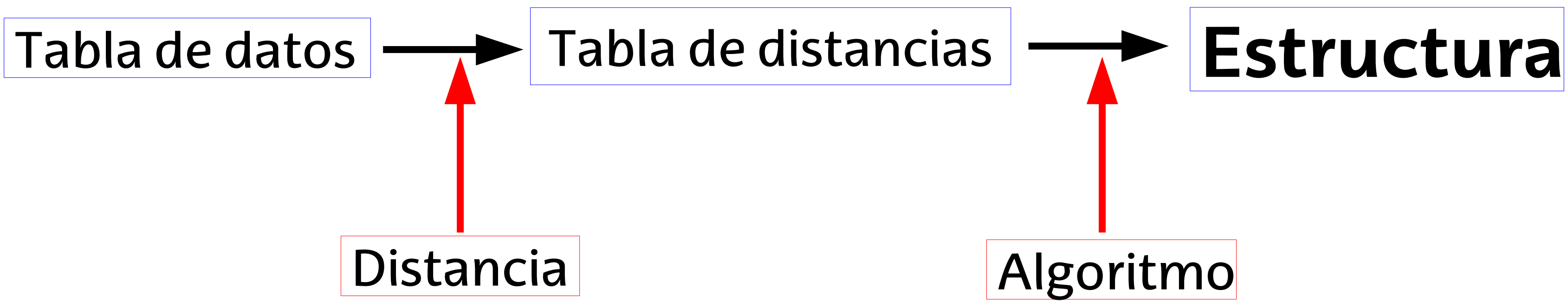
Clustering: Estructuras



Clustering: Estructuras

- Las estructuras de *cluster* más habituales que buscan los algoritmos de *clustering* son:

- **Partición**
- **Jerarquía**

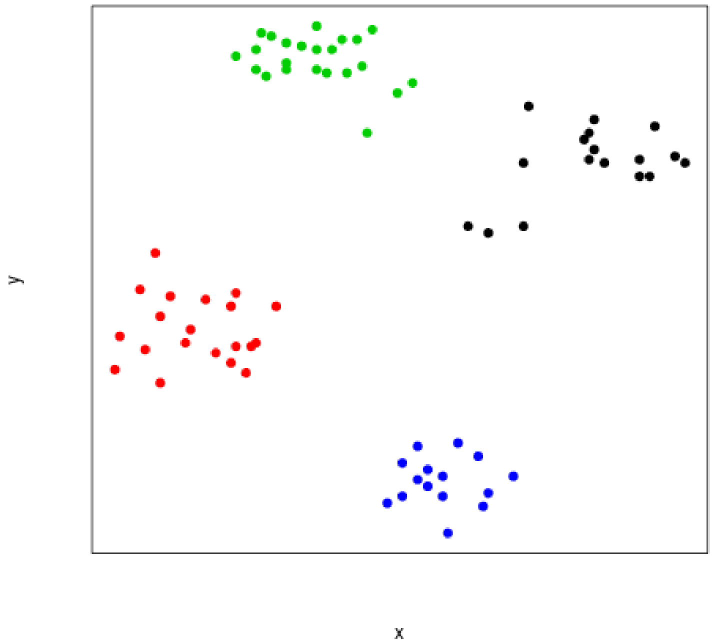
- Otras estructuras son:

- Recubrimiento
- Partición difusa
- Pirámide

Clustering: Estructuras

» La estructura de *partición* consiste en un conjunto de clases (*clusters*, subconjuntos) <u>no vacías</u> del conjunto de objetos tal que todo objeto <u>pertenece a únicamente una sola clase (*cluster*)</u>

- El conjunto vacío no es una clase.
- La unión de las clases es el conjunto total de objetos.
- La intersección de dos clases es vacía.



Clustering: Estructuras

» El resultado de realizar una *partición* es la definición de una **nueva variable**, **cualitativa**.

