$ son (representaciones de) objetos,
- una función de similitud sim , es un mapping

$$sim : \mathcal{O} \times \mathcal{O} \mapsto [0, 1]$$

tal que:

- ➊ valores de $sim(o_1, o_2)$ cercanos a 1, indican que los objetos o_1 y o_2 son similares.
- ➋ valores de $sim(o_1, o_2)$ cercanos a 0, indican poca similitud entre o_1 y o_2 .

Medidas de Similitud y Disimilitud de Objetos

Medidas de Similitud

- Cuantifican la **semejanza** entre objetos.
- Ejemplos: **Matching Simple**, coeficiente de **Jaccard**, Similitud **Coseno**.

Medidas de Similitud y Disimilitud de Objetos

Medidas de Similitud

- Cuantifican la **semejanza** entre objetos.
- Ejemplos: **Matching Simple**, coeficiente de **Jaccard**, Similitud **Coseno**.

Medidas de Disimilitud

- Cuantifican cuan **diferentes** son dos objetos.
- Ejemplos: distintas distancias: **Euclídea**, **Minkowski** y sus variantes.

Medidas de Similitud y Disimilitud de Objetos

Medidas de Similitud

- Cuantifican la **semejanza** entre objetos.
- Ejemplos: **Matching Simple**, coeficiente de **Jaccard**, Similitud **Coseno**.

Medidas de Disimilitud

- Cuantifican cuan **diferentes** son dos objetos.
- Ejemplos: distintas distancias: **Euclíadiana**, **Minkowski** y sus variantes.

Medidas de Proximidad

Referencian tanto a las medidas de **similitud** como **disimilitud**.

Propiedades de las medidas de similitud

- ① $sim(p, q) = 1$ (o **máxima semejanza**) sólo si $p = q$
- ② $sim(p, q) = sim(q, p)$ para todo p y q (**Simetría**)

donde $sim(p, q)$ es la similitud (o semejanza) entre los puntos (objetos de datos), p y q .

Medidas de Similitud basadas en Conjuntos

Idea

Dos objetos $o_i, o_j \in \mathcal{O}$ son representados por los conjuntos O_i, O_j (por ej. si o_i es un documento, O_i podría ser el conjunto de sus términos). Las similitudes se basan en distintas ponderaciones de la intersección de conjuntos.

Medidas de Similitud basadas en Conjuntos

Idea

Dos objetos $o_i, o_j \in \mathcal{O}$ son representados por los conjuntos O_i, O_j (por ej. si o_i es un documento, O_i podría ser el conjunto de sus términos). Las similitudes se basan en distintas ponderaciones de la intersección de conjuntos.

Ejemplos:

- Matching simple: $\varphi_{matsim}(O_i, O_j) = \frac{|O_i \cap O_j|}{\max(|O_i|, |O_j|)}.$
- Coeficiente de Jaccard: $\varphi_{jacc}(O_i, O_j) = \frac{|O_i \cap O_j|}{|O_i \cup O_j|}.$

Similitud

Similitud para objetos representados con valores binarios

		q	
		1	0
		M_{11}	M_{10}
p	1	M_{01}	M_{00}
	0		

M_{xy} = número de atributos donde p es x y q es y

Similitud para objetos representados con valores binarios

		<i>q</i>
	1	1 0
<i>p</i>	1	M_{11} M_{10}
	0	M_{01} M_{00}

M_{xy} = número de atributos donde p es x y q es y

- ① Matching Simple:

$$\text{sim}(p, q) = \frac{M_{00} + M_{11}}{M_{00} + M_{11} + M_{01} + M_{10}}$$

- ② Coeficiente de Jaccard:

$$\text{sim}(p, q) = \frac{M_{11}}{M_{11} + M_{01} + M_{10}}$$

Medidas de Similitud Geométricas

Idea

Dos objetos $o_i, o_j \in \mathcal{O}$ son comparados usando sus representaciones vector $\vec{o_i}, \vec{o_j}$, y su similitud se estima en base a la amplitud del ángulo formado por ambos vectores.

Medidas de Similitud Geométricas

Idea

Dos objetos $o_i, o_j \in \mathcal{O}$ son comparados usando sus representaciones vector $\vec{o_i}, \vec{o_j}$, y su similitud se estima en base a la amplitud del ángulo formado por ambos vectores.

Ejemplo:

La función de **Similitud Coseno** $sim_{cos} : \mathbb{R}^m \times \mathbb{R}^m \mapsto [0, 1]$

$$sim_{cos}(\vec{p}, \vec{q}) = \frac{\vec{p} \cdot \vec{q}}{\|\vec{p}\| \|\vec{q}\|}$$

Propiedades de las medidas de distancia

- ① a) $\forall o_i, o_j \in \mathcal{O} : dis(o_i, o_j) \geq 0$
b) $\forall o_i, o_j \in \mathcal{O} : dis(o_i, o_j) = 0$ sólo si $o_i = o_j$ (**positividad**)
- ② $\forall o_i, o_j \in \mathcal{O} : dis(o_i, o_j) = dis(o_j, o_i)$ (**simetría**)
- ③ $\forall o_i, o_j, o_k \in \mathcal{O} : dis(o_i, o_j) + dis(o_j, o_k) \geq dis(o_i, o_k)$
(**desigualdad triangular**)

donde $dis(p, q)$ es la distancia (o disimilitud) entre los puntos (objetos de datos), p y q .

Una distancia que satisface estas propiedades es una **métrica**.

Distancia Euclídea

Sean $p = \langle p_1, p_2, \dots, p_k \rangle$ y $q = \langle q_1, q_2, \dots, q_k \rangle$ dos **objetos** (representados como **vectores**), la **distancia euclídea** es:

$$dis(p, q) = \sqrt{\sum_{j=1}^k (p_j - q_j)^2} \quad (1)$$

Es decir, es la distancia **ordinaria** entre los dos puntos geométricos (como se deriva del **teorema de Pitágoras**)

Distancias

Distancia de Minkowski

$$dis(p, q) = \left(\sum_{j=1}^k |p_j - q_j|^r \right)^{\frac{1}{r}} \quad (2)$$

donde $r \geq 1$

- ① $r = 2$ (Distancia *Euclídea*)
- ② $r = 1$ (Distancia *Manhattan* o *Hamming*)

$$dis(p, q) = \sum_{j=1}^k |p_j - q_j| \quad (3)$$

- ③ $r \rightarrow \infty$ (Distancia *Cheby Shov* (*dominante* o *máxima*))

$$dis(p, q) = \max_{1 \leq j \leq k} |p_j - q_j| \quad (4)$$

¿qué es el Análisis de Clusters?

Definición

Análisis de Clusters

Proceso que divide los datos en **grupos (clusters)** que tienen un **significado**, que son **útiles**, o ambos.

- Grupos **significativos** ⇒ Los grupos deberían capturar la **estructura natural** de los datos.
- Grupos **útiles** ⇒ Los grupos sirven de base para otras técnicas de análisis y procesamiento de datos.

¿qué es el Análisis de Clusters?

Análisis de Clusters

Grupos significativos

- Estos grupos mejoran nuestro **entendimiento** de los datos y las clases subyacentes.
- Rol fundamental en Biología, Recuperación de Información, Meteorología, Psicología y Medicina, y Negocios.

Grupos útiles

- Hincapié en encontrar **prototipos de clusters** (objetos de datos representativos de los otros objetos del cluster).
- Rol fundamental en **resumir** grandes conjuntos de datos, **compresión** de imagen y sonido, y búsqueda NN eficiente.

¿qué es el Análisis de Clusters?

Definición (más operativa)

Análisis de Clusters

Encontrar grupos de objetos tal que los de un mismo grupo sean **similares** (o están **relacionados**) y sean **diferentes** (o están **poco relacionados**) con los objetos de los otros grupos.

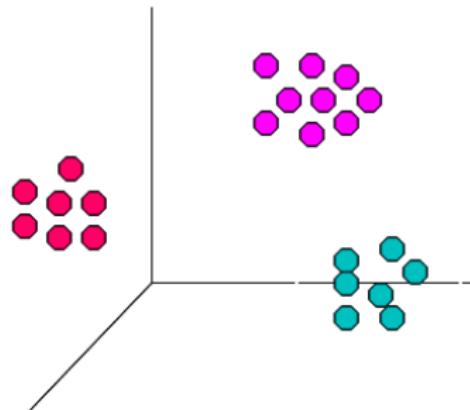
- También conocida como **categorización no supervisada**.
- Áreas conectadas (pero no iguales) al Análisis de Cluster:
 - **Particionado** ⇒ usualmente relacionado al particionado de grafos en **subgrafos**.
 - **Segmentado** ⇒ división de grupos mediante técnicas muy simples (ejemplo: segmentado de imágenes basado en el color y la intensidad de los pixels, o de personas de acuerdo a su ingreso).

¿qué es el Análisis de Clusters?

