



# *Técnicas NO Supervisadas*

## *Minería de Datos*

PhD Jose Ricardo Zapata  
< joser.zapata@upb.edu.co >

UNIVERSIDAD PONTIFICIA BOLIVARIANA  
FACULTAD TIC

# AGENDA



1. Método Particional
2. Método Jerárquico
3. Método Probabilístico
4. Redes Neuronales
5. Reglas de Asociación: A priori
6. Votación para selección de factores

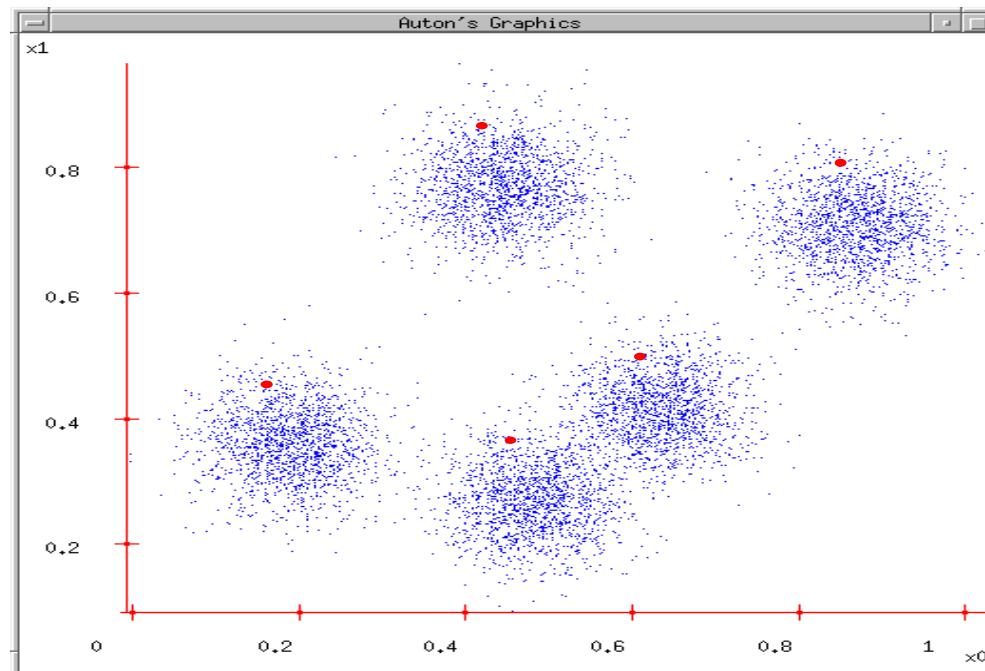
# AGENDA



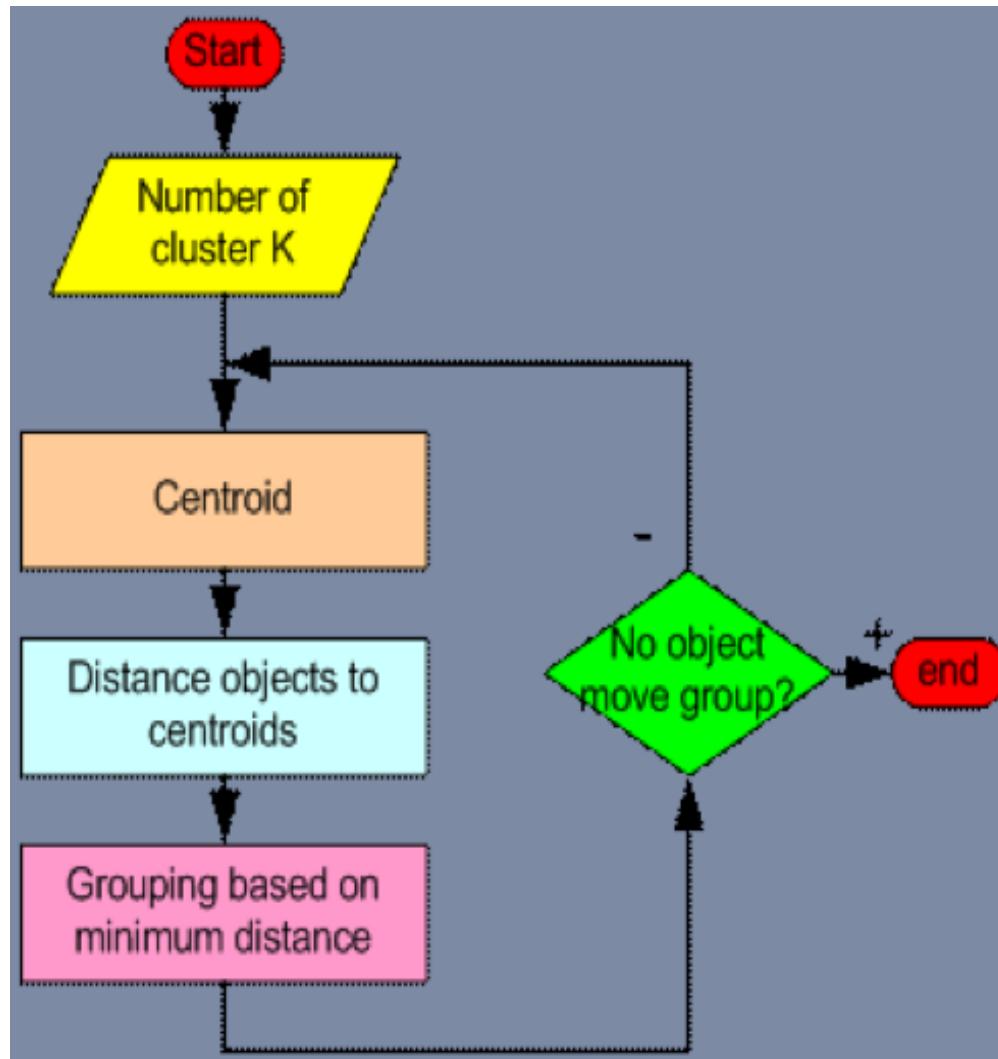
1. Método Particional
2. Método Jerárquico
3. Método Probabilístico
4. Redes Neuronales
5. Reglas de Asociación: A priori
6. Votación para selección de factores

# *Métodos Particionales*

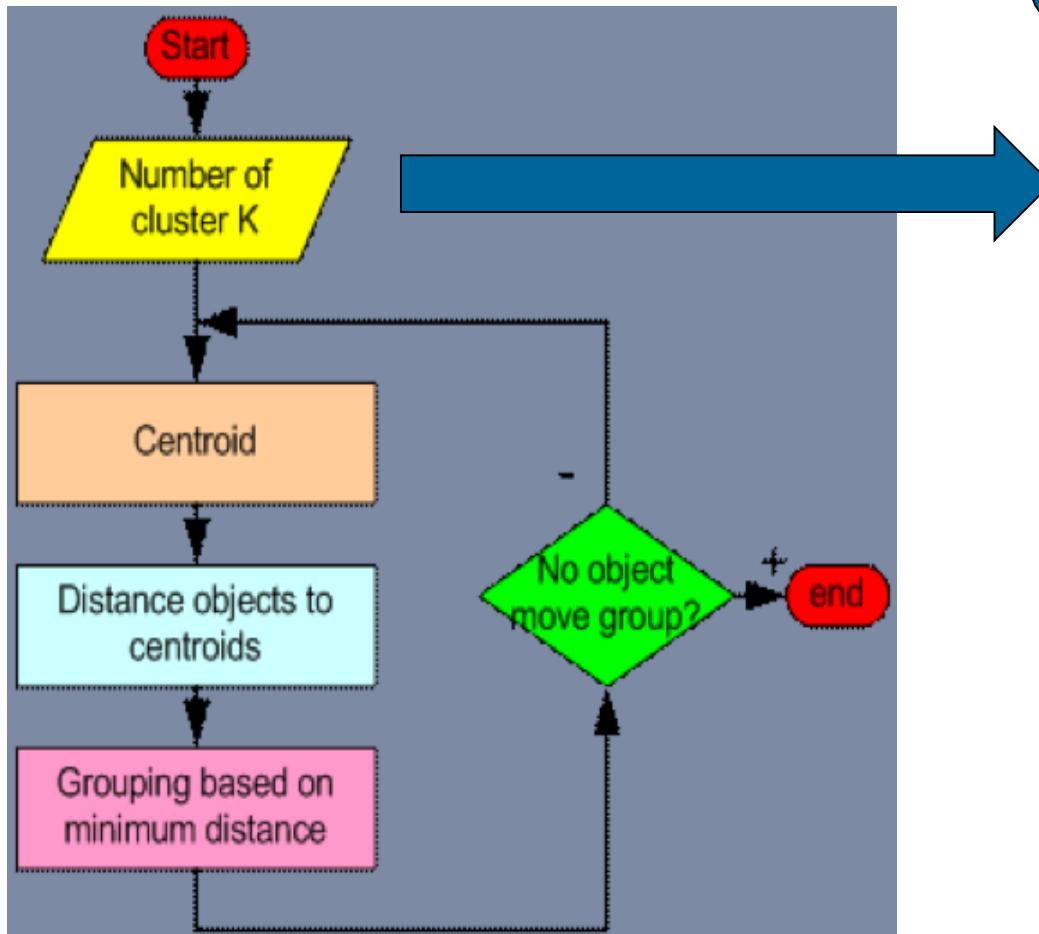
- Divide el conjunto de datos en un número predefinido de grupos. K-Means es el método más comúnmente utilizado, la idea del método es definir  $k$  centroides, uno por clúster, y los datos son asociados al centroide más cercano.



# *K-Means*



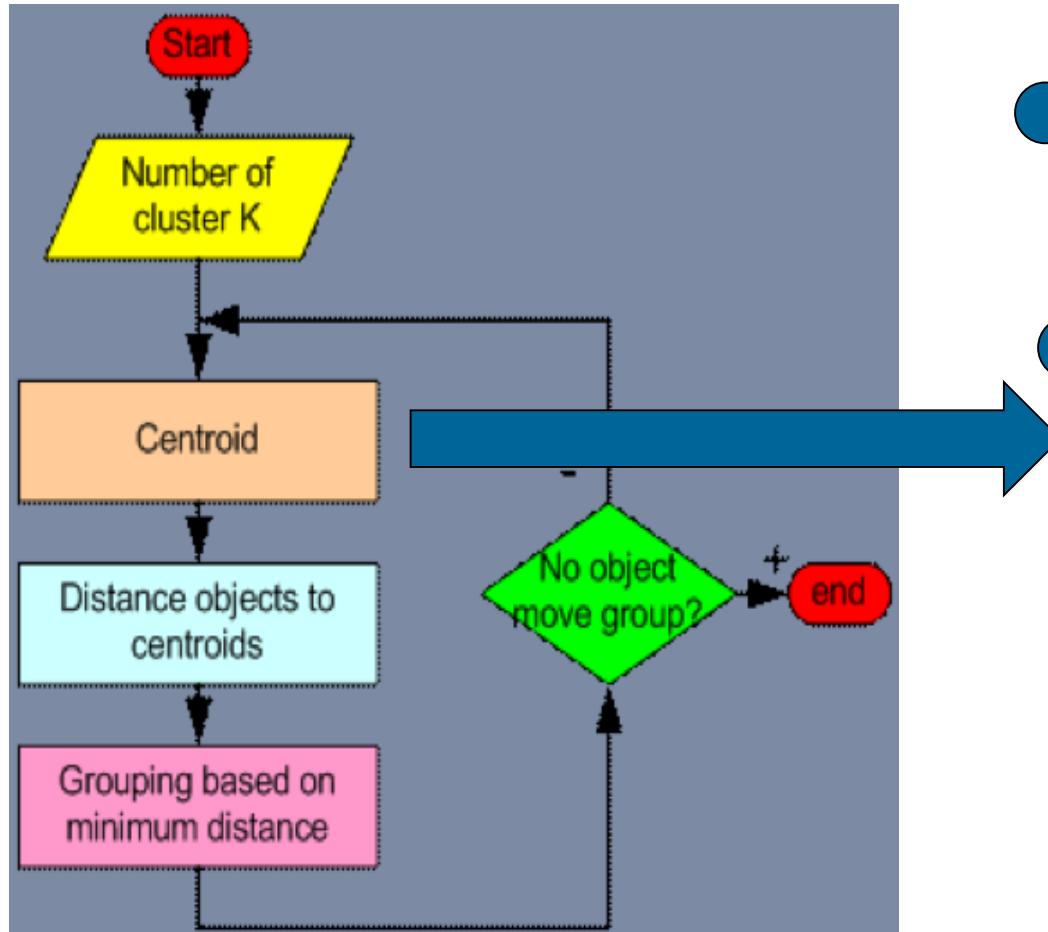
# *K-Means*



● Se debe especificar por adelantado cuántos grupos se van a crear

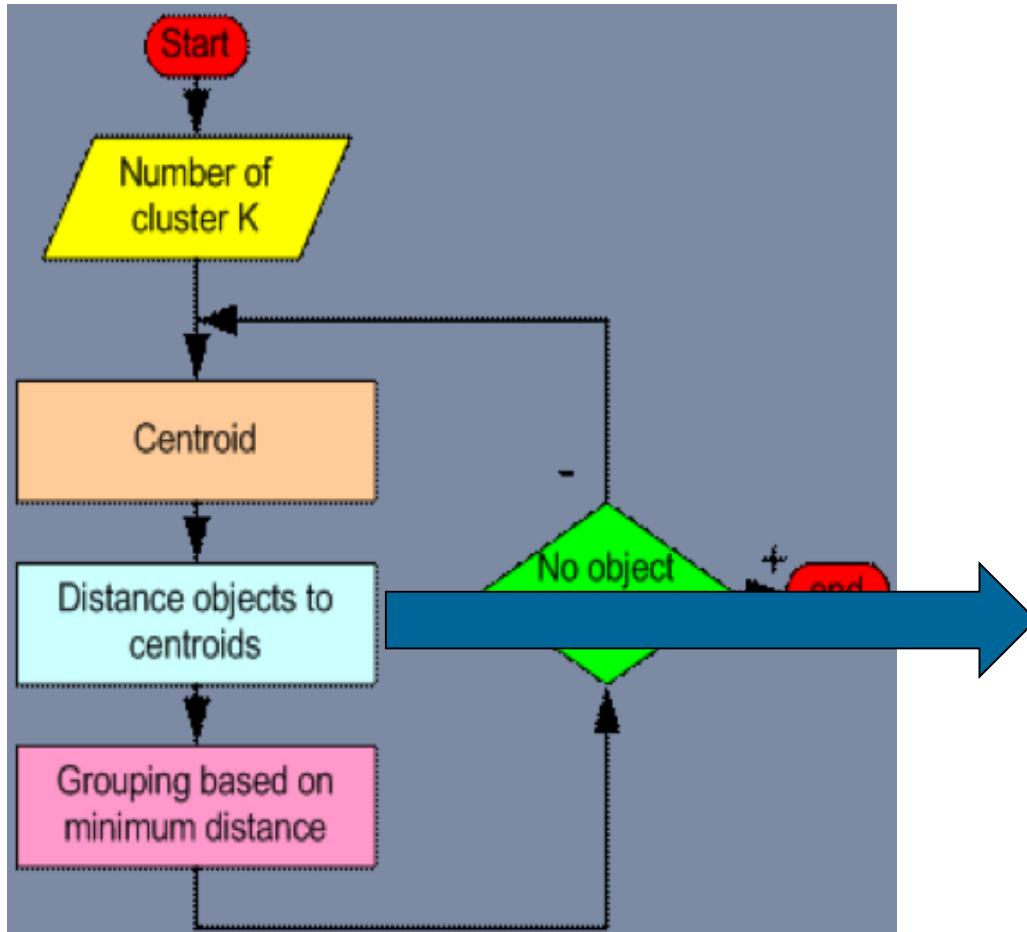
↓  
**Parámetro K**

# *K-Means*



- Se determinan los centroides de los grupos.
- La primera vez se seleccionan aleatoriamente k datos que representan el centro o media de cada clúster

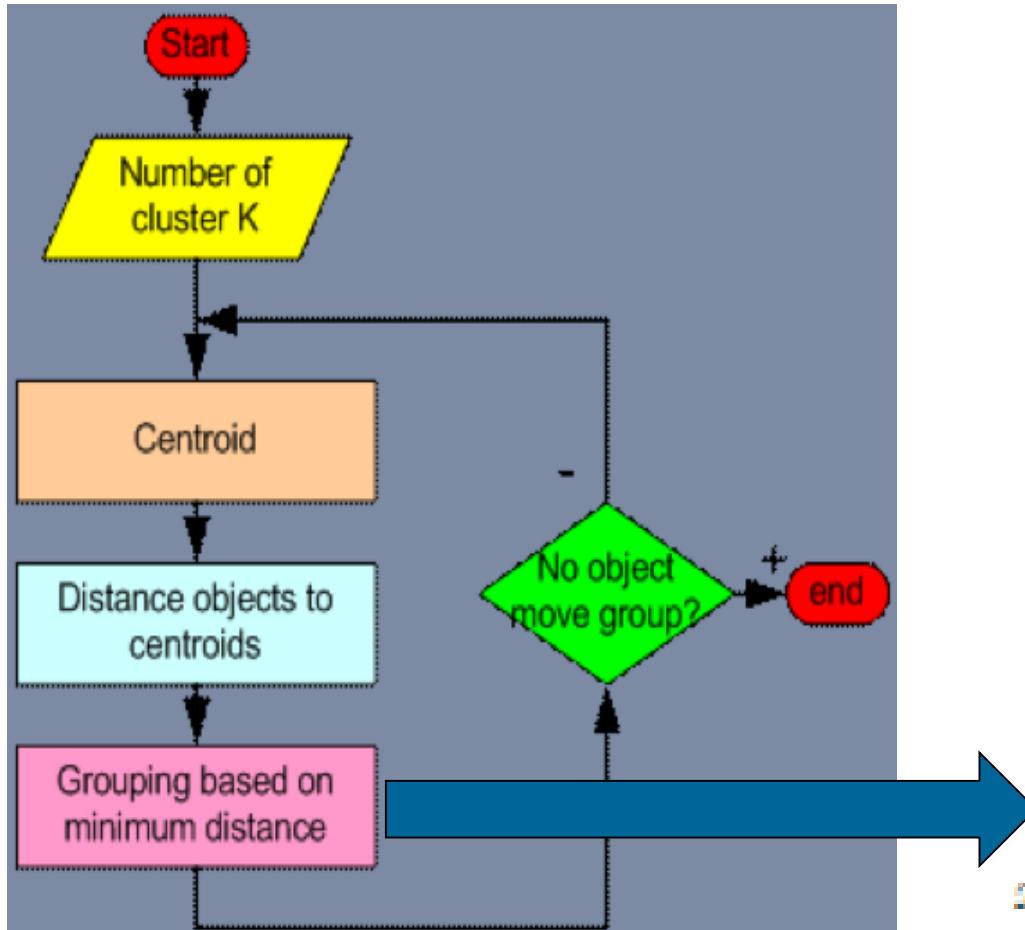
# K-Means



Se calcula la distancia de los datos a cada centroide.

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^n (x_{il} - x_{jl})^2}$$

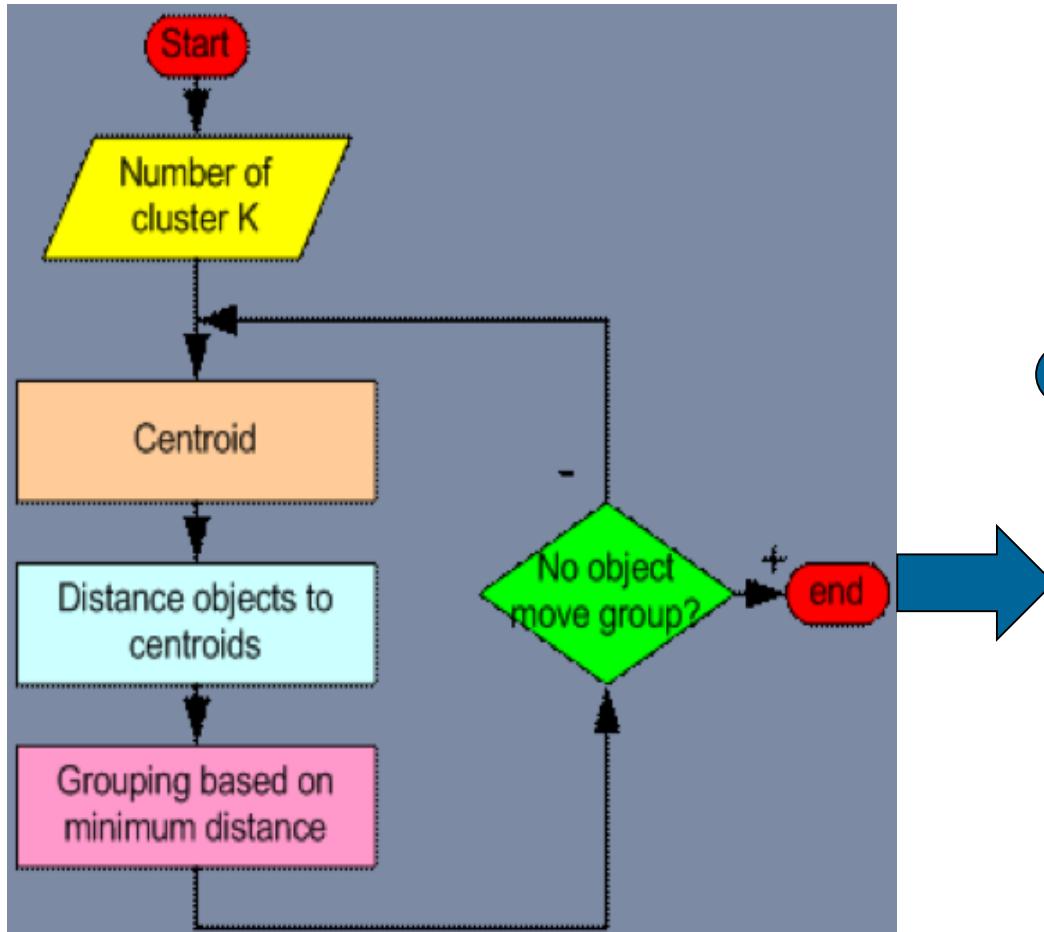
# K-Means



Cada dato es asignado al centro del clúster más cercano.

$$x_j \in c_p \text{ if } \|x_j - v_p\| \leq \|x_j - v_q\|$$

# *K-Means*

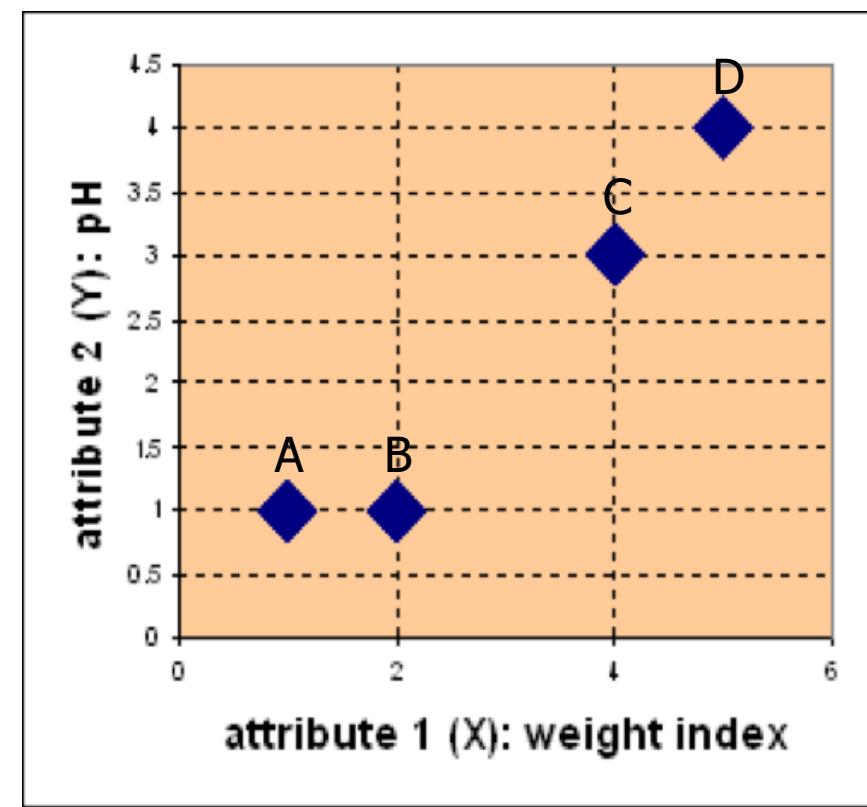


La iteración continúa hasta que se repite la asignación de los mismos ejemplos a los mismos clústers, ya que los puntos centrales de los clústers se han estabilizado.

# *K-Means – Ejemplo*

Se tienen 4 tipos de medicinas, cada una con 2 atributos (pH y weight). Se requiere agrupar las medicinas en 2 clusters.

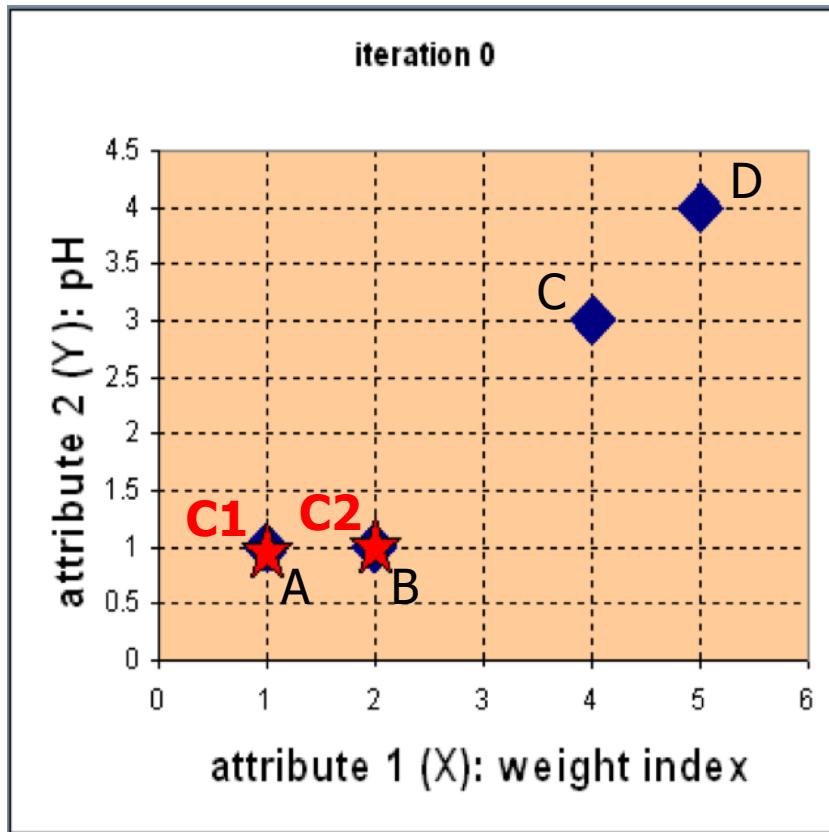
<b>Medicina</b>	<b>Weight</b>	<b>pH</b>
A	1	1
B	2	1
C	4	3
D	5	4



# *K-Means – Ejemplo*

- Seleccionar centroides aleatorios

$$c_1 = A, c_2 = B$$



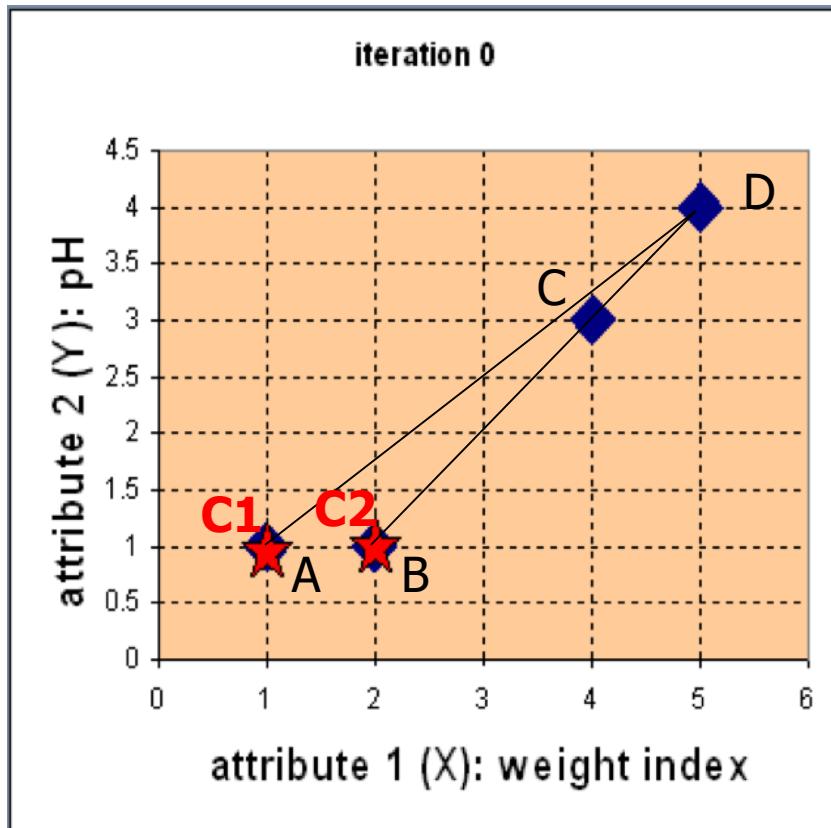
Medicina	Weight	pH
A	1	1
B	2	1
C	4	3
D	5	4

$$c_1 = (1, 1) \quad group - 1$$

$$c_2 = (2, 1) \quad group - 2$$

# *K-Means – Ejemplo*

- Seleccionar centroides aleatorios  $c_1 = A, c_2 = B$



Medicina	Weight	pH
A	1	1
B	2	1
C	4	3
D	5	4

$$\mathbf{c}_1 = (1, 1) \quad group - 1$$
$$\mathbf{c}_2 = (2, 1) \quad group - 2$$

- Calcular las distancias de cada dato a cada centroide

Euclidean distance

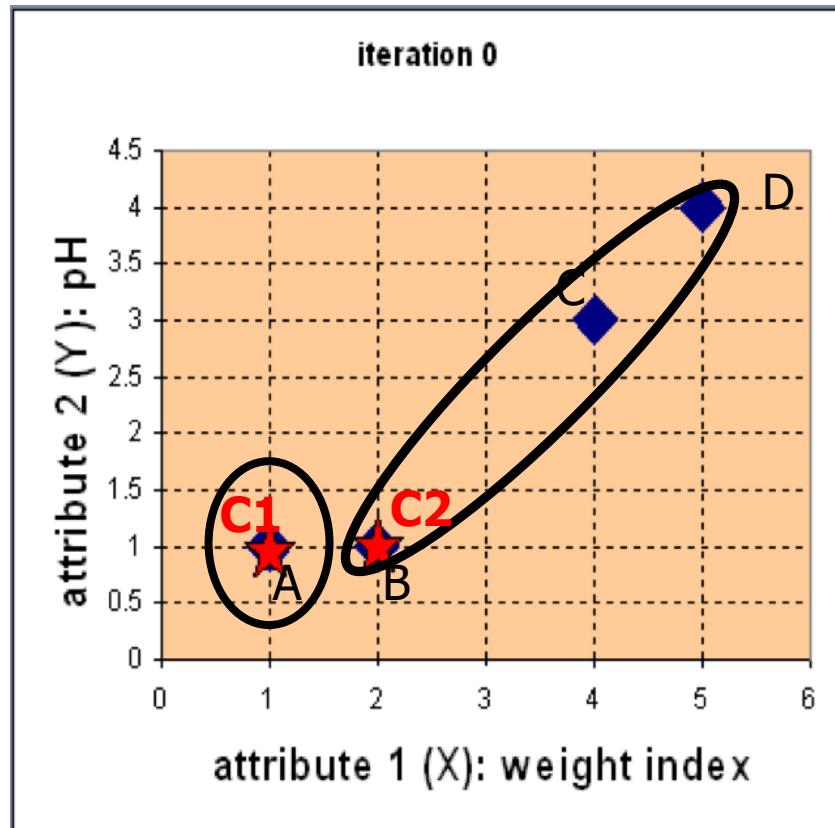
$$d(D, c_1) = \sqrt{(5-1)^2 + (4-1)^2} = 5$$

$$d(D, c_2) = \sqrt{(5-2)^2 + (4-1)^2} = 4.24$$

$$\mathbf{D}^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \begin{matrix} A \\ B \\ C \\ D \end{matrix} \begin{matrix} c_1 \\ c_2 \end{matrix}$$

# *K-Means – Ejemplo*

- Seleccionar centroides aleatorios  $c_1 = A, c_2 = B$



$$c_1 = (1, 1) \quad group - 1$$
$$c_2 = (2, 1) \quad group - 2$$

- Calcular las distancias de cada dato a cada centroide

Euclidean distance

$$d(D, c_1) = \sqrt{(5-1)^2 + (4-1)^2} = 5$$

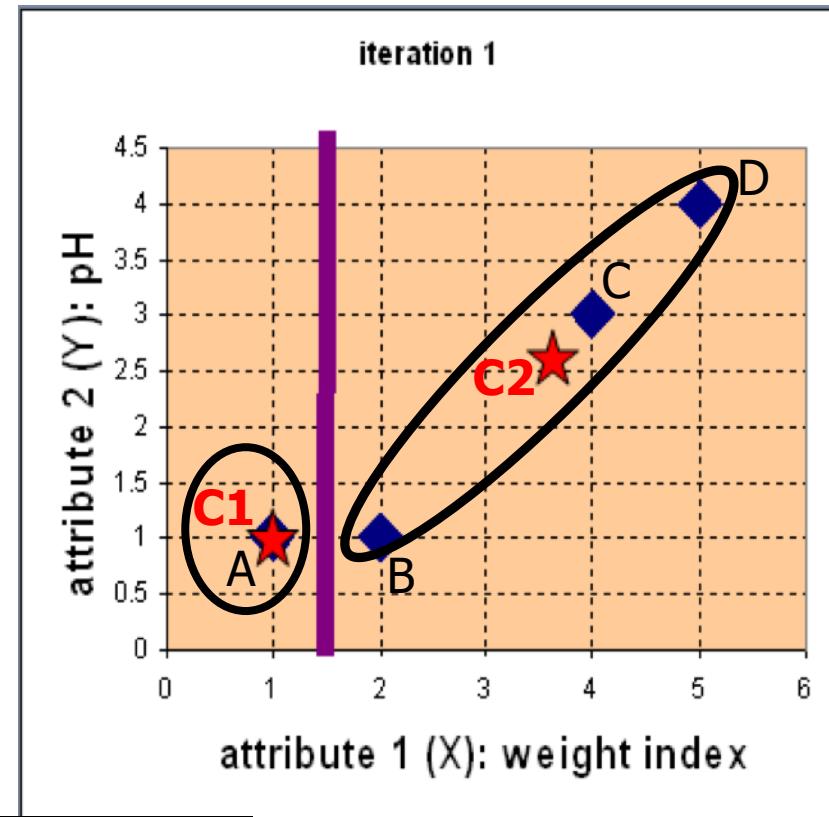
$$d(D, c_2) = \sqrt{(5-2)^2 + (4-1)^2} = 4.24$$

$$\mathbf{D}^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \begin{matrix} A \\ B \\ C \\ D \end{matrix} \begin{matrix} c_1 \\ c_2 \end{matrix}$$

- Asignar cada dato al clúster más cercano.

# *K-Means – Ejemplo*

- Calcular los nuevos centroides con los miembros de cada grupo.

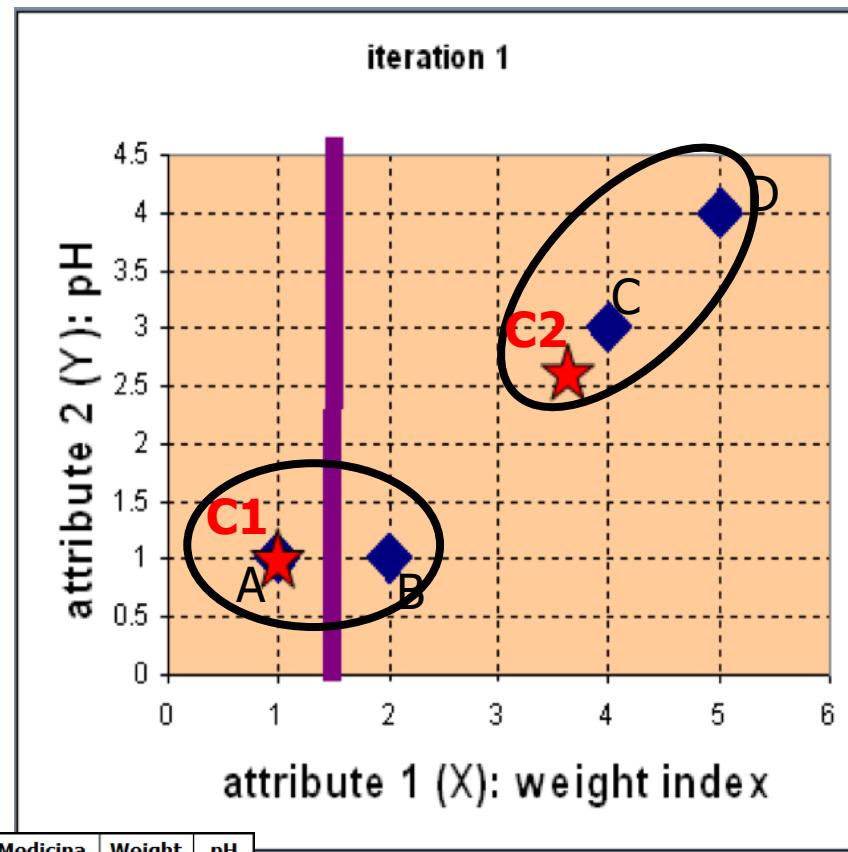


Medicina	Weight	pH
A	1	1
B	2	1
C	4	3
D	5	4

$$c_1 = (1, 1)$$
$$c_2 = \left( \frac{2+4+5}{3}, \frac{1+3+4}{3} \right)$$
$$= (11/3, 8/3)$$
$$= (3.67, 2.67)$$

# *K-Means - Ejemplo*

- Calcular la distancia de cada dato a cada centroide



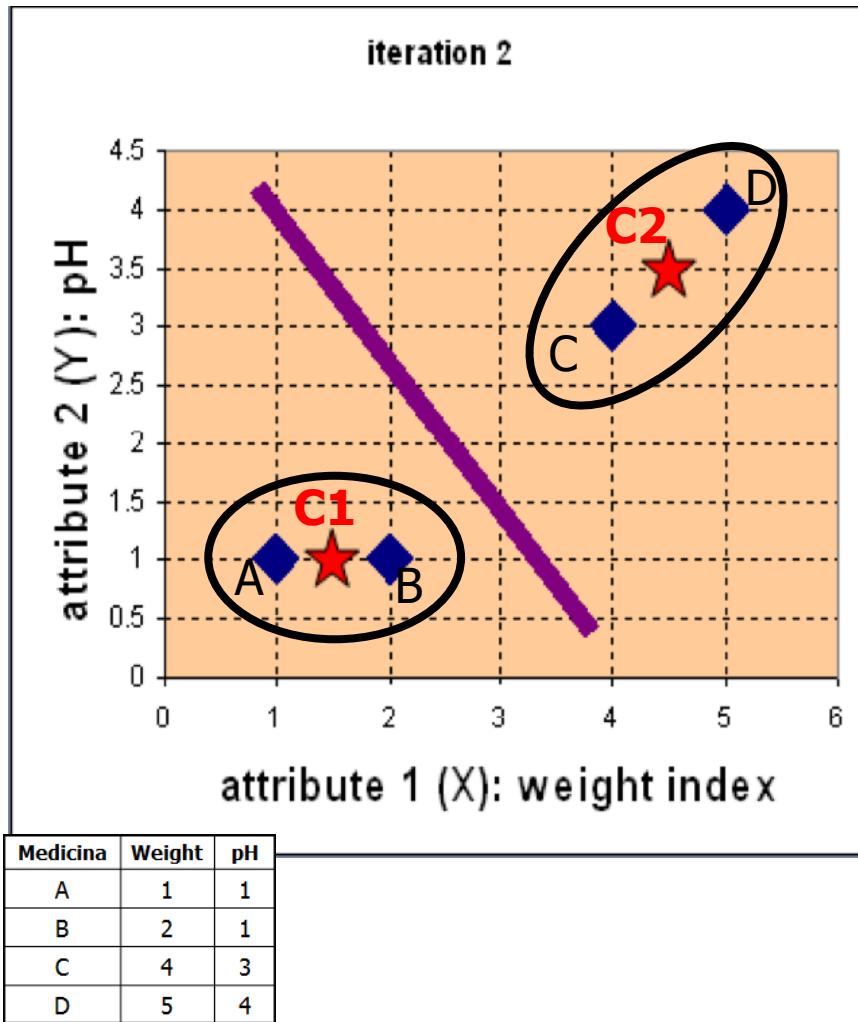
A	B	C	D
0	1	3.61	5
3.14	2.36	0.47	1.89

$$\mathbf{D}^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \quad \mathbf{c}_1 = (1, 1) \quad \text{group - 1}$$
$$\mathbf{c}_2 = \left( \frac{11}{3}, \frac{8}{3} \right) \quad \text{group - 2}$$

- Asignar cada dato al clúster más cercano.

# *K-Means – Ejemplo*

- Calcular los nuevos centroides con los miembros de cada grupo.

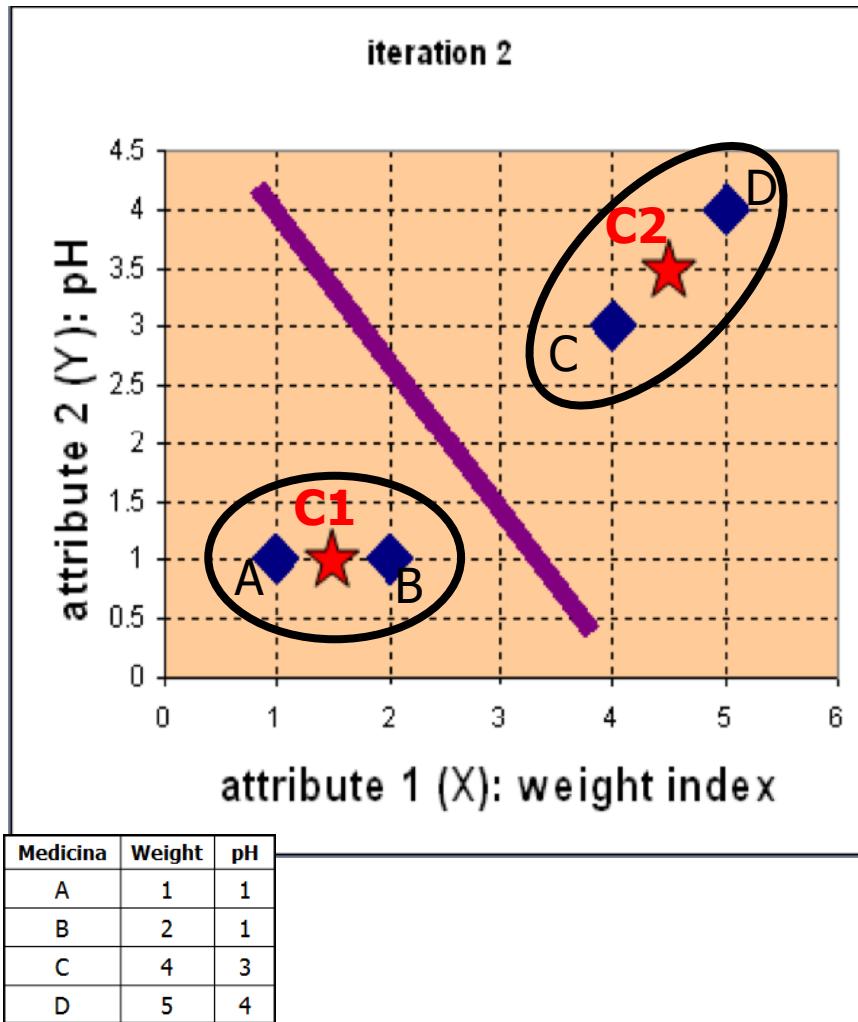


$$c_1 = \left( \frac{1+2}{2}, \frac{1+1}{2} \right) = \left( 1\frac{1}{2}, 1 \right)$$

$$c_2 = \left( \frac{4+5}{2}, \frac{3+4}{2} \right) = \left( 4\frac{1}{2}, 3\frac{1}{2} \right)$$

# *K-Means – Ejemplo*

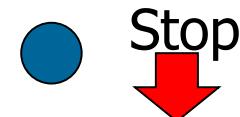
- Calcular la distancia de cada dato a cada centroide



A	B	C	D
---	---	---	---

$$\mathbf{D}^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \quad \mathbf{c}_1 = (1\frac{1}{2}, 1) \quad \text{group - 1}$$
$$\mathbf{c}_2 = (4\frac{1}{2}, 3\frac{1}{2}) \quad \text{group - 2}$$

Asignar cada dato al clúster más cercano.



No se modifican los grupos

# *K-Means – Limitaciones*

- Sensible a los centroides iniciales, ya que converge a óptimos locales.
- Requiere especificar el número de clusters.
- Se afecta por datos “ruidosos”.
- No es aplicable a datos categóricos.

# *K-Means - Modificaciones*

---

- K-Medoids: resistencia a ruido y datos atípicos.
- K-Modes: extensión para datos categóricos

# *Simulación en WEKA*

- La empresa de software para Internet “Memolum Web” quiere extraer tipologías de empleados, con el objetivo de hacer una política de personal más fundamentada y seleccionar a qué grupos incentivar.
- Las variables que se recogen de las fichas de los 15 empleados de la empresa son:
  - Sueldo: sueldo anual en euros.
  - Casado: si está casado o no.
  - Coche: si viene en coche a trabajar (o al menos si lo aparcá en el parking de la empresa).
  - Hijos: si tiene hijos.
  - Alq/Prop: si vive en una casa alquilada o propia.
  - Sindic: si pertenece al sindicato revolucionario de Internet
  - Bajas/Año: media del nº de bajas por año
  - Antigüedad: antigüedad en la empresa
  - Sexo: H: hombre, M: mujer.
- Se intenta extraer grupos de entre estos quince empleados.

# *Simulación en WEKA*

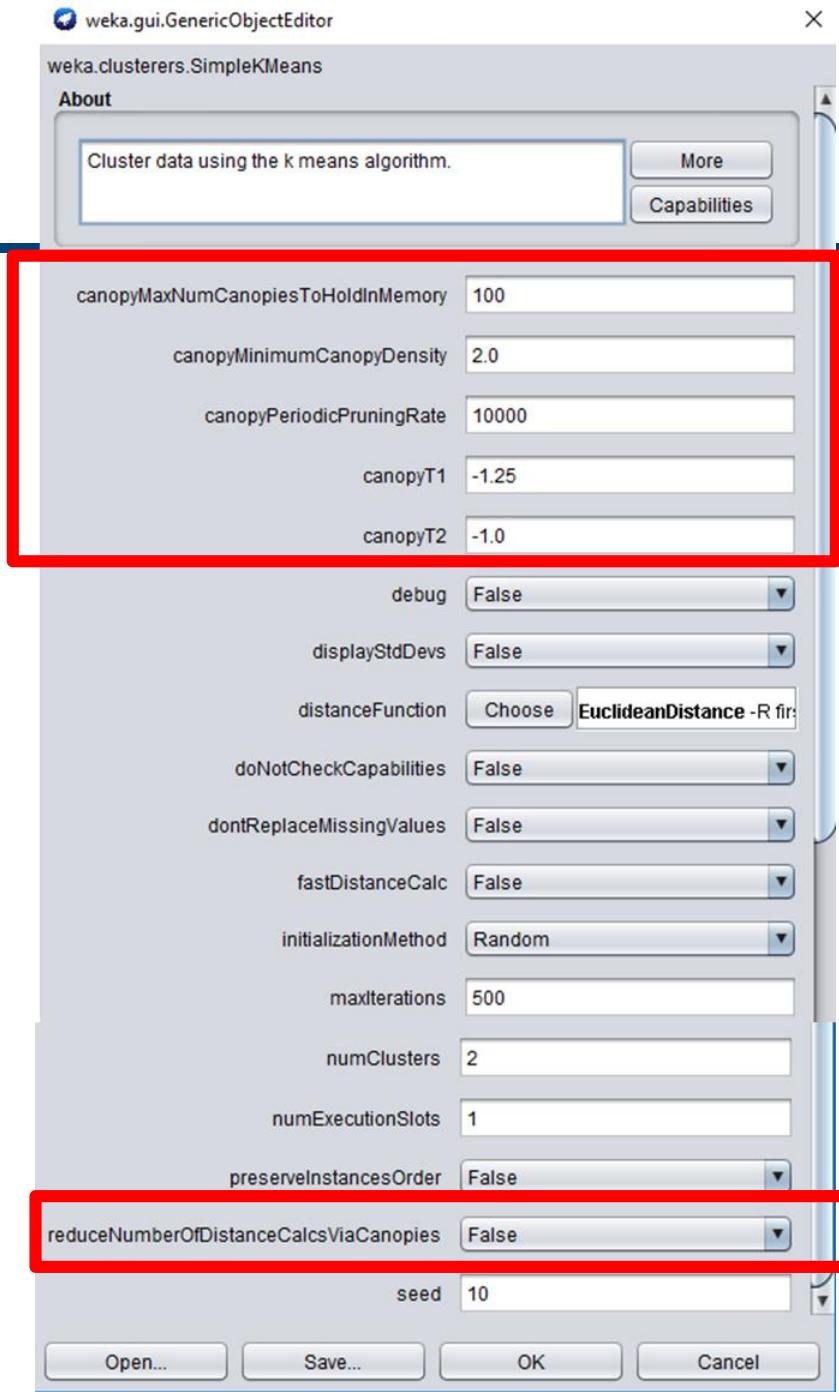
File: empleados.arff

```
@relation empleados
```

```
@attribute Sueldo numeric  
@attribute Casado {Sí,No}  
@attribute Coche {No,Sí}  
@attribute Hijos numeric  
@attribute Alq/Prop {Alquiler,Prop}  
@attribute Sindic. {No,Sí}  
@attribute Bajas/Año numeric  
@attribute Antigüedad numeric  
@attribute Sexo {H,M}
```

```
@data
```

```
10000,Sí,No,0,Alquiler,No,7,15,H  
20000,No,Sí,1,Alquiler,Sí,3,3,M  
15000,Sí,Sí,2,Prop,Sí,5,10,H  
30000,Sí,Sí,1,Alquiler,No,15,7,M  
10000,Sí,Sí,0,Prop,Sí,1,6,H  
40000,No,Sí,0,Alquiler,Sí,3,16,M  
25000,No,No,0,Alquiler,Sí,0,8,H  
20000,No,Sí,0,Prop,Sí,2,6,M  
20000,Sí,Sí,3,Prop,No,7,5,H  
30000,Sí,Sí,2,Prop,No,1,20,H  
50000,No,No,0,Alquiler,No,2,12,M  
8000,Sí,Sí,2,Prop,No,3,1,H  
20000,No,No,0,Alquiler,No,27,5,M  
10000,No,Sí,0,Alquiler,Sí,0,7,H  
8000,No,Sí,0,Alquiler,No,3,2,H
```



## Canopy

Preprocesamiento para acelerar el clustering, eliminando registros que son muy parecidos (distancia menor a T2).

- displayStdDevs: mostrar las desviaciones estándar de los atributos.
- distanceFunction: definir la distancia
- dontReplaceMissingValues: reemplazar valores faltantes con la media/moda.
- fastDistanceCalc: no calcular la cohesion
- InitializationMethod
- maxIterations: definir el número máximo de iteraciones.
- numClusters: configurar el número k de clústers.
- preserveInstancesOrder: preservar el orden de las instancias.

# *Simulación en WEKA: SimpleKmeans*

Cluster centroids:

Attribute	Full Data (15)	Cluster#		
		0 (6)	1 (5)	2 (4)
Sueldo	21066.6667	29166.6667	16600	14500
Casado	No	No	Si	Si
Coche	Si	No	Si	Si
Hijos	0.7333	0.1667	1.8	0.25
Alq/Prop	Alquiler	Alquiler	Prop	Alquiler
Sindic.	No	Si	No	No
Bajas/Año	5.2667	6.1667	3.4	6.25
Antigüedad	8.2	8.3333	8.4	7.75
Sexo	H	M	H	H

Clustered Instances		
0	6	( 40%)
1	5	( 33%)
2	4	( 27%)

# Simulación en WEKA: SimpleKmeans

Cluster centroids:				
Attribute	Full Data (15)	Cluster# 0 (6)	1 (5)	2 (4)
Sueldo	21066.6667 +/-12308.3402	29166.6667 +/-12812.7541	16600 +/-8820.4308	14500 +/-10376.2549
Casado	No Sí No	No 0 ( 0%) 6 (100%)	Si 5 (100%) 0 ( 0%)	Si 2 ( 50%) 2 ( 50%)
Coche	Si No Sí	No 3 ( 50%) 3 ( 50%)	Si 0 ( 0%) 5 (100%)	Si 1 ( 25%) 3 ( 75%)
Hijos	0.7333 +/-1.0328	0.1667 +/-0.4082	1.8 +/-1.0954	0.25 +/-0.5
Alq/Prop	Alquiler Alquiler Prop	Alquiler 5 ( 83%) 1 ( 16%)	Prop 0 ( 0%) 5 (100%)	Alquiler 4 (100%) 0 ( 0%)
Sindic.	No No Sí	Si 2 ( 33%) 4 ( 66%)	No 3 ( 60%) 2 ( 40%)	No 3 ( 75%) 1 ( 25%)
Bajas/Año	5.2667 +/-7.106	6.1667 +/-10.2648	3.4 +/-2.6077	6.25 +/-6.5
Antigüedad	8.2 +/-5.4406	8.3333 +/-4.8442	8.4 +/-7.2319	7.75 +/-5.3774
Sexo	H H M	M 1 ( 16%) 5 ( 83%)	H 5 (100%) 0 ( 0%)	H 3 ( 75%) 1 ( 25%)

# Simulación en WEKA: SimpleKmeans

**Weka Explorer**

Preprocess Classify Cluster Associate Select attributes Visualize

Clusterer

Choose SimpleKMeans -M -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Cluster mode

Use training set

Supplied test set Set...

Percentage split % 66

Classes to clusters evaluation

(Nom) Sexo

Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

06:36:52 - SimpleKMeans  
06:37:33 - SimpleKMeans  
06:39:18 - SimpleKMeans

View in main window  
View in separate window  
Save result buffer  
Delete result buffer

Load model  
Save model  
Re-evaluate model on current test set

Visualize cluster assignments  
Visualize tree

**Weka Clusterer Visualize: 06:39:18 - SimpleKMeans (empleados)**

X: Instance\_number (Num)  
Y: Sueldo (Num)  
Colour: Cluster (Nom)  
Select Instance

Re... Clear Open Save Jitter

Plot: empleados\_clustered

Number of iterations: 3  
Within cluster sum of square

Cluster centroids:

Attribute	Full Data (15)
Sueldo	21066.6667 291
Casado	No
Coche	Si
Hijos	0.7333
Alq/Prop	Alquiler
Sindic.	No
Bajas/Año	5.2667
Antigüedad	8.2

Class colour

cluster0 cluster1 cluster2

1 (full training data) : 0 seconds  
on training set ===

The scatter plot displays salary (Sueldo) on the Y-axis (ranging from 0 to 50,000) against instance number (Instance\_number) on the X-axis (ranging from 0 to 14). The data points are colored according to their cluster assignment: cluster 0 (green), cluster 1 (red), and cluster 2 (blue). A legend at the bottom right identifies the colors for each cluster. The plot shows a clear separation between the three clusters based on salary levels.

# *Simulación en WEKA: SimpleKmeans*

```
@relation empleados_clustered
```

```
@attribute Instance_number numeric
```

```
@attribute Sueldo numeric
```

```
@attribute Casado {Sí,No}
```

```
@attribute Coche {No,Sí}
```

```
@attribute Hijos numeric
```

```
@attribute Alq/Prop {Alquiler,Prop}
```

```
@attribute Sindic. {No,Sí}
```

```
@attribute Bajas/Año numeric
```

```
@attribute Antigüedad numeric
```

```
@attribute Sexo {H,M}
```

```
@attribute Cluster {cluster0,cluster1,cluster2}
```

```
@data
```

```
0,10000,Sí,No,0,Alquiler,No,7,15,H,cluster2
```

```
1,20000,No,Sí,1,Alquiler,Sí,3,3,M,cluster0
```

```
2,15000,Sí,Sí,2,Prop,Sí,5,10,H,cluster1
```

```
3,30000,Sí,Sí,1,Alquiler,No,15,7,M,cluster2
```

```
4,10000,Sí,Sí,0,Prop,Sí,1,6,H,cluster1
```

```
5,40000,No,Sí,0,Alquiler,Sí,3,16,M,cluster0
```

```
6,25000,No,No,0,Alquiler,Sí,0,8,H,cluster0
```

```
7,20000,No,Sí,0,Prop,Sí,2,6,M,cluster0
```

```
8,20000,Sí,Sí,3,Prop,No,7,5,H,cluster1
```

```
9,30000,Sí,Sí,2,Prop,No,1,20,H,cluster1
```

```
10,50000,No,No,0,Alquiler,No,2,12,M,cluster0
```

```
11,8000,Sí,Sí,2,Prop,No,3,1,H,cluster1
```

```
12,20000,No,No,0,Alquiler,No,27,5,M,cluster0
```

```
13,10000,No,Sí,0,Alquiler,Sí,0,7,H,cluster2
```

```
14,8000,No,Sí,0,Alquiler,No,3,2,H,cluster2
```

# *Simulación en WEKA: SimpleKmeans -> Clasificación*

- Seleccionar un método de clasificación para aprender el conjunto de entrenamiento arrojado por Kmeans.

```
@relation empleados_clustered
```

```
@attribute Instance_number numeric REMOVE
```

```
@attribute Sueldo numeric
```

```
@attribute Casado {Sí,No}
```

```
@attribute Coche {No,Sí}
```

```
@attribute Hijos numeric
```

```
@attribute Alq/Prop {Alquiler,Prop}
```

```
@attribute Sindic. {No,Sí}
```

```
@attribute Bajas/Año numeric
```

```
@attribute Antigüedad numeric
```

```
@attribute Sexo {H,M}
```

```
@attribute Cluster {cluster0,cluster1,cluster2}
```

```
@data
```

```
0,10000,Sí,No,0,Alquiler,No,7,15,H,cluster2  
1,20000,No,Sí,1,Alquiler,Sí,3,3,M,cluster0  
2,15000,Sí,Sí,2,Prop,Sí,5,10,H,cluster1  
3,30000,Sí,Sí,1,Alquiler,No,15,7,M,cluster2  
4,10000,Sí,Sí,0,Prop,Sí,1,6,H,cluster1  
5,40000,No,Sí,0,Alquiler,Sí,3,16,M,cluster0  
6,25000,No,No,0,Alquiler,Sí,0,8,H,cluster0  
7,20000,No,Sí,0,Prop,Sí,2,6,M,cluster0  
8,20000,Sí,Sí,3,Prop,No,7,5,H,cluster1  
9,30000,Sí,Sí,2,Prop,No,1,20,H,cluster1  
10,50000,No,No,0,Alquiler,No,2,12,M,cluster0  
11,8000,Sí,Sí,2,Prop,No,3,1,H,cluster1  
12,20000,No,No,0,Alquiler,No,27,5,M,cluster0  
13,10000,No,Sí,0,Alquiler,Sí,0,7,H,cluster2  
14,8000,No,Sí,0,Alquiler,No,3,2,H,cluster2
```

# *Simulación en WEKA: SimpleKmeans -> Clasificación*

- Predecir el tipo de empleado de las siguientes personas:

```
@relation empleados_tested
```

```
@attribute Sueldo numeric  
@attribute Casado {Sí,No}  
@attribute Coche {No,Sí}  
@attribute Hijos numeric  
@attribute Alq/Prop {Alquiler,Prop}  
@attribute Sindic. {No,Sí}  
@attribute Bajas/Año numeric  
@attribute Antigüedad numeric  
@attribute Sexo {H,M}  
@attribute Cluster {cluster0,cluster1,cluster2}
```

```
@data  
25000,No,No,0,Alquiler,Sí,0,8,H,?  
10000,Sí,Sí,0,Prop,Sí,1,6,H, ?  
10000,No,Sí,0,Alquiler,Sí,0,7,H, ?
```

# *Simulación en WEKA: SimpleKmeans -> Titanic*

- Datos del hundimiento del Titanic. Los datos se encuentran en el fichero "titanic.arff" y corresponden a las características de los 2.201 pasajeros del Titanic. Estos datos son reales y se han obtenido de: "*Report on the Loss of the 'Titanic' (S.S.)*" (1990), *British Board of Trade Inquiry Report\_ (reprint)*, Gloucester, UK: Allan Sutton Publishing.
- Los atributos son:
  - Clase (0 = tripulación, 1 = primera, 2 = segunda, 3 = tercera)
  - Edad (1 = adulto, 0 = niño)
  - Sexo (1 = hombre, 0 = mujer)
  - Sobrevivió (1 = sí, 0 = no)

# *Simulación en WEKA: SimpleKmeans -> Titanic*

@relation titanic

@attribute Clase {0,1,2,3}  
@attribute Edad {0,1}  
@attribute Sexo {0,1}  
@attribute Sobrevivió? {0,1}

@data

1,1,1,1  
1,1,1,1  
1,1,1,1  
1,1,1,1

- Qué información se puede descubrir al construir 4 grupos?
- Qué información se puede descubrir al construir 6 grupos?
- Qué información se puede descubrir al construir 10 grupos?

# *Simulación en WEKA: SimpleKmeans -> Titanic*

Cluster centroids:

Attribute	Full Data (2201)	Cluster#					
		0 (314)	1 (508)	2 (673)	3 (254)	4 (387)	5 (65)
<hr/>							
Clase	0	2	3	0	0	3	3
Edad	1	1	1	1	1	1	0
Sexo	1	1	0	1	1	1	1
Sobrevivió?	0	0	1	0	1	0	0

- Las únicas personas de tercera clase que sobrevivieron eran mujeres adultas.

Time taken to build model (full training data) : 0.02 seconds

==== Model and evaluation on training set ===

Clustered Instances

0	314 ( 14%)
1	508 ( 23%)
2	673 ( 31%)
3	254 ( 12%)
4	387 ( 18%)
5	65 ( 3%)

- Clase (0 = tripulación, 1 = primera, 2 = segunda, 3 = tercera)
- Edad (1 = adulto, 0 = niño)
- Sexo (1 = hombre, 0 = mujer)
- Sobrevivió (1 = sí, 0 = no)

# AGENDA

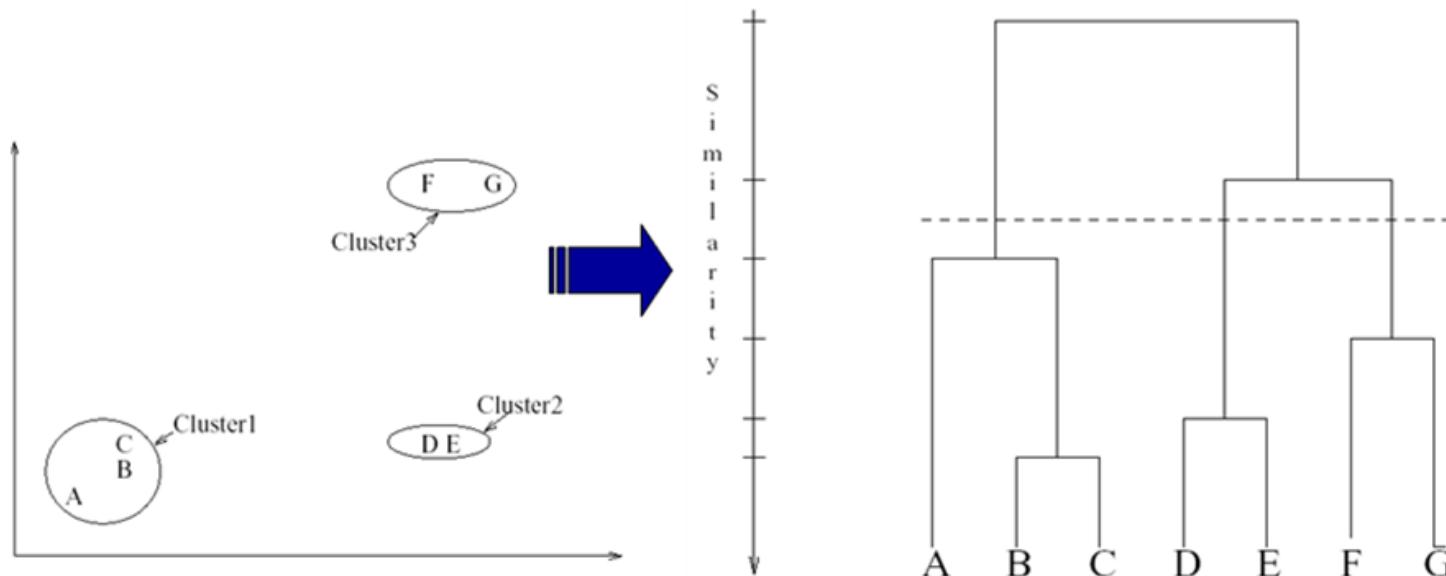


1. Método Particional
- 2. Método Jerárquico**
3. Método Probabilístico
4. Redes Neuronales
5. Reglas de Asociación: A priori
6. Votación para selección de factores

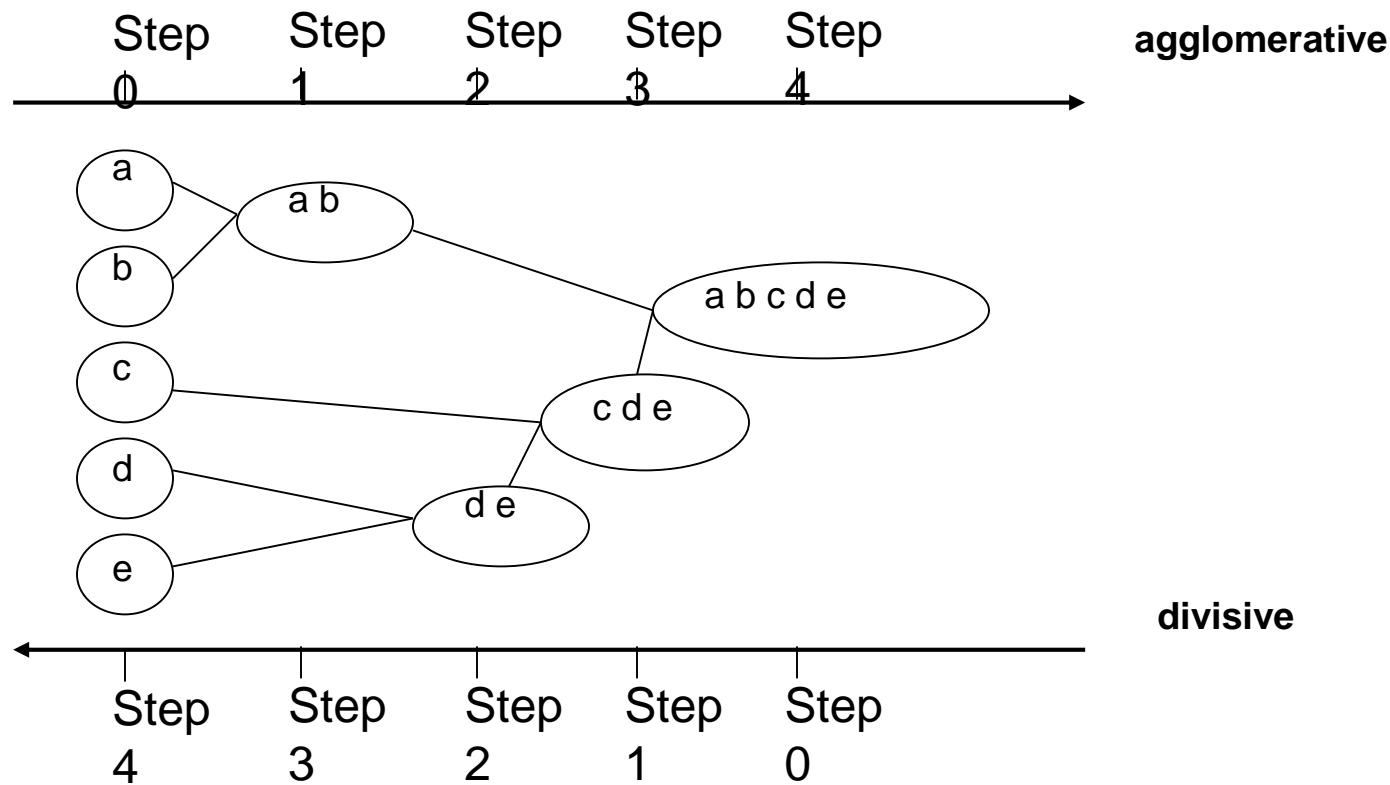
# Métodos Jerárquicos



Busca subgrupos recursivamente creando un dendograma o árbol binario de acuerdo a una matriz de proximidad. No requiere el número k de clústers como entrada, pero requiere una condición de finalización.



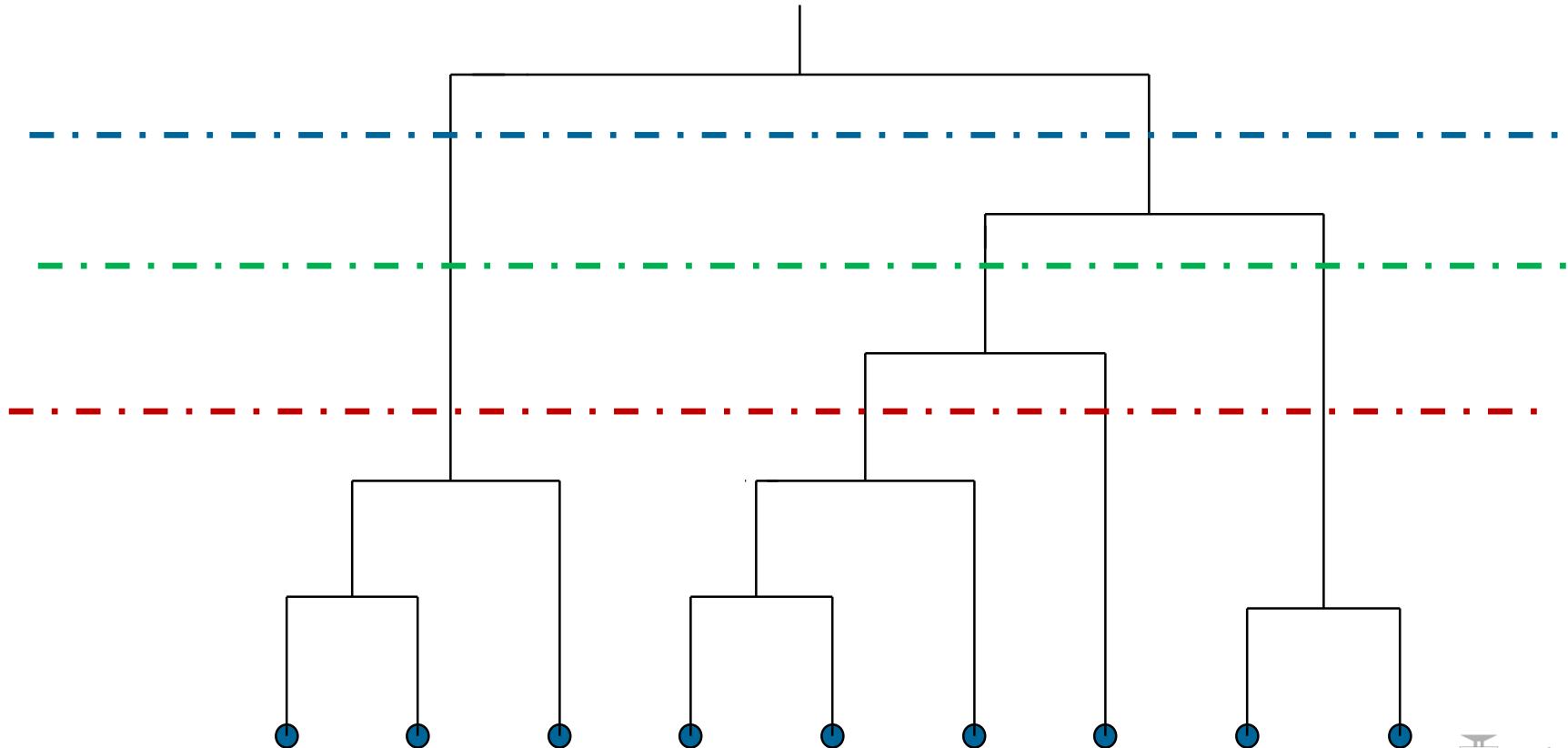
# Métodos Jerárquicos



# Métodos Jerárquicos



Las estructuras de clustering son obtenidas cortando el dendograma a un nivel específico, de manera que los componentes conectados se encuentran en el mismo clúster.



# *Métodos Jerárquicos*

- **Single-linkage**

La distancia entre clusters es igual a la distancia más corta entre cualquier par de miembros de los grupos.



- Single-Link Method / Nearest Neighbor

# *Métodos Jerárquicos*

- **Complete-linkage**

La distancia entre clusters es igual a la distancia más larga entre cualquier par de miembros de los grupos.

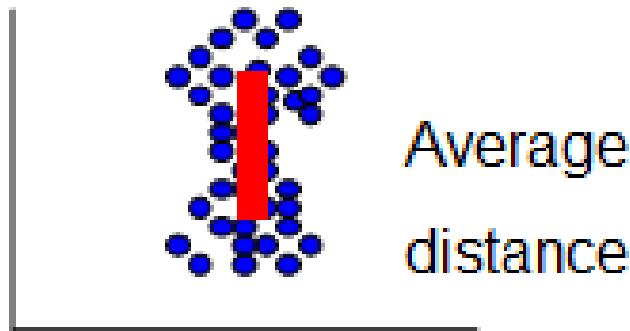


Complete-Link / Furthest Neighbor

# *Métodos Jerárquicos*

- **Average-linkage**

La distancia entre clusters es igual a la distancia promedio de los miembros de los grupos.



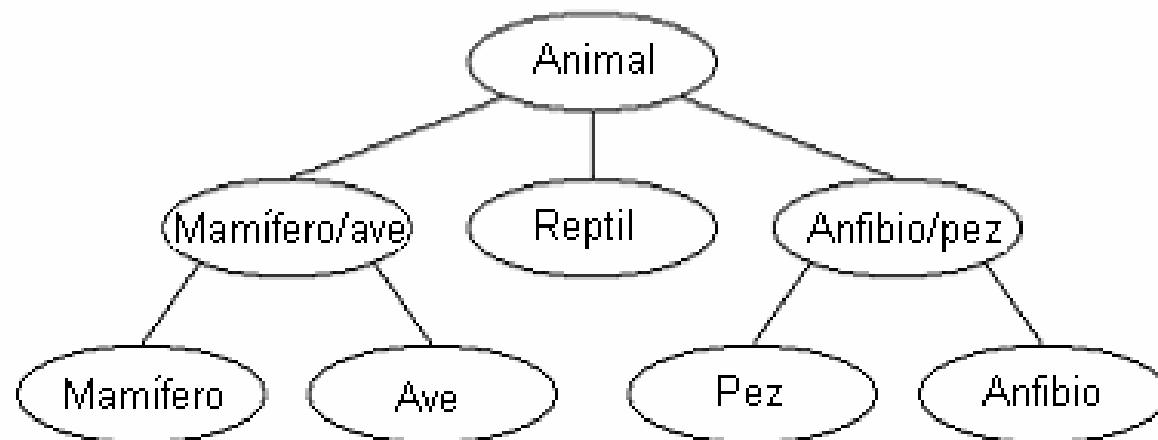
Average of all cross-cluster pairs.

# *COBWEB: Clustering conceptual*

Datos de Entrada

Nombre	Recubierto de	Cavidades corazón	temperatura del cuerpo	Fertilización
Mamífero	Pelo	Cuatro	Regulada	Interna
Pez	Escamas	Dos	Sin regulación	Externa
Anfibio	Piel húmeda	Tres	Sin regulación	Externa
Ave	Plumas	Cuatro	Regulada	Interna
Reptil	Piel dura	Cuatro Imperfectas	Sin regulación	Interna

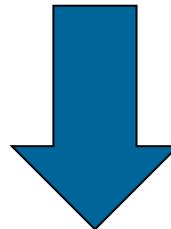
Árbol Generado



# *COBWEB: Clustering conceptual*

- COBWEB forma los conceptos por agrupación de ejemplos con atributos similares.

Cluster



Distribución de probabilidad sobre el espacio de los valores de los atributos, generando un árbol de agrupación jerárquica.

# COBWEB

- Algoritmo jerárquico incremental. Al principio, el árbol consiste en un **único nodo raíz**. Las instancias se van añadiendo una a una y el árbol se va actualizando en cada paso.
- En el árbol **cada nodo es un concepto** que tiene una descripción probabilística de ese concepto que resume los objetos clasificados bajo cada nodo.

$$(P(C_i))$$

Probabilidad del concepto

$$(P(A_i = V_{ij} | C_k)).$$

Probabilidades condicionales de pares  
atributos-valor dado el concepto

- La actualización del árbol consiste en encontrar el mejor sitio donde incluir la nueva instancia, operación que puede necesitar de la reestructuración de todo el árbol (generar un nuevo nodo o la fusión de nodos existentes).

# COBWEB

- Utilidad de categoría:

$$CU = \frac{\sum_{k=1}^n P(C_k) \left[ \sum_i \sum_j P(A_i = V_{ij} | C_k)^2 - \sum_i \sum_j P(A_i = V_{ij})^2 \right]}{n}$$

Calidad general de una partición de instancias en un segmento

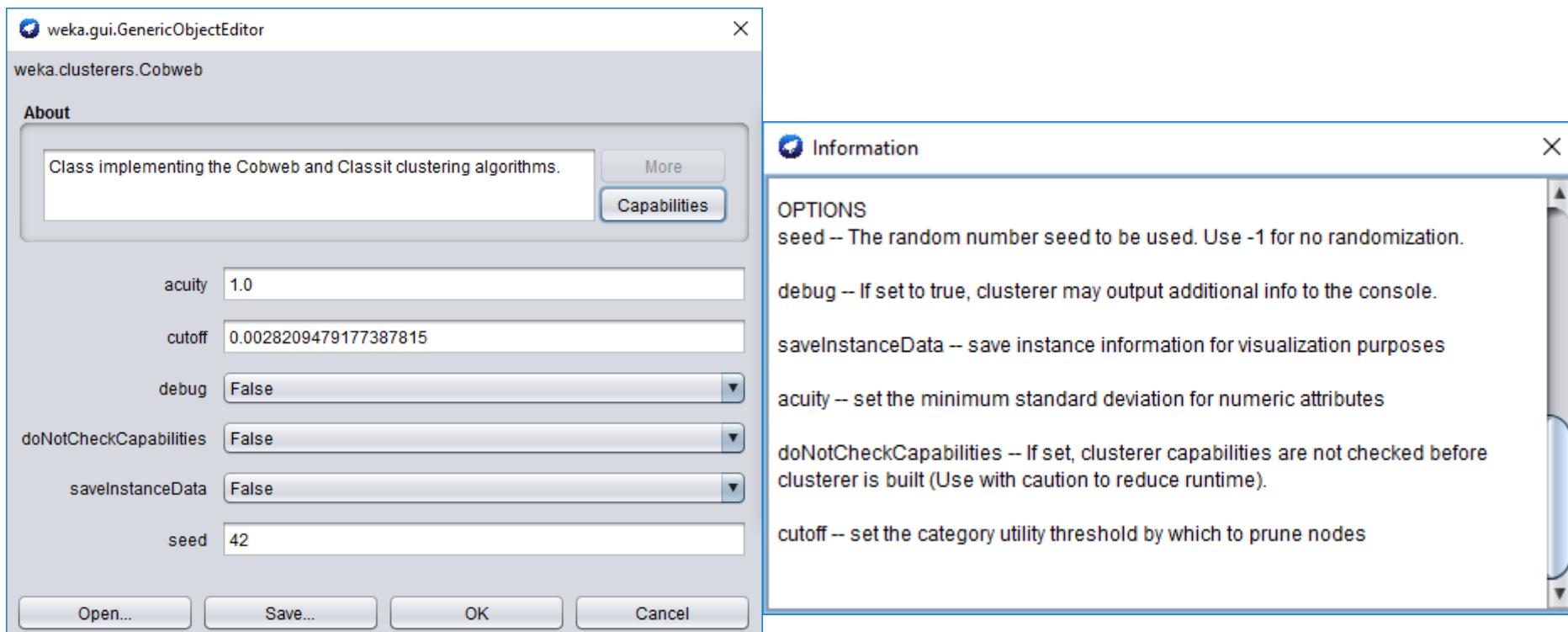
- Acuity: valor de varianza mínimo para atributos numéricos.

- Cut-off: Indica el grado de mejoría que se debe producir en la utilidad de categoría para que la instancia sea tenida en cuenta de manera individual. **Este valor se utiliza para evitar el crecimiento desmesurado del número de clusters.**

# *Simulación en WEKA*

- La empresa de software para Internet “Memolum Web” quiere extraer tipologías de empleados, con el objetivo de hacer una política de personal más fundamentada y seleccionar a qué grupos incentivar.
- Las variables que se recogen de las fichas de los 15 empleados de la empresa son:
  - Sueldo: sueldo anual en euros.
  - Casado: si está casado o no.
  - Coche: si viene en coche a trabajar (o al menos si lo aparcá en el parking de la empresa).
  - Hijos: si tiene hijos.
  - Alq/Prop: si vive en una casa alquilada o propia.
  - Sindic: si pertenece al sindicato revolucionario de Internet
  - Bajas/Año: media del nº de bajas por año
  - Antigüedad: antigüedad en la empresa
  - Sexo: H: hombre, M: mujer.
- Se intenta extraer grupos de entre estos quince empleados.

# COBWEB



- **Acuity:** valor de varianza mínimo en atributos numéricos.
- **Cut-off:** Indica el grado de mejoría que se debe producir en la utilidad de categoría para que la instancia sea tenida en cuenta de manera individual

# COBWEB

Clustered Instances

```

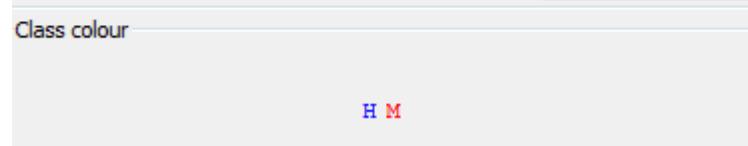
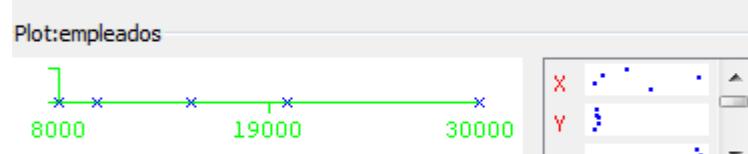
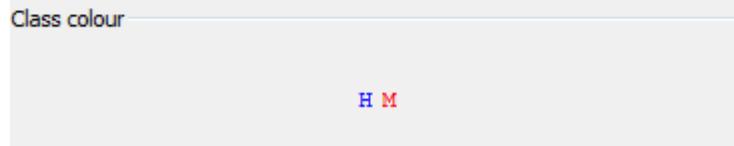
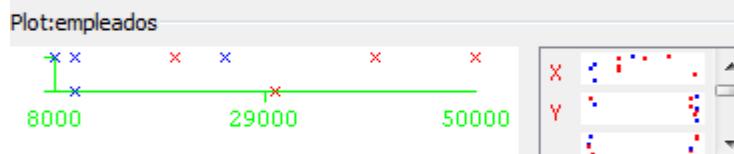
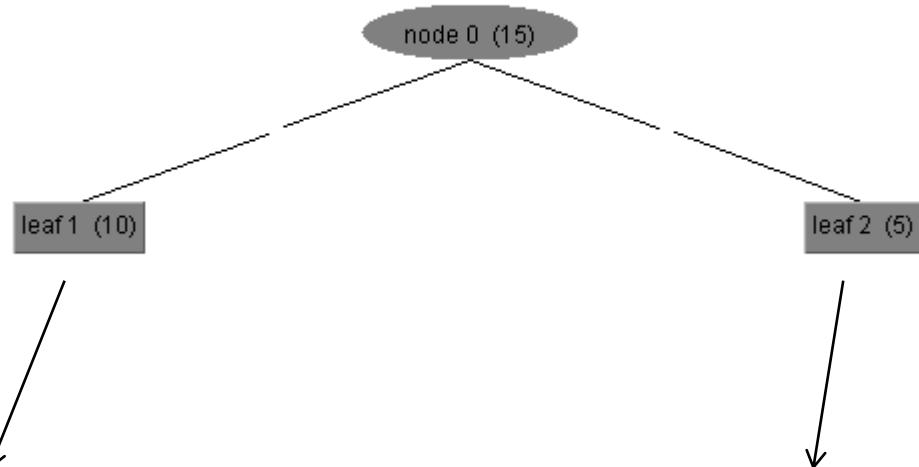
1      10 ( 67%)
2      5 ( 33%)

```

```

node 0 [15]
|   leaf 1 [10]
node 0 [15]
|   leaf 2 [5]

```



# COBWEB

```
@data
0,10000,Sí,No,0,Alquiler,No,7,15,H,cluster1
1,20000,No,Sí,1,Alquiler,Sí,3,3,M,cluster1
2,15000,Sí,Sí,2,Prop,Sí,5,10,H,cluster2
3,30000,Sí,Sí,1,Alquiler,No,15,7,M,cluster1
4,10000,Sí,Sí,0,Prop,Sí,1,6,H,cluster2
5,40000,No,Sí,0,Alquiler,Sí,3,16,M,cluster1
6,25000,No,No,0,Alquiler,Sí,0,8,H,cluster1
7,20000,No,Sí,0,Prop,Sí,2,6,M,cluster1
8,20000,Sí,Sí,3,Prop,No,7,5,H,cluster2
9,30000,Sí,Sí,2,Prop,No,1,20,H,cluster2
10,50000,No,No,0,Alquiler,No,2,12,M,cluster1
11,8000,Sí,Sí,2,Prop,No,3,1,H,cluster2
12,20000,No,No,0,Alquiler,No,27,5,M,cluster1
13,10000,No,Sí,0,Alquiler,Sí,0,7,H,cluster1
14,8000,No,Sí,0,Alquiler,No,3,2,H,cluster1|
```

```
|resultados,H,S,C,ON,TELITUPLA,O,IS,ON,8000,4|
|resultados,H,S,O,IS,TELITUPLA,O,IS,ON,00001,31|
|resultados,H,S,2,AS,ON,TELITUPLA,O,ON,ON,00005,IS|
|resultados,H,T,CON,GOYA,S,IS,8000,11|
```

# AGENDA



1. Método Particional
2. Método Jerárquico
- 3. Método Probabilístico**
4. Redes Neuronales
5. Reglas de Asociación: A priori
6. Votación para selección de factores

# *Métodos Probabilísticos*

- Asumen que los datos son generados de acuerdo a una distribución de probabilidad.

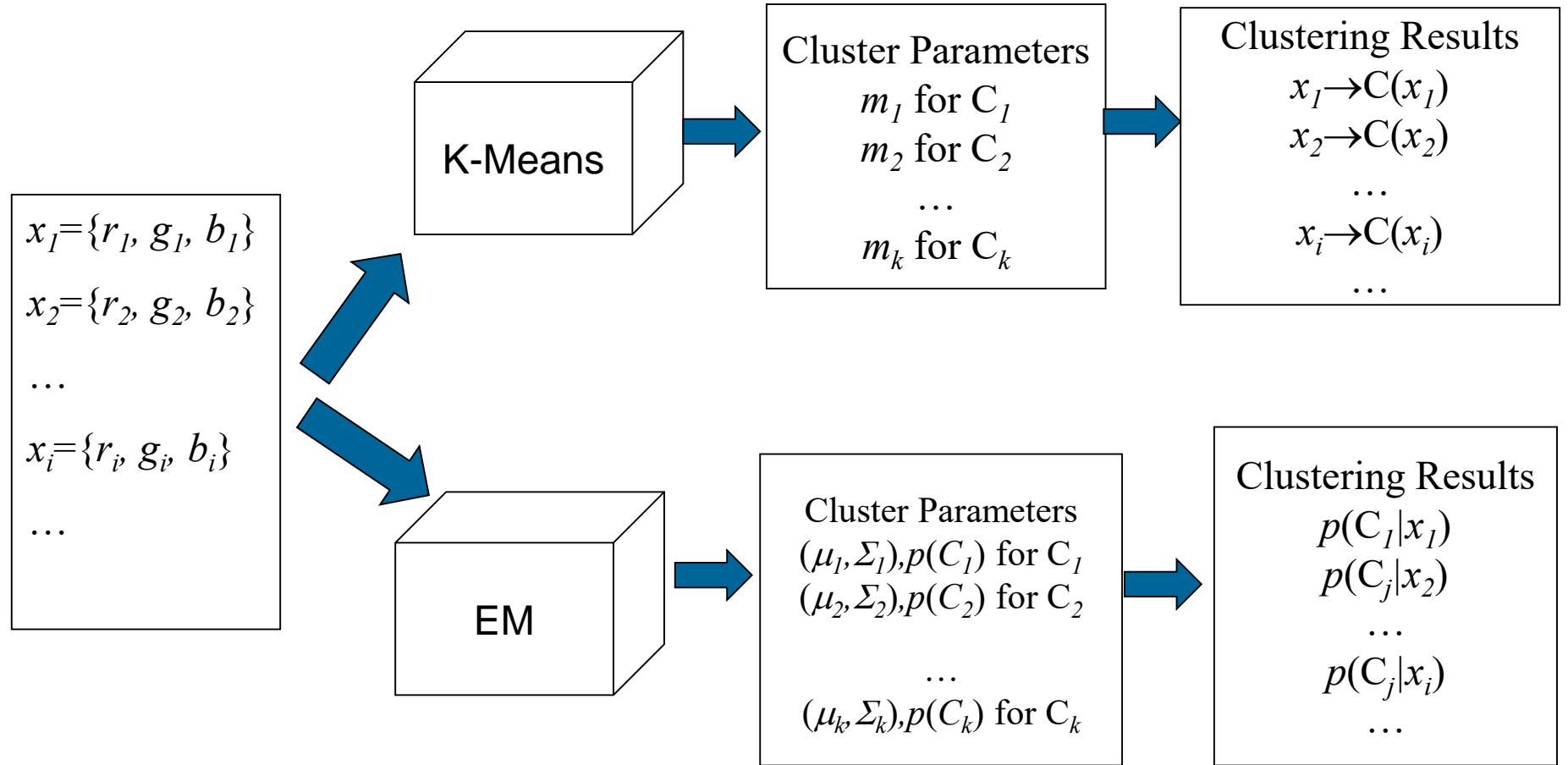
Se tienen  $k$  distribuciones de probabilidad que representan los  $k$  clusters.

No depende del orden de los ejemplos.

Se busca el grupo de clusters más probables dados los datos.

Los ejemplos tienen ciertas probabilidades de pertenecer a un cluster.

# *K-means vs EM*



# *Algoritmo EM – Expectation Maximization*

El algoritmo empieza adivinando los parámetros de las distribuciones y, a continuación, los utiliza para calcular las probabilidades de que cada objeto pertenezca a un cluster y usa esas probabilidades para re-estimar los parámetros de las probabilidades, hasta converger.

1. **Expectation:** cálculo de las probabilidades de los grupos o los valores esperados de los grupos.
2. **Maximization:** cálculo de los valores de los parámetros de las distribuciones, maximizando la verosimilitud de las distribuciones dados los datos.



**Log-likelihood:** maximiza cómo de bien encajan los datos sobre la distribución que los representa.

# *Algoritmo EM*

- Boot Step:

- Initialize  $K$  clusters:  $C_1, \dots, C_K$   
 $(\mu_j, \Sigma_j)$  and  $P(C_j)$  for each cluster  $j$ .

- Iteration Step:

- Estimate the cluster of each data

→ **Expectation**

$$p(C_j | x_i)$$

- Re-estimate the cluster parameters

→ **Maximization**

$$(\mu_j, \Sigma_j), p(C_j) \text{ For each cluster } j$$

# *Algoritmo EM*

- Boot Step:

- Initialize  $K$  clusters:  $C_1, \dots, C_K$   
 $(\mu_j, \Sigma_j)$  and  $P(C_j)$  for each cluster  $j$ .

- Iteration Step:

- Estimate the cluster of each data

→ **Expectation**

$$p(C_j | x_i) = \frac{p(x_i | C_j) \cdot p(C_j)}{p(x_i)} = \frac{p(x_i | C_j) \cdot p(C_j)}{\sum_j p(x_i | C_j) \cdot p(C_j)}$$

- Re-estimate the cluster parameters

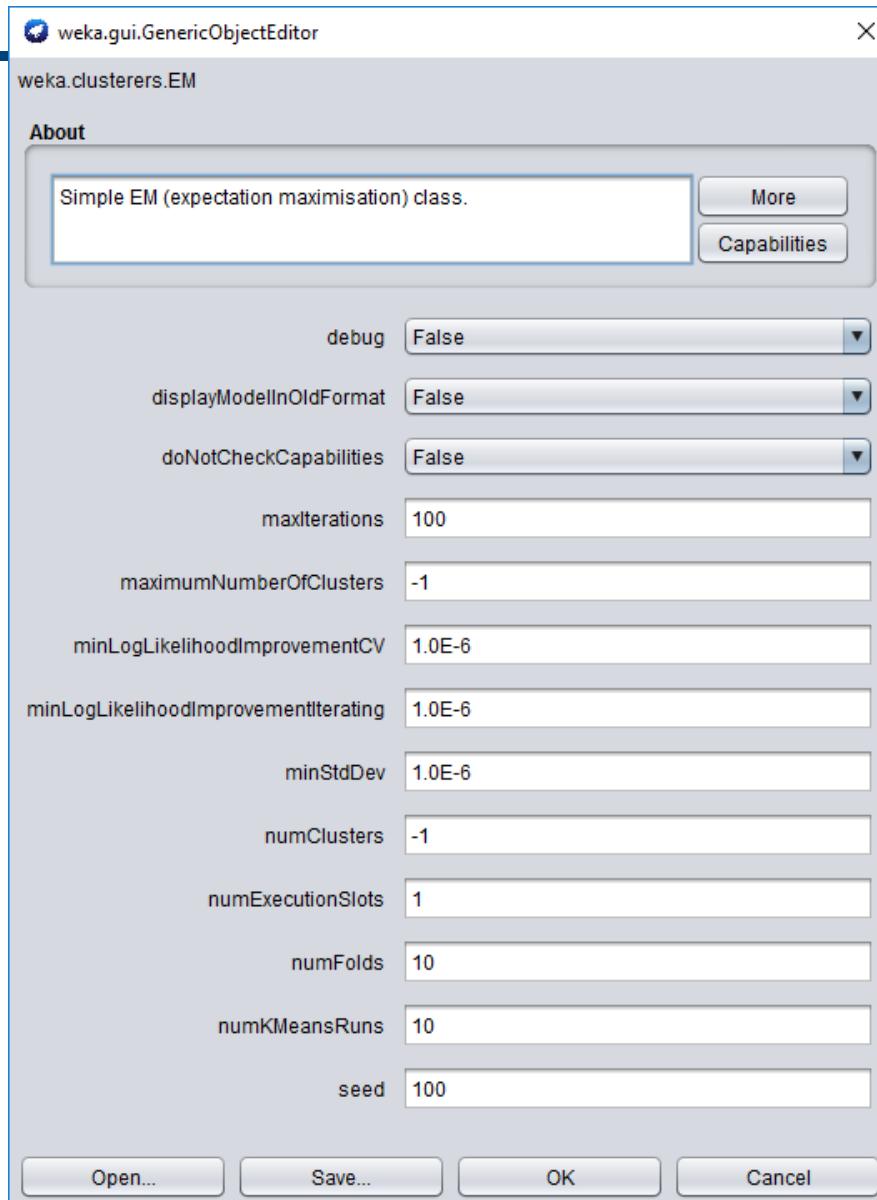
→ **Maximization**

$$\mu_j = \frac{\sum_i p(C_j | x_i) \cdot x_i}{\sum_i p(C_j | x_i)} \quad \Sigma_j = \frac{\sum_i p(C_j | x_i) \cdot (x_i - \mu_j) \cdot (x_i - \mu_j)^T}{\sum_i p(C_j | x_i)} \quad p(C_j) = \frac{\sum_i p(C_j | x_i)}{N}$$

# *Simulación en WEKA*

- La empresa de software para Internet “Memolum Web” quiere extraer tipologías de empleados, con el objetivo de hacer una política de personal más fundamentada y seleccionar a qué grupos incentivar.
- Las variables que se recogen de las fichas de los 15 empleados de la empresa son:
  - Sueldo: sueldo anual en euros.
  - Casado: si está casado o no.
  - Coche: si viene en coche a trabajar (o al menos si lo aparcá en el parking de la empresa).
  - Hijos: si tiene hijos.
  - Alq/Prop: si vive en una casa alquilada o propia.
  - Sindic: si pertenece al sindicato revolucionario de Internet
  - Bajas/Año: media del nº de bajas por año
  - Antigüedad: antigüedad en la empresa
  - Sexo: H: hombre, M: mujer.
- Se intenta extraer grupos de entre estos quince empleados.

# EM



# EM

	Cluster	
Attribute	0 (0.61)	1 (0.39)
<hr/>		
Sueldo		
mean	24480.7952	15748.2561
std. dev.	12823.4476	7684.6102
Casado		
Si	2.4052	6.5948
No	8.7303	1.2697
[total]	11.1355	7.8645
Coche		
No	4.3246	1.6754
Si	6.8109	6.1891
[total]	11.1355	7.8645
Hijos		
mean	0.2132	1.5436
std. dev.	0.4096	1.0987
Alq/Prop		
Alquiler	9.1001	1.8999
Prop	2.0354	5.9646
[total]	11.1355	7.8645
Sindic.		
No	5.1818	4.8182
Sí	5.9537	3.0463
[total]	11.1355	7.8645
Bajas/Año		
mean	6.2062	3.8031
std. dev.	8.4439	2.4415

## Clustered Instances

0	9	( 60%)
1	6	( 40%)

# *EM*

```
@data
0,10000,Si,No,0,Alquiler,No,7,15,H,cluster1
1,20000,No,Sí,1,Alquiler,Sí,3,3,M,cluster0
2,15000,Sí,Sí,2,Prop,Sí,5,10,H,cluster1
3,30000,Sí,Sí,1,Alquiler,No,15,7,M,cluster0
4,10000,Sí,Sí,0,Prop,Sí,1,6,H,cluster1
5,40000,No,Sí,0,Alquiler,Sí,3,16,M,cluster0
6,25000,No,No,0,Alquiler,Sí,0,8,H,cluster0
7,20000,No,Sí,0,Prop,Sí,2,6,M,cluster0
8,20000,Sí,Sí,3,Prop,No,7,5,H,cluster1
9,30000,Sí,Sí,2,Prop,No,1,20,H,cluster1
10,50000,No,No,0,Alquiler,No,2,12,M,cluster0
11,8000,Sí,Sí,2,Prop,No,3,1,H,cluster1
12,20000,No,No,0,Alquiler,No,27,5,M,cluster0
13,10000,No,Sí,0,Alquiler,Sí,0,7,H,cluster0
14,8000,No,Sí,0,Alquiler,No,3,2,H,cluster0
```

# Comparación

```
@data  
0,10000,Si,No,0,Alquiler,No,7,15,H,cluster1  
1,20000,No,Si,1,Alquiler,Si,3,3,M,cluster0  
2,15000,Si,Si,2,Prop,Si,5,10,H,cluster1  
3,30000,Si,Si,1,Alquiler,No,15,7,M,cluster0  
4,10000,Si,Si,0,Prop,Si,1,6,H,cluster1  
5,40000,No,Si,0,Alquiler,Si,3,16,M,cluster0  
6,25000,No,No,0,Alquiler,Si,0,8,H,cluster0  
7,20000,No,Si,0,Prop,Si,2,6,M,cluster0  
8,20000,Si,Si,3,Prop,No,7,5,H,cluster1  
9,30000,Si,Si,2,Prop,No,1,20,H,cluster1  
10,50000,No,No,0,Alquiler,No,2,12,M,cluster0  
11,8000,Si,Si,2,Prop,No,3,1,H,cluster1  
12,20000,No,No,0,Alquiler,No,27,5,M,cluster0  
13,10000,No,Si,0,Alquiler,Si,0,7,H,cluster0  
14,8000,No,Si,0,Alquiler,No,3,2,H,cluster0
```

```
@data  
0,10000,Si,No,0,Alquiler,No,7,15,H,cluster1  
1,20000,No,Si,1,Alquiler,Si,3,3,M,cluster1  
2,15000,Si,Si,2,Prop,Si,5,10,H,cluster2  
3,30000,Si,Si,1,Alquiler,No,15,7,M,cluster1  
4,10000,Si,Si,0,Prop,Si,1,6,H,cluster2  
5,40000,No,Si,0,Alquiler,Si,3,16,M,cluster1  
6,25000,No,No,0,Alquiler,Si,0,8,H,cluster1  
7,20000,No,Si,0,Prop,Si,2,6,M,cluster1  
8,20000,Si,Si,3,Prop,No,7,5,H,cluster2  
9,30000,Si,Si,2,Prop,No,1,20,H,cluster2  
10,50000,No,No,0,Alquiler,No,2,12,M,cluster1  
11,8000,Si,Si,2,Prop,No,3,1,H,cluster2  
12,20000,No,No,0,Alquiler,No,27,5,M,cluster1  
13,10000,No,Si,0,Alquiler,Si,0,7,H,cluster1  
14,8000,No,Si,0,Alquiler,No,3,2,H,cluster1
```

EM

K-means

```
@data  
0,10000,Si,No,0,Alquiler,No,7,15,H,cluster1  
1,20000,No,Si,1,Alquiler,Si,3,3,M,cluster0  
2,15000,Si,Si,2,Prop,Si,5,10,H,cluster1  
3,30000,Si,Si,1,Alquiler,No,15,7,M,cluster0  
4,10000,Si,Si,0,Prop,Si,1,6,H,cluster1  
5,40000,No,Si,0,Alquiler,Si,3,16,M,cluster0  
6,25000,No,No,0,Alquiler,Si,0,8,H,cluster0  
7,20000,No,Si,0,Prop,Si,2,6,M,cluster0  
8,20000,Si,Si,3,Prop,No,7,5,H,cluster1  
9,30000,Si,Si,2,Prop,No,1,20,H,cluster1  
10,50000,No,No,0,Alquiler,No,2,12,M,cluster0  
11,8000,Si,Si,2,Prop,No,3,1,H,cluster1  
12,20000,No,No,0,Alquiler,No,27,5,M,cluster0  
13,10000,No,Si,0,Alquiler,Si,0,7,H,cluster0  
14,8000,No,Si,0,Alquiler,No,3,2,H,cluster0
```

Cobweb

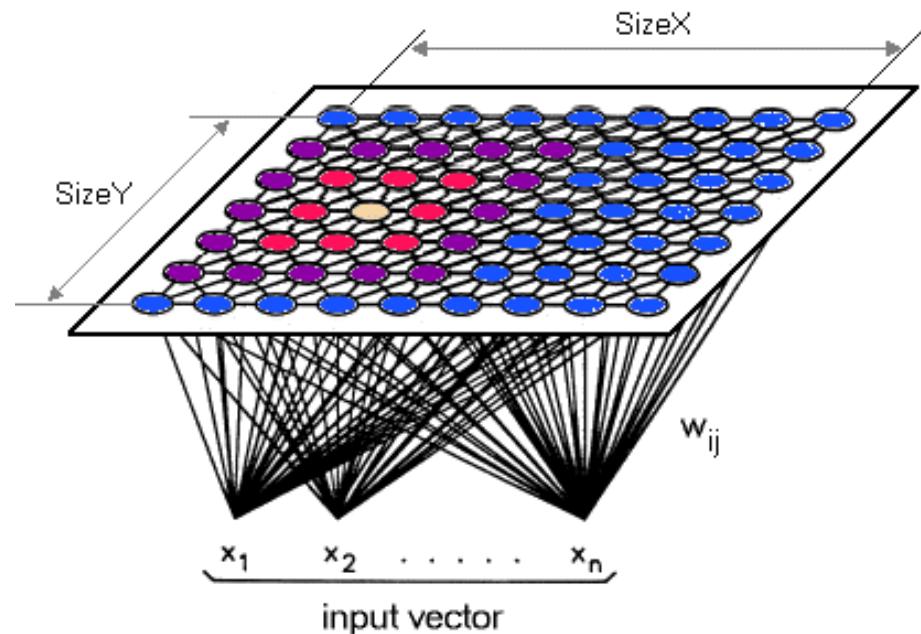
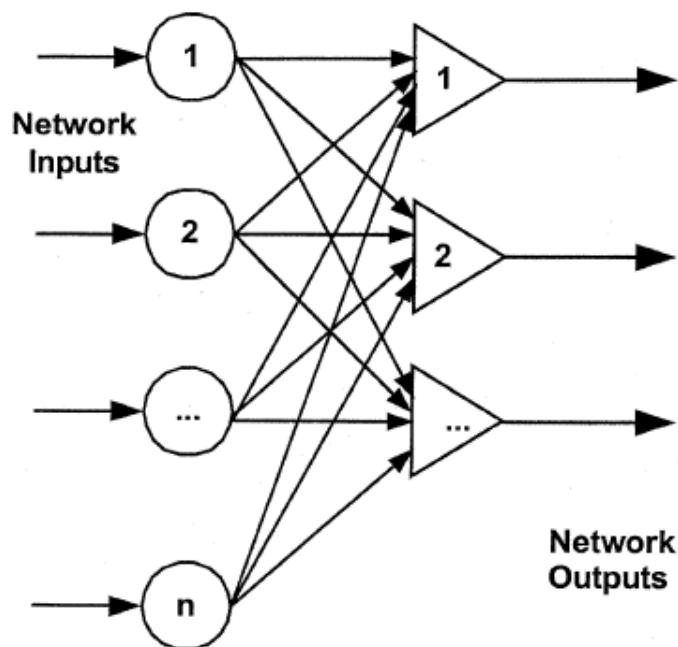
# AGENDA



1. Método Particional
2. Método Jerárquico
3. Método Probabilístico
- 4. Redes Neuronales**
5. Reglas de Asociación: A priori
6. Votación para selección de factores

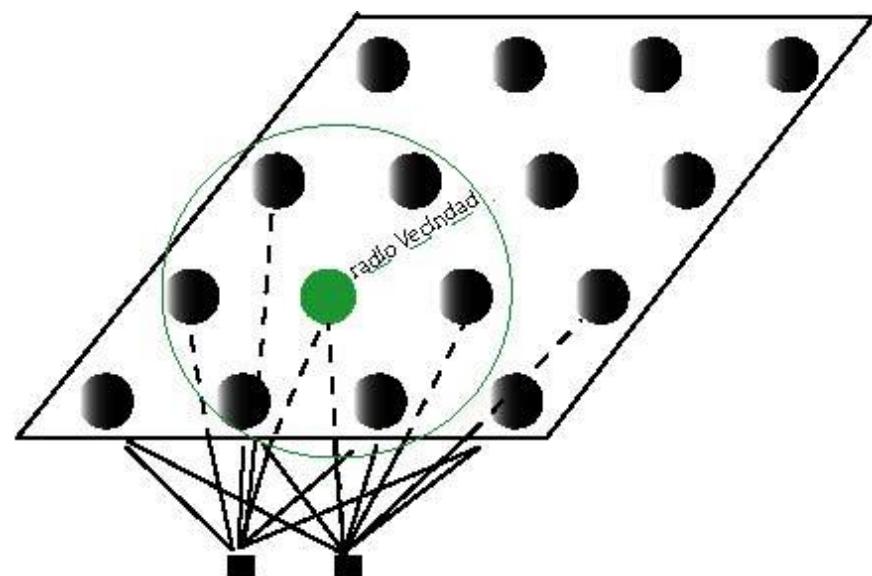
# SOM- Mapas Auto\_Organizados

Los Mapas Autoorganizativos o Self-Organizing Maps es un modelo de red neuronal que procesa los datos formando un mapa (usualmente bidimensional) donde casos similares se “mapean” en regiones cercanas. De esta manera “vecindad” significa “similaridad”.

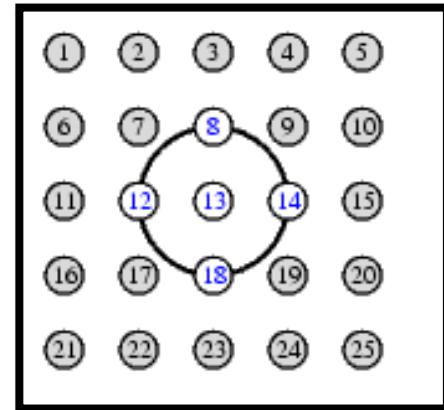
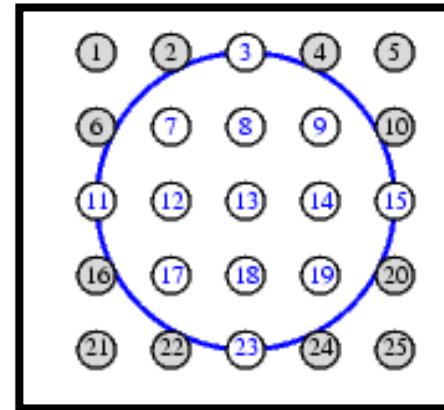
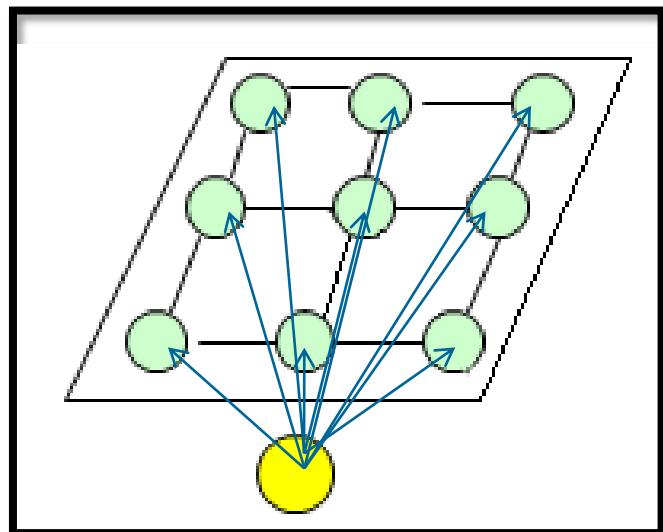


# *SOM- Mapas Auto\_Organizados*

1. Se inicializan los pesos de cada nodo (por ej. aleatoriamente.)
2. Se presenta una entrada a la red.
3. Se busca el nodo ganador
4. Se actualizan los pesos del nodo ganador y de sus vecinos.
5. Se vuelve al paso 2 hasta que se satisface el criterio de detención impuesto.



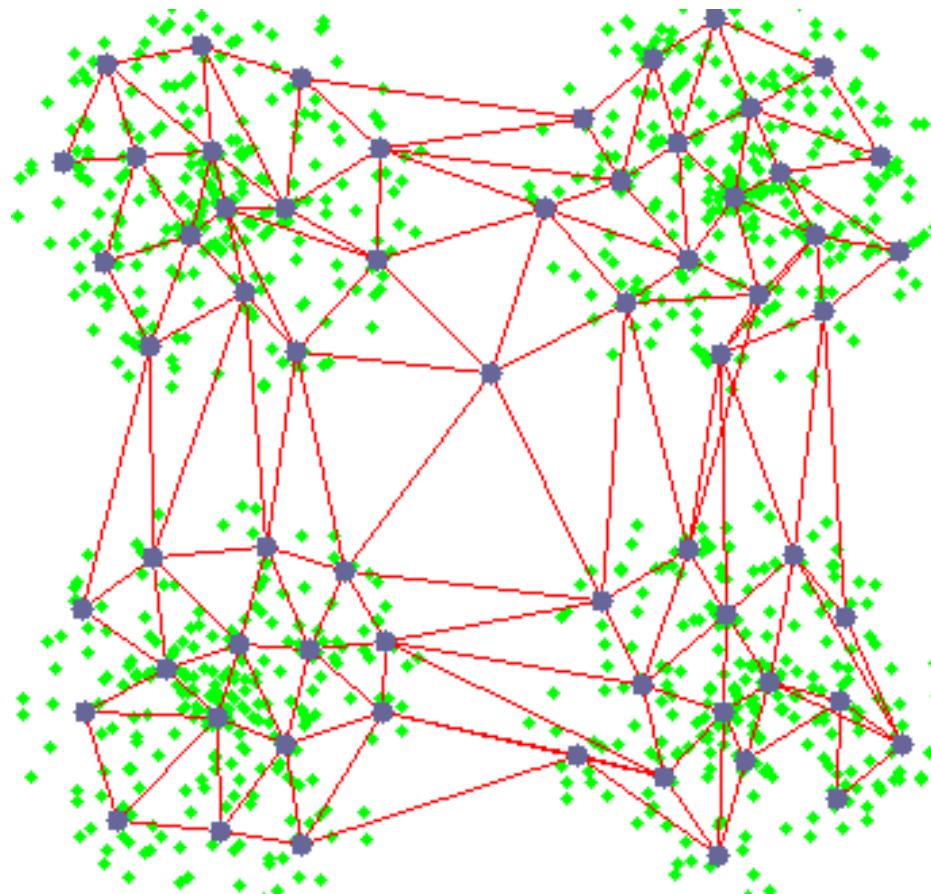
# SOM- Mapas Auto\_Organizados



Vector siendo expuesto a la malla.

$$w_q^{s+1} = w_q^s + \alpha^s (X_i - w_q^s).$$

# *SOM- Mapas Auto\_Organizados*



# *SOM- Mapas Auto\_Organizados*

---

- Diseño y configuración de la red:

- Selección de la dimensión de la red
- Selección de la cantidad de neuronas
- Selección de una medida de similaridad
- Inicialización aleatoria de pesos
- Selección del coeficiente de aprendizaje
- Elección de una vecindad inicial.

# AGENDA

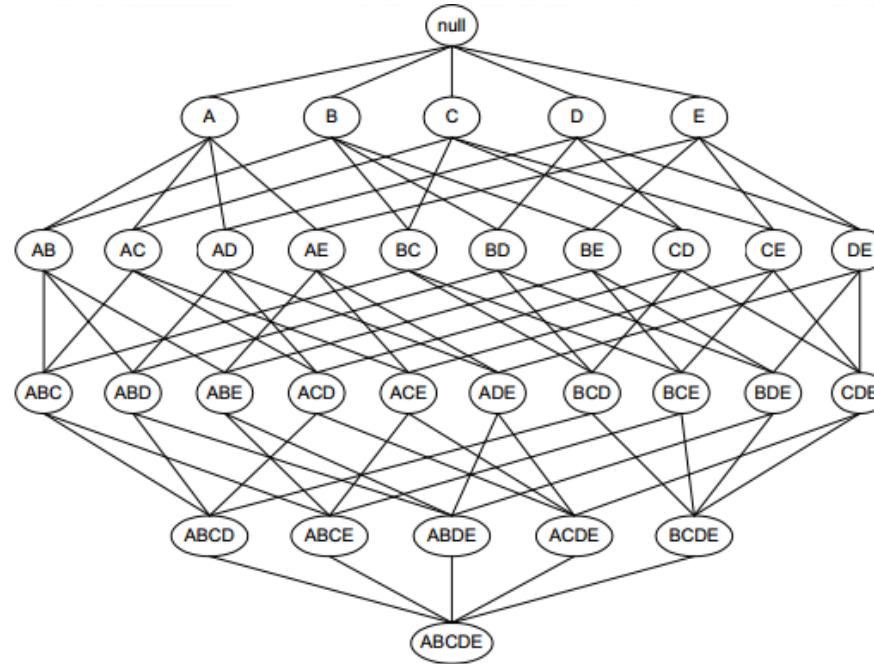


1. Método Particional
2. Método Jerárquico
3. Método Probabilístico
4. Redes Neuronales
- 5. Reglas de Asociación: A priori**
6. Votación para selección de factores

# *Apriori*

A priori realiza las búsqueda de reglas de asociación válidas según la frecuencia de los elementos bajo el siguiente supuesto:

**“Si un conjunto de elementos es frecuente, también lo son todos sus subconjuntos”**

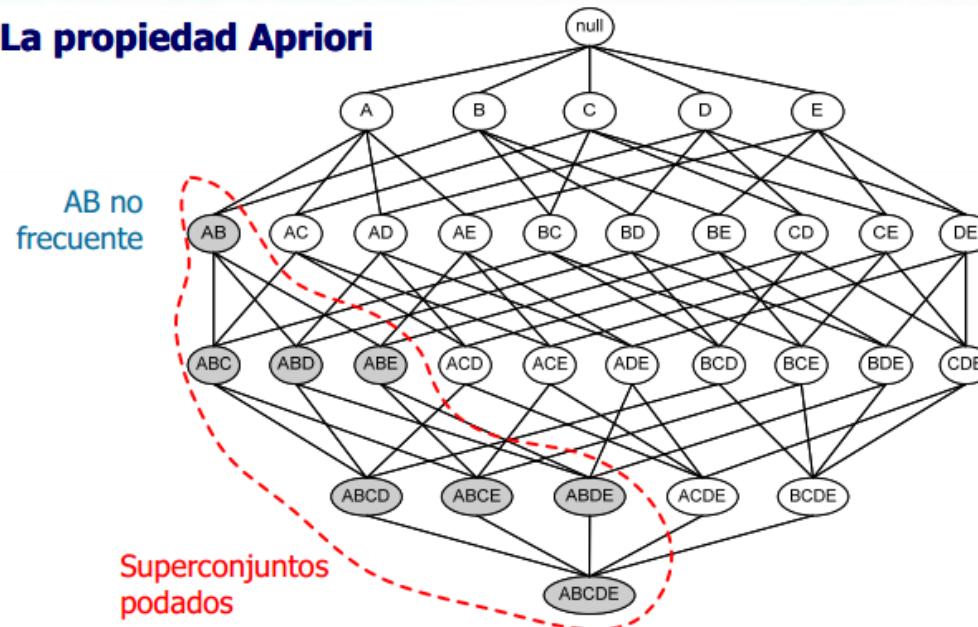


# *Apriori*

Apriori realiza las búsquedas de reglas de asociación válidas según la frecuencia de los elementos bajo el siguiente supuesto:

**“Si un conjunto de elementos es frecuente, también lo son todos sus subconjuntos”**

**La propiedad Apriori**



**Supp(elementos) >= Min\_supp**

# *Simulación en WEKA: A priori → Titanic*

- Datos del hundimiento del Titanic. Los datos se encuentran en el fichero "titanic.arff" y corresponden a las características de los 2.201 pasajeros del Titanic. Estos datos son reales y se han obtenido de: "*Report on the Loss of the 'Titanic' (S.S.)*" (1990), *British Board of Trade Inquiry Report\_ (reprint)*, Gloucester, UK: Allan Sutton Publishing.
- Los atributos son:
  - Clase (0 = tripulación, 1 = primera, 2 = segunda, 3 = tercera)
  - Edad (1 = adulto, 0 = niño)
  - Sexo (1 = hombre, 0 = mujer)
  - Sobrevivió (1 = sí, 0 = no)

# *Simulación en WEKA: A priori → Titanic*

---

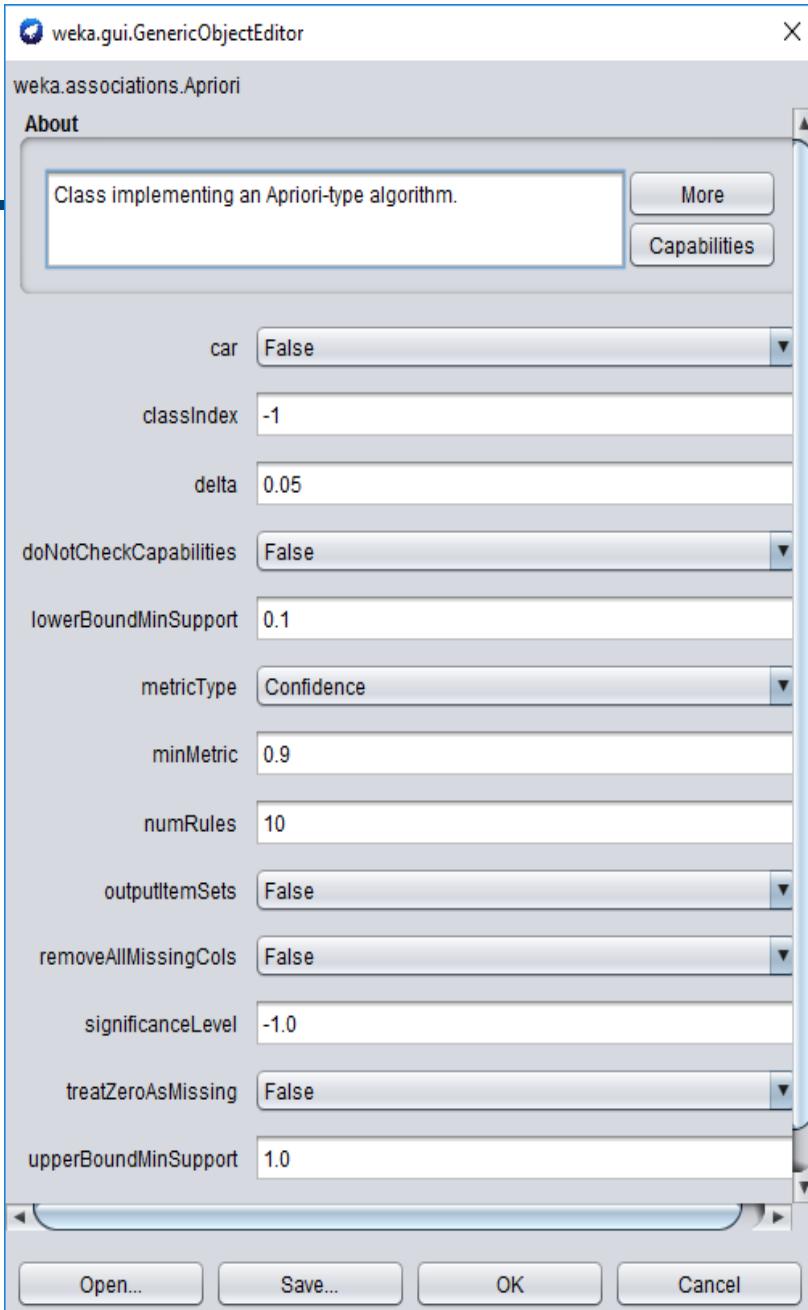
@relation titanic

@attribute Clase {0,1,2,3}  
@attribute Edad {0,1}  
@attribute Sexo {0,1}  
@attribute Sobrevivió? {0,1}

@data

1,1,1,1  
1,1,1,1  
1,1,1,1  
1,1,1,1

# Apriori



- **car** – True si se tiene un atributo de clase.
- **classIndex** – Posición del atributo de clase (-1 si está al final).
- **delta** – decrecimiento del soporte.
- **lowerBoundMinSupport** – límite inferior para el soporte
- **metricType** – Medida de evaluación de las reglas.
- **minMetric** – Valor mínimo para el evaluación de las reglas.
- **numRules** – cantidad de reglas a buscar.
- **outputItemSets** – True para conjuntos de elementos en el consecuente.
- **removeAllMissingCols** – True para eliminar comlumnas con datos faltantes.
- **significanceLevel** – Nivel de significancia para la confianza.
- **upperBoundMinSupport** – Límite superior para el soporte. Iterativamendt decrede.

# *Simulación en WEKA: A priori → Titanic*

1. Sexo=1 Sobrevivió?=0 1364 => Edad=1 1329 conf: (0.97)
2. Sobrevivió?=0 1490 => Edad=1 1438 conf: (0.97)
3. Sexo=1 1731 => Edad=1 1667 conf: (0.96)
4. Edad=1 Sobrevivió?=0 1438 => Sexo=1 1329 conf: (0.92)
5. Sobrevivió?=0 1490 => Sexo=1 1364 conf: (0.92)
6. Sobrevivió?=0 1490 => Edad=1 Sexo=1 1329 conf: (0.89)
7. Edad=1 Sexo=1 1667 => Sobrevivió?=0 1329 conf: (0.8)
8. Edad=1 2092 => Sexo=1 1667 conf: (0.8)
9. Sexo=1 1731 => Sobrevivió?=0 1364 conf: (0.79)
10. Sexo=1 1731 => Edad=1 Sobrevivió?=0 1329 conf: (0.77)

# AGENDA



1. Método Particional
2. Método Jerárquico
3. Método Probabilístico
4. Redes Neuronales
5. Reglas de Asociación: A priori
- 6. Votación para selección de factores**

# *Selección de Factores*

---

- Análisis de Componentes Principales
- Árboles de Decisión
- Análisis de Correlaciones
- Reglas de Asociación
- Regresión

# *EJERCICIOS*

# *Elecciones 2014*

- Se ha tomado una muestra de 500 posibles votantes con respecto a las elecciones presidenciales 2014 en Colombia primera vuelta, para ello se ha elegido al azar a las personas en el departamento de Nariño.
- Los atributos son:

Variable	Valores
Sexo	0 = Mujer; 1 = Hombre
Edad	0 = Entre 18 y 25; 1 = Entre 26 y 40; 2 = Mayor de 40
Etnia	0 = Mestizo; 1 = Indigena; 2 = Afro
Partido_politico	0 = De gobierno; 1 = Centro Democratico; 2 = Conservador; 3 = Izquierda; 4 = Sin Partido
Estrato	0; 1; 2; 3; 4; 5
Elección	0 = Santos; 1 = Zuluaga; 2 = Ramirez; 3 = Peñalosa; 4 = Lopez

● Cómo extraer información de esos datos?

# *Elecciones 2014*

@relation elecciones2014

@attribute Sexo {0,1}  
@attribute Edad {0,1,2}  
@attribute Etnia {0,1,2}  
@attribute Partido\_politico {0,1,2,3,4}  
@attribute Estrato {0,1,2,3,4,5}  
@attribute Eleccion {0,1,2,3,4}

@data

	<b>Variable</b>	<b>Valores</b>
0,0,0,0,0,0	Sexo	0 = Mujer; 1 = Hombre
1,0,1,1,0,1	Edad	0 = Entre 18 y 25; 1 = Entre 26 y 40; 2 = Mayor de 40
1,2,2,2,1,1	Etnia	0 = Mestizo; 1 = Indigena; 2 = Afro
1,0,0,0,1,1	Partido_politico	0 = De gobierno; 1 = Centro Democratico; 2 = Conservador; 3 = Izquierda; 4 = Sin Partido
1,1,2,2,1,2	Estrato	0; 1; 2; 3; 4; 5
0,0,2,0,1,0	Eleccion	0 = Santos; 1 = Zuluaga; 2 = Ramirez; 3 = Peñalosa; 4 = Lopez
1,0,2,0,1,0		
0,0,2,0,1,0		

# Elecciones 2014 -> Análisis de Clustering

Cluster centroids:

Attribute	Full Data (505)	Cluster#				
		0 (121)	1 (156)	2 (147)	3 (24)	4 (57)
<hr/>						
Sexo	1	1	1	0	1	1
Edad	1	0	1	1	0	2
Etnia	0	0	2	0	2	2
Partido_politico	1	1	2	4	0	4
Estrato	2	5	4	2	1	2
Eleccion	1	1	2	4	0	4

		Variable	Valores
Clustered Instance		Sexo	0 = Mujer; 1 = Hombre
0	121 ( 24%)	Edad	0 = Entre 18 y 25; 1 = Entre 26 y 40; 2 = Mayor de 40
1	156 ( 31%)	Etnia	0 = Mestizo; 1 = Indigena; 2 = Afro
2	147 ( 29%)	Partido_politico	0 = De gobierno; 1 = Centro Democratico; 2 = Conservador; 3 = Izquierda; 4 = Sin Partido
3	24 ( 5%)	Estrato	0; 1; 2; 3; 4; 5
4	57 ( 11%)	Elección	0 = Santos; 1 = Zuluaga; 2 = Ramirez; 3 = Peñalosa; 4 = Lopez

# *Elecciones 2014 -> Asociación*

Best rules found:

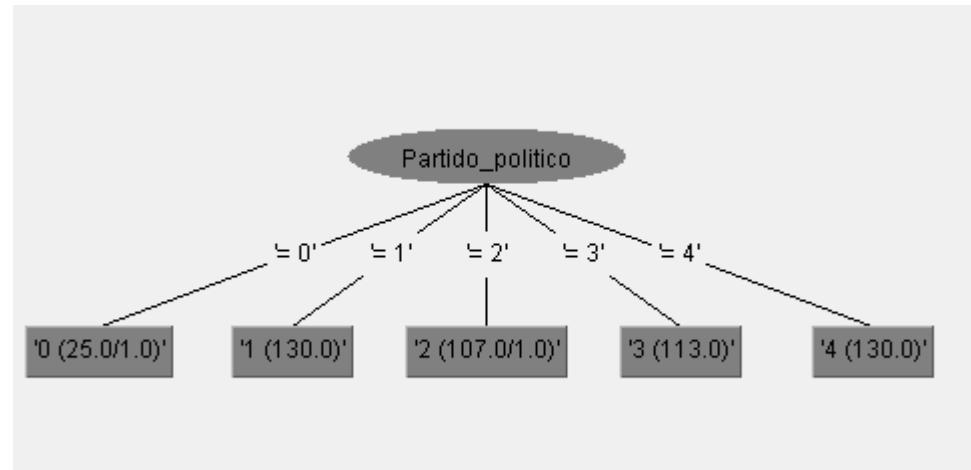
1. Partido\_politico=1 130 ==> Eleccion=1 130 conf: (1)
2. Partido\_politico=4 130 ==> Eleccion=4 130 conf: (1)
3. Partido\_politico=4 Estrato=2 130 ==> Eleccion=4 130 conf: (1)
4. Partido\_politico=3 113 ==> Eleccion=3 113 conf: (1)
5. Edad=1 Partido\_politico=3 113 ==> Eleccion=3 113 conf: (1)
6. Edad=2 Partido\_politico=4 111 ==> Eleccion=4 111 conf: (1)
7. Edad=2 Partido\_politico=4 Estrato=2 111 ==> Eleccion=4 111 conf: (1)
8. Sexo=1 Partido\_politico=1 108 ==> Eleccion=1 108 conf: (1)
9. Sexo=0 Partido\_politico=4 107 ==> Eleccion=4 107 conf: (1)
10. Sexo=0 Partido\_politico=4 Estrato=2 107 ==> Eleccion=4 107 conf: (1)

Variable	Valores
Sexo	0 = Mujer; 1 = Hombre
Edad	0 = Entre 18 y 25; 1 = Entre 26 y 40; 2 = Mayor de 40
Etnia	0 = Mestizo; 1 = Indigena; 2 = Afro
Partido_politico	0 = De gobierno; 1 = Centro Democratico; 2 = Conservador; 3 = Izquierda; 4 = Sin Partido
Estrato	0; 1; 2; 3; 4; 5
Elección	0 = Santos; 1 = Zuluaga; 2 = Ramirez; 3 = Peñalosa; 4 = Lopez

# Elecciones 2014 -> Predicción

==== Confusion Matrix ====

a	b	c	d	e	<- classified as
24	0	0	0	0	a = 0
1	130	1	0	0	b = 1
0	0	106	0	0	c = 2
0	0	0	113	0	d = 3
0	0	0	0	130	e = 4



Variable	Valores
Sexo	0 = Mujer; 1 = Hombre
Edad	0 = Entre 18 y 25; 1 = Entre 26 y 40; 2 = Mayor de 40
Etnia	0 = Mestizo; 1 = Indigena; 2 = Afro
Partido_politico	0 = De gobierno; 1 = Centro Democratico; 2 = Conservador; 3 = Izquierda; 4 = Sin Partido
Estrato	0; 1; 2; 3; 4; 5
Elección	0 = Santos; 1 = Zuluaga; 2 = Ramirez; 3 = Peñalosa; 4 = Lopez

# Zoo -> Análisis Descriptivo

```
@RELATION zoo
```

```
@ATTRIBUTE animal {aardvark,antelope,bass,bear,boar,buffalo,calf,carp,catfish,cavy,cheetah,chicken,chub,  
@ATTRIBUTE hair {false, true}  
@ATTRIBUTE feathers {false, true}  
@ATTRIBUTE eggs {false, true}  
@ATTRIBUTE milk {false, true}  
@ATTRIBUTE airborne {false, true}  
@ATTRIBUTE aquatic {false, true}  
@ATTRIBUTE predator {false, true}  
@ATTRIBUTE toothed {false, true}  
@ATTRIBUTE backbone {false, true}  
@ATTRIBUTE breathes {false, true}  
@ATTRIBUTE venomous {false, true}  
@ATTRIBUTE fins {false, true}  
@ATTRIBUTE legs INTEGER  
@ATTRIBUTE tail {false, true}  
@ATTRIBUTE domestic {false, true}  
@ATTRIBUTE catsize {false, true}
```