1. Aprendizaje supervisado VS no supervisado

Aprendizaje estadístico → dos categorías:

- supervisado:
- no supervisado:

Relación entre las variables:

- Clustering
- PCA: resumir varias variables en un pequeño nº de ellas.

2. Agrupamiento o clustering (muy útil para la segmentación en marketing).

Conjunto amplio de técnicas para encontrar subgrupos o clústeres en un conjunto de datos. La similitud o diferencia entre observaciones debe definirse según el dominio y el conocimiento de los datos.
La agrupación busca subgrupos homogéneos entre las observaciones.
Otra manera de buscar la homogeneidad: densidad de las relaciones mediante la maximización de modularidad.
Otras metodologías buscan grupos no superpuestos y particiones superpuestas existentes.

3. KMEANS

Enfoque simple y elegante para particionar un conjunto de datos en K Clusters distintos y no superpuestos.

¿Cómo especificamos el número deseado de cúmulos KMEANS?

- Especificamos el nº deseado de cúmulos K.
- El algoritmo KMEANS asignará cada observación exacta a cada uno de los cúmulos K.

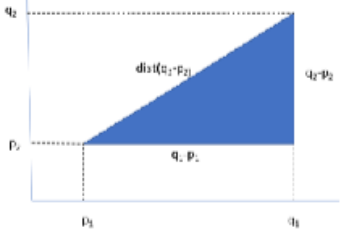
Buena agrupación → es aquella para la cual la variación dentro del clúster es lo más pequeña posible. La variación dentro del clúster para el grup oCk es una medida W(Ck) de la cantidad en la que las observaciones dentro de un clúster difieren entre sí.
Buscamos una división que haga que la variación interna sea la mínima → la definimos con la distancia euclidiana al cuadrado.

$$J(c,\mu) = \sum_{1=1}^{m} \|x^i - \mu_c(i)\|^2$$

Enfoque simple → permite encontrar esa división. Los centroides son la media de las observaciones en un cúmulo.

$$dist\ (p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}$$

Disminuye la variación interna y mejora hasta que no haya cambios, pero termina en un óptimo local.



4. Limitaciones del KMEANS

El modelo K-means clustering presenta dos limitaciones principales:

- la elección del número de grupos K.
- la posibilidad de que la solución encontrada sea un óptimo local en lugar de un óptimo global.

Para aplicar el modelo, es necesario estandarizar la escala de las diferentes características.

- La elección de K no es simple y existen metodologías para comparar y validar los resultados.
- El problema del óptimo local puede resolverse ejecutando el algoritmo varias veces con diferentes centroides y seleccionando el que tenga la medida de variación interna más baja.

Es una buena práctica estandarizar los datos usando la misma escala o al menos intentar que sean lo más homogéneos posible y hacerlos centrados en la media.

5. Ejemplos del clustering

Muchos de los ejemplos mezclan la clasificación y la agrupación. Para clasificar algo usamos algoritmos de aprendizaje supervisado.

- Segmentación de marketing.
- noticias falsas
- detección de fraudes
- las imágenes médicas segmentadas.