Definición (más operativa)

Análisis de Clusters

Encontrar grupos de objetos tal que los de un mismo grupo sean similares (o están relacionados) y sean diferentes (o están poco relacionados) con los objetos de los otros grupos.

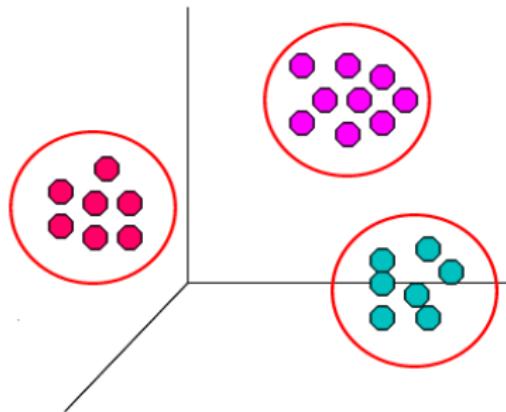


¿qué es el Análisis de Clusters?

Definición (más operativa)

Análisis de Clusters

Encontrar **grupos** de objetos tal que los de un mismo grupo sean similares (o están relacionados) y sean diferentes (o están poco relacionados) con los objetos de los otros grupos.

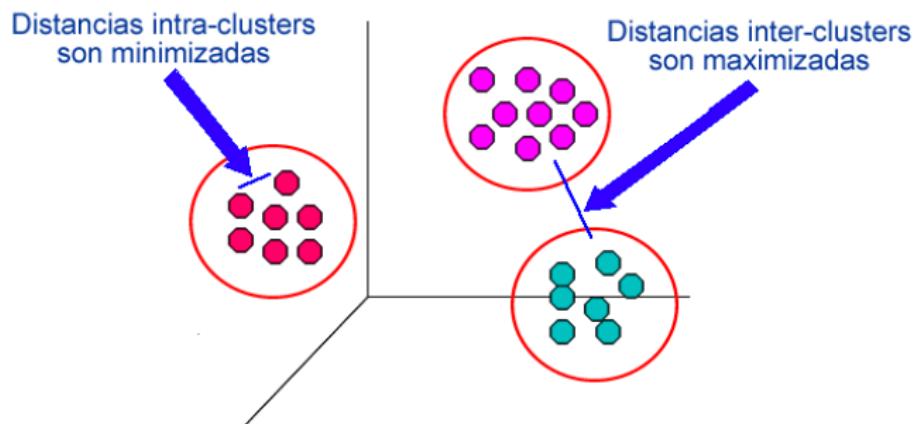


¿qué es el Análisis de Clusters?

Definición (más operativa)

Análisis de Clusters

Encontrar grupos de objetos tal que los de un mismo grupo sean similares (o están relacionados) y sean **diferentes** (o están **poco relacionados**) con los objetos de los **otros grupos**.



¿qué es el Análisis de Clusters?

La noción de cluster es ambigua....



¿Cuántos Clusters?

¿qué es el Análisis de Clusters?

La noción de cluster es ambigua....



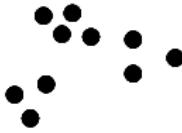
¿Cuántos Clusters?



Dos Clusters

¿qué es el Análisis de Clusters?

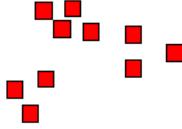
La noción de cluster es ambigua....



¿Cuántos Clusters?



Cuatro Clusters

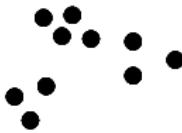


Dos Clusters



¿qué es el Análisis de Clusters?

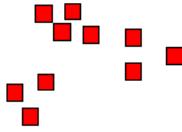
La noción de cluster es ambigua....



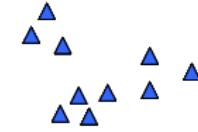
¿Cuántos Clusters?



Cuatro Clusters



Dos Clusters



Seis Clusters

Tipos de clusterings (agrupamientos)

Un **clustering (agrupamiento)** es un conjunto de clusters.

Principal **distinción** entre tipos de agrupamientos:

Clustering Particional

Los objetos de datos se dividen en subconjuntos (clusters) no solapados, tal que cada objeto pertenece a exactamente un subconjunto.

Clustering Jerárquico

Conjunto de clusters anidados organizados como un árbol jerárquico.

Tipos de Clustering y de Clusters

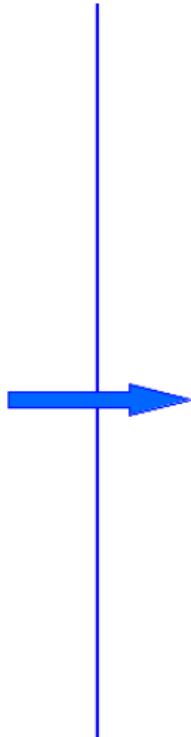
Clustering Particional



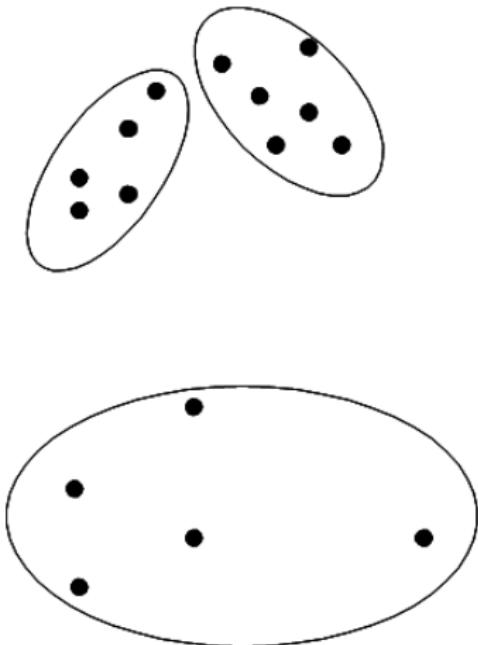
Puntos Originales

Tipos de Clustering y de Clusters

Clustering Particional

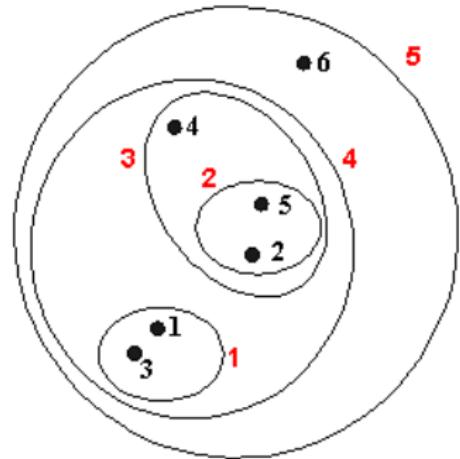
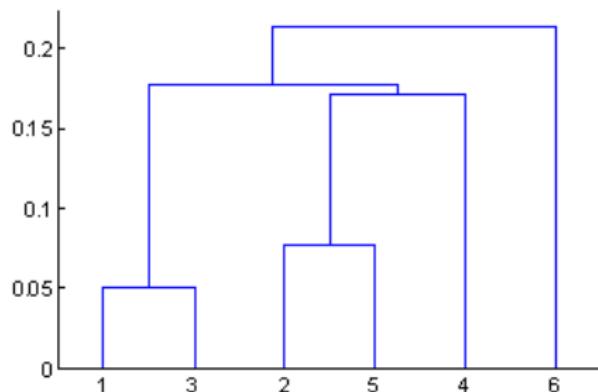


