Universidad de Santiago de Chile
Facultad de Ingeniería
Depto. de Ingeniería Informática.

# *Análisis de Datos*
# *Capítulo III*
# *"Análisis de Agrupamientos"*

Profesor: Dr. Max Chacón.

## Objetivos

- **Establecer diferencias entre agrupamientos jerárquicos y no jerárquicos**
- **Comprender los conceptos de similaridad en espacios *n*-dimensionales como un concepto de distancia**
- **Comprender la estructuración de un agrupamiento jerárquico**
- **Cuantificar las medidas de similaridad y su aplicación a la agrupación**
- **Comprender los algoritmos básicos de los agrupamientos**
- **Comprender las medidas de calidad para evaluar agrupamientos.**

## 3.1. Medidas de Similaridad

- La medida fundamental para el agrupamiento, es la similaridad (asociación, proximidad) o la distancia en $\Re^n$.

- *Similaridades:*

- Una similaridad debe cumplir las condiciones de una distancia (una distancia corresponde a una disimilaridad):

- *No-negatividad: $d(x,y) \geq 0$*

- *La distancia de una instancia (observación) así misma es cero, $d(x,x) = 0$*

- *Simetría: $d(x,y) = d(y,x)$*

- *Desigualdad Triangular:*

$$d(x,y) \leq d(x,z) + d(z,y)$$

**Las medidas de similaridad más conocidas son las de distancia.**

**Para dos vectores $\vec{x}$ e $\vec{y} \in \Re^n$**

$$\|\vec{x} - \vec{y}\| = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}$$

**dependiendo del valor de *p* se generan los siguientes casos particulares**

**- *p=1* Distancia de Manhattan (block):**

$$\|\vec{x} - \vec{y}\| = \sum_{i=1}^{n} |x_i - y_i|$$

**- *p=2* Distancia Euclidiana:** $\|\vec{x} - \vec{y}\| = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$

**- $p \to \infty$ Distancia de Schebyshev:**

$$\|\vec{x} - \vec{y}\| = \max_{i=1,2\ldots n} |x_i - y_i|$$

***Otras sin ponderar:***

***Distancia de Camberra:***

$$D_{Camb} = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i + y_i|}$$

***- Distancias de formas cuadráticas, la más general es la distancia de Mahalanobis:***

$$\|\vec{x}-\vec{y}\| = (\vec{x}-\vec{y})^T M^{-1}(\vec{x}-\vec{y})$$

***con M una matriz definida positiva.***

- ***La principal característica de esta distancia es que representa las interrelaciones entre las características individuales. Pero no es fácil obtener un escalamiento adecuado cuando las componentes están representadas en rangos diferentes.***

***-Distancia de Bhattaharaya***

***Con:*** $M^{-1} = \begin{bmatrix} \sigma_1^{-2} & 0\ldots & 0 \\ 0 & \sigma_2^{-2}\ldots & 0 \\ 0 & 0 & \sigma_n^{-2} \end{bmatrix}$ $\|\vec{x}-\vec{y}\| = \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{\sigma_i^2}$

***Si M=I la distancia es la Euclidiana.***

***- Distancia de Hamming (caso binario)***

***Sea x e y dos vectores binarios del mismo largo (n).***

$$\sum_{i=1}^{n} x_i \oplus y_i$$

***Es importante notar que todas las funciones de distancia tienden a un modelo de región convexa en el espacio n-dimensional de las características.***

***Medidas de Correlación:*** $\dfrac{(\vec{x}-\bar{\vec{x}})'(\vec{y}-\bar{\vec{y}})}{\left\|\vec{x}-\bar{\vec{x}}\right\|\left\|\vec{y}-\bar{\vec{y}}\right\|}$

***- Correlación de Pearson (r) [-1,1]***

$$r(\vec{x},\vec{y}) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

***•Correlación de Spearman (r) (variables cuantitativas y ordinales) [-1,1]***

$$\rho(\vec{x},\vec{y}) = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

***•Para calcular d se parean las dos variables, luego una es ordenadas en rangos (ej: se ordena x, menor valor de x se asigna 1, al segundo 2, etc), la segunda variable (y) se le asignan los rangos correspondientes a sus valores, el valor d será la diferencia entre los rangos de x e y. Cuando hay valores iguales se toma el promedio entre rangos consecutivos.***

- *Correlación de Cramer (V)* *[0,1]*

$$V(\vec{x},\vec{y}) = \sqrt{\frac{\chi^2}{n(q-1)}}$$

*$c_2$: diferencias al cuadrado entre valores observados y esperados de la tabla de contingencia general, tamaño r*x*s.*

