Minería de Datos, Aprendizaje no supervisado: Agrupamiento

Maria Amparo Vila Miranda
vila@decsai.ugr.es

Departamento de Ciencias de la
Computación e Inteligencia Artificial
Universidad de Granada

Noviembre 2019

Esquema del tema

Introducción

Concepto Básicos

Definición

Agrupamiento *Clasificación no supervisada de patrones (observaciones, datos o vectores de características) en grupos (clusters).*

Este problema ha sido tratado en en muchos contextos y por investigadores de muchas disiciplinas (Biología, Psicología, Análisis Económico, Sociología etc..),

Introducción

Concepto Básicos

Definición

Agrupamiento *Clasificación no supervisada de patrones (observaciones, datos o vectores de características) en grupos (clusters).*

Este problema ha sido tratado en en muchos contextos y por investigadores de muchas disiciplinas (Biología, Psicología, Análisis Económico, Sociología etc..), Otra alternativa

Definición

Agrupamiento *Proceso de clasificar en grupos un conjunto de items sin tener una información previa acerca de su estructura.*

Introducción

Concepto Básicos

Los items están representados por un vector de datos, estas medidas también se suelen denominar *factores*, *componentes* o simplemente *variables.*

Introducción

Concepto Básicos

Los items están representados por un vector de datos, estas medidas también se suelen denominar *factores*, *componentes* o simplemente *variables.*

> *Intuitivamente dos items pertencientes a un agrupamiento válido deben ser más parecidos entre sí que aquellos que estén en grupos distintos y partiendo de esta idea se desarrollan las técnicas de agrupamiento.*

Introducción

Concepto Básicos

Históricamente:

- Final de los 60:dentro del ámbito del Análisis de Datos y de la Taxonomía Numérica
- Años 70 y 80 el agrupamiento se incluye en la Inteligencia Artificial dentro del Aprendizaje no Supervisado
- A partir de los 90: la Minería de Datos recoje el Agrupamiento. Recupera las técnicas y metodologías adaptándolas grandes masas de datos

Introducción

Concepto Básicos

Históricamente:

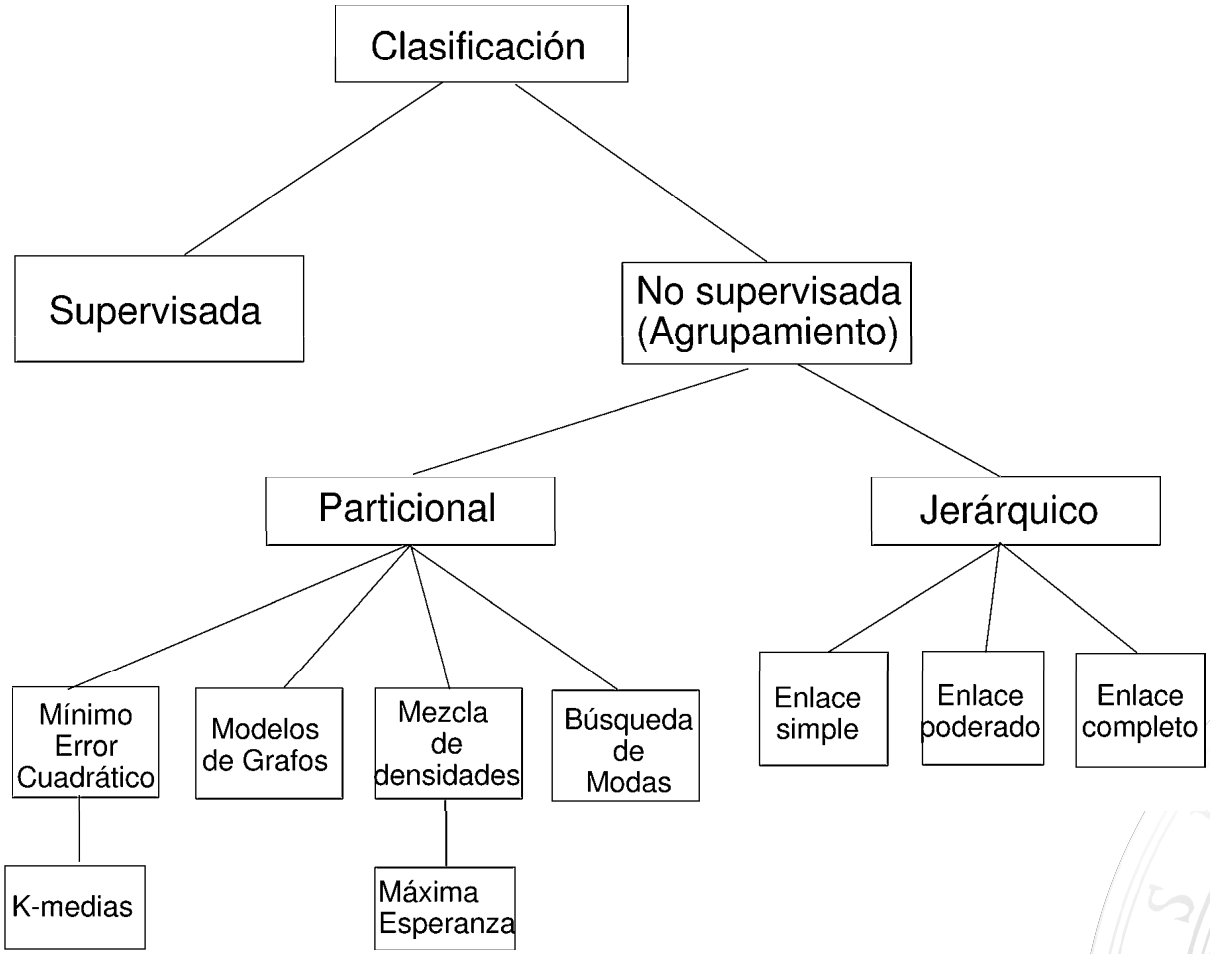
- Final de los 60:dentro del ámbito del Análisis de Datos y de la Taxonomía Numérica
- Años 70 y 80 el agrupamiento se incluye en la Inteligencia Artificial dentro del Aprendizaje no Supervisado
- A partir de los 90: la Minería de Datos recoje el Agrupamiento. Recupera las técnicas y metodologías adaptándolas grandes masas de datos

Las técnicas de agrupamiento dependen de:

- Qué clase de problemas se estén resolviendo.
- Cómo sean los datos de partida
- Qué medidas de parecido (semejanza) se estén utilizando

Introducción

Clasificación de las técnicas de Agrupamiento



Introducción

Clasificación de las técnicas de Agrupamiento

Clasificación supervisada y no supervisada .

-En el caso de que sea supervisada sabemos a que grupo pertence cada patron, entonces lo que se desea es encontrar un conjunto de "criterios", probablemente reglas, que nos permitan, dado un nuevo item, situarlo en un grupo.

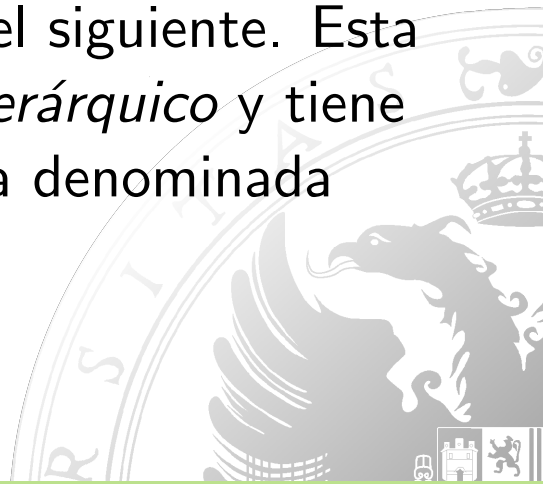
- En el caso de la clasificación no supervisada no se tiene tanta información acerca de los grupos, a veces no se sabe siquiera cuantos grupos hay, los que se trata entonces es de encontrar un agrupamiento que reuna en el mismo grupo los items mas "parecidos".

Introducción

Clasificación de las técnicas de Agrupamiento

Agrupamiento particional y agrupamiento jerárquico .

- Cuando los grupos a obtener son disjuntos y cubren todo el conjunto de items se dice que el agrupamiento es *particional*.
- Cuando se obtiene una jerarquía de agrupamientos particionales "anidados ", de tal manera que cada grupo de un nivel se divide en varios en el nivel siguiente. Esta estructura se denomina *agrupamiento jerárquico* y tiene una representación gráfica muy intuitiva denominada *dendrograma*.



Introducción

Clasificación de las técnicas de Agrupamiento

Modelos de Grafos :

Si se considera que los datos están representados mediante un grafo donde los vértices son los items o patrones y las aristas son conexiones entre ellos.

Modelos relacionales .

Cuando se considera que los grupos deben ser "cohesionados" de manera que los items de un mismo grupo estén más cercanos a entre sí y la distancia entre grupos sea la mayor posible, Uno de los más extendido es el de *mínimos cuadrados*, donde el criterio de cohesión se obtiene como la suma total de la distancia de cada item al punto medio (*centroide*) El *método de las k-medias* se basa en este enfoque.

Modelos de Análisis de densidad .

Cuando se considera que un grupo es una región del espacio donde la densidad de items es muy alta y que está rodeada de una región de baja densidad.

Introducción

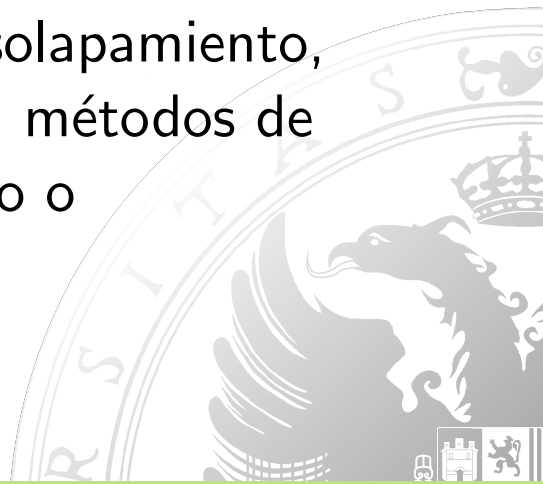
Clasificación de las técnicas de Agrupamiento

Modelos de Análisis de densidad .

Cuando se considera que un grupo es una región del espacio donde la densidad de items es muy alta y que está rodeada de una región de baja densidad.

Agrupamientos exclusivos, agrupamientos no exclusivos .

Todos los enfoques comentados en los párrafos anteriores parten de la hipótesis de no-solapamiento, cuando se relaja esta hipótesis aparecen métodos de agrupamiento que admiten solapamiento o no-exclusivos.



Introducción

Clasificación de las técnicas de Agrupamiento

Agrupamientos difusos .

Los métodos de agrupamiento no-exclusivos que han tenido más éxito son los que suponen que los grupos son *conjuntos difusos* de forma que un item puede pertenecer a diversos grupos con un nivel de pertencencia a cada uno.

Introducción

Clasificación de las técnicas de Agrupamiento

Agrupamientos difusos .

Los métodos de agrupamiento no-exclusivos que han tenido más éxito son los que suponen que los grupos son *conjuntos difusos* de forma que un item puede pertenecer a diversos grupos con un nivel de pertencencia a cada uno.

Agrupamientos aglomerativos y divisivos .

Si se parte de un agrupamiento en el que cada item es un grupo y se van construyendo nuevas soluciones uniendo grupos en otros más amplios, se tiene un algoritmo de tipo *aglomerativo*, si el proceso es el contrario, tendremos un algoritmo *divisivo*.

Se trata de una filosofía de actuación que se aplica fundamentalmente en agrupamientos jerárquicos

Los datos de partida

La información de partida puede estar representada de dos formas:

Los datos de partida

La información de partida puede estar representada de dos formas:

- Por medio de un **Dataset**

items\variables	V_1	V_2	$\cdots\cdots$	V_N
o_1	d_{11}	d_{12}	$\cdots\cdots$	d_{1N}
$\vdots$	$\vdots$	$\vdots$	$\cdots\cdots$	$\vdots$
o_M	d_{M1}	d_{M2}	$\cdots\cdots$	d_{MN}

Los datos de partida

La información de partida puede estar representada de dos formas:

- Por medio de un **Dataset**

items\variables	V_1	V_2	$\cdots\cdots$	V_N
o_1	d_{11}	d_{12}	$\cdots\cdots$	d_{1N}
$\vdots$	$\vdots$	$\vdots$	$\cdots\cdots$	$\vdots$
o_M	d_{M1}	d_{M2}	$\cdots\cdots$	d_{MN}

- Por medio de una **Matriz de proximidad**, $n \times n$ donde el valor de la casilla ik representa una medida de la similaridad o de la distancia entre el item i y el k.

Los datos de partida

La información de partida puede estar representada de dos formas:

- Por medio de un **Dataset**

items\variables	V_1	V_2	$\cdots\cdots$	V_N
o_1	d_{11}	d_{12}	$\cdots\cdots$	d_{1N}
$\vdots$	$\vdots$	$\vdots$	$\cdots\cdots$	$\vdots$
o_M	d_{M1}	d_{M2}	$\cdots\cdots$	d_{MN}

- Por medio de una **Matriz de proximidad**, $n \times n$ donde el valor de la casilla ik representa una medida de la similaridad o de la distancia entre el item i y el k.

  Habitualmente la matriz de proximidad se calcula a partir del data set; pero en ciertas aplicaciones psicométricas y sociológicas es posible que los datos se recojan directamente en términos de concordancias

Tipos de datos

Existen distintos tipos de variables para representar un item:

Cuantitativas se dividen en:

Con valores continuos por ejemplo el peso de una persona, o el nivel de sodio en un suelo.

Con valores discretos por ejemplo el número de ordenadores de un centro.

Variables cualitativos se dividen en:

Con valores nominales o no ordenados por ejemplo el color de un suelo, o el diagnóstico de un enfermo.

Con valores ordinales por ejemplo el rango de un militar o el nivel de gravedad de un enfermo. Un caso importante particular son las*binarias* que representan la presencia o ausencia de una determinada característica.

