COMPUTACIÓN

TEC | Tecnológico de Costa Rica

Aprendizaje automático

No supervisado - Clustering

María Auxiliadora Mora Cross, agosto de 2023

Índice

- Introducción
- Tipos de agrupamientos (*clustering*)
- El algoritmo K-Means
- Criterio de convergencia
- Función distancia
- Ventajas y desventajas
- Evaluación

Aprendizaje supervisado y no supervisado

- Aprendizaje supervisado
  - **Descubre patrones** en los datos que relacionan los atributos de los ejemplos con una clase meta (por ejemplo en clasificación).
  - Los patrones luego son **utilizados para predecir la clase a la que pertenece una nueva instancia**.
- Aprendizaje no supervisado
  - Los **datos no contienen el atributo meta**.
  - Los algoritmos exploran los datos para encontrar estructuras y relaciones intrínsecas en ellos.

Clustering

- Es una técnica para encontrar grupos de instancias similares a partir de los datos (*clusters*).
  - Mantiene instancias de datos que son similares (cerca) en un grupo y datos que son diferentes (alejados) en grupos diferentes.
  - Objetivo maximizar la similitud dentro de la clase y minimizar esta entre clases.

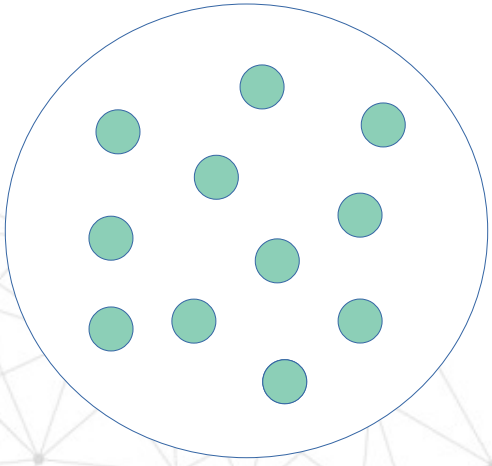
Ejemplos de uso

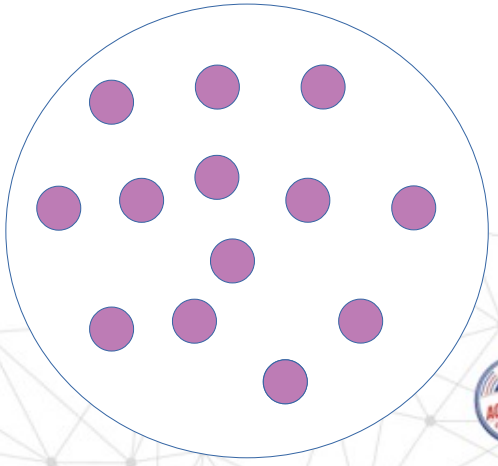
- Calcular las épocas del año, seca y lluviosa (basado en temperatura y humedad).
- Grupos de personas de tallas similares para diseñar tamaños estándar de una prenda de vestir.
- Mercadeo de productos, segmentación de clientes.
- Inclinación en votos presidenciales (indecisos).
- Dado una colección de textos, se desea organizarlos de acuerdo a similitudes en el contenido.
  - Clustering es una de las técnicas más utilizadas para minería de datos.

Diferentes tipos de *clusters*

Bien separados

- Cada punto está más **cerca de todos los puntos de su grupo** que de cualquier punto de otro grupo
- A veces se usa un **umbral** para especificar que todos los objetos en un clúster deben estar lo suficientemente cerca entre sí.
- Esta definición idealista de un clúster se satisface solo cuando los datos contienen ***clusters naturales*** que están bastante lejos unos de otros.