Puntos Originales

Un Clustering
Particional

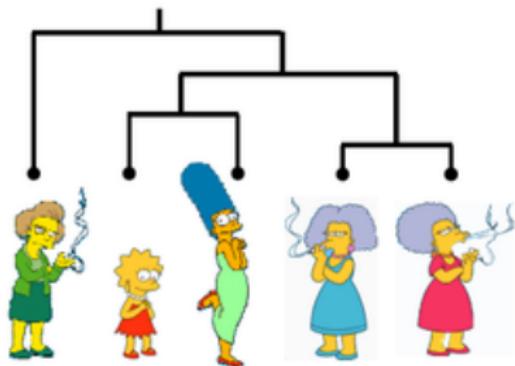
Tipos de Clustering y de Clusters

Clustering Jerárquico

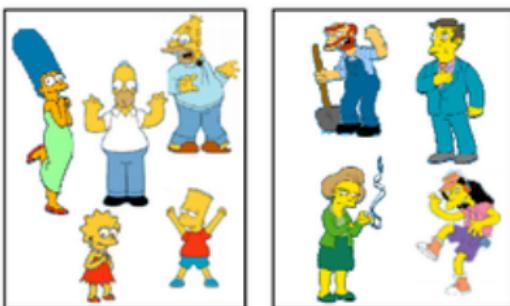


Clustering Jerárquico vs Particional

Hierarchical

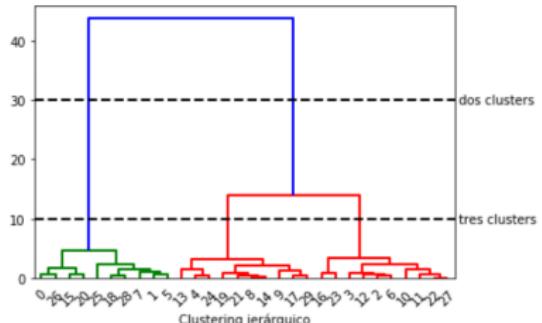
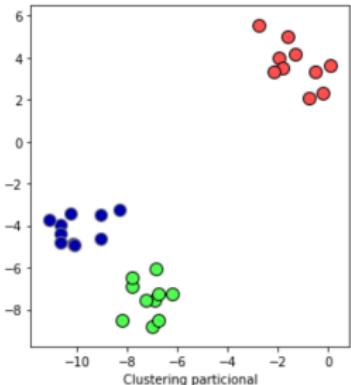
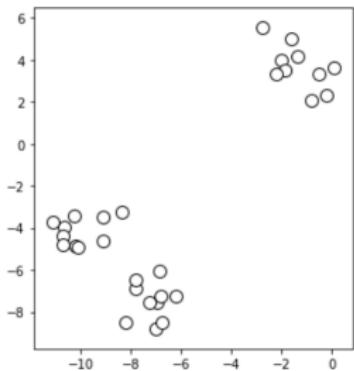


Partitional



Tipos de Clustering y de Clusters

Clustering Jerárquico vs Particional



Definiendo agrupamiento particional formalmente...

Dada una colección \mathcal{O} de objetos

Encontrar una partición \mathcal{C} de \mathcal{O} , $\mathcal{C} = \{C_1, \dots, C_k\}$ que cumple:

- ① $C_i \cap C_j = \emptyset, \forall C_i \neq C_j, C_i, C_j \in \mathcal{C}$
- ② $\bigcup_{k=1}^{i=1} C_i = \mathcal{O}$
- ③ $\forall C_i \in \mathcal{C}, C_i \neq \emptyset$

Otras distinciones de los agrupamientos

Exclusivo vs no exclusivo (NE)

En agrupamientos NE los objetos de datos pueden pertenecer a múltiples clusters.

Difuso vs no difuso

Agrupamiento **difuso**:

- Un objeto pertenece a cada cluster con un peso $w_i \in [0, 1]$
- Pesos deben sumar 1.
- Clustering probabilístico tiene características similares.

Parcial vs completo

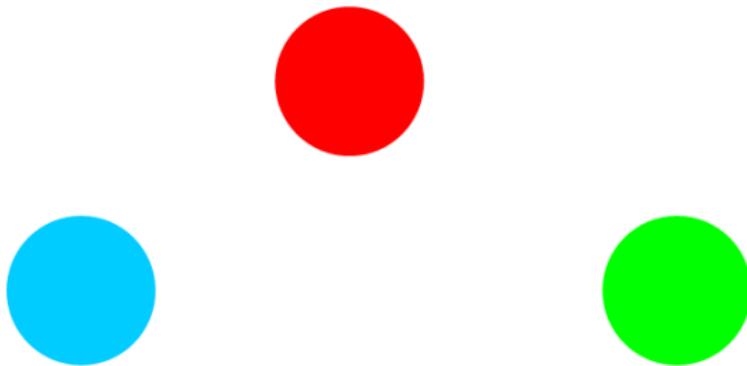
En agrupamientos parciales algunos puntos pueden quedar sin clasificar.

Tipos de Grupos (clusters)

Existen **distintas concepciones** respecto a lo que constituye un grupo:

- Grupos bien separados
- Grupos basados en centros
- Grupos basados en grafos (o contiguidad)
- Grupos basados en densidad
- Grupos conceptuales
- Grupos definidos por una función objetivo

Grupos bien separados

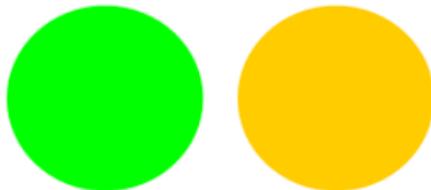
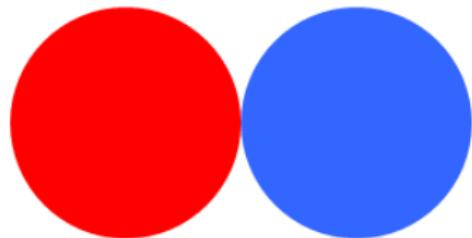


Definición

Cada objeto es **más cercano** (o más similar) a cada objeto en su cluster que a cualquier objeto en otro cluster.

- Sólo se cumple con grupos **bien separados**.
- Los clusters pueden tener **cualquier forma**.

Grupos basados en centros



Definición

Cada objeto es **más cercano** (o más similar) al **centro** de su cluster **que al centro** de cualquier **otro cluster**.

- También llamados basados en **prototipos** (**centroide** o **medoide**).
- Tienden a ser **globulares**.

Grupos basados en grafos (contiguidad)

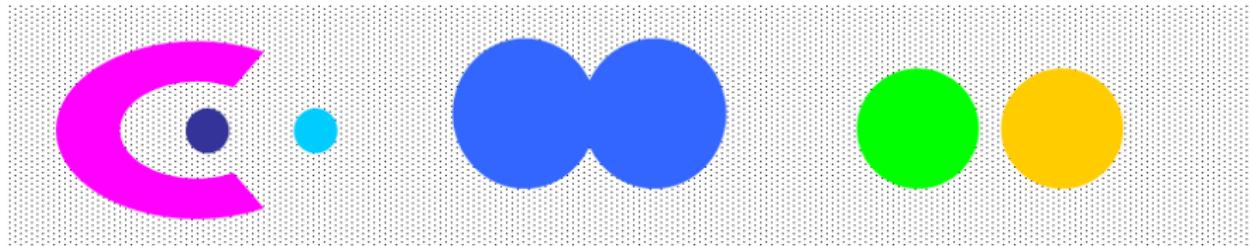


Definición

Cada objeto es **más cercano** (o más similar) a **al menos un punto** en su cluster **que a cualquier otro punto en otro cluster**.

- Útiles cuando los grupos son **irregulares** (o “superpuestos”).
- Problema: **ruido** puede unir clusters distintos.

Grupos basados en densidad

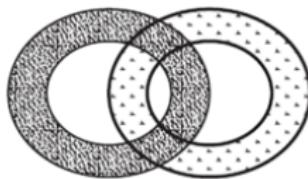
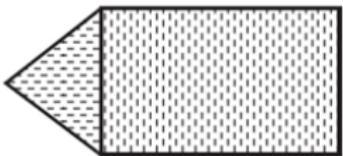


Definición

Los grupos son regiones de **alta densidad** separados por regiones de **baja densidad**.

- Útiles cuando los grupos son **irregulares** (o “superpuestos”), cuando hay **ruido** y “outliers”.

Grupos conceptuales (o propiedad compartida)



Definición

Los objetos del grupo comparten alguna **propiedad general** que se deriva del conjunto completo de objetos.

- Ejemplos: a) un área triangular (cluster) es adyacente a una rectangular (cluster). b) dos círculos superpuestos.
- Conceptos muy sofisticados ⇒ reconocimiento de patrones.

Grupos definidos por una **función objetivo**

- ➊ **Idea:** encontrar clusters que minimizan o maximizan una función objetivo.
- ➋ Proceso: enumerar todas las maneras posibles de dividir los puntos en clusters y evaluar “cuan buenos” son los agrupamientos resultantes usando la función objetivo dada (problema NP duro).
- ➌ En la práctica, se usan algoritmos de optimización que logran soluciones **buenas** aunque no necesariamente óptimas.

K-means

- Algoritmo de agrupamiento **particional**.
- Cada grupo tiene asociado un **centroide** (punto central).
- Cada punto es asignado al grupo del **centroide más cercano (CMC)**.
- El número de grupos, K , debe ser especificado.
- El algoritmo básico es muy simple.

Algoritmos iterativos: *K*-means

***K*-means**

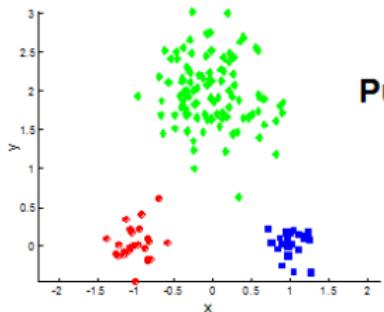
- Algoritmo de agrupamiento **particional**.
- Cada grupo tiene asociado un **centroide** (punto central).
- Cada punto es asignado al grupo del **centroide más cercano (CMC)**.
- El número de grupos, K , debe ser especificado.
- El algoritmo básico es muy simple.

