Técnicas multivariantes en ecología (I): La clasificación

José Antonio Palazón Ferrando
y
José Francisco Calvo Sendín

Depto. Ecología e Hidrología

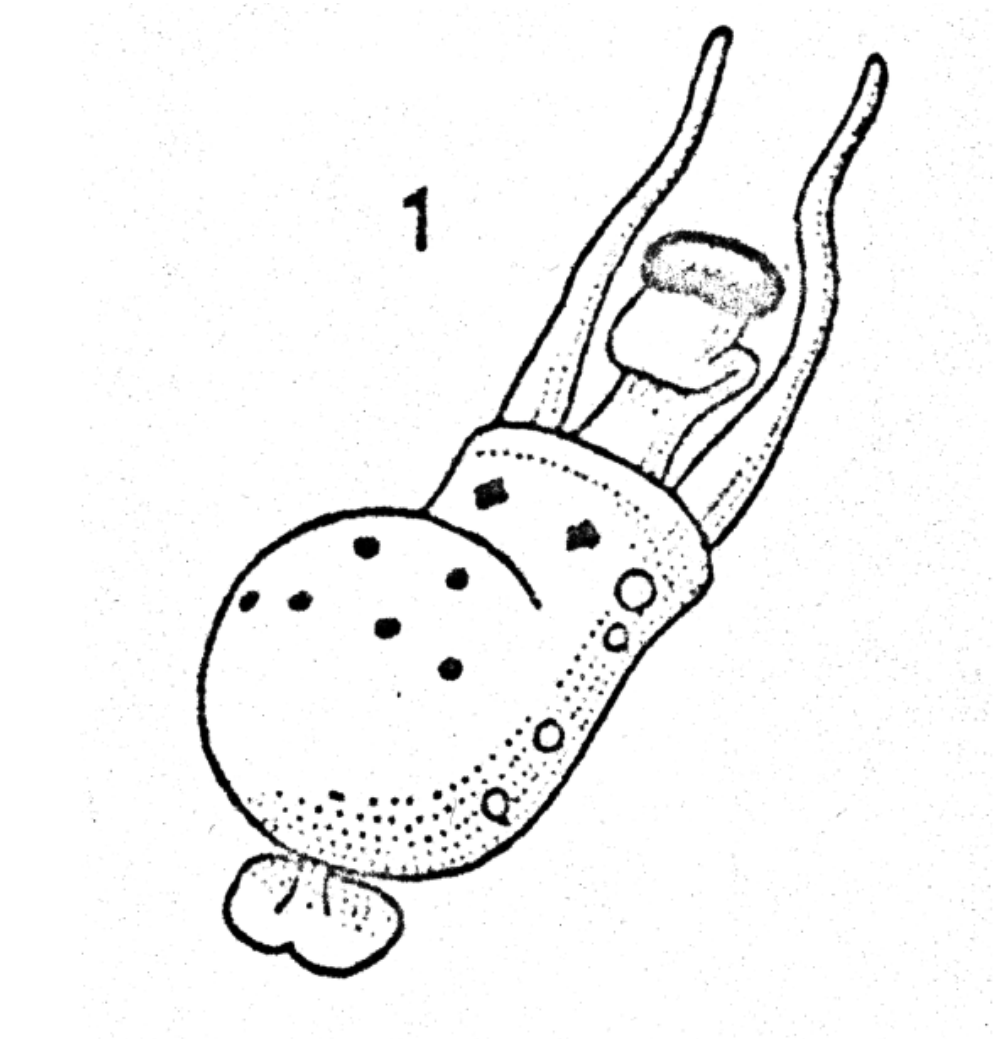
Universidad de Murcia

Ecol. Met. y Cuant., curso 2006–07

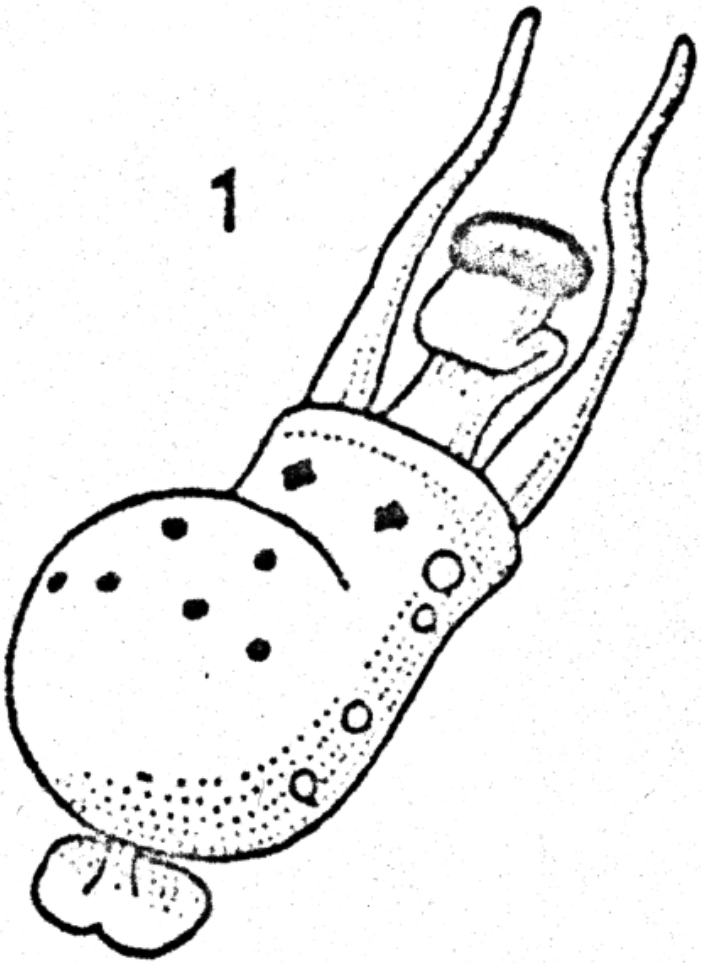
Sokal, *Numerical taxonomy*

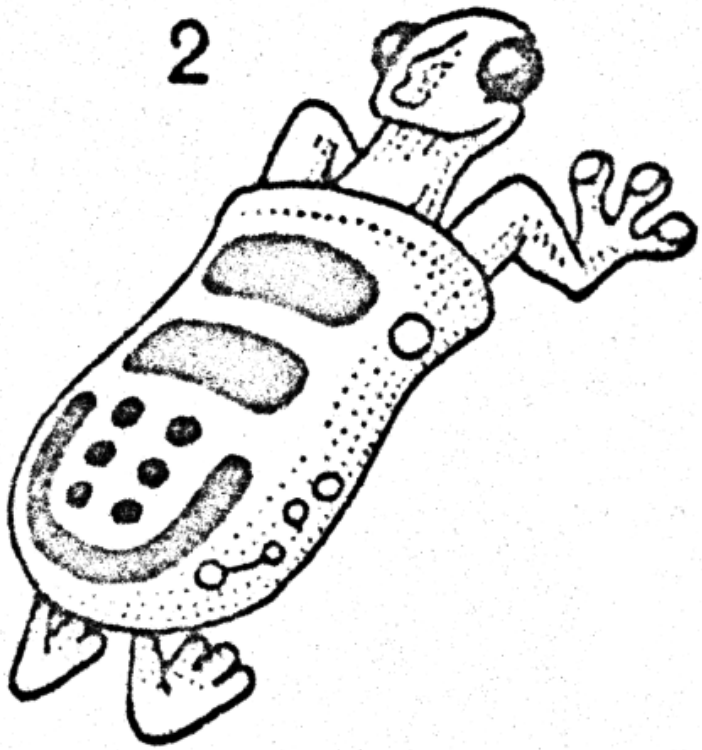
The computer has made possible to consider large numbers of characteristics in classifying many phenomena, notably living organisms, fossil organisms and eve imaginariy organisms.

Sokal, *Numerical taxonomy*

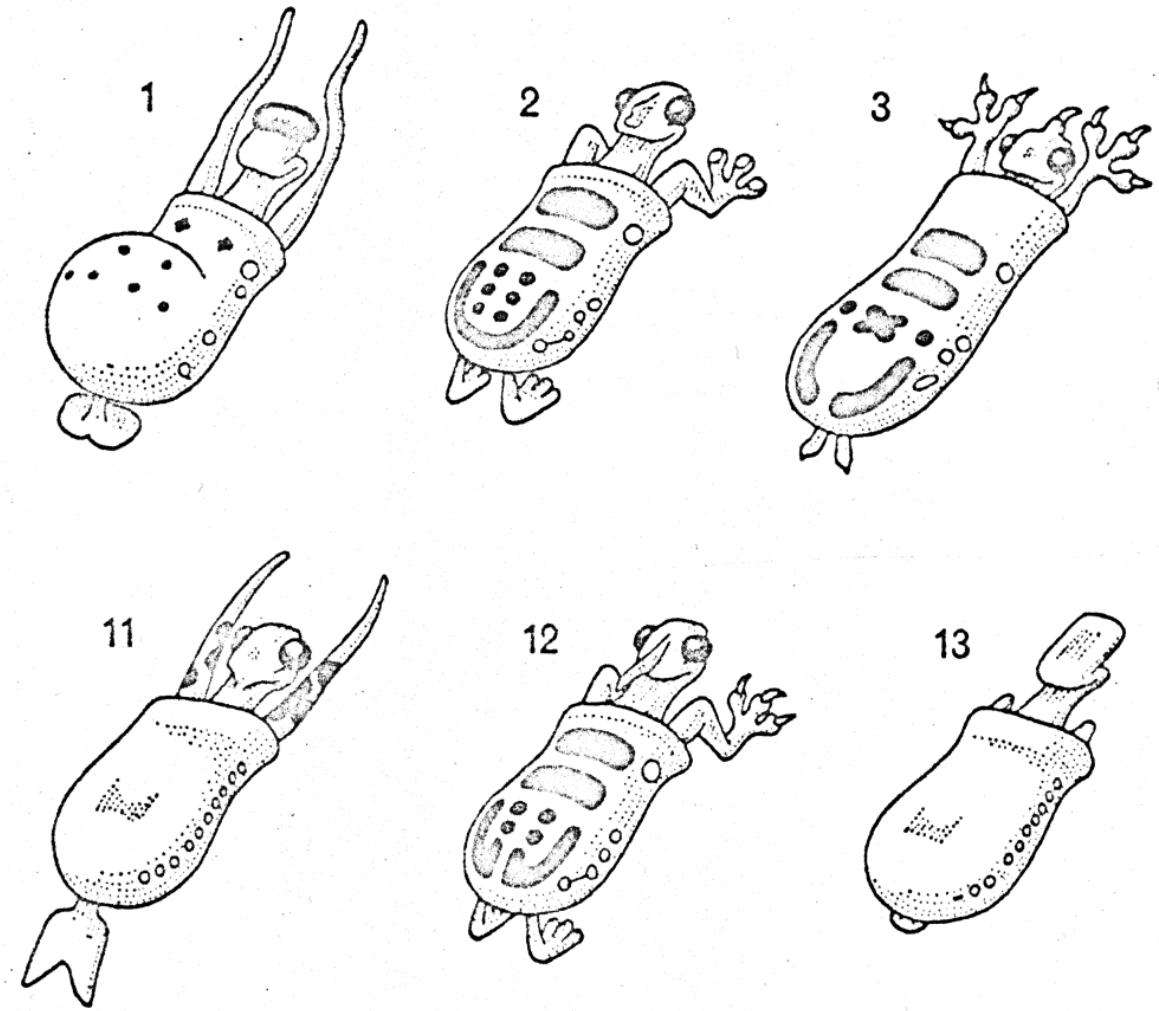


Sokal, *Numerical taxonomy*





Sokal, *Numerical taxonomy*



Organismos

Organismos

Paisajes

De la observación

- Observaciones: objetos de estudio
- Medidas: Procedimiento (tipos de variables)
- Objetividad: Regla de repetibilidad
- Sistema de referencia objetivo para los casos observados

