Clustering

Introducción

Christian Oliva Moya

Dpto. de Ingeniería Informática, Escuela Politécnica Superior

Universidad Autónoma de Madrid

28049 Madrid, Spain

v1 - 15/02/2024

Grado en **Ingeniería Informática** UAM

Máster en **Inteligencia Computacional y Sistemas Interactivos** UAM

Doctorando en **Ingeniería Informática y Telecomunicaciones** UAM

Profesor Ayudante Dpto. Ingeniería Informática

Grupo de Neurocomputación Biológica

christian.oliva@uam.es

Investigación

Interpretabilidad-Explicabilidad de Redes Neuronales Profundas (XAI)

Implementación de mecanismos experimentales

Aplicación de Inteligencia Artificial y XAI en Mercados Financieros

Contenido del bloque

1. Introducción
    a. Métricas de distancia
    b. Taxonomía de los algoritmos de clustering
    c. Objetivos
2. Clustering Jerárquico (Aglomerativo)
3. Clustering Particional (Basado en centroides)
4. Clustering EM (Mezcla de Gaussianas)
5. Clustering basado en Densidades (DBSCAN)

Introducción (1)

¿Qué es el clustering?

Introducción (1)

¿Qué es el clustering?

¿Qué diferencias hay entre aprendizaje supervisado y no supervisado?

Introducción (1)

¿Qué es el clustering?

¿Qué diferencias hay entre aprendizaje supervisado y no supervisado?

¿Cuál es el objetivo de los algoritmos de clustering?

Introducción (1)

¿Qué es el clustering?

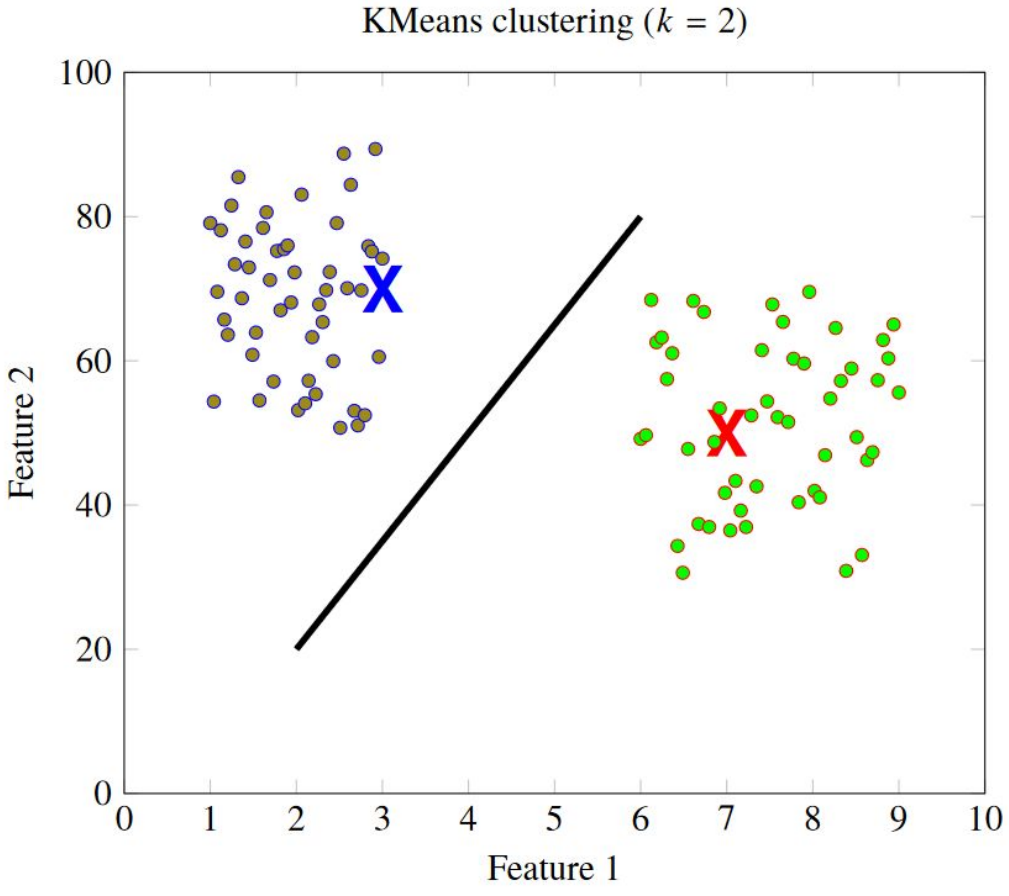
¿Qué diferencias hay entre aprendizaje supervisado y no supervisado?