Algoritmo

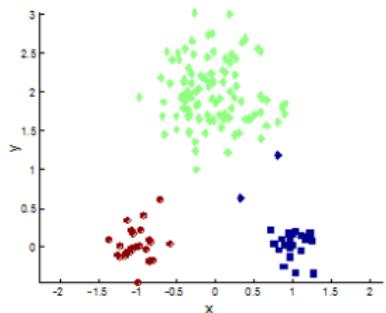
- 1 Seleccionar K puntos como los centroides iniciales
- 2 **repetir**
- 3 Formar K grupos asignando todos los puntos al CMC
- 4 Recacular el centroide de cada grupo
- 5 **hasta que** los centroides no cambien

Algoritmos iterativos: *K*-means

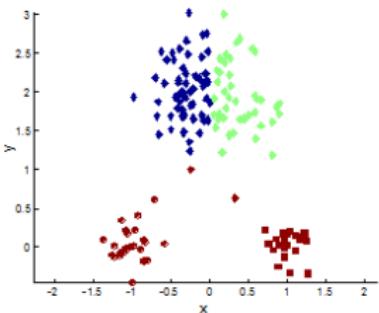
Dos agrupamientos diferentes generados por *K*-means



Puntos originales



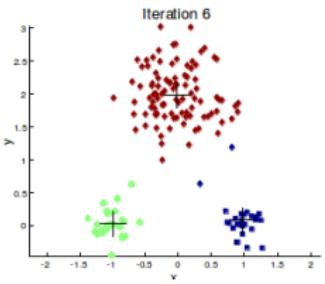
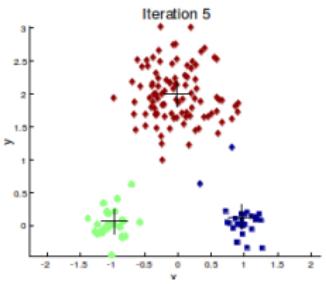
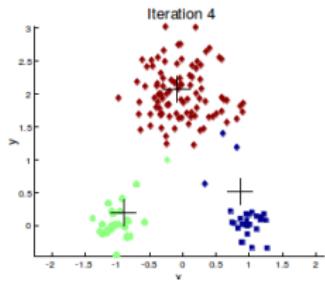
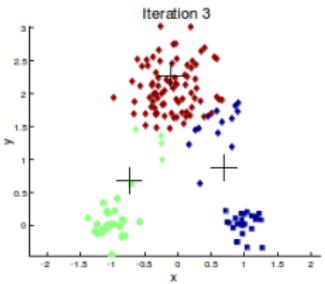
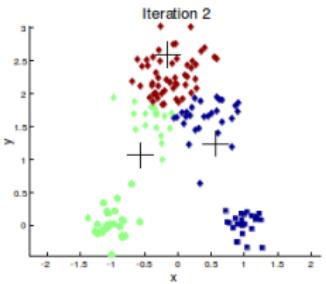
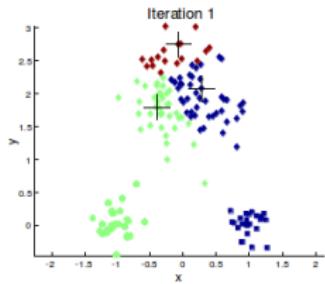
Clustering óptimo



Clustering sub-óptimo

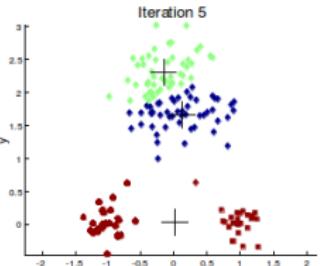
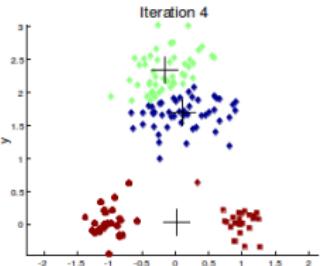
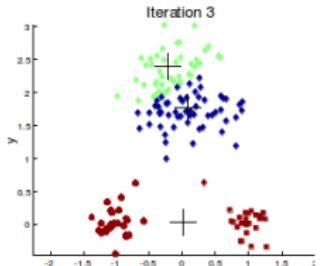
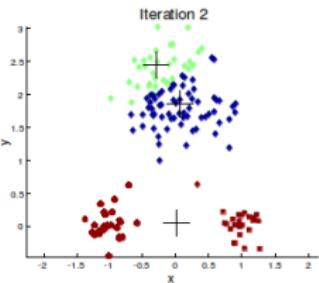
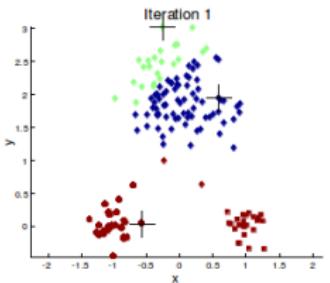
Algoritmos iterativos: *K*-means

Importancia de los centroides iniciales (1)



Algoritmos iterativos: *K*-means

Importancia de los centroides iniciales (2)



Algoritmos iterativos: K-means

Importancia de los centroides iniciales (3)

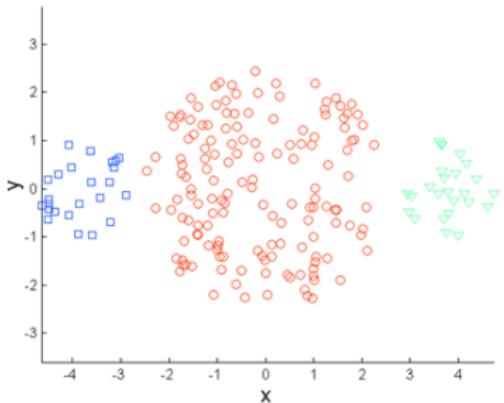
- El centroide depende de la función de distancia
- La cercanía es medida por la distancia Euclídea, similitud coseno, correlación, etc.
- Media de los puntos o punto más representativo (centroide/medoide)
- Encontrar un buen centroide puede ser un problema NP-duro

Limitaciones de k-means

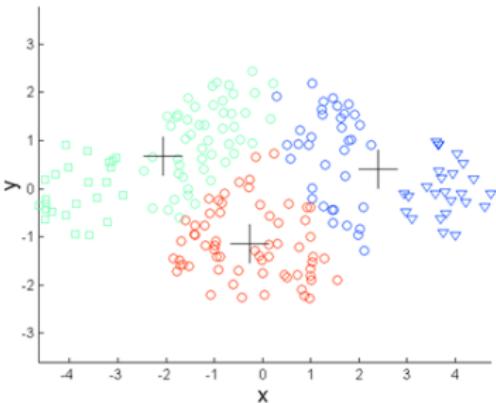
- Grupos de diferentes tamaños (desbalance)
- Grupos con diferentes densidades
- Grupos no globulares
- Datos con outliers

Algoritmos iterativos: *K*-means

Grupos de diferentes tamaños



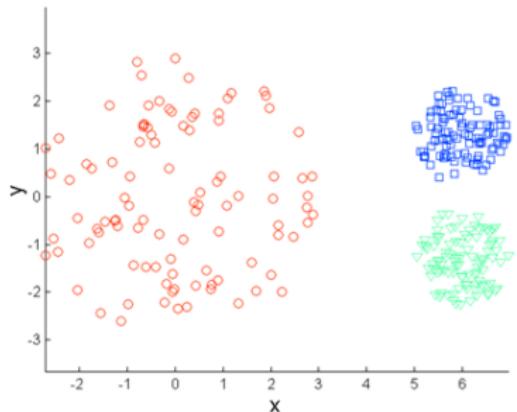
Puntos Originales



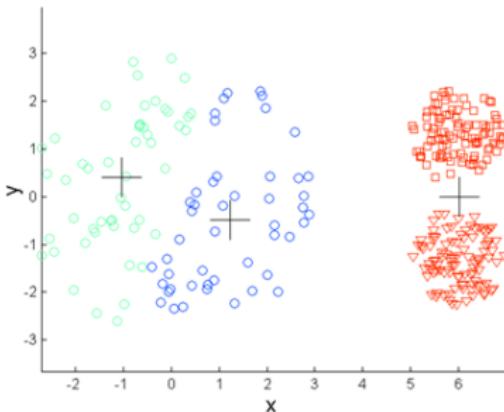
k-means ($k=3$)