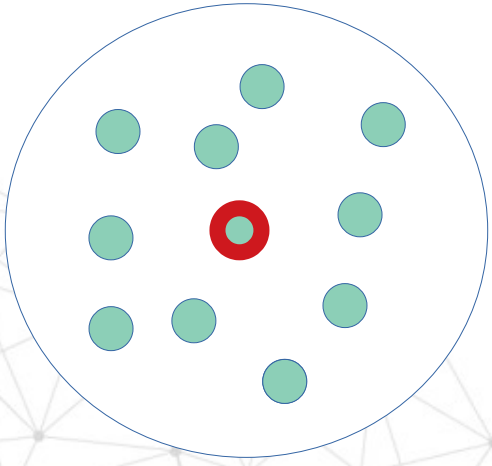


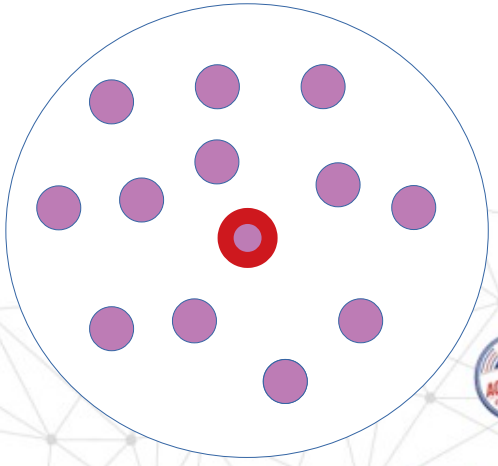


Diferentes tipos de *clusters*

Basado en prototipo

- Cada objeto está más cerca **(es más similar) al prototipo** que define el clúster que al prototipo de cualquier otro clúster.
- Para los datos con atributos continuos, el prototipo de un clúster es a menudo un **centroide**, es decir, el promedio de todos los puntos en el grupo.
- Cuando un centroide no es significativo, el prototipo es a menudo un medoide, es decir, el punto más representativo del clúster.
- El prototipo puede considerarse como el punto central del grupo.

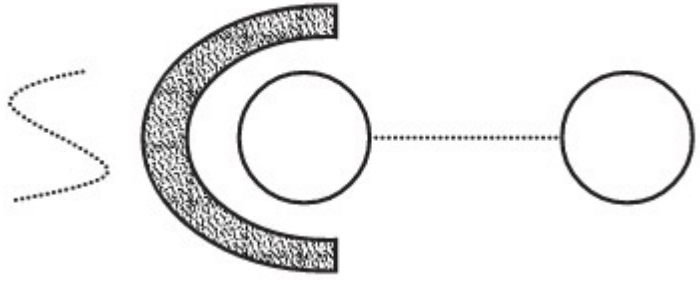




Diferentes tipos de *clusters*

Basado en grafos

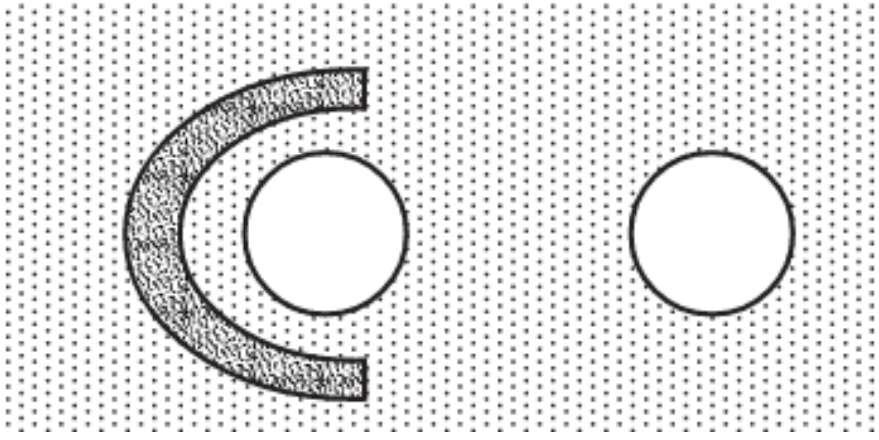
- Los **datos se representan como un grafo**, donde los nodos son objetos y los enlaces representan conexiones entre objetos
- Entonces, un agrupamiento puede definirse como un grupo de objetos que están conectados entre sí, pero que **no tienen conexión también con objetos fuera del grupo**.
- Un **ejemplo** de clúster basados en grafos son los basados en la contigüidad, donde dos objetos están conectados solo si están dentro de una distancia específica entre sí.