¿Cuál es el objetivo de los algoritmos de clustering?

¿Qué significa proximidad?

Introducción (2)

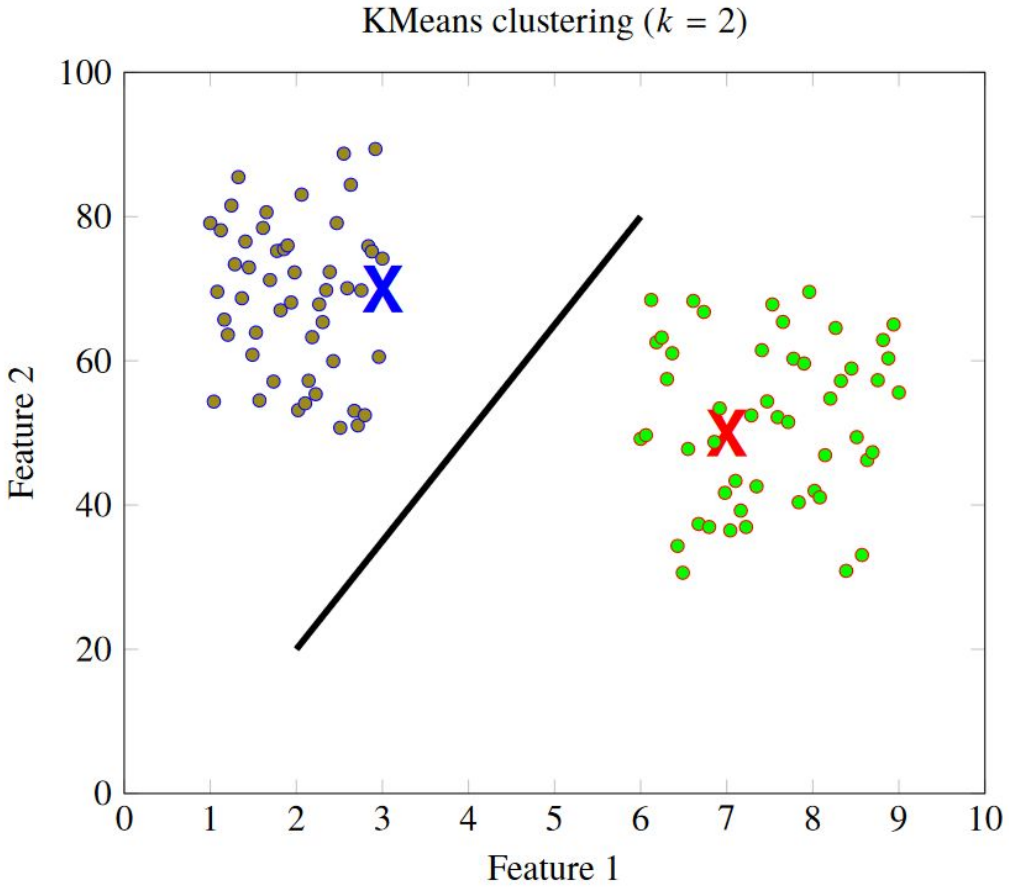
¿Esto es clustering?



