# Clustering

## Hierarchical algorithms

Christian Oliva Moya

Dpto. de Ingeniería Informática, Escuela Politécnica Superior

Universidad Autónoma de Madrid

28049 Madrid, Spain

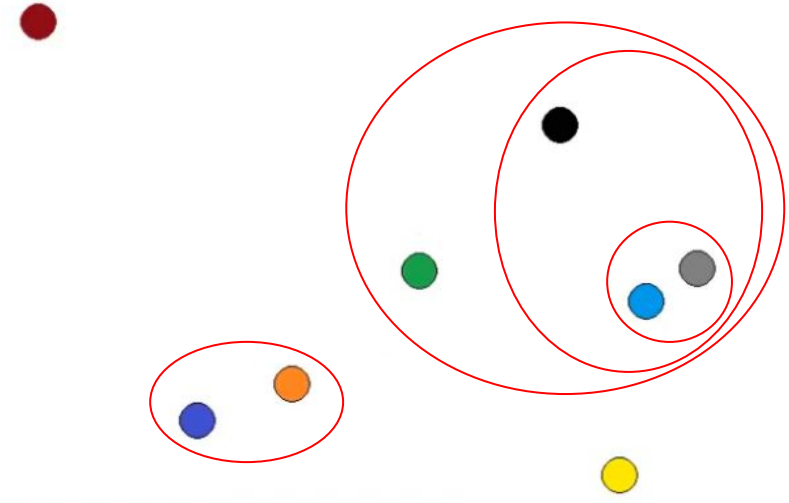
v1 - 15/02/2024

# Introducción (1)

- Hierarchical clustering (Connectivity clustering)

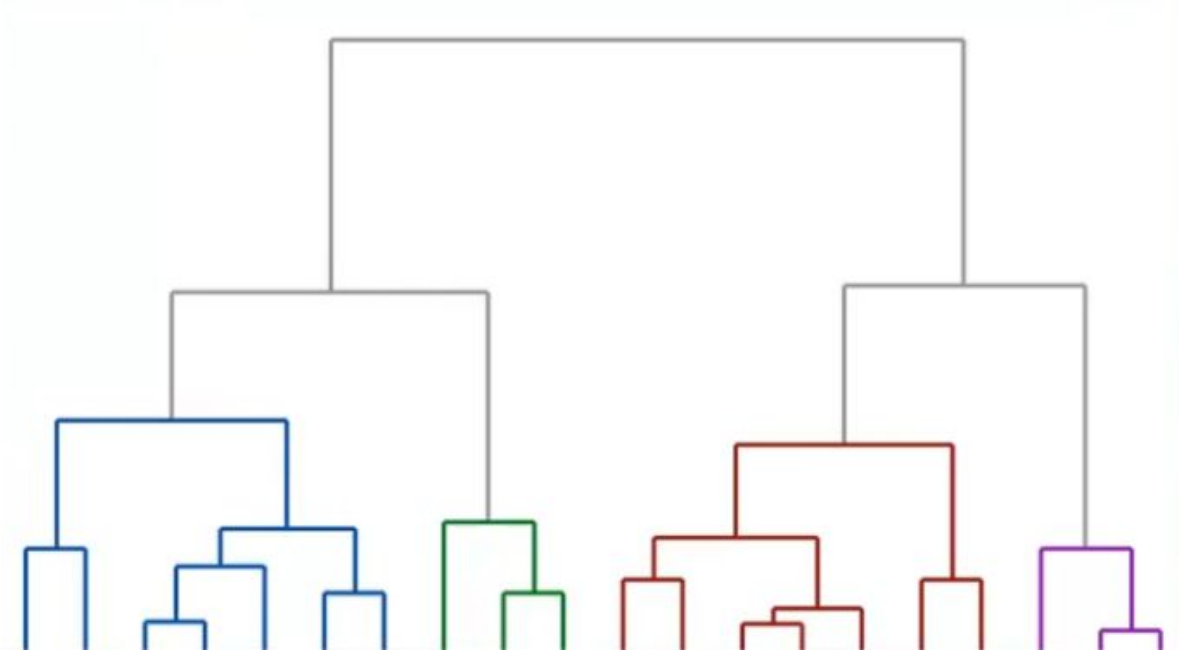
  Los datos se van conectando (agrupando) de forma escalonada