Indices de proximidad: distancias y semejanzas

Proximidad

⋆*Problema*

Construir una matriz de proximidad a partir de un dataset

Definición

Indice de proximidad *Consideremos un conjunto de n patrones que notaremos por* $i, l, k.. \in I = \{1, 2, ..., n\}$ *,decimos que* $d : I \times I \longrightarrow R$ *es un índice de proximidad si y sólo si verifica:*

1.1.1 *Para medidas de disimilaridad o distancia:* $\forall i \in I \; d(i, i) = 0$

1.2 *Para medidas de similaridad:* $\forall i \in I \; d(i, i) \geq max_{k \in I} d(i, k)$

2. $d(i, k) = d(k, i) \; \forall i, k \in I$

3. $d(i, k) \geq 0 \forall i, k \in I$

Indices de proximidad: distancias y semejanzas

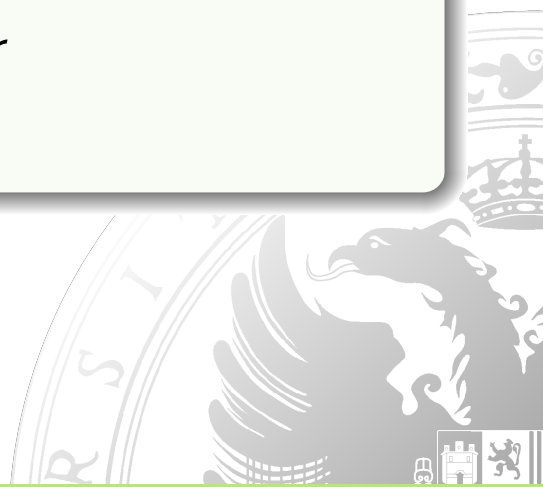
Funciones de distancia

Los índices de proximidad son una generalización de otros conceptos más conocidos

Definición

Decimos que el índice de proximidad d es una **Función de Distancia** *si y sólo si verifica:*

1. *Las propiedades 1.a, 2 y 3 de la definición anterior*
2. $d(i,k) \leq d(i,l) + d(l,k)\ \forall i,l,k \in I$



Indices de proximidad: distancias y semejanzas

Funciones de distancia

Distintas funciones de distancia

NOMBRE	EXPRESION
Euclídea o norma-l_2	$d_2(i,k) = [\sum_{j=1}^{m}(x_{ij} - x_{kj})^2]^{1/2} = [(\mathbf{x_i} - \mathbf{x_k})^{\mathbf{T}}(\mathbf{x_i} - \mathbf{x_k})]$
Manhattan o norma-l_1	$d_1(i,k) = \sum_{j=1}^{m} \|(x_{ij} - x_{kj}\|$
norma del supremo	$d_\infty(i,k) = sup_{j\in\{1,2..m\}}\|(x_{ij} - x_{kj}\|$
Minkowski o norma-l_p	$d_p(i,k) = \sum_{j=1}^{m}[\|x_{ij} - x_{kj}\|^p]^{1/p}$
Distancia de Mahalanobis	$d_M = [(\mathbf{x_i} - \mathbf{x_k})^{\mathbf{T}}\mathbf{\Sigma^{-1}}(\mathbf{x_i} - \mathbf{x_k})]$ Σ es la covarianza muestral o una matriz de covarianza intra-grup

Indices de proximidad: distancias y semejanzas

Funciones de distancia

Indices de proximidad: distancias y semejanzas

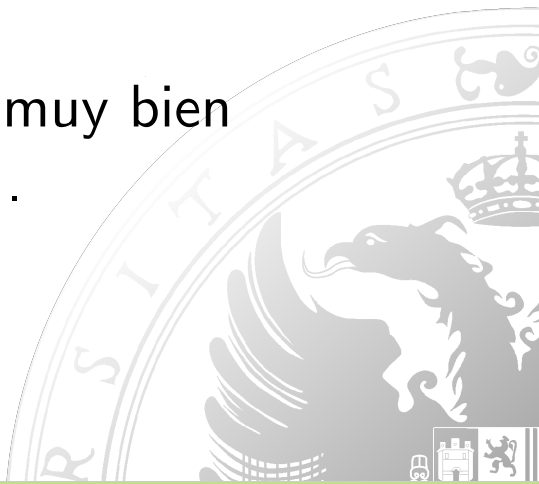
Funciones de distancia

- La distancia d_p generaliza a las otras,la distancia de Mahalanobis también generaliza la distancia euclídea.

Indices de proximidad: distancias y semejanzas

Funciones de distancia

- La distancia d_p generaliza a las otras,la distancia de Mahalanobis también generaliza la distancia euclídea.
- Existen también funciones de distancia basadas en la distancia de dos distribuciones de probabilidad y funciones de distancia basadas en el coeficiente de correlación que se aplican al espacio de variables, no al de items.
- La distancia Euclídea es la más intuitiva y trabaja muy bien cuando se tienen grupos "compactos" y "aislados".



Indices de proximidad: distancias y semejanzas

Funciones de distancia

Indices de proximidad: distancias y semejanzas

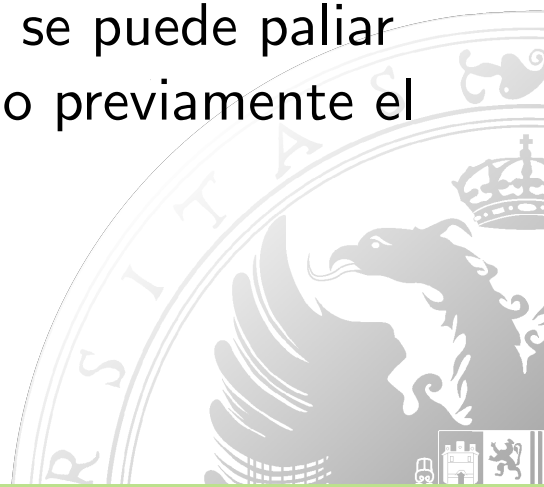
Funciones de distancia

- El principal inconveniente de todas las métricas de Minkowski , en general, es que su uso da un gran peso a las variables con valores muy grandes, este problema se soluciona normalizando previamente las variables.

Indices de proximidad: distancias y semejanzas

Funciones de distancia

- El principal inconveniente de todas las métricas de Minkowski , en general, es que su uso da un gran peso a las variables con valores muy grandes, este problema se soluciona normalizando previamente las variables.
- Otro problema que presentan las variables continuas es la posible existencia de correlación entre ellos, este problema se puede paliar utilizando la distancia de Mahalanobis o reduciendo previamente el espacio.



Indices de proximidad: distancias y semejanzas

Indices de semejanza

Definición

Dado un índice de proximidad s decimos que es una **función de semejanza** *si y sólo si verifica:*

1. $\forall i \in I \; d(i,i) = 1$
2. *Las propiedades 2 y 3 de la definición de proximidad*

Indices de proximidad: distancias y semejanzas

Indices de semejanza

Definición

Dado un índice de proximidad s decimos que es una **función de semejanza** *si y sólo si verifica:*

1. $\forall i \in I \; d(i,i) = 1$
2. *Las propiedades 2 y 3 de la definición de proximidad*

Se puede obtener un índice de semejanza a partir de una distancia:

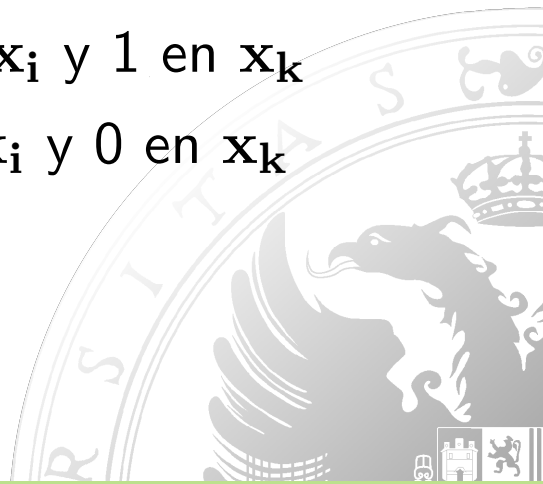
$$\forall i,k \in I \; s(i,k) = 1 - (d(i,k)/D) \text{ siendo } D = max_{i,k} d(i,k)$$

Indices de proximidad: distancias y semejanzas

Indices de semejanza

La mayoría de los índices de semejanza, no basados en distancia, se han definido para items cuyas variables son binarios. Si consideramos los items $\mathbf{x_i}$ y $\mathbf{x_k}$, formados por m variables binarias,

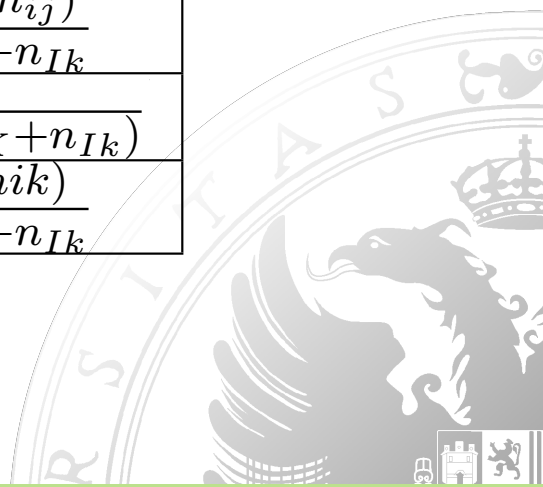
- n_{IK} número de variables que toman el valor 1 en $\mathbf{x_i}$ y $\mathbf{x_k}$
- n_{iK} número de variables que toman el valor 0 en $\mathbf{x_i}$ y $\mathbf{x_k}$
- n_{iK} número de variables que toman el valor 0 en $\mathbf{x_i}$ y 1 en $\mathbf{x_k}$
- n_{ik} número de variables que toman el valor 1 en $\mathbf{x_i}$ y 0 en $\mathbf{x_k}$



Indices de proximidad: distancias y semejanzas

Indices de semejanza

NOMBRE	EXPRESION
Indice de Jaccard	$\frac{n_{IK}}{n_{IK}+n_{iK}+n_{Ik}}$
Indice de acoplamiento simple	$\frac{n_{IK}+n_{ik}}{m}$
Indice de Russell	$\frac{n_{IK}}{m}$
Indice de Dice	$\frac{2n_{IK}}{2n_{IK}+n_{iK}+n_{Ik}}$
	$\frac{2(n_{IK}+n_{ij})}{m+n_{iK}+n_{Ik}}$
	$\frac{n_{IK}}{n_{IK}+2(n_{iK}+n_{Ik})}$
	$\frac{(n_{IK}+nik)}{m+n_{iK}+n_{Ik}}$



Indices de proximidad: distancias y semejanzas

Indices de semejanza

La medida del coseno

Se basa en la representación de cada documento como un vector de frecuencias de aparición de términos y calcula en coseno del ángulo que forman ambos vectores. De forma que si $t_1 = (t_{11}...t_{1d})$ y $t_2 = (t_{21}...t_{2d})$ son dos vectores de documentos en un espacio d-dimensional, entonces:

$$cos(t_1, t_2) = (t_1 \odot t_2)/|t_1||t2|$$

donde $\odot$ representa el producto escalar y $|.|$ el módulo, es decir:

$$cos(t_1, t_2) = \frac{\sum_{j=1}^{d} t_{1j} t_{2j}}{\sqrt{\sum_{j=1}^{d} t_{1j}^2} \sqrt{\sum_{j=1}^{d} t_{2j}^2}} \quad (1)$$

Indices de proximidad: distancias y semejanzas

Consideraciones importantes

Tanto las distancias como las semejanzas se utilizan para obtener la matriz de proximidad de un conjunto de items que es el punto de partida para un proceso de agrupamiento.

Indices de proximidad: distancias y semejanzas

Consideraciones importantes

Tanto las distancias como las semejanzas se utilizan para obtener la matriz de proximidad de un conjunto de items que es el punto de partida para un proceso de agrupamiento.

Cada una de los enfoques corresponde a un tipo de variables

- Las distancias se utilizarán en presencia de variables continuas, y pueden usarse con valores enteros e incluso ordinales asimilables a enteros, con cuidado.
- Los índices de semejanza son adecuadas cuando se trabaje con factores binarios y pueden utilizarse con variables nominales no ordinales tranformándoles en un conjunto de factores binarios.
- Cuando se trata de valores nominales se pueden utilizar relaciones de semejanza previas difusas

Indices de proximidad: distancias y semejanzas

Consideraciones importantes

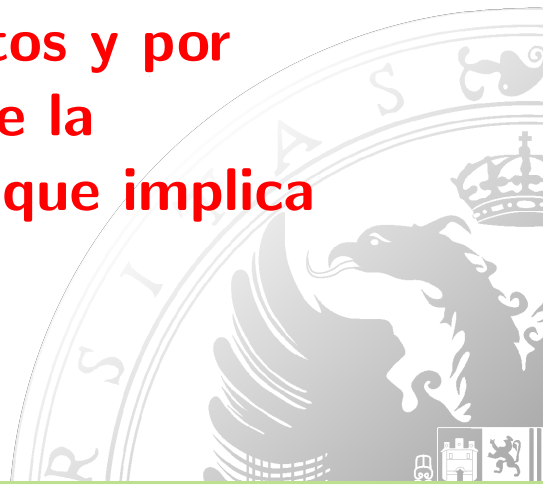
Es importante tener en cuenta que puede ser problemático mezclar enfoques directamente, cuando se tienen varios tipos de variables. Habrá que establecer combinaciones de distancias y/o semejanzas convenientemente normalizadas

Indices de proximidad: distancias y semejanzas

Consideraciones importantes

Es importante tener en cuenta que puede ser problemático mezclar enfoques directamente, cuando se tienen varios tipos de variables. Habrá que establecer combinaciones de distancias y/o semejanzas convenientemente normalizadas

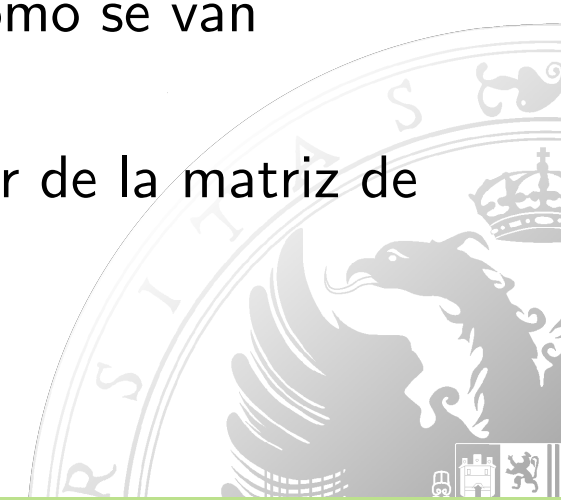
Preparación de datos y selección de distancia son cruciales en los procesos de agrupamiento. Es habitual que los resultados dependan mucho de estos dos puntos y por tanto de cada tipo de problema. Normalmente la selección de ambas cosas es un proceso largo que implica varios ensayos



Técnicas de agrupamiento jerárquicas

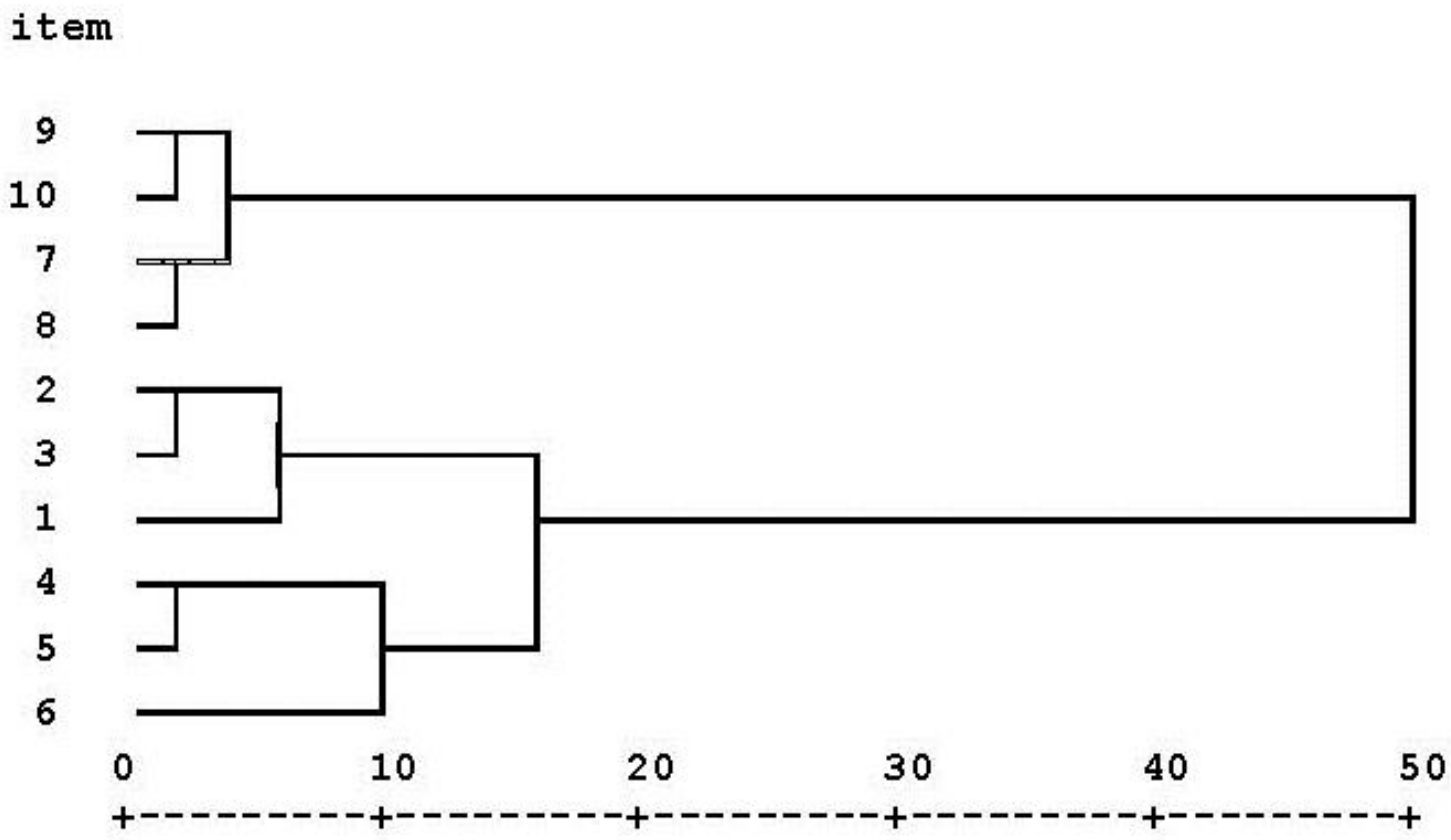
Ideas básicas

- Un agrupamiento jerárquico es una sucesión de particiones "anidadas"
- Cada grupo de items perteneciente a una determinada partición está totalmente incluido en algún grupo de la partición siguiente
- Esta estructura tiene una representación gráfica muy intuitiva que se denomina *Dendrograma*. Donde se presenta cómo se van uniendo los distintos patrones en grupos
- Obviamente el criterio de unión se obtiene a partir de la matriz de distancia, mediante procesos algoritmicos



Técnicas de agrupamiento jerárquicas

Ejemplo de dendrograma



Técnicas de agrupamiento jerárquicas

Algoritmos

La mayoría de los algoritmos són de tipo aglomerativo, partiendo de una partición en la que cada item es un grupo se van obteniendo nuevas particiones uniendo grupos entre sí.