Ω	X_1		X_j		X_p	C
ω_1	x_{11}		x_{1j}		x_{1p}	c_1
ω_i	x_{i1}		x_{ij}		x_{ip}	c_i
ω_n	x_{n1}		x_{nj}		x_{np}	c_n

Clustering: Estructuras

- Se trata de buscar una buena *partición* (una buena *variable cualitativa*).
  - Una buena partición es un conjunto de clases (*clusters*) de objetos que son similares entre sí (distancias pequeñas dentro de cada clase), y diferentes de los objetos de otras clases (*clusters*).

$$\Sigma_i \Sigma_{i'} d(\omega_i, \omega_{i'}) = \Sigma_k \Sigma_i \Sigma_{i'} d_j(\omega_i^k, \omega_{i'}^k) + \Sigma_k \Sigma_{k' \neq k} \Sigma_i \Sigma_{i'} d_j(\omega_i^k, \omega_{i'}^{k'})$$

  - Una buena *variable cualitativa* es aquélla que está *bien correlacionada* con las variables originales. Siendo r_j la correlación entre X_j y la nueva variable, un criterio es buscar aquella variable que maximice $\Sigma_j r_j$.

Clustering: Estructuras

- Si las variables X_j son cuantitativas, r_j puede ser la *razón de correlación* entre la variable cuantitativa X_j y la nueva variable, cualitativa, C:

$$R_j = 1 - \left(\Sigma_k \, n_k Var_k(X_j) \,/\, nVar(X_j) \right)$$

con $Var_k(X_j)$, la varianza de X_j en la clase c_k de C.

- Como $nVar(X_j) = \Sigma_i \Sigma_{i'} \, d_j(\omega_i, \omega_{i'}) \,/\, 2n$, siendo d_j la *distancia euclidiana cuadrática* en la variable X_j , y

$$\Sigma_i \Sigma_{i'} d(\omega_i, \omega_{i'}) = \Sigma_i \Sigma_{i'} \Sigma_j d_j(\omega_i, \omega_{i'})$$

se tiene que:

$$\Sigma_j r_j = 1 - \left(\Sigma_k \Sigma_i \Sigma_{i'} d(\omega_i^k, \omega_{i'}^k) / \Sigma_i \Sigma_{i'} d(\omega_i, \omega_{i'}) \right)$$

Clustering: Estructuras

- **Resumen**: Minimizar la *suma de distancias euclidianas cuadráticas* dentro de las clases es **equivalente** a maximizar la suma de correlaciones de las variables cuantitativas con la variable cualitativa asociada a la partición (tomando las variables *independiente*mente); y viceversa.
- Cuando se usa otra distancia se suele plantear el problema de minimizar la *suma de distancias* dentro de las clases:

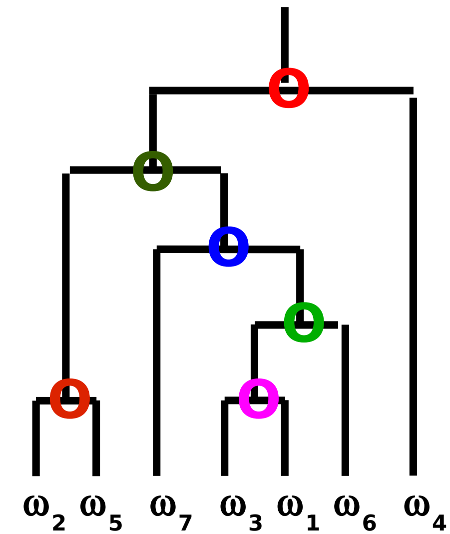
$$min\ \Sigma_k \Sigma_i \Sigma_{i'} d(\omega_i^k, \omega_{i'}^k)$$

para realizar una buena estructura de partición, pero en estos casos no se hace ninguna referencia a la relación entre las variables.

Clustering: Estructuras

» La estructura de *jerarquía* (*árbol* en teoría de grafos) consiste en un conjunto de clases (*clusters*, subconjuntos) <u>no vacías</u> del conjunto de objetos al que pertenecen:

- el conjunto o clase total (raíz)
- las clases singulares (hojas)
- y tal que si de dos clases son de la jerarquía, su intersección es o bien vacía u bien una de ellas (una contenida en la otra).

