Universidad Nacional Autónoma de México
Facultad de Ingeniería

Clustering Jerárquico

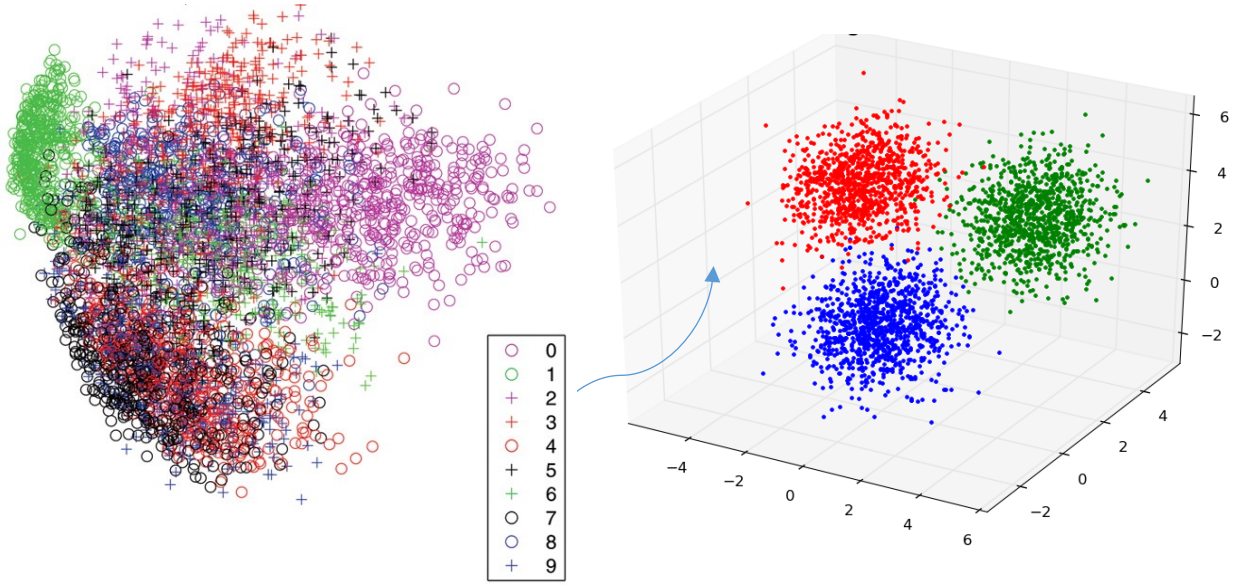
Guillermo Molero-Castillo
guillermo.molero@ingenieria.unam.edu

Noviembre, 2020

Clustering

Contexto

- La **Inteligencia Artificial** aplicada a la definición de cluster consiste en la segmentación y delimitación de grupos de elementos, que son unidos por características comunes que éstos comparten (aprendizaje no supervisado).
- El objetivo es dividir una población heterogénea de datos en un número de grupos naturales (regiones o segmentos homogéneos), de acuerdo a la **similitud de sus elementos**.

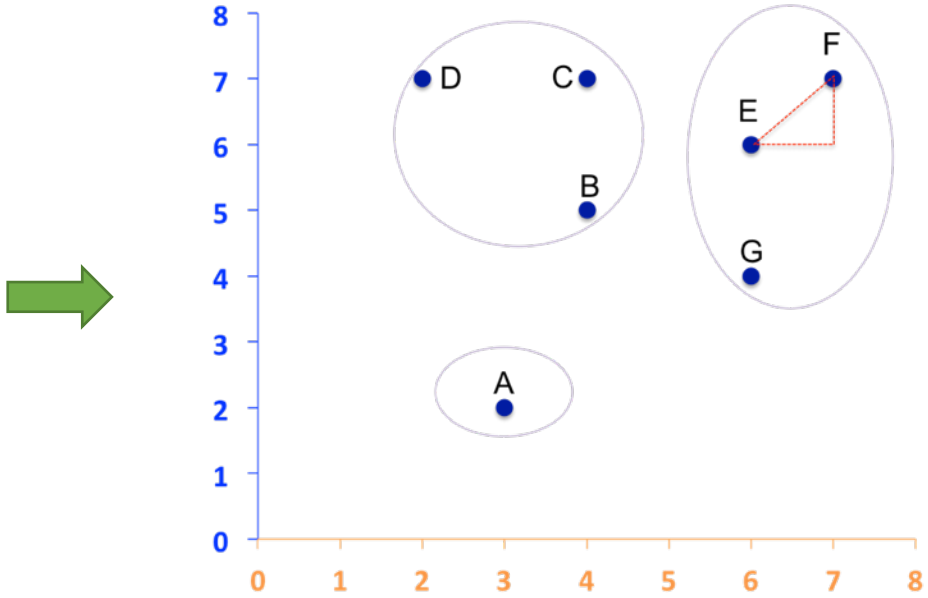


Los grupos nacen a partir de los datos y se descubren una serie de patrones ocultos en éstos.

Contexto

- Para hacer **clustering** es necesario saber el **grado de similitud** entre los elementos.
- La forma de hacer esto es utilizando las distancias.

Sujeto	Lealtad a la tienda (x)	Lealtad a la marca (y)
A	3	2
B	4	5
C	4	7
D	2	7
E	6	6
F	7	7
G	6	4

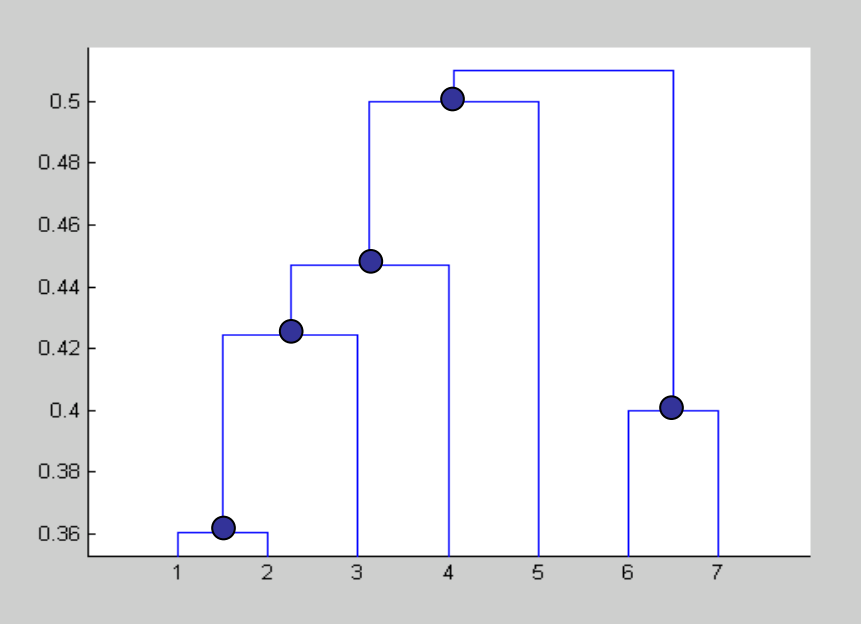


Métodos para hacer clustering

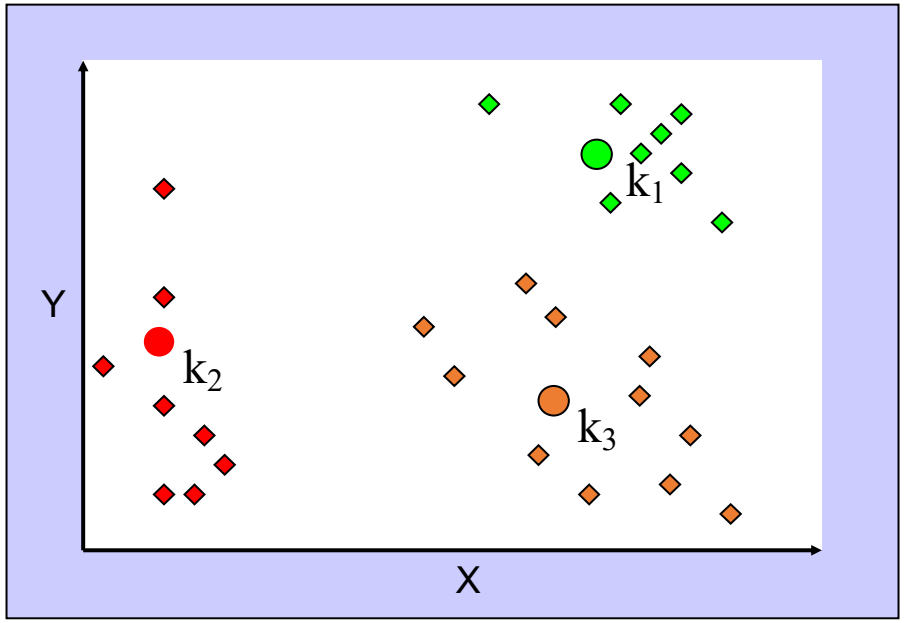
Jerárquico. Organiza los elementos, de manera recursiva, en una estructura en forma de árbol (dendrograma).

Particional. Organiza los registros dentro de k grupos.

Jerárquico



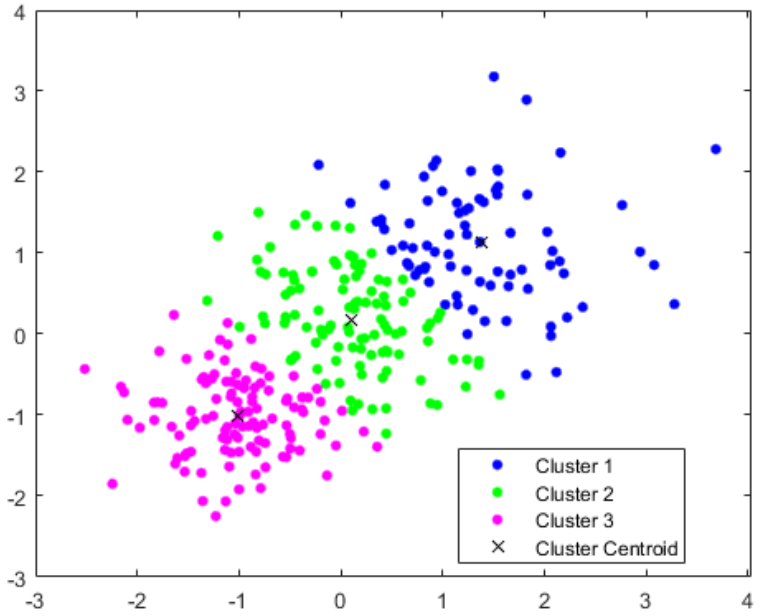
Particional



Métodos para hacer clustering

Jerárquico

Particional



Los **métodos particionales** tienen ventajas en aplicaciones que involucran gran cantidad de datos, para los cuales, la construcción de un árbol puede resultar complejo.