  Ejemplo:



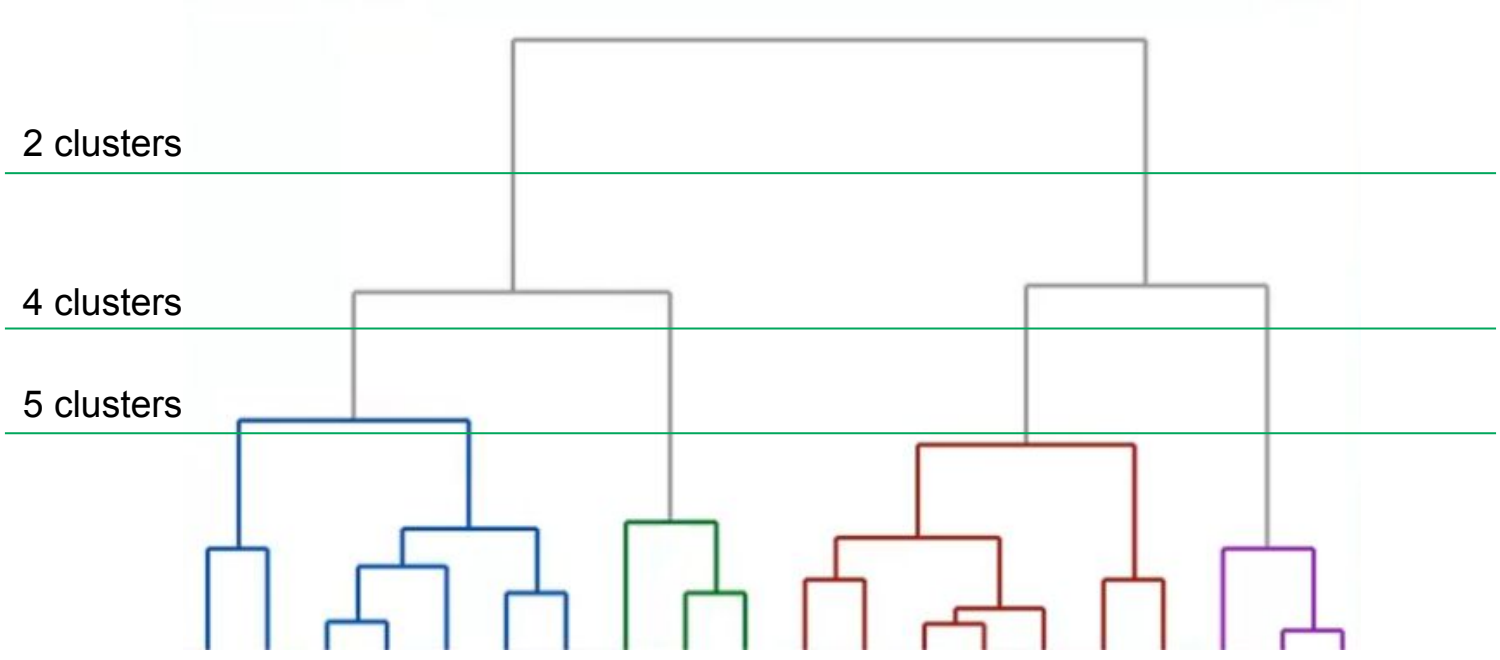
# Introducción (2)