Introducción (3)

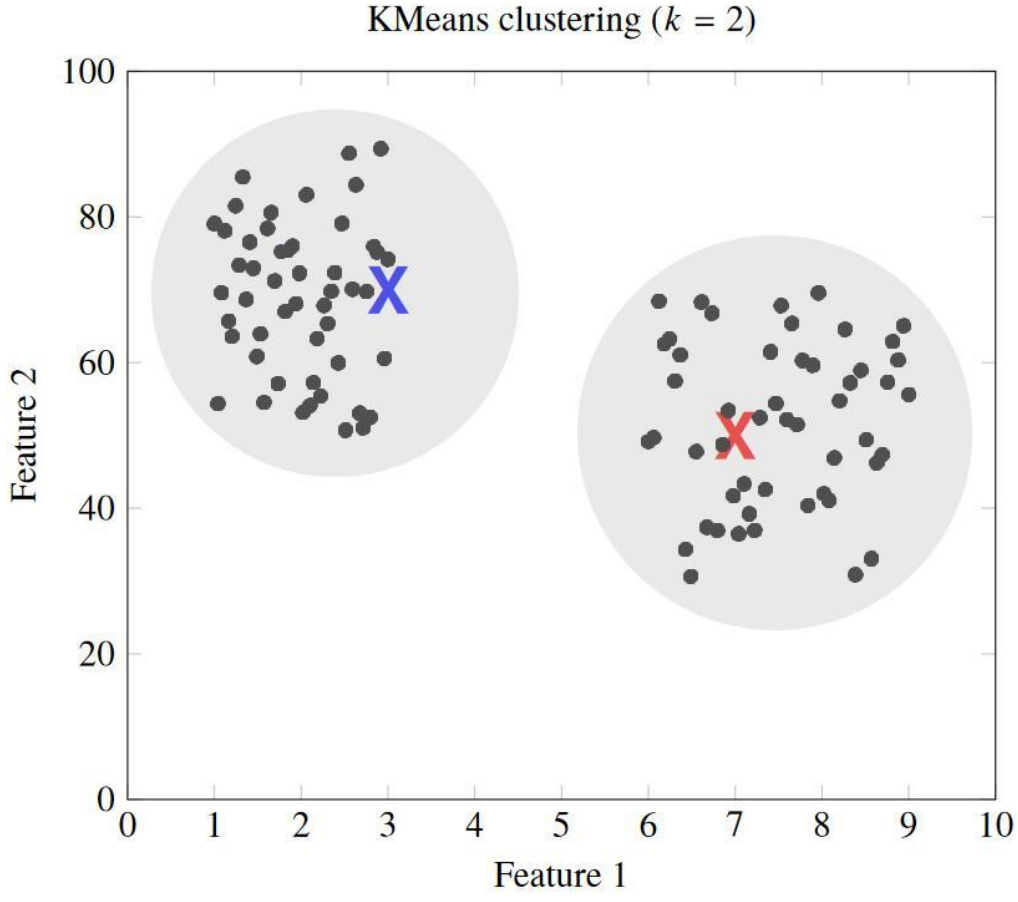
¿Esto es clustering?

NO! Es un clasificador lineal



Introducción (4)

Esto SÍ es clustering



Introducción (5)

Clasificación	Clustering
Los datos están etiquetados (supervisado)	No hay etiquetas (no supervisado)
	Los nuevos datos se agrupan según una métrica de proximidad.
Objetivo:	Objetivo:
Etiquetar correctamente datos no vistos	Identificar estructuras en los datos

Introducción (6) – Definición

- Clustering ≡ organización de una "colección de patrones en grupos basados en la similitud" (Jain, Murty, et al. 1999)
- Según (Jain and Dubes 1988), un cluster es:
  - Conjunto de objetos similares
  - **Conjunto de puntos similares en el entorno** | la distancia entre dos puntos en un grupo es menor que la distancia entre cualquier punto en el grupo y cualquier punto de otros grupos
  - Región densamente conectadas en un espacio multidimensional separado de otros por puntos débilmente conectados

Introducción (7) – Objetivos

- **Escalabilidad**
  El algoritmo debe ser escalable para grandes conjuntos de datos
- **Robustez**
  Los outliers deben detectarse con precisión
- **Independencia del orden**
  Diferente orden en los datos de entrada no deben conducir a diferentes resultados finales
- **Mínima elección de parámetros definidos por el usuario**
  Reducir la carga de configuración humana

Métricas de distancia (1) – Definición

Definition ($d(x_i, x_j)$)

The distance between two instances x_i and x_j, which is a metric distance measure if it satifies the following properties:

- ✓ *Triangle inequality*

$$d(x_i, x_j) \leq d(x_i, x_k) + d(x_k, x_j), \ \forall x_i, x_j, x_k \in \mathcal{S}$$

- ✓ $d(x_i, x_j) = 0 \rightarrow x_i = x_j \ \forall x_i, x_j \in \mathcal{S}$

Métricas de distancia (2) – Distancia Minkowski

- Distancia Minkowski

$$d(x_i, x_j) = (|x_{i,1} - x_{j,1}|^g + |x_{i,2} - x_{j,2}|^g + \ldots + |x_{i,p} - x_{j,p}|^g)^{1/g}$$

$$x_{i,k} \in [a, b] \subset R$$

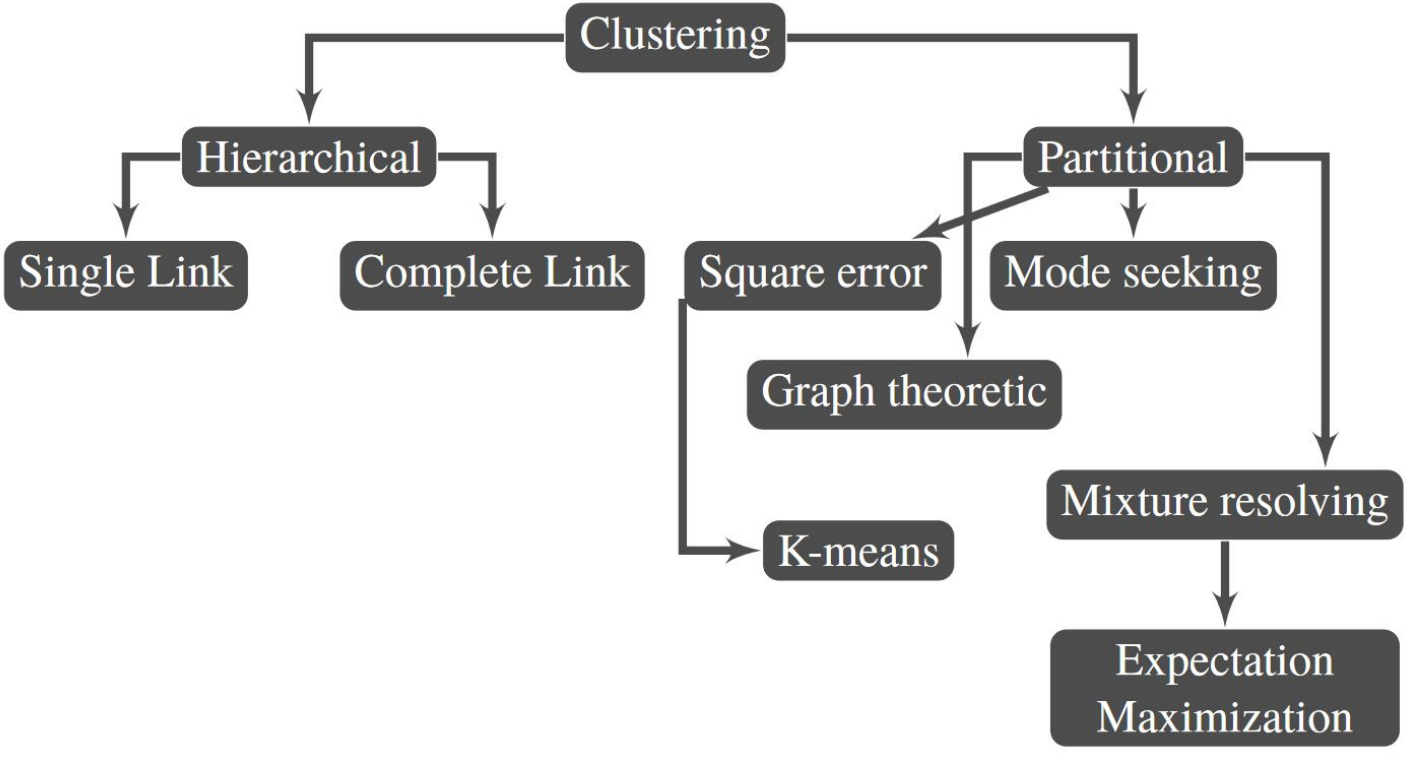
Si g = 2 ⇒ Euclidean

Si g = 1 ⇒ Manhattan

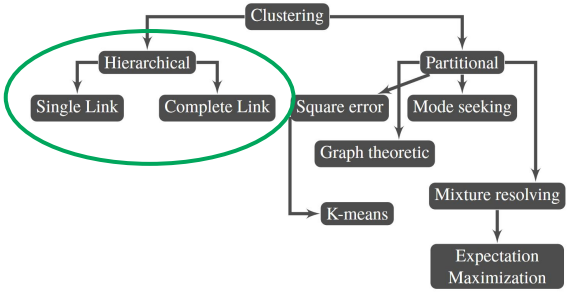