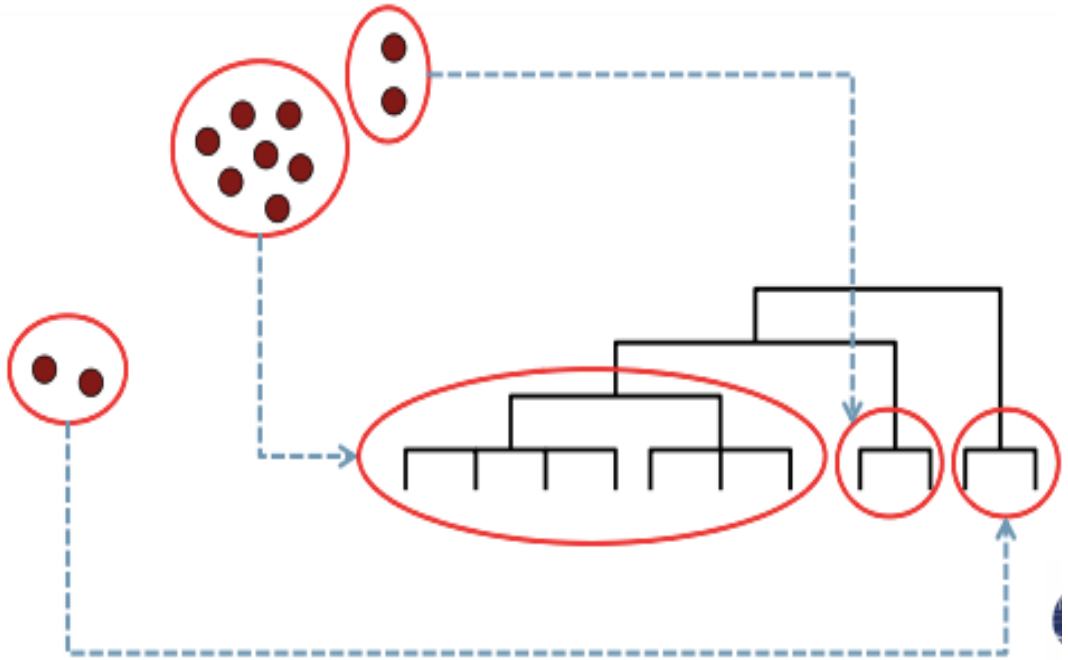
Aplicaciones

- **Marketing.** Para caracterizar y descubrir segmentos de clientes con fines de marketing.
- **Biología.** Para organizar diferentes especies de plantas y animales.
- **Bibliotecas.** Para agrupar libros a través de temas o autores.
- **Seguro.** Para reconocer a los clientes, sus pólizas e identificar los fraudes.
- **Urbanismo.** Para organizar tipos de viviendas y analizar sus valores en función de su ubicación geográfica.
- **Otras.** Estudios demográficos, regiones afectadas por terremotos, identificación de zonas peligrosas, regionalizaciones climáticas, comunidades de usuarios para los sistemas de recomendación, entre otros.

Clustering Jerárquico

Clustering Jerárquico

El algoritmo de **clustering jerárquico** construye un árbol que representa las relaciones de similitud entre los distintos elementos.



Pros

- Facilidad de manejo de los datos.
- No se asume un número particular de grupos.

Contras

- Una vez que se toma la decisión de combinar dos grupos, no se puede regresar atrás.
- Lento para grandes conjuntos de datos, $O(n^2 \log(n))$.

Pasos para formar grupos (clústeres)

Son tres los pasos necesarios:

1. Utilizar un método para medir la similitud de los elementos.
2. Utilizar un método para agrupar a los elementos.
3. Utilizar un método para decidir la cantidad adecuada de grupos.

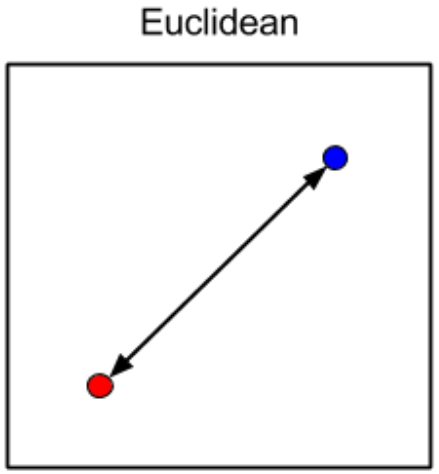
Métodos para medir la similitud

- Algunas métricas conocidas:

  - Distancia Euclidiana o Euclídea
  - Distancia de Chebyshev
  - Distancia de Manhattan o Geometría del taxista
  - Distancia de Minkowsky

Métodos para medir la similitud

Dimensiones:



$$dist(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

$$dist(p,q) = \sqrt{(p_1 - q_1)^2}$$ **1 dimensión**

$$dist(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$$ **2 dimensiones**

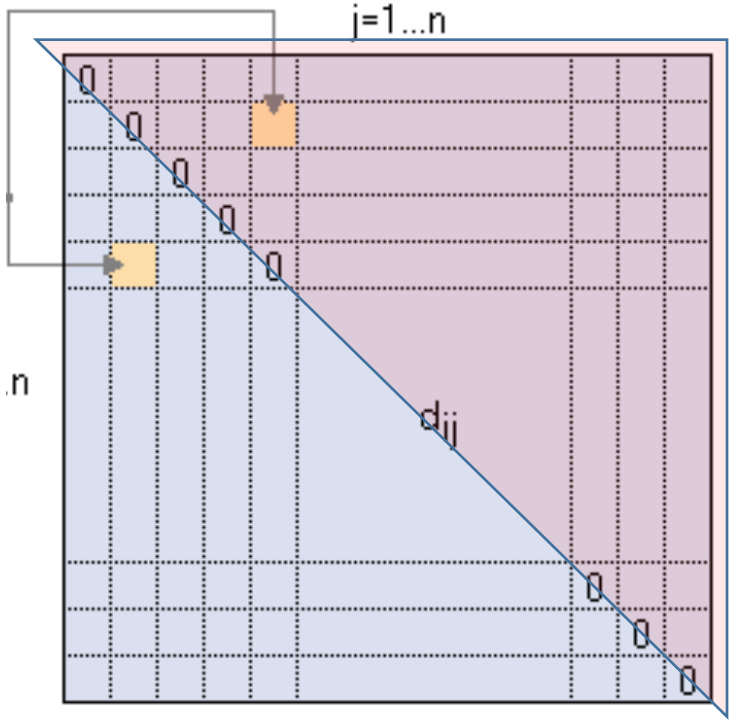
$$dist(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}$$ **3 dimensiones**

$$dist(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2 + \ldots + (p_n - q_n)^2}$$ ***n* dimensiones**

Métodos para medir la similitud

Matriz de similitudes

$$dist(p,q) = dij = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

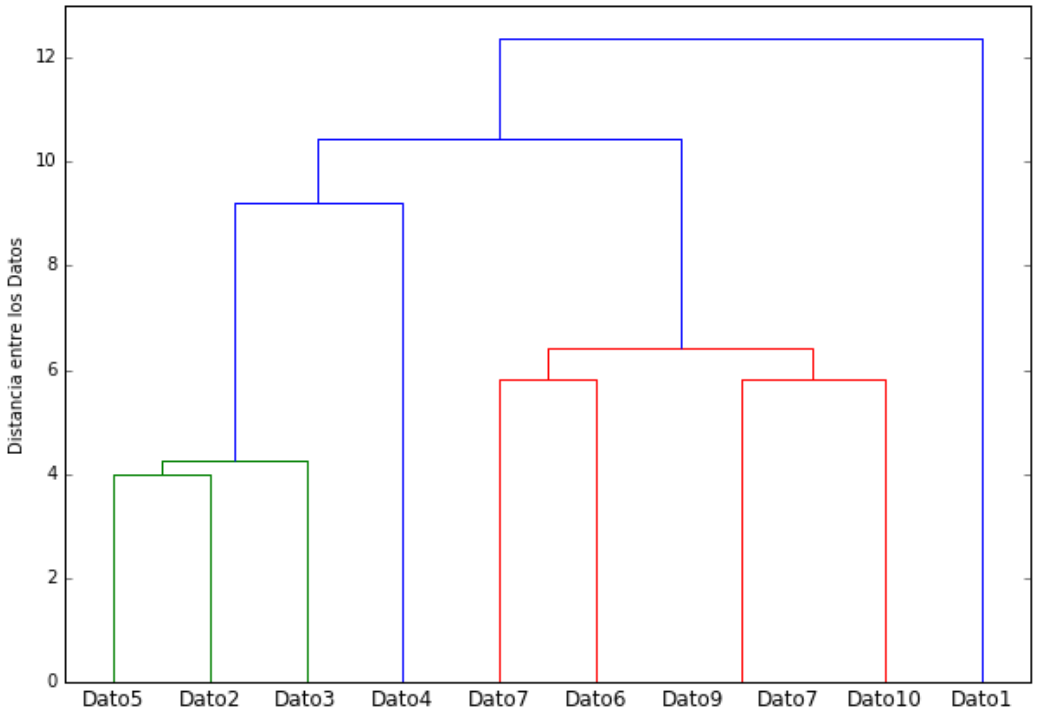


Algoritmo Ascendente Jerárquico

Algoritmo Ascendente Jerárquico

Consiste en agrupar en cada paso aquellos 2 elementos (cluster) más cercanos. De esta manera se va construyendo una estructura en forma de árbol.

El proceso concluye cuando se forma un único cluster (grupo).