El resultado de los algoritmos jerárquicos es un dendrograma



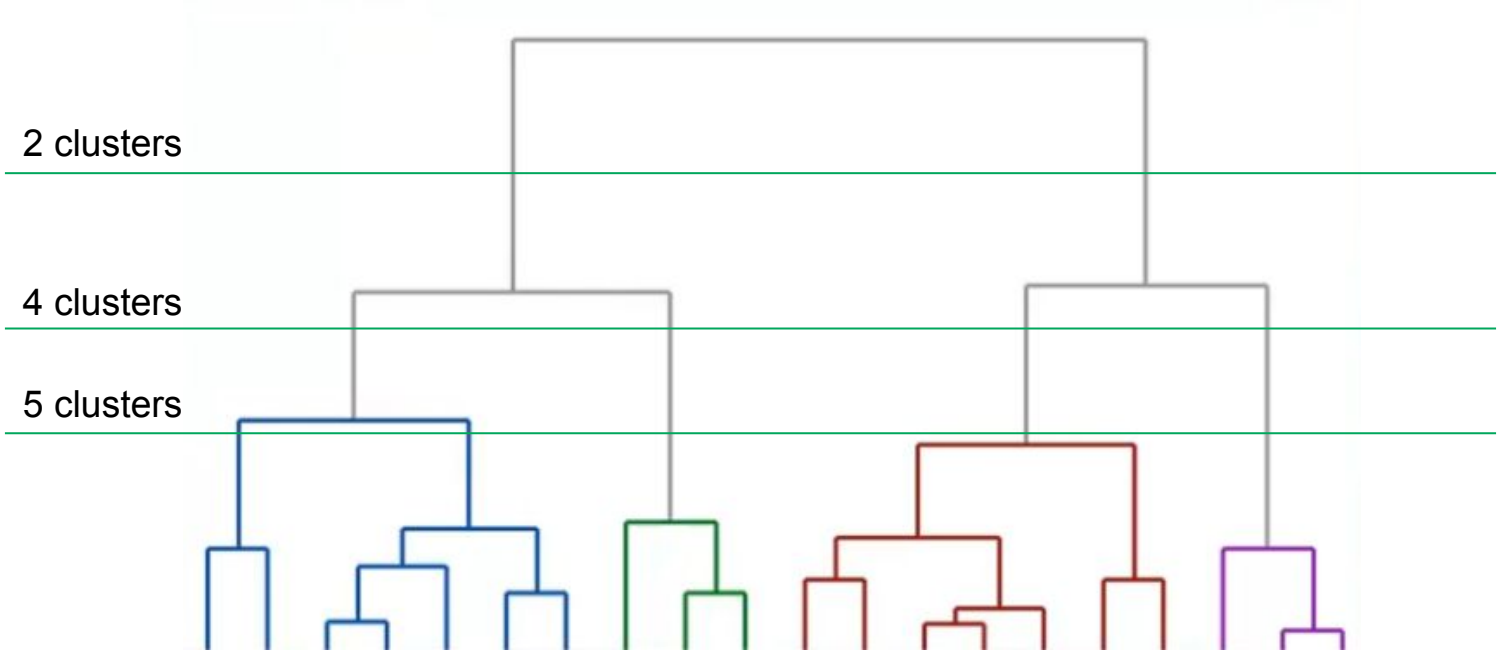
# Introducción (3)