*q: mínimo entre las filas y columnas de la tabla de contingencia min{r,s}.*

*n: total de casos.*

*Otros:*

- *Correlación de Kendall (t)*
- *Coeficiente de Goodman-Kruskal (g)*

*Similaridad de variables binarias:*

*Se usa como base la tabla de contingencia de 2*x*2*

| | x | ¬x |
|---|---|---|
| y | a | b |
| ¬y | c | d |

- *Euclidiana binaria:* $Eb(x,y) = \sqrt{a+c}$

- *Diferencia de tamaño:* $Dt(x,y) = \frac{(b-c)^2}{(a+b+c+d)^2}$

*-Diferencia de tamaño:*

$$Dt(x,y) = \frac{(b-c)^2}{(a+b+c+d)^2}$$

*-Diferencia de configuración:*

$$- DCb(x,y) = \frac{bc}{(a+b+c+d)^2}$$

*[0,1]*

*-Varianza binaria:*

$$-Vb(x,y) = \frac{b+c}{4(a+b+c+d)^2}$$

*- Dispersión: Db(x,y) =*

$$\frac{ad-bc}{(a+b+c+d)^2}$$

*[0,1]*

*- Coeficiente Pi: p(x,y) =*

$$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

*esta similaridad es la versión binaria del coeficiente de correlación de Pearson.*

- *Coeficiente de Hamann: H(x,y) =*

$$\frac{(a+d)-(b+c)}{(a+b+c+d)} \qquad [-1,1]$$

-*Coeficiente de Jaccard: J(x,y) =* $\dfrac{a}{a+b+c}$

- *Ochiai: Och(x,y)=* $\dfrac{a}{\sqrt{(a+b)(a+c)}}$ *[0,1]*

- *versión binaria del coseno.*

**Para un conjunto de características son ordenadas en una matriz de similaridad (o disimilaridad) que será una matriz triangular (pxp) donde los pares (fila, columna) representan las relaciones entre las características del conjunto de datos.**

**Las matrices de similaridad son el insumo para aplicar los algoritmos de agrupamientos y contienen toda la información necesaria respecto de los objetos o patrones que serán agrupados.**

**Los principales objetivos del análisis de agrupamiento son:**

- ***Exploración de Datos.***
- ***Reducción de Datos.***
- ***Predicciones basadas en grupos establecidos.***

Los métodos de agrupamientos se pueden dividir en:

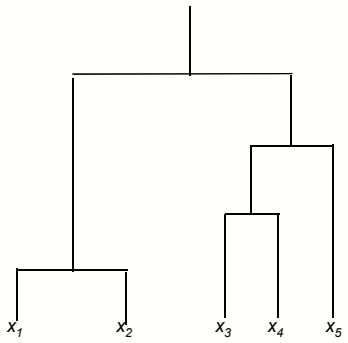
- ***Agrupamiento Jerárquico***
- ***Agrupamiento no Jerarquico.***

## 3.2. Agrupamiento Jerárquico

La metodología jerárquica trabaja con los datos de entrada previamente normalizados y dispuestos sobre vectores o matrices de datos.

Se intenta formar estructuras en forma de árboles que establecen las relaciones entre los datos.

La forma de representación se denomina árbol jerárquico o dendrograma.



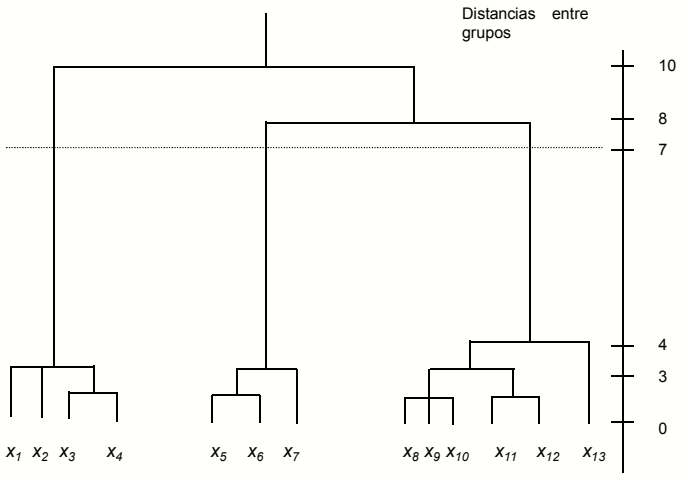

Las entidades de la raíz representan toda la colección de datos indistintamente.