Algoritmos iterativos: *K*-means

Grupos de diferentes densidades



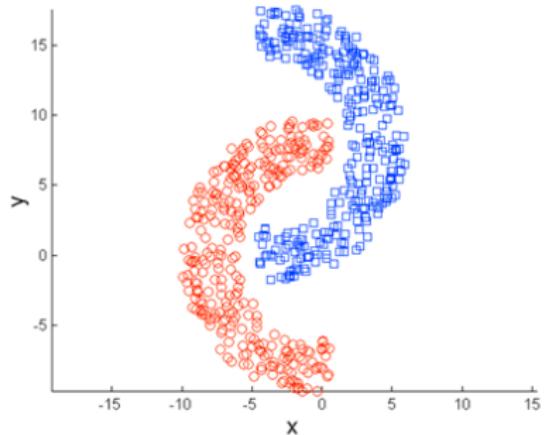
Puntos originales



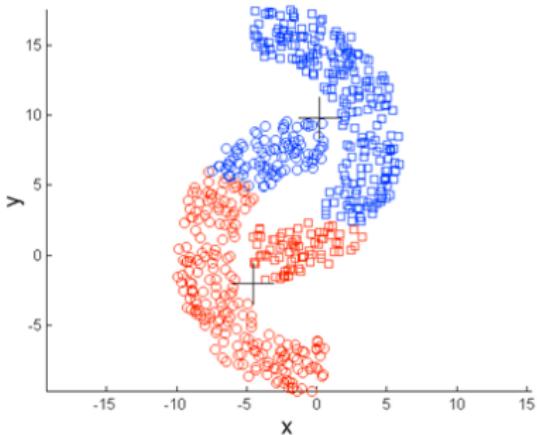
K-means ($k=3$)

Algoritmos iterativos: *K*-means

Grupos no globulares



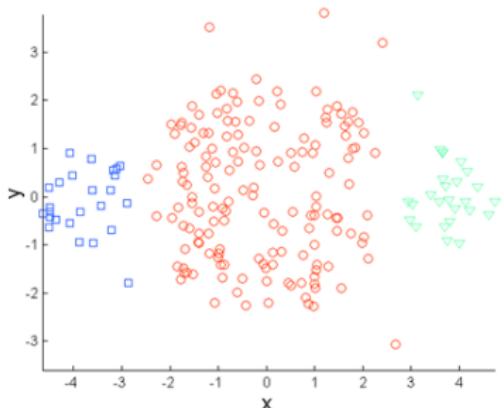
Puntos originales



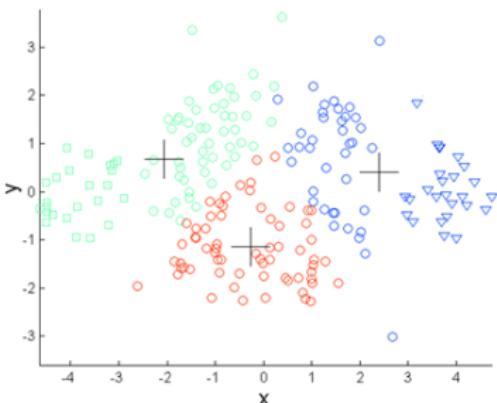
K-means ($k=2$)

Algoritmos iterativos: *K*-means

Grupos con outliers



Puntos Originales



k-means (k=3)

Algoritmo basado en densidad: DBSCAN

DBSCAN

- Algoritmo de agrupamiento **particional** basado en densidad.
- Se determinan los grupos (automáticamente) en base a las regiones densas de datos.
- La *densidad en el punto p* es el número de puntos dentro del círculo con radio **Eps**.
- Una *región densa* es un círculo con radio **Eps** que contiene al menos **MinPts** puntos.
- Puntos en regiones de baja densidad son catalogados como **ruido** (**outliers**).
- Puede producir agrupamientos incompletos.

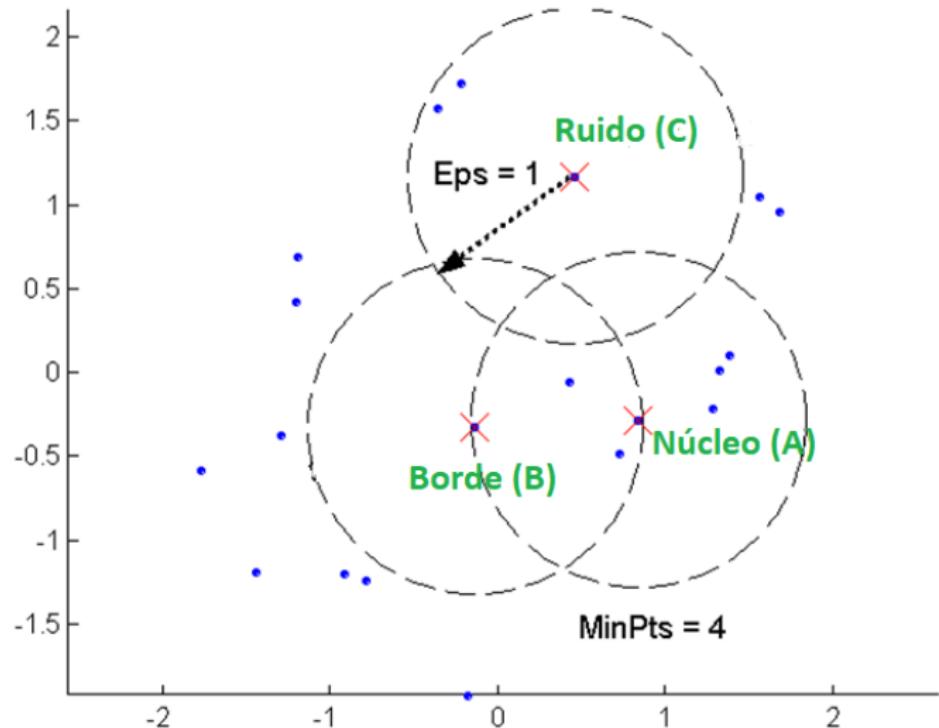
Algoritmo basado en densidad: DBSCAN

Caracterización de los puntos

- A** Puntos del **núcleo**: puntos del interior del grupo con vecindario $MinPts$ en un círculo de radio Eps .
- B** Puntos del **borde**: no son puntos del núcleo pero forman parte del vecindario de aquellos (incluso pueden caer en vecindarios de otros núcleos).
- C** Puntos **ruido**: puntos que no cumplen con la caracterización de ser núcleo ni borde.

Algoritmo basado en densidad: DBSCAN

Puntos en DBSCAN

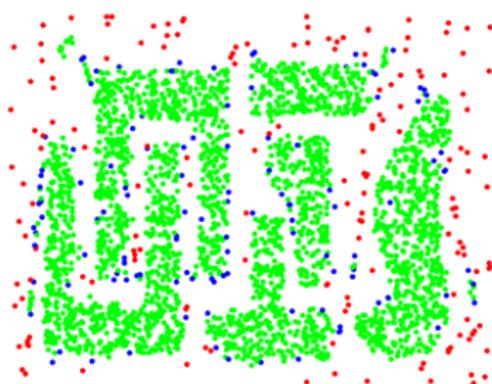


Algoritmo basado en densidad: DBSCAN

Puntos en DBSCAN



Puntos originales



Puntos Núcleos, bordes
y ruido.

Eps=10 MinPts=4

Algoritmo basado en densidad: DBSCAN

DBSCAN

Algoritmo

- 1 Etiquetar los puntos como *núcleo*, *borde* o *ruido*
- 2 Eliminar los puntos *ruido*
- 3 Por cada punto *núcleo* p que no ha sido asignado a un grupo
 - 4 Crear un arco que conecte ese punto con otros dentro del vecindario con radio Eps
 - 5 Crear un grupo con todos los puntos conectados
- 6 Asignar cada punto *borde* a uno de los grupos de sus núcleos asociados.

Algoritmo basado en densidad: DBSCAN

Fortalezas y Limitaciones de DBSCAN

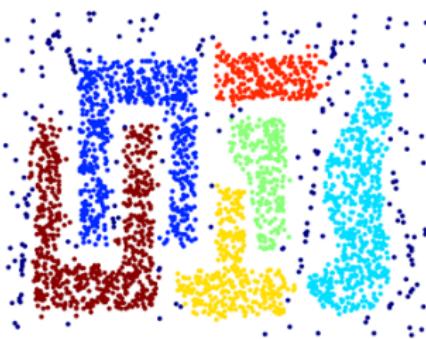
- Puede manejar grupos de diferentes tamaños y formas
- Es resistente a los outliers
- No puede manejar grupos con distintas densidades
- No suele utilizarse con datos de alta dimensionalidad (cálculo de densidad)
- Suele ser costoso debido al cálculo de los vecindarios

Algoritmo basado en densidad: DBSCAN

Grupos de diferentes tamaños y formas



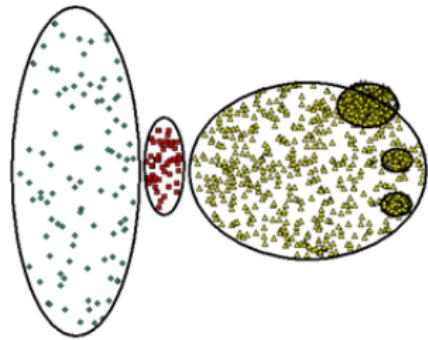
Puntos originales



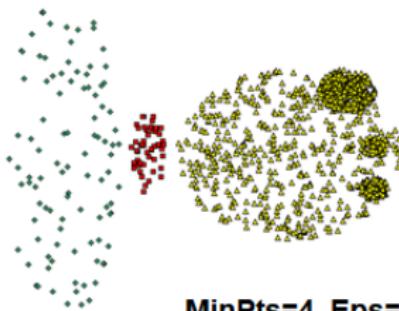
DBSCAN (6 grupos)

Algoritmo basado en densidad: DBSCAN

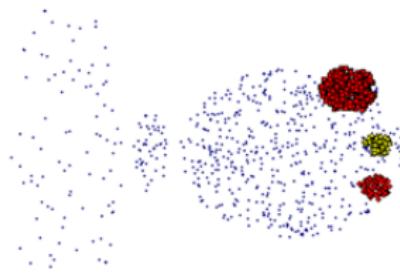
Grupos de diferentes densidades y alta dim



Puntos originales



MinPts=4, Eps=9.75



MinPts=4, Eps=9.92

Validación de los grupos (o agrupamientos)

Evaluación (o validación) de grupos

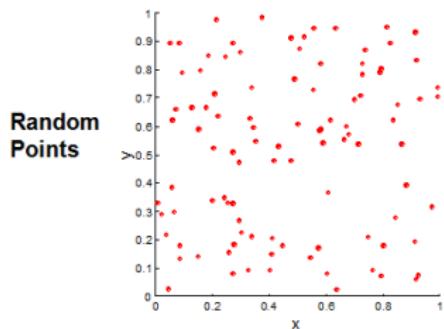
Parte **fundamental** aunque poco explorada del análisis de grupos (cluster analysis).

Incluye

- Determinar la **tendencia de clustering**.
- Determinar el **número correcto de clusters**.
- Evaluar **cuan bien** los resultados del análisis de clusters (AC) se ajustan a los datos **sin** referencia a información externa.
- Comparar los resultados del AC con resultados conocidos **externamente**.
- Comparar dos conjuntos de clusters para determinar cual es mejor.

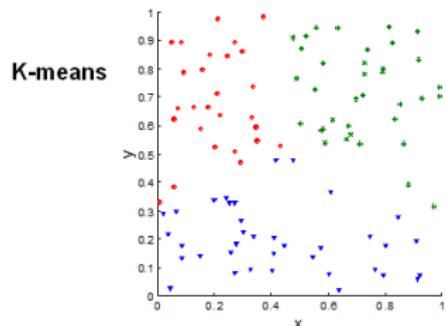
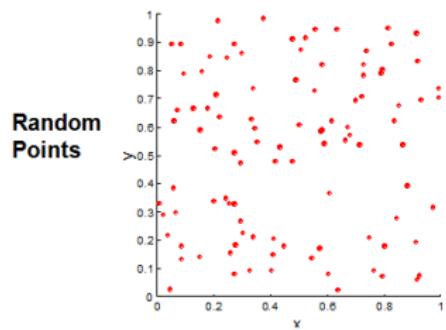
Validación de los grupos (o agrupamientos)

El mejor agrupamiento...



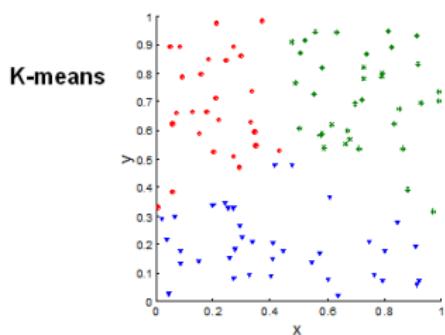
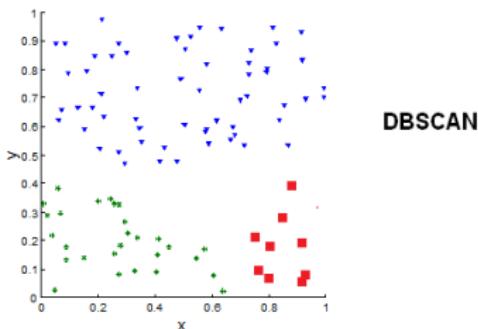
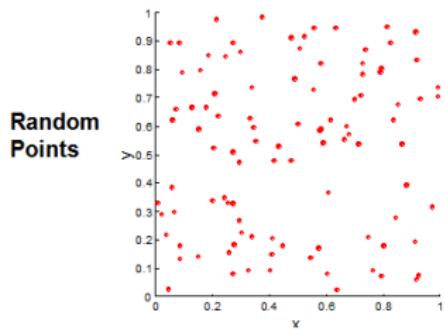
Validación de los grupos (o agrupamientos)

El mejor agrupamiento...



Validación de los grupos (o agrupamientos)

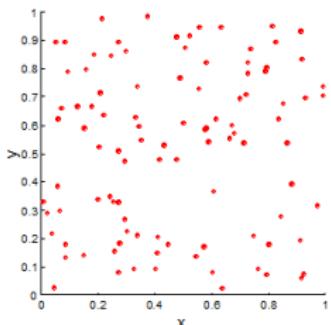
El mejor agrupamiento...



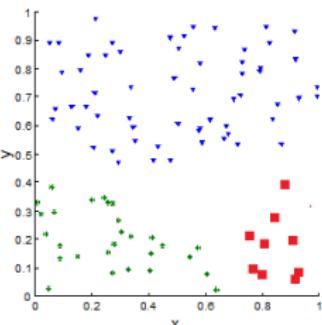
Validación de los grupos (o agrupamientos)

El mejor agrupamiento...

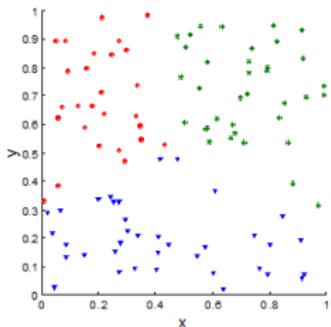
Random Points



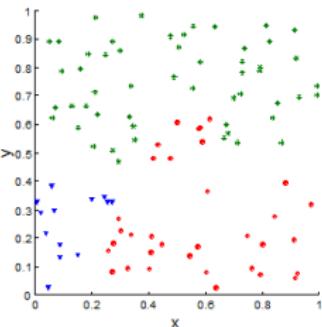
DBSCAN



K-means



Complete Link



Medidas de Validación de agrupamientos (MVA)

Las MVA's se dividen en 3 grandes grupos

Medidas de Validación de agrupamientos (MVA)

Las MVA's se dividen en 3 grandes grupos

Internas (o no supervisadas)

Miden las “bondades” de la estructura de un agrupamiento sin recurrir a ningún tipo de información externa. Estas medidas (o índices) suelen ser referenciados como **internas** dado que sólo usan información presente en el conjunto de datos.

Medidas de Validación de agrupamientos (MVA)

Las MVA's se dividen en 3 grandes grupos

Internas (o no supervisadas)

Miden las “bondades” de la estructura de un agrupamiento sin recurrir a ningún tipo de información externa. Estas medidas (o índices) suelen ser referenciados como **internas** dado que sólo usan información presente en el conjunto de datos.

Externas (o supervisadas)

Miden el grado de concordancia entre la estructura de los grupos descubiertos y alguna estructura externa al conjunto de datos (de ahí su nombre).

Medidas de Validación de agrupamientos (MVA)

Las MVA's se dividen en 3 grandes grupos

Internas (o no supervisadas)

Miden las “bondades” de la estructura de un agrupamiento sin recurrir a ningún tipo de información externa. Estas medidas (o índices) suelen ser referenciados como **internas** dado que sólo usan información presente en el conjunto de datos.

Externas (o supervisadas)

Miden el grado de concordancia entre la estructura de los grupos descubiertos y alguna estructura externa al conjunto de datos (de ahí su nombre).

Relativas

Compara agrupamientos o grupos particulares usando alguna de las dos medidas previas.

Medidas de Validez Internas (MVI)

Medidas de Validez Internas (MVI)

Las diferentes MVIs intentan identificar propiedades estructurales específicas de los agrupamientos como **cohesión, separación, densidad** o alguna combinación de estas propiedades.

- La familia de *índices de Dunn*
- el *índice de Davies-Bouldin*
- el *coeficiente de Silueta* (Silhouette Coefficient)
- la *Medida-Λ*
- la *Medida de Densidad Esperada* $\bar{\rho}$

MVIs, cohesión y separación

Las MVIs, suelen expresar la validez de un cluster global de K clusters como:

$$\text{validez}_{total} = \sum_{i=1}^K w_i \text{validez}(C_i)$$

y la función de *validez* suele ser alguna forma de **cohesión**, **separación** o una **combinación** de éstas.

