Instituto Politécnico Nacional

Centro de Investigación en Computación

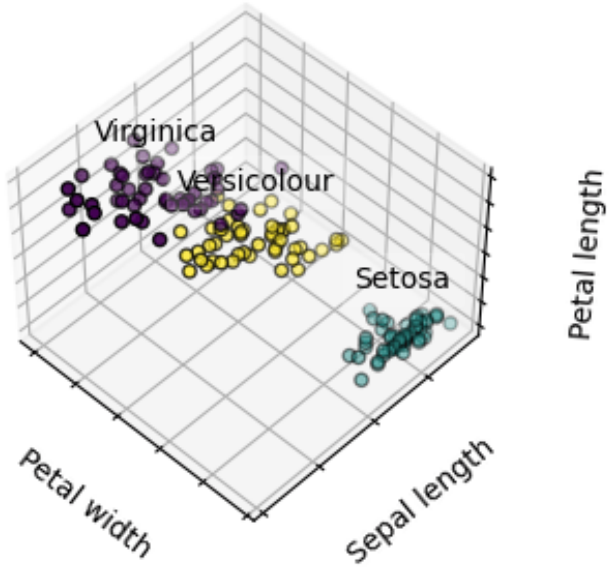
Agrupamiento de textos

Presenta: Erick Quintana Martínez

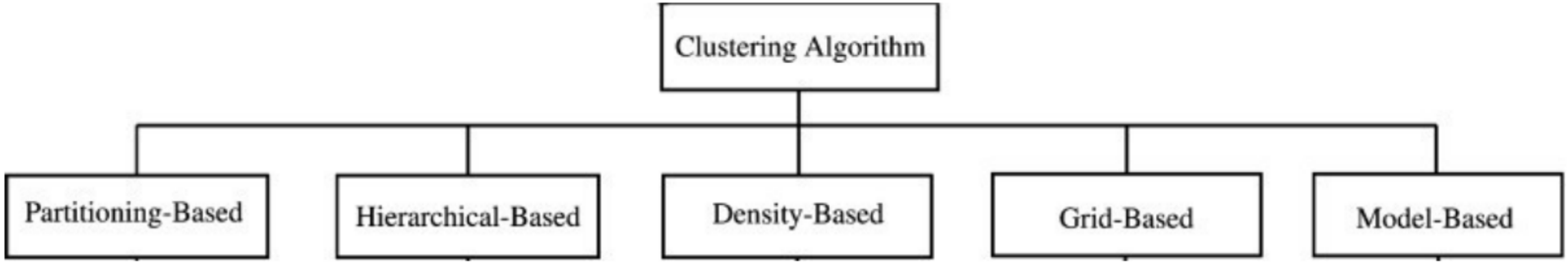
Ciudad de México, 7 de diciembre de 2022

Clustering

- Una forma de aprendizaje no supervisado es la agrupación (en inglés, clustering).
- Consiste en la división de los datos en grupos de objetos similares.