Diferentes tipos de clústeres

Basado en densidad

- Un *cluster* es una región densa de objetos que está rodeada por una región de baja densidad.
- Ejemplo: DBSCAN



Algoritmo K-Means

- *Cluster* de particionamiento basado en prototipos.
- Definición: sea un conjunto de n instancias de datos dadas por

$$D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\},$$

donde $\mathbf{x}_i = (a_1, a_2, \ldots, a_m)$ es un vector en R^m, y m es el número de atributos en los datos.

- El algoritmo particiona los datos en k *clusters*
  - Cada *cluster* tiene un centroide.
  - k es especificado por el usuario.

Algoritmo K-Means

Dado k, el algoritmo trabaja como sigue:

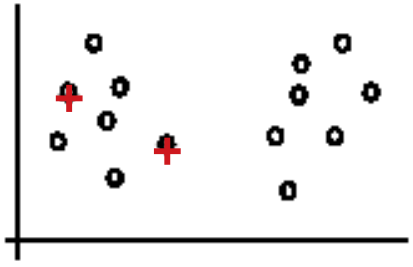
1) Se seleccionan aleatoriamente k centroides.

2) Se asocia cada instancia al centroide más cercano utilizando la distancia Euclidiana.

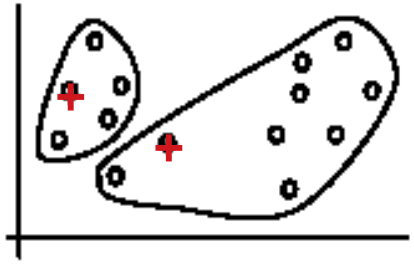
3) Se recalculan los centroides de acuerdo al conjunto de puntos en cada clúster.

4) Si el criterio de convergencia no se ha cumplido se vuelve al paso 2).

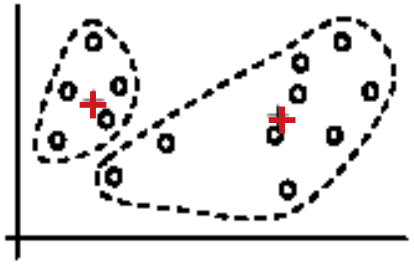
Algoritmo K-Means (cont.)



(A). Random selection of k centers



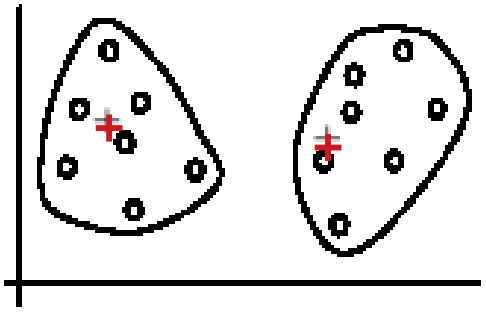
Iteration 1: (B). Cluster assignment



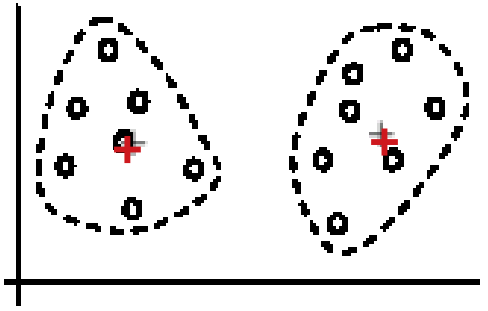
(C). Re-compute centroids

Imagen: Bing Liu (2019)

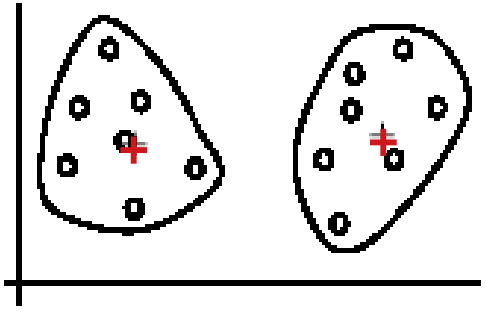
Algoritmo K-Means (cont.)