Enfoque basado en grafos

Se considera que cada item es un vértice de un grafo y se van generando particiones, conectando los vertices de menor distancia, aparecen dos formas:

Agrupamiento de enlace simple *(Single-link clustering)* Los grupos se obtienen buscando las componentes conexas de grafo y se termina cuando todos los vértices están conectados.

Agrupamiento de enlace completo *(Complete-link clustering)* Los grupos se obtienen buscando los subgrafos completamente conectados (cliques)

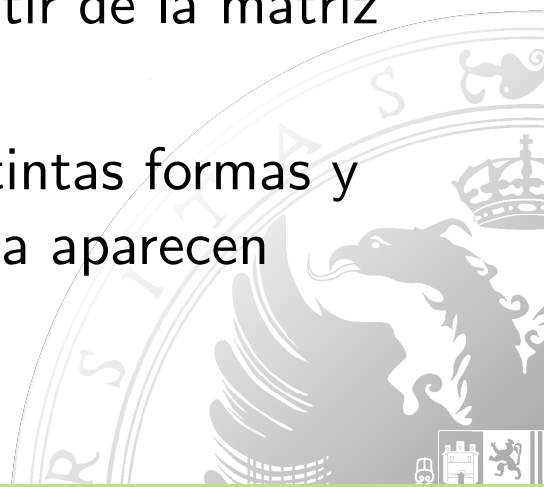
Técnicas de agrupamiento jerárquicas

Algoritmos

⋆ *Algoritmo de Jhonson*

Ideas básicas:

- Realiza sucesivas transformaciones de la matriz distancia, reduciendo la dimension de la misma siempre que se forme un nuevo grupo.
- La idea es que se trabaje con una *matriz de distancia entre grupos*, y que esta se vaya calculando iterativamente a partir de la matriz de la etapa anterior.
- La distancia entre grupos se puede calcular de distintas formas y que dependiento de cómo se calcule dicha distancia aparecen distintas formas de agrupamiento.

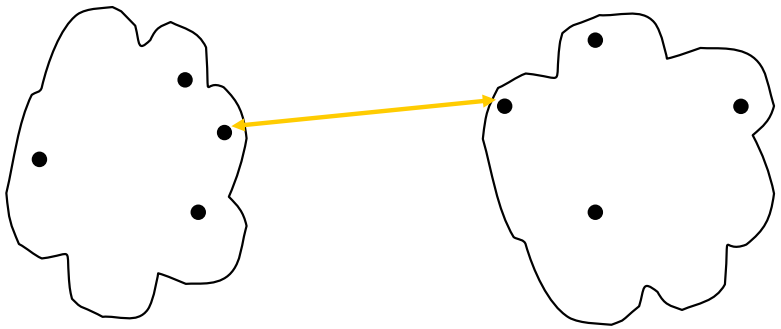


Técnicas de agrupamiento jerárquicas

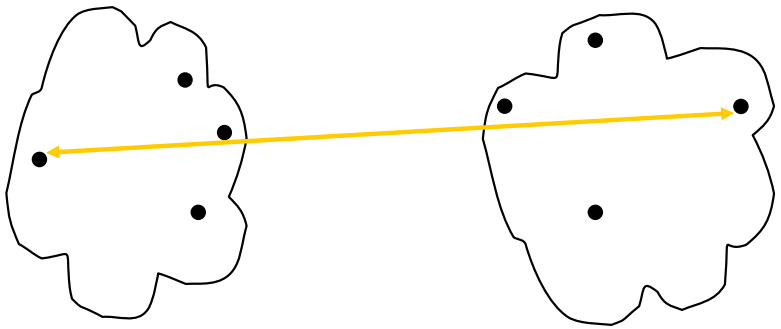
Algoritmos

FORMAS DE CALCULAR LA PROXIMIDAD ENTRE GRUPOS

MINIMO



MAXIMO

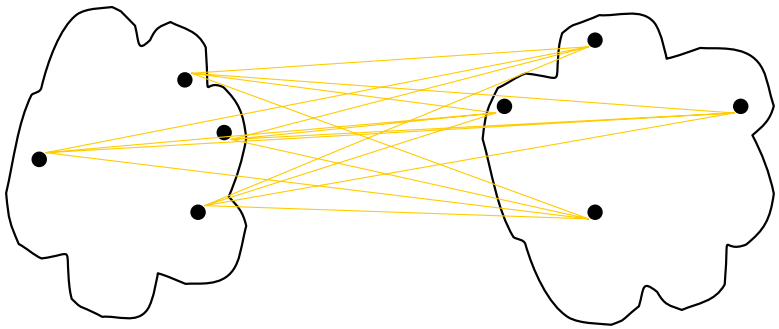


Técnicas de agrupamiento jerárquicas

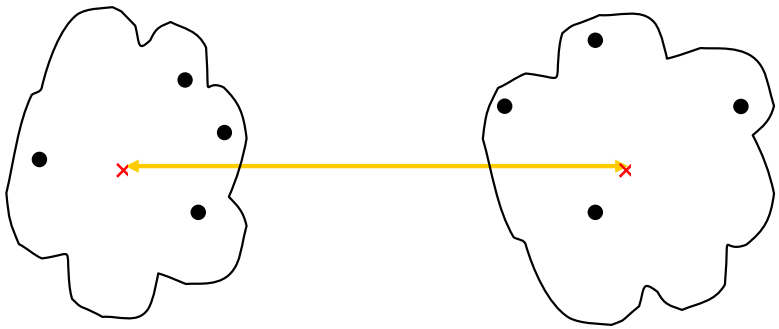
Algoritmos

FORMAS DE CALCULAR LA PROXIMIDAD ENTRE GRUPOS

MEDIA DE GRUPOS



DISTANCIA ENTRE CENTROIDES



Técnicas de agrupamiento jerárquicas

Algoritmos

⋆ *Algoritmo de Jhonson. Forma general de Lance y William*

1. Sean m=0, $D_m = D$, matriz de distancia de partida, $\mathcal{C}_m = \{\{1\}, .., \{n\}\}$ el agrupamiento inicial, $L(m) = 0$ el nivel al cual se hace este agrupamiento.
2. Sean R y S aquellos grupos de $\mathcal{C}_m$ que tienen distancia mínima:
   - $L(m+1) = D_m(R, S)$
   - Formar un nuevo grupo $K = R \cup S$. Hacer $\mathcal{C}_{m+1} = \mathcal{C}_m \cup (R \cup S) - R - S$ y transformar la matriz D_m de la siguiente manera.
     - Eliminar la fila y columna de S y asignar la fila y columna de R a K.
     - Para todo T perteciente a $\mathcal{C}_m$ distinto de K, hacer:

$$D_{m+1}(K, T) = (2$$

$$a(R)D_m(R, T) + a(S)D_m(S, T) + bD_m(R, S) + c|D_m(R, T) - D_m(S, T) \quad (3$$

3. Hacer $m = m + 1$
4. Si se han unido todos los items parar, en caso contrario ir a 2

Técnicas de agrupamiento jerárquicas:

Algoritmos

⋆ *Algoritmo de Jhonson. Forma general de Lance y William*

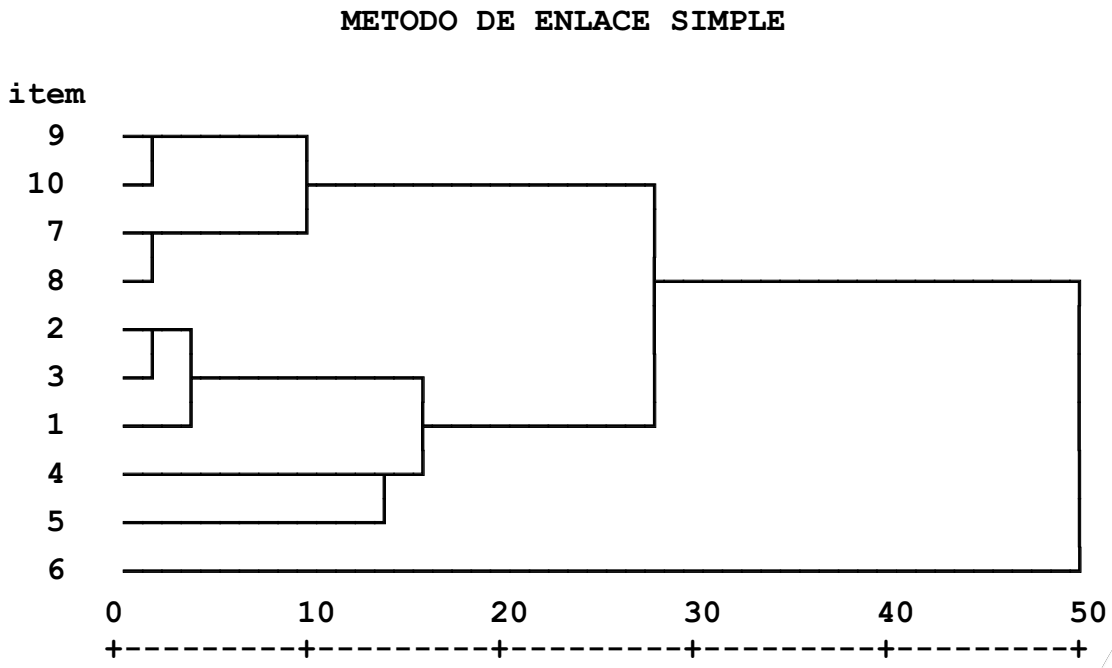
Sea n_Y el numero de elementos que tiene el grupo Y,la tabla nos muestra los coeficientes de la expresión anterior para los algoritmos de aglomerativos de agrupamiento jerárquico más conocidos

METODO	a(R)	a(S)
Enlace Simple	$1/2$	$1/2$
Enlace completo	$1/2$	$1/2$
Media de grupos	n_R/n_K	n_S/n_K
Centroide	$n_R/(n_R+n_T)$	$n_S/(n_s+n_T)$
Método de Ward	$(n_s+n_T)/(n_K+n_T)$	$(n_R+n_T)/(n_K+n_T)$

METODO	b	c
Enlace Simple	0	$-1/2$
Enlace completo	0	$1/2$
Media de grupos	0	0
Centroide	$-(n_R n_S)/n_K^2$	0

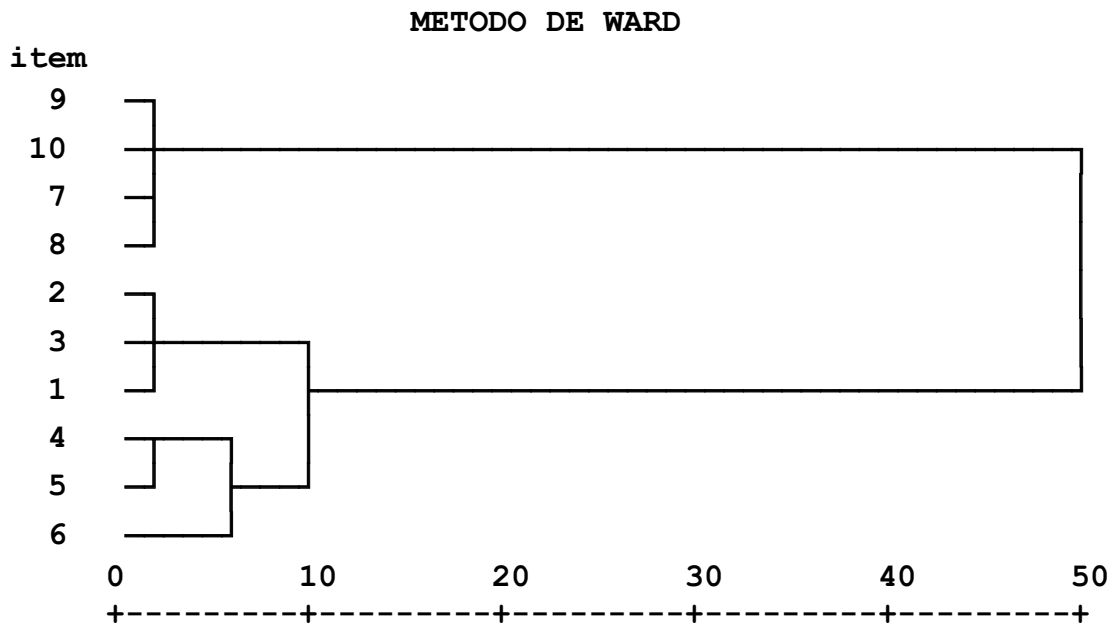
Técnicas de agrupamiento jerárquicas

Algoritmos



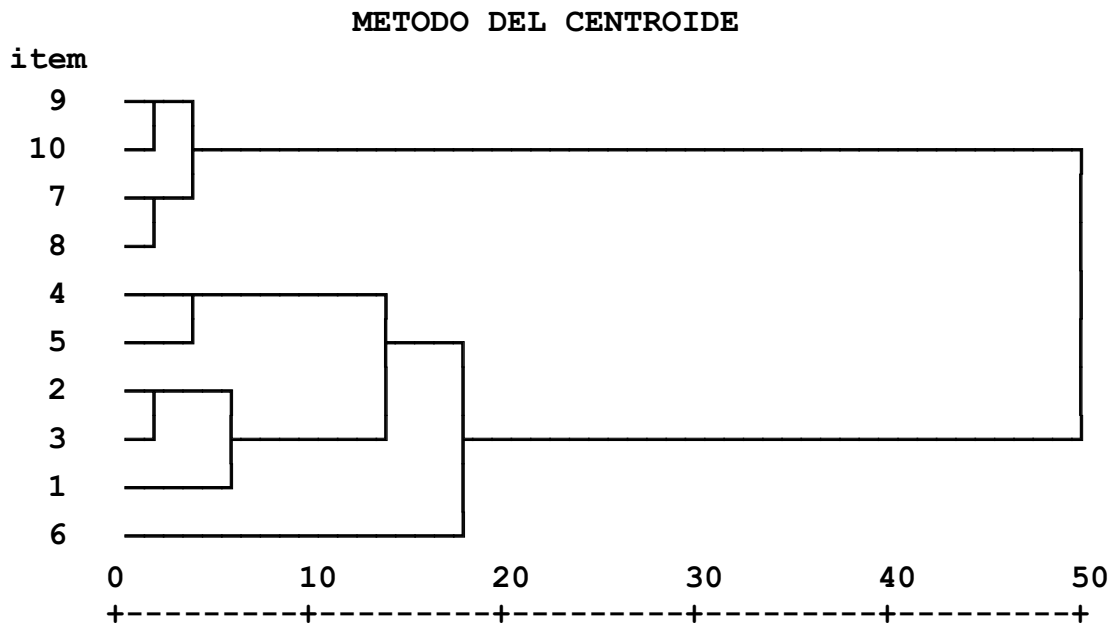
Técnicas de agrupamiento jerárquicas

Algoritmos



Técnicas de agrupamiento jerárquicas

Algoritmos



Técnicas de agrupamiento particional

Ideas iniciales

Definición

Dados n items representados en un espacio d-dimensional donde esté definida una distancia, determinar una partición de los mismos en K subconjuntos o grupos tales que los items situados en un grupo se parezcan más entre sí que al resto de los situados en grupos diferentes.