Métricas de distancia (3)

- Ejemplos de distancias en python:
  - Notebook 01_ejemplo_distancias.ipynb

Taxonomía de los algoritmos de clustering (1)



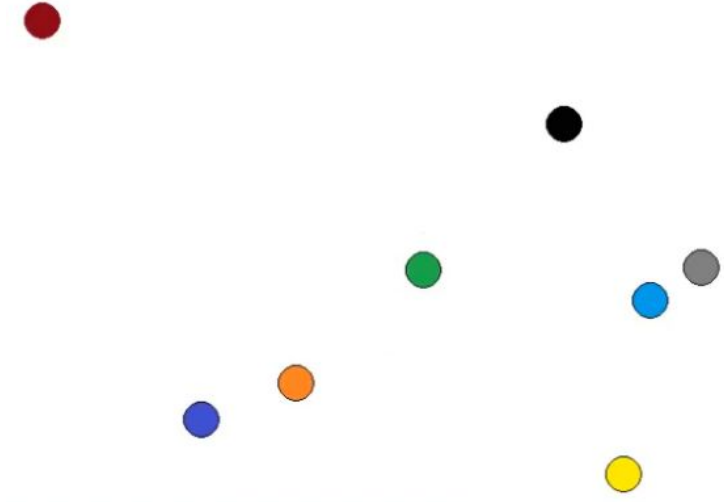
Taxonomía de los algoritmos de clustering (2)



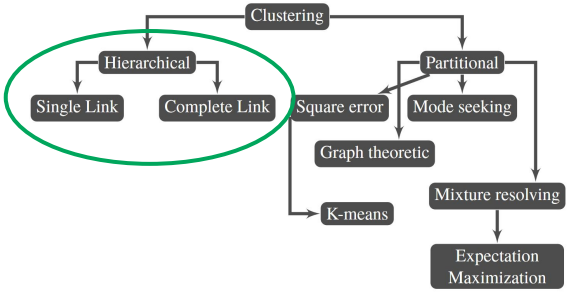
- Hierarchical clustering (Connectivity clustering)

  Los datos se van conectando (agrupando) de forma escalonada

  Datos iniciales:



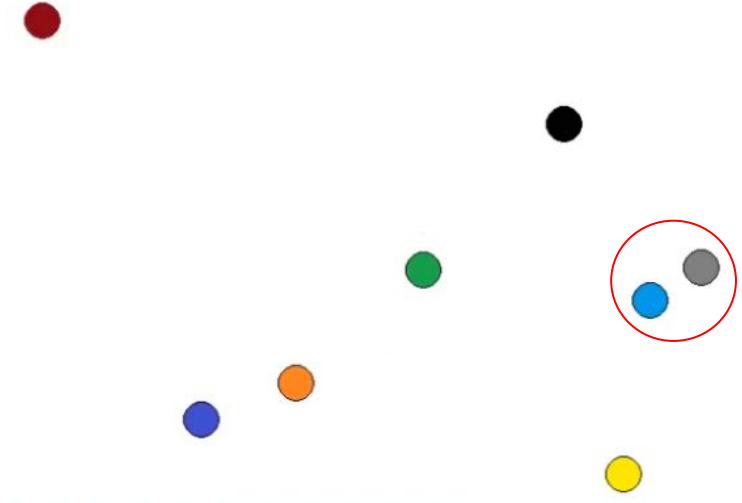
Taxonomía de los algoritmos de clustering (3)