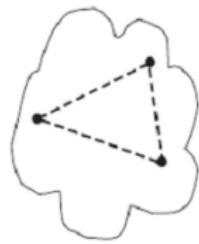
Cohesión

Mide cuán estrechamente relacionados están los objetos en un cluster.

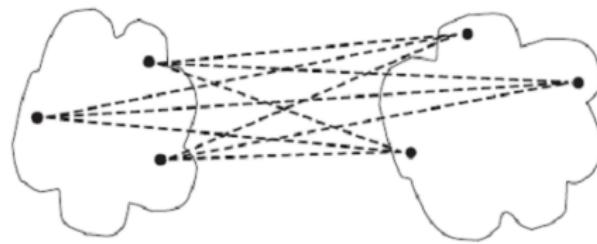
Separación

Mide cuán distintos (bien-separados) está un cluster de otro.

Cohesión y separación basada en grafos

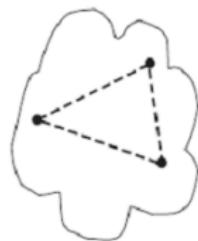


Cohesión

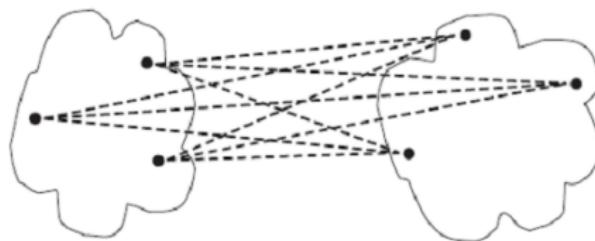


Separación

Cohesión y separación basada en grafos



Cohesión



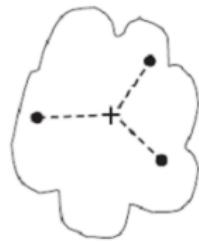
Separación

$$\text{cohes}(C_i) = \sum_{x \in C_i, y \in C_i} \text{proximidad}(x, y)$$

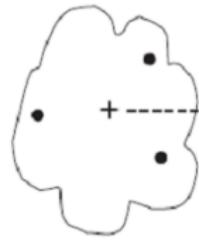
$$\text{separ}(C_i, C_j) = \sum_{x \in C_i, y \in C_j} \text{proximidad}(x, y)$$

la función de *proximidad* puede ser **similitud**, **(dis)similitud** (o distancia) o una función simple de estas cantidades.

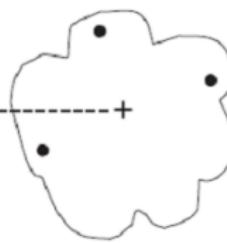
Cohesión y separación basada en prototipos



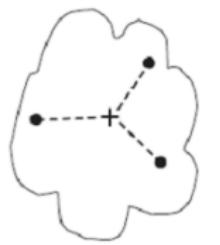
Cohesión



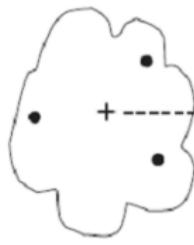
Separación



Cohesión y separación basada en prototipos



Cohesión



Separación

$$\text{cohes}(C_i) = \sum_{x \in C_i} \text{proximidad}(x, c_i)$$

$$\text{separ}(C_i, C_j) = \text{proximidad}(c_i, c_j)$$

$$\text{separ}(C_i) = \text{proximidad}(c_i, c)$$

La familia de Índices de Dunn

Identifica conjuntos de clusters **compactos y bien separados**.

- Sea O el conjunto de objetos y $\mathcal{C} = \{C_1, \dots, C_k\}$ un agrupamiento de O ,
- $\delta : \mathcal{C} \times \mathcal{C} \rightarrow \mathbb{R}$ una **distancia de cluster a cluster**.
- $\Delta : \mathcal{C} \rightarrow \mathbb{R}$ una medida de **diámetro de clusters**.

Todas la medidas de la forma:

$$I(C) = \frac{\min_{i \neq j} \delta(C_i, C_j)}{\max_{1 \leq i \leq k} \Delta(C_i)} \quad (5)$$

son llamadas **índices de Dunn**. Valores grandes de $I(C)$ corresponden a una buena partición en grupos.

► Ejemplos de δ y Δ

El Índice de Davies-Bouldin

Combina el grado de dispersión en los clusters y la separación entre clusters de un clustering C .

- Sea $\delta : \mathcal{C} \times \mathcal{C} \rightarrow \mathbb{R}$ una **distancia de cluster a cluster**,
- $s : \mathcal{C} \rightarrow \mathbb{R}$ una medida de la **dispersión** de los objetos dentro de un cluster.

El Índice de Davies-Bouldin se calcula como:

$$DB(C) = \frac{1}{k} \cdot \sum_{i=1}^k R_i(C), \text{ con} \quad (6)$$

$$R_i(C) = \max_{\substack{j=1, \dots, n \\ i \neq j}} R_{ij}(C) \text{ y } R_{ij}(C) = \frac{(s(C_i) + s(C_j))}{\delta(C_i, C_j)}$$

El Índice de Davies-Bouldin

Valores **pequeños** de *DB* corresponden a **buenos clusters**, dado que los clusters serán compactos y sus centros estarán distantes unos de otros.

Para nuestro análisis definimos

$$s(C_i) = \frac{1}{|C_i|} \sum_{x \in C_i} \|x - c_i\|$$

y

$$\delta = \|c_i - c_j\|$$

Una MVI informativa: el Coeficiente de Silueta

Componente fundamental de esta medida: fórmula para determinar el coeficiente de silueta de un objeto arbitrario i :

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

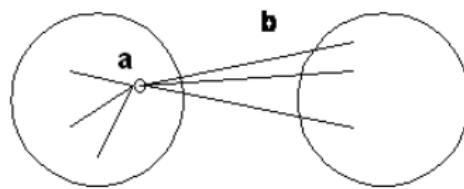
con $-1 \leq s(i) \leq 1$.

Una MVI informativa: el Coeficiente de Silueta

Componente fundamental de esta medida: fórmula para determinar el coeficiente de silueta de un objeto arbitrario i :

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

con $-1 \leq s(i) \leq 1$.



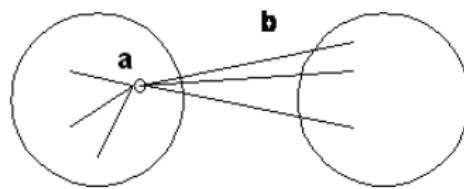
- $a(i)$ es la distancia promedio de i a los restantes objetos de su cluster.
- $b(i)$ es la distancia promedio de i a todos los objetos del cluster más cercano.

Una MVI informativa: el Coeficiente de Silueta

Componente fundamental de esta medida: fórmula para determinar el coeficiente de silueta de un objeto arbitrario i :

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

con $-1 \leq s(i) \leq 1$.



- $a(i)$ es la distancia promedio de i a los restantes objetos de su cluster.
- $b(i)$ es la distancia promedio de i a todos los objetos del cluster más cercano.
- Se busca que $s(i)$ sea tan cercano a 1 como sea posible

Una MVI informativa: el Coeficiente de Silueta

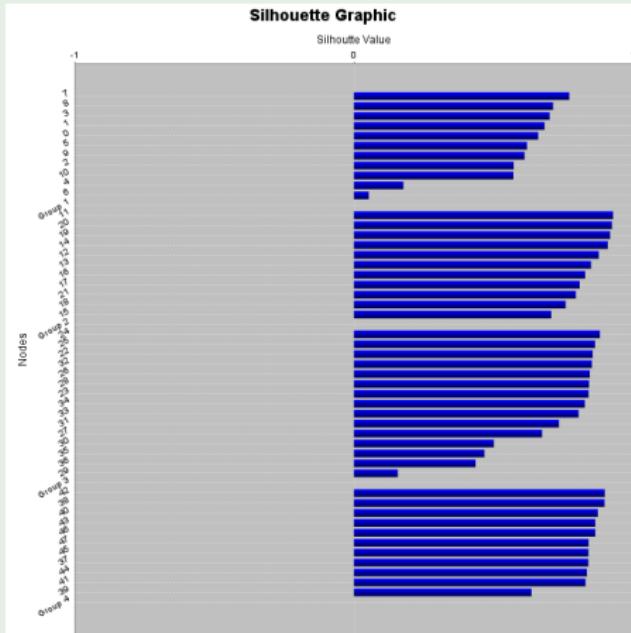
Combina ideas de **cohesión** y **separación**, pero para puntos individuales, grupos y agrupamientos...