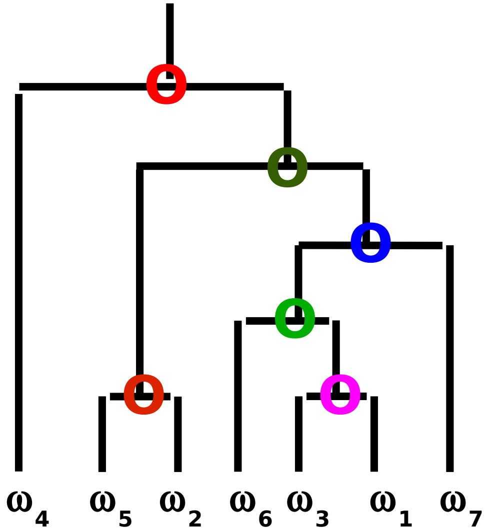


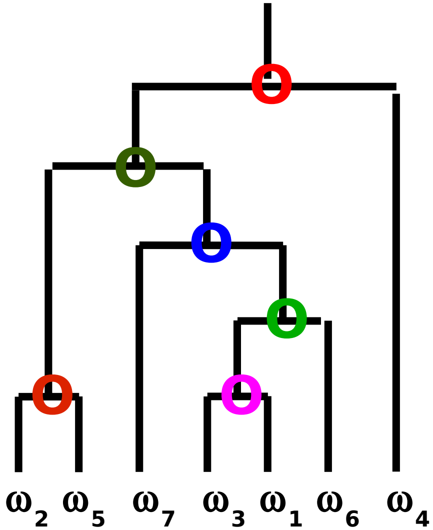
- $\{\omega_2\},\{\omega_5\},\{\omega_7\},\{\omega_3\},\{\omega_1\},\{\omega_6\},\{\omega_4\}$,
- $\{\omega_1,\omega_2,\omega_3,\omega_4,\omega_5,\omega_6,\omega_7\}$,
- $\{\omega_2,\omega_5\},\{\omega_1,\omega_3\},\{\omega_1,\omega_3,\omega_6\}$, $\{\omega_1,\omega_3,\omega_6,\omega_7\},\{\omega_1,\omega_2,\omega_3,\omega_5,\omega_6,\omega_7\}$

Clustering: Estructuras

›› El orden de aparición de los objetos (hojas) en el árbol no tiene ningún significado.

- Ambos árboles representan la misma jerarquía



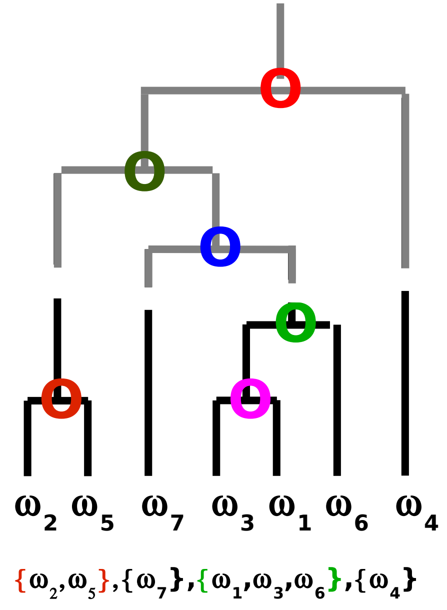


- $\{\omega_2\},\{\omega_5\},\{\omega_7\},\{\omega_3\},\{\omega_1\},\{\omega_6\},\{\omega_4\},$
- $\{\omega_1,\omega_2,\omega_3,\omega_4,\omega_5,\omega_6,\omega_7\},$
- $\{\omega_2,\omega_5\},\{\omega_1,\omega_3\},\{\omega_1,\omega_3,\omega_6\},$
  $\{\omega_1,\omega_3,\omega_6,\omega_7\},\{\omega_1,\omega_2,\omega_3,\omega_5,\omega_6,\omega_7\}$

Clustering: Estructuras

›› La <u>poda</u> de ramas del árbol (*jerarquía*) determina una *partición* del conjunto de objetos

- La poda de las hojas determina la partición más fina (la de las clases singulares $\{\omega_i\}$).
- La poda desde la raíz determina la partición menos fina (la del conjunto entero Ω).
- Ambas podas son triviales y no aportan información. Interesan otras podas.



Universidad del País Vasco Euskal Herriko Unibertsitatea

KISA

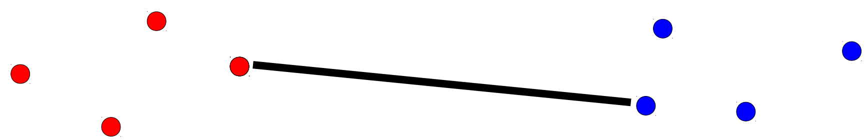
Clustering: Estructuras

- Las longitudes de las ramas del árbol se usan para representar gráficamente las distancias entre los objetos, y las distancias entre clases de objetos (*clusters*): *dendrograma* (jerarquía valorada o indexada)
- La *altura* a la que se fusionan dos clases de objetos representa la *distancia entre ambas clases*.
- La distancia entre clases debe estar basada en la distancia entre objetos. La distancia entre dos clases singulares debe coincidir con la distancia entre los dos objetos correspondientes.

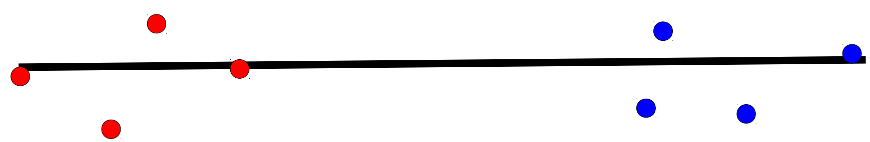
Clustering: Estructuras

›› <u>Distancias entre clases</u> de objetos (*clusters*):

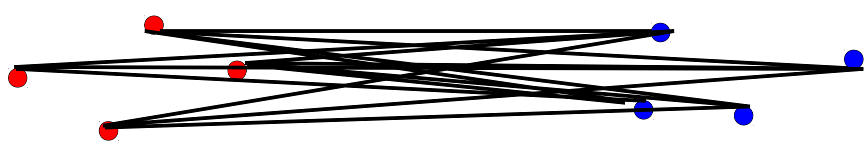
- Mínima



- Máxima



- Media



Clustering: Estructuras

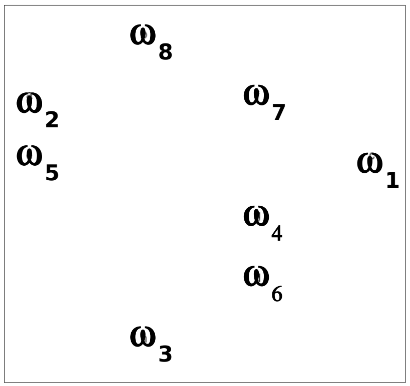
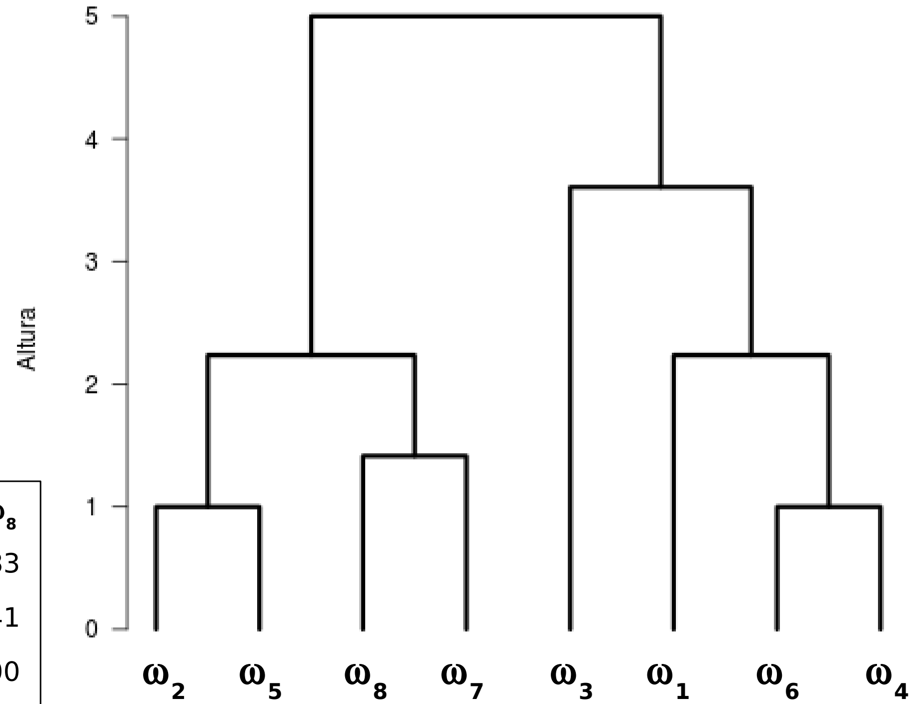


Tabla de distancias

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
ω_1	0.00	3.16	3.61	1.41	3.00	2.24	1.41	2.83
ω_2	3.16	0.00	4.12	2.83	1.00	3.61	2.00	1.41
ω_3	3.61	4.12	0.00	2.24	3.16	1.41	4.12	5.00
ω_4	1.41	2.83	2.24	0.00	2.24	1.00	2.00	3.16
ω_5	3.00	1.00	3.16	2.24	0.00	2.83	2.24	2.24
ω_6	2.24	3.61	1.41	1.00	2.83	0.00	3.00	4.12
ω_7	1.41	2.00	4.12	2.00	2.24	3.00	0.00	1.41
ω_8	2.83	1.41	5.00	3.16	2.24	4.12	1.41	0.00



Distancia entre clases de objetos con el el criterio del máximo.

Universidad del País Vasco Euskal Herriko Unibertsitatea

KISA

Clustering: Estructuras

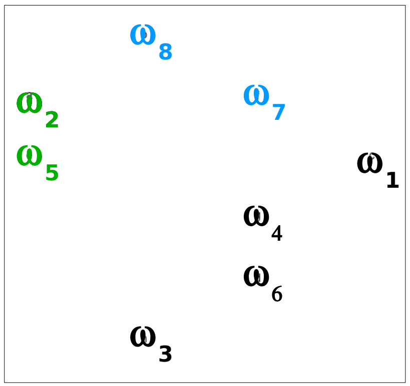
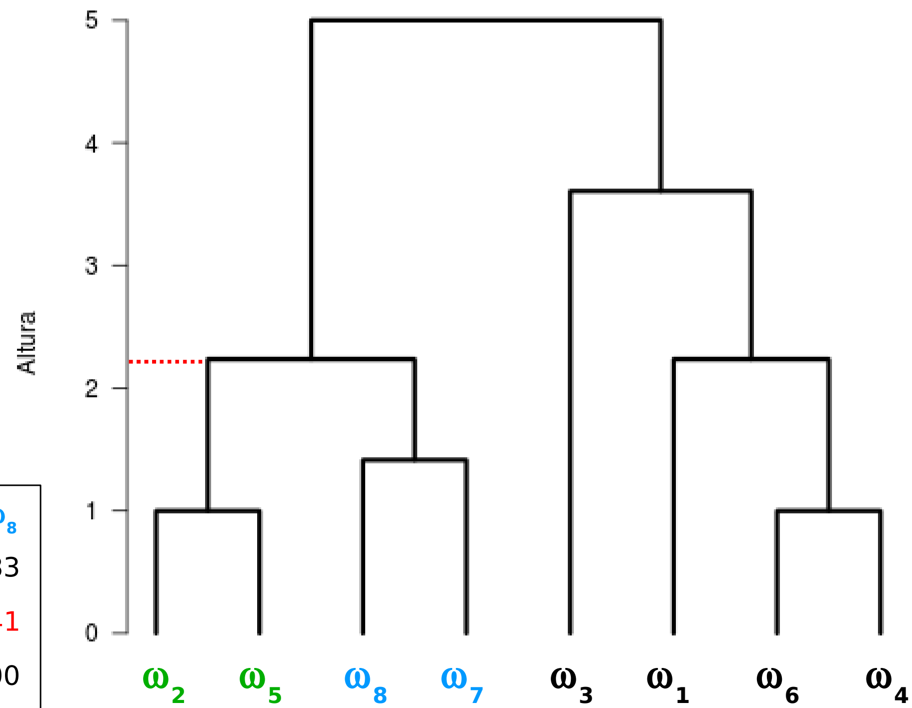


Tabla de distancias

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
ω_1	0.00	3.16	3.61	1.41	3.00	2.24	1.41	2.83
ω_2	3.16	0.00	4.12	2.83	1.00	3.61	2.00	1.41
ω_3	3.61	4.12	0.00	2.24	3.16	1.41	4.12	5.00
ω_4	1.41	2.83	2.24	0.00	2.24	1.00	2.00	3.16
ω_5	3.00	1.00	3.16	2.24	0.00	2.83	2.24	2.24
ω_6	2.24	3.61	1.41	1.00	2.83	0.00	3.00	4.12
ω_7	1.41	2.00	4.12	2.00	2.24	3.00	0.00	1.41
ω_8	2.83	1.41	5.00	3.16	2.24	4.12	1.41	0.00



$$d(\{\omega_2,\omega_5\},\{\omega_8,\omega_7\})$$

$$= \max(d(\omega_2,\omega_8),d(\omega_2,\omega_7),d(\omega_5,\omega_8),d(\omega_5,\omega_7))$$

$$= 2.24$$

Universidad del País Vasco Euskal Herriko Unibertsitatea

KISA

Clustering: Estructuras

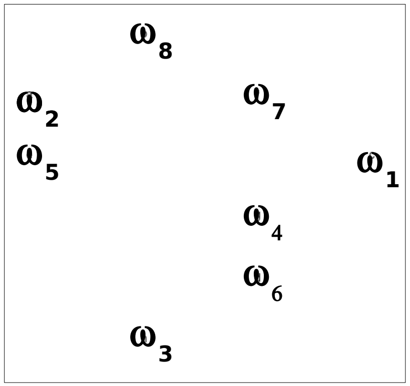
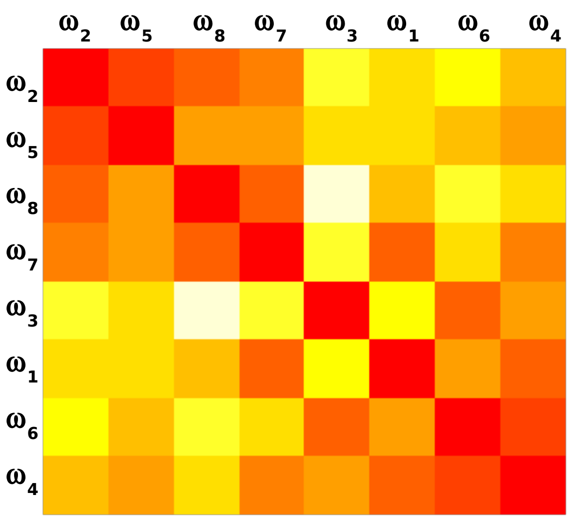
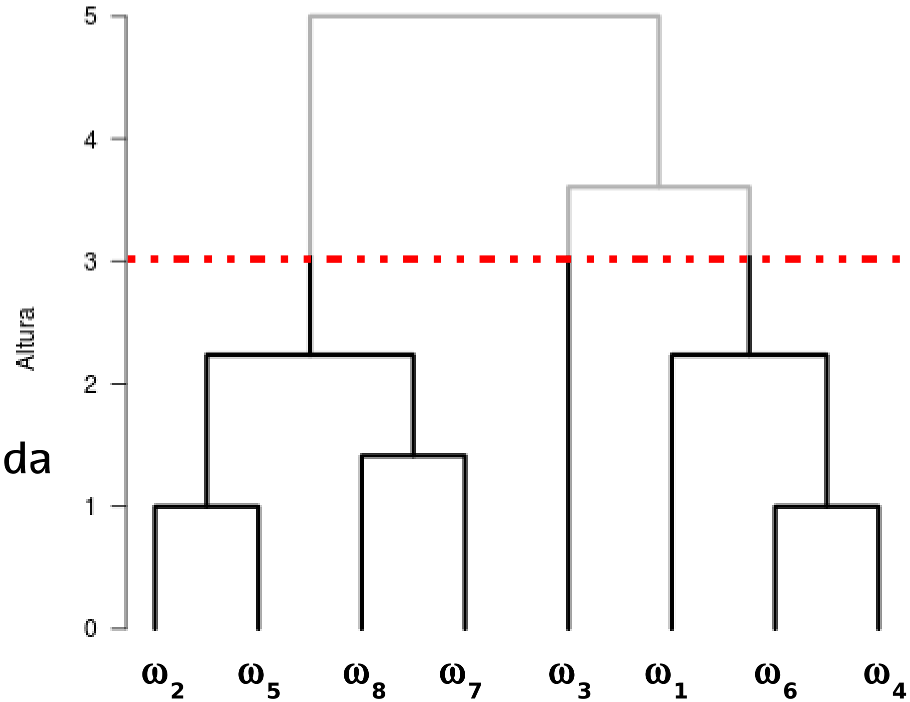


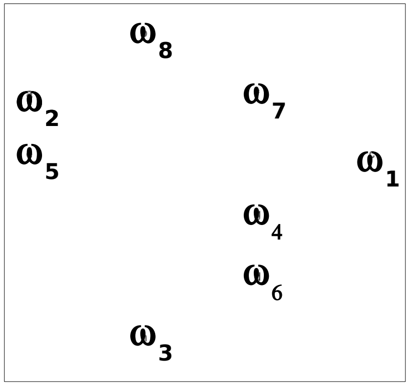
Tabla de distancias reordenada

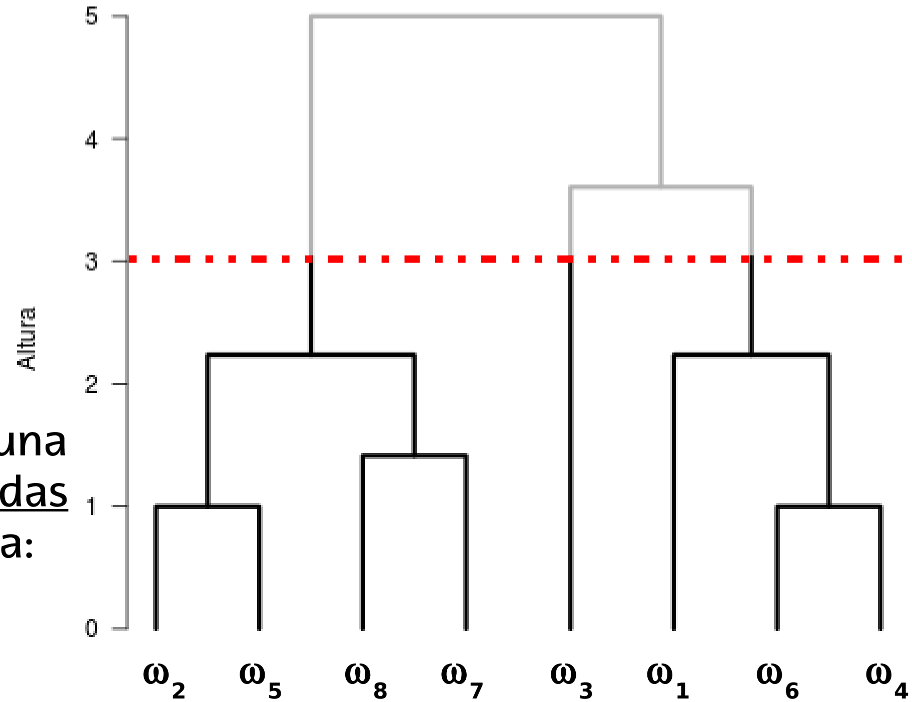




- Poda del árbol de acuerdo a una altura
- Partición: $\{\omega_2,\omega_5,\omega_7,\omega_8\}$, $\{\omega_3\}$, $\{\omega_1,\omega_6,\omega_4\}$
- La distancia entre las clases es mayor que 3

Clustering: Estructuras





- Las podas por alturas determinan una sucesión de <u>particiones encajadas</u> desde la más fina hasta la menos fina:
  - $\{\omega_2,\omega_5,\omega_7,\omega_8,\omega_3,\omega_1,\omega_6,\omega_4\}$
  - $\{\omega_2,\omega_5,\omega_7,\omega_8\},\{\omega_3,\omega_1,\omega_6,\omega_4\}$
  - $\{\omega_2,\omega_5,\omega_7,\omega_8\},\{\omega_3\},\{\omega_1,\omega_6,\omega_4\}$
  - $\{\omega_2,\omega_5\},\{\omega_7,\omega_8\},\{\omega_3\},\{\omega_1\},\{\omega_6,\omega_4\}$
  - $\{\omega_2,\omega_5\},\{\omega_7\},\{\omega_8\},\{\omega_3\},\{\omega_1\},\{\omega_6,\omega_4\}$
  - $\{\omega_2\},\{\omega_5\},\{\omega_7\},\{\omega_8\},\{\omega_3\},\{\omega_1\},\{\omega_6\},\{\omega_4\}$