Aprendizaje Automático

Métodos de Aprendizaje No Supervisado
Primera Parte

2023-2Q

Aprendizaje Supervisado vs. No Supervisado

Aprendizaje Supervisado

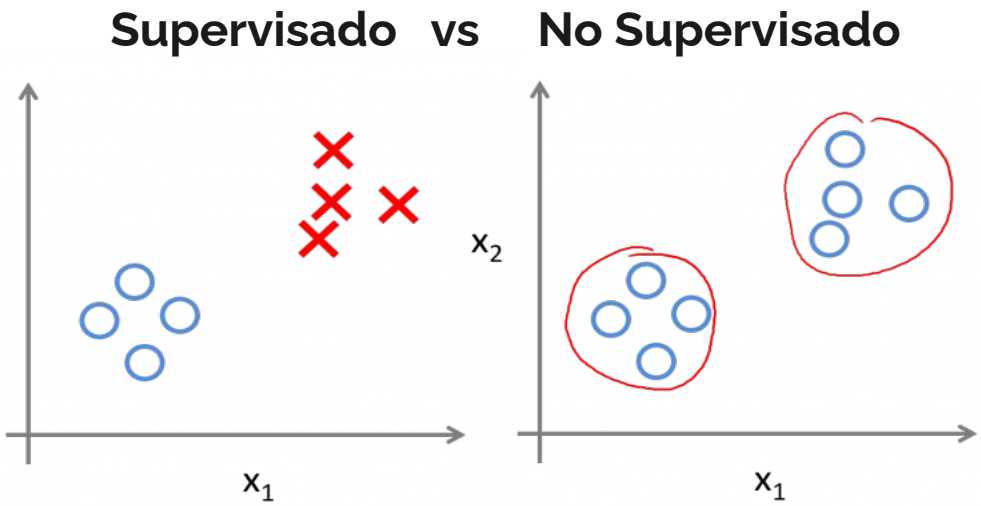
Construyen modelos de predicción basándose en el conocimiento de la variable Respuesta

Aprendizaje NO Supervisado

Construyen modelos de predicción cuando la variable Respuesta no es una información disponible.

Aprendizaje No Supervisado

Los **datos** de entrenamiento **no están anotados.**



Consiste en analizar y entender las relaciones existentes entre las variables observadas.

CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

SUPERVISED

Predict a category

CLASSIFICATION

«Divide the socks by color»

Predict a number

REGRESSION

«Divide the ties by length»

Data is not labeled in any way

UNSUPERVISED

Divide by similarity

CLUSTERING

«Split up similar clothing into stacks»

Find hidden dependencies

DIMENSION REDUCTION (generalization)

«Make the best outfits from the given clothes»

Métodos No Supervisados

Algunos métodos comunes de aprendizaje no supervisado son:

- **Clustering o agrupamiento**
- **Reducción de dimensionalidad**
- **Asociación**

Clustering

El análisis de clusters es una técnica para resolver problemas de clasificación no supervisada.

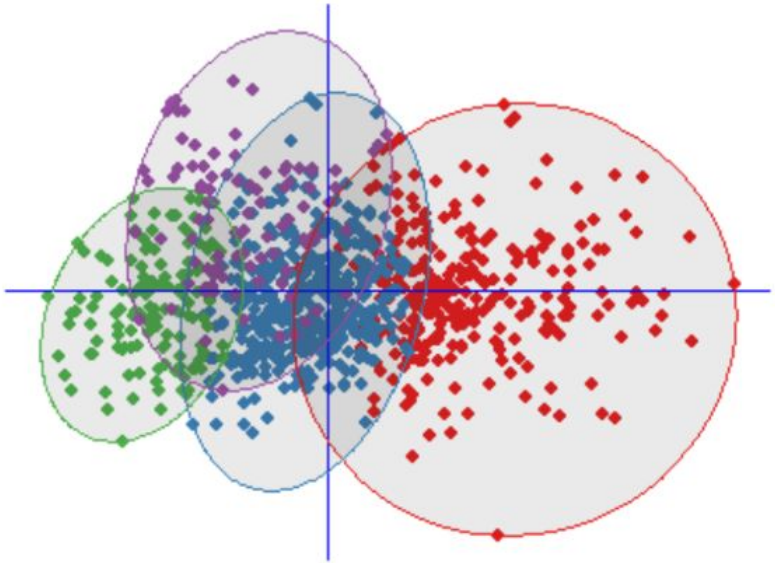
Clustering o agrupamiento:

Se agrupan las observaciones de acuerdo a algún criterio.