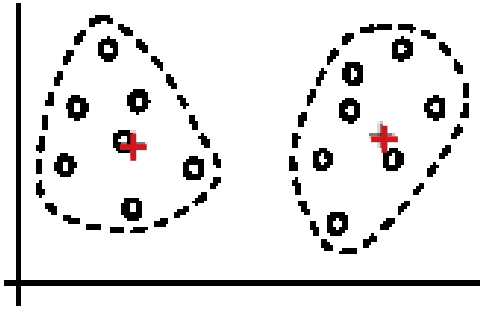
Iteration 2: (D). Cluster assignment



(E). Re-compute centroids



Iteration 3: (F). Cluster assignment



(G). Re-compute centroids

Imagen: Bing Liu (2019)

Criterios de convergencia

- Re-asignación mínima (o ninguna) de datos a diferentes clústeres.
- Mínimo cambio en el centroide (o ninguno, en caso de K-Means).
- Mínima disminución en la suma de cuadrados del error (o Sum of squared errors SSE).

$$SSE = \sum_{j=1}^{k} \sum_{\mathbf{x} \in C_j} dist(\mathbf{x}, \mathbf{m}_j)^2$$

Con: $\boldsymbol{C_i}$ el *jésimo* cluster, $\mathbf{m}_j$ el centroide de C_j, y *dist*(**x**, $\mathbf{m}_j$) es la distancia entre el punto **x** y el centroide $\mathbf{m}_j$.

Cálculo de centroides

La media (m_j) de cada *cluster* (C_j) en el Espacio Euclideano se define como:

$$m_j = \frac{1}{p_j} \sum_{X_i \in C_j} x_i$$

Con: p_j la cantidad de puntos en C_j y x_i puntos en C_j.

Cálculo de las distancias

- Distancia Euclideana de un punto (x_i) al centroide (m_j) del *cluster* (C_j) se define como:

$$dist(x_i, m_j) = \left\| x_i - m_j \right\| = \sqrt{(x_{i1} - m_{j1})^2 + (x_{i2} - m_{j2})^2 + \ldots + (x_{ia} - m_{j1a})^2}$$

- La distancia Euclideana es la más utilizada.
- Diferentes funciones distancia pueden ser aplicadas.

Cálculo de las distancias (cont.)

- La distancia de Manhattan:

$$dist\left(x_i, x_j\right) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ir} - x_{jr}|$$

- Distancia Euclídea con **pesos**:

$$dist(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{w_1 (x_{i1} - x_{j1})^2 + w_2 (x_{i2} - x_{j2})^2 + \ldots + w_r (x_{ir} - x_{jr})^2}$$

- Distancia de Chebychev: Qué tan diferentes son dos puntos en cualquiera de sus atributos.

$$dist(\mathbf{x}_i, \mathbf{x}_j) = \max(|x_{i1} - x_{j1}|, |x_{i2} - x_{j2}|, \ldots, |x_{ir} - x_{jr}|)$$

Ventajas y desventajas de k-means

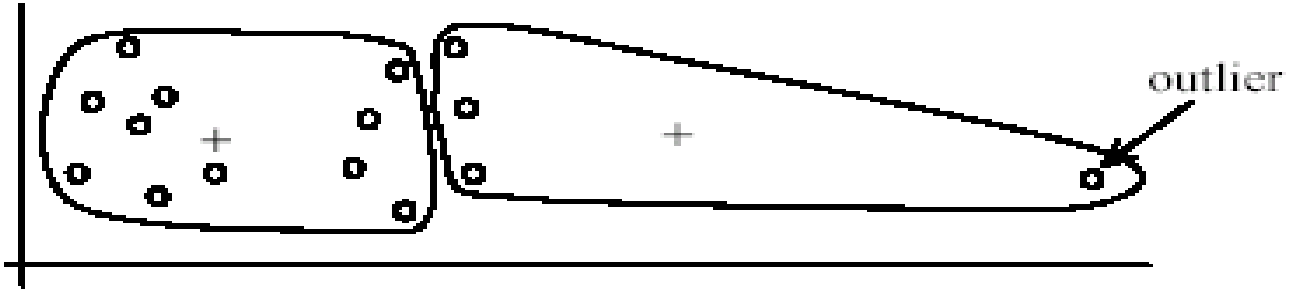
Ventajas

- Fácil de implementar y entender.
- Es el algoritmo más utilizado para *clustering*.
- Eficiente:
  - Complejidad en tiempo de O(tkn)
    - con n el número de puntos, k el número de grupos y t el número de iteraciones.
  - Como k y t son pequeños el algoritmo se considera lineal.