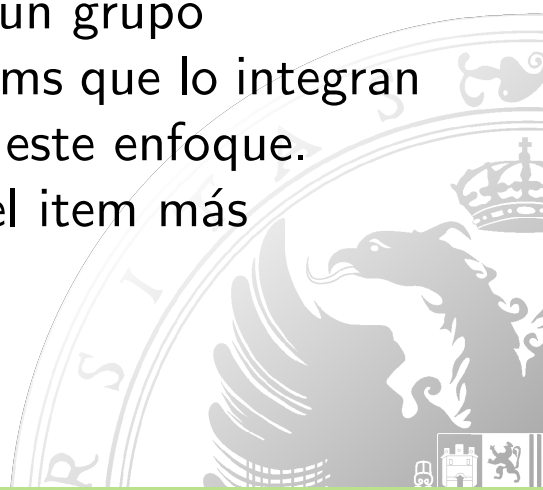
El **número** K *de grupos a generar, puede estar definido o no.La* **similaridad** *o coherencia de un grupo y de un conjunto de grupos se mide según distintos criterios*

Técnicas de agrupamiento particional

Ideas iniciales

⋆ *Criterios globales*

- Suponen que cada grupo está representado por un prototipo y asignan cada item al grupo cuyo prototipo esté más cercano.
- Se usan en este enfoque medidas de coherencia basadas en la distancia de cada item a su prototipo y dependiendo de la distancia que se considere aparecerán distintas medidas.
  - Para datos con atributos continuos, el prototipo de un grupo habitualmente tiene como valores la media de los items que lo integran (*centroide*). El *método de las k-medias* pertenece a este enfoque.
  - En el caso de variables categóricas se suele utilizar el item más representativo del grupo (*medoide*)

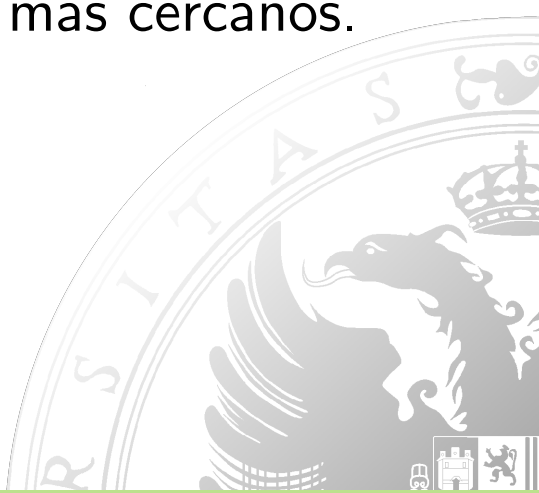


Técnicas de agrupamiento particional

Ideas iniciales

⋆ *Criterios locales*

- Forman los grupos utilizando la estructura local de los datos.
- Ejemplos de este enfoque son los métodos basados en la identificación de regiones de alta densidad de puntos o aquellos que asignan al mismo grupo un patrón y sus k-vecinos más cercanos.
- Uno de los métodos más conocidos es *DBSCAN*



Técnicas de agrupamiento particional

El método de las k-medias

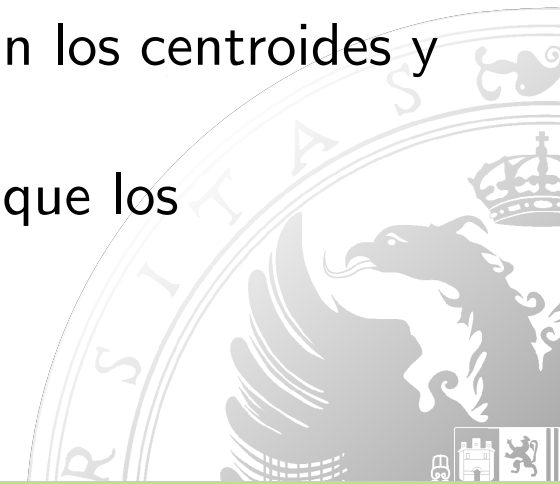
⋆ *Ideas básicas*

Parámetros iniciales :

El número de grupos K y K centroides iniciales

Proceso básico .

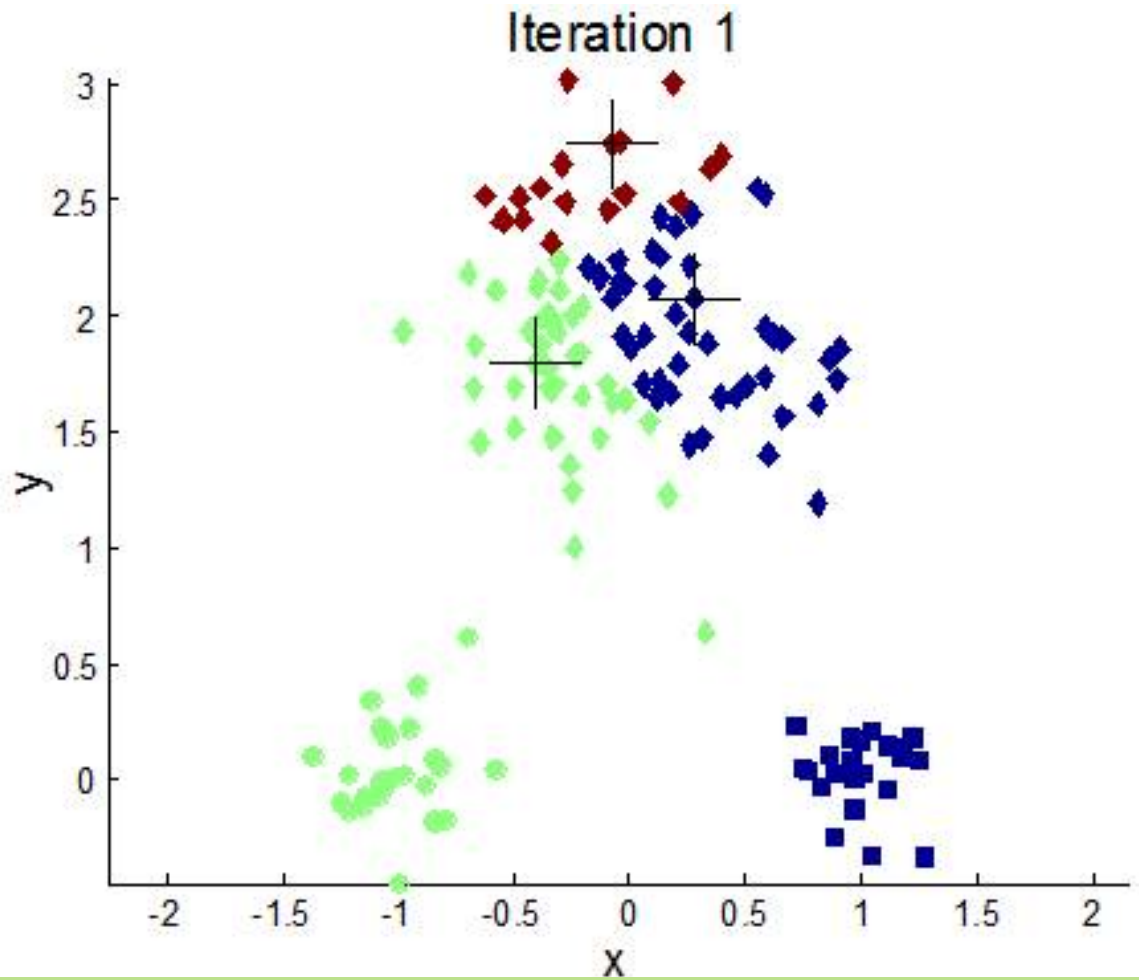
1. Se asigna entonces cada punto a su centroide más cercano y así se obtienen los grupos iniciales.
2. A partir de estos grupos se recalculan los centroides y se hace una nueva reasignación.
3. El proceso se vuelve a repetir hasta que los centroides no cambian.



