

### Preliminares: Distancias y medidas de Proximidad

Nos interesa establecer alguna medida de la proximidad entre diferentes tipos de observaciones multivariadas.

Es una distancia que aplica a dos vectores, estos vectores, que pueden ser dos observaciones distintas de una variable multivariada o bien, dos filas, de una matriz de datos.

Vamos a extender nuestro concepto de distancia, proximidad o similaridad.

#### Concepto de distancia:

En un conjunto de elementos  $X$  se define distancia o métrica como una función matemática aplicada a un par de elementos de  $X$  que satisface simultáneamente las siguientes condiciones:

$$d: X \times X \rightarrow R$$

- No negatividad:  $d(a, b) \geq 0 \quad \forall (a, b) \in X \times X$
- Simetría:  $d(a, b) = d(b, a) \quad \forall a, b \in X \times X$
- Desigualdad triangular:  $d(a, c) \leq d(a, b) + d(b, c) \quad \forall a, b, c \in X$
- $\forall a \in X \quad d(a, a) = 0$
- Si  $a$  y  $b$  son tales que  $d(a, b) = 0$ , entonces  $a=b$ .

Un espacio métrico es un espacio vectorial  $X$  en el cual se ha definido una función que satisface las condiciones de distancia.  $(X, d)$

Vamos a analizar diferentes tipos de observaciones.

## Primero: Vectores de observaciones continuas

### A. Medidas de Disimilaridad

Consideremos que tenemos dos vectores  $x$  e  $y$  de dimensión  $k$  que contienen valores de variables continuas.

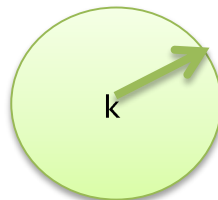
#### a) Distancia Minkowski o norma $L_p$

$$D_p(\mathbf{x}, \mathbf{y}) = \left( \sum_{i=1}^k (x_i - y_i)^p \right)^{1/p}$$

Para el caso particular de  $p=2$ , tenemos la distancia euclídea.

$$D_2 = \sqrt{\sum_{i=1}^k (x_i - y_i)^2}$$

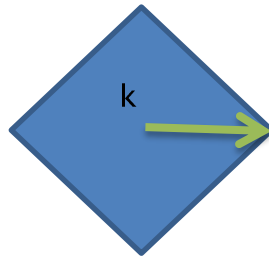
En este caso, todos los puntos que están a distancia  $k$  de un punto dado, conforman una circunferencia de radio  $k > 0$ .



#### b) Distancia Manhattan o City-Block $p=1$ :

$$D_1 = \sum_{i=1}^M |x_i - y_i|$$

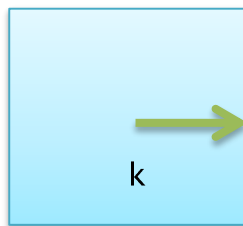
En este caso, todos los puntos que están a distancia  $k$  de un punto dado, conforman un rombo o cuadrado de diagonal  $2k$ .



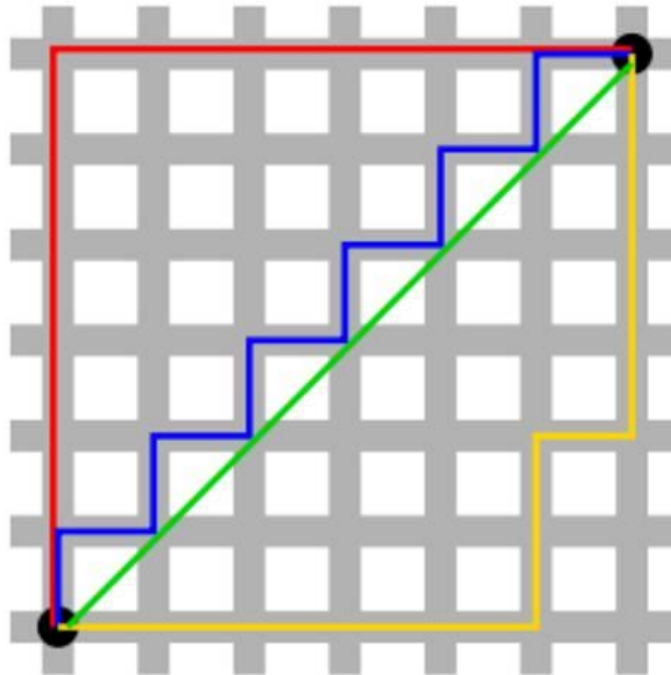
c) **Distancia Chebychev**,  $p=\infty$

$$d_{\infty} = \max_{1 \leq i \leq M} |x_i - y_i|$$

En este caso, todos los puntos que están a distancia  $k$  de un punto dado, conforman un rombo o cuadrado de lado  $2k$ .



Plano de Manhattan. La distancia euclidiana (segmento verde), no se corresponde con el «camino más corto» ente dos puntos de dicha ciudad, además de no ser único.



### Ejemplo 1

Consideremos las siguientes dos observaciones correspondientes a tres corredores para los cuales se han registrado los tiempos empleados para correr cada uno de los cuatro tramos de 4km en que se subdividió la carrera:

Corredor	x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	x <sub>4</sub>
<b>A</b>	10	10	13	12
<b>B</b>	12	12	14	15
<b>C</b>	11	10	14	13

Apliquemos las distancias definidas entre cada par de corredores:

#### **Distancias euclídeas**

$$d_2(A,B) = \sqrt{(10-12)^2 + (10-12)^2 + (13-14)^2 + (12-15)^2} = \sqrt{18}$$

$$d_2(A,C) = \sqrt{(10-11)^2 + (10-10)^2 + (13-14)^2 + (12-13)^2} = \sqrt{3}$$

$$d_2(B,C) = \sqrt{(12-11)^2 + (12-10)^2 + (14-14)^2 + (15-13)^2} = \sqrt{9} = 3$$

### Distancias de Manhattan

$$d_1(A,B) = |10-12| + |10-12| + |13-14| + |12-15| = 8$$

$$d_1(A,C) = |10-11| + |10-10| + |13-14| + |12-13| = 3$$

$$d_1(B,C) = |12-11| + |12-10| + |14-14| + |15-13| = 5$$

### Distancias de Chebychev

$$d_1(A,B) = \max\{|10-12|, |10-12|, |13-14|, |12-15|\} = 3$$

$$d_1(A,C) = \max\{|10-11|, |10-10|, |13-14|, |12-13|\} = 1$$

$$d_1(B,C) = \max\{|12-11|, |12-10|, |14-14|, |15-13|\} = 2$$

Algunas versiones más generalizadas incluyen pesos  $w_i$  para cada variable  $i=1,\dots,k$ .

e) De esta manera la distancia ponderada Minkowski, será:

$$D_p(\mathbf{x}, \mathbf{y}) = \left( \sum_{i=1}^M w_i (x_i - y_i)^p \right)^{1/p}$$

f) **Distancias cuadráticas.** Si consideramos que  $Q=(Q_{ij})$  es una matriz cuadrada  $k \times k$  definida positiva de pesos entonces la distancia cuadrática entre  $\mathbf{x}$  y  $\mathbf{y}$  está dada por:

$$D_Q(\mathbf{x}, \mathbf{y}) = [(\mathbf{x} - \mathbf{y})' Q (\mathbf{x} - \mathbf{y})]^{1/2} = \sqrt{\sum_{i=1}^k \sum_{j=1}^k (x_i - y_i) Q_{ij} (x_j - y_j)}$$

Un caso particular de esta distancia es cuando  $Q=V^{-1}$ , siendo  $V$  la matriz de covarianza entre  $x$  y  $y$ . En este caso la distancia coincide con la **distancia Mahalanobis**.

**g) Distancia Camberra:**

$$D_{\text{Can}}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^M \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

Cuando ocurra que simultáneamente  $x_i$  y  $y_i$  sean ambos ceros entonces el  $i$ -ésimo término de la suma se considera como cero.

h) **Distancia de Bray Curtis** (1957) se aplica en diversos campos del conocimiento pero fundamentalmente en biología, ecología y ciencias del ambiente.

$$d^{\text{BCD}}(i, j) = \frac{\sum_{k=0}^{n-1} |y_{i,k} - y_{j,k}|}{\sum_{k=0}^{n-1} (y_{i,k} + y_{j,k})}$$

$D^{\text{BCD}}$  (Bray Curtis dissimilarity)  $D^{\text{BCS}}$  (Bray Curtis simmilarity) están vinculados de la siguiente manera:  $D^{\text{BCD}} + D^{\text{BCD}} = 1$

## B. Medidas de similitud

**h) Medida de correlación:** Llamada también medida de separación angular o de producto interno normalizado

$$s(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^M (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^M (x_i - \bar{x})^2 \sum_{i=1}^M (y_i - \bar{y})^2}} = \frac{(\mathbf{x} - \bar{\mathbf{x}})'(\mathbf{y} - \bar{\mathbf{y}})}{\|\mathbf{x} - \bar{\mathbf{x}}\| \|\mathbf{y} - \bar{\mathbf{y}}\|}$$

i) **Medida de Tanimoto.** Se define por

$$s_T(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}'\mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x}'\mathbf{y}}$$

Puede ser usada también para datos nominales.

**La medida de Tanimoto.** Sean  $X$  y  $Y$  dos conjuntos de cardinalidad  $n_x$  y  $n_y$  respectivamente. Sea  $n_{X \cap Y}$  que representa la cardinalidad de  $X \cap Y$ , entonces la medida de Tanimoto se define por

$$d_T = \frac{n_{X \cap Y}}{n_X + n_Y - n_{X \cap Y}}$$

Es decir la medida de Tanimoto es la razón del número de elementos que los vectores tienen en común entre el número de elementos distintos.

## Segundo: Vectores de variables categóricas o nominales

Consideremos ahora vectores de dimensión  $k$  cuyas coordenadas asumen valores en el conjunto finito  $\Omega = \{0, 1, 2, \dots, m-1\}$ , donde  $m$  es un entero positivo

## B. Medidas de similitud

Constituyen una extensión de las distancias. En general satisfacen todas las condiciones de distancia excepto la desigualdad triangular.