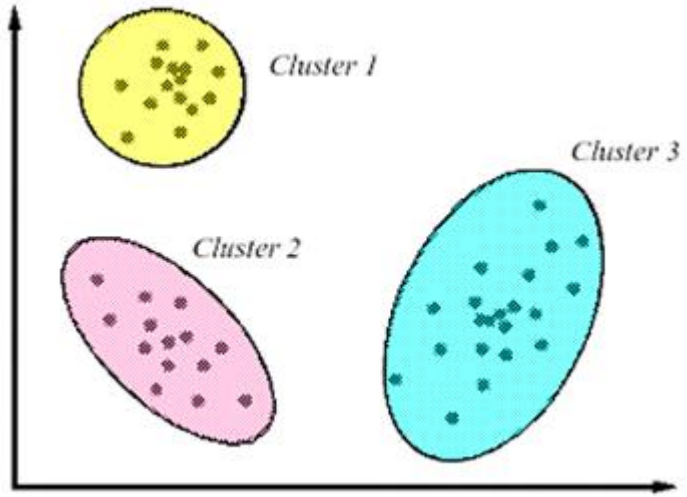


Tipos de algoritmos de clustering



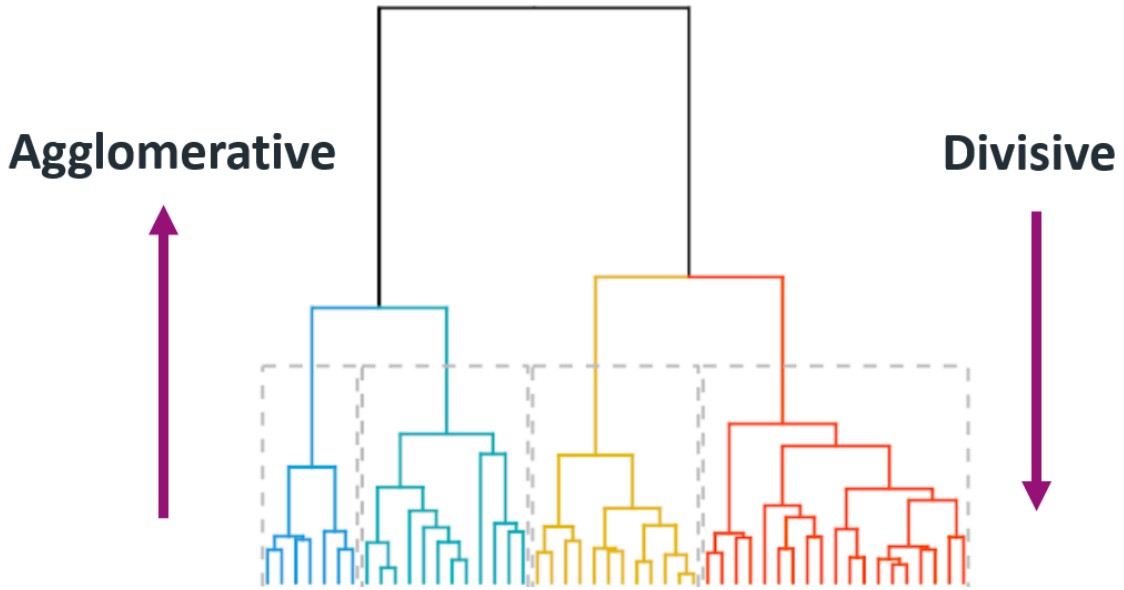
Basados en particiones

- Los algoritmos de partición dividen los datos en particiones, donde cada partición representa un grupo.
- En estos algoritmos cada clúster debe contener al menos un objeto, y cada objeto debe pertenecer exactamente a un grupo.



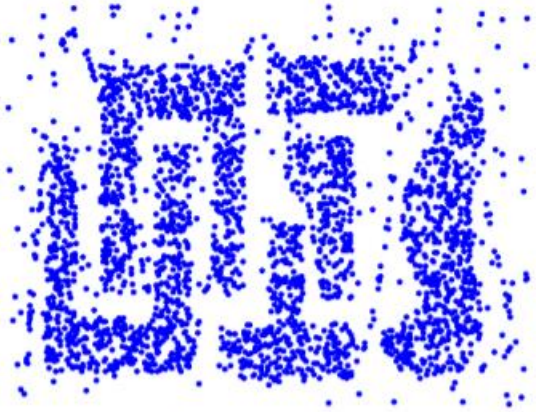
Basados en jerarquía

- Pueden ser aglomerantes o divisivos.
- El conjunto de datos está representado por un dendrograma, donde los datos individuales se presentan mediante nodos hoja.

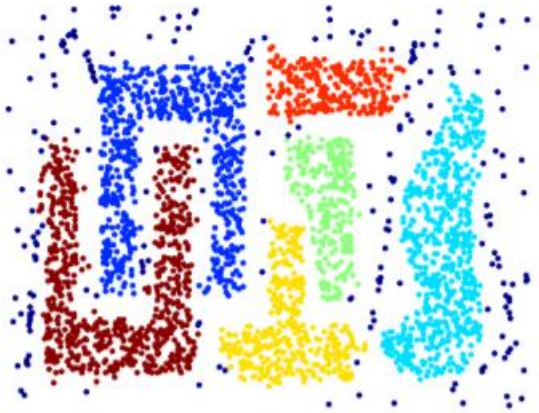


Basados en densidad

- En este algoritmo, los elementos se separan en función de sus regiones de densidad, conectividad y frontera.
- Un clúster se define como un grupo de elementos considerados como vecinos dependiendo de su densidad como grupo, y crece en cualquier dirección hacia donde conduce la densidad.