Técnicas de agrupamiento particional

El método de las k-medias

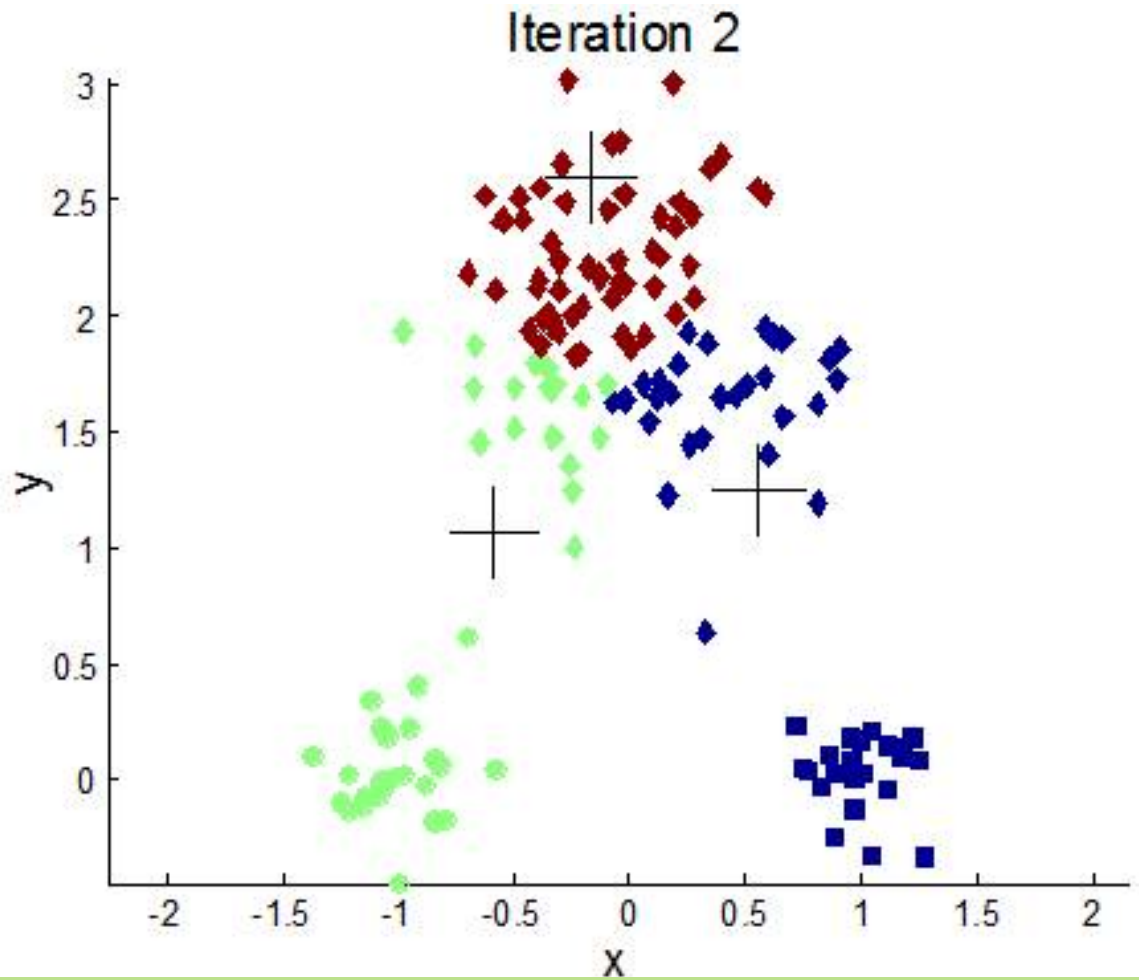
⋆ *Funcionamiento*



Técnicas de agrupamiento particional

El método de las k-medias

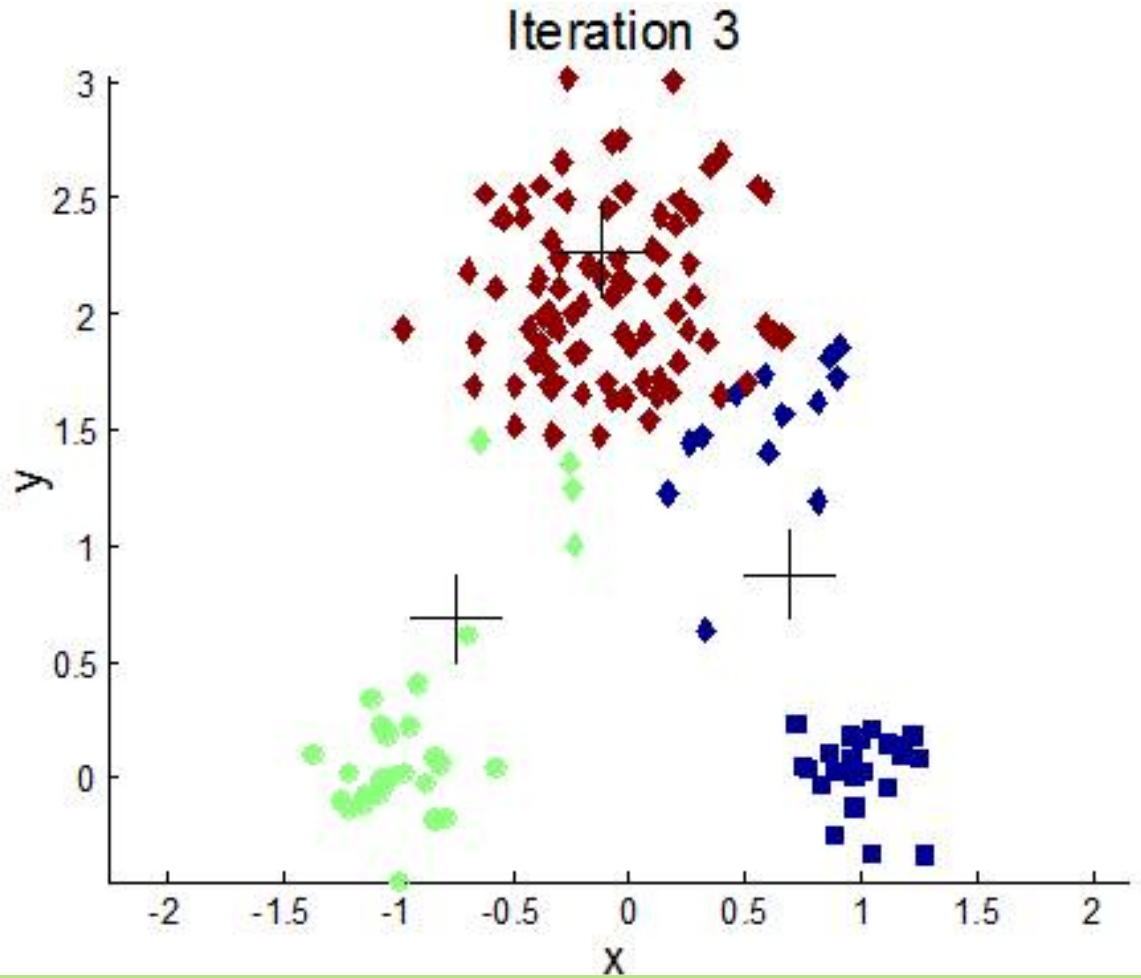
- ⋆ *Funcionamiento*



Técnicas de agrupamiento particional

El método de las k-medias

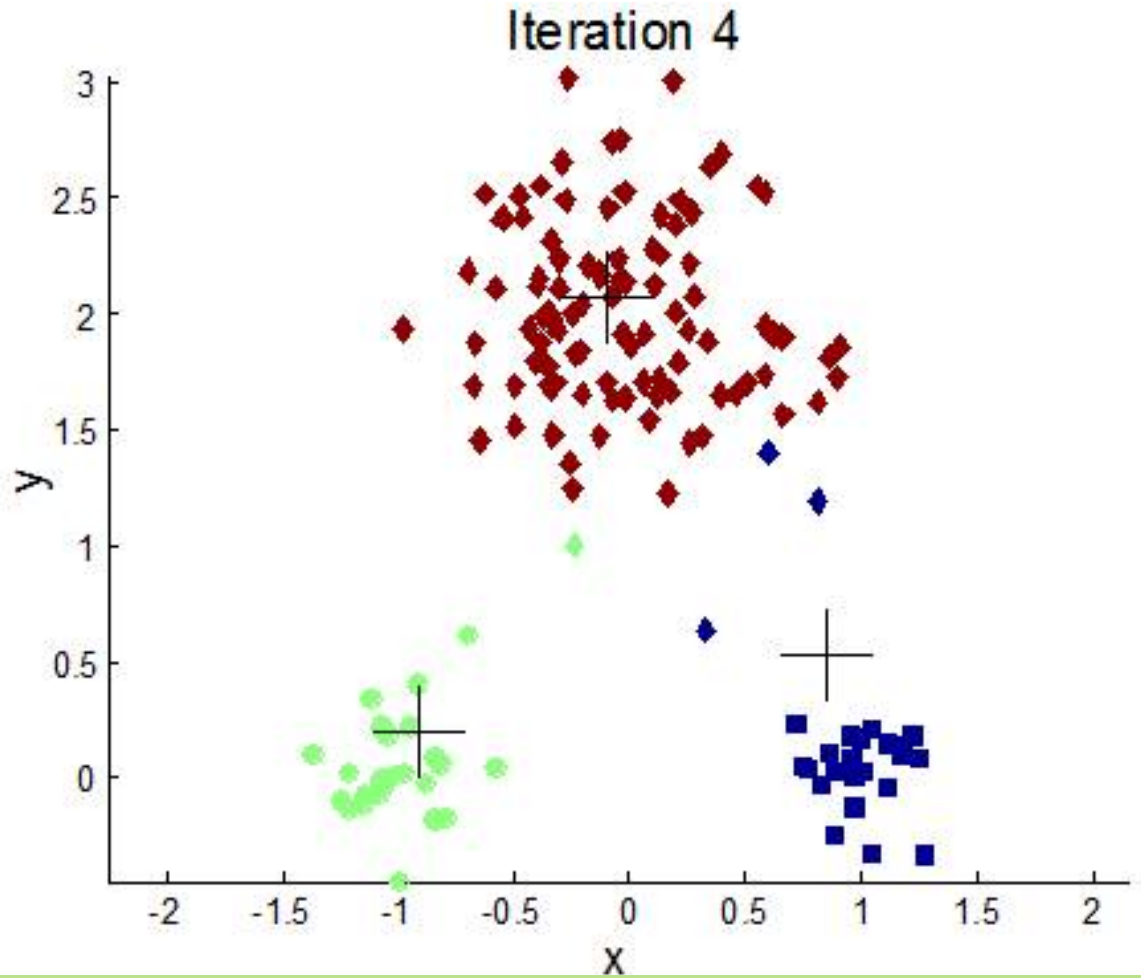
- ⋆ *Funcionamiento*



Técnicas de agrupamiento particional

El método de las k-medias

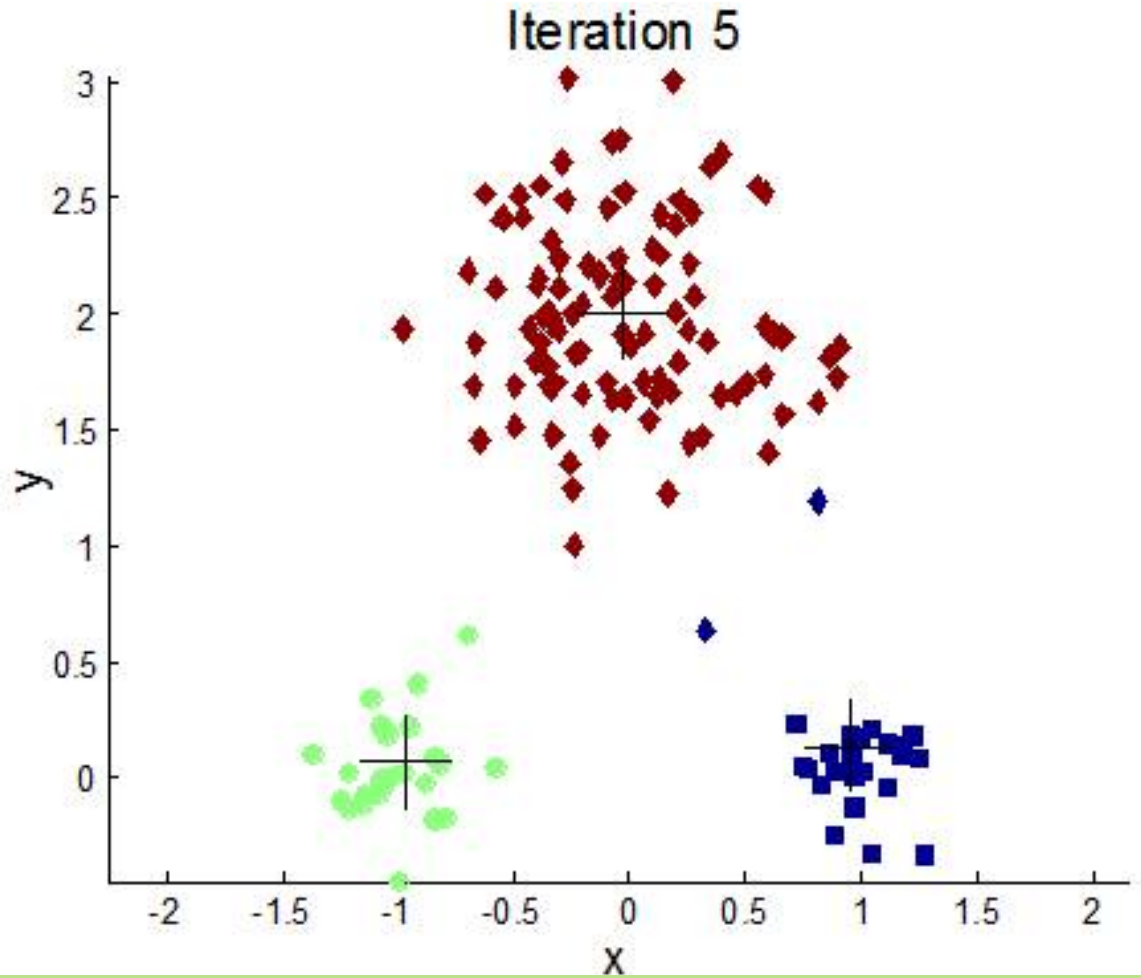
- ⋆ *Funcionamiento*



Técnicas de agrupamiento particional

El método de las k-medias

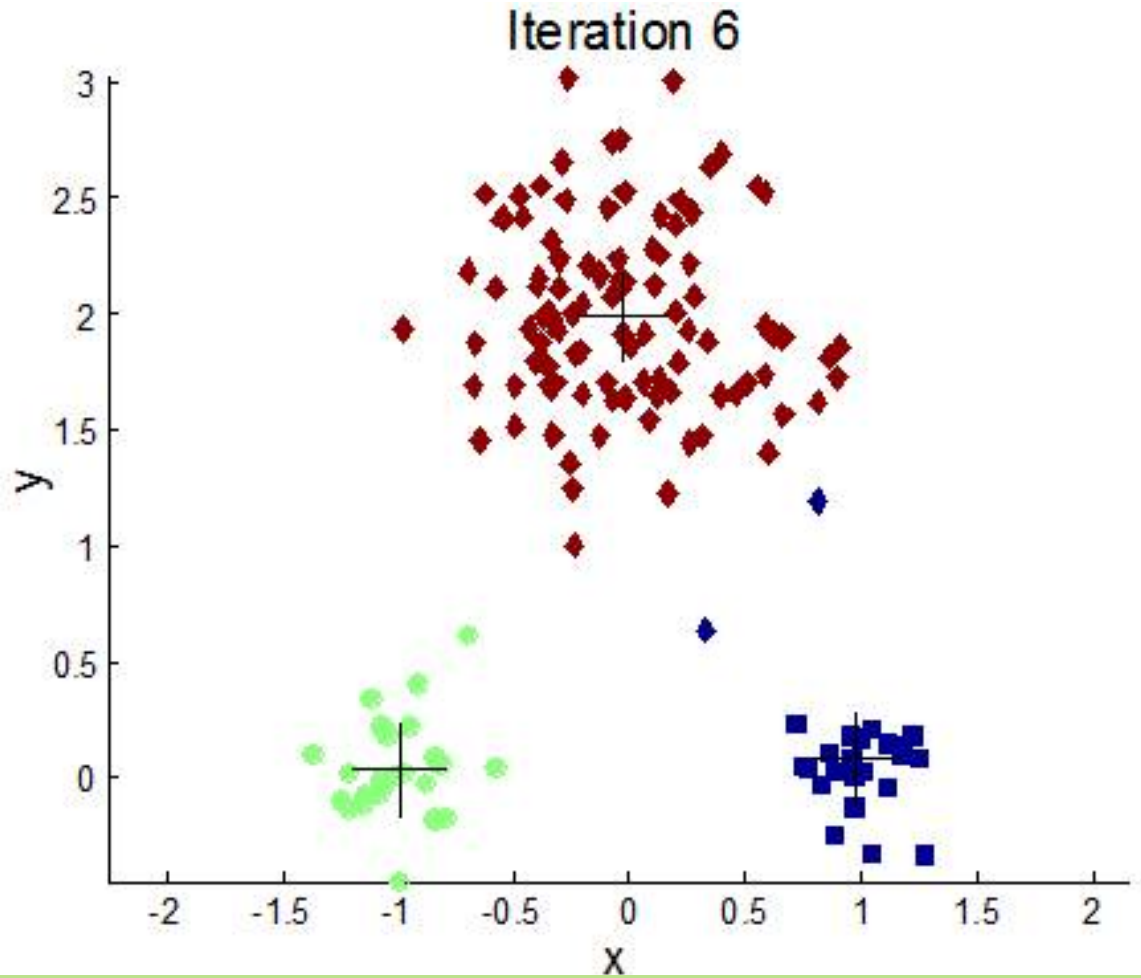
- ⋆ *Funcionamiento*



Técnicas de agrupamiento particional

El método de las k-medias

⋆ *Funcionamiento*



Técnicas de agrupamiento particional

El método de las k-medias

⋆ *Descripción formal del algoritmo*

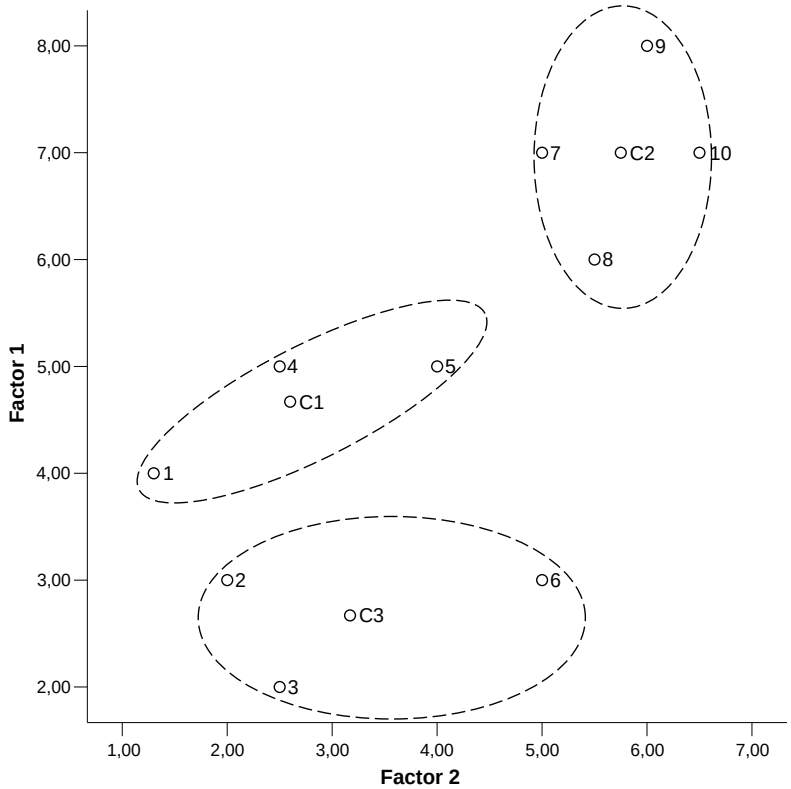
1. Sean $\{x_1 ... x_n\}$ n items definidos en un espacio d-dimensional E, con matriz de datos x_{il}, $i = \{1, ..n\}$ y $l = \{1, ..d\}$ y distancia $p(.,.)$. Elegir K y un conjunto $c_1 ... c_K$ de centroides iniciales. Sea $\{G_1, ...G_k\}$ el conjunto de grupos que vamos a obtener, inicialmente $G_j = \emptyset \; \forall j \in \{1, ...K\}$
2. $\forall i \in \{1, ..n\}$:
   - calcular $j_i \in \{1, ..K\}$ tal que: $p(x_i, c_{j_i}) = min_{j \in \{1,..k\}}(p(x_i, c_j))$
   - Hacer $G_{j_i} = G_{j_i} \cup \{x_i\}$
3. Obtener los nuevos centroides haciendo:
   $\forall j \in \{1, ..K\} \; \forall l \in \{1, .., d\} \; cn_{jl} = \sum_{x_i \in G_j} x_{il}/|G_j|$
4. Si $cn_j = c_j \forall j \in \{1, ..K\}$, parar. En caso contrario:
   - Hacer $cn_j = c_j \forall j \in \{1, ..K\}$
   - Hacer $G_j = \emptyset \; \forall j \in \{1, ...K\}$

   Ir a 2.

Técnicas de agrupamiento particional

El método de las k-medias

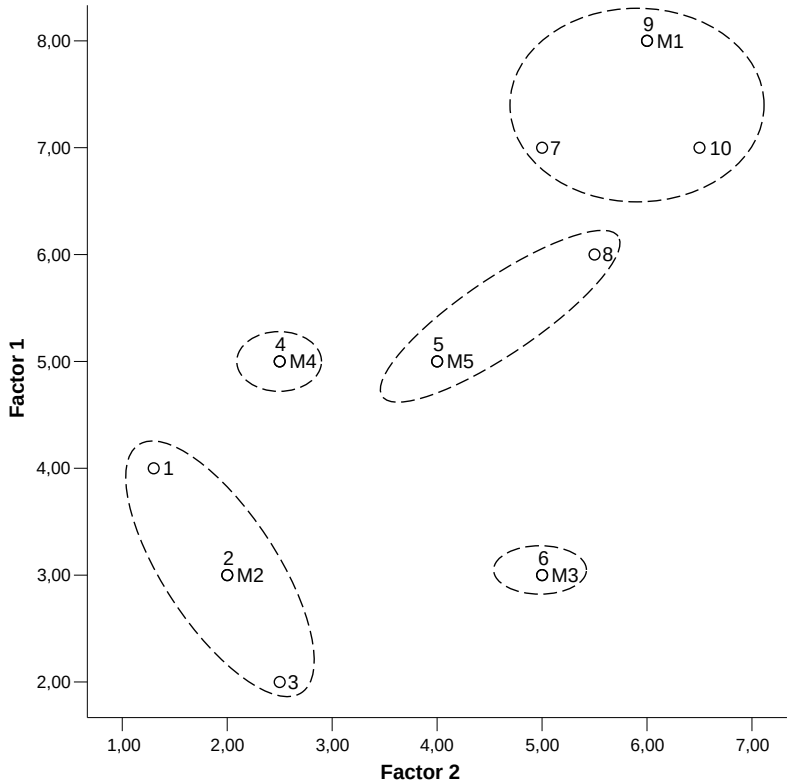
⋆ *Problemas de aplicacion. Resultados con tres grupos*



Técnicas de agrupamiento particional

El método de las k-medias

⋆ *Problemas de aplicacion. Resultados con cinco grupos*



Técnicas de agrupamiento particional

El método de las k-medias: problemas de aplicacion

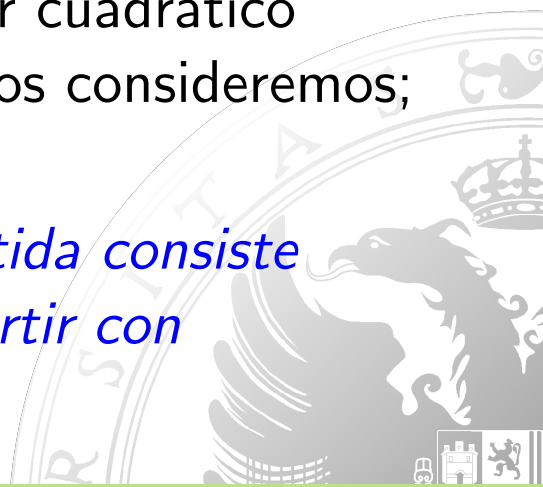
El método de las K-medias es fuertmente dependiente los parámetros de entrada.