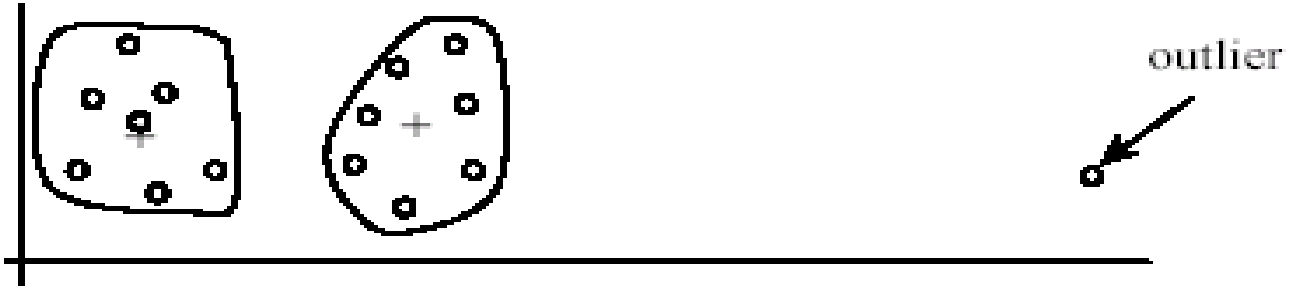
Desventajas

- Puede converger a mínimos locales.
- Lento en conjuntos de datos grandes.
- El algoritmo es sensible a los valores atípicos, a los centroides iniciales aleatorios y la distribución de los datos.

Desventaja: Valores atípicos



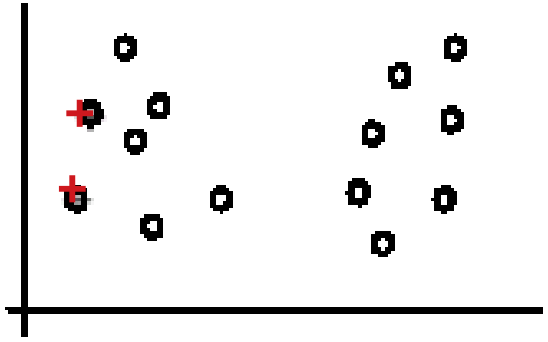
(A): Undesirable clusters



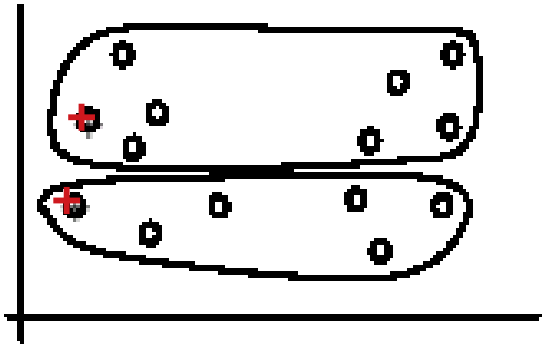
(B): Ideal clusters

Fuente: Bing Liu (2019)

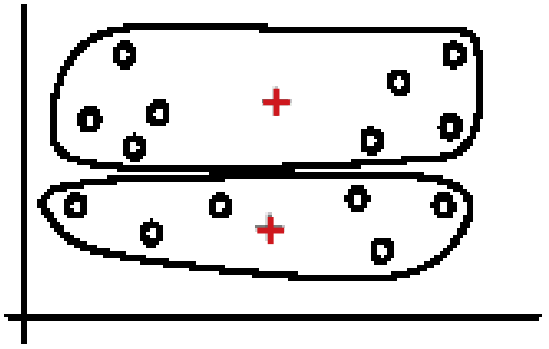
Desventaja: selección aleatoria de centroides iniciales



(A). Random selection of seeds (centroids)