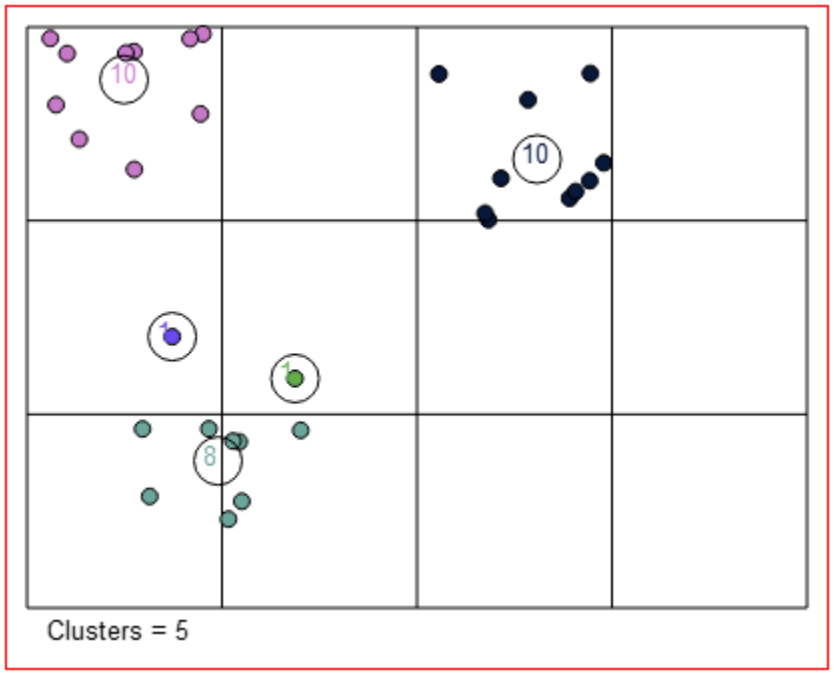
Original Points



Clusters

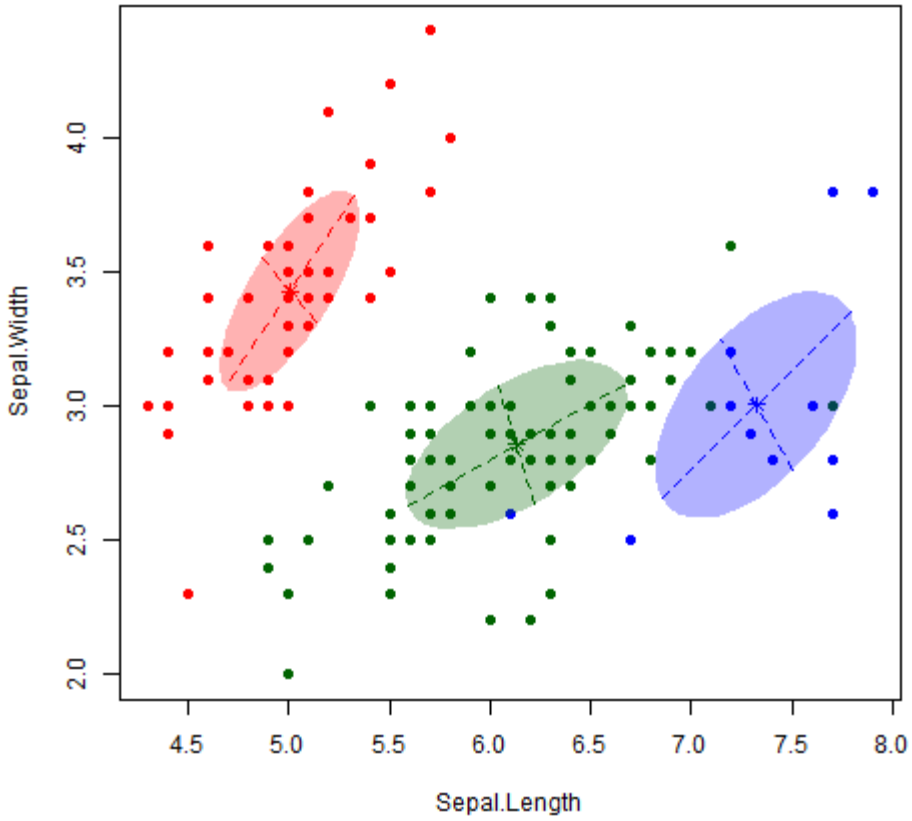
Basados en cuadriculas

- El enfoque de agrupamiento basado en cuadrículas difiere de los algoritmos de agrupamiento convencionales en que no se ocupa de los puntos de datos sino del espacio de valor que rodea los puntos de datos.