Una buena medida de la bondad del agrupamiento es justamente la suma total de la proximidad que se minimiza:

$$SSE = \sum_{j=1}^{j=K} \sum_{x_i \in G_j} p(x_i, c_j)/n$$

Si trabajamos con distancia euclídea, tenemos el error cuadrático global. Este valor será tanto menor cuanto más grupos consideremos; pero esta opción puede no ser la más adecuada

> *Un procedimiento para fijar los parámetros de partida consiste en realizar un agrupamiento jerárquico previo y partir con alguno de los agrupamientos que proporcione*



Técnicas de agrupamiento particional

El método de las k-medias: problemas de aplicacion

⋆ *Ejemplo*

Agrupamiento en tres grupos

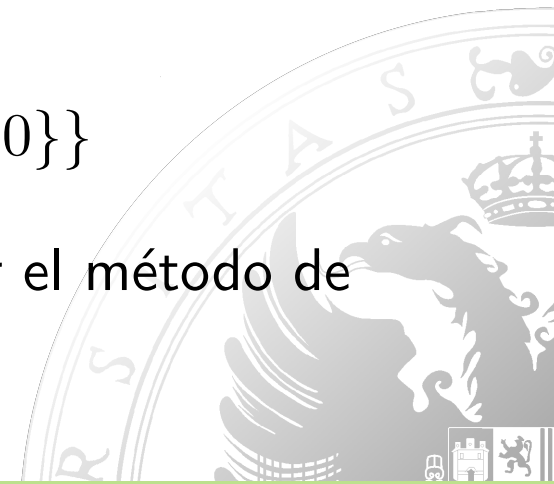
$$P_1 = \{\{1, 2, 3\}, \{4, 5, 6\}, \{7, 8, 9, 10\}$$

proporcionado por el enfoque jerárquico usando el método de enlace completo.

Agrupamiento en cinco grupos

$$P_2 = \{\{1, 2, 3\}, \{4\}, \{5\}, \{6\}, \{7, 8, 9, 10\}\}$$

que es uno de los agrupamientos resultado de aplicar el método de enlace simple



Técnicas de agrupamiento particional

El método de las k-medias: problemas de aplicacion

⋆ *Ejemplo*

La tabla muestra los valores de SSE para cada caso:

- El valor disminuye con el número de grupos

- la elección inicial usando un agrupamiento obtenido por el medio del enfoque jerárquico mejora esta medida de bondad.

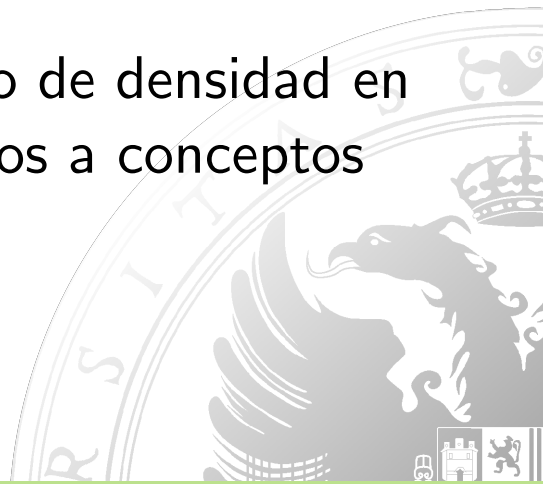
N. de grupos	Sin selección previa	Con selección previa
3	1.083	0.980
5	0.650	0.590

Técnicas de agrupamiento particional

DBSCAN: un método basado en el análisis de densidad

Ideas básicas

- Los métodos de agrupamiento basados en la densidad analizan regiones del espacio de alta densidad que están separadas por otras de baja densidad.
- Todos los métodos se basan en el concepto de *densidad de una región*.
- Existen diversos formas para establecer el concepto de densidad en una región de un espacio, algunos de ellos asociados a conceptos estadisticos



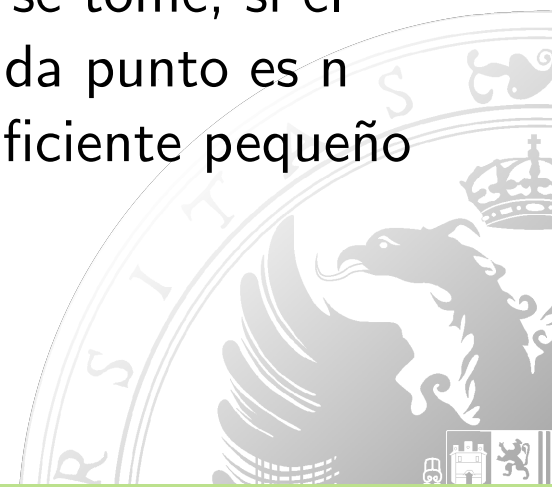
Técnicas de agrupamiento particional

DBSCAN: un método basado en el análisis de densidad

Ideas básicas

El enfoque de DBSCAN es el *densidad basada en centros*:

- La densidad se estima para un punto concreto contando el número de puntos que caen dentro de un entorno centrado en el mismo y de un radio fijado (eps).
- La densidad de cada punto depende del radio que se tome, si el radio es suficientemente grande, la densidad de cada punto es n (número total de puntos); por el contrario si es suficiente pequeño la densidad es de 1 en cada punto.



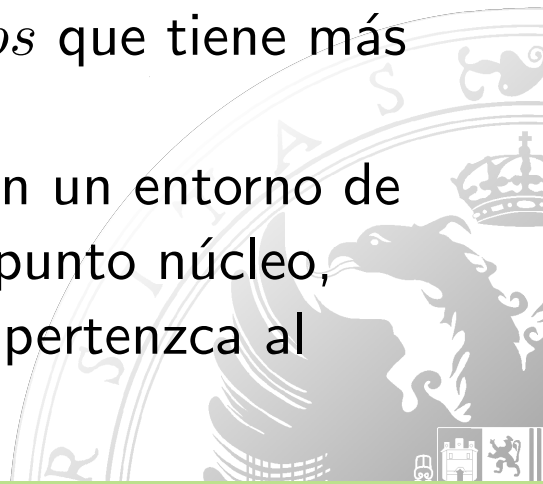
Técnicas de agrupamiento particional

DBSCAN: descripción

1. Fijar
   - 1.1 Un valor de radio eps
   - 1.2 Un número de puntos mínimo: $MinPt$ adecuado para que se considere que un entorno tiene densidad suficiente para formar parte de un grupo
2. Todo punto del espacio de patrones se puede clasificar en como:

   Punto Núcleo. Son aquellos puntos que se considera pertenecen al interior de un grupo y se definen como aquellos que son centro de un entorno de radio eps que tiene más de $MinPt$ puntos.

   Punto Frontera. Son aquellos que se encuentran en un entorno de radio eps que tiene como centro un punto núcleo, puede ocurrir que un punto frontera pertenzca al entorno de varios puntos núcleo.



Técnicas de agrupamiento particional

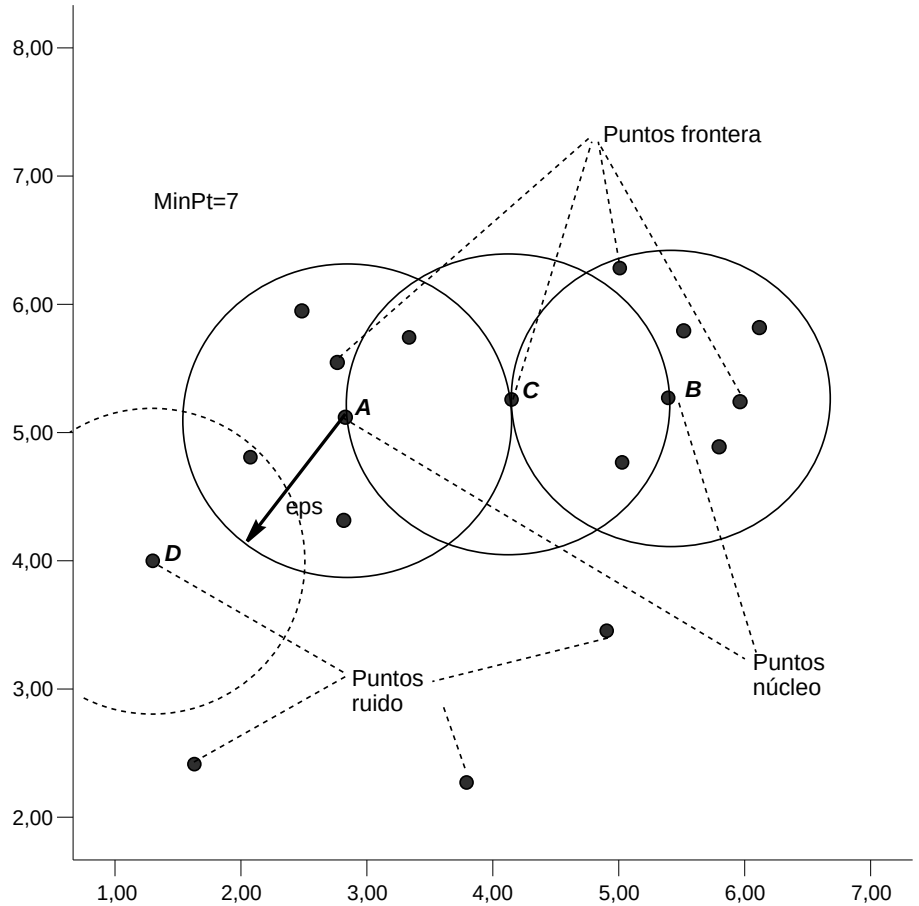
DBSCAN: descripción

2. Todo punto del espacio de patrones se puede clasificar en como:

   Punto ruido. Son los puntos que no son núcleo, ni frontera, se supone que va a estar en regiones muy poco densas y que no van a formar parte de ningún grupo.

Técnicas de agrupamiento particional

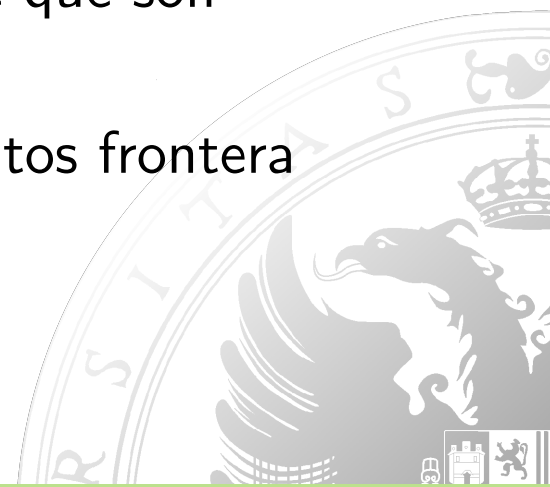
DBSCAN: descripción



Técnicas de agrupamiento particional

DBSCAN: descripción

- Se ponen en un mismo grupo todos los puntos núcleo que distan entre sí menos de eps, utilizando un criterio de transitividad:
  Si la distancia entre dos puntos núcleo $f(n_1, n_2) \leq eps$ y para otro punto núcleo $f(n_2, n_3) \leq eps$ entonces n_1, n_2 y n_3 pertencen al mismo grupo.
  Se dice entonces que n_1 y n_2 (o n_2 y n_3) son *directamente densidad alcanzables* mientras que n_1 y n_3 se dice que son *densidad alcanzables*
- También se asignan al mismo grupo todos los puntos frontera asociados a cada punto núcleo.
- Se eliminan todos los puntos ruido.

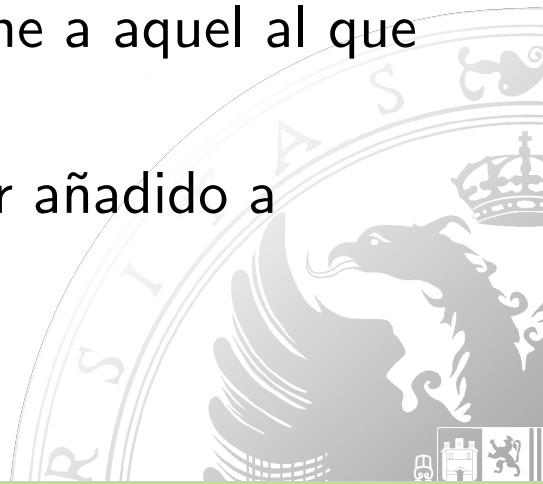


Técnicas de agrupamiento particional

DBSCAN: descripción

DBSCAN funciona realmente de forma iterativa, fijados sus parámetros eps y $MinPt$:

- Va considerando punto a punto, estableciendo su entorno de radio eps y viendo si es o no núcleo, en el caso de que lo sea se construye un grupo con dicho entorno
- Se buscan otros núcleo sean densidad alcanzables a partir de él, si existe alguno, el grupo generado inicialmente se une a aquel al que pertenzca este núcleo.
- El proceso termina cuando ningún punto puede ser añadido a ningún grupo.



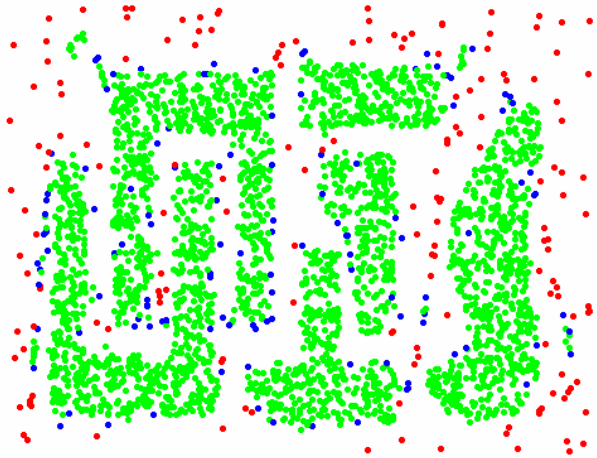
Técnicas de agrupamiento particional

DBSCAN

DBSCAN es un algoritmo potente y sencillo que puede optimizarse para espacios de baja dimensionalidad.
Produce grupos complejos para los cuales no hay ninguna hipótesis de "centralidad" o "globularidad" como en el caso del método de las k-medias.