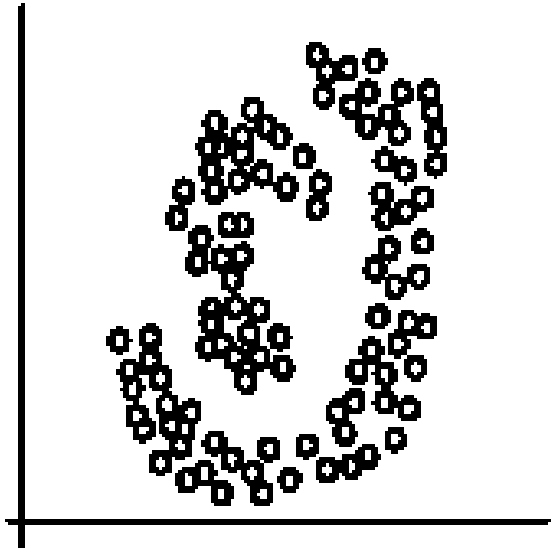
(B). Iteration 1



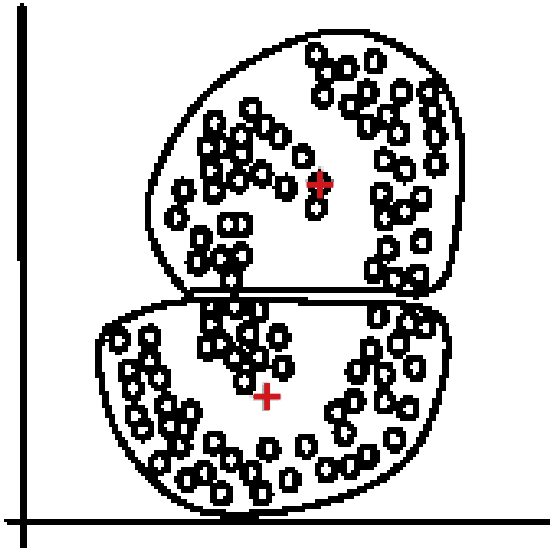
(C). Iteration 2

Fuente: Bing Liu (2019)

Desventaja: distribución de los datos



(A): Two natural clusters

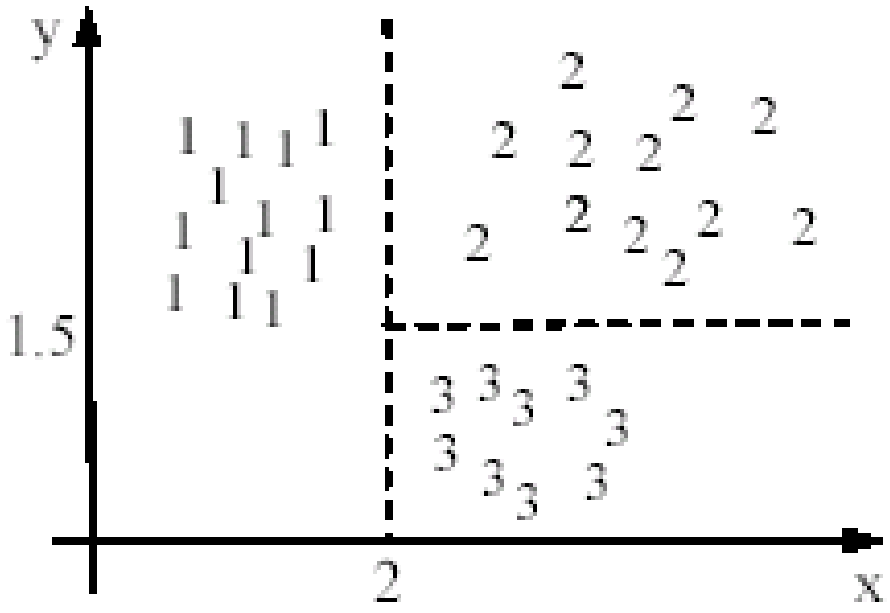


(B): *k*-means clusters

Fuente: Bing Liu (2019)