Clustering o agrupamiento

¿Qué hacen?

Agrupan objetos en conglomerados o clusters de forma tal que el grado de asociación o similitud entre miembros del mismo cluster sea lo más fuerte posible.



Cluster basados en prototipos

Un cluster es un conjunto de objetos en el cual cada objeto está más cerca (o es más similar) al **prototipo** que define al cluster que al prototipo que define cualquier otro cluster.

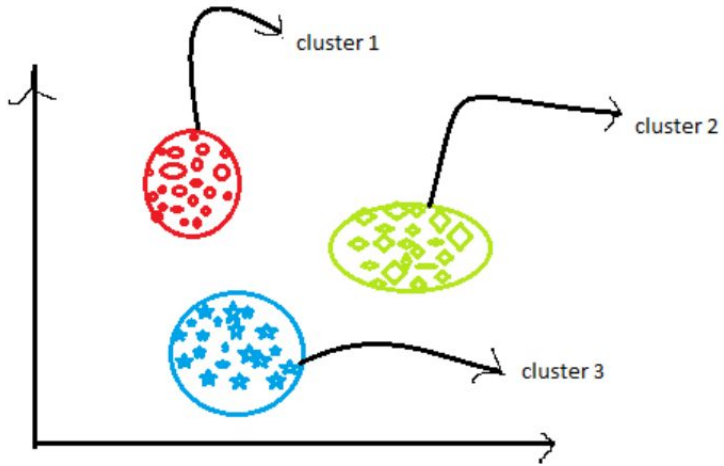
- **Atributos continuos:** el prototipo de un cluster es usualmente el centroide.
- **Atributos categóricos:** el prototipo es el objeto más representativo del cluster.

K-means

¿Qué es la agrupación de k-medias?

Es un método de agrupación que tiende a dividir sus datos en particiones llamadas clústeres. Se asignan cada uno de los puntos de datos al grupo con la media más cercana.

k representa el **número de clusters**, es decir, k grupos.



Algoritmo K-medias

Dado un conjunto de observaciones $\{x_1, x_2, \ldots x_n\}$, donde cada observación es un vector real de dimensión p.

El algoritmo K-medias construye una partición de las observaciones en K conjuntos ($K \leq n$) que minimiza la distancia de los elementos dentro de cada grupo: $S = \{S_1, S_2, \ldots, S_k\}$

$$\arg\min_{S} \sum_{i=1}^{K} \sum_{x_j \in S_i} |x_j - \mu_i|^2$$

donde μ_i es la media de puntos en S_i

Algoritmo K-medias

Agrupa el conjunto de datos en K conjuntos no solapados

- Definir K
- n observaciones
- p variables
- $\{C_1, C_2, \ldots, C_k\}$ son los conjuntos donde están los índices de las observaciones tal que

$$C_1 \cup C_2, \cdots \cup C_K = \{1, \ldots, n\}$$
$$C_i \cap C_j = \emptyset, \ \forall i \neq j$$

La idea de K-medias

Un buen agrupamiento es aquel que la variación dentro de un mismo cluster es pequeña.

Pero...

¿Cómo medimos la variación?

Medida la variación W

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,j \in C_k} \sum_{l=1}^{p} (x_{il} - x_{jl})^2$$

- |Ck | el número de elementos de la clase k.
- p es la cantidad de variables.

Medida la variación W

Queremos minimizar

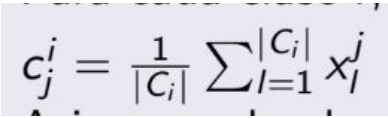
$$min_{C_1,\ldots,C_K} \sum_{k=1}^{K} W(C_k)$$

K- medias, es entonces un problema de optimización no lineal.

El Algoritmo

Dado K

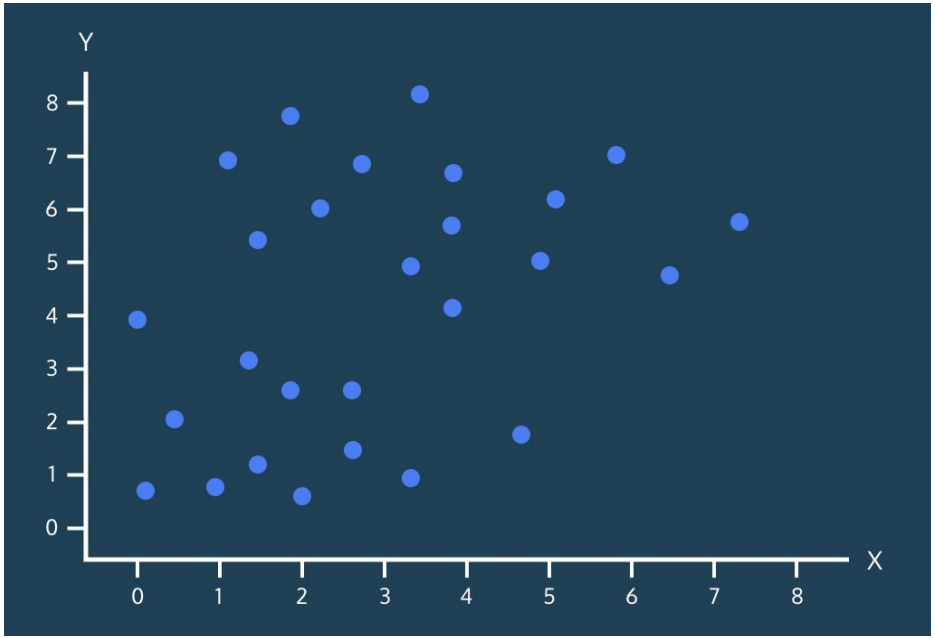
1. Asignar aleatoriamente un número de 1 a K a cada una de las observaciones
2. Realizar los siguientes pasos hasta que la asignación de clusters se mantenga estable de una iteración a otra.
   a. Para cada clase i, calcular el centroide $c^i = (c^i_1, c^i_2, , c^i_p)$ donde



   b. Asignar cada observación al cluster cuyo centroide está más cerca en distancia euclídea

Algoritmo K-medias

1. **Inicialización**
2. **Asignación**
3. **Actualizar**
4. **Repetir**



Clasificación utilizando K-medias

Observar que

- Cuando de un paso a otro no hay modificaciones, significa que se está en presencia de un mínimo local.
- El método modifica la clasificación si existe una mejoría, por lo tanto siempre converge a una clasificación mejorada.

Link visualización

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

Clasificación utilizando K-medias

Inconveniente

El resultado final de la clasificación depende fuertemente de la asignación de clases que se haya utilizado en el primer paso del algoritmo.

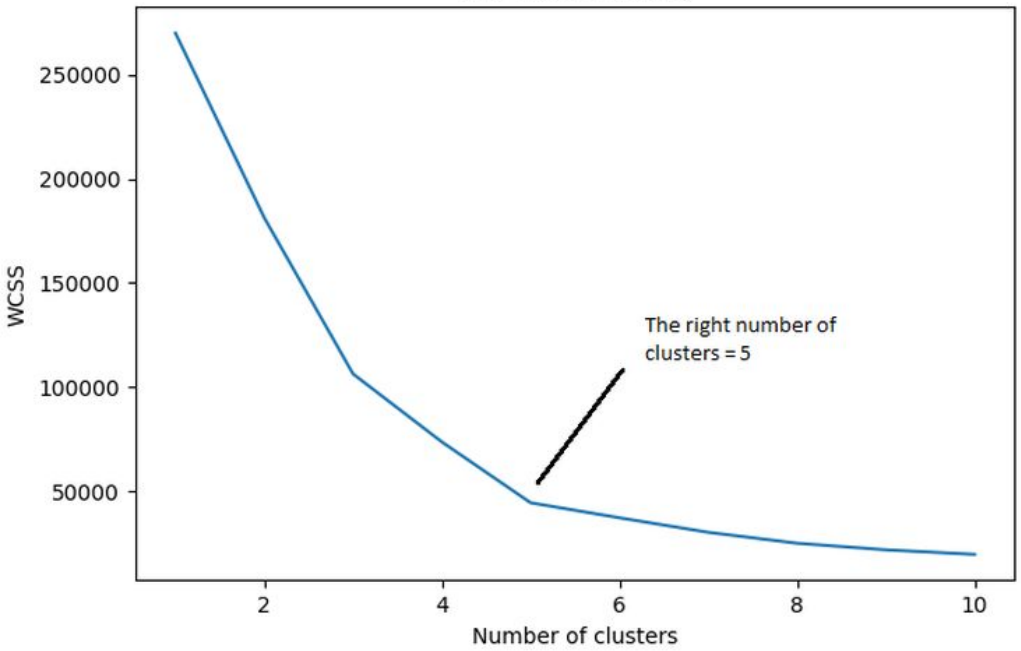
Clasificación utilizando K-medias

Para solucionarlo

- Aplicar el método varias veces.
- Elegir la clasificación que minimiza $\sum_{k=1}^{K} W(C_k)$

Método del codo

Busca el número adecuado de clústeres identificando el punto donde la disminución de la variabilidad se detiene abruptamente, formando un **"codo"** en el gráfico.

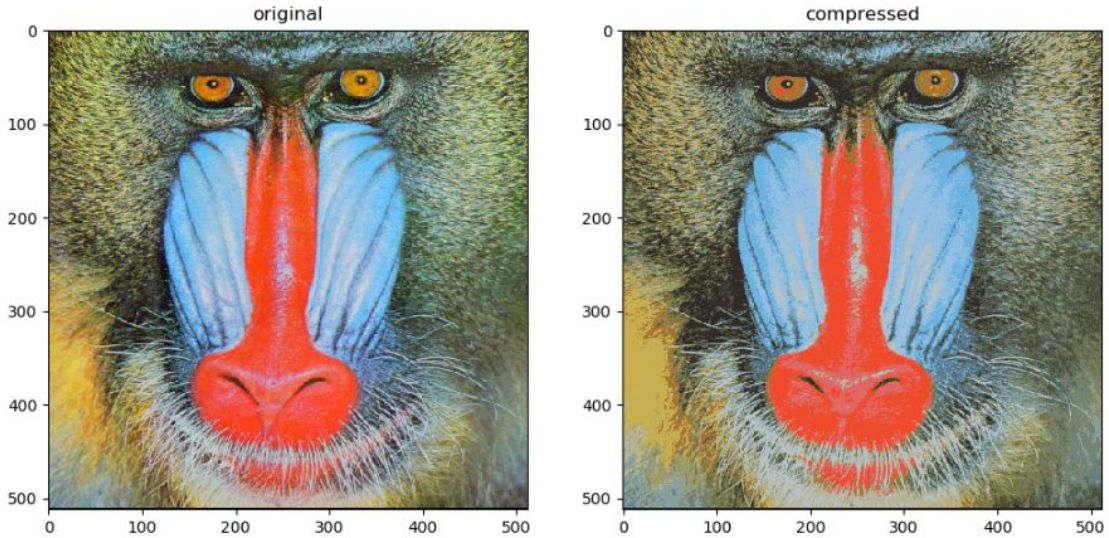


Usos de K-medias

Reducir el tamaño del archivo de una imagen sin reducir significativamente su calidad.

Usos de K-medias

Reduciendo de 16,77 millones de colores a 16 colores para ver qué tan buena es la compresión



¡Fin! ¿Alguna pregunta?