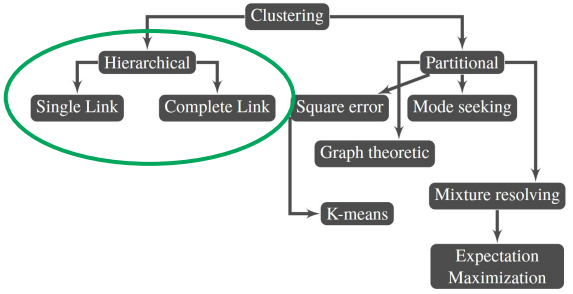
- Hierarchical clustering (Connectivity clustering)

  Los datos se van conectando (agrupando) de forma escalonada

  Paso 1:



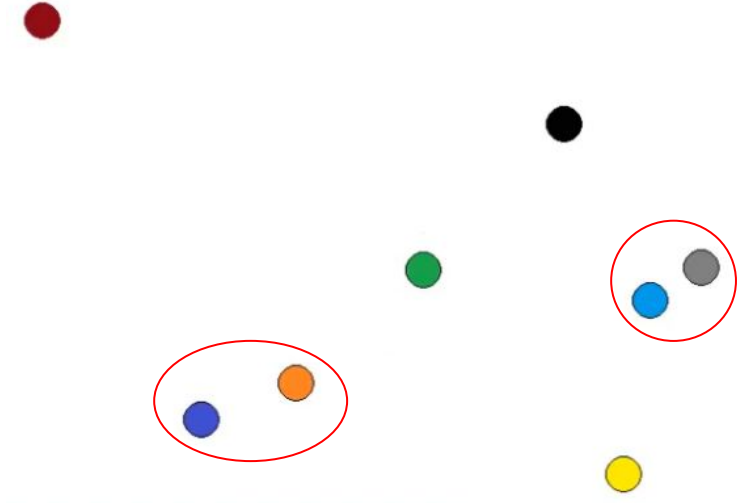
Taxonomía de los algoritmos de clustering (4)



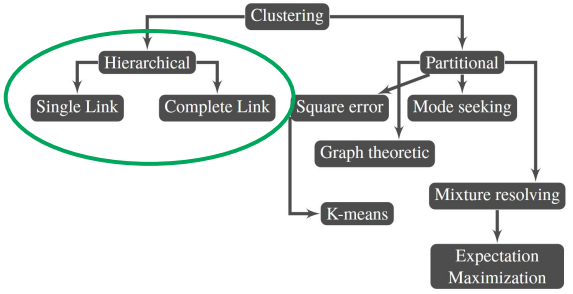
- Hierarchical clustering (Connectivity clustering)

  Los datos se van conectando (agrupando) de forma escalonada

  Paso 2:



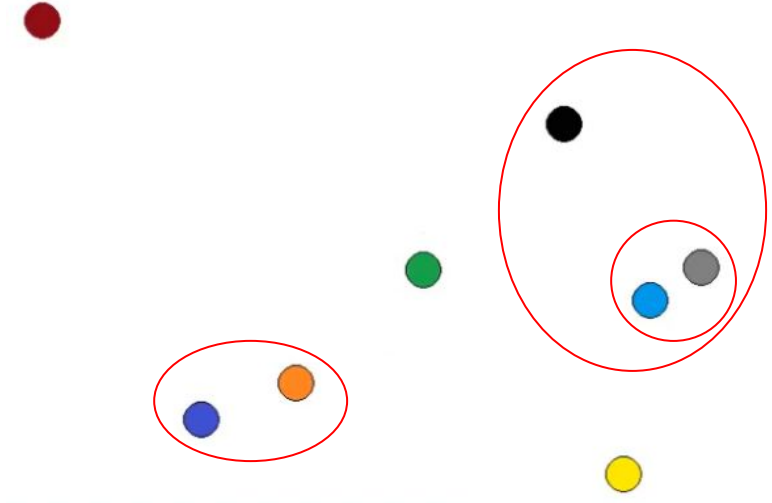
Taxonomía de los algoritmos de clustering (5)