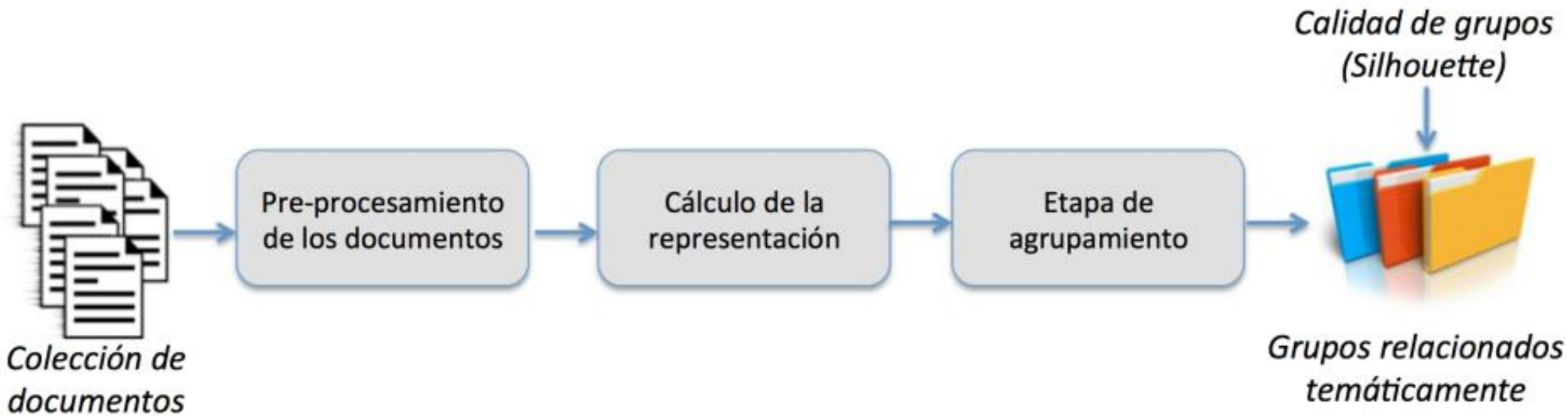
Basados en modelos

- Intentar optimizar el ajuste entre los dados datos y algún modelo matemático



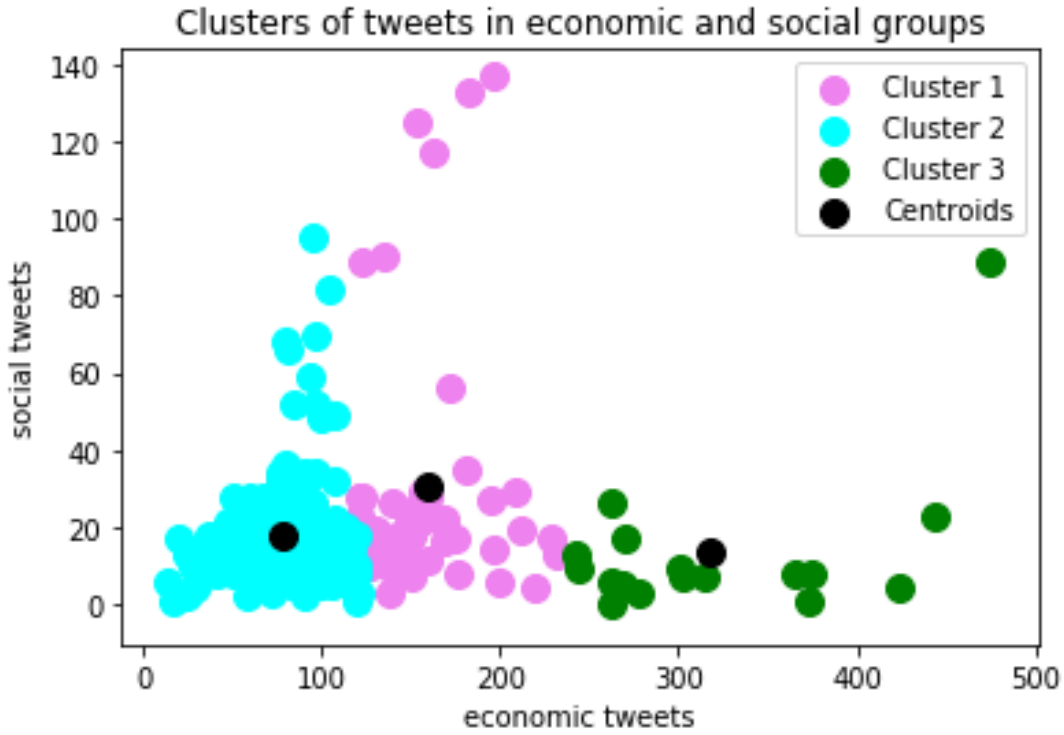
Agrupación de textos