En la dirección ascendente aparecen las relaciones que se van estableciendo entre los datos para formar gr***upos.***

Un árbol jerárquico es una secuencia anidada de particiones de los individuos en ***g*** grupos, donde ***g*** varía de ***1*** a ***n***.

Las particiones se desarrollan en orden ascendente.

Un árbol es una familia de grupos donde cada rama contiene cierto número de nodos. Cada nodo desciende de una rama y así sucesivamente. Ninguna línea que forme el árbol debe interceptarse.



Al considerar una cierta distancia, como por ejemplo 7, se observan tres grupos bien definidos. Los individuos se encuentran muy cercanos unos con otros dentro de cada agrupación.

- Primero se agrupan todos los individuos a distancias menores que 3.
- Luego los individuos o grupos a distancias entre 3 y 4.
- Se repite el proceso anterior para distancias superiores.

Método principal para la construcción de los árboles jerárquicos. También se denomina agrupamiento por vecinos.

Algoritmo:

1. Comenzar con ***n*** grupos, donde cada uno de los grupos sólo contiene un individuo.
2. Unir los dos individuos más cercanos (Ej: individuos ***i*** y ***j***, en grupo simple ***k***). Por lo tanto ahora se encuentran ***(n-1)*** grupos.
3. La diferencia entre este nuevo grupo a cualquier otro individuo ***t***, es definida como: $\boldsymbol{min(\|\vec{x}_k - \vec{x}_t\|)}$.

Algoritmo:

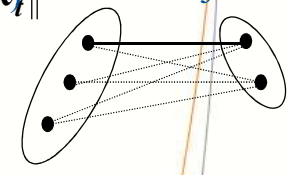
4. Unir los dos grupos más cercanos, pero considerando el grupo formado en el paso ii).
5. Construir una nueva diferencia entre los grupos que quedaron al realizar los pasos anteriores esto es (***n-2***) grupos.
6. Continuar combinando los grupos, siempre reduciendo el número de los grupos en uno y la diferencia resultante entre los nuevos grupos es definida nuevamente por los grupos más cercanos.
7. Repetir los pasos anteriores hasta obtener la cantidad de grupos deseados.

Notar que la agrupación se va realizando al reducir los grupos de uno en uno (unión simple).

Los diferentes algoritmos que se obtienen con esta técnica sólo varían en la forma de seleccionar las distancias (paso iii).

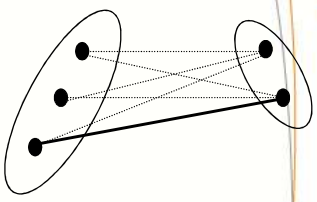
En general existen tres formas de seleccionar las distancias:

Distancia mínima min( $\|\vec{x}_k - \vec{x}_t\|$ Conjunto $r$



Conjunto $\kappa$

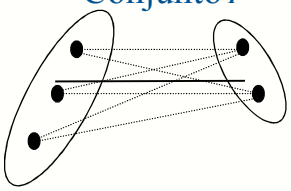
Distancia mínima max( $\|\vec{x}_k - \vec{x}_l\|$ Conjunto $T$



Conjunto $K$

Distancia promedio $\bar{d}_{kt} = \frac{1}{car(K)car(T)} \sum_{\substack{x_k \in K \\ x_t \in T}} \|\vec{x}_k - \vec{x}_t\|$

Conjunto $T$



Conjunto $K$

Estos algoritmos son denominados “algoritmos aglomerativos”. Puesto que usan una serie de uniones entre los vecinos próximos, comenzando desde ***n*** grupos hasta terminar formando sólo un grupo de los ***n*** individuos.

## 3.3. Algoritmo de la *k* medias

Los métodos no jerárquicos se caracterizan porque el número de grupos es determinado previamente e ingresado como parámetro al sistema de clasificación (ingresado por el usuario).

Las ideas centrales son:

- Escoger una partición inicial de los datos y luego alterar los miembros de los grupos para obtener nuevas (mejores) agrupaciones.
- Escoger una partición y asignar a la partición un representante llamado *centroide* de la misma.

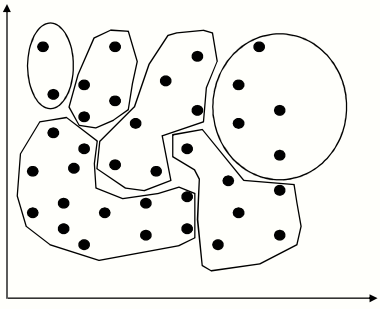
Básicamente existen dos formas para escoger el centroide:

- Determinar el promedio entre los integrantes de cada grupo y el dato que más se asemeje a él, exponerlo como centroide del grupo.
- Calcular la media entre los integrantes de cada grupo y este valor exponerlo como centroide.

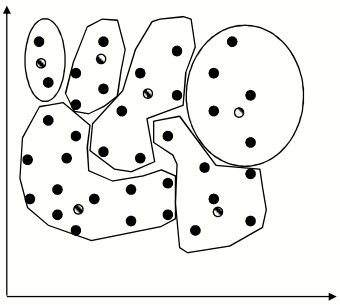
La elección adecuada dependerá del problema al cual se enfrente y del algoritmo que se escoja para efectuar el agrupamiento.

El creador del método MacQueen (1967) usa el término de *k-medias* para denotar a un proceso que forma ***k*** grupos usando distancias mínimas entre los ***n*** datos de entrada y los centroides o medias de cada grupo.

i. Formar ***k*** agrupaciones con los datos de entrada en forma aleatoria.



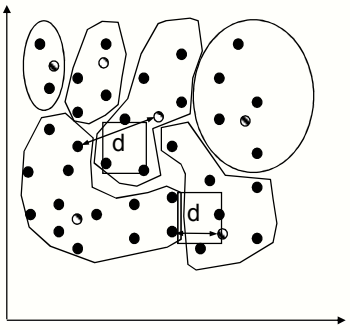
ii. Determinar el centroide $c_k$ de cada grupo calculando la media entre los integrantes del grupo .



iii. Calcular la distancia entre un patrón ***xi*** y los centroides $c_k$ , $d_{ik}$ para $k=1,2,...K$.

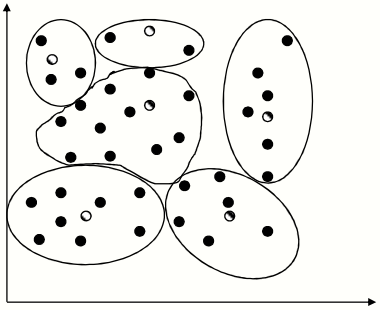
iv. Asignar el patrón $x_i$ al grupo cuya distancia al centroide fue menor. Esto es:

$$k_0 = \arg\left( \min_{k=1...K} (d_{ik}) \right)$$



v. Después de cada asignación, recalcular el centroide del grupo al cual se le adhirió un nuevo integrante o perdió uno.

vi. Repetir los pasos iii) al v) hasta que no ocurra ningún cambio en el sentido de que no emigren datos a otros grupos o no se muevan los centroides.



El esfuerzo total desde la configuración inicial hasta la agrupación final, está dado por los siguientes valores:

- ***k(2n – k)*** iteraciones.
- ***(k –1)(2n – k)*** comparaciones.
- ***n – k*** actualizaciones de los centroides.

Los estudios de la influencia del ordenamiento inicial de los datos muestran que este orden tiene un efecto pequeño al tratarse de agrupaciones muy separadas.

*Usando conjuntos de prueba, los autores muestran que el orden en los datos de entrada no altera más que un 0.07% el agrupamiento entre un tipo de ordenamiento inicial y otro.*

*El método converge a un mínimo local.*

**MacQueen propuso una variación al algoritmo de las k-medias, basado en la adaptación del número de grupos desde la asignación inicial hasta los resultados finales.**

*Usa como punto de partida la agrupación realizada por el algoritmo de las k-medias.*

*Adicionalmente al parámetro k usa los parámetros de Cohesión Co y de Refinamiento Re.*

## 3.4. Algoritmo de las *k* Medias adaptivo

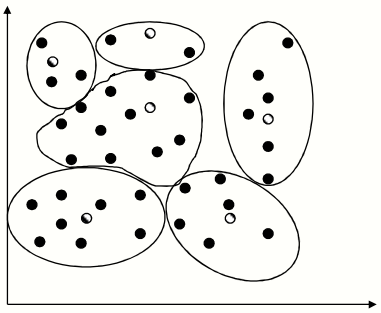
*Sea* $\vec{c}_i$ *el centroide del grupo i .* $Co = \min_{\forall i,j} \|\vec{c}_i - \vec{c}_j\|$

*El parámetro de cohesión será:* $Re = \bar{d}_{ij}$

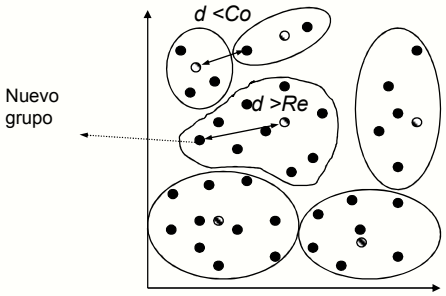
*El parámetro de Refinamiento será:*
*(promedio entre centroides)*

*Algoritmo*

***El primer paso corresponde al último paso de las k-medias.***

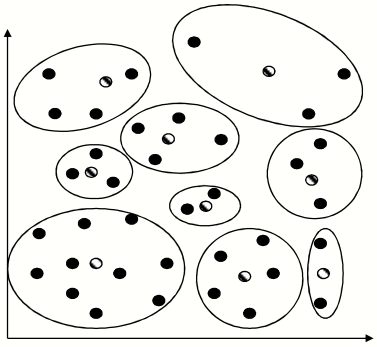


i. *Se recalculan las distancias de los centroides a cada sujeto (**pt**).*
   - *Si **d(ci,pt)<Co** ⇒ se juntan al grupos, asignar los restantes patrones a los grupo más cercanos.*
   - *Si **d(ci,pl)≥Re** ⇒ se genera un nuevo grupo, con un solo patron inicialmente*




ii. *Después de reasignar patrones se recalculan los centroides de cada uno de los grupos.*

iii. *Repetir los pasos I y II hasta que no existan más modificaciones*



iv. Al permitir que los grupos con centroides cercanos se unan, el método evita crear distinciones que dividan artificialmente a los grupos.

v. Mediante la creación de nuevos grupos es posible observar una mejor distribución de los grupos, puesto que puntos alejados no son forzados a pertenecer a algún grupo establecido.

vi. Al utilizar esta modificación al algoritmo de las *k-medias puro* se elimina la restricción de obtener *k* grupos y la cantidad de grupos en general aumenta, pero las distribuciones son mejores.

vii. Al ejecutar este algoritmo con distintos valores para los parámetros ***C*** y ***R***, se puede observar que la relación de los parámetros R y C es importante y se logran mejores agrupaciones cuando ***R*** > ***C.***

## 3.5. Evaluación de los métodos de agrupamientos

***En general los métodos no-jerárquicos se pueden evaluar usando algún criterio de optimalidad (minimización) para las estructuras de grupos resultantes.***

***La forma más común es adoptar una suma de las varianzas ponderadas.***

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij}\left\|\vec{x}_j - \vec{c}_i\right\|^2 = \sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij} d_{ij}^2$$

Donde:

- *k* número de grupos.
- *n* número de patrones.
- *uij* son los elementos de la matriz de particiones ***U*** que almacena los resultados de los agrupamientos al interior de los grupos.
- $\vec{c}_j$ centroide del grupo *j*.

***Los elementos de esta matriz satisfacen las siguientes condiciones:***

- $u_{ij} \in \{0,1\}$
- $o < \sum_{j=1}^{n} u_{ij} < n$ ***para i=1,…,k***
- $\sum_{i=1}^{k} u_{ij} = 1$ ***para j=1,…,n***

***Existen varios índices para evaluar la clasificación. Mediante estos índices es posible obtener el número k que se considera como un parámetro inicial en el método de las k medias puros.***

***También es posible usar estos métodos u otros similares para decidir el nivel de corte de los algoritmos aglomerativos.***

- ***Fukuyama y Sugeno, 1989***

$$Ind_{FS}(U) = \sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij}\left(\left\|\vec{x}_j - \vec{c}_i\right\|^2 - \left\|\vec{c}_i - \vec{c}\right\|^2\right)$$

***Con $\vec{c}$ el centroide de la totalidad de los datos.***

- ***Xie y Beni, 1991***

$$Ind_{XB}(U) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij}\left\|\vec{x}_j - \vec{c}_i\right\|^2}{n \ min_{i,j}\left\|\vec{c}_i - \vec{c}_j\right\|}$$

Existe una infinidad de variaciones de estos métodos y sistemas híbridos que incluyen clasificación Bayesiana, conjuntos difusos, conjuntos rugosos y otros.

Una aplicación muy utilizada actualmente es la introducción de una variable de contexto que es asociada a cada patrón.

Con esto es posible realizar el agrupamiento en función de las variables de contexto. En este caso el proceso de minería basado en el agrupamiento actúa como un filtro al focalizarse en un conjunto específico de datos.

También es posible relacionar los patrones con características semánticas para realizar búsquedas inteligentes en texto.