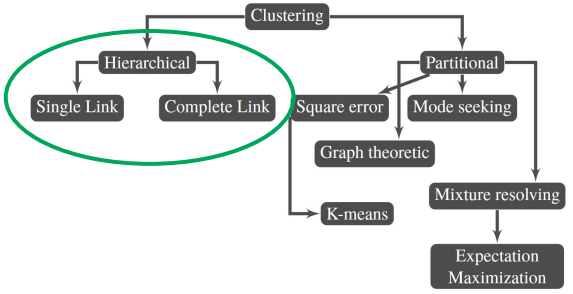
- Hierarchical clustering (Connectivity clustering)

  Los datos se van conectando (agrupando) de forma escalonada

  Paso 3:



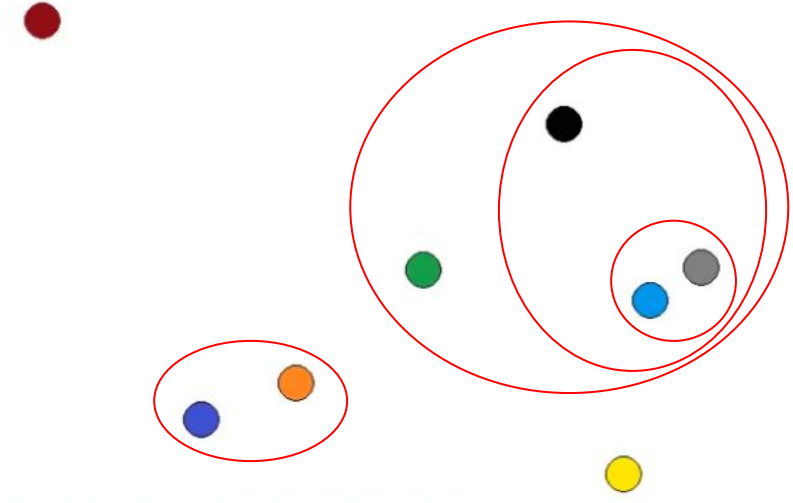
Taxonomía de los algoritmos de clustering (6)



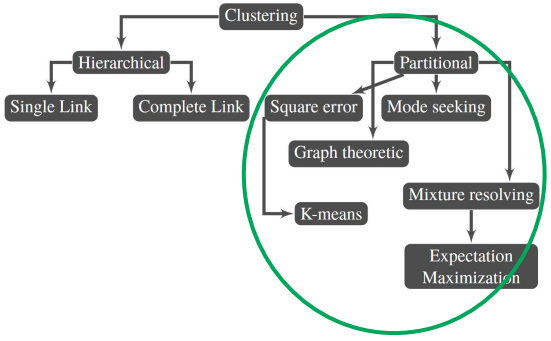
- Hierarchical clustering (Connectivity clustering)

  Los datos se van conectando (agrupando) de forma escalonada

  Paso 4:



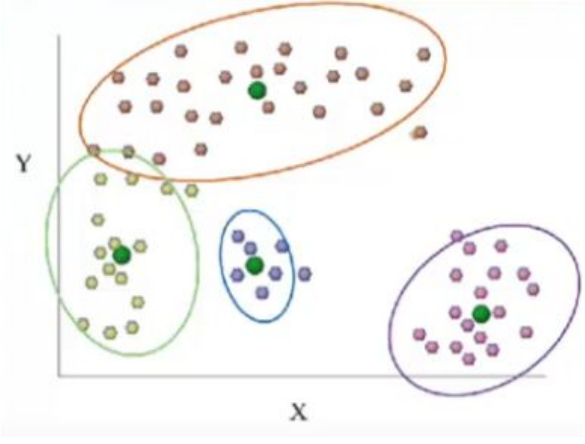
Taxonomía de los algoritmos de clustering (7)



- Partitional clustering (Centroid clustering)

  Se deben definir cuántos grupos se van a formar

  Se basan en centroides:



Objetivos del curso

- Conocer la mecánica de los algoritmos jerárquicos (connectivity)
- Conocer la mecánica de los algoritmos basados en centroides (partitional)
- Conocer la mecánica de los algoritmos basados en mezclas de gaussianas (EM)
- Aprender a utilizar las librerías típicas en Python