De la observación

- Observaciones: objetos de estudio
- Medidas: Procedimiento (tipos de variables)
- Objetividad: Regla de repetibilidad
- Sistema de referencia objetivo para los casos observados y por tanto puedo compararlos

De la observación

- Observaciones: objetos de estudio
- Medidas: Procedimiento (tipos de variables)
- Objetividad: Regla de repetibilidad
- Sistema de referencia objetivo para los casos observados y por tanto puedo compararlos

De la observación

- Observaciones: objetos de estudio
- Medidas: Procedimiento (tipos de variables)
- Objetividad: Regla de repetibilidad
- Sistema de referencia objetivo para los casos observados y por tanto puedo compararlos

De la observación

- Observaciones: objetos de estudio
- Medidas: Procedimiento (tipos de variables)
- Objetividad: Regla de repetibilidad
- Sistema de referencia objetivo para los casos observados y por tanto puedo compararlos

De la teoría

- Conceptos: abstacción
- Realidad compleja: Multiples variables
- Procesos y patrones

De la teoría

- Conceptos: abstacción
- Realidad compleja: Multiples variables
- Procesos y patrones

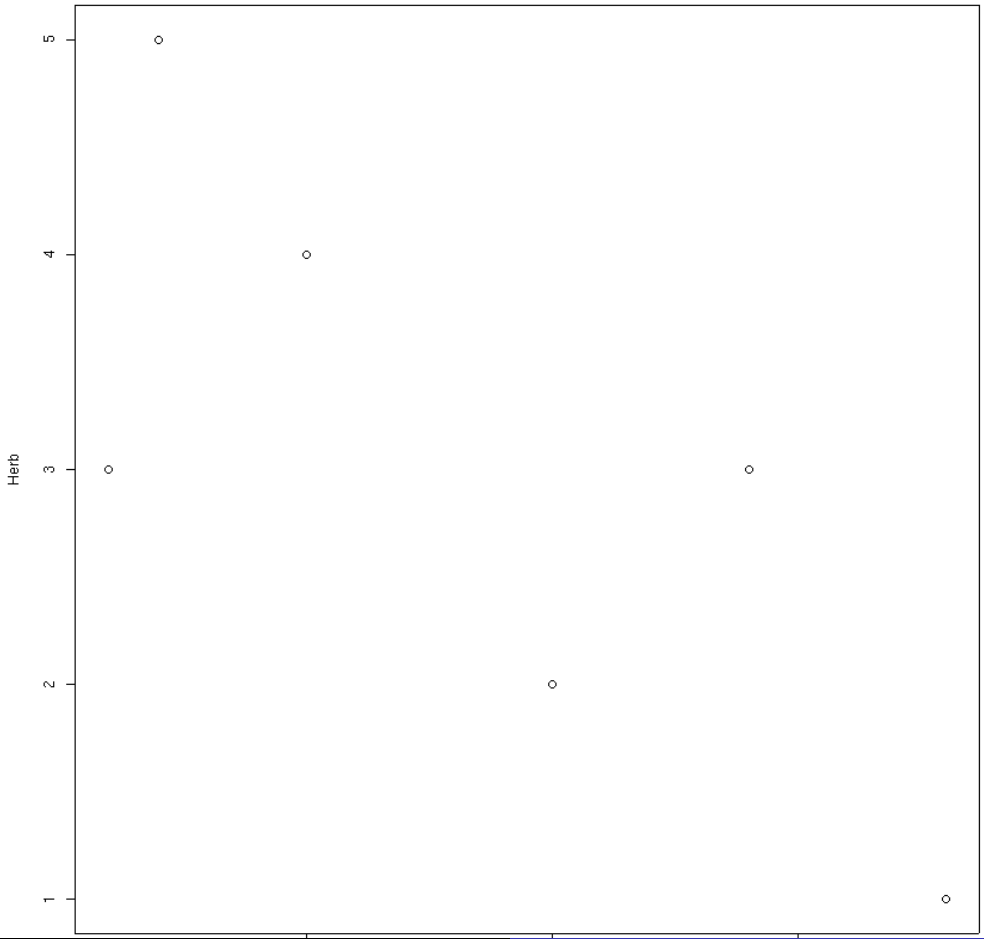
De la teoría

- Conceptos: abstacción
- Realidad compleja: Multiples variables
- Procesos y patrones

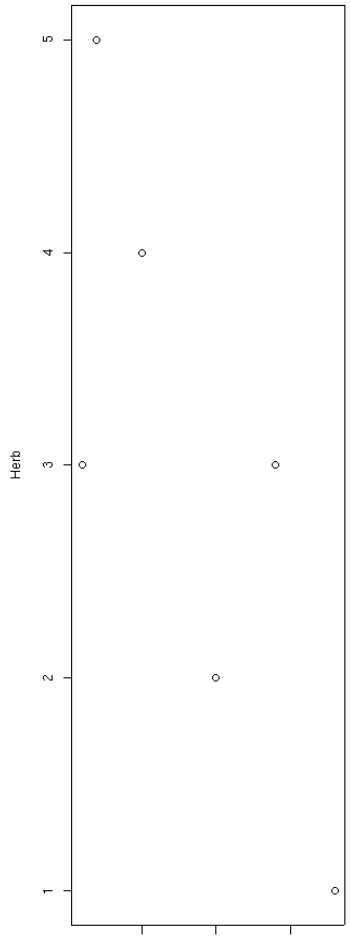
Dado un conjunto de datos ...

	Arb	**Herb**
1	18	1
2	14	3
3	10	2
4	5	4
5	2	5
6	1	3

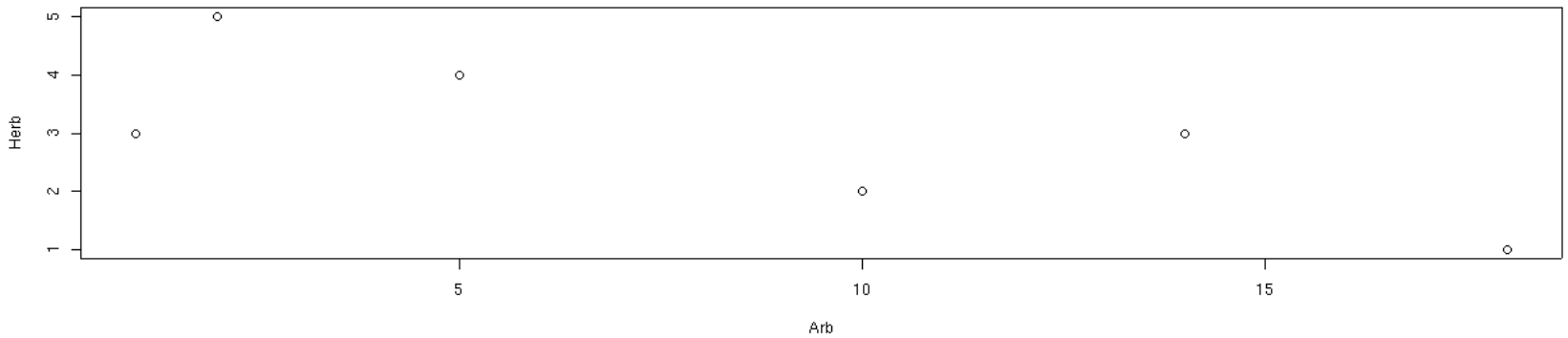
. . . ¿cuál es más real?



. . . ¿cuál es más real?



. . . ¿cuál es más real?

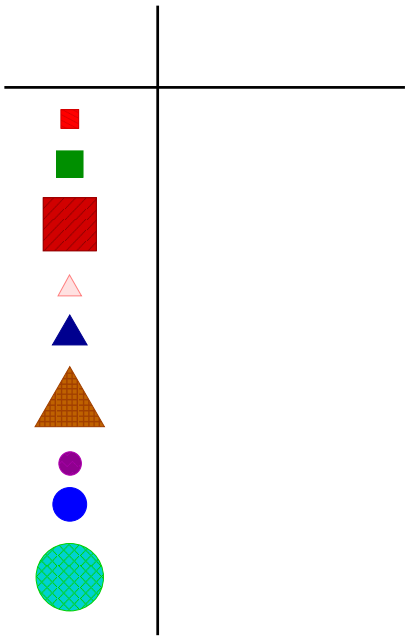


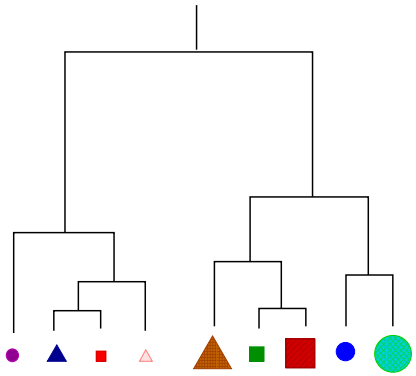
Principales problemas a resolver con técnicas multivariantes