Algoritmo Ascendente Jerárquico

Pseudocódigo

1 **Calcular** la matriz de similitud/distancias
2 **Inicialización:** Cada elemento, un cluster
3 **Repetir**
4 Combinar los dos clusters más cercanos
5 Actualizar la matriz de similitud/distancias
6 **Hasta** que sólo quede un cluster

Given:

A set X of objects $\{x_1,\ldots,x_n\}$

A distance function $dist(c_1,c_2)$

for $i = 1$ to n

$c_i = \{x_i\}$

end for

$C = \{c_1,\ldots,c_n\}$

$l = n+1$

while C.size > 1 **do**

- (c_{min1},c_{min2}) = minimum $dist(c_i,c_j)$ for all c_i,c_j in C
- remove c_{min1} and c_{min2} from C
- add $\{c_{min1},c_{min2}\}$ to C
- $l = l + 1$

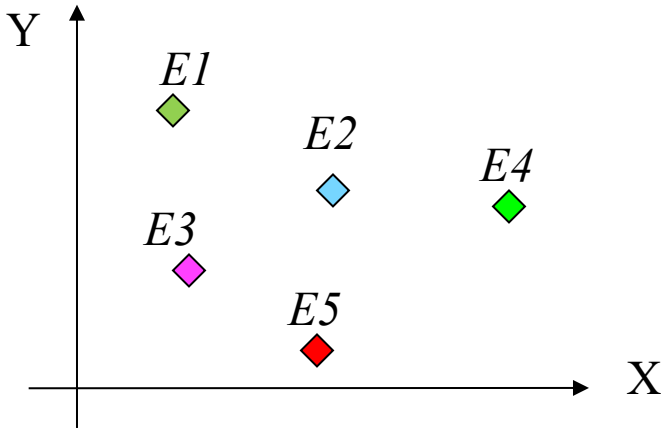
end while

Algoritmo Ascendente Jerárquico

Procedimiento

1 Construir una matriz de distancias entre los elementos (objetos). **Por ejemplo, euclidiana.**

	V1	V2	V3	…	V10
E1	1	4	0.5		3
E2	2	6	0.7		2
E3	4	4	0.4		4
E4	6	2	0.2		2
E5	7	2	0.2		2



Si son 5 elementos, la matriz de distancias será 5 x 5.

Distancia entre E2 y E1

$$\begin{bmatrix} dist(E1,E1) & dist(E1,E2) & \dots & dist(E1,E5) \\ dist(E2,E1) & dist(E2,E2) & \dots & dist(E2,E5) \\ dist(E3,E1) & dist(E3,E2) & \dots & dist(E3,E5) \\ dist(E4,E1) & dist(E4,E2) & \dots & dist(E4,E5) \\ dist(E5,E1) & dist(E5,E2) & \dots & dist(E5,E5) \end{bmatrix}$$

Algoritmo Ascendente Jerárquico

Procedimiento

1 Construir matriz de distancias entre los elementos (objetos).

	V1	V2	V3	…	V10
E1	1	4	0.5		3
E2	2	6	0.7		2
E3	4	4	0.4		4
E4	6	2	0.2		2
E5	7	2	0.2		2

$$dist(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

d	1	2	3	4	5
1	0	5	6	10	13
2	5	0	1	5	8
3	6	1	0	4	7
4	10	5	4	0	3
5	13	8	7	3	0

Algoritmo Ascendente Jerárquico

Procedimiento

2 Cada elemento representa un grupo (cluster).

Nota. Si hay un error en algún paso no se puede regresar atrás …

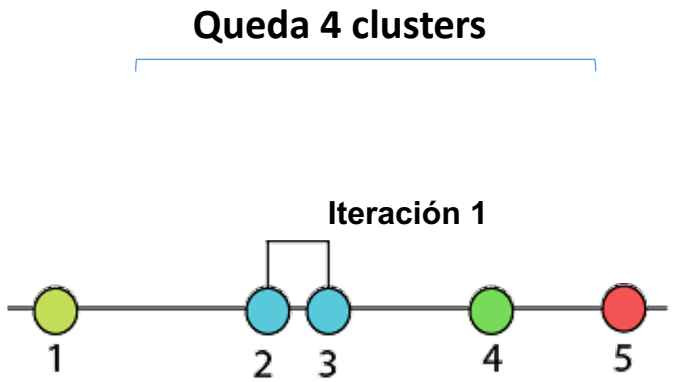
d	1	2	3	4	5
1	0	5	6	10	13
2		0	1	5	8
3			0	4	7
4				0	3
5					0

Algoritmo Ascendente Jerárquico

Procedimiento

3 Se encuentra el par más cercano de elementos (grupos) y se forma un único grupo.

- Menor distancia entre 2 objetos (grupos).



d	1	2	3	4	5
1	0	5	6	10	13
2		0	1	5	8
3			0	4	7
4				0	3
5					0

Algoritmo Ascendente Jerárquico

Procedimiento

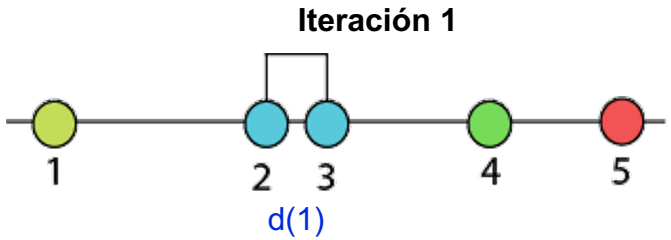
4 Se actualizan las distancias entre el nuevo grupo y los grupos anteriores.

d	1	2	3	4	5
1	0	5	6	10	13
2		0	1	5	8
3			0	4	7
4				0	3
5					0

Iteración 1

Se promedian las nuevas distancias

d	1	(2-3)	4	5
1	0	5.5	10	13
(2-3)		0	4.5	7.5
4			0	3
5				0



Algoritmo Ascendente Jerárquico

Procedimiento

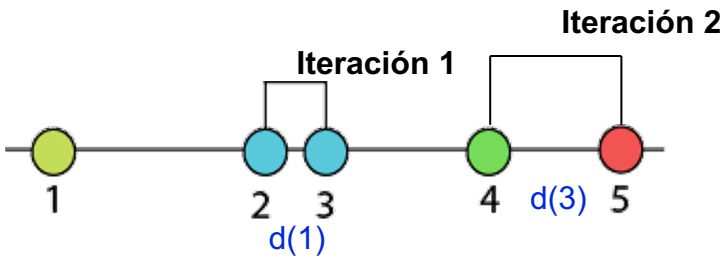
5 Se repiten los **pasos 3 y 4** hasta que todos los elementos se agrupen en un solo cluster.

d	1	(2-3)	4	5
1	0	5.5	10	13
(2-3)		0	4.5	7.5
4			0	3
5				0

Iteración 2

Se promedian las nuevas distancias

d	1	(2-3)	(4-5)
1	0	5.5	11.5
(2-3)		0	6
(4-5)			0



Algoritmo Ascendente Jerárquico

Procedimiento

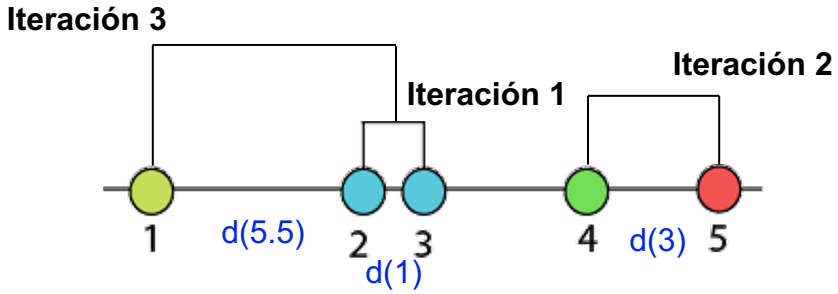
5 Se repiten los **pasos 3 y 4** hasta que todos los elementos se agrupen en un solo cluster.

d	1	(2-3)	(4-5)
1	0	5.5	11.5
(2-3)		0	6
(4-5)			0

Iteración 3

Se promedian las nuevas distancias

d	(1-2-3)	(4-5)
(1-2-3)	0	8.75
(4-5)		0



Algoritmo Ascendente Jerárquico

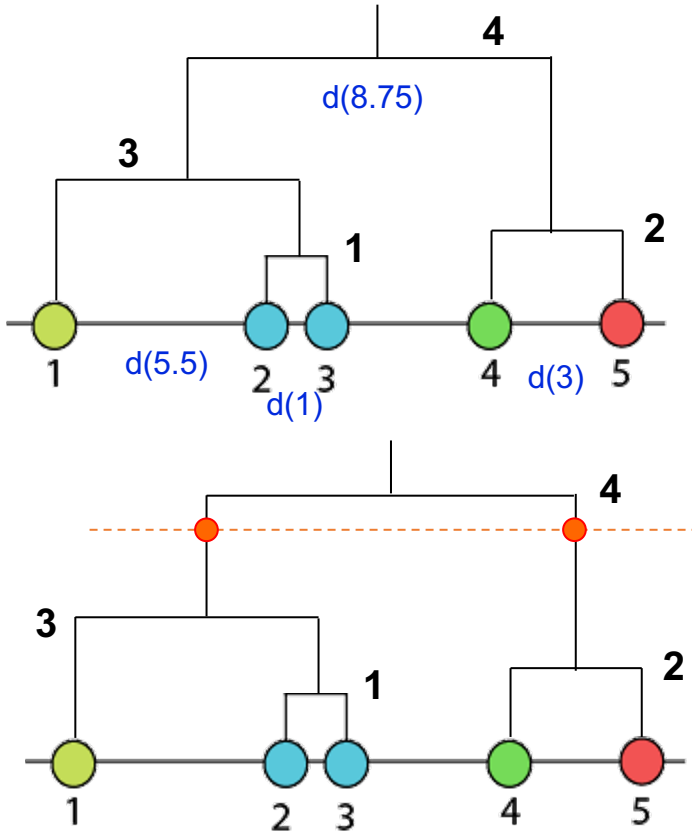
Procedimiento

5 Se repiten los **pasos 3 y 4** hasta que todos los elementos se agrupen en un solo cluster.

d	(1-2-3)	(4-5)
(1-2-3)	0	8.75
(4-5)		0

Iteración 4

d	(1-2-3-4-5)
(1-2-3-4-5)	0



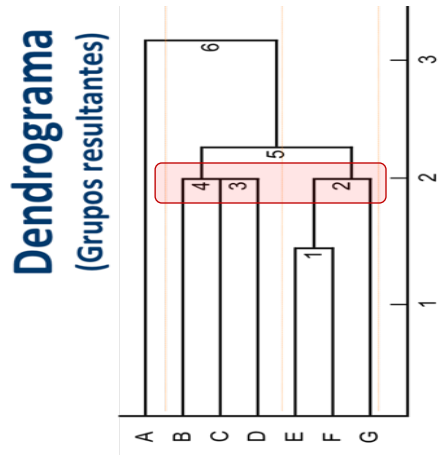
Algoritmo Ascendente Jerárquico

Tipos de representación (árbol)

Sujetos	A	B	C	D	E	F	G
A	---						
B	3.16	---					
C	5.10	2.00	---				
D	5.10	2.83	2.00	---			
E	5.00	2.24	2.24	4.12	---		
F	6.40	3.61	3.00	5.00	1.41	---	
G	3.61	2.24	3.61	5.00	2.00	3.16	---