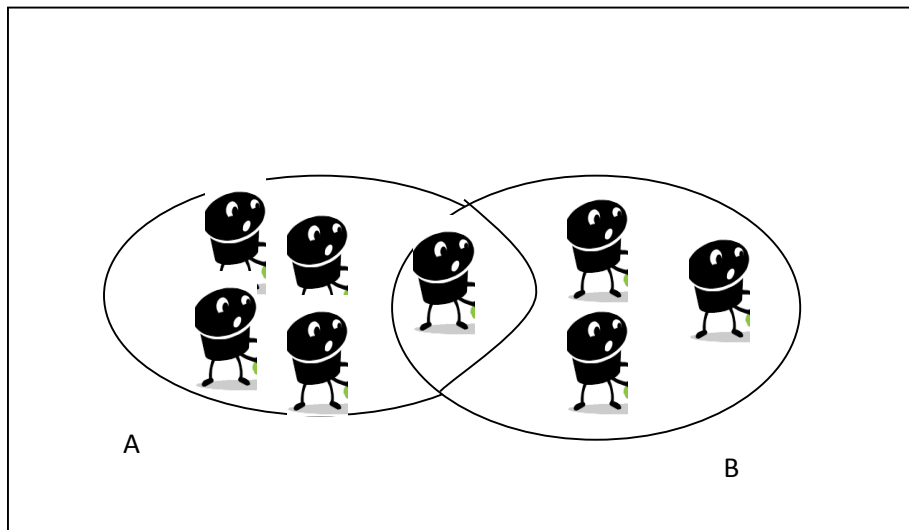
j) **Medida de correlación de Pearson:** Llamada también medida de separación angular o de producto interno normalizado

$$s(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^M (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^M (x_i - \bar{x})^2 \sum_{i=1}^M (y_i - \bar{y})^2}} = \frac{(\mathbf{x} - \bar{\mathbf{x}})'(\mathbf{y} - \bar{\mathbf{y}})}{\|\mathbf{x} - \bar{\mathbf{x}}\| \|\mathbf{y} - \bar{\mathbf{y}}\|}$$

Si este coeficiente se calcula en lugar de sobre las observaciones, sobre sus rangos, recibe el nombre de **coeficiente de correlación de Spearman**.

### Ejemplo 2:

$$d_r(A, B) = 1 / (5 + 4 - 1) = 1/8$$



En el caso particular de variables binarias. Las entradas de los dos vectores se pueden resumir en la siguiente tabla

		Y	
		0	1
X	0	a	b



	1	c	d
--	---	---	---

Donde **a** representa el número de posiciones donde los vectores X y Y coinciden en tomar el valor 0. Similarmente **d** representa el número de coincidencias donde X y Y valen ambos 1. Mientras que **c** y **d** representan el número de no coincidencias.

La medida de Tanimoto se reduce a

$$d_T = \frac{a + d}{a + 2(c + b) + d}$$

Se pueden considerar además las siguientes medidas

ii) **El coeficientes de coincidencias simple.** Definido por:

$$DCS = \frac{a + d}{a + b + c + d}$$

### Ejemplo 3

A y B son dos sujetos distintos a los que se les ha observado presencia o ausencia de hipertensión arterial, obesidad, acv y dbt.

	hta	obesidad	acv	dbt
A	1	1	1	0
B	1	1	0	1

Queremos calcular la similitud entre dos pacientes en cuyas historias clínicas se han registrado los siguientes datos:

	A			
		0	1	

B	0	0(a)	1(b)	1
	1	1(c)	2(d)	3
		1	3	4

$$dT(A,B) = \frac{a+d}{a+2(c+b)+d} = 2/6 = 1/3$$

$$DCS(A,B) = \frac{a+d}{a+b+c+d} = 2/4 = 1/2$$

Ahora que tenemos un concepto más amplio de medidas de distancia o similaridad, introduzcamos una nueva técnica de análisis multivariado.

k) **Coefficiente de Jaccard**: indica la proporción de factores en los cuales dos individuos

no coinciden, ignorando el (0,0)  $D_{jacc} = \frac{b+c}{b+c+d}$

l) **Coefficiente Phi** es la versión del coeficiente de correlación de Pearson para variables

binarias:  $PHI(X, Y) = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$

Este coeficiente toma valores entre 0 y 1.

m) **Coefficiente de Anderberg** evalúa la capacidad predictiva de una variable sobre la otra.

$$And(X, Y) = \frac{t_1+t_2}{2(a+b+c+d)}$$

donde:

$$t_1 = \max(a, b) + \max(a, c) + \max(b, d) + \max(c, d)$$

$$y \ t_2 = \max(a + c, b + d) + \max(a + b, c + d)$$

n) **Coeficiente de Dice**( también conocido como Sorensen o Czekanowsky):

$$\text{Dice}(X, Y) = \frac{2a}{b + c + 2a}$$

o) **Coeficiente de Ochiai**:

$$\text{Och}(X, Y) = \frac{a}{\sqrt{(a + b)(a + c)}}$$

## INTRODUCCION AL ANALISIS DE CLUSTER (Conglomerados)

### Objetivos:

- **Obtener una representación “compacta” de los datos, para:**
- **Generar una clasificación**
- **Describir los datos**

Dado un conjunto de  $n$  objetos (animales, plantas, minerales...), cada uno de los cuales viene descrito por un conjunto de  $p$  características o variables, deducir una división útil en un número de clases. *Se han de determinar tanto el número de clases como las propiedades de dichas clases.*

El Análisis de Clusters o de conglomerados es una **técnica de Análisis Exploratorio** de Datos para resolver problemas de clasificación. Su objeto consiste en ordenar objetos u observaciones: personas, cosas, animales, plantas, variables, etc, ... en grupos o conglomerados o clusters, de forma tal que **el grado de asociación/similitud entre miembros del mismo cluster sea más fuerte que el grado de asociación/similitud entre miembros de diferentes clusters**. Cada cluster se describe como la clase a la que sus miembros pertenecen.

El análisis de cluster es un método que ***permite descubrir asociaciones y estructuras en los datos que no son evidentes a priori pero que pueden ser útiles una vez que se han encontrado***. Los resultados de un Análisis de Clusters pueden contribuir a la definición formal de un esquema de clasificación tal como una taxonomía para un conjunto de objetos, a sugerir modelos estadísticos para describir poblaciones, a asignar nuevos individuos a las clases para diagnóstico e identificación, etc ...

Si por ejemplo, un investigador ha recogido datos en un cuestionario puede enfrentarse a un número elevado de observaciones que no tendra sentido a menos que clasifique en grupos manejables.

El Análisis Cluster es una técnica de reducción de datos mediante la reducción de la población en subgrupos más manejables. Se aplica en psicología para la clasificación o descripción de tipologías personales como también en la segmentación del mercado.

Como muchos procedimientos multivariados, también conlleva inicialmente a una pérdida de información. Se recurre a técnicas de agrupamiento cuando **no se conoce una estructura de agrupamiento de los datos “a priori”** y el objetivo operacional es *identificar el agrupamiento natural de las observaciones.*

Las técnicas de clasificación basadas en agrupamientos implican la distribución de las unidades de estudio en clases o categorías de manera tal que cada clase (o conglomerado) reúne unidades cuya similitud es máxima bajo algún criterio.

Es decir los objetos en un mismo grupo comparten el mayor número permisible de características y los objetos en diferentes grupos tienden a ser distintos.

### **Análisis de cluster por individuos o variables**

Generalmente lo que se pretende agrupar son individuos, pero existen algunas circunstancias en las que es interesante agrupar variables para intentar buscar variables de comportamiento similar. Para ello, ***la metodología es la misma que para el análisis cluster por individuos y simplemente tendremos que transponer la matriz de datos y aplicar el método general.***

Para agrupar objetos (casos o variables) es necesario seguir algún **algoritmo**. Los algoritmos o métodos de agrupamiento permiten identificar clases existentes en relación a un conjunto dado de atributos o características.

El agrupamiento logrado dependerá de:

- ♥ el algoritmo utilizado
- ♥ la distancia seleccionada
- ♥ De la cantidad de grupos deseados, si existe esta información.
- ♥ De las variables recogidas en la base

- ♥ De si las variables han sido estandarizadas o no.

Existe una medida que define de alguna forma la complejidad del problema:

$$P(N, K) = \frac{1}{K!} \sum_{m=1}^K (-1)^{K-m} C_K^m m^N$$

Donde N es la Cantidad de objetos y K la cantidad de clusters.

Para K=3 y N=30  $P(N,K) = 2 * 10^{14}$

Los algoritmos de clasificación pueden dividirse en no jerárquicos y jerárquicos.

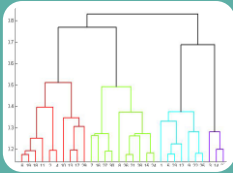
- Dentro de los Métodos jerárquicos hay a su vez dos tipos:
  - Ascendentes o Aglomerativos
  - Descendentes o de Difusión
- Entre los Métodos no jerárquicos o de partición el mas conocido es:
  - K medias
- También existen los llamados Métodos mixtos

Existen diferentes alternativas para la Validación de los clusters:

- **Criterios externos**: Comparan la clusterización con alguna segmentación previa de referencia.
- **Criterios internos**: Analizan la significatividad de los clusters solo considerando los datos usados en la clusterización.
- **Criterios relativos**: Comparan la clusterización con otras resultantes de segmentaciones alternativas

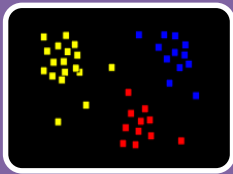


## Metodos de AGrupamiento



### Jerárquicos

- pretenden encontrar particiones jerarquizadas
- consecutivamente más finas (o menos finas), luego los objetos son unidos (o separados) en grupos paso por paso.



### No Jerárquicos

- se desea obtener una única **descomposición o partición del conjunto original** de objetos.
- en base a la optimización de una función objetivo

### Ventajas de los métodos de clasificación jerárquicos

- Sugiere el número de clusters.
- Establece una jerarquía de clusters.
- El dendograma permite la visualización del proceso y la solución final del problema planteado.

### Desventajas de los métodos de clasificación jerárquicos

- Resulta costoso en grandes bases de datos.
- Es lento

Muchas veces, informaciones preliminares disponibles o resultados de experimentos pilotos, pueden orientar al experimentador o usuario en la selección del **número de clases**. Otras veces, se conoce algún valor máximo para el número de clases, y entonces el algoritmo se implementa especificando dicho valor y luego, en relación con los resultados obtenidos, se vuelven a realizar agrupamientos. Las técnicas de clasificación jerárquicas son generalmente del tipo **no supervisadas**.

En el análisis de conglomerados de casos o registros individuales se parte de una matriz de datos  $n \times p$  (supongamos  $p$  mediciones o variables en cada uno de los  $n$  objetos estudiados), que luego es transformada en una matriz de distancia ( $n \times n$ ) donde el elemento  $i,j$ -ésimo mide la distancia entre pares de objetos  $i$  y  $j$  para  $i,j=1,\dots,n$ .

Cuando se disponen de numerosas variables para realizar el agrupamiento, es común utilizar - antes del análisis de conglomerados- técnicas de reducción de dimensión tal como **Análisis de Componentes Principales** para obtener un número menor de variables capaces de expresar la variabilidad en los datos. Esta técnica puede facilitar la interpretación de los agrupamientos obtenidos.

En la práctica, se recomienda aplicar varios algoritmos de agrupamiento y de selección o combinación de variables para cada conjunto de datos. Seleccionando, finalmente, desde los agrupamientos realizados la interpretación más apropiada.

Para comparar varios agrupamientos alternativos, suele utilizarse el **coeficiente de correlación cofenética**. Este coeficiente indica la correlación de las distancias definidas por la métrica de árbol binario con las distancias originales entre objetos, luego se espera que el agrupamiento con mayor coeficiente sea el que mejor describe el agrupamiento natural de los datos.

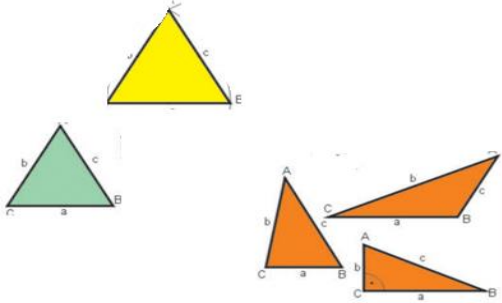

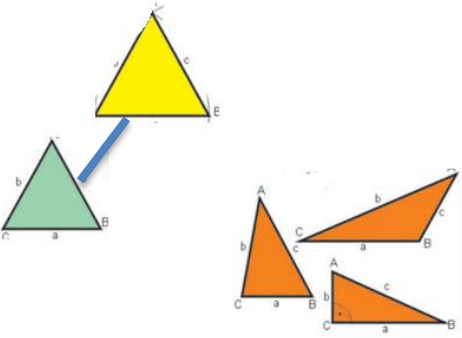

Es importante remarcar que los procedimientos de agrupamiento producen resultados exitosos cuando la matriz de datos tiene una estructura que es posible interpretar desde el problema que originó la recolección de la información. Por ello, logrados los grupos es importante caracterizar los mismos a través de diversas medidas resumen para favorecer la interpretación del agrupamiento final.

Podemos encontrarnos dos tipos fundamentales de métodos de clasificación:

<b>Jerárquicos y No Jerárquicos.</b>
--------------------------------------

En los primeros, la clasificación resultante tiene un número creciente de clases anidadas mientras que en el segundo las clases no son anidadas.



<b>Aglomerativos</b>	<b>Divisivos</b>
Se parte de tantas clases como objetos tengamos que clasificar	Se parte de una única clase formada por todos los objetos
En pasos sucesivos vamos obteniendo clases de objetos similares	En pasos sucesivos va dividiendo en clases sucesivamente.
	
	

### Métodos de agrupamiento jerárquicos

Los métodos jerárquicos producen agrupamientos de tal manera que un conglomerado puede estar contenido completamente dentro de otro, pero no está permitido otro tipo de superposición entre ellos.

Los resultados de agrupamientos jerárquicos se muestran en un dendrograma (diagramas de árboles en dos dimensiones), en el que se pueden observar las uniones y/o divisiones que se van realizando en cada nivel del proceso de construcción de conglomerados.

Las ramas en el árbol representan los conglomerados. Las ramas se unen en un nodo cuya posición a lo largo del eje de distancias indica el nivel en el cual la fusión ocurre. El nodo donde todas las entidades forman un único conglomerado, se denomina *nodo raíz*.

Una de las principales características de los procedimientos de agrupamiento jerárquicos aglomerativos es que la ubicación de un objeto en un grupo no cambia, o sea, que una vez que un objeto se ubicó en un conglomerado, no se lo reubica.

Este objeto puede ser fusionado con otros pertenecientes a algún otro conglomerado, para formar un tercero que incluye a ambos.

Los procedimientos jerárquicos descritos anteriormente no realizan ninguna acción diferencial con observaciones aberrantes. Si una observación rara fue clasificada en etapas tempranas del procedimiento en algún grupo, esta permanecerá ahí en la configuración final.

Algunos experimentadores, usan la técnica de la perturbación (introducción de errores en los datos y reagrupamiento bajo la nueva situación) para probar la estabilidad de la clasificación jerárquica.

### Pasos de la clasificación jerárquica:

Los pasos de la clasificación jerárquica son los siguientes:

<b>1.- Decidir qué datos tomamos para cada uno de los casos.</b>
------------------------------------------------------------------

En primer lugar tendremos que estudiar el tipo de variables con la que trabajar. Generalmente tomaremos varias variables todas del mismo tipo, es decir todas continuas, o todas categóricas, etc.); suele ser difícil mezclar tipos distintos.

Evidentemente, sobre cualquier individuo es posible encontrar un gran número de variables, pero esto no siempre es útil, ya que la inclusión de variables irrelevantes no puede ser contrastada por el análisis de cluster y además aumenta la posibilidad de

errores en la conclusión final. Por ello **se deben de eliminar las variables irrelevantes en base al objetivo de la investigación.**

También es interesante el **tipificar las variables**. Si las variables están medidas en diferentes unidades o escalas, la comparación entre unas variables u otras será difícil. Por ello se suelen tipificar los datos, de tal manera que obtengamos que todas las variables tengan media 0 y desviación típica 1 y además que no existan unidades entre los valores

**2.- Elegimos una medida de la distancia entre los objetos a clasificar, es decir entre los clusters o clases iniciales.**

De tal manera que existirán muchos tipos diferentes de distancias y similitudes y dependiendo de cada circunstancia se elegirán una u otra.

**3.- Buscamos que clusters son más similares.**

Utilizando para ello la minimización de la distancia seleccionada y el método de distancia al cluster elegido. Una vez unidos dos objetos, no se separarán durante el resto del proceso.

**4.- Juntamos estos dos clusters en un nuevo cluster que tenga al menos 2 objetos, de forma que el número de clusters decrece en una unidad.**

**5.- Seleccionamos la técnica de cluster y calculamos la distancia entre este nuevo cluster y el resto.**

No es necesario recalcular todas las distancias, solamente las del nuevo cluster con los anteriores.

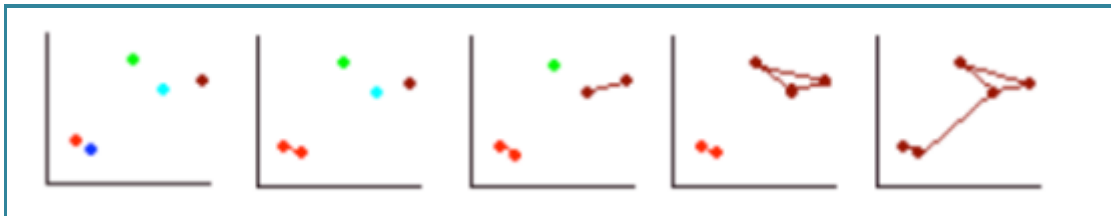
**6.- Repetimos desde el paso 3 hasta que todos los objetos estén en un único cluster.**

Luego seleccionamos la cantidad de clusters adecuada a la respuesta a nuestro problema, es decir en que paso de la técnica nos detenemos.

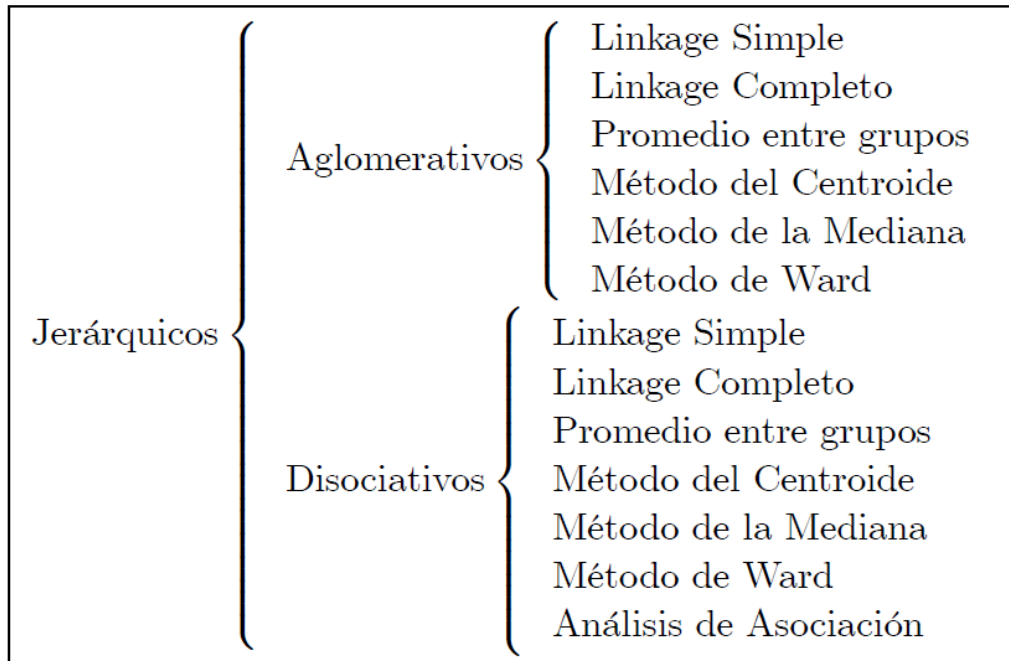
### 7.- Interpretación de los resultados

Una vez determinados los grupos, corresponderá al investigador de cada campo, psicológico, sociólogo, pedagogo..., investigar los grupos y el por qué de su formación y sacar las conclusiones relevantes de este, así como las características en las que se diferencian cada conglomerado.