- Agrupamiento de texto puede describirse intuitivamente como hallazgo, dado un conjunto de vectores de datos en un espacio multidimensional.



Agrupación de textos

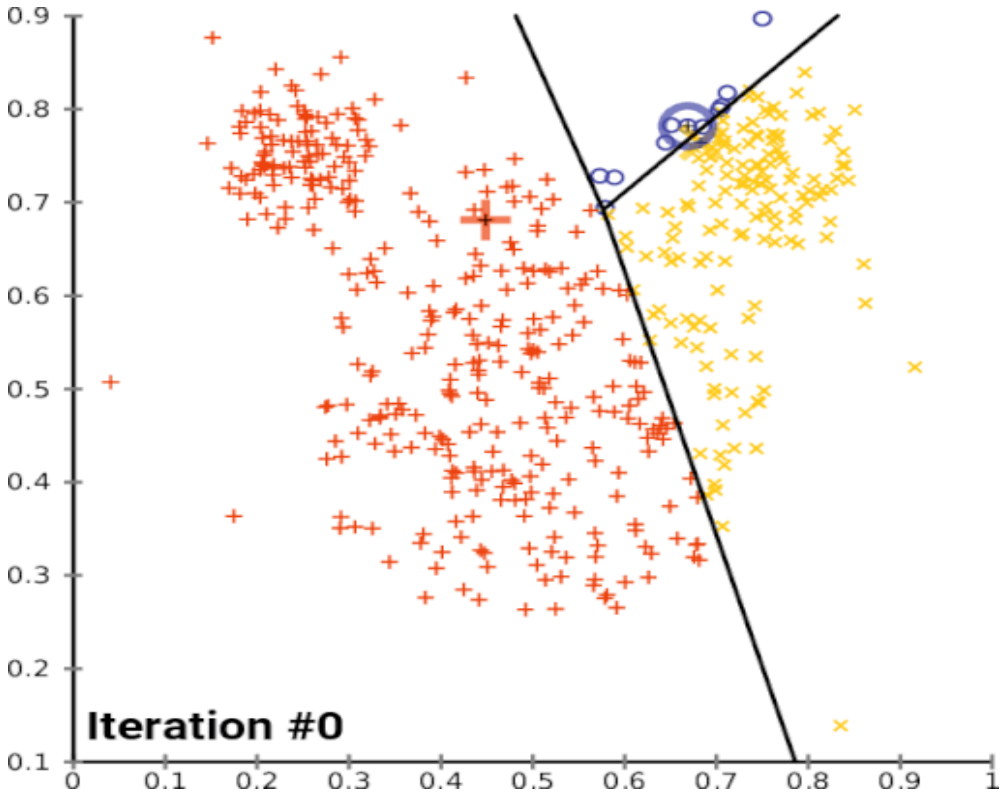
- Descubre automáