# Clustering jerárquico

Christopher Expósito Izquierdo
Airam Expósito Márquez
Israel López Plata
Belén Melián Batista
J. Marcos Moreno Vega

{**cexposit, aexposim, ilopezpl, mbmelian, jmmoreno**}**@ull.edu.es**

Departamento de Ingeniería Informática y de Sistemas
Universidad de La Laguna

# Índice

# CLUSTERING JERÁRQUICO

# Clustering jerárquico

## Observaciones

- Una de las mayores dificultades al agrupar elementos es encontrar el número apropiado de clusters.
- Los métodos jerárquicos construyen una estructura en la que los elementos se agrupan en subconjuntos cada vez mayores hasta que todos pertenecen al mismo conjunto.
- De esta forma, no se muestra un agrupamiento sino las relaciones de proximidad que existen entre los elementos.

## Procedimientos básicos

- **Aglomerativos.** Inicialmente se forman clusters individuales, cada uno de los cuales contiene a un único elemento. En cada iteración se unen los dos clusters más próximos. El procedimiento finaliza cuando solo haya un cluster.

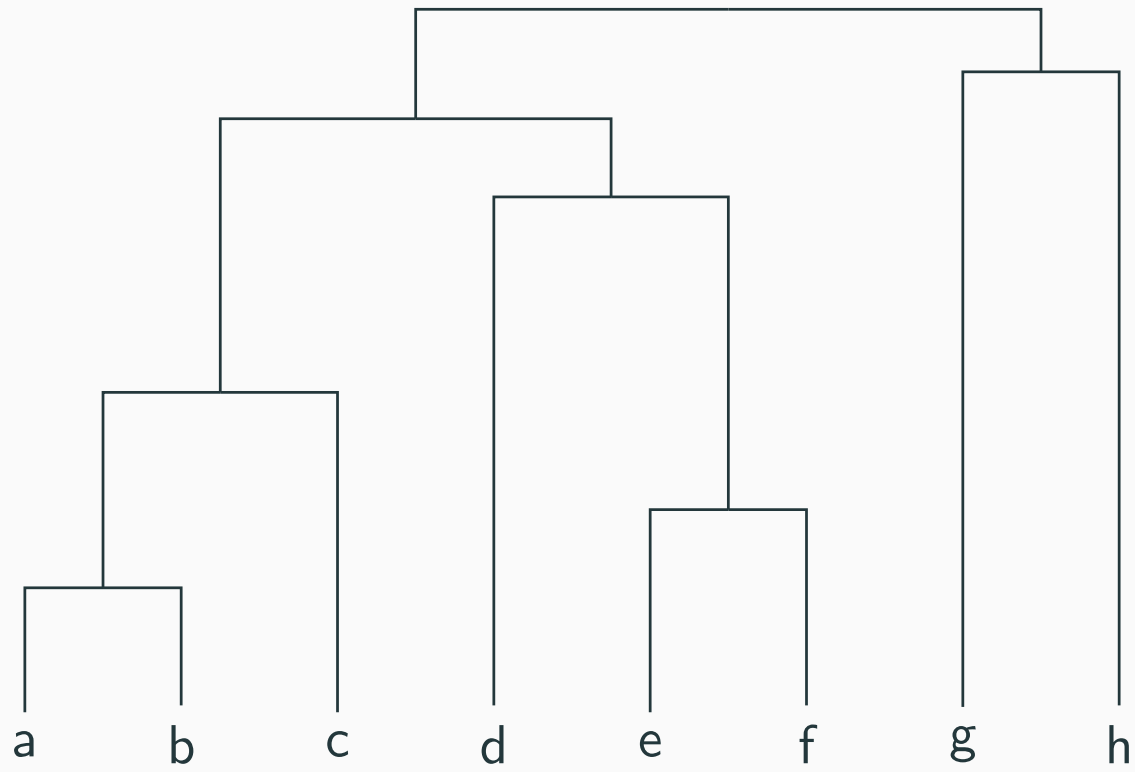
- **Divisivos.** Se parte de un único cluster al que pertenecen todos los elementos. En cada iteración se escoge un cluster y se divide.

Debe decidirse qué cluster se selecciona para dividir y cómo se divide. El procedimiento finaliza cuando hayan tanto clusters como elementos.