Los pasos se resumen en el siguiente diagrama:



**Importante:** Los distintos métodos o algoritmos dependen del método utilizado en el paso 5 para calcular la distancia entre clusters. Es necesario resaltar, que los distintos métodos para el cálculo de las distancias entre clusters producen distintas clasificaciones, por lo que no existe una clasificación correcta única.



### EL DENDOGRAMA: se utiliza para representar la clasificación jerárquica

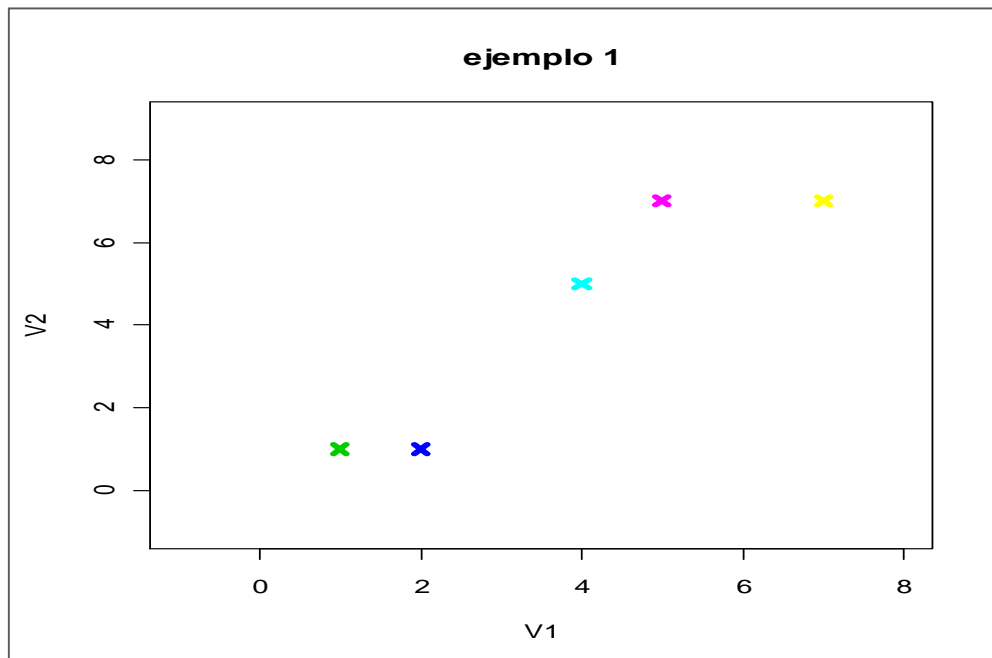
Un dendograma es una representación gráfica en forma de árbol que resume el proceso de agrupación en un análisis de clusters. Los objetos similares se conectan mediante enlaces cuya posición en el diagrama está determinada por el nivel de similitud/disimilitud entre los objetos.

Para entender la construcción de un dendograma y su significado utilizaremos un ejemplo sencillo que lo ilustre.

Consideremos un **primer ejemplo** sencillo con solo 5 objetos y dos variables:

objeto	V1	V2
1	1	1
2	2	1
3	4	5
4	7	7
5	5	7

Representemos los puntos en un diagrama euclídeo.



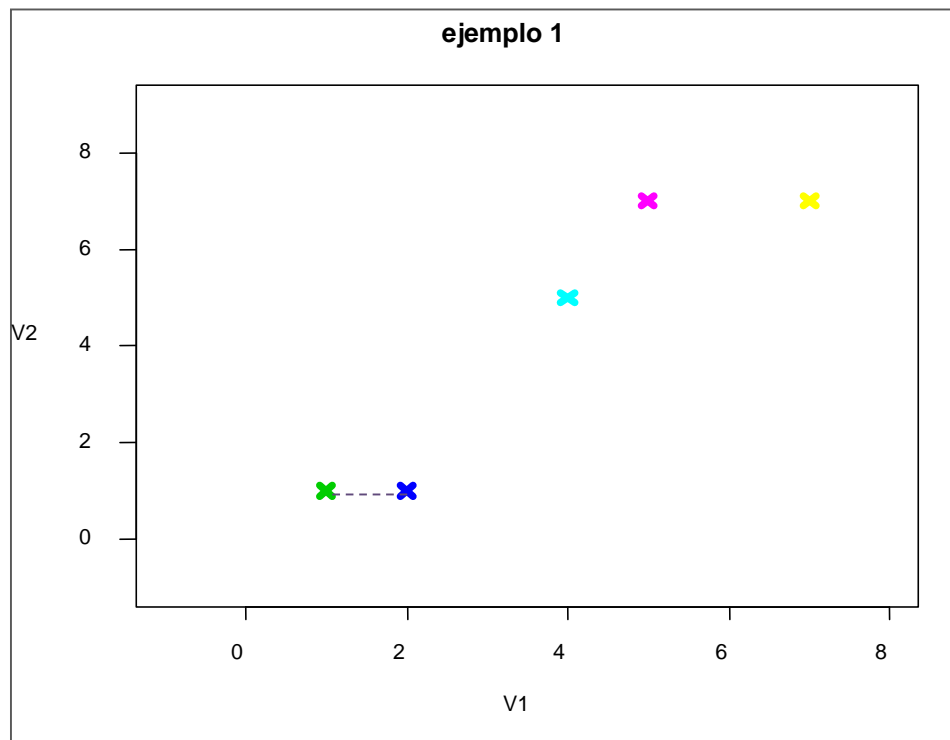
A partir de estos datos consideramos la matriz de distancias euclídeas entre los

	1	2	3	4
2	1,000			
3	5,000	4,472		
4	7,210	6,710	2,236	
5	8,490	7,810	3,610	2,000

Inicialmente tenemos 5 clusters, uno por cada uno de los objetos a clasificar.

De acuerdo con la matriz de distancias, los objetos (clusters) más similares son el 1 y el 2 (con distancia 1), por lo tanto los fusionamos construyendo un nuevo cluster que contiene los objetos 1 y 2. Llamaremos A al cluster.

El problema ahora es medir la distancia de este cluster al resto de los objetos/clusters.



Para este ejemplo lo que haremos será tomar como representante del grupo el centroide de los puntos que forman el cluster, es decir, el punto que tiene como coordenadas las medias de los valores de las variables para sus componentes, es decir, las coordenadas de A son

$$A = ((1+2)/2; (1+1)/2) = (1,5; 1).$$

Tendríamos entonces la siguiente tabla de datos cluster.

objeto	V1	V2
A	1,5	1
3	4	5
4	5	7
5	7	7

A partir de estas coordenadas calculamos la nueva matriz de distancias entre los clusters que tenemos en este momento.

	<b>A</b>	<b>3</b>	<b>4</b>
<b>3</b>	4,720		
<b>4</b>	6,950	2,230	
<b>5</b>	8,140	3,610	2,000

Ahora los clusters más similares son el 4 y el 5 (con distancia 2) que se deben fusionar en un nuevo cluster, al que llamaremos B.

El centroide de este nuevo cluster es el punto (6, 7).

La nueva tabla de datos es ahora:

<b>objeto</b>	<b>V1</b>	<b>V2</b>
<b>A</b>	1,5	1
<b>3</b>	4	5
<b>B</b>	6	7

Recalculando como antes la matriz de distancias euclídeas tenemos:

	<b>A</b>	<b>B</b>	<b>3</b>
<b>A</b>	0		
<b>B</b>	7,5	0	
<b>3</b>	4,7	2,8	0

La distancia más pequeña está ahora entre el cluster B y el 3 (distancia 2.8) que se fusionan en uno nuevo que denominamos C. Los valores medios (centroide) son ahora:  
 $v1=(4+5+7)/3=5,3$ ;  $v2=(5+7+7)/3=6,3$

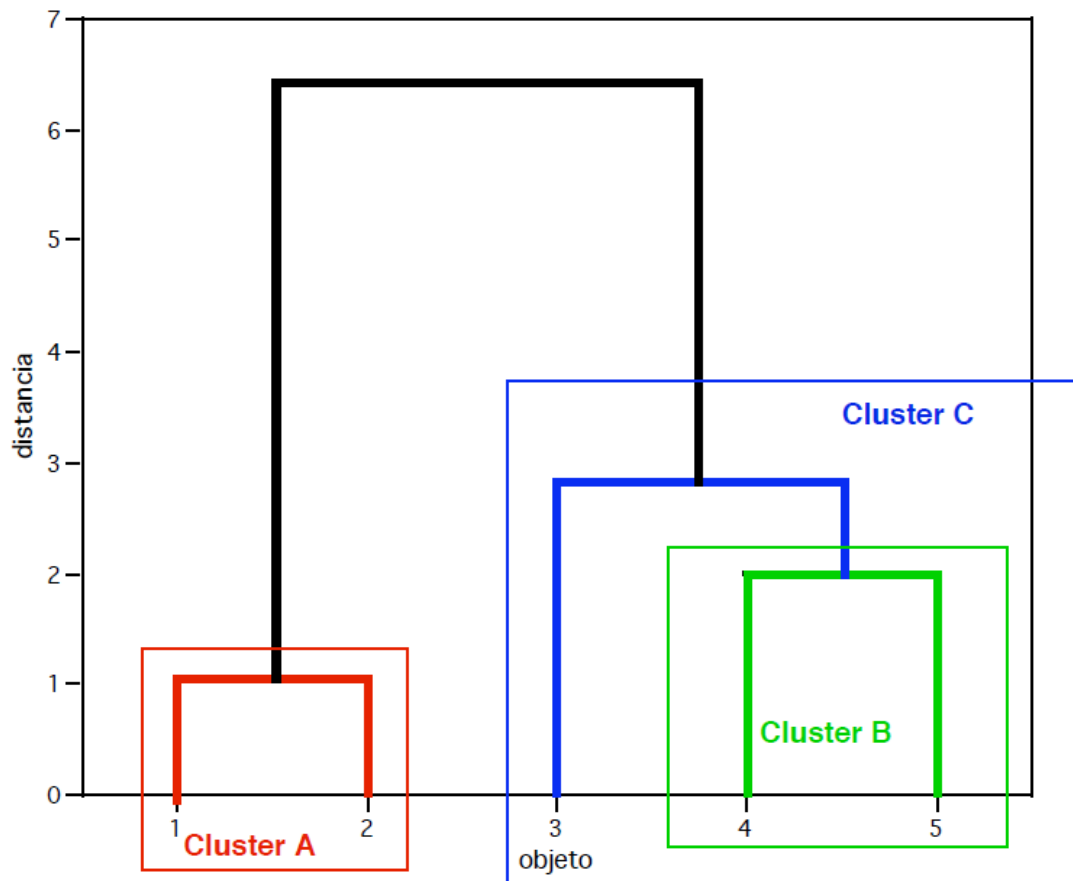
La tabla de datos es ahora:

	<b>V1</b>	<b>V2</b>
<b>A</b>	1,5	1
<b>C</b>	5,3	6,3

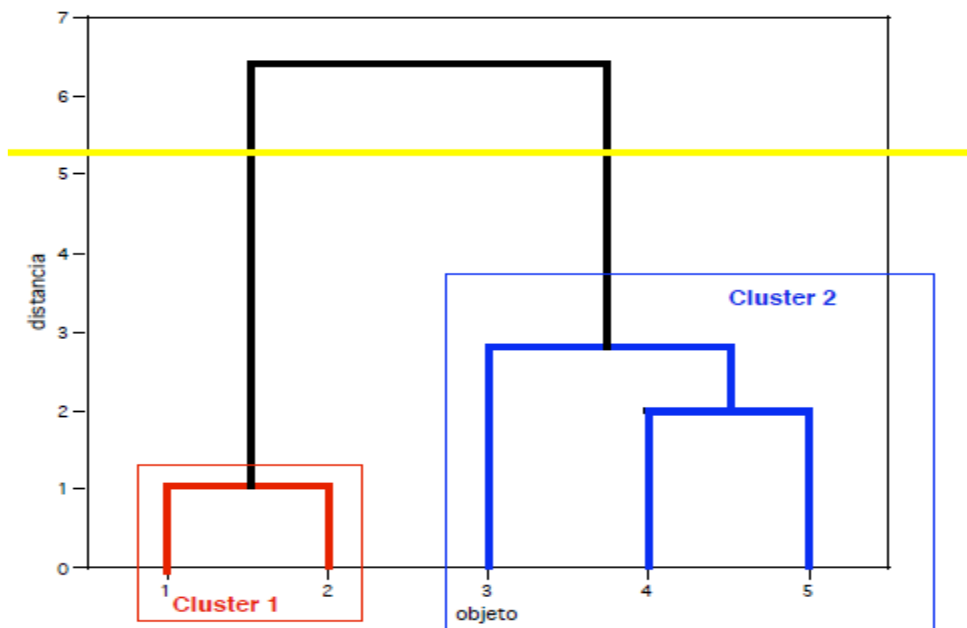


Y nos han quedado ahora solamente dos clusters con distancia 6.4 que se fusionarán en el paso siguiente terminando el proceso.

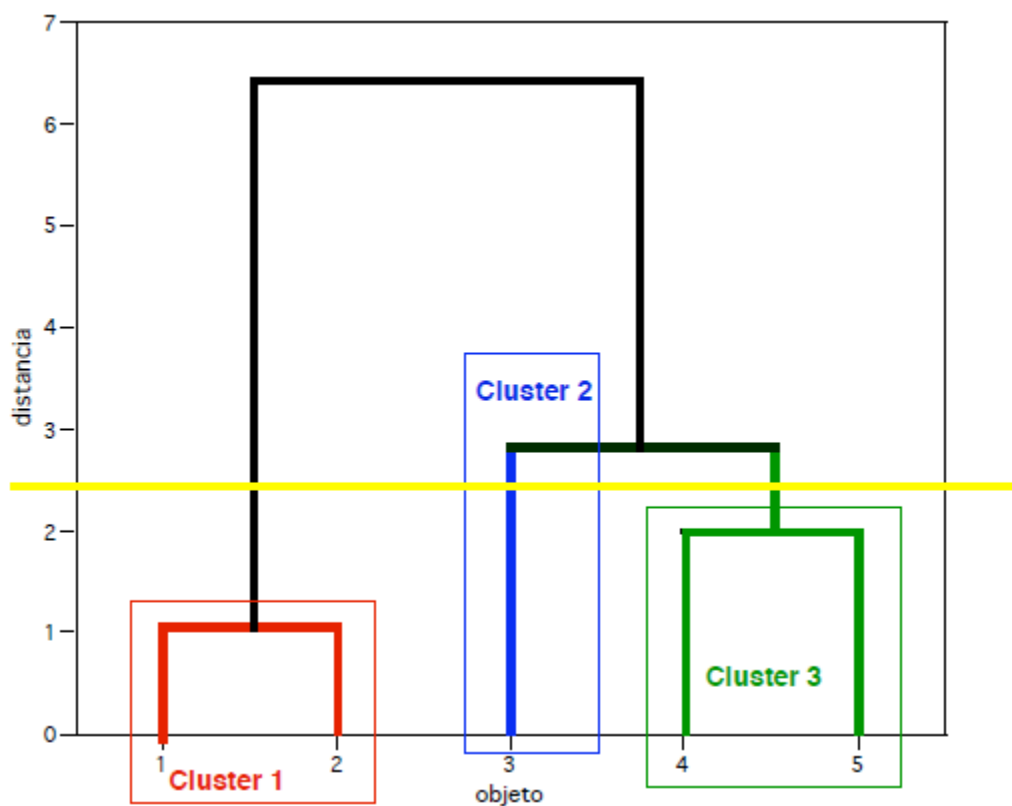
El proceso completo de fusiones puede resumirse mediante un dendograma.



En el gráfico parece evidente que tenemos 2 clusters, el que habíamos denominado A y el que habíamos denominado C. En general, si cortamos el dendograma mediante una línea horizontal como en el gráfico siguiente, determinamos el número de clusters en que dividimos el conjunto de objetos.



Cortando el dendrograma como en la figura anterior obtendríamos 2 clusters. Si lo cortamos como en la figura siguiente obtendríamos 3.



La decisión sobre el **número óptimo de clusters** es subjetiva, especialmente cuando se incrementa el número de objetos ya que si se seleccionan demasiado pocos, los clusters

resultantes son heterogéneos y artificiales, mientras que si se seleccionan demasiados, la interpretación de los mismos suele ser complicada.

Como ayuda a la decisión sobre el número de clusters se suelen representar los distintos pasos del algoritmo y la distancia a la que se produce la fusión. En los primeros pasos el salto en las distancias será pequeño, mientras que en los últimos el salto entre pasos será mayor. El punto de corte será aquel en el que comienzan a producirse saltos bruscos.

En el ejemplo, el salto brusco se produce entre los pasos 3 y 4, luego el punto óptimo es el 3, en el que había 2 clusters.

#### ALGORITMOS PARA EL ANALISIS DE CLUSTER:

##### DISTINTAS FORMAS DE MEDIR LA DISTANCIA ENTRE CLUSTERS

Como ya indicamos, existen diversas formas de medir la distancia entre clusters que producen diferentes agrupaciones y diferentes dendogramas. No hay un criterio para seleccionar cuál de los algoritmos es el mejor. La decisión es normalmente subjetiva y depende del método que mejor refleje los propósitos de cada estudio particular.

#### Veamos algunos métodos para medir distancias entre clusters:

##### MÉTODO DE LA MEDIA (AVERAGE LINKAGE)

En el método de la media, la distancia entre clusters se calcula como la distancia media (o promedio de las distancias) entre pares de observaciones, una de cada cluster.

Si P y Q son un cluster y R es otro cluster, para medir la distancia entre los dos clusters:

$$d(R, \{P,Q\}) = 0,5d(R,P) + 0,5d(R,Q)$$

Para el ejemplo anterior con matriz de distancias:

	1	2	3	4
2	1,000			
3	5,000	4,472		
4	7,210	6,710	2,236	
5	8,490	7,810	3,610	2,000

después de agrupar el 1 y el 2 en el cluster A, calculamos las distancias del cluster A a los puntos 3, 4 y 5.

La matriz de distancias es entonces:

distancias	Elem.1	Elem.2	promedio	Distancia al cluster formado por {1,2}
3	5,0	4,5	$(5+4,5)/2$	4.75
4	8,5	7,8	$(8,5+7,8)/2$	8,15
5	7,2	6,7	$(7,2+6,7)/2$	6,95

La nueva matriz de distancias es entonces:

	A	3	4
3	4,75		
4	8,15	2,23	
5	6,95	3,61	2,0

De nuevo, la distancia más pequeña es entre 4 y 5, por lo que los fusionamos en un cluster que denominamos B.

Calculamos la distancia entre B y el resto, es decir, con A y con 3.

Entre A y B, buscamos las distancias entre todos los pares de puntos y calculamos la media

		B	
		4	5
A	1	8,5	7,2
	2	7,8	6,7

La media de los 4 valores de distancias calculados es  $(8,5+7,2+7,8+6,7)/4=7.55$ .

	A	B
B	7,55	
3	4,75	2,9

El valor más pequeño es 2.9, luego juntamos B con 3 en C y el proceso termina.

El dendograma obtenido sería muy similar al del ejemplo anterior con la única difiere ligeramente en las distancias a las que se fusionan los clusters.

**OBSERVACION:** en el proceso se han utilizado solamente las distancias, de forma que para este procedimiento no es necesario disponer de los valores originales de las variables, basta con cualquiera de las matrices de distancias que utilizábamos en capítulos anteriores.

### CARACTERISTICAS

- ✓ Proporciona clusters que no resultan ni demasiado grandes ni demasiado pequeños.
- ✓ Pueden utilizarse medidas de similitud.

- ✓ Tiende a fusionar clusters con varianzas pequeñas y tiende a proporcionar clusters con la misma varianza.
- ✓ Buena representación gráfica de los resultados.

### METODO DEL VECINO MÁS PRÓXIMO

En el método del vecino más próximo la distancia entre dos clusters es el **mínimo de las distancias entre un objeto cualquiera de uno de los clusters y un objeto cualquiera del otro**.

$$d(R, \{P, Q\}) = \min\{d(R, P), d(R, Q)\}$$

Utilizando nuevamente el ejemplo con matriz de distancias:

	1	2	3	4	5
1	0.0				
2	1.0	0.0			
3	5.0	4.5	0.0		
4	8.5	7.8	3.6	0.0	
5	7.2	6.7	2.2	2.0	0.0

después de agrupar el 1 y el 2 en el cluster A, calculamos las distancias del nuevo cluster A a los clusters 3, 4 y 5.

$$\text{Dist}(A, 3) = \min\{d(1, 3); d(2, 3)\} = \min\{5; 4,5\} = 4,5$$

$$\text{Dist}(A, 4) = \min\{d(1, 4); d(2, 4)\} = \min\{7,8; 8,5\} = 7,8$$

$$\text{Dist}(A, 5) = \min\{d(1, 5); d(2, 5)\} = \min\{7,2; 6,7\} = 6,7$$

la matriz de distancias es entonces:

	A	3	4	5
A	0			
3	4,5	0		
4	7,8	3,6	0	
5	6,7	2,2	2	0

De nuevo, la distancia más pequeña es entre 4 y 5, por lo que los fusionamos en un

cluster que denominamos B.

Calculamos la distancia entre B y el resto, es decir, con A y con 3.

Entre A y B, buscamos las distancias entre todos los pares de puntos y calculamos el mínimo.

		B	
		4	5
A	1	8,5	7,2
	2	7,8	6,7

El mínimo de los 4 valores es 6,7.

La distancia entre B y 3 es 2,2; entonces la matriz de distancias queda:

	A	B	3
A	0		
B	6,7	0	
3	4,5	2,2	0

El valor más pequeño es 2.2, luego juntamos B con 3 en C.

El dendograma obtenido sería muy similar al de los ejemplos anteriores con la única difiere ligeramente en las distancias a las que se fusionan los clústeres.

Obsérvese que en el proceso se han utilizado solamente las distancias, de forma que para este procedimiento basta con cualquiera de las matrices de distancias que utilizábamos en capítulos anteriores.

## CARACTERÍSTICAS

- ⌚ Resulta útil para detectar outliers (estarán entre los últimos en unirse a la jerarquía).
- ⌚ Pueden usarse medidas de la similitud.
- ⌚ Tiende a construir clústeres demasiado grandes.

### METODO DEL VECINO MAS LEJANO (COMPLETE LINKAGE)

En el método del vecino más lejano la distancia entre dos clusters es el máximo de las distancias entre un objeto de uno de los clusters y un objeto del otro cluster.

$$d(R, \{P, Q\}) = \max\{d(R, P), d(R, Q)\}$$

#### CARACTERISTICAS

- ★ Resulta útil para detectar outliers.
- ★ Pueden usarse medidas de la similitud.
- ★ Tiende a construir clústeres pequeños y compactos.

### MÉTODO DEL CENTROIDE

El método del centroide es el que se utilizó en el ejemplo ilustrativo para la construcción del dendograma. La distancia entre dos clusters se calcula como la distancia entre los centroides de los mismos, por tanto es necesario disponer de los valores originales de las variables.

#### CARACTERISTICAS

- ☹ **Las variables deben estar en escala de intervalo.**
- ☹ Las distancias entre grupos se calculan como las distancias entre los vectores medios de cada uno de los grupos.
- ☹ Si los tamaños de los dos grupos a mezclar son muy diferentes, entonces el centroide del nuevo grupo será muy próximo al de mayor tamaño y probablemente estará dentro de este grupo.



## MÉTODO DE WARD (MÉTODO DE VARIANZA MINIMA)

La distancia entre dos clusters se calcula como la suma de cuadrados entre grupos en el ANOVA sumando para todas las variables (ahora vamos a desarrollar bien este concepto!!). En cada paso se minimiza la suma de cuadrados dentro de los clusters sobre todas las particiones posibles obtenidas fusionando dos clusters del paso anterior. Las sumas de cuadrados son más fáciles de entender cuando se expresan como porcentaje de la suma de cuadrados total.

Los cálculos son más complejos que para los casos anteriores.

Hasta acá hemos tratado con métodos aglomerativos; considerando diferentes distancias y técnicas.

## Veamos ahora... Cómo funcionan los métodos divisivos??

Los métodos divisivos trabajan en la dirección opuesta que los aglomerativos. Parten de un gran cluster que contiene a todos los elementos y buscan subdividirlo en clusters más pequeños. Si disponemos de  $n$  elementos, y los queremos partir en dos subclusters, disponemos de  $2^{n-1} - 1$  posibles particiones.

n=2	 	 / 
n=3	  	 /  
		  / 
		 /  

Para cada una de estas posibles particiones, deberíamos calcular alguna medida que nos indique la eficiencia de la misma. **Si tenemos muchos objetos esto puede resultar complejo.**

Sin embargo, para el caso de  $p$  variables binarias existen métodos computacionalmente simples y eficientes que se conocen como **métodos monocategóricos**.

Estos dividen los clústeres de acuerdo con la presencia o ausencia de cada una de las características; de esta forma en cada subdivisión los clusters contienen individuos con ciertos atributos, todos presentes o todos ausentes dentro del cluster.

El término monocategórico se refiere al uso de una única variable para definir la partición en cada etapa del proceso.

En contraposición los métodos policategóricos basan en más de una variable la partición de cada etapa del proceso.

La elección de la variable en la cual se basara la siguiente partición depende de la optimización de un criterio que contemple al mismo tiempo: la homogeneidad de los nuevos conglomerados y la asociación de las variables.

Esto tiende a minimizar el número de particiones que deben realizarse.

Lance y Williams(1968) propusieron como criterio de homogeneidad el índice  $C$ (de caos):

$$C = pn \log(n) - \sum_{k=1}^p [f_k \log f_k - (n - f_k) \log(n - f_k)]$$

Donde:

$f_k$  es el número de individuos que tienen el atributo  $k$

$n$ : el el número de individuos

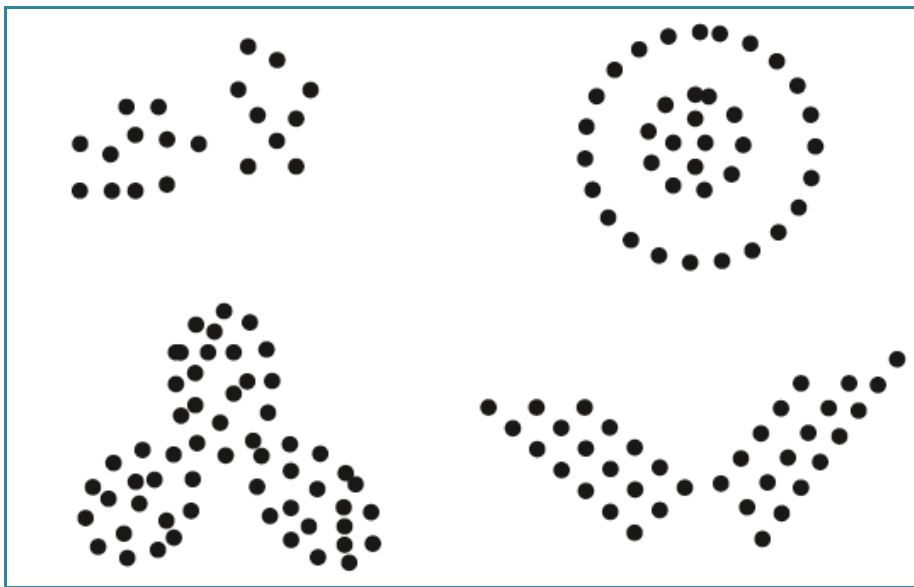
$p$ : es la cantidad de variables consideradas sobre cada individuo

A pesar de que los métodos divisivos son menos usados que los aglomerativos; tienen la ventaja de revelar la estructura principal de los datos.

## CUANTOS CLUSTERS??

La **decisión sobre el número óptimo de clusters es subjetiva**, especialmente cuando se incrementa el número de objetos ya que si se seleccionan demasiado pocos, los clusters resultantes son heterogéneos y artificiales, mientras que si se seleccionan demasiados, la interpretación de los mismos suele ser complicada

## NÚMERO DE CLUSTERS : SON 4 O SON 9???



.Como ayuda a la decisión sobre el número de clusters se suelen representar los distintos pasos del algoritmo y la distancia a la que se produce la fusión. En los primeros pasos el salto en las distancias será pequeño, mientras que en los últimos el salto entre pasos será mayor. **El punto de corte será aquel en el que comienzan a producirse saltos bruscos.**

Algunos analistas recomiendan aplicar varios algoritmos de agrupamiento y de selección o combinación de variables para cada conjunto de datos. Seleccionando, finalmente, desde los agrupamientos realizados la interpretación más apropiada.

InfoStat provee automáticamente el valor del coeficiente de correlación cofenética el cual puede ser usado para seleccionar uno de varios agrupamientos alternativos. **Este coeficiente indica la correlación de las distancias definidas por la métrica de árbol binario con las distancias originales entre objetos, luego se espera que el agrupamiento con mayor coeficiente sea el que mejor describe el agrupamiento natural de los datos.**

Es importante destacar que los procedimientos de agrupamiento producen resultados exitosos cuando la matriz de datos tiene una estructura que es posible interpretar desde el problema que originó la recolección de la información. Por ello, logrados los grupos es importante caracterizar los mismos a través de diversas medidas resumen para favorecer la interpretación del agrupamiento final.

## Métodos de partición no jerárquicos

### El más conocido y utilizado es el Algoritmo K means(K medias)

En qué consiste??

**Se debe fijar el número de conglomerados deseados K.**

El algoritmo se implementa en 4 pasos :

#### 1- Se divide a los n objetos en k subconjuntos no vacíos

Como elegir los k subconjuntos??

- Opción 1: asignándolos aleatoriamente.
- Opción 2: tomando los centros de los grupos como los puntos más alejados entre sí.
- Opción 3: utilizando información previa disponible.

#### 2- Se calcula el centroide (punto medio) del clúster

Esta asignación es secuencial, cada vez que se reasigna un elemento se vuelve a calcular el centroide del nuevo clúster.

**3- Se reasigna cada objeto al centroide de clúster más cercano**

**3- Se vuelve al paso 2, hasta que no se convenga realizar nuevas asignaciones.**

Esto sucede cuando no se puede mejorar el criterio de optimalidad establecido.

### Ventajas

- ♥ Relativamente eficiente
- ♥ Generalmente termina con un óptimo local.

### Limitaciones

- ♥ Solo es aplicable cuando la media está definida.
- ♥ Se necesita especificar K de antemano.

Un criterio de optimalidad muy difundido es minimizar la suma de cuadrados dentro del grupo, es decir minimizar las distancias al centroide del grupo de los elementos.

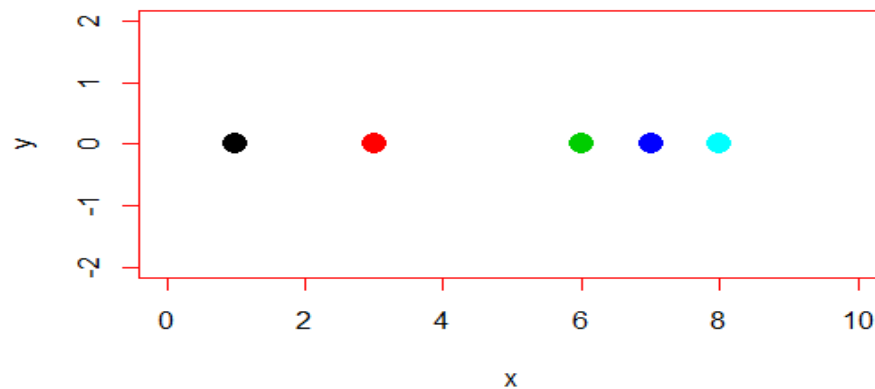
$$SCDG(\text{suma de cuadrados dentro del grupo}) = \sum_{k=1}^K \sum_{j=1}^p \sum_{i=1}^{n_k} (x_{ijk} - \bar{x}_{jk})^2$$

Otra forma de expresar esta sumatoria es:

$$\text{mín. SCDG} = \text{mín} \sum_{j=1}^p \sum_{k=1}^K n_k s_{jk}^2$$

Veamos un pequeño ejemplo univariado y luego extendamos este concepto a vectores multivariados:

Consideremos las observaciones: { 1-3- 6-7-8 }



Queremos encontrar dos clusters.

Elegimos como centroides a los puntos mas alejados. {1,8}

Calculamos la distancia de todos los puntos a los centroides elegidos:

observación	Distancia a {1}	Distancia a {8}	Grupo asignado
1	0	7	1
2	1	6	1
6	5	2	8
7	6	1	8
8	7	0	8

Calculamos los nuevos centroides:

$G_1: \{1,2\} \rightarrow$  Centroide: 1.5

$G_2: \{6,7,8\} \rightarrow$  Centroide: 7

Calculamos las distancias a los nuevos centroides:

observación	Distancia a {1.5}	Distancia a {7}	Grupo asignado
1	0.5	6	1
2	0.5	5	1
6	4.5	1	8
7	5.5	0	8
8	6.5	1	8

Si bien al cambiar los centroides cambiaron las distancias, no cambio la clasificación original.

Como es un caso sencillo, no tiene sentido buscar el criterio de optimalidad.

Analicemos la metodología para vectores:

El criterio de optimalidad ahora puede expresarse de la siguiente manera:

$$\min \sum_{k=1}^K \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)' (x_{ik} - \bar{x}_k) = \min \sum_{k=1}^K \sum_{i=1}^{n_k} d^2(i, k) =$$

$$\min \sum_{k=1}^K \sum_{i=1}^{n_k} Tr(W)$$

Siendo:

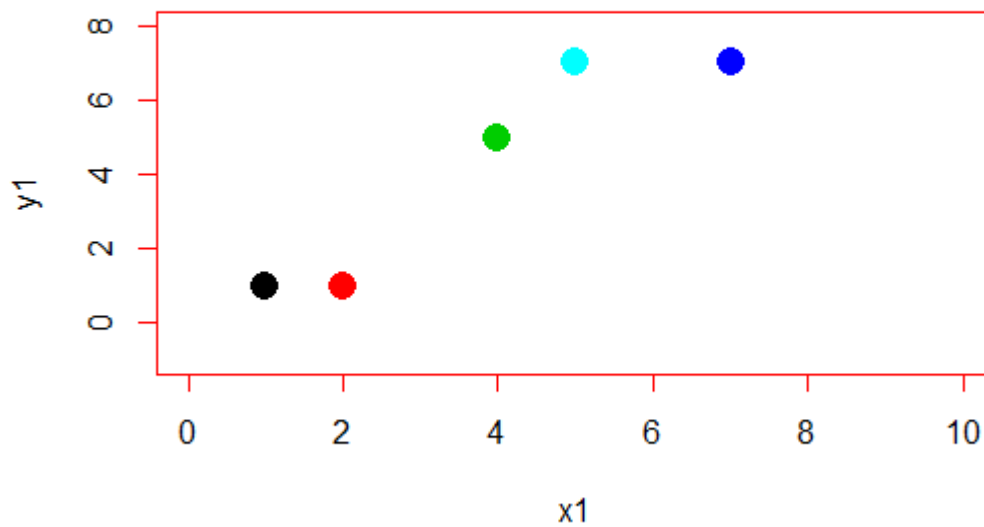
$$W = \sum_{k=1}^K \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k) (x_{ik} - \bar{x}_k)'$$

Ilustremos estas consideraciones para p=2

Tomemos los puntos del algoritmo jerarquice descripto:

objeto	V1	V2
1	1	1
2	2	1
3	4	5
4	7	7
5	5	7

Los graficamos y elegimos la distancia euclidea por simplicidad:



Elegimos dos centroides al azar entre los puntos:

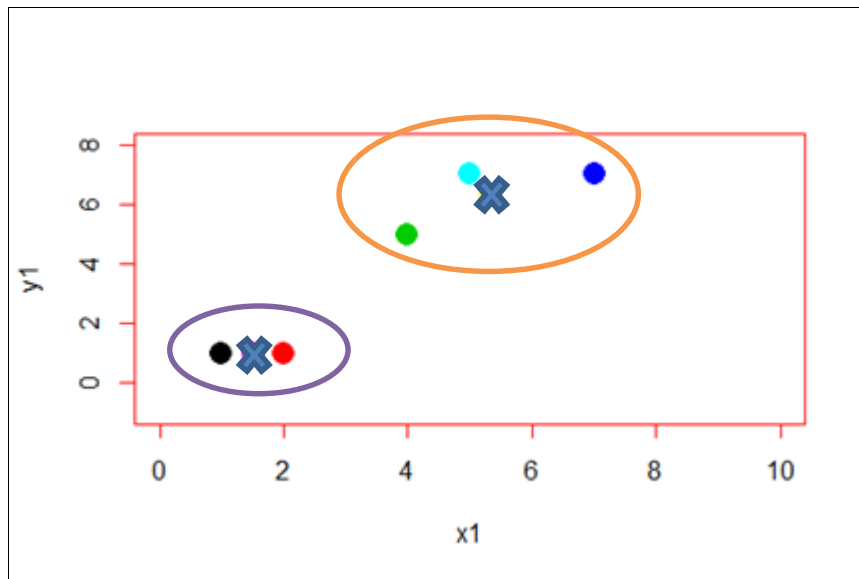
$C_{g1} = (2,1)$  y  $C_{g2} = (7,7)$

Calculamos las distancias de los puntos a los dos centroides elegidos:

objeto	V1	V2	distG1	distG2	Grupo de pertenencia
1	1	1	1	7.211	1
2	2	1	0	6.708	1
3	4	5	4.472	2.23	2
4	7	7	7.81	2	2
5	5	7	6.708	0	2

Calculamos ahora los nuevos centroides de los grupos 1 y 2:





En cada grupo aparece una crucecita, que representa el centroide.

$C_{g1} = (1.5, 1)$  y  $C_{g2} = (16/3, 19/3)$

objeto	V1	V2	distG1	distG2	Grupo de pertenencia
1	1	1	0.5	6.872	G1
2	2	1	0.5	6.289	G1
3	4	5	4.717	1.886	G2
4	7	7	8.139	1.795	G2
5	5	7	6.946	0.745	G2

Nuevamente las distancias cambian pero no el grupo de pertenencia.

Para determinar el número de grupos suele utilizarse el contraste:

$$F = \frac{SCDG(G) - SCDG(G + 1)}{SCDG(G + 1)/(n - G - 1)}$$

Que se compara con una  $F_{p,p(N-K-1); 0.95}$

¿Qué sucedería en nuestro ejemplo si proponemos que sean 3 los conglomerados...es decir  $k=3$ ?

Proponemos tres centroides, elegidos aleatoriamente:

$$C_{g1} = (1,1) ; C_{g2} = (4,5) \text{ y } C_{g3} = (7,7)$$

Calculamos las distancias de los puntos a los tres centroides elegidos y los clasificamos en el grupo del que están a menor distancia:

objeto	V1	V2	distG1	distG2	distG3	Grupo de pertenencia
1	1	1	0	5	8.485	G1
2	2	1	1	4.472	7.81	G1
3	4	5	5	0	3.606	G2
4	7	7	8.485	3.606	0	G3
5	5	7	7.211	2.236	2	G3

Calculamos los nuevos centroides:

$$C_{g1} = (1.5,1) ; C_{g2} = (4,5) \text{ y } C_{g3} = (6,7)$$

Estudiamos las distancias de los puntos a estos nuevos centroides:

objeto	V1	V2	distG1	distG2	distG3	Grupo de pertenencia
1	1	1	0.5	5	7.81	G1
2	2	1	0.5	4.472	7.211	G1
3	4	5	4.717	0	2.828	G2
4	7	7	8.139	3.606	1	G3
5	5	7	6.946	2.236	1	G3

Como no han cambiado las clasificaciones no se realizan movimientos.

La pregunta que queda pendiente es si conviene realizar la partición en 2 o en 3 clusters.

Utilizaremos para ello el criterio de F propuesto:

$$F = \frac{SCDG(G) - SCDG(G + 1)}{SCDG(G + 1)/(n - G - 1)}$$

$$SCDG(3) = 0.5^2 + 0.5^2 + 0^2 + 1^2 + 1^2 = 2.5$$

$$SCDG(2) = 0.5^2 + 0.5^2 + 1.886^2 + 1.795^2 + 0.745^2 = 7.834$$

$$\text{Entonces } F = F' = \frac{7.834 - 2.5}{2.5 / (5 - 3 - 1)} = 2.134$$

Comparamos este valor con el cuantil 0.95 de la distribución F de Snedecor  $F_{0.95, 2, 2} = 19$

Como  $2.134 < 19$  no resulta significativa la mejora de considerar 3 clusters en lugar de considerar 2.

### Características del método

- No se satisface el criterio de optimización globalmente, solo produce un óptimo local.
- El algoritmo de k-means es computacionalmente rápida.
- Puede trabajar bien con datos faltantes (missing values).
- Es sensible a la presencia de “outliers”.

### Consideraciones Finales

#### I- Como elegir la medida de proximidad o distancia??

- ✪ La elección de esta medida depende en primera instancia de la naturaleza de los datos.
- ✪ Bajo ciertas circunstancias conviene discretizar el análisis de alguna variable continua o bien categorizarla. Por ejemplo, si la mayoría de las variables son categóricas y la variable edad, que es de escala continua, me parece relevante para el análisis, se podría subdividir las categorías de edad...existen diferentes formas de hacer esta subdivisión, en un estudio médico podría considerarse la subdivisión en función del riesgo...en un estudio de mercado en función de los niveles de disposición del dinero...etc.
- ✪ La selección de la distancia y de la técnica, pueden cambiar la disposición de los clusters.

## II- Como tratar los valores perdidos (missing data)?

- ⌚ La forma más simple, aunque no siempre implica la mejor, es utilizar únicamente los registros completos. Esto, sin embargo, puede reducir drásticamente la información disponible para el estudio.
- ⌚ Otra estrategia alternativa es utilizar el coeficiente de similitud Gower, si hay al menos una variable disponible del registro; de esta forma las variables no disponibles en ambos registros se desconsideran en el cálculo.
- ⌚ En otras técnicas multivariadas existe la alternativa es estimar los datos faltantes mediante algún resumen estadístico de los datos disponibles en los restantes registros( como la media o la mediana por ejemplo); esta estrategia no debe utilizarse en análisis de clusters.

## III- Algoritmos de Optimización:

- ⌚ Una vez lograda una partición en conglomerados cabe preguntarse si la lograda es o no la mejor cauterización. Para optimizar la partición es posible utilizar técnicas.
- ⌚ Se elige un criterio a optimizar y se cambiará la partición sólo si mejora el criterio elegido respecto de la anterior. Existen algoritmos ascendentes y descendentes.
- ⌚ La esencia de los algoritmos es: encontrar alguna partición inicial de n objetos en g grupos, calcular el cambio en la función objetivo de mover cada uno de los elementos de grupo. Efectuar el cambio en caso de que algún movimiento produjera una mejora del criterio. Se repite el procedimiento hasta que ningún cambio produzca mejora en el criterio objetivo.
- ⌚ La partición inicial puede hacerse en base a un conocimiento anterior, al azar o mediante algún algoritmo de los detallados anteriormente para clusterizar.

### Ejemplo 1: razas de perros

Se ha registrado sobre 27 razas de perros, el tamaño, el peso, la inteligencia, la afectividad, la agresividad y la función.

Intentamos agrupar a estos perros en un número de clusters que de sentido a la agrupación.

raza	tamaño	peso	veloci	intelig	afectivd	agresiv	función
Basset	1	1	1	1	1	3	caza
beaucheron	3	2	3	2	3	3	utilitario
Bóxer	2	2	2	2	3	3	compañía
Buldog	1	1	1	2	3	1	compañía
bulmastif	3	3	1	3	1	3	utilitario
caniche	1	1	2	3	3	1	compañía
chihuahua	1	1	1	1	3	1	compañía
Cocker	2	1	1	2	3	3	compañía
Collie	3	2	3	2	3	1	compañía
dalmata	2	2	2	2	3	1	compañía
doberman	3	2	3	3	1	3	utilitario
Dogo	3	3	3	1	1	3	utilitario
foxhound	3	2	3	1	1	3	caza
foxterrier	1	1	2	2	3	3	compañía
Galgo	3	2	3	1	1	1	caza
Gascon	3	2	2	1	1	3	caza

labrador	2	2	2	2	3	1	caza
Masa	3	2	3	3	3	3	utilitario
Mastin	3	3	1	1	1	3	utilitario
pekines	1	1	1	1	3	1	compañía
Podb	2	2	2	3	3	1	caza
Podf	3	2	2	2	1	1	caza
pointer	3	2	3	3	1	1	caza
Setter	3	2	3	2	1	1	caza
sanbernardo	3	3	1	2	1	3	utilitario
Teckel	1	1	1	2	3	1	compañía
terranova	3	3	1	2	1	1	utilitario

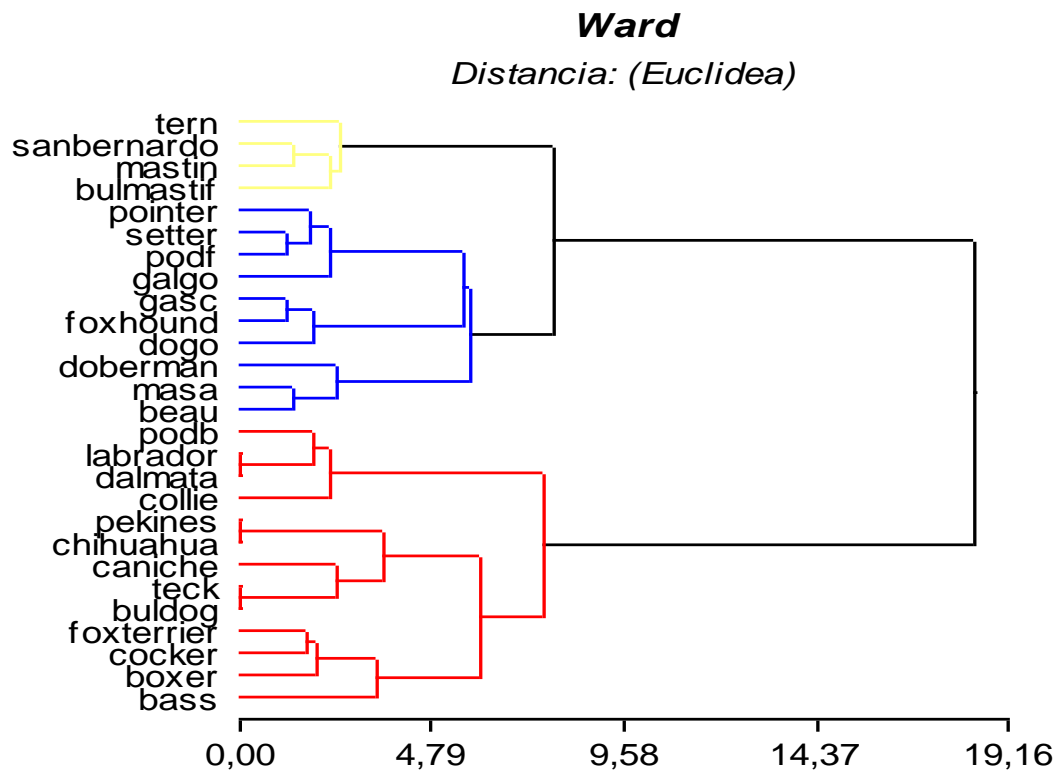
#### a) distancias euclídeas y método de Ward

k=3 **Ward** Distancia: (Euclídea) Correlación cofenética= 0,650

Conglomerado	caza	compañía	utilitario	Total
congl.1	3	10	0	13
congl.2	6	0	4	10
congl.3	0	0	4	4
Total	9	10	8	27

Se aprecia que esta clusterización está bastante vinculada a la función de los animales, dado que todos los del tercer conglomerado son utilitarios, casi todos los del primer conglomerado son de compañía y el segundo conglomerado tiene mayoría de caza.

El correspondiente dendrograma se muestra a continuación:

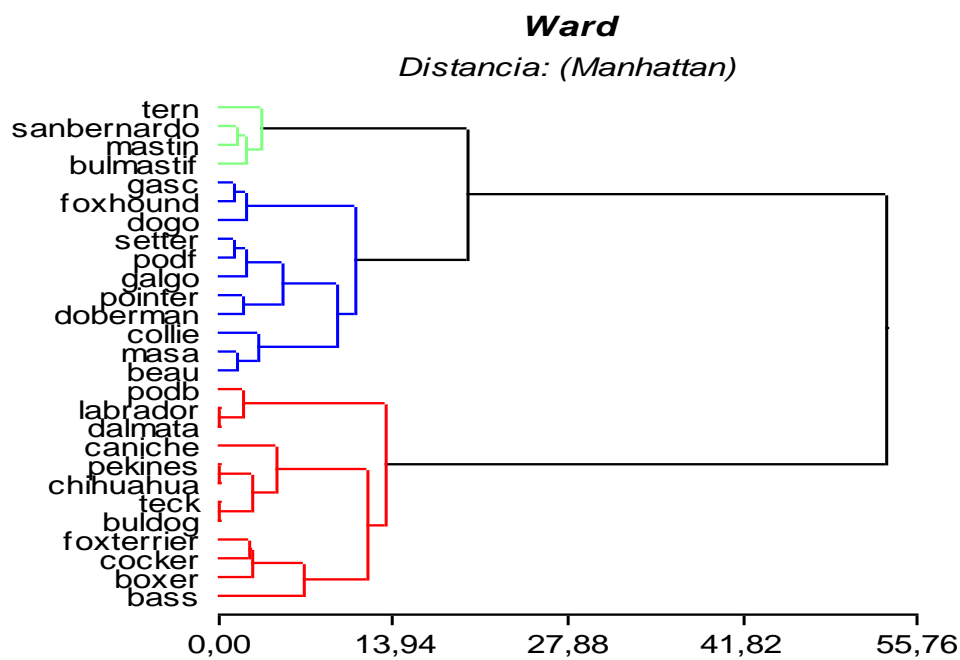


Si utilizamos distancias de Manhattan, para la clusterización, la clasificación resulta muy similar a la anterior y mejora el coeficiente de correlación cofenética.

Distancia: (Manhattan)

Correlación cofenética= 0,690

Probamos con  $k > 3$  y no mejora la correlación cofenética al mismo tiempo se pierde interpretabilidad de los clusters.



Conglomerado	caza	compañía	utilitario	Total
congl.1	3	9	0	12
congl.2	6	1	4	11
congl.3	0	0	4	4
Total	9	10	8	27

Utilizando average linkage y distancia euclídea

**K=3** Correlación cofenética= 0,728

*Mejora la correlación cofenética!!*

Conglomerado	caza	compañía	utilitario	Total
congl.1	1	8	0	9
congl.2	6	2	3	10
congl.3	2	0	5	8
Total	9	11	7	27