DBSCAN: Core, Border and Noise Points

Original Points

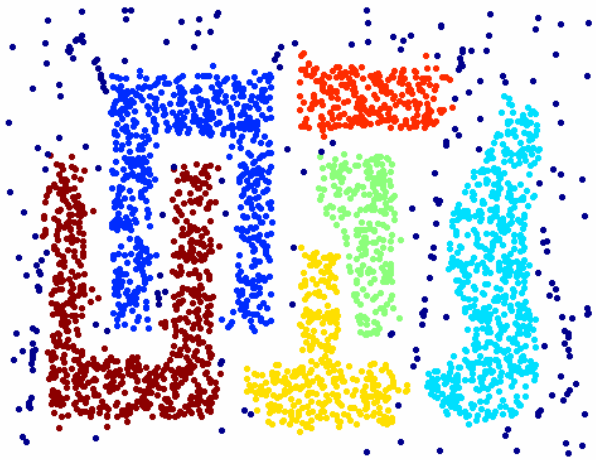


Point types: core, border and noise

Eps = 10, MinPts = 4

When DBSCAN Works Well

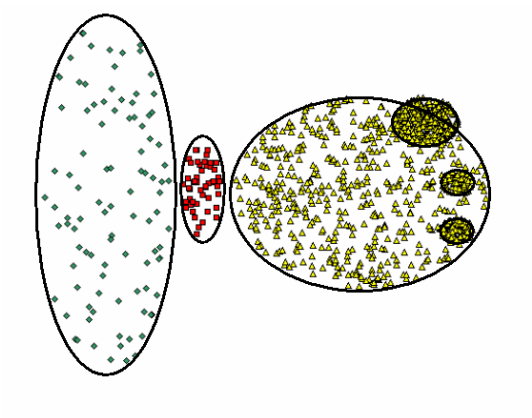
Original Points



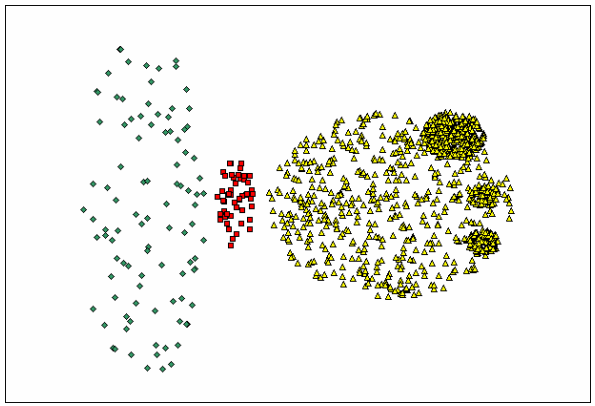
Clusters

- Resistant to Noise
- Can handle clusters of different shapes and sizes

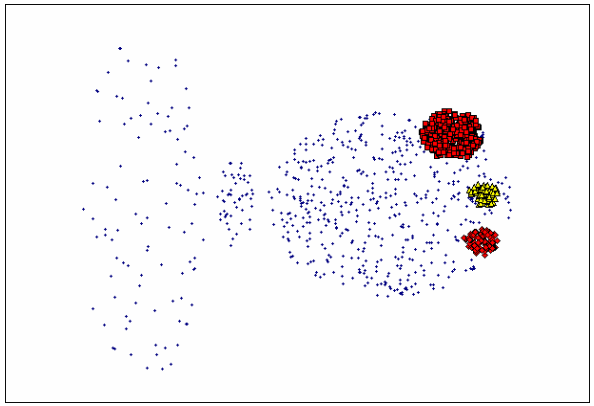
When DBSCAN Does NOT Work Well



Original Points



(MinPts=4, Eps=9.75).



(MinPts=4, Eps=9.92)

- **Varying densities**
- **High-dimensional data**

Técnicas de agrupamiento particional

Problemas en la aplicación de DBSCAN

- Tiene una gran dependencia de los parámetros de partida eps y $MinPt$.
- Puede ocurrir que haya zonas del espacio de patrones muy densas, con grupos que incluyan muchos puntos y otras con una densidad más baja que sin embargo presenten grupos menos densos. Si se toman los mismos parámetros para ambas zonas los puntos de la segunda se considerarán como ruido, con lo que se pierde mucha información.
- Existen generalizaciones de DBSCAN que permiten evitar en lo posible estos problemas, **OPTICS** debido a Ankerst et. al trabaja con regiones de densidad variable, considerando valores de eps menores.

Validación de agrupamientos

Concepto Básicos

Problema

Como evaluar la "bondad" de un resultados de agrupamiento

- Hay que tener en cuenta que, en principio no se conoce nada acerca del problema. Estamos en un entorno *no supervisado*
- Queremos evaluar para:
  - No confundir grupos con "ruido"
  - Comparar distintos algoritmos de agrupamiento
  - Comparar dos conjuntos de clusters
  - Comparar dos clusters

Validación de agrupamientos

Distintos aspectos de la validación de agrupamientos

1. Determinación de la tendencia de agrupamiento de un conjunto de datos, es decir , distinguir si la estructura no aleatoria realmente existe en los datos.
2. Comparación de los resultados de un análisis cluster con una partición conocida previamente.
3. Evaluación de cómo los resultados de un análisis cluster se ajustan a los datos sin referencia a información externa .
4. Comparación de los resultados de dos conjuntos diferentes de análisis cluster para determinar cuál es mejor.
5. Determinación del número "correcto" de grupos

Para 2, 3 , y 4 , podemos distinguir , además, si queremos evaluar la totalidad de la agrupación o simplemente grupos individuales .

Validación de agrupamientos

Medidas para evaluar agrupamientos

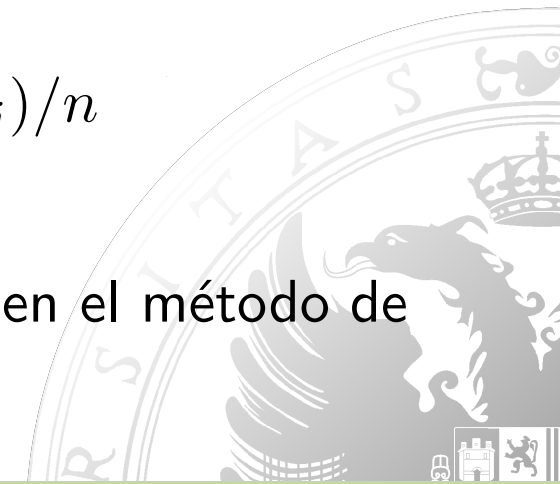
Son medidas que se usan para determinar distintos aspectos de la validez de un cluster

Las hay de distintos tipos:

No supervisadas Se utilizan para medir la bondad de un agrupamiento sin tener ninguna información adicional al respecto. El ejemplo más clasico es:

$$SSE = \sum_{j=1}^{j=K} \sum_{x_i \in G_j} p(x_i, c_j)/n$$

que se aplica como medida de bondad en el método de las k-medias.



Validación de agrupamientos

Medidas para evaluar agrupamientos

No supervisadas Se dividen en dos clases:

- *Medidas de cohesión* miden cómo de compactos son los grupos
- *Medidas de separación* Miden cómo están de separados los grupos.

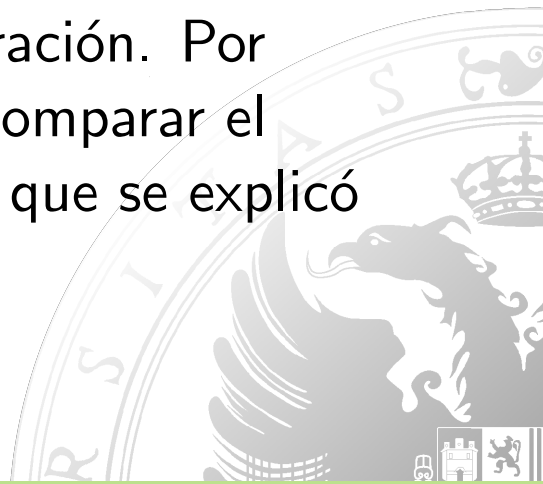
Estas medidas se denominan también **Indices internos** ya que sólo utilizan los datos del problema

Validación de agrupamientos

Medidas para evaluar agrupamientos

Supervisadas Son aquellas que miden la adecuación del agrupamiento obtenido con una particón ya existente. Las medidas supervisadas se denominan también **Indices externos** ya que utilizan información no presente en el data set.

Relativas Comparan diferentes agrupamientos o grupos, pueden ser supervisadas o no pero siempre se formularan de forma relativa con el obetivo de comparación. Por ejemplo el uso de la medida SSE para comparar el proceso de selección de grupos iniciales que se explicó en el método de las k-medias.



Validación de agrupamientos

Validación no supervisada utilizando cohesión y separación

- Las medidas de cohesión y separación se calculan de forma diferente según se hayan obtenido los grupos mediante técnicas basadas en prototipos o se hayan usado otras técnicas.
  - Cuando no se dispone de prototipo tenemos:

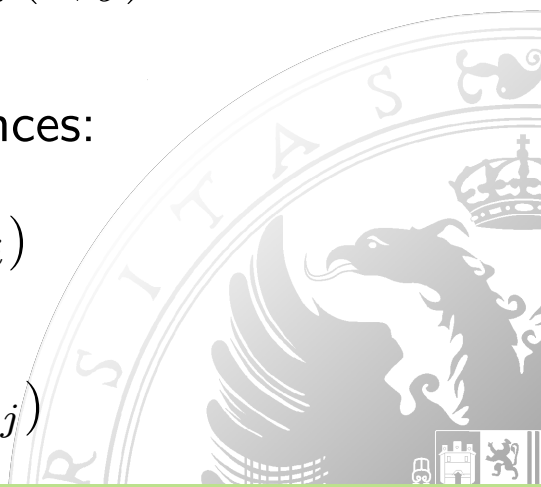
$$\text{cohesion}(C_i) = \sum_{x,y \in C_i} \text{proximity}(x, y)$$

$$\text{separacion}(C_i, C_j) = \sum_{x \in C_i, y \in C_j} \text{proximity}(x, y)$$

  - Cuando tenemos un centro c_i de cada grupos, entonces:

$$\text{cohesion}(C_i) = \sum_{x \in C_i} \text{proximity}(x, c_i)$$

$$\text{separacion}(C_i, C_j) = \text{proximity}(c_i, c_j)$$



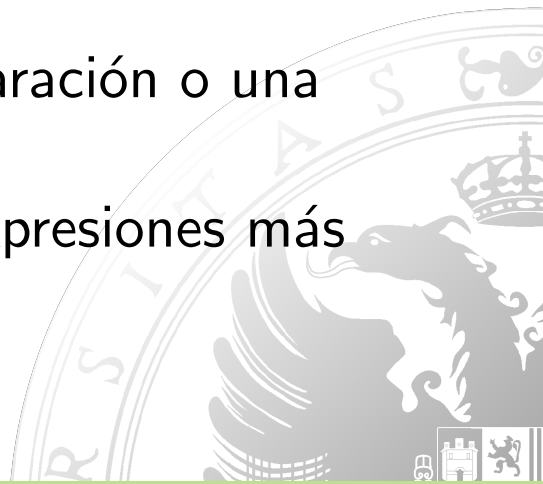
Validación de agrupamientos

Validación no supervisada utilizando cohesión y separación

- En general se considera que la medida de un agrupamiento es la suma ponderada de la medida de us grupos. Es decir, la validez total de un agrupamiento de K grupos es:

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

La validez puede ser una medida de cohesión, separación o una combinación de ambos
Los pesos pueden ser, 1 el tamaño del cluster o expresiones más sofisticadas.



Validación de agrupamientos

Validación no supervisada. Coeficiente de silueta (Silhouette Coefficient)

- Combina las ideas de separación y cohesión, se calcula para cada punto de la siguiente forma:
- Para cada punto i
  1. Se calcula la distancia media de i a los elementos de su grupo. Sea a_i
  2. Se calcula la media de la distancia de i a los elemento de cada grupo que no es el suyo y se hace el mínimo de todas estas medias. Sea b_i
  3. El coeficiente de silueta para i viene dado por:

$$s_i = \frac{(b_i - a_i)}{max(a_i, b_i)}$$

- El coeficiente varía entre -1, y 1. Es deseable un coeficiente positivo y cerca del 1.
- El coeficiente de silueta de un grupo o un agrupamiento se calcula mediante la media de los correspondientes coeficientes.

Validación de agrupamientos

Validación supervisada

Existen dos enfoques

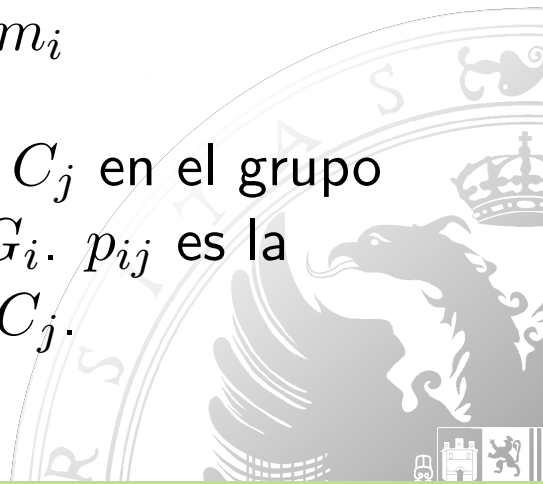
1.- Orientado a clasificación

Miden cómo se ajusta la partición obtenida en un agrupamiento a una clasificación previamente dada.

Sean $\{G_i; i \in \{1, ..K\}\}$ los grupos obtenidos y $\{C_j; j \in \{1, ..L\}\}$ las clases a comparar, sea m el número total de puntos. Definimos:

$$\forall i \in \{1, ..K\}; j \in \{1, ..L\}\} \; p_{ij} = m_{ij}/m_i$$

donde m_{ij} es el número de items que hay de la clase C_j en el grupo G_i y m_i es el número de items que hay en el grupo G_i. p_{ij} es la probabilidad de que un miembro de G_i pertenezca a C_j.



Validación de agrupamientos

Validación supervisada: orientada a la clasificación

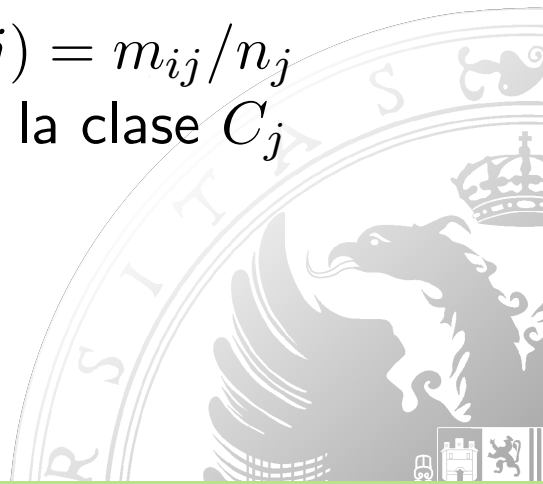
A partir de p_{ij} se definen:

Entropía $\forall i \in \{1, ..K\}\ e_i = -\sum_{j=1}^{L} p_{ij} log_2 p_{ij}$ es la entropía de un cluster $e = \frac{\sum_{i=1}^{K} m_i e_i}{m}$ es la entropía total

Purity La "purity" de un grupo es $p_i = max_j(p_{ij})$ la total del agrupamiento es $purity = \frac{\sum_{i=1}^{K} m_i p_i}{m}$

Precisión y Recall $precision(i,j) = p_{ij}$; $recall(i,j) = m_{ij}/n_j$ donde n_j es el número de elementos de la clase C_j

F-medida $F(i,j) = \frac{2 \times precision(i,j) \times recall(i,j)}{precision(i,j) + recall(i,j)}$



Validación de agrupamientos

Validación supervisada

2.- Orientado a similaridad

La idea básica se construir las matrices de incidencia del agrupamiento:

$$IG_{ij} = \begin{cases} 1 & \text{si i j están en el mismo cluster} \\ 0 & \text{en caso contrario} \end{cases}$$

y de la clasificación.

$$IC_{ij} = \begin{cases} 1 & \text{si i j están en la misma clase} \\ 0 & \text{en caso contrario} \end{cases}$$

y establecer medidas de coincidencia entre ambas
Una de las más comunes es calcular la **correlación** entre ambas matrices.