Se obtiene el clustering cortando el dendrograma al nivel deseado



# Introducción (4)

El algoritmo de clustering se basa en una métrica de similitud



# Introducción (6)

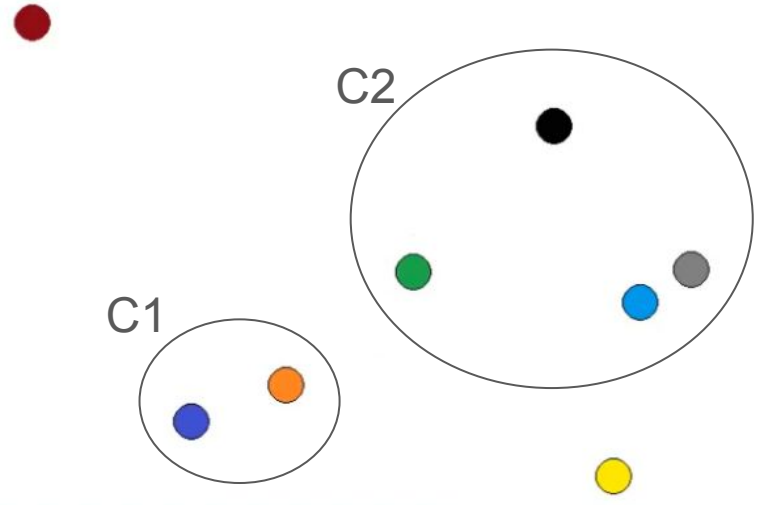
- Ventajas
  - Particiones múltiples: no es necesario especificar el número de clusters, el usuario puede elegir el número de particiones partiendo el dendrograma.
  - El dendrograma da mucha información para interpretar los datos.
- Desventajas
  - Imposibilidad de backtracking: no se pueden deshacer los pasos anteriores. Si dos puntos son agrupados y vemos que la conexión no es buena, no se puede deshacer la conexión.
  - Baja escalabilidad: la complejidad de los algoritmos jerárquicos es al menos $O(N^2)$

# Clustering aglomerativo (1)

- Inicialmente, cada punto forma un singleton cluster
- El objetivo es ir conectando clusters cuya distancia sea mínima
- Estrategias de enlazado (Linkage):
  - Single-link: distancia mínima entre puntos
  - Average-link: distancia media entre puntos
  - Complete-link: distancia máxima entre puntos
  - Centroid-link: distancia entre los centroides de los clusters
  - Otros

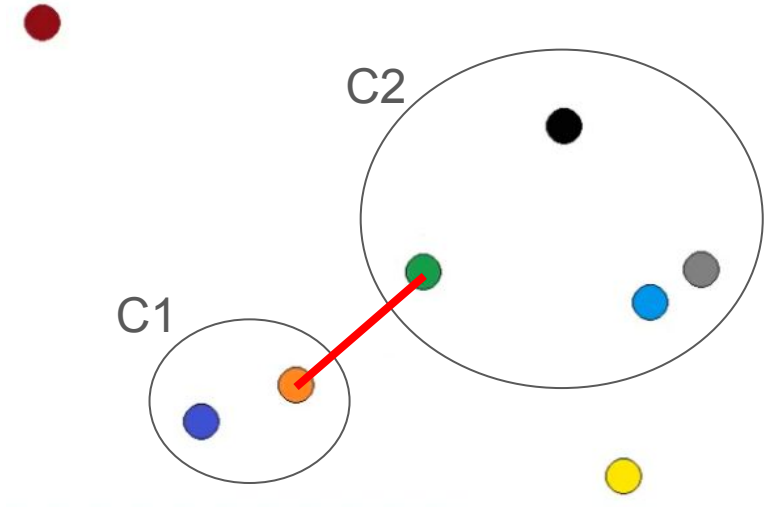
# Clustering aglomerativo (2) - Ejemplo de Linkage

- ¿Cómo defines la distancia entre los clusters C1 y C2?



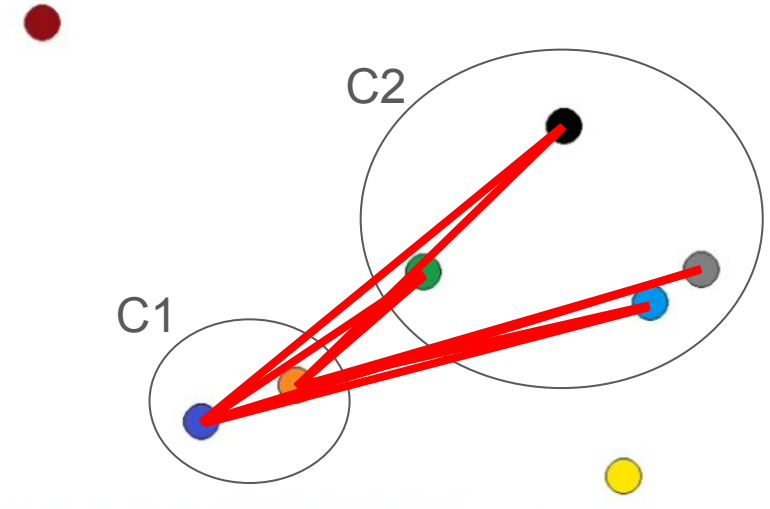
# Clustering aglomerativo (3) - Ejemplo de Linkage

- Single Linkage: distancia mínima: dist(C1, C2).min()



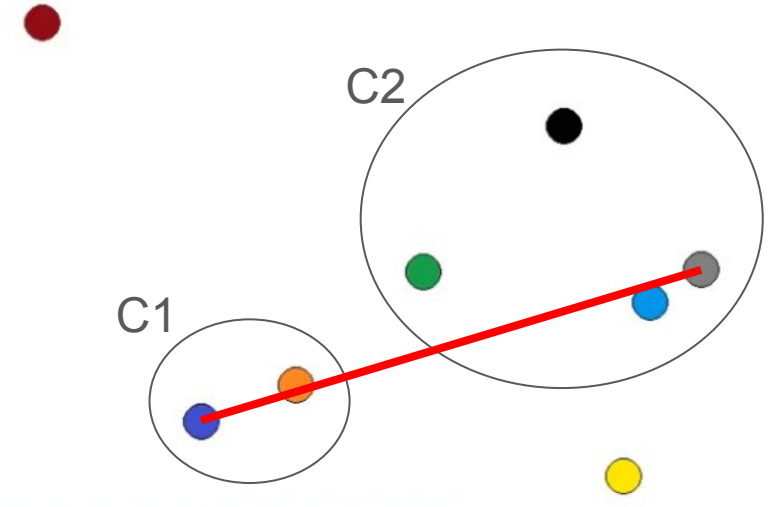
# Clustering aglomerativo (4) - Ejemplo de Linkage

- Average Linkage: distancia promedio: dist(C1, C2).mean()



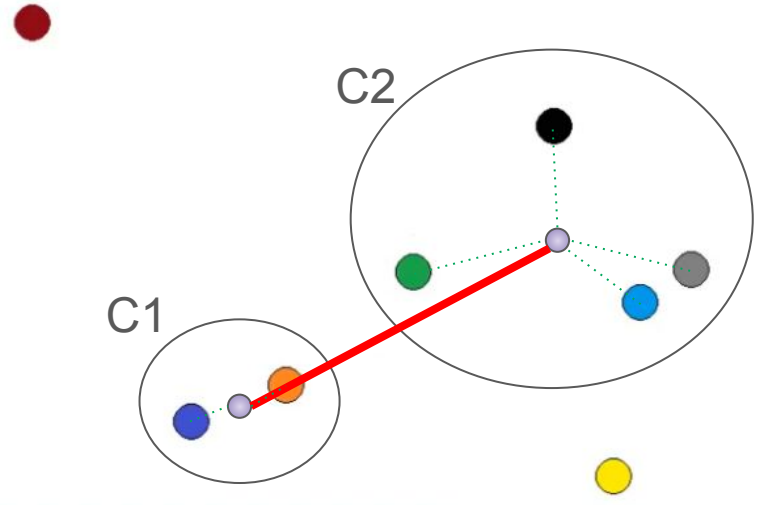
# Clustering aglomerativo (5) - Ejemplo de Linkage

- Complete Linkage: distancia máxima: dist(C1, C2).max()



# Clustering aglomerativo (6) - Ejemplo de Linkage

- Centroid Linkage: distancia entre centroides: dist(C1.mean(), C2.mean())



# Clustering aglomerativo (7)

- Implementación de clustering aglomerativo con centroides:

  Notebook 02_ejemplo_clustering_aglomerativo.ipynb

  https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html

# Práctica

- La práctica la podemos ir empezando ya. La tenéis en el siguiente notebook:

  practica_clustering.ipynb