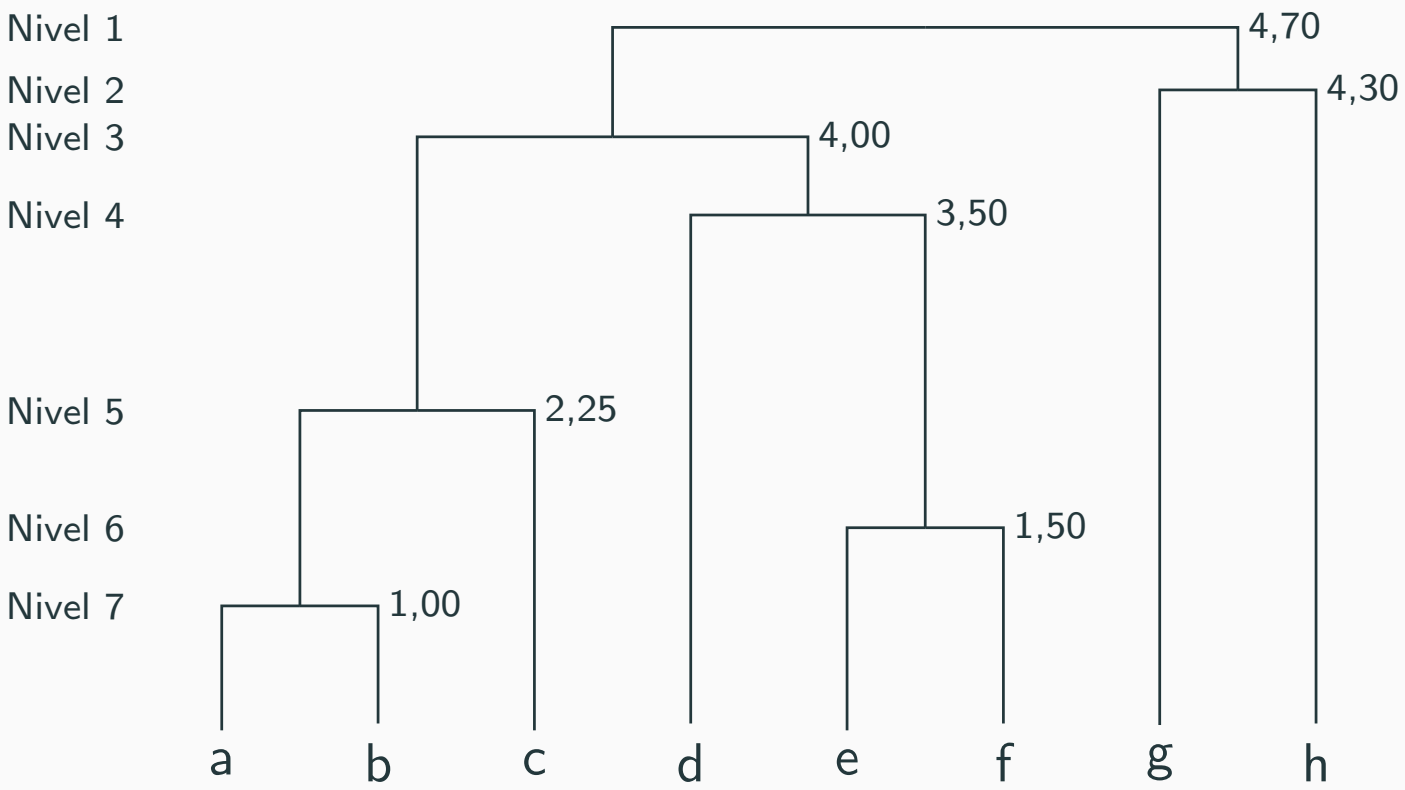
## Dendrograma (i)

- El clustering jerárquico suele representarse a través de un dendograma, que muestra en qué orden se han unido los cluster y cuál es el grado de proximidad que tienen los clusters que se unen.
- Los nodos hojas del dendograma se corresponden con los elementos individuales.
- En el nodo raíz se representa el cluster al que pertenecen todos los elementos.
- El resto de nodos se corresponde con los clusters que se van formando.