Validación de agrupamientos

Validación supervisada. Orientado a similaridad

Otra alternativa. Sean la cantidades de parejas de puntos que coinciden o difieren, según la tabla:

	Igual cluster	Diferente cluster
Igual clase	f_{11}	f_{10}
Diferente clase	f_{01}	f_{00}

Se definen:

Coeficiente de Jaccard $\frac{f_{11}}{f_{10}+f_{01}+f00}$

Estadistico de Rand $\frac{f_{11}+f_{00}}{f_{10}+f_{01}+f00+f_{11}}$

Extensiones de los métodos estudiados

Extensiones en agrupamiento jerárquico

⋆ **Problema**

Las técnicas de agrupamiento jerárquico siempre han tenido el inconveniente suelen ser muy costosas desde el punto de vista computacional. Complejidad mínima $O(n^2)$ Este inconveniente se agrava ya que, en su versión clásica, , estas técnicas obtienen el conjunto de todos los posibles agrupamientos, desde el inicial con n grupos hasta el último con un sólo grupo. Complejidad en el peor caso $O(2^n)$

Las primeras versiones que tratan de evitar estos inconvenientes son los algoritmos llamados **DIANA** y **AGNES** (Kauffman y Rousseeuw 1990), el primero con un enfoque divisivo y el segundo aglomerativo, en ambos se especifica el número de grupos que se desea como una condición de terminación

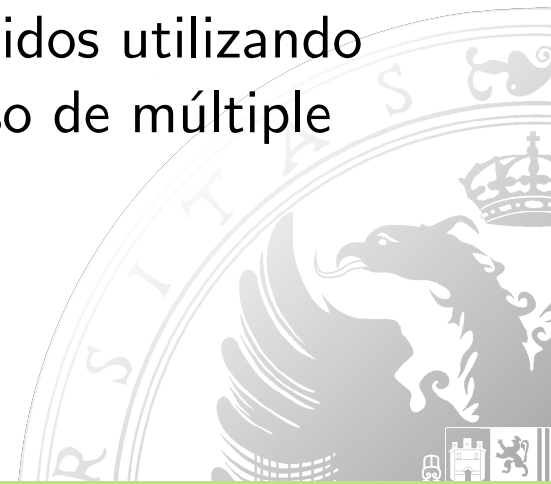
Extensiones de los métodos estudiados

Resultados recientes en agrupamiento jerárquico

⋆ **Problema**

También se considera un gran inconveniente de los métodos jerárquico clásicos el hecho de que, cuando se toma la decisión de unir dos grupos en los enfoques aglomerativos (dividir un grupo en los divisivos), no se puede volver atrás

La solución es mejorar la calidad de los grupos obtenidos utilizando otras técnicas de agrupamiento, realizando un proceso de múltiple fases. Entre las soluciones más populares están

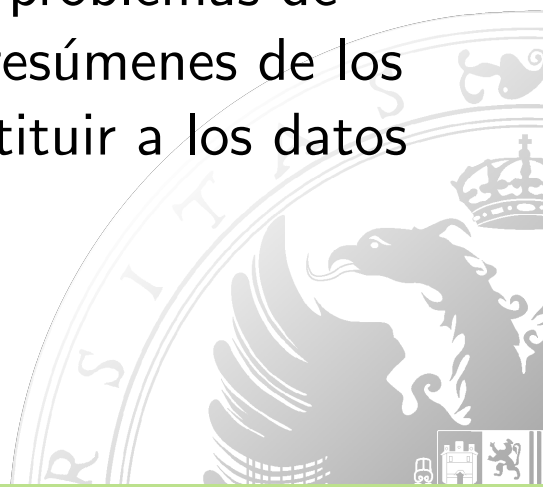


Extensiones de los métodos estudiados

Resultados recientes en agrupamiento jerárquico

BIRCH Balanced Iterative Reducing and Clustering using Hierarchies

(Zhang et. al 1966) que inicialmente particiona los patrones de forma jerárquica utilizando estructuras de árboles y posterioremente aplica otros algoritmos de agrupamiento para refinar el resultado. Este algoritmo es facilmente escalable y muy usado en problemas de Minería de Datos se basa en el uso de resúmenes de los datos (estadísticos suficientes) para sustituir a los datos originales.



Extensiones de los métodos estudiados

Resultados recientes en agrupamiento jerárquico

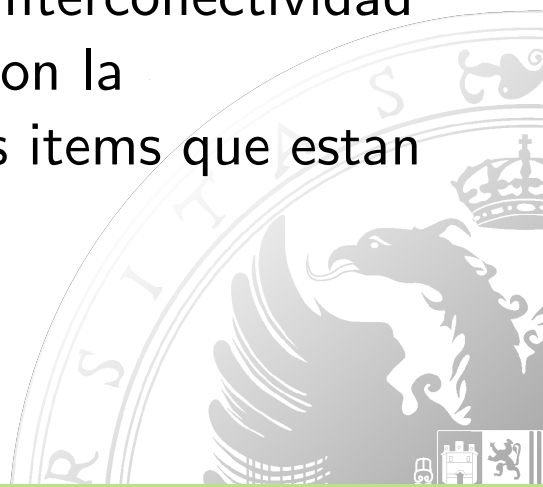
CURE **C**lustering **U**sing **RE**presentatives (Guha et. al.1998) Este método emplea de agrupamiento jerárquico que se encuentra a mitad de camino entre los métodos de enlace ponderado y de enlace completo, ya que en lugar de utilizar un único punto, o todos ellos para representar un grupo, se elige un número fijo de puntos representativos. Estos puntos se generan elegiendo primeramente un conjunto de puntos bien distribuidos sobre el grupo y posteriormente reduciendo la distancia de estos al centro del grupo un determinado factor (factor de reducción). Los grupos con puntos representativos más cercanos se unen en cada paso del algoritmo.

Extensiones de los métodos estudiados

Resultados recientes en agrupamiento jerárquico

ROCK Otro algoritmo aglomerativo desarrollado por los mismos autores que está orientado al uso de atributos categóricos

CHAMELEON (Karpys et. al. 1999) explora un modelo dinámico de agrupamiento jerárquico. En su proceso de agrupamiento dos grupos se unen si la interconectividad y la cercaní entre ellos se corresponde con la conectividad interna y la cercanía de los items que estan en los grupos.



Extensiones de los métodos particionales. El método de las k-medias difuso

Introducción a las técnicas de agrupamiento difuso

⋆ Ideas Básicas

- En todos los métodos de agrupamiento clásicos se ha supuesto, al menos implicitamente, la hipótesis de que el agrupamiento es exclusivo, es decir los patrones se particionan en conjuntos disjuntos.
- Si los grupos son compactos y están bien separados esta es la mejor opción, pero el problema aparece cuando los grupos tienen puntos comunes e incluso se solapan. Tenemos entonces conjuntos cuyas fronteras están mal definidas o "borrosas".
- La teoría de subconjuntos difusos (fuzzy sets),permite que un patrón pertenezca a un grupo con un cierto "grado de pertenencia".

Extensiones de los métodos particionales. El método de las k-medias difuso

Introducción a las técnicas de agrupamiento difuso

⋆ Ideas Básicas

- Cada grupo difuso $C_j;\ j \in \{1, .., K\}$ tiene asociada una "función de pertenencia:

$$C_j : X \longrightarrow [0, 1],$$

  siendo $X = \{x_1, ...x_N\}$ el espacio de patrones.
- El valor $u_{ij} = C_j(x_i)$ mide el grado de pertenencia del punto x_i al grupo C_j. Los valores u_{ij} constituyen la "matriz de pertenencia" que notaremos U.
- Es habitual imponer la condición *de partición difusa* o *posibilística*:

$$\sum_{j=1}^{K} u_{ij} = 1, \forall i \in \{1, ..N\}, \quad \max_{j \in \{1...K\}} u_{ij} = 1, \forall i \in \{1, ..N\},$$

Extensiones de los métodos particionales. El método de las k-medias difuso

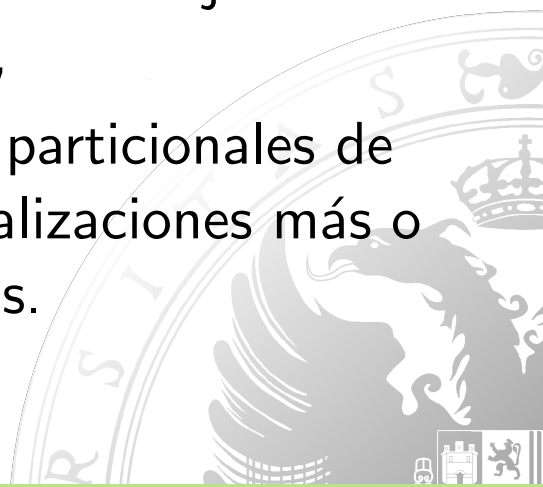
⋆ Ideas Básicas

- El grado de pertenencia no tiene el mismo sentido que una probabilidad. El grado de pertenencia se puede interpretar como el grado de compatibilidad del punto x_i con el grupo C_j, entendido este como el resultado de una propiedad (o un conjunto de propiedades) expresadas de forma imprecisa.
- Este enfoque es muy útil cuando se intenta una interpretación de los grupos, ya que, en muchos casos, las descripciones de los grupos obtenidos en un problema concreto serán de tipo impreciso por serlo las etiquetas que los caracterizan.

Extensiones de los métodos particionales. El método de las k-medias difuso

Introducción a las técnicas de agrupamiento difuso

- Por ejemplo si se intenta agrupar un conjunto de coches itentado obtener los de gama "alta", "media" o "utilitario" o si se intenta agrupar un conjunto de parcelas atendiendo a las prácticas de cultivo que realizan sobre ellas.
- Casi todos los algoritmos de agrupamiento basados en conjuntos difusos hacen uso del concepto de partición difusa,
- Existen distintos enfoques para diseñar algoritmos particionales de tipo difuso, la mayor parte de los cuales son generalizaciones más o menos directas directas del método de las k-medias.



El método de las k-medias difuso

Descripción del método

1. Seleccionar una partición difusa inicial de N objetos en K grupos seleccionando una matriz de pertenencia U.
2. Calcular los "centros" de los grupos difusos asociados a U mediante la expresión: $c_j = \sum_{i=1}^{N} u_{ij} x_i$
3. Calcular el valor óptimo de:

$$E^2(U) = \sum_{i=1}^{N} \sum_{j=1}^{K} u_{ij} ||x_i - c_k||^2 \qquad (4)$$

   Tenemos un problema de optimización sobre los valores de pertenencia u_{ij}, sujetos a:
   $\sum_{j=1}^{K} u_{ij} = 1;\ u_{ij} \geq 0$ o bien $max_{j \in \{1..K\}} u_{ij} = 1;\ u_{ij} \geq 0$
4. Repetir desde el paso 2 hasta que los valores de U no cambien significativamente.

El método de las k-medias difuso

Descripción del método

⋆ **Variantes**

• Que el centro de un grupo difuso no es su media sino su valor más representativo es decir:

$$\forall j \in \{1..K\} \; ; \; c_j = x_{lj} \mid u_{lj} = max_{i \in \{1,..n\}} u_{ij}$$

• Otras funciones de distancia más generales.

La distancia entre dos puntos asociada al grupo C_j:

$$\forall x, y \; d_j(x, y) = min(u_j(x), u_j(y))||y - x||^2$$

La distancia se transforma en:

$$S^2(U) = \sum_{i=1}^{N} \sum_{j=1}^{K} d_j(x_i, c_j) \tag{5}$$

Si $u_j(c_j) = 1$ obtenemos la expresión inicial

• Otras variantes utilizan una función de distancia adaptativa

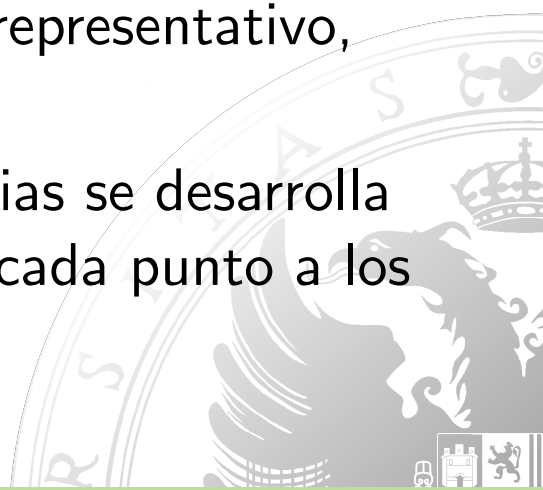
Extensiones del método de las k-medias.

Los métodos de k-medoides

Problema

El método de las k-medias supone que el espacio de items es continuo y por tanto que el centroide puede ser un posible item. Esto solo pasa cuando todos los datos son numéricos y continuos.

- Una alternativa al uso del centroide que toma como prototipo de cada grupo un punto del mismo que se considera representativo, Este punto que se denomina *medoide*
- Todo el proceso iterativo del método de las k-medias se desarrolla ahora minimizando las sumas de las distancias de cada punto a los medoides considerados.

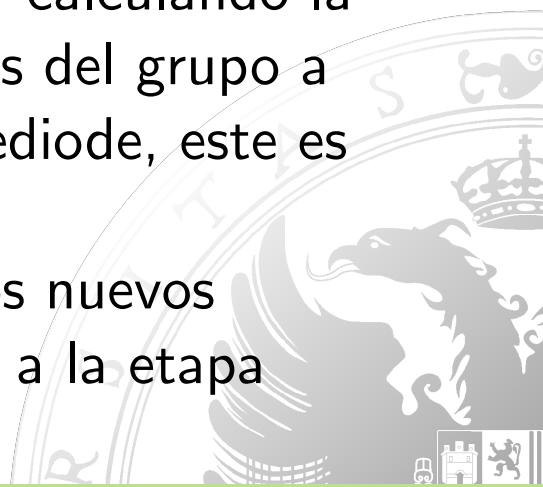


Extensiones del método de las k-medias.

Los métodos de k-medoides

PAM (Kauffmann y Rousseau 1990)

- Se parte de una selección inicial de k medoides. Como en el caso de las k-medias, estos generan una partición en k grupos del conjunto total de items.
- Para cada grupo obtenido se intenta reemplazar el medoide asociado a el por algún punto del mismo grupo que sea más idóneo. Esto se hace considerando cada punto del grupo y calculando la distancia total del resto los elementos del grupo a dicho punto, si esta mejora la del mediode, este es sustituido por el punto en cuestión.
- Se recalculan los grupos en base a los nuevos medoides seleccionado. Y se vuelven a la etapa anterior con ellos.



Extensiones del método de las k-medias.

Los métodos de k-medoides

CLARA (Kauffmann y Rousseau 1990)

· PAM no se adapta bien a grandes bases de datos, la alternativa propuesta es CLARA, basado en un proceso de muestreo.

· CLARA realiza suscesivos muestreos y aplica PAM a cada uno de ellos, los conjuntos de medoides obtenidos se aplican al conjunto total, seleccionando que el nos dé menor distancia global.

· Con estos puntos de partida se genera un nuevo proceso de muestreo y una nueva iteración. La complejidad de cada iteración es ahora de $O(kS^2 + k(n-k)$, donde S es el tamaño de la muestra.

· La efectividad de CLARA depende del tamaño del la muestra, si bien es fácilmente escalable y puede trabajar

Extensiones del método de las k-medias.

Los métodos de k-medoides

CLARANS (NG y Han 1994)

· Se considera la búsqueda de los medoides óptimos como un proceso de búsqueda en un árbol donde cada nodo es un conjunto de k medoides.

· Considerado un nodo, el agrupamiento obtenido reemplazando un medoide del mismo por algún otro punto se denomina entorno.

· El proceso prueba una serie de entornos generando puntos aleatoriamente, si encuentra un entorno mejor que el agrupamiento considerado, el algoritmo se mueve a este nodo y comienza de nuevo a probar, en caso contrario se considera que se ha llegado a un óptimo local.

Extensiones del método de las k-medias.

Los métodos de k-medoides

CLARANS (NG y Han 1994) · Cuando se llega a un óptimo local, el algoritmo comienza con un nuevo conjunto de nodos obtenidos por medio de un muestreo aleatorio y una aplicación de PAM.

· El algoritmo termina cuando se han alcanzado un número suficiente de mínimos locales (datos experimentales recomiendan 2 como este número).

·La complejidad de CLARANS es de $O(n^2)$. Ester, Kriegel and XU 1995 han mejorado este algoritmo mediante el uso de R^*-árboles.