

Clustering

Partitional algorithms

Christian Oliva Moya

Dpto. de Ingeniería Informática, Escuela Politécnica Superior

Universidad Autónoma de Madrid

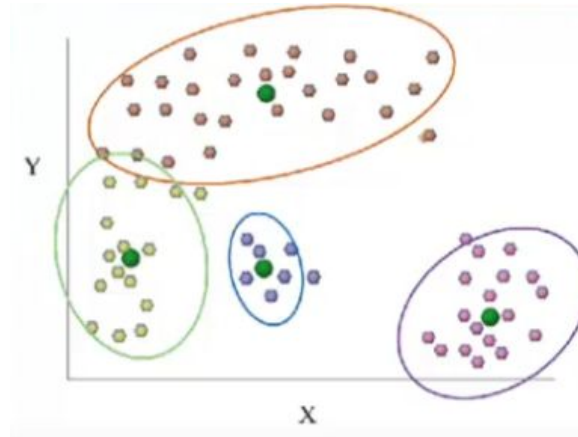
28049 Madrid, Spain

Introducción

- **Partitional clustering** (Centroid clustering)

Se deben definir cuántos grupos se van a formar

Se basan en centroides:



K-means (1)

Input S: conjunto de todos los posibles datos (N datos, D atributos)

Input K: número de clusters (hiperparámetro)

Output: K clusters $\{S_1, S_2, \dots, S_k\}$

inicializar los K centroides

while no se cumpla la condición de parada **do**

Asignar cada dato x_i al centroide más cercano

Actualizar los centroides según cierta operación

end

N patterns

$S = S_1 \cup S_2 \cup S_3 \cdots \cup S_k$

$S_i \cap S_j = \emptyset, \forall i \neq j$

$x_i \in S_j$

K-means (2)

Input S: conjunto de todos los posibles datos (N, D)

Input K: número de clusters

Output: K clusters $\{S_1, S_2, \dots, S_k\}$

inicializar los K centroides

while no se cumpla la condición de parada **do**

Asignar cada dato x_i al centroide más cercano

Actualizar los centroides según cierta operación

end

¿Cómo inicializar los K centroides?

¿Cuál es la condición de parada?

¿Cómo actualizar los centroides?

K-means (3)

¿Cómo inicializar los K centroides?

Se seleccionan **puntos aleatorios del conjunto de datos**

¿Cuál es la condición de parada?

Que en una iteración del algoritmo **no se modifique ningún centroide**

¿Cómo actualizar los centroides?

Actualizar el centroide por el **promedio de los puntos del cluster**

K-means (4)

- Notebook 03_fundamentos_kmeans.ipynb

K-means (5)

- Ventajas de K-means
 - Baja complejidad, fácil interpretación e implementación
 - Centroides como representantes de cada cluster
- Desventajas de K-means
 - Alta sensibilidad a la partición inicial
 - Alta sensibilidad a datos ruidosos y outliers
 - Solamente se puede calcular si la media está bien definida (datos numéricos)
 - Necesario definir correctamente el valor K

Problemas de inicialización en K-means: K-means++

- K-means es **extremadamente sensible a la inicialización de los centroides** y tiene una baja tasa de convergencia
- Medidas para solventarlo:
 - Desarrollar una **heurística para la selección de los centroides**. Por ejemplo, elegir el segundo como el más lejano al primero, elegir el tercero como el más lejano a ambos, etc.
 - Probar **múltiples inicializaciones** y escoger la que mejor resultado da.
 - Usar otros procedimientos de inicialización como **K-means++** (Pena et al. 1999)