## Dendrograma (ii)



## Dendrograma. Cómo se construye

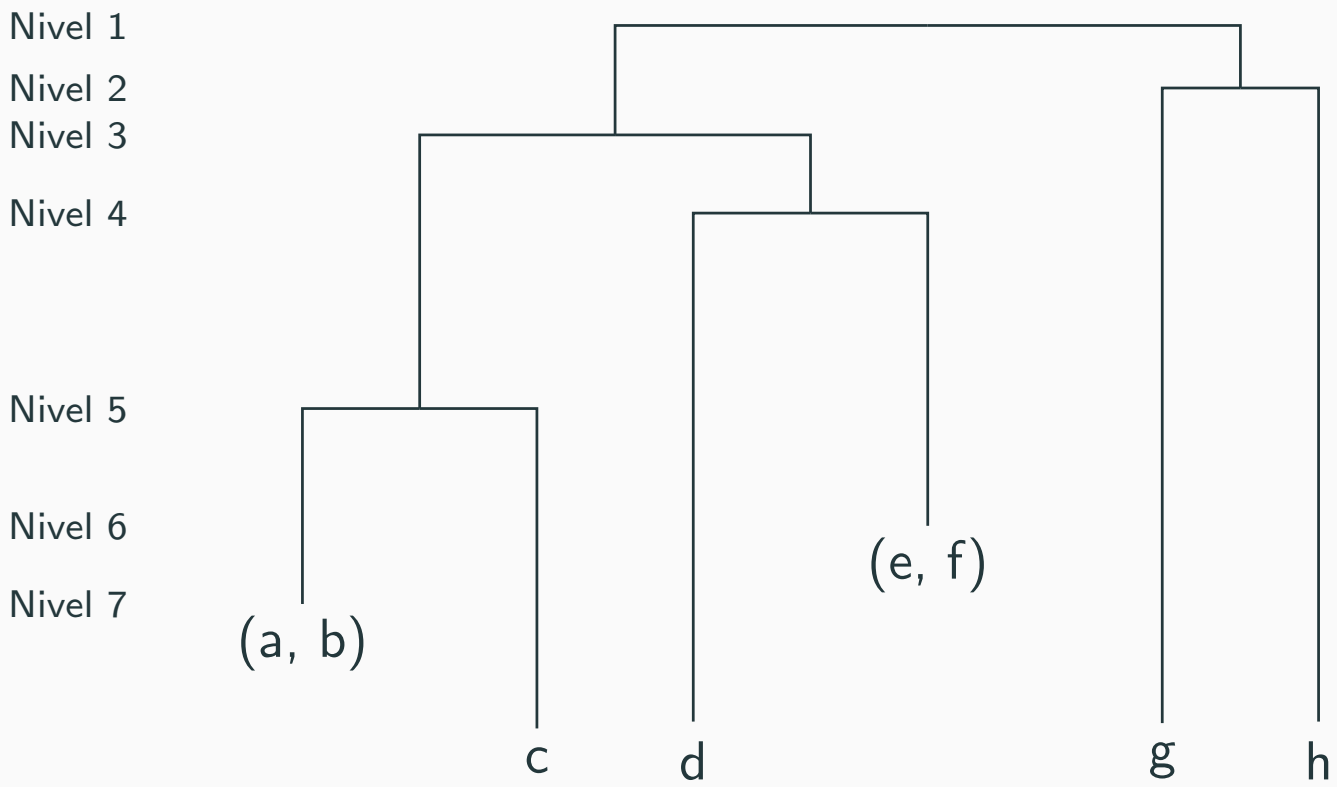


## Dendrograma. Cómo se genera un agrupamiento (i)

- El dendrograma puede usarse para generar diferentes agrupamientos.
- Para ello, se selecciona un nivel y se poda el dendrograma descartando los hijos de los nodos con nivel igual o superior al nivel seleccionado. Los nodos hojas del árbol resultante dan el agrupamiento buscado.
- Dependiendo del nivel seleccionado se obtienen agrupamientos con clusters más o menos compactos.

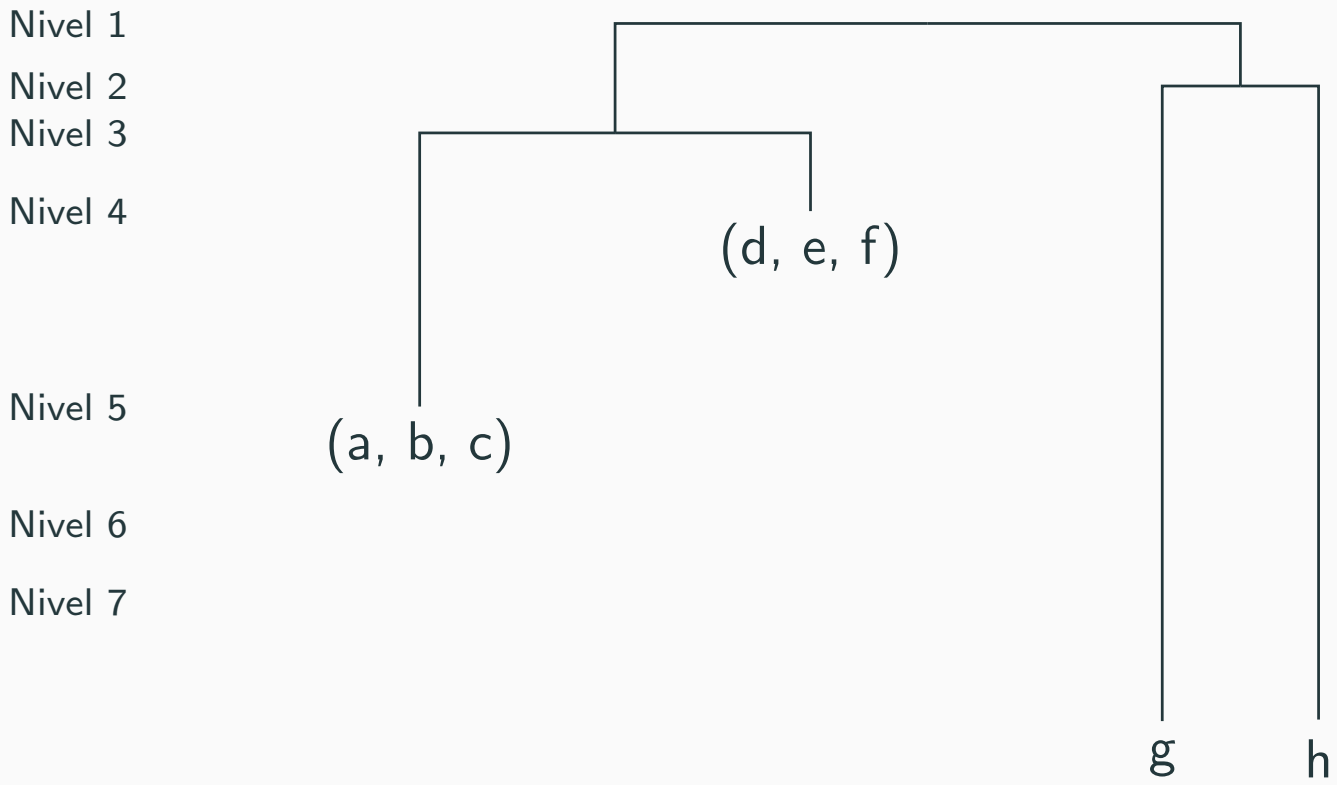
## Dendrograma. Cómo se genera un agrupamiento (ii)

- Nivel seleccionado = 6



## Dendrograma. Cómo se genera un agrupamiento (iii)

- Nivel seleccionado = 4



# Clustering jerárquico

**Algorithm 1:** Algoritmo aglomerativo básico

`Matriz de proximidad`

Obtener la proximidad entre cada par de elementos;

**repeat**

    Unir los dos clusters más próximos;

    Actualizar la matriz de proximidad;

**until** *Solo hay un cluster*;

## Medidas de proximidad entre clusters (i)

- **Enlace simple (single link).** La proximidad entre dos clusters se define como la proximidad entre los dos elementos más próximos que pertenecen a clusters diferentes.
- **Enlace completo (complete link).** La proximidad entre dos clusters se define como la proximidad entre los dos elementos menos próximos que pertenecen a clusters diferentes.
- **Promedio del grupo (group average).** La proximidad entre dos clusters se define como la proximidad promedio entre todos los pares de elementos que pertenecen a clusters diferentes.

## Medidas de proximidad entre clusters (ii)

- **Proximidad basada en centroides.** La proximidad entre dos clusters se define como la proximidad entre los centroides de cada cluster.
- Tras unir dos clusters debe obtenerse el centroide del nuevo cluster. Algunas alternativas para ello son:
  - Obtener el centroide como el punto central del nuevo cluster.
  - Obtener el centroide como la suma ponderada de los centroides de los clusters que se unen usando la siguiente expresión:

$$c = \left( \frac{N_1 \cdot x_1 + N_2 \cdot y_1}{N_1 + N_2}, \frac{N_1 \cdot x_2 + N_2 \cdot y_2}{N_1 + N_2}, \ldots, \frac{N_1 \cdot x_n + N_2 \cdot y_n}{N_1 + N_2} \right)$$

  con $(x_i, x_2, \ldots, x_n)$ e $(y_1, y_2, \ldots, y_n)$ los centroides de los cluster y $N_1$ e $N_2$ el número de elementos en cada uno de ellos.

## Método de Ward

- **Método de Ward.** La medida de proximidad usada es la suma de errores al cuadrado.

De esta manera, en cada etapa se unen los dos clusters que dan lugar al cluster con menor suma de errores al cuadrado.

# BIBLIOGRAFÍA

# Bibliografía

Tan, P.-N., Steinbach, M., and Kumar, V.
**Introduction to Data Mining.**
Addison-Wesley, 2006.

# Licencia

**Esta obra está bajo una licencia de Creative Commons.**
**Reconocimiento - No comercial - Compartir igual**