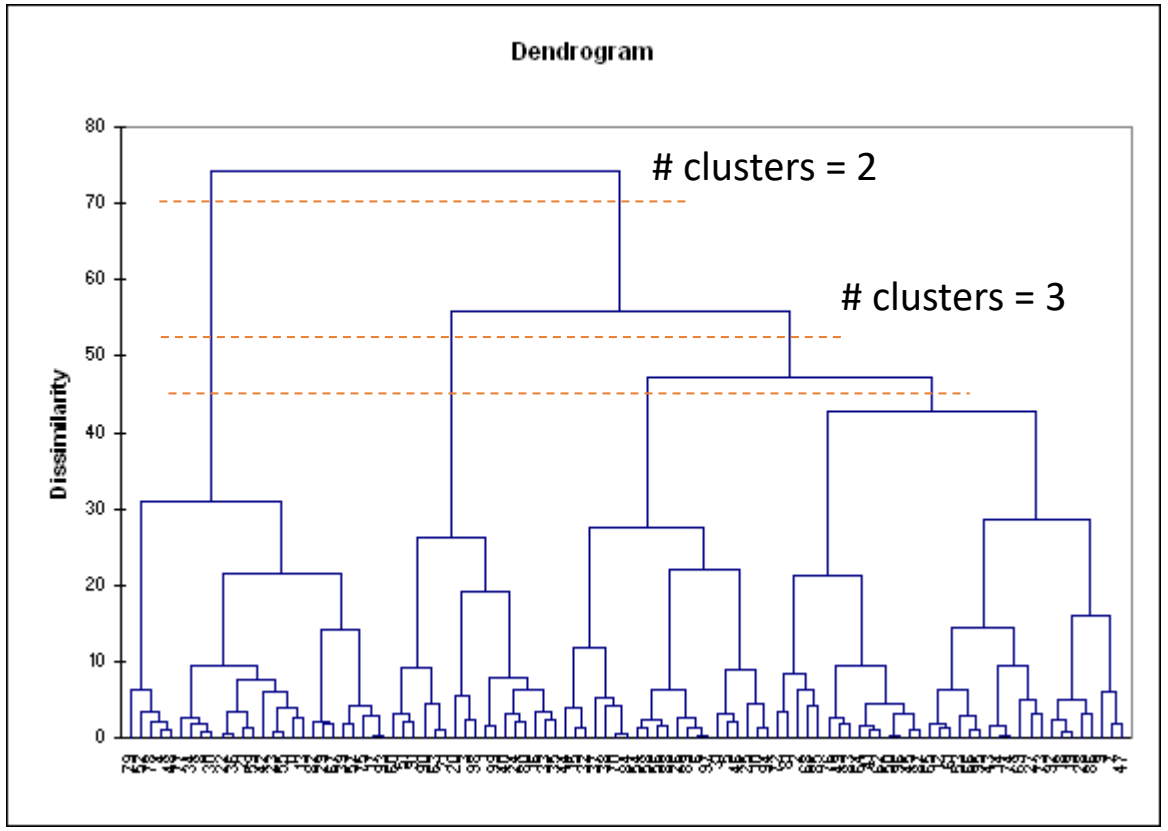
Paso	Distancia mínima entre sujetos	Sujetos
	Solución inicial	
1	1.414	E-F
2	2.000	E-G
3	2.000	C-D
4	2.000	B-C
5	2.236	B-E
6	3.162	A-B

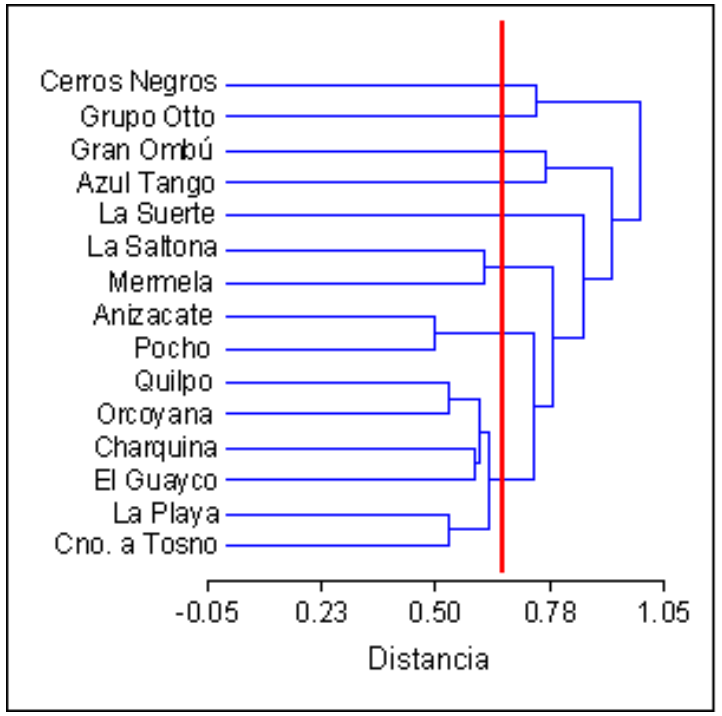
...



Algoritmo Ascendente Jerárquico

Método para decidir la cantidad de grupos





Práctica 5
Clustering Jerárquico

-Distancias Euclidianas-

Práctica

```
DatosEmp <- read.table("/Users/guille/Documents/1 FI-UNAM/1 Cursos/2021-1/1 IA2021-1/2 CasosPracticos/3 Similitudes/Empleados.txt", header=T, sep="\t")

DatosEmp
```

	ID	Salario	Casado	Coche	Hijos	Vivienda	Sindicato	FaltasAno	Antiguedad	Sexo
1	E1	10000	1	0	0	0	0	7	15	1
2	E2	20000	0	1	1	0	1	3	3	0
3	E3	15000	1	1	2	1	1	5	10	1
4	E4	30000	1	1	1	0	0	15	7	0
5	E5	10000	1	1	0	1	1	1	6	1
6	E6	40000	0	1	0	0	1	3	16	0
7	E7	25000	0	0	0	0	1	0	8	1
8	E8	20000	0	1	0	1	1	2	6	0
9	E9	20000	1	1	3	1	0	7	5	1
10	E10	30000	1	1	2	1	0	1	20	1
11	E11	45000	0	0	0	0	0	2	12	0
12	E12	8000	1	1	2	1	0	3	1	1
13	E13	20000	0	0	0	0	0	27	5	0
14	E14	10000	0	1	0	0	1	0	7	1
15	E15	8000	0	1	0	0	0	3	2	1

1 Obtención de una matriz de distancias

```
DistEuclidiana <- round(dist(DatosEmp[2:10], method = "euclidean"), 2)

DistEuclidiana
```

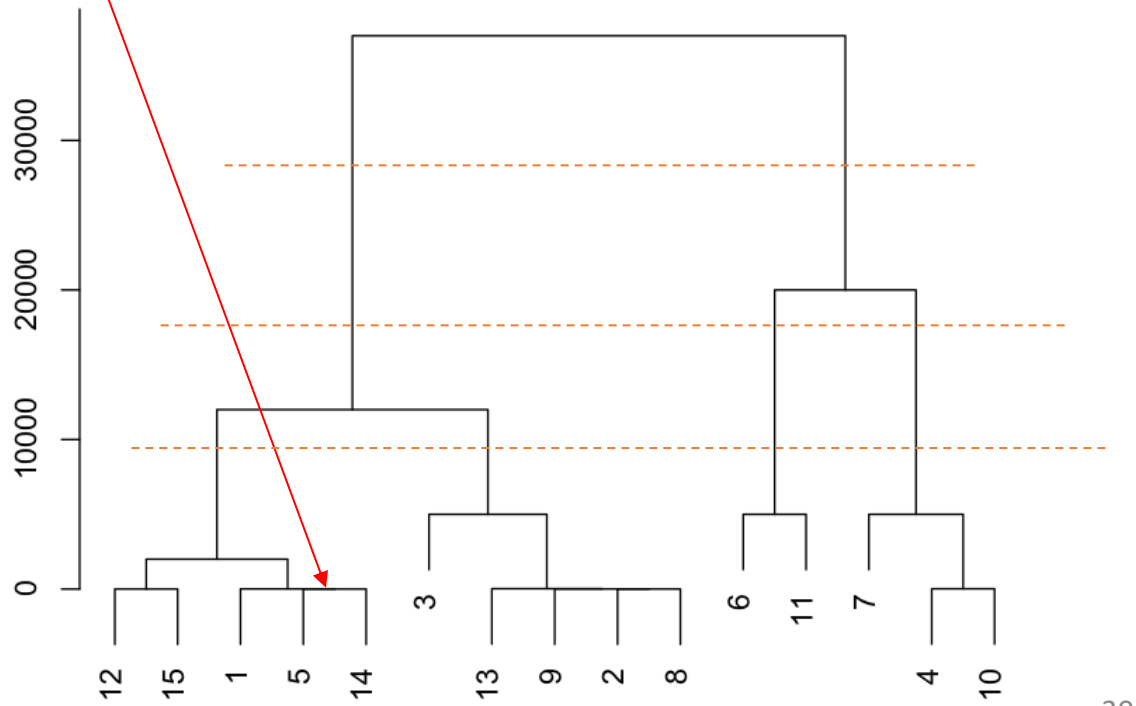
```
          1        2        3        4        5        6        7        8        9       10       11       12       13       14
2  10000.01
3   5000.00  5000.01
4  20000.00 10000.01 15000.00
5     10.95 10000.00  5000.00 20000.01
6  30000.00 20000.00 25000.00 10000.01 30000.00
7  15000.00  5000.00 10000.00  5000.02 15000.00 15000.00
8  10000.01     3.46  5000.00 10000.01 10000.00 20000.00  5000.00
9  10000.01     5.29  5000.00 10000.00 10000.00 20000.00  5000.01     6.16
10 20000.00 10000.01 15000.00    19.18 20000.01 10000.00  5000.02 10000.01 10000.01
11 35000.00 25000.00 30000.00 15000.01 35000.00  5000.00 20000.00 25000.00 25000.00 15000.00
12  2000.05 12000.00  7000.01 22000.00  2000.01 32000.00 17000.00 12000.00 12000.00 22000.01 37000.00
13 10000.03    24.15  5000.05 10000.01 10000.03 20000.02  5000.07    25.08    20.32 10000.05 25000.01 12000.02
14    10.77 10000.00  5000.00 20000.01     2.00 30000.00 15000.00 10000.00 10000.00 20000.00 35000.00  2000.01 10000.04
15  2000.05 12000.00  7000.01 22000.00  2000.01 32000.00 17000.00 12000.00 12000.00 22000.01 37000.00     2.65 12000.02  2000.01
```

2 Obtención de grupos

```
Jerarquico <- hclust(DistEuclidiana)

plot(Jerarquico)
```

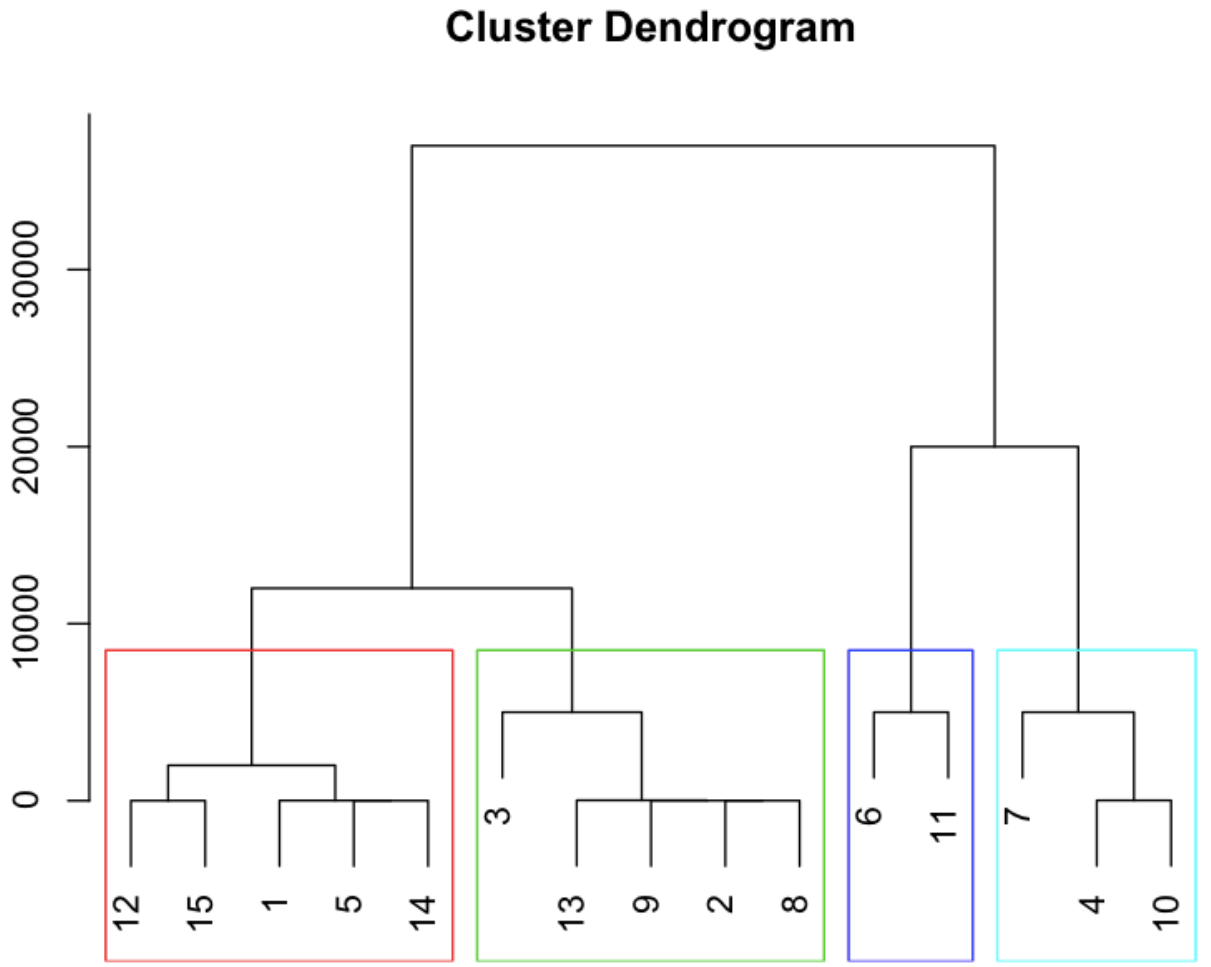
```
         1        2        3        4        5        6        7        8        9       10       11       12       13       14
2  10000.01
3   5000.00  5000.01
4  20000.00 10000.01 15000.00
5     10.95 10000.00  5000.00 20000.01
6  30000.00 20000.00 25000.00 10000.01 30000.00
7  15000.00  5000.00 10000.00  5000.02 15000.00 15000.00
8  10000.01     3.46  5000.00 10000.01 10000.00 20000.00  5000.00
9  10000.01     5.29  5000.00 10000.00 10000.00 20000.00  5000.01     6.16
10 20000.00 10000.01 15000.00    19.18 20000.01 10000.00  5000.02 10000.01 10000.01
11 35000.00 25000.00 30000.00 15000.01 35000.00  5000.00 20000.00 25000.00 25000.00 15000.00
12  2000.05 12000.00  7000.01 22000.00  2000.01 32000.00 17000.00 12000.00 12000.00 22000.01 37000.00
13 10000.03    24.15  5000.05 10000.01 10000.03 20000.02  5000.07    25.08    20.32 10000.05 25000.01 12000.02
14    10.77 10000.00  5000.00 20000.01     2.00 30000.00 15000.00 10000.00 10000.00 20000.00 35000.00  2000.01 10000.04
15  2000.05 12000.00  7000.01 22000.00  2000.01 32000.00 17000.00 12000.00 12000.00 22000.01 37000.00     2.65 12000.02  2000.01
```

	ID	Salario	Casado	Coche	Hijos	Vivienda	Sindicato	FaltasAno	Antiguedad	Sexo
1	E1	10000	1	0	0	0	0	7	15	1
2	E2	20000	0	1	1	0	1	3	3	0
3	E3	15000	1	1	2	1	1	5	10	1
4	E4	30000	1	1	1	0	0	15	7	0
5	E5	10000	1	1	0	1	1	1	6	1
6	E6	40000	0	1	0	0	1	3	16	0
7	E7	25000	0	0	0	0	1	0	8	1
8	E8	20000	0	1	0	1	1	2	6	0
9	E9	20000	1	1	3	1	0	7	5	1
10	E10	30000	1	1	2	1	0	1	20	1
11	E11	45000	0	0	0	0	0	2	12	0
12	E12	8000	1	1	2	1	0	3	1	1
13	E13	20000	0	0	0	0	0	27	5	0
14	E14	10000	0	1	0	0	1	0	7	1
15	E15	8000	0	1	0	0	0	3	2	1



3 Definición de grupos

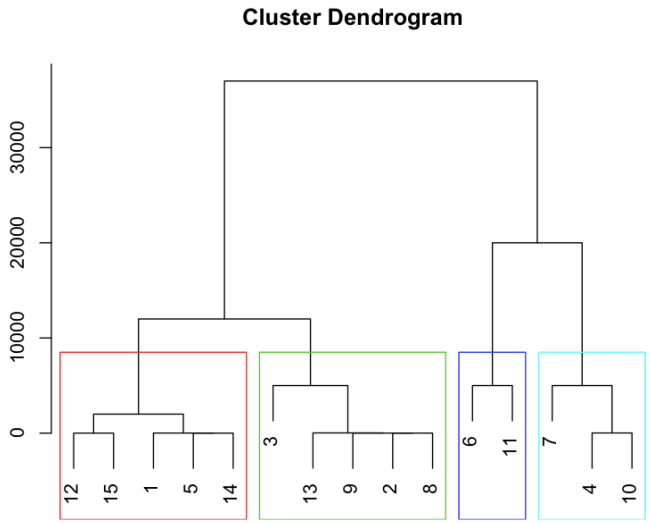
```
plot(Jerarquico)
rect.hclust(Jerarquico, k = 4, border = 2:6)
```




```
Clusters <- cutree(Jerarquico, 4)

Clusters
[1] 1 2 2 3 1 4 3 2 2 3 4 1 2 1 1

table(Clusters)

1 2 3 4
5 5 3 2
```

3 Definición de grupos

```
require(rattle)
Centros <- centers.hclust(DatosEmp [2:10], Jerarquico, nclust=4)
Centros
```



	Salario	Casado	Coche	Hijos	Vivienda	Sindicato	FaltasAnno	Antiguedad	Sexo
[1,]	9200.00	0.6000000	0.8000000	0.4	0.4000000	0.4000000	2.800000	6.20000	1.0000000
[2,]	19000.00	0.4000000	0.8000000	1.2	0.6000000	0.6000000	8.800000	5.80000	0.4000000
[3,]	28333.33	0.6666667	0.6666667	1.0	0.3333333	0.3333333	5.333333	11.66667	0.6666667
[4,]	42500.00	0.0000000	0.5000000	0.0	0.0000000	0.5000000	2.500000	14.00000	0.0000000

4 Interpretación

```
      Salario    Casado     Coche Hijos  Vivienda Sindicato FaltasAnno Antiguedad      Sexo
[1,]  9200.00 0.6000000 0.8000000   0.4 0.4000000 0.4000000   2.800000    6.20000 1.0000000
[2,] 19000.00 0.4000000 0.8000000   1.2 0.6000000 0.6000000   8.800000    5.80000 0.4000000
[3,] 28333.33 0.6666667 0.6666667   1.0 0.3333333 0.3333333   5.333333   11.66667 0.6666667
[4,] 42500.00 0.0000000 0.5000000   0.0 0.0000000 0.5000000   2.500000   14.00000 0.0000000
```

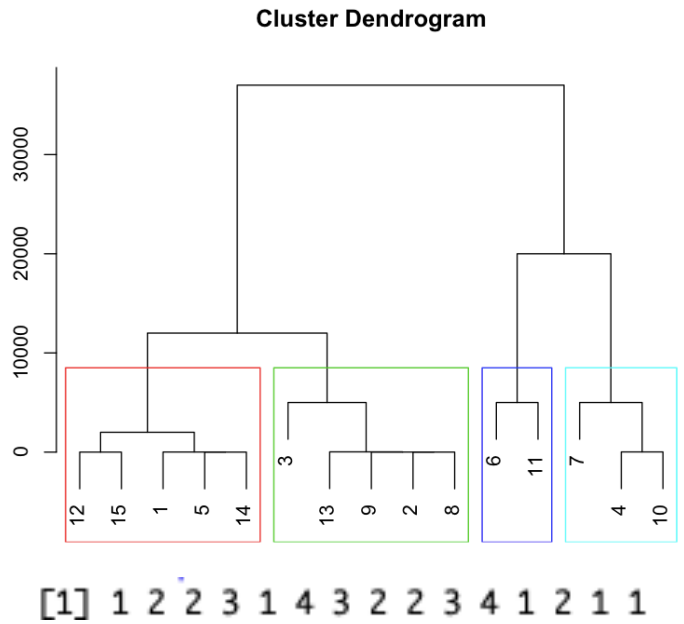
Cluster 1: 5 empleados

Salario :	9200
Casado :	Si = 0.6 / No = 0.4
Coche :	Si = 0.8 / No = 0.2
Hijos :	0.4
Vivienda :	Prop = 0.4 Alquiler = 0.6
Sindicato :	Si = 0.4 / No = 0.6
Faltas/Año :	2.8 (3)
Antigüedad :	6.2 (6)
Sexo :	M = 1

Cluster 2: 5 empleados

Salario :	19000
Casado :	Si = 0.4 / No = 0.6
Coche :	Si = 0.8 / No = 0.2
Hijos :	1.2
Vivienda :	Prop = 0.6 Alquiler = 0.4
Sindicato :	Si = 0.6 / No = 0.4
Faltas/Año :	8.8 (9)
Antigüedad :	5.8 (6)
Sexo :	M = 0.4 / F = 0.6

...



- **Cluster 1 [5 elementos –1, 5, 12, 14, 15–]**. Empleados con salario promedio de $9200, casados en su mayoría (60%), con coche (80%) y casi sin hijos. No tienen vivienda propia en su mayoría (60%), no sindicalizados en su mayoría (60%), con algunas faltas al año (3), una antigüedad promedio de 6 años y todos varones (100%).

Práctica 5
Clustering Jerárquico

-Distancias Chebyshev-

1 Obtención de una matriz de distancias (Chebyshev)

```
DistChebyshev <- dist(DatosEmp[2:10], method = "maximum")

DistChebyshev
```

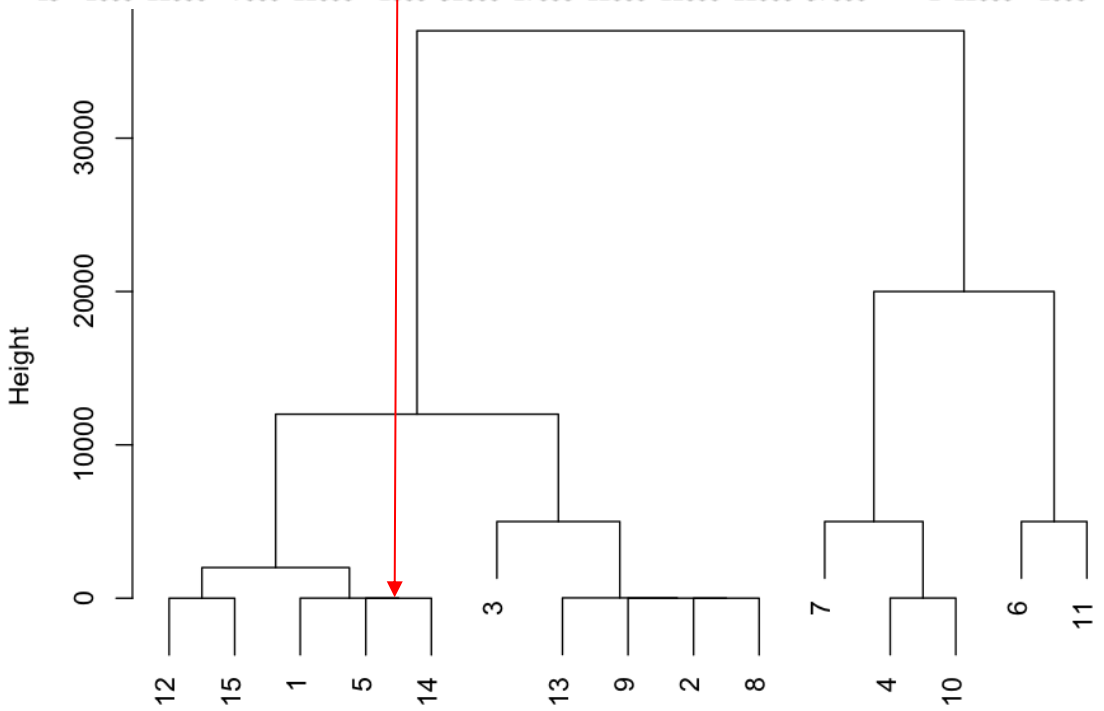
```
      1     2     3     4     5     6     7     8     9    10    11    12    13    14
2  10000
3   5000  5000
4  20000 10000 15000
5      9 10000  5000 20000
6  30000 20000 25000 10000 30000
7  15000  5000 10000  5000 15000 15000
8  10000     3  5000 10000 10000 20000  5000
9  10000     4  5000 10000 10000 20000  5000     5
10 20000 10000 15000    14 20000 10000  5000 10000 10000
11 35000 25000 30000 15000 35000  5000 20000 25000 25000 15000
12  2000 12000  7000 22000  2000 32000 17000 12000 12000 22000 37000
13 10000    24  5000 10000 10000 20000  5000    25    20 10000 25000 12000
14     8 10000  5000 20000     1 30000 15000 10000 10000 20000 35000  2000 10000
15  2000 12000  7000 22000  2000 32000 17000 12000 12000 22000 37000     2 12000  2000
```

2 Obtención de grupos

```
Jerarquico <- hclust(DistChebyshev)

plot(Jerarquico)
```

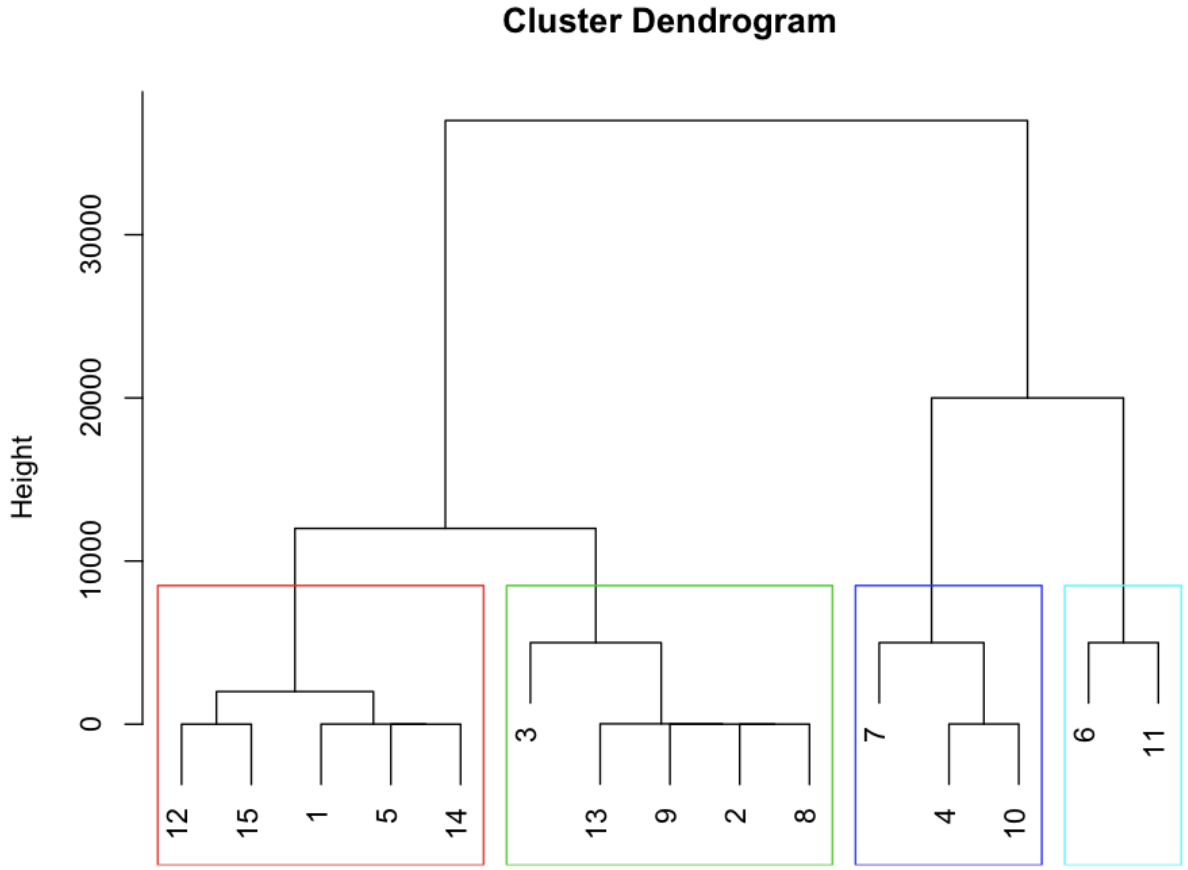
	ID	Salario	Casado	Coche	Hijos	Vivienda	Sindicato	FaltasAno	Antiguedad	Sexo
1	E1	10000	1	0	0	0	0	7	15	1
2	E2	20000	0	1	1	0	1	3	3	0
3	E3	15000	1	1	2	1	1	5	10	1
4	E4	30000	1	1	1	0	0	15	7	0
5	E5	10000	1	1	0	1	1	1	6	1
6	E6	40000	0	1	0	0	1	3	16	0
7	E7	25000	0	0	0	0	1	0	8	1
8	E8	20000	0	1	0	1	1	2	6	0
9	E9	20000	1	1	3	1	0	7	5	1
10	E10	30000	1	1	2	1	0	1	20	1
11	E11	45000	0	0	0	0	0	2	12	0
12	E12	8000	1	1	2	1	0	3	1	1
13	E13	20000	0	0	0	0	0	27	5	0
14	E14	10000	0	1	0	0	1	0	7	1
15	E15	8000	0	1	0	0	0	3	2	1

```
      1     2     3     4     5     6     7     8     9    10    11    12    13    14
2  10000
3   5000  5000
4  20000 10000 15000
5      9 10000  5000 20000
6  30000 20000 25000 10000 30000
7  15000  5000 10000  5000 15000 15000
8  10000     3  5000 10000 10000 20000  5000
9  10000     4  5000 10000 10000 20000  5000     5
10 20000 10000 15000    14 20000 10000  5000 10000 10000
11 35000 25000 30000 15000 35000  5000 20000 25000 25000 15000
12  2000 12000  7000 22000  2000 32000 17000 12000 12000 22000 37000
13 10000    24  5000 10000 10000 20000  5000    25    20 10000 25000 12000
14     8 10000  5000 20000     1 30000 15000 10000 10000 20000 35000  2000 10000
15  2000 12000  7000 22000  2000 32000 17000 12000 12000 22000 37000     2 12000  2000
```



3 Definición de grupos

```
plot(Jerarquico)
rect.hclust(Jerarquico, k = 4, border = 2:6)
```



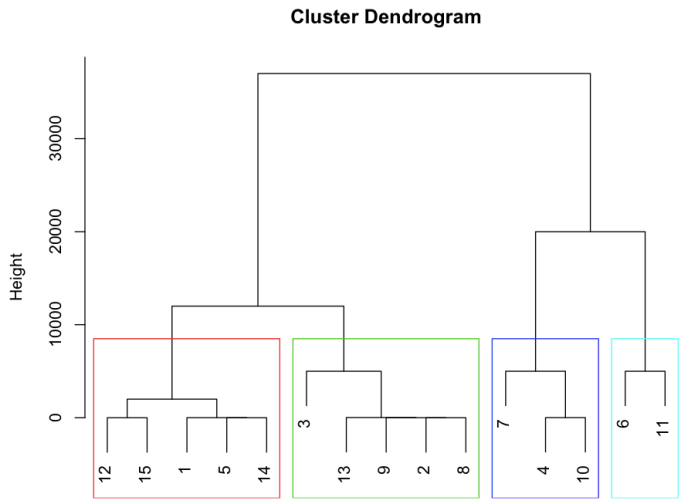

```
Clusters <- cutree(Jerarquico, 4)

Clusters
```

```
[1] 1 2 2 3 1 4 3 2 2 3 4 1 2 1 1
```

```
table(Clusters)
```

```
1 2 3 4
5 5 3 2
```

3 Definición de grupos




```
require(rattle)
Centros <- centers.hclust(DatosEmp [2:10], Jerarquico, nclust=4)
Centros
```

```
      Salario    Casado     Coche Hijos  Vivienda Sindicato FaltasAnno Antiguedad      Sexo
[1,]  9200.00 0.6000000 0.8000000   0.4 0.4000000 0.4000000   2.800000    6.20000 1.0000000
[2,] 19000.00 0.4000000 0.8000000   1.2 0.6000000 0.6000000   8.800000    5.80000 0.4000000
[3,] 28333.33 0.6666667 0.6666667   1.0 0.3333333 0.3333333   5.333333   11.66667 0.6666667
[4,] 42500.00 0.0000000 0.5000000   0.0 0.0000000 0.5000000   2.500000   14.00000 0.0000000
```

- **Cluster 2 [5 elementos –2, 3, 8, 9, 13–]**. Empleados con salario promedio de $19000, solteros en su mayoría (60%), con coche (80%) y con un hijo en promedio. Tienen vivienda propia en su mayoría (60%), sindicalizados en su mayoría (60%), con la mayor cantidad de faltas al año (9), una antigüedad promedio de 6 años y en su mayoría varones (60%).

4 Interpretación

```
     Salario    Casado     Coche Hijos  Vivienda Sindicato FaltasAnno Antiguedad      Sexo
[1,]  9200.00 0.6000000 0.8000000   0.4 0.4000000 0.4000000   2.800000    6.20000 1.0000000
[2,] 19000.00 0.4000000 0.8000000   1.2 0.6000000 0.6000000   8.800000    5.80000 0.4000000
[3,] 28333.33 0.6666667 0.6666667   1.0 0.3333333 0.3333333   5.333333   11.66667 0.6666667
[4,] 42500.00 0.0000000 0.5000000   0.0 0.0000000 0.5000000   2.500000   14.00000 0.0000000
```

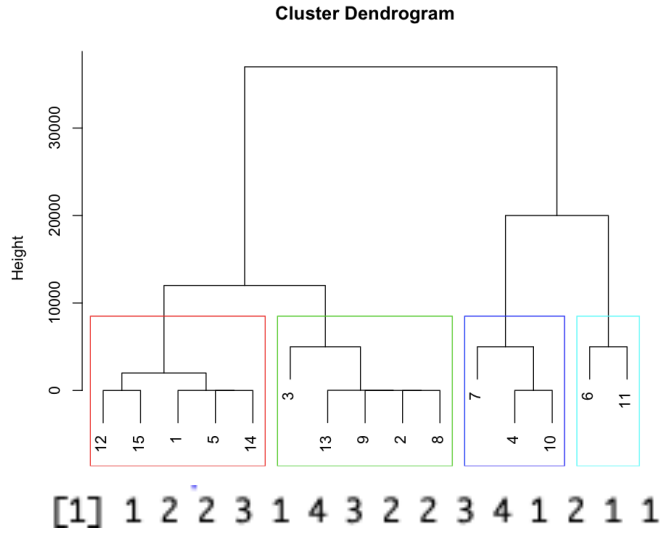
Cluster 1: 5 empleados

Salario :	9200
Casado :	Si = 0.6 / No = 0.4
Coche :	Si = 0.8 / No = 0.2
Hijos :	0.4
Vivienda :	Prop = 0.4 Alquiler = 0.6
Sindicato :	Si = 0.4 / No = 0.6
Faltas/Año :	2.8 (3)
Antigüedad :	6.2 (6)
Sexo :	M = 1

Cluster 2: 5 empleados

Salario :	19000
Casado :	Si = 0.4 / No = 0.6
Coche :	Si = 0.8 / No = 0.2
Hijos :	1.2
Vivienda :	Prop = 0.6 Alquiler = 0.4
Sindicato :	Si = 0.6 / No = 0.4
Faltas/Año :	8.8 (9)
Antigüedad :	5.8 (6)
Sexo :	M = 0.4 / F = 0.6

...



- **Cluster 2 [5 elementos –2, 3, 8, 9, 13–]**. Empleados con salario promedio de $19000, solteros en su mayoría (60%), con coche (80%) y con un hijo en promedio. Tienen vivienda propia en su mayoría (60%), sindicalizados en su mayoría (60%), con la mayor cantidad de faltas al año (9), una antigüedad promedio de 6 años y en su mayoría mujeres (60%).

Práctica 5
Clustering Jerárquico

-Distancias Manhattan-

1 Obtención de una matriz de distancias (Manhattan)

```
DistManhattan <- dist(DatosEmp[2:10], method = "manhattan")

DistManhattan
```

```
       1     2     3     4     5     6     7     8     9    10    11    12    13    14
2  10021
3   5012  5013
4  20019 10018 15017
5     18 10009  5010 20019
6  30009 20014 25013 10024 30015
7  15016  5011 10012  5021 15006 15013
8  10019     6  5011 10018 10003 20012  5007
9  10015    12  5009 10014 10011 20022  5017    12
10 20015 10024 15015    30 20017 10012  5019 10020 10022
11 35010 25013 30012 15021 35012  5007 20008 25009 25019 15015
12  2022 12007  7012 22021  2010 32021 17016 12011 12009 22021 37018
13 10032    29  5034 10017 10032 20037  5032    29    27 10047 25032 12034
14    18 10009  5012 20019     4 30013 15002 10005 10015 20019 35010  2014 10032
15  2019 12004  7015 22020  2009 32016 17011 12008 12012 22024 37013     5 12029  2009
```

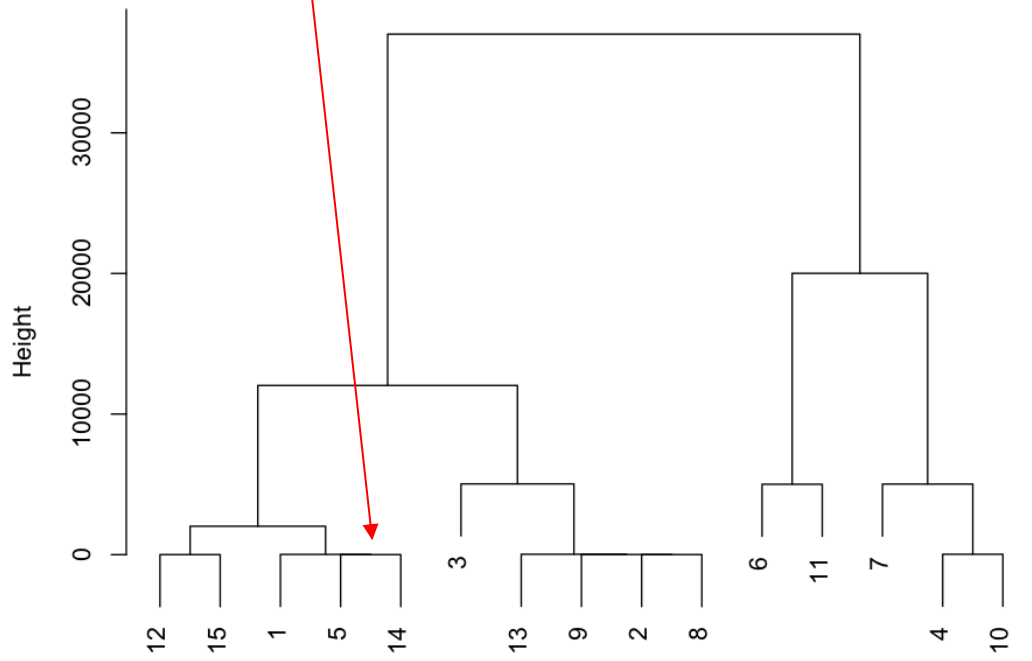
Práctica

2 Obtención de grupos

```
Jerarquico <- hclust(DistManhattan)

plot(Jerarquico)
```

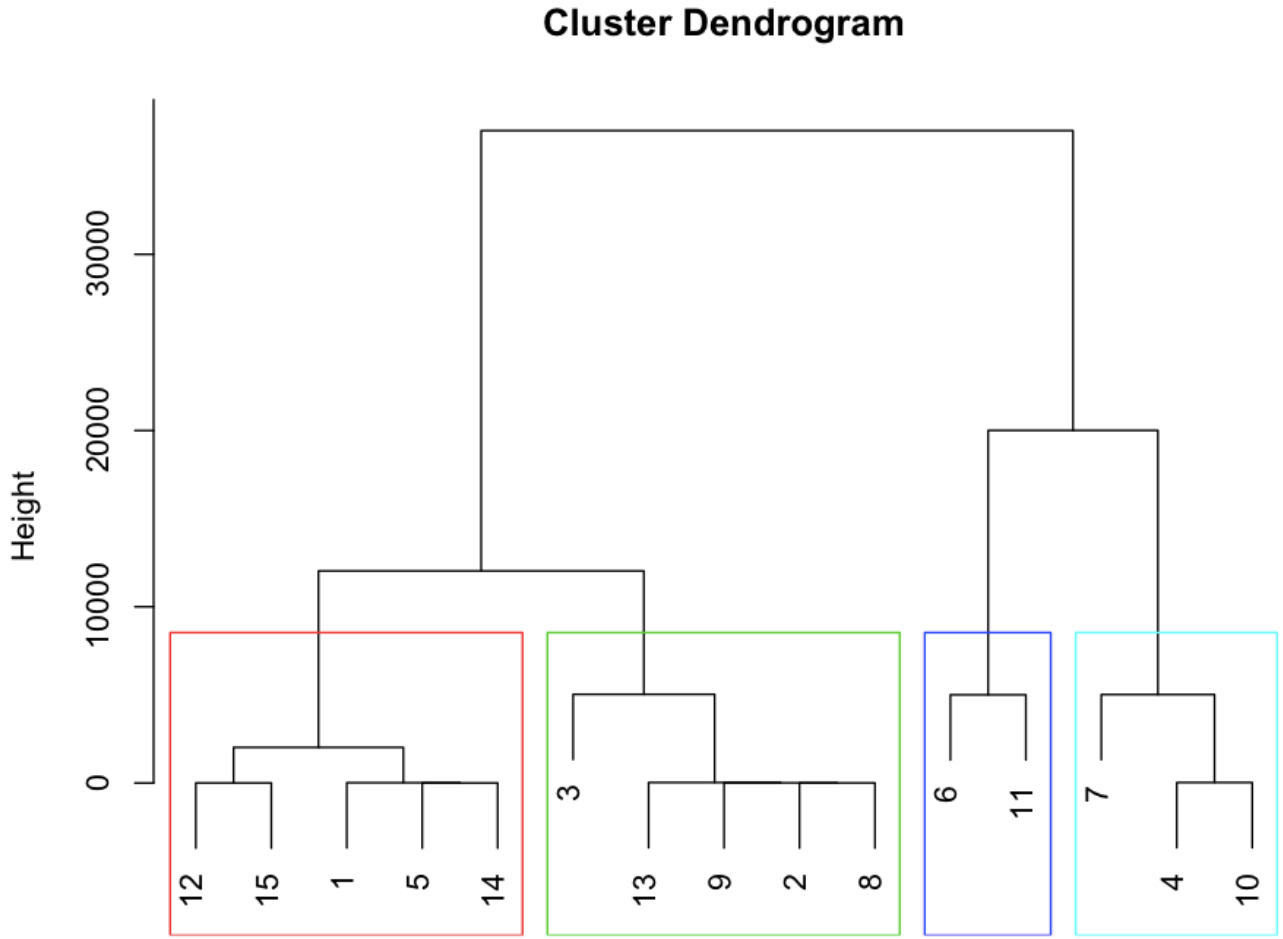
	ID	Salario	Casado	Coche	Hijos	Vivienda	Sindicato	FaltasAno	Antiguedad	Sexo
1	E1	10000	1	0	0	0	0	7	15	1
2	E2	20000	0	1	1	0	1	3	3	0
3	E3	15000	1	1	2	1	1	5	10	1
4	E4	30000	1	1	1	0	0	15	7	0
5	E5	10000	1	1	0	1	1	1	6	1
6	E6	40000	0	1	0	0	1	3	16	0
7	E7	25000	0	0	0	0	1	0	8	1
8	E8	20000	0	1	0	1	1	2	6	0
9	E9	20000	1	1	3	1	0	7	5	1
10	E10	30000	1	1	2	1	0	1	20	1
11	E11	45000	0	0	0	0	0	2	12	0
12	E12	8000	1	1	2	1	0	3	1	1
13	E13	20000	0	0	0	0	0	27	5	0
14	E14	10000	0	1	0	0	1	0	7	1
15	E15	8000	0	1	0	0	0	3	2	1

```
      1     2     3     4     5     6     7     8     9    10    11    12    13    14
2  10021
3   5012  5013
4  20019 10018 15017
5     18 10009  5010 20019
6  30009 20014 25013 10024 30015
7  15016  5011 10012  5021 15006 15013
8  10019     6  5011 10018 10003 20012  5007
9  10015    12  5009 10014 10011 20022  5017    12
10 20015 10024 15015    30 20017 10012  5019 10020 10022
11 35010 25013 30012 15021 35012  5007 20008 25009 25019 15015
12  2022 12007  7012 22021  2010 32021 17016 12011 12009 22021 37018
13 10032    29  5034 10017 10032 20037  5032    29    27 10047 25032 12034
14    18 10009  5012 20019     4 30013 15002 10005 10015 20019 35010  2014 10032
15  2019 12004  7015 22020  2009 32016 17011 12008 12012 22024 37013     5 12029  2009
```



3 Definición de grupos

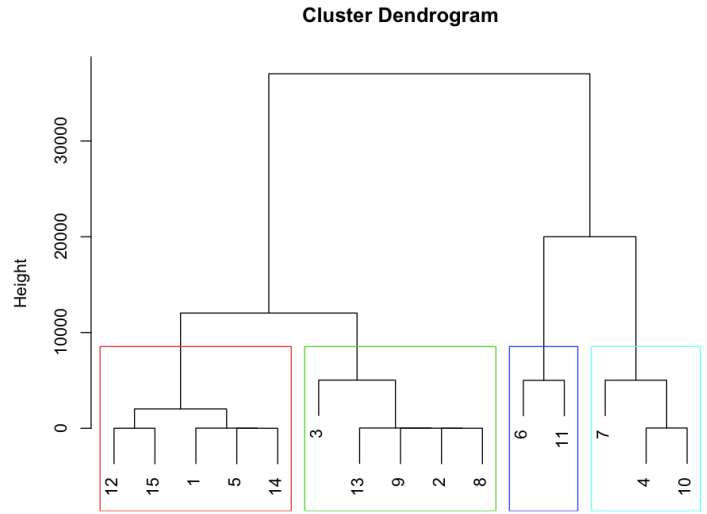
```
plot(Jerarquico)
rect.hclust(Jerarquico, k = 4, border = 2:6)
```




```
Clusters <- cutree(Jerarquico, 4)

Clusters
[1] 1 2 2 3 1 4 3 2 2 3 4 1 2 1 1

table(Clusters)

1 2 3 4
5 5 3 2
```

3 Definición de grupos

```
require(rattle)
Centros <- centers.hclust(DatosEmp [2:10], Jerarquico, nclust=4)
Centros
```



	Salario	Casado	Coche	Hijos	Vivienda	Sindicato	FaltasAnno	Antiguedad	Sexo
[1,]	9200.00	0.6000000	0.8000000	0.4	0.4000000	0.4000000	2.800000	6.20000	1.0000000
[2,]	19000.00	0.4000000	0.8000000	1.2	0.6000000	0.6000000	8.800000	5.80000	0.4000000
[3,]	28333.33	0.6666667	0.6666667	1.0	0.3333333	0.3333333	5.333333	11.66667	0.6666667
[4,]	42500.00	0.0000000	0.5000000	0.0	0.0000000	0.5000000	2.500000	14.00000	0.0000000

- **Cluster 3 [5 elementos –4, 7 10–]**. Empleados con salario promedio de $28333, casados en su mayoría (67%), con coche en su mayoría (67%) y con un hijo. No tienen vivienda propia en su mayoría (67%), no sindicalizados en su mayoría (67%), con varias faltas al año (5), con una antigüedad promedio de 12 años y la mayoría varones (67%).

4 Interpretación

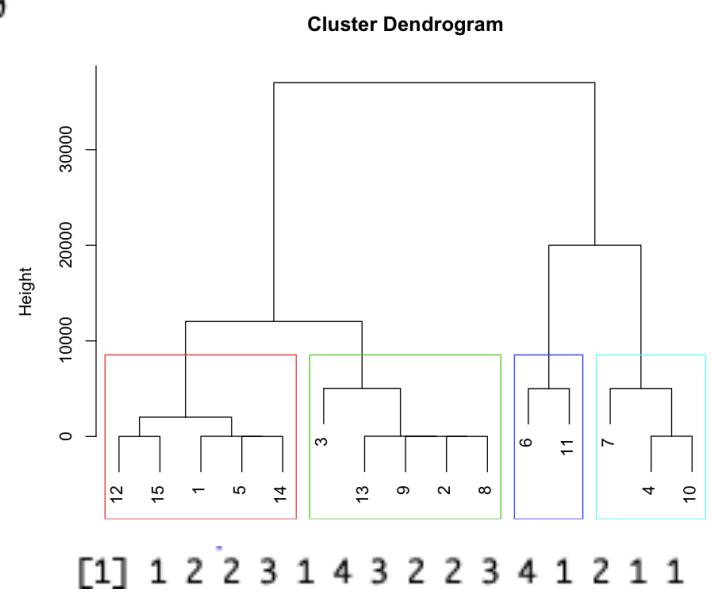
```
     Salario    Casado     Coche Hijos  Vivienda Sindicato FaltasAnno Antiguedad      Sexo
[1,]  9200.00 0.6000000 0.8000000   0.4 0.4000000 0.4000000   2.800000    6.20000 1.0000000
[2,] 19000.00 0.4000000 0.8000000   1.2 0.6000000 0.6000000   8.800000    5.80000 0.4000000
[3,] 28333.33 0.6666667 0.6666667   1.0 0.3333333 0.3333333   5.333333   11.66667 0.6666667
[4,] 42500.00 0.0000000 0.5000000   0.0 0.0000000 0.5000000   2.500000   14.00000 0.0000000
```

Cluster 3: 3 empleados

Salario :	28333
Casado :	Si = 0.67 / No = 0.33
Coche :	Si = 0.67 / No = 0.33
Hijos :	1
Vivienda :	Prop = 0.33 Alquiler = 0.67
Sindicato :	Si = 0.33 / No = 0.67
Faltas/Año :	5.3 (5)
Antigüedad :	11.6 (12)
Sexo :	M = 0.67 / F = 0.33

Cluster 4: 2 empleados

Salario :	42500
Casado :	No = 1
Coche :	Si = 0.5 / No = 0.5
Hijos :	0
Vivienda :	Alquiler = 1
Sindicato :	Si = 0.5 / No = 0.5
Faltas/Año :	2.5 (3)
Antigüedad :	14
Sexo :	F = 1



- **Cluster 3 [5 elementos –4, 7, 10–]**. Empleados con salario promedio de $28333, casados en su mayoría (67%), con coche en su mayoría (67%) y con un hijo. No tienen vivienda propia en su mayoría (67%), no sindicalizados en su mayoría (67%), con varias faltas al año (5), con una antigüedad promedio de 12 años y la mayoría varones (67%).