Es posible calcular la silueta de:

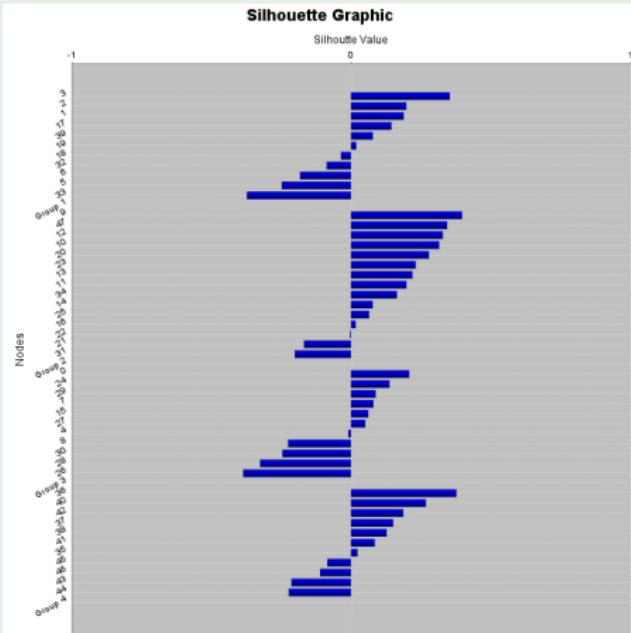
- un **grupo**: es el promedio de los coeficientes de silueta de sus objetos.
- un **agrupamiento**: es el promedio de los coeficientes de silueta de sus grupos.

Una MVI informativa: el Coeficiente de Silueta

Agrupamiento bueno



Agrupamiento malo



Medidas de Validez Externas (**MVE**)

Las MVEs evalúan un agrupamiento usando las medidas clásicas para evaluar un modelo de clasificación (categorización supervisada)

Medidas de Validez Externas (**MVE**)

Las MVEs evalúan un agrupamiento usando las medidas clásicas para evaluar un modelo de clasificación (categorización supervisada)

- Entropía
- Pureza
- Precisión y Recall
- Medida F (F -measure)

Medidas de Validez Externas (MVE)

Las MVEs evalúan un agrupamiento usando las medidas clásicas para evaluar un modelo de clasificación (categorización supervisada)

- Entropía
- Pureza
- Precisión y Recall
- Medida F (F -measure)

Notación

- Sea O el conjunto de objetos y $\mathcal{C} = \{C_1, \dots, C_k\}$ un agrupamiento de O .
- Sea $\mathcal{C}^* = \{C_1^*, \dots, C_l^*\}$ la categorización humana de O tomada como referencia.

Entropía

- Medida de Teoría de la Información. Cuantifica la incertidumbre de una fuente de información S que emite símbolos S_1, \dots, S_k con probabilidades $P(S_1), \dots, P(S_k)$.
- La entropía de la fuente de información (FI) S es:

$$H(S) = - \sum_{i=1}^k P(S_i) \log_2 P(S_i)$$

- Idea: cada cluster es una FI que “emite” números de clases de la categorización de referencia:

$$H(C_j) = - \sum_{|C_j \cap C_i^*| \neq 0} \frac{|C_j \cap C_i^*|}{|C_j|} \log_2 \left(\frac{|C_j \cap C_i^*|}{|C_j|} \right)$$

Entropía

- En pocas palabras, la entropía de un cluster mide en qué medida el cluster consiste de objetos de una **única clase** (clusters **puros**)
- La entropía del agrupamiento completo es

$$H(\mathcal{C}) = \sum_{C_j \in \mathcal{C}} \frac{|C_j|}{|D|} H(C_j)$$

- El clustering “perfecto” tiene una entropía de 0.
- Atención!!!:** el caso “patológico” (cada objeto en un cluster distinto) también tiene entropía 0.

Medidas de Validez Externas (MVE)

Pureza

- Otra medida del grado en que un cluster contiene objetos de una única clase.
- Pureza de un cluster C_j

$$Pur(C_j) = \max_{C_i^*} \frac{|C_j \cap C_i^*|}{|C_j|}$$

- Pureza de un agrupamiento completo

$$Pur(\mathcal{C}) = \sum_{C_j \in \mathcal{C}} \frac{|C_j|}{|D|} Pur(C_j)$$

Precisión y Recall

- La **precisión** de un cluster j con respecto a una clase i , $prec(i, j)$, indica qué proporción de los elementos del cluster pertenecen a la clase i . Se define como:

$$|C_j \cap C_i^*| / |C_j|$$

- El **recall** de un cluster j con respecto a una clase i , $rec(i, j)$, indica qué proporción de los elementos de la clase i pertenecen al cluster j . Se define como $|C_j \cap C_i^*| / |C_i^*|$.

Precisión

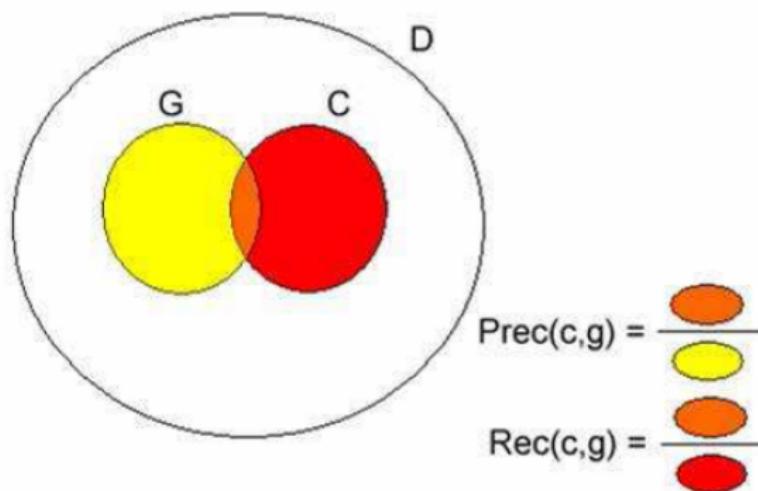
- La **precisión** de un cluster j con respecto a una clase i , $prec(i, j)$, indica qué **proporción** de los elementos del **cluster** pertenecen a la clase i . Se define como:

$$prec(i, j) = |C_j \cap C_i^*| / |C_j|$$

Precisión

- La **precisión** de un cluster j con respecto a una clase i , $prec(i, j)$, indica qué **proporción** de los elementos del **cluster** pertenecen a la clase i . Se define como:

$$prec(i, j) = |C_j \cap C_i^*| / |C_j|$$



Recall

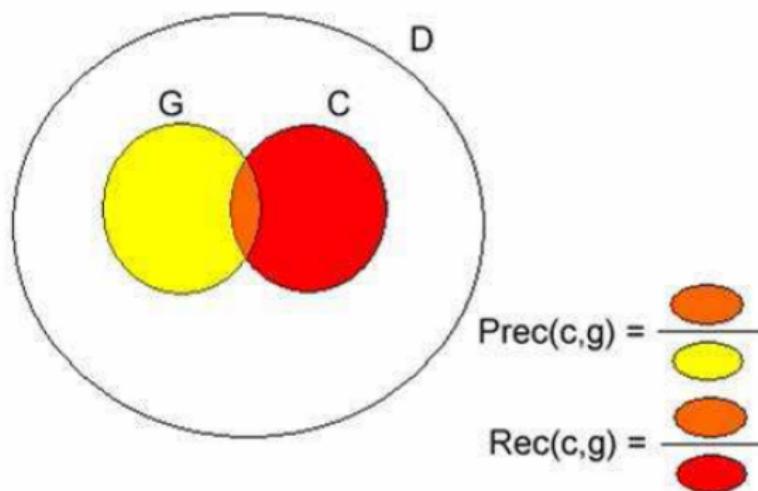
- El **recall** de un cluster j con respecto a una clase i , $rec(i, j)$, indica qué **proporción** de los elementos de la **clase i** pertenecen al cluster j . Se define como

$$rec(i, j) = |C_j \cap C_i^*| / |C_i^*|$$

Recall

- El **recall** de un cluster j con respecto a una clase i , $rec(i, j)$, indica qué **proporción** de los elementos de la **clase i** pertenecen al cluster j . Se define como

$$rec(i, j) = |C_j \cap C_i^*| / |C_i^*|$$



La medida *F* (*F-measure*)

- Combina **precisión** y **recall**.
- Indica en qué medida un cluster contiene **sólo** objetos de una clase y **todos** los elementos de esa clase.
- La medida *F* del cluster *j* con respecto a una clase *i* es:

$$F_{i,j} = \frac{2 \cdot \text{prec}(i,j) \cdot \text{rec}(i,j)}{\text{prec}(i,j) + \text{rec}(i,j)}$$

- La medida *F* total es:

$$F = \sum \frac{|C_i^*|}{|D|} \cdot \max_{j=1,\dots,k} \{F_{i,j}\}$$

Ejemplos de δ y Δ

Para nuestros análisis hemos usado:

$$\delta(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i, y \in C_j} d(x, y)$$

y

$$\Delta(C_i) = 2 \left(\frac{\sum_{x \in C_i} d(x, c_i)}{|C_i|} \right)$$

donde $d : D \times D \rightarrow R$ es una función que mide la distancia entre objetos y c_i denota el centroide de un cluster C_i .