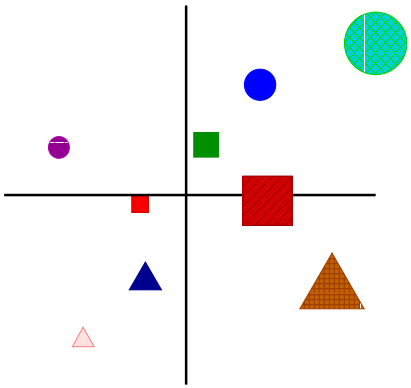
Los datos

La clasificación

La ordenación

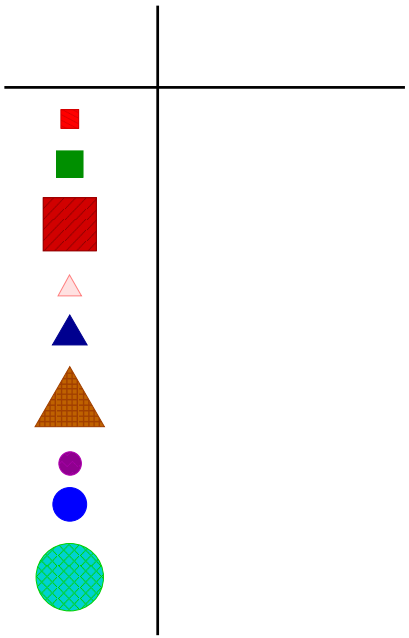




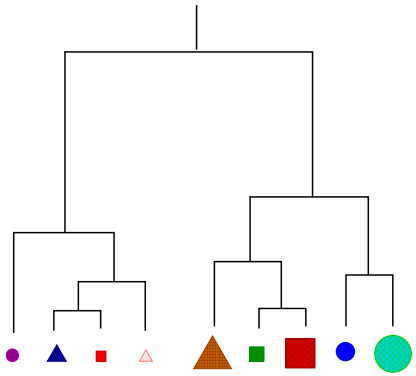


Principales problemas a resolver con técnicas multivariantes

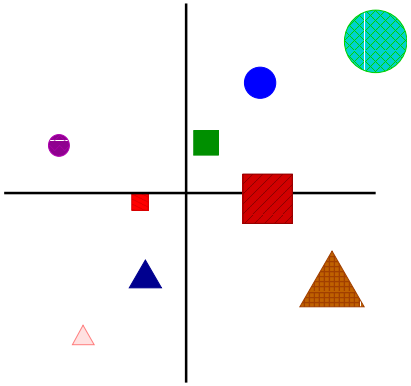
Los datos



La clasificación

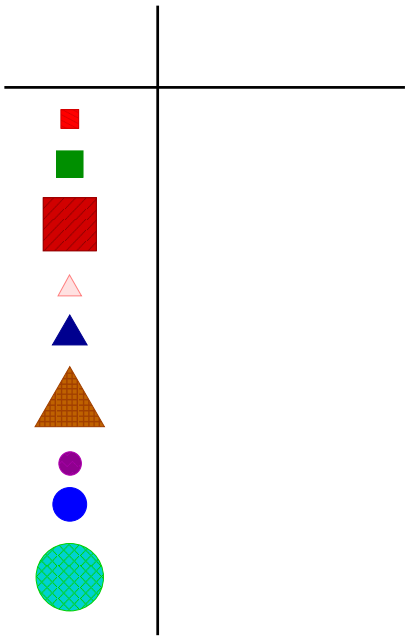


La ordenación

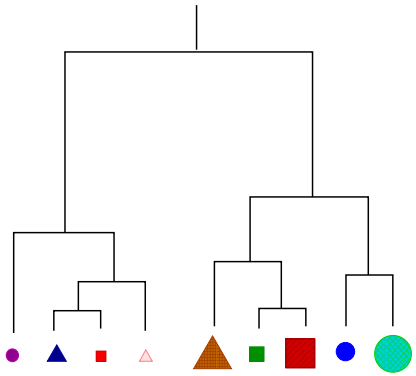


Principales problemas a resolver con técnicas multivariantes

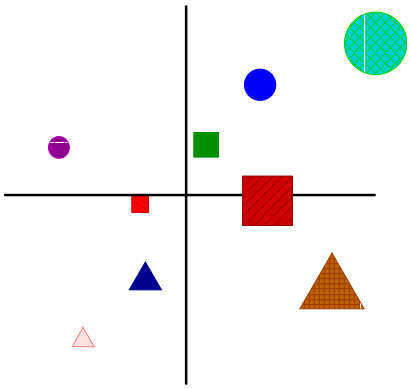
Los datos



La clasificación



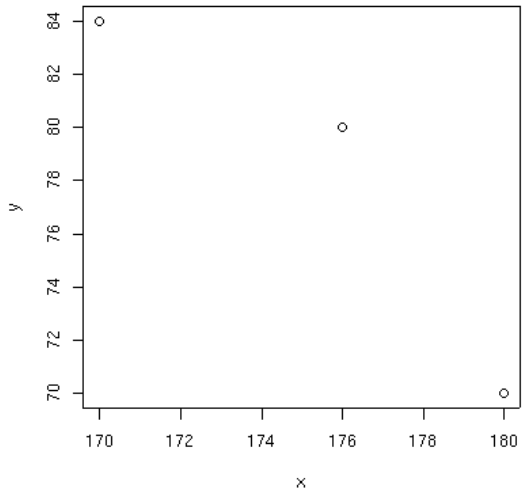
La ordenación



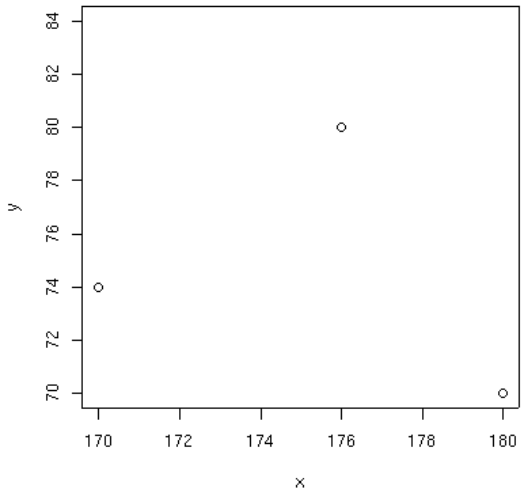
¿A quién me parezco más?

- Tres mejor que dos...

170	84
176	80
180	70



170	74
176	80
180	70

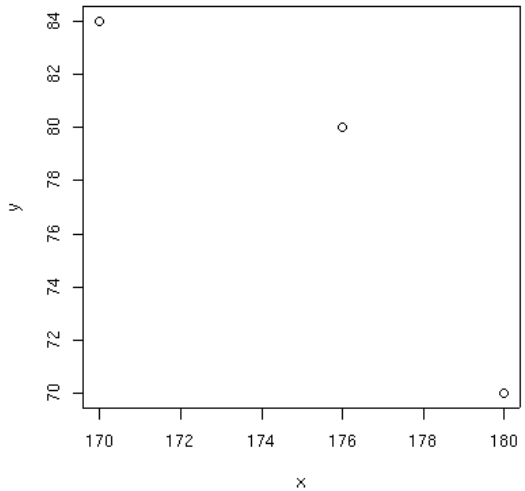


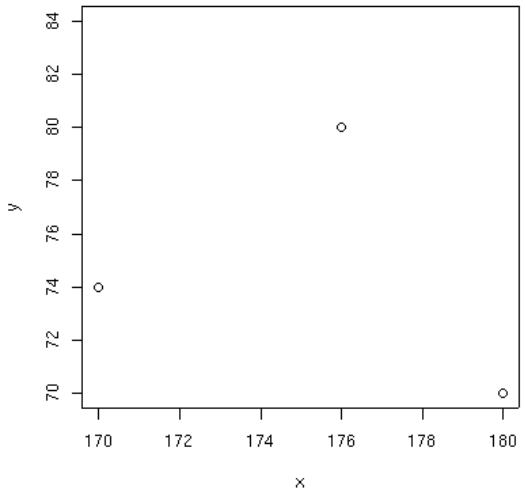
¿A quién me parezco más?

- Tres mejor que dos...

170	84
176	80
180	70

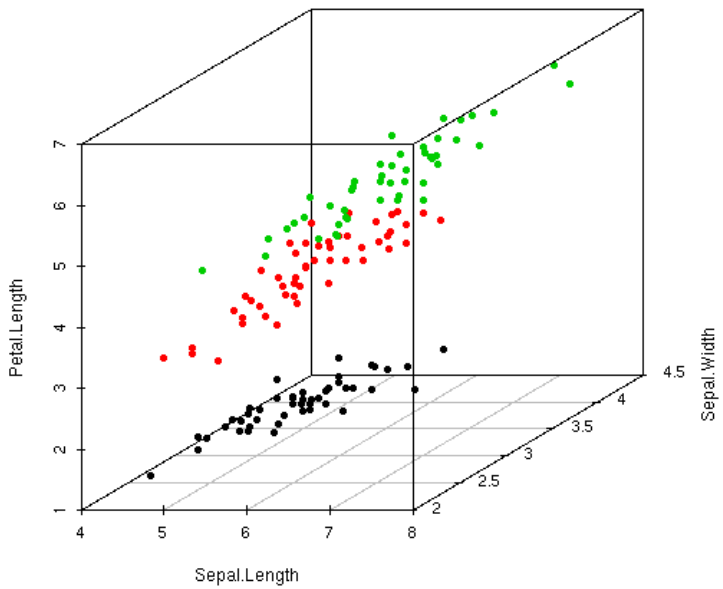
170	74
176	80
180	70

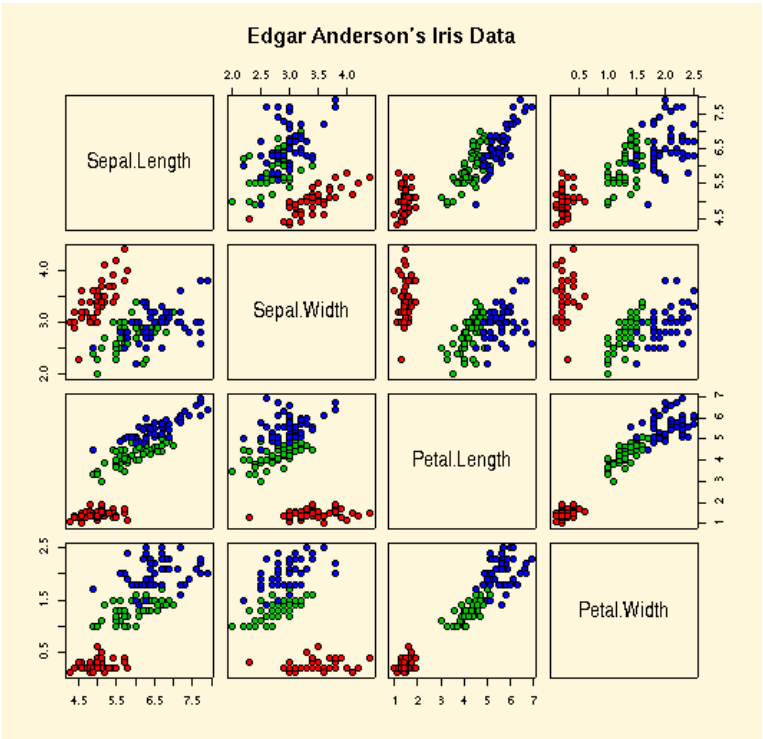




Las cosas se complican en la realidad . . .

. . . cuando el número de variables > 3





¿Qué variables puedo analizar?

- cualitativas
- cuantitativas
- una sóla naturaleza u origen
- más de un origen o naturaleza
- mixtas

Datos, datos, datos

- Ecología metodológica y cuantitativa
- Censos de la población murciana
- Datos biométricos de alumnos de Ecología metodológica y cuantitativa (2003–04)
- Hábitos y dietas de rapaces

Índices, similaridades y distancias

- Datos de presencia–ausencia

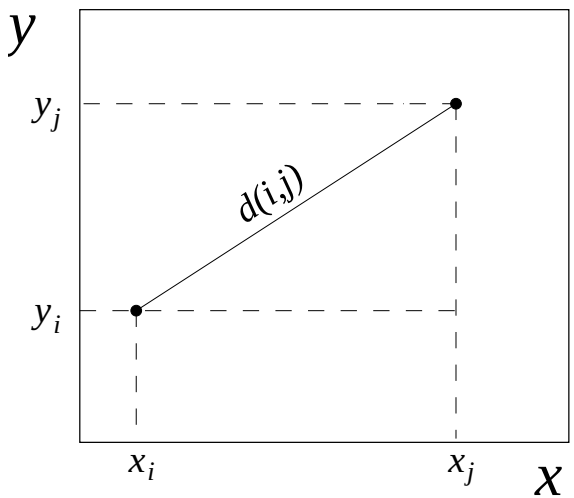
		Objeto j		
		1	0	
Objeto i	1	a	b	$a+b$
	0	c	d	$c+d$
		$a+c$	$b+d$	$a+b+c+d$

$$I_J = \frac{a}{a+b+c} \qquad I_{CS} = \frac{a+d}{a+b+c+d}$$

- Transformación de similaridad a distancia euclídea

$$d_J = \sqrt{1-I_J} \qquad d_{CS} = \sqrt{1-I_{CS}}$$

Distancia entre dos puntos



$$d(i,j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

$$d(i,j) = \sqrt{\sum_{k=1}^{p} (x_{i,k} - x_{j,k})^2}$$

Matriz de distancias o disimiliaridades

	1	2	3	4	5
1	0.00				
2	1.00	0.00			
3	2.83	2.24	0.00		
4	7.07	6.40	4.24	0.00	
5	7.62	7.28	5.10	2.83	0.00

Tipos de clasificación

- Posesión de criterios de grupo *a priori*
  - Automática
  - Supervisada
- Estructura de grupos y subgrupos
  - Particiones
  - Jerárquicas
- ¿Qué es el grupo inicial?
  - Aglomerativas
  - Divisivas

Objetivos

- Dividir al conjunto de observaciones en grupos más o menos homogéneos
- Fijar a priori el número de grupos deseado
- Obtener como resultado una variable cualitativa que indique la pertenecia al grupo

$k - means$: El procedimiento

1. Seleccionar el número de grupos que ha de tener la partición.
2. Elegir de forma aleatoria tantos líderes como grupos.
3. Calcular la distancia de todos los objetos a clasificar a cada uno de los líderes.
4. Asignar cada objeto a su líder más próximo.
5. Si no se han realizado cambios en la asignación a los líderes, o se ha pasado por este punto más del número fijado de veces ir al paso 8.
6. Cacular el centroide de cada grupo y declararlo nuevo líder del grupo.
7. Ir al paso 3.
8. La clasificación ha terminado: cada objeto pertenece al grupo al que ha sido asignado.

$k - means$: El procedimiento

1. Seleccionar el número de grupos que ha de tener la partición.
2. Elegir de forma aleatoria tantos líderes como grupos.
3. Calcular la distancia de todos los objetos a clasificar a cada uno de los líderes.
4. Asignar cada objeto a su líder más próximo.
5. Si no se han realizado cambios en la asignación a los líderes, o se ha pasado por este punto más del número fijado de veces ir al paso 8.
6. Cacular el centroide de cada grupo y declararlo nuevo líder del grupo.
7. Ir al paso 3.
8. La clasificación ha terminado: cada objeto pertenece al grupo al que ha sido asignado.

$k - means$: El procedimiento

1. Seleccionar el número de grupos que ha de tener la partición.
2. Elegir de forma aleatoria tantos líderes como grupos.
3. Calcular la distancia de todos los objetos a clasificar a cada uno de los líderes.
4. Asignar cada objeto a su líder más próximo.
5. Si no se han realizado cambios en la asignación a los líderes, o se ha pasado por este punto más del número fijado de veces ir al paso 8.
6. Cacular el centroide de cada grupo y declararlo nuevo líder del grupo.
7. Ir al paso 3.
8. La clasificación ha terminado: cada objeto pertenece al grupo al que ha sido asignado.

$k - means$: El procedimiento

1. Seleccionar el número de grupos que ha de tener la partición.
2. Elegir de forma aleatoria tantos líderes como grupos.
3. Calcular la distancia de todos los objetos a clasificar a cada uno de los líderes.
4. Asignar cada objeto a su líder más próximo.
5. Si no se han realizado cambios en la asignación a los líderes, o se ha pasado por este punto más del número fijado de veces ir al paso 8.
6. Cacular el centroide de cada grupo y declararlo nuevo líder del grupo.
7. Ir al paso 3.
8. La clasificación ha terminado: cada objeto pertenece al grupo al que ha sido asignado.

$k - means$: El procedimiento

1. Seleccionar el número de grupos que ha de tener la partición.
2. Elegir de forma aleatoria tantos líderes como grupos.
3. Calcular la distancia de todos los objetos a clasificar a cada uno de los líderes.
4. Asignar cada objeto a su líder más próximo.
5. Si no se han realizado cambios en la asignación a los líderes, o se ha pasado por este punto más del número fijado de veces ir al paso 8.
6. Cacular el centroide de cada grupo y declararlo nuevo líder del grupo.
7. Ir al paso 3.
8. La clasificación ha terminado: cada objeto pertenece al grupo al que ha sido asignado.

$k - means$: El procedimiento

1. Seleccionar el número de grupos que ha de tener la partición.
2. Elegir de forma aleatoria tantos líderes como grupos.
3. Calcular la distancia de todos los objetos a clasificar a cada uno de los líderes.
4. Asignar cada objeto a su líder más próximo.
5. Si no se han realizado cambios en la asignación a los líderes, o se ha pasado por este punto más del número fijado de veces ir al paso 8.
6. Cacular el centroide de cada grupo y declararlo nuevo líder del grupo.
7. Ir al paso 3.
8. La clasificación ha terminado: cada objeto pertenece al grupo al que ha sido asignado.

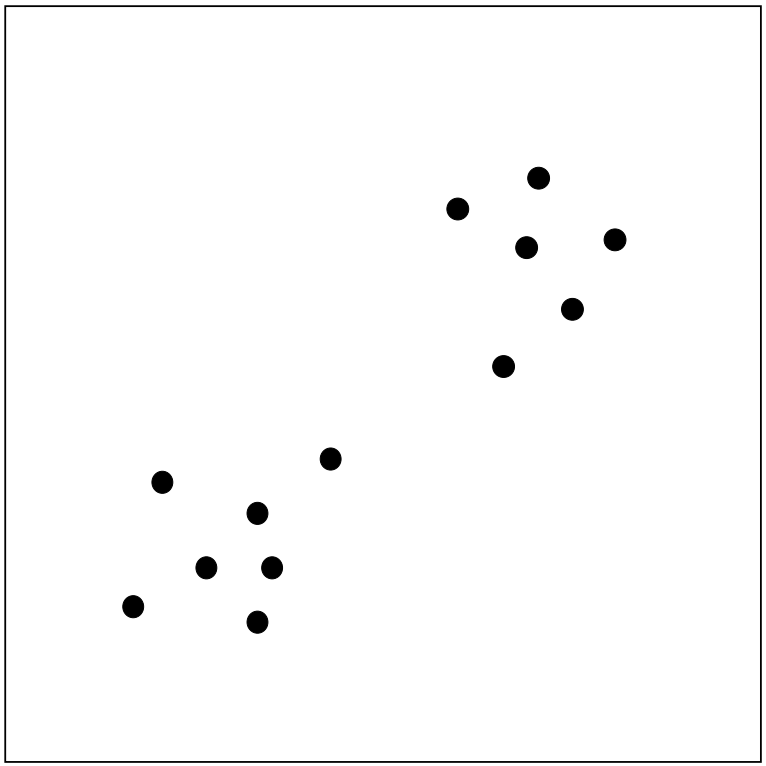
$k - means$: El procedimiento

1. Seleccionar el número de grupos que ha de tener la partición.
2. Elegir de forma aleatoria tantos líderes como grupos.
3. Calcular la distancia de todos los objetos a clasificar a cada uno de los líderes.
4. Asignar cada objeto a su líder más próximo.
5. Si no se han realizado cambios en la asignación a los líderes, o se ha pasado por este punto más del número fijado de veces ir al paso 8.
6. Cacular el centroide de cada grupo y declararlo nuevo líder del grupo.
7. Ir al paso 3.
8. La clasificación ha terminado: cada objeto pertenece al grupo al que ha sido asignado.

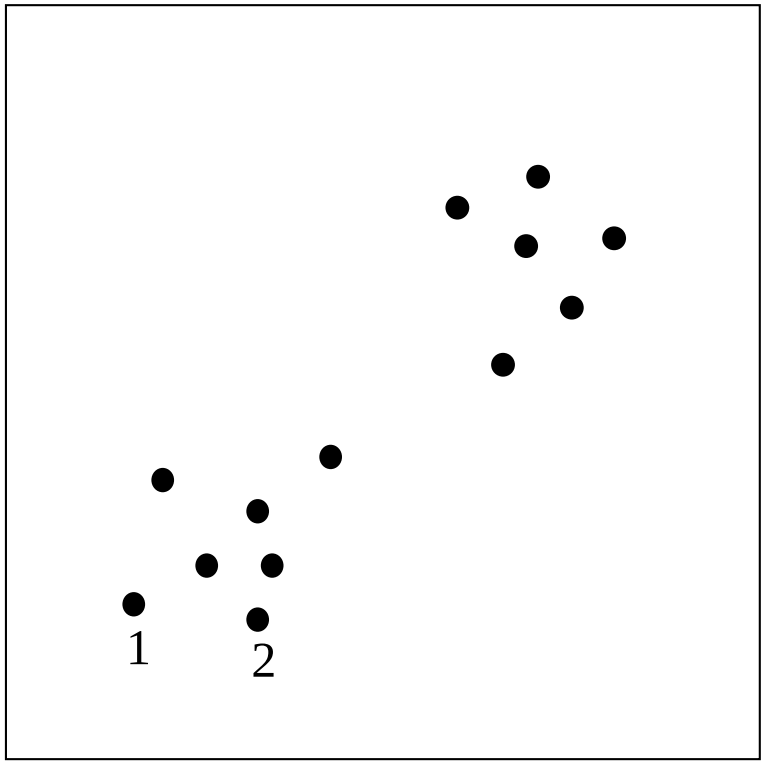
$k - means$: El procedimiento

1. Seleccionar el número de grupos que ha de tener la partición.
2. Elegir de forma aleatoria tantos líderes como grupos.
3. Calcular la distancia de todos los objetos a clasificar a cada uno de los líderes.
4. Asignar cada objeto a su líder más próximo.
5. Si no se han realizado cambios en la asignación a los líderes, o se ha pasado por este punto más del número fijado de veces ir al paso 8.
6. Cacular el centroide de cada grupo y declararlo nuevo líder del grupo.
7. Ir al paso 3.
8. La clasificación ha terminado: cada objeto pertenece al grupo al que ha sido asignado.

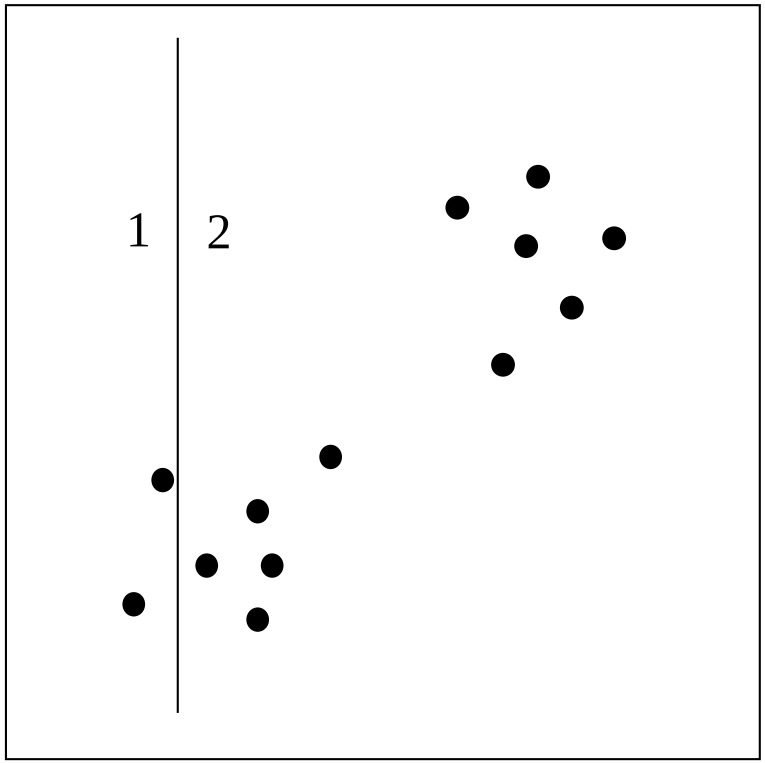
Las observaciones



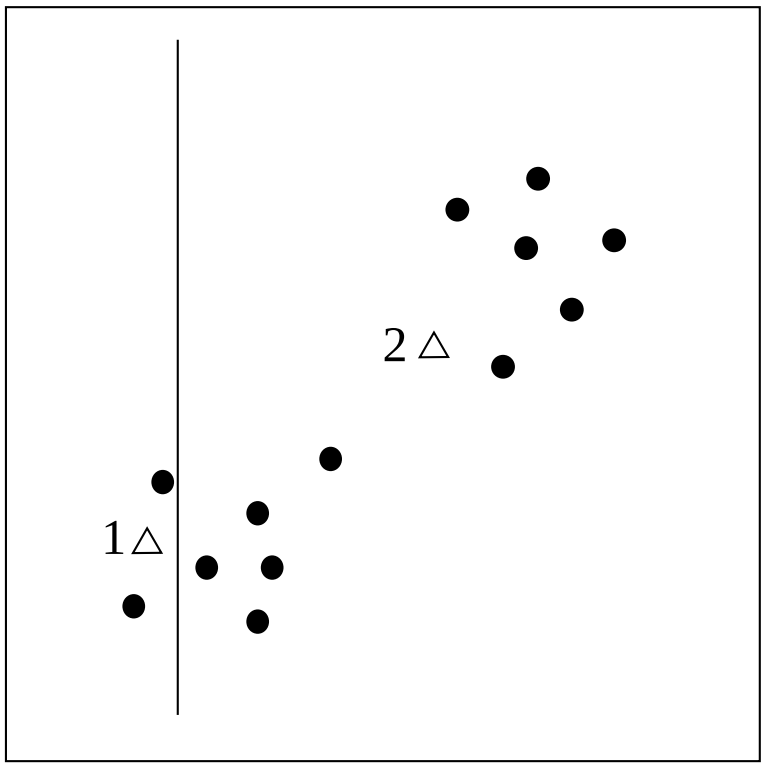
Los lideres



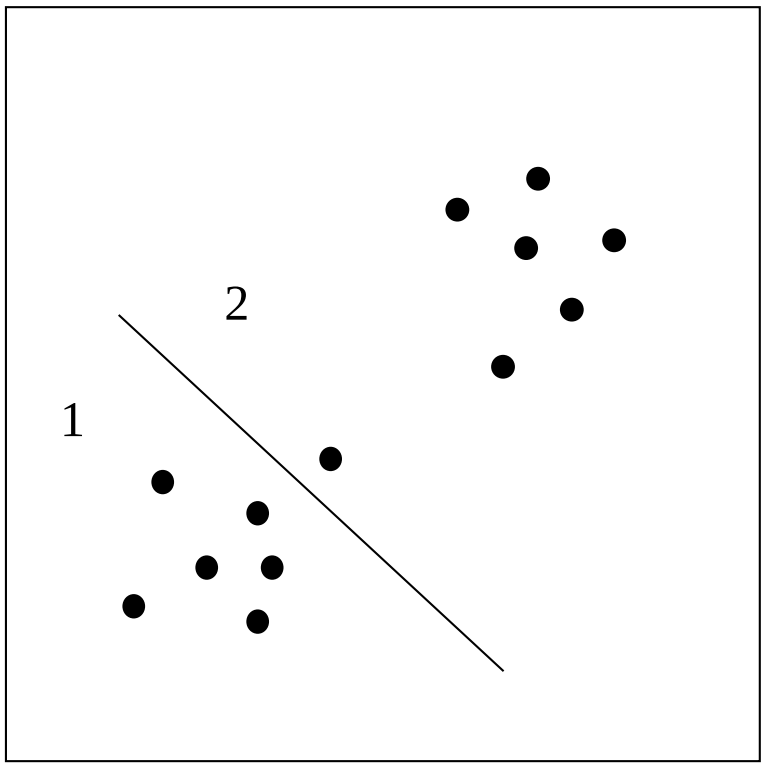
La primera partición



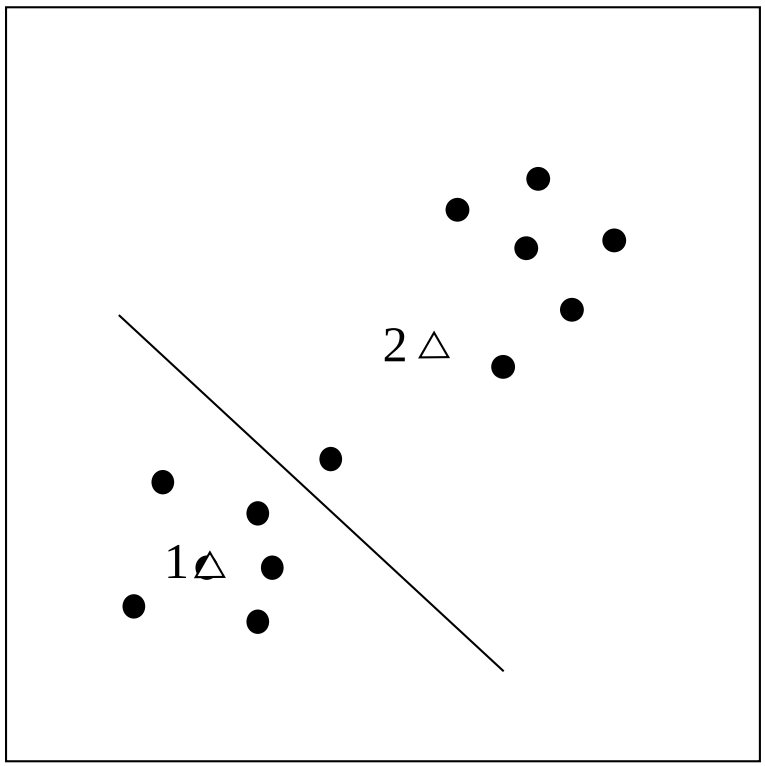
Los centroides de la 1ª partición



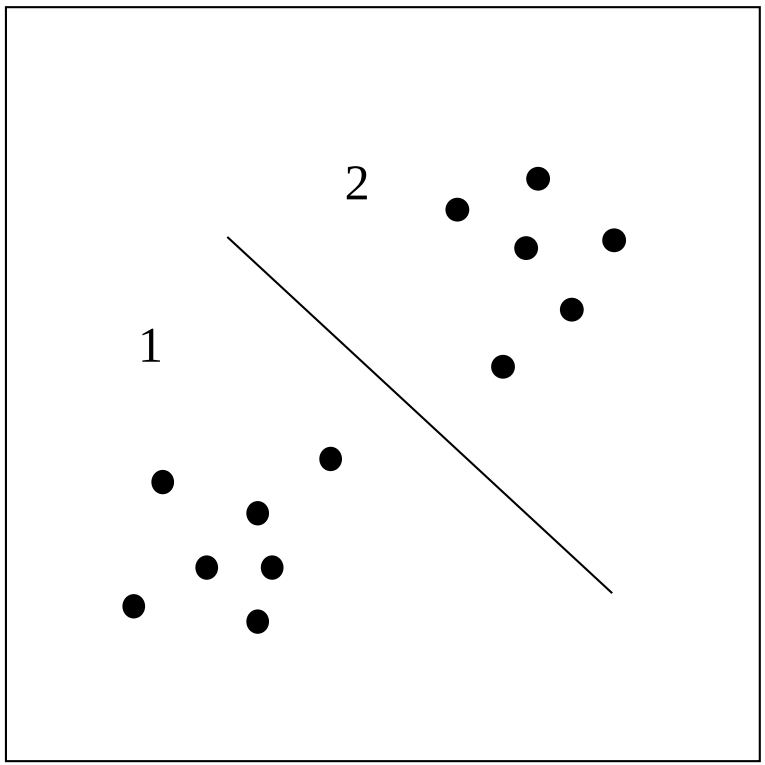
La segunda partición



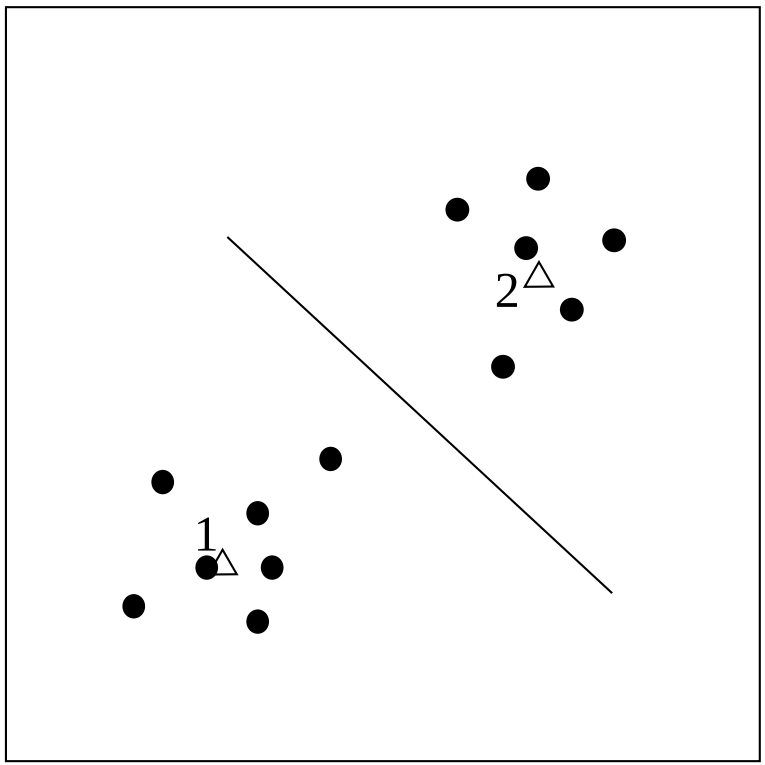
Los centroides de la 2^a partición



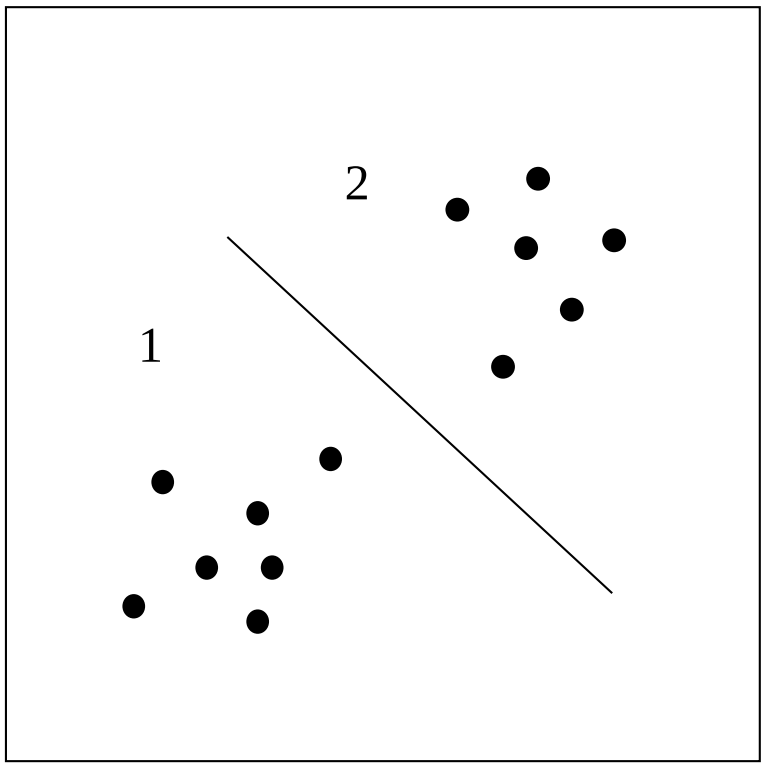
La tercera partición



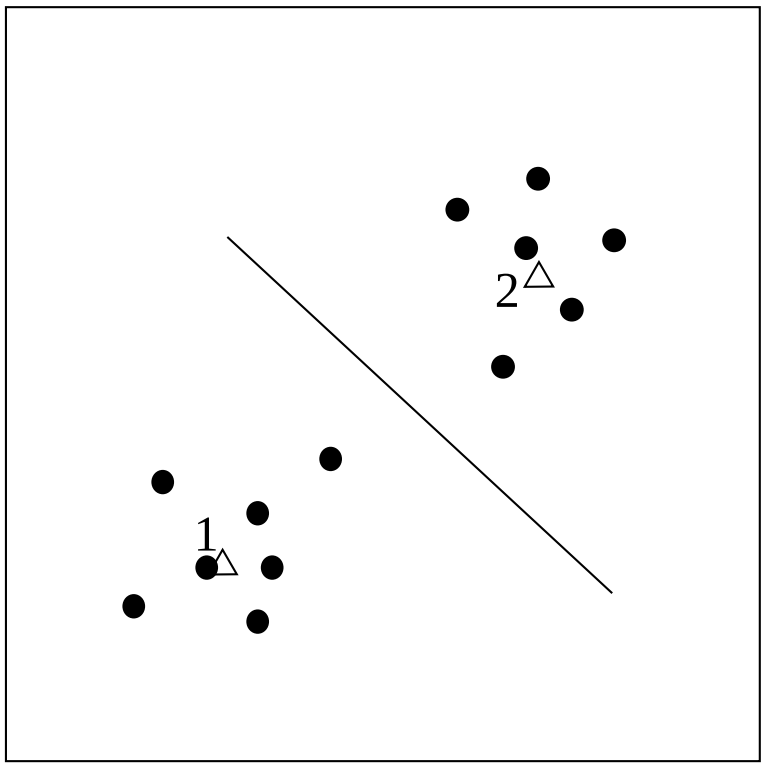
Los centroides de la 3ª partición



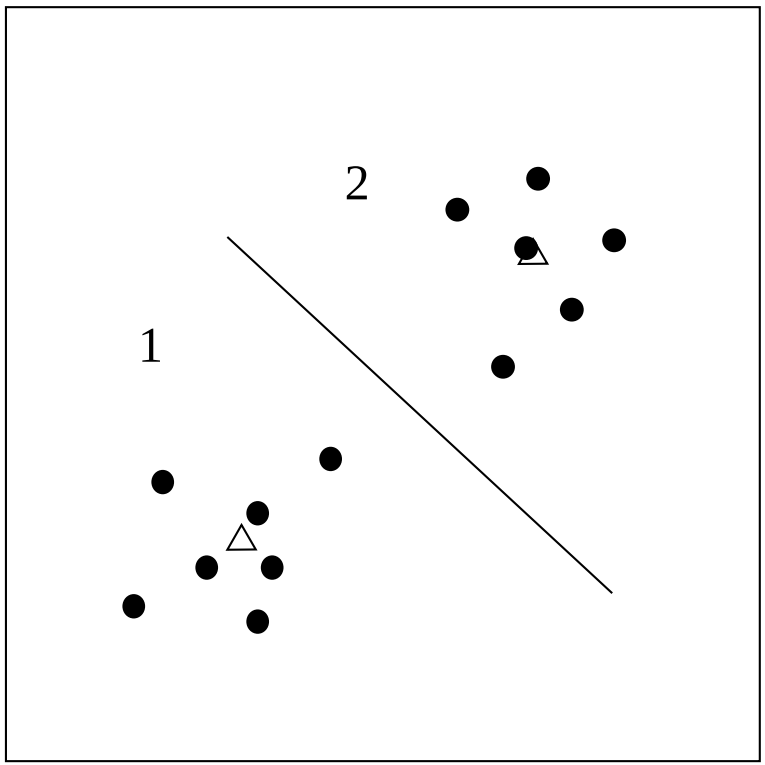
La cuarta partición



Los centroides de la 4^a partición



La partición definitiva



Archivo Editar Ver Terminal Solapas Ayuda

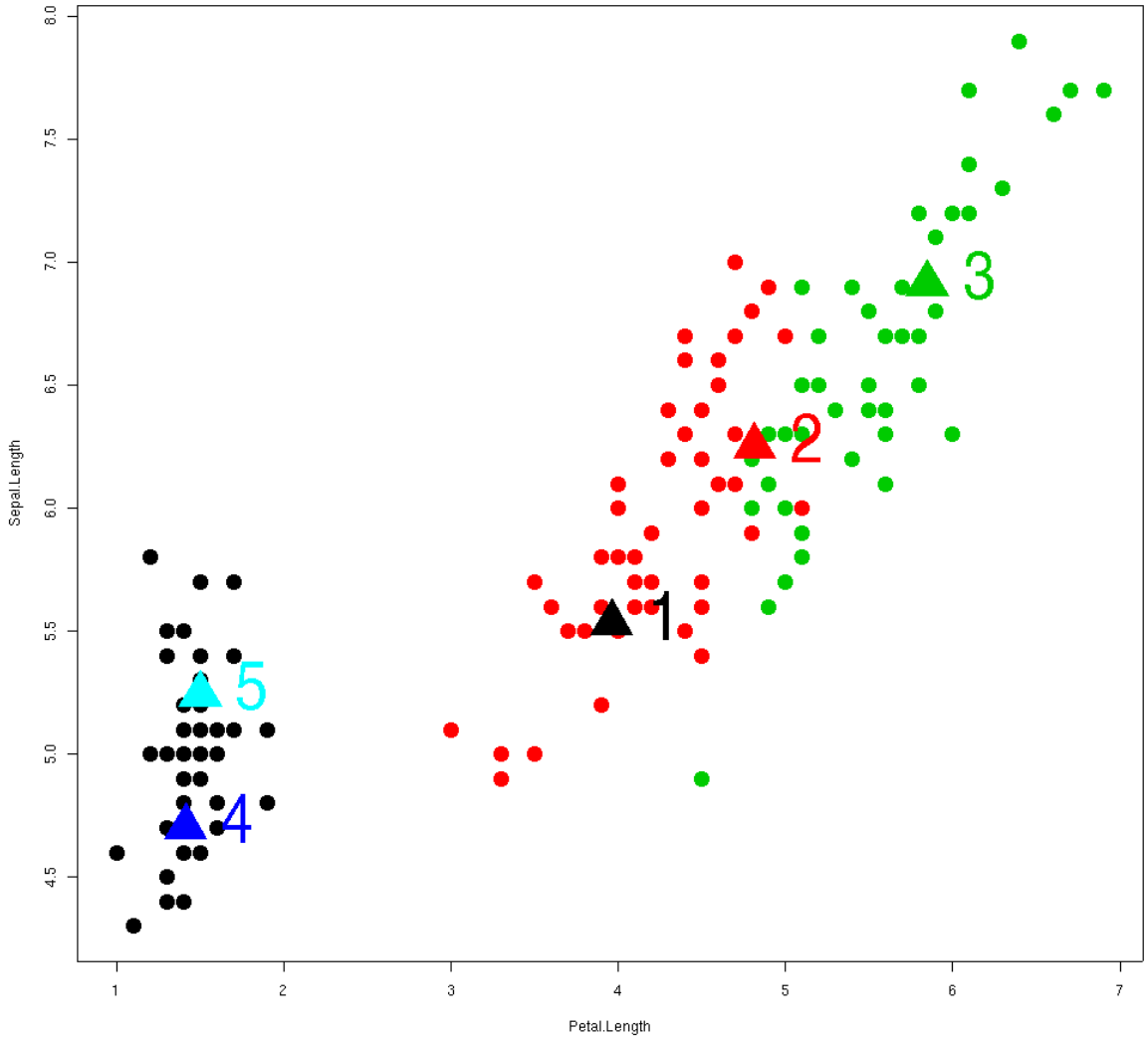
```
> kmeans(iris[,1:4],5)
$cluster
  [1] 1 5 5 5 1 1 5 1 5 5 1 5 5 5 1 1 1 1 1 1 1 1 5 1 5 5 1 1 1 5 5 1 1 1 5 5 1
 [38] 1 5 1 1 5 5 1 1 5 1 5 1 5 2 2 2 3 2 3 2 3 2 3 3 3 3 2 3 2 3 3 2 3 2 3 2 2
 [75] 2 2 2 2 2 3 3 3 3 2 3 2 2 2 3 3 3 2 3 3 3 3 3 2 3 3 4 2 4 2 4 4 3 4 4 4 2
[112] 2 4 2 2 2 2 4 4 2 4 2 4 2 4 4 2 2 4 4 4 4 4 2 2 4 4 2 2 4 4 4 2 4 4 4 2 2
[149] 2 2

$centers
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     5.242857    3.667857     1.500000   0.2821429
2     6.264444    2.884444     4.886667   1.6666667
3     5.532143    2.635714     3.960714   1.2285714
4     7.014815    3.096296     5.918519   2.1555556
5     4.704545    3.122727     1.413636   0.2000000

$withinss
[1]  4.630714 17.014222  9.749286 15.351111  3.114091

$size
[1] 28 45 28 27 22

>
> 
```


Sepal.Length
Petal.Length
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
1
2
3
4
5
6
7
1
2
3
4
5

Objetivos

- Realizar una clasificación a partir de una matriz de distancias
- Otener una jerarquía de agrupación que se representa mendiante un dedrograma
- Determinar el número de grupos:
  - A priori: criterio de distancia
  - A posteriori: en relación a la estructura del dendrograma

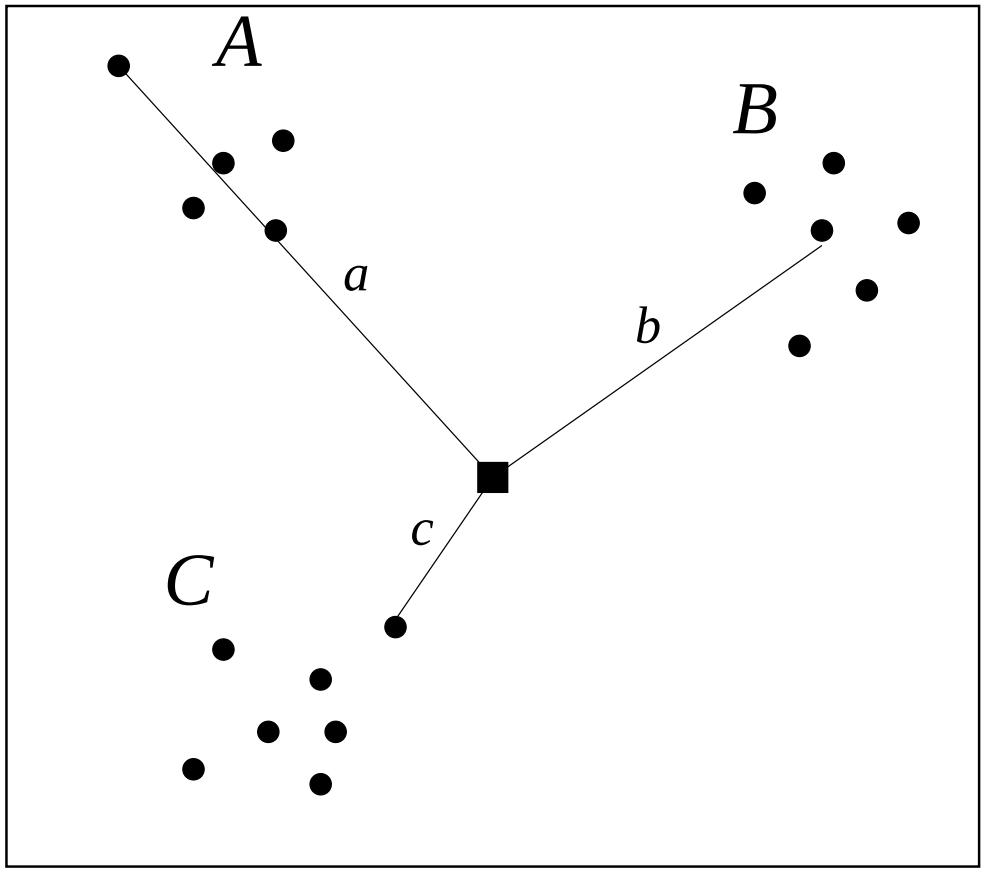
Clasificación primer paso: matriz de distancias

	1	2	3	4	5
1	0.00				
2	1.00	0.00			
3	2.83	2.24	0.00		
4	7.07	6.40	4.24	0.00	
5	7.62	7.28	5.10	2.83	0.00

Clasificación jerárquica

	1	2	3	4	5
1	0.00				
2	1.00	0.00			
3	2.83	**2.24**	0.00		
4	7.07	**6.40**	4.24	0.00	
5	7.62	**7.28**	5.10	2.83	0.00

Criterios de agregación



Algunos criterios de agregación:

- *Single linkage or nearest neighbor method*: distancia al vecino más próximo
- *Complete linkage or diameter method*: distancia al vecino más alejado
- *McQuitty's or WPGMA method: weighted pair group method using median average*
- *Average linkage (group average or UPGMA) method*: distancia media

Algunos criterios de agregación:

- *Median (Gower's or WPGMC) method: weighted pair group method using median centroid*
- *Centroid or UPGMC method: unweighted pair group method using median centroid*
- *Ward's min. variance or error sum of squares method*: Crirterio de mínima varianza de Ward

Clasificación según el vecino más próximo

	1,2	3	4	5
1,2	0.00			
3	2.24	0.00		
4	6.40	4.24	0.00	
5	7.28	5.10	2.83	0.00

Clasificación según el vecino más próximo

	1,2	3	4	5
1,2	0.00			
3	2.24	0.00		
4	6.40	**4.24**	0.00	
5	7.28	**5.10**	2.83	0.00

Clasificación según el vecino más próximo

	1,2,3	4	5
1,2,3	0.00		
4	4.24	0.00	
5	5.10	2.83	0.00

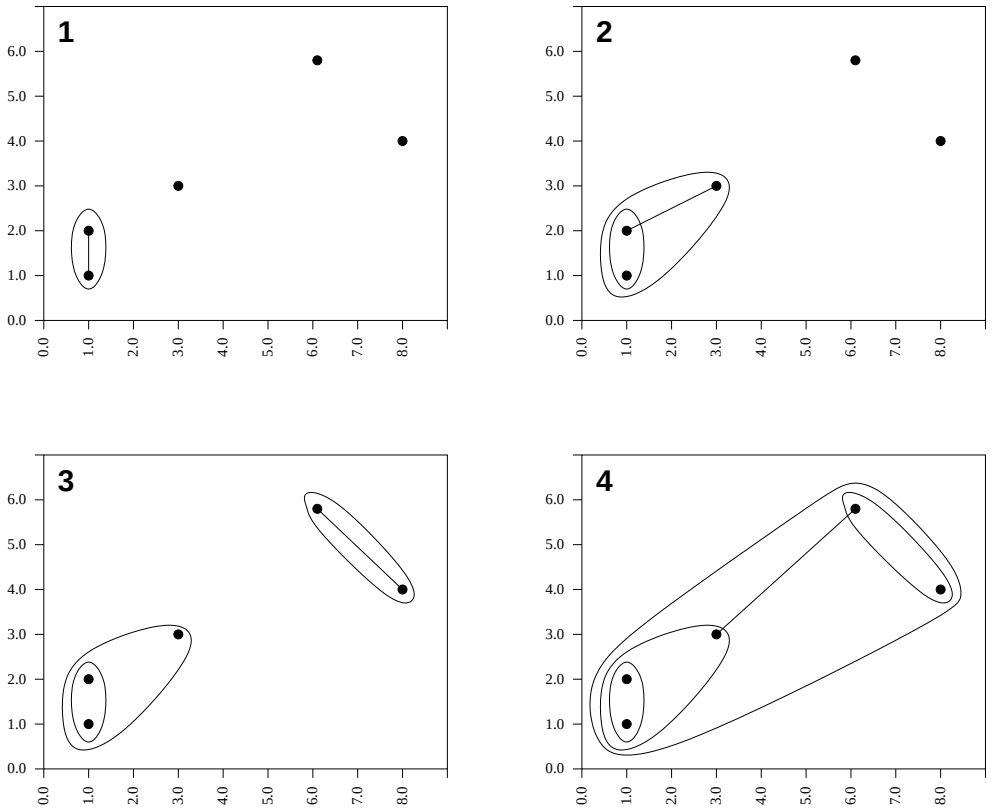
Clasificación según el vecino más próximo

	1,2,3	4	5
1,2,3	0.00		
4	**4.24**	0.00	
5	5.10	2.83	0.00

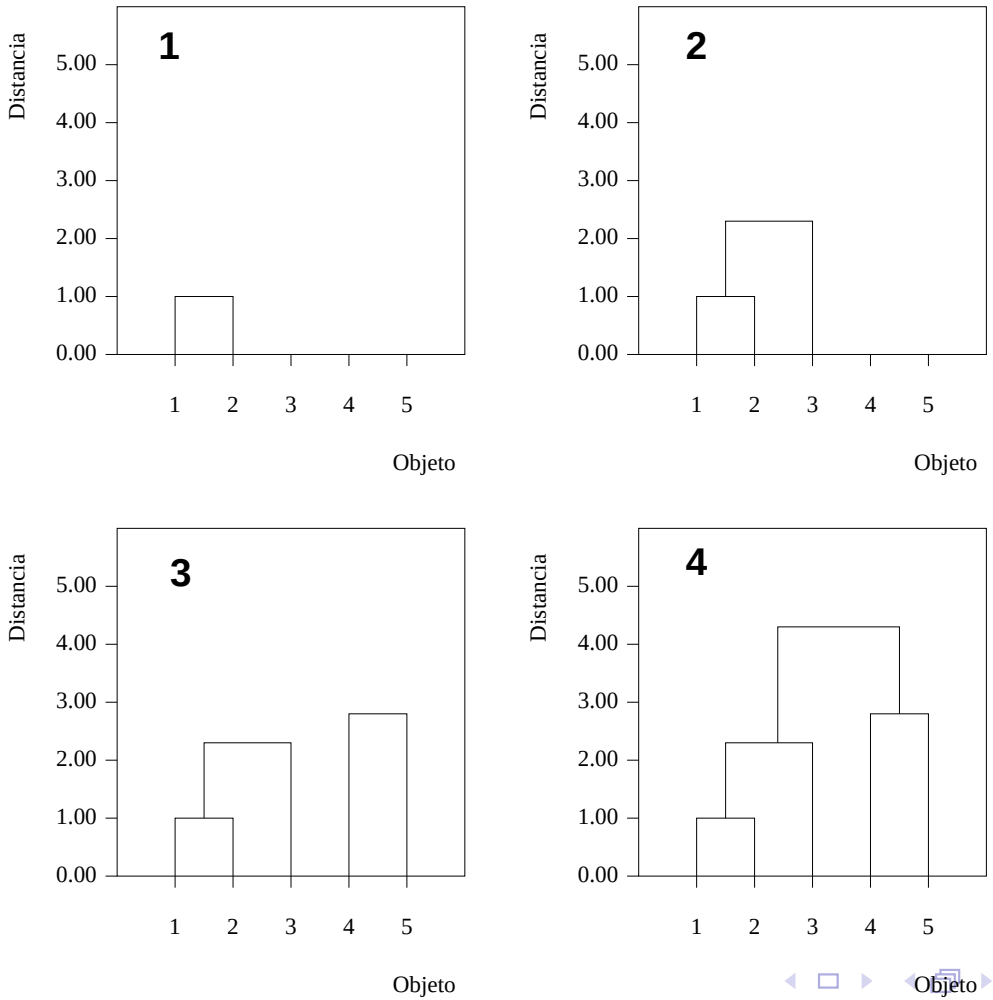
Clasificación según el vecino más próximo

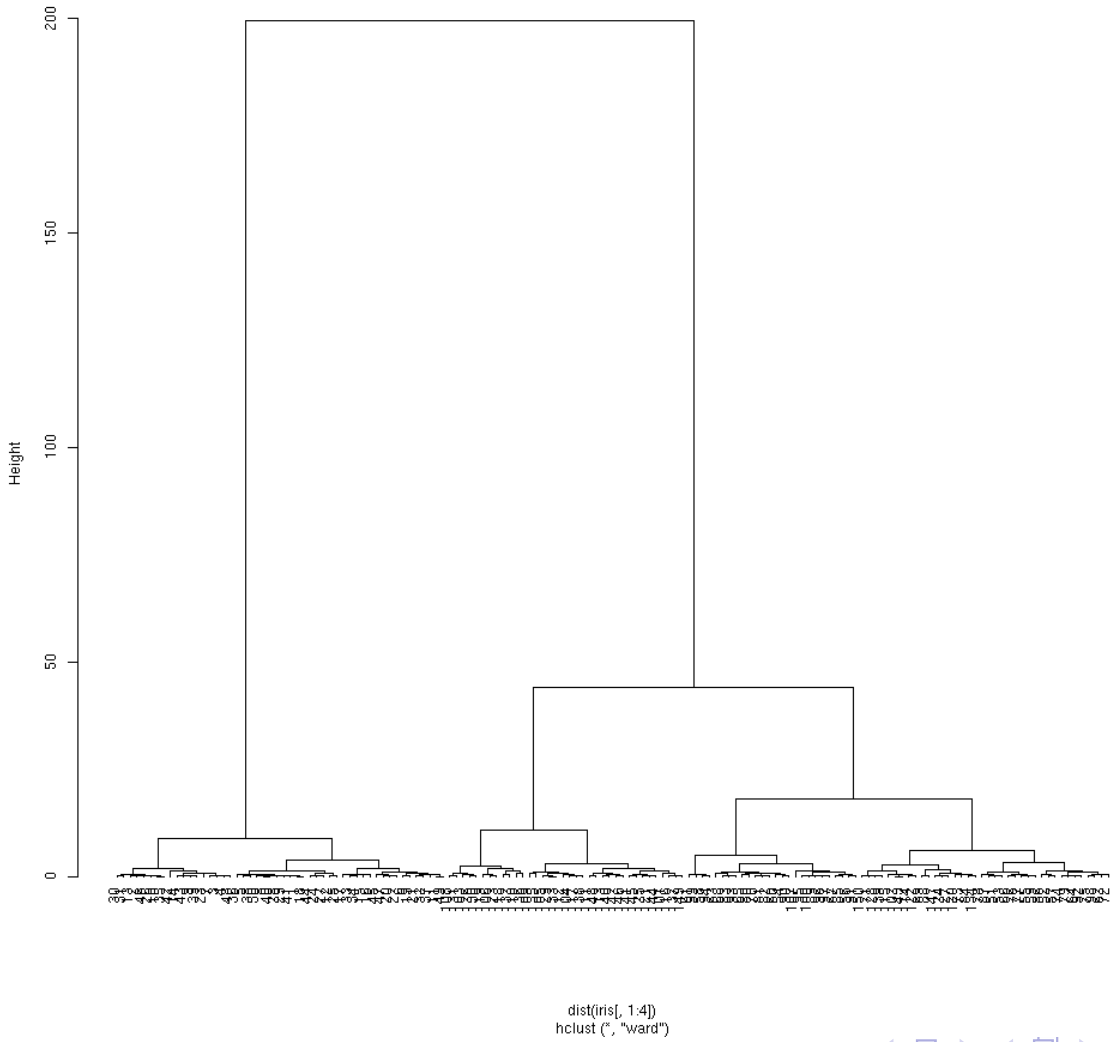
	1,2,3	4,5
1,2,3	0.00	
4,5	4.24	0.00

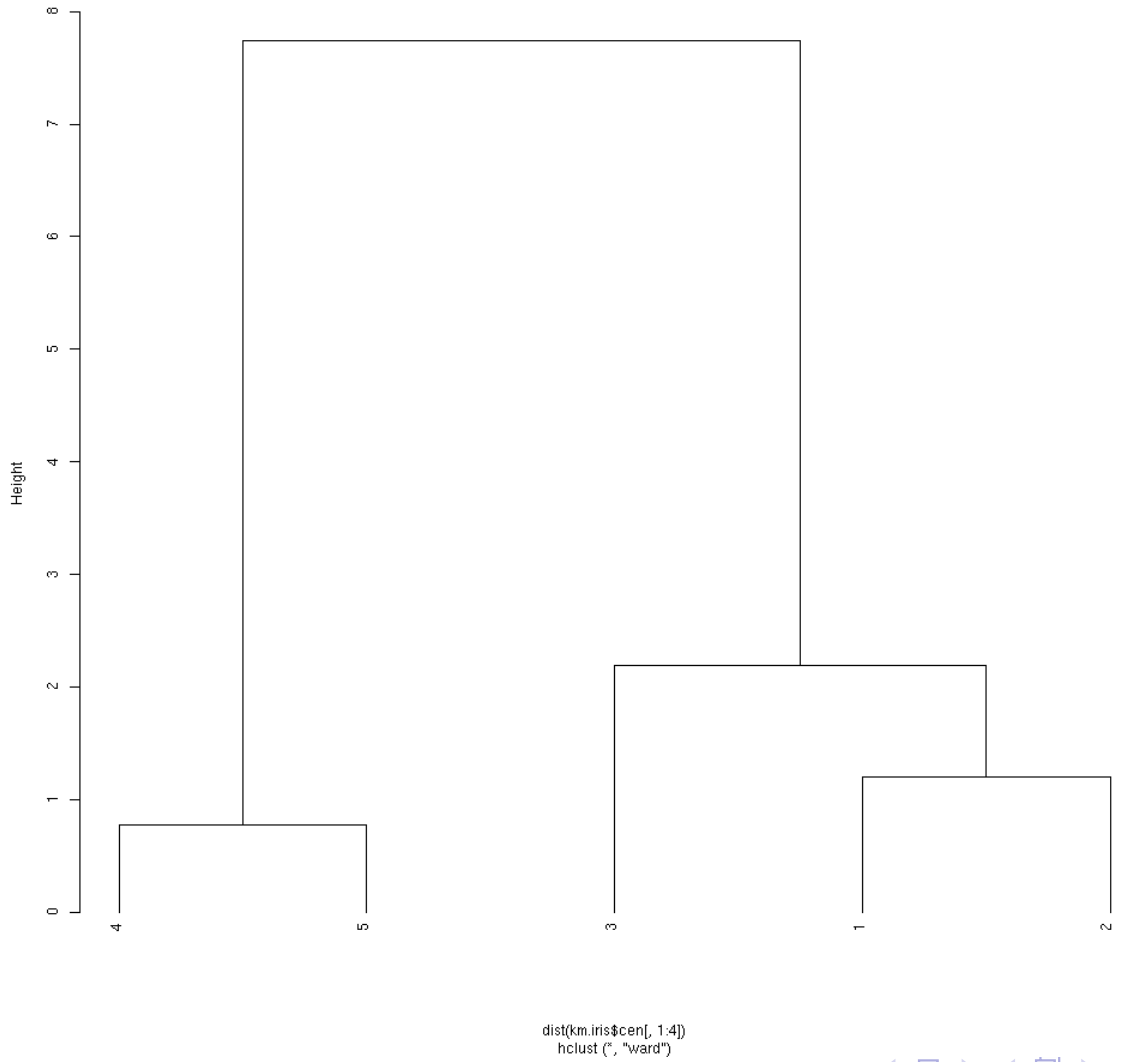
Clasificación mediante el criterio de agregación del vecino más próximo



Construcción del dendrograma




Height
0
50
100
150
200
dist(iris[, 1:4])
hclust (*, "ward")


Height
0
1
2
3
4
5
6
7
8
4
5
3
1
2
dist(km.iris$cen[, 1:4])
hclust (*, "ward")

Archivo Editar Ver Terminal Solapas Ayuda

```
> str(hclust(dist(km.iris$cen[,1:4]),method="ward"))
List of 7
 $ merge      : int [1:4, 1:2] -4 -1 -3 1 -5 -2 2 3
 $ height     : num [1:4] 0.775 1.206 2.195 7.745
 $ order      : int [1:5] 4 5 3 1 2
 $ labels     : chr [1:5] "1" "2" "3" "4" ...
 $ method     : chr "ward"
 $ call       : language hclust(d = dist(km.iris$cen[, 1:4]), method = "ward")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
> hcl.iris$mer
     [,1] [,2]
[1,]   -4   -5
[2,]   -1   -2
[3,]   -3    2
[4,]    1    3
> hcl.iris$he
[1] 0.775340 1.205803 2.195435 7.745059
>
>
>
>
>
>
```