Clústeres como parte del flujo de trabajo

- Generar modelos de clasificación a partir del clúster



El *clustering* genera ejemplos clasificados

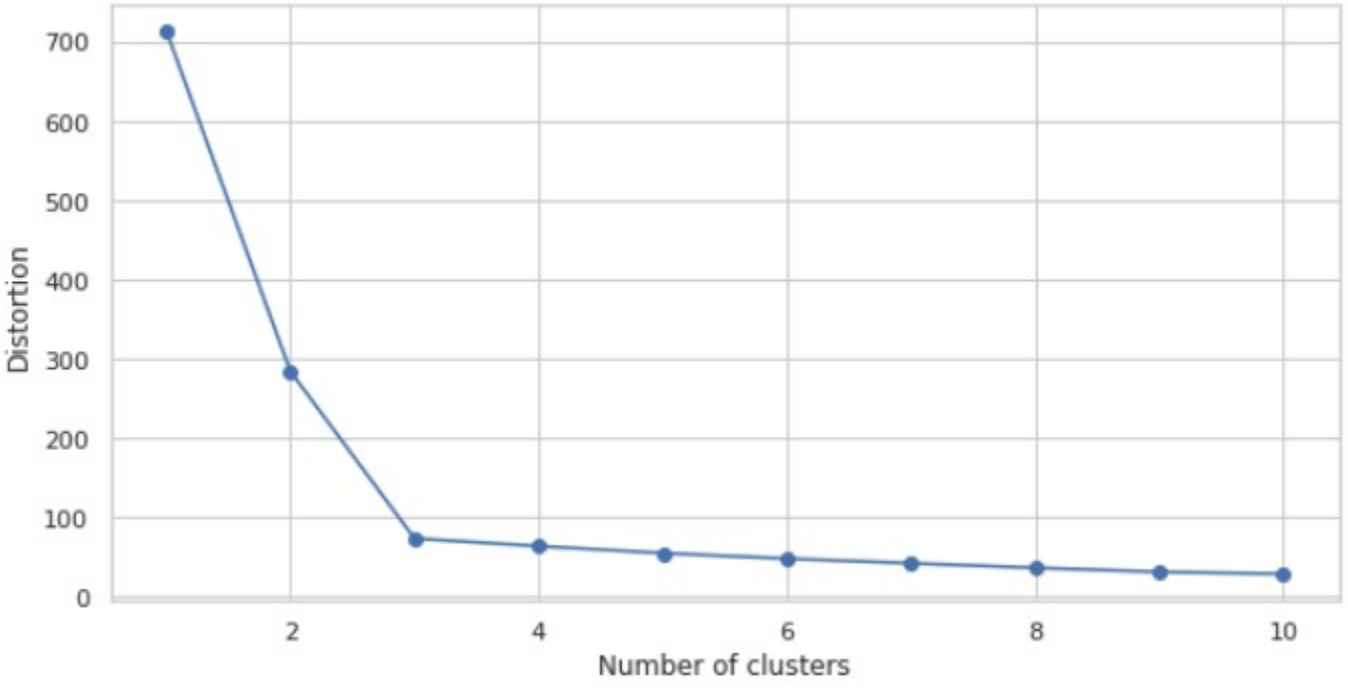
Evaluación

- La calidad del cluster es muy difícil de evaluar automáticamente.
  - No hay puntos de comparación
- Algunos métodos utilizados:
  - Evaluación de usuario
  - Cohesión interna
  - Separación entre clústeres
  - Evaluación indirecta

¿Cómo escoger el k en K-Means? Método del codo

- Intuitivamente, si k aumenta, la distorsión dentro del *cluster* disminuirá. Esto se debe a que las muestras estarán más cerca de los centroides a los que están asignadas.
- La idea es identificar el valor de k donde la distorsión comienza a disminuir más rápidamente, lo que será más claro si trazamos la distorsión para diferentes valores de k
- Cómo funciona:
  - Suma de los cuadrados de las distancias de todos los puntos de un clúster.
  - Usar diferentes valores de k.
  - Graficar los valores de k.

Método del codo



Referencias

- Harrington Peter (2012). Machine Learning in Action. ManningPublications Co. USA.
- Bing Liu (2019). Data Mining and Text Mining. Recuperado de https://www.cs.uic.edu/~liub/teach/cs583-fall-05/CS583-unsupervised-learning.ppt. Department of Computer Science. University of Illinois at Chicago (UIC)
- Tan, *P.*, Steinbach, M., Karpatne A., Kumar, V.(2018). Introduction to Data Mining (Second Edition). Recuperado de https://www-users.cs.umn.edu/~kumar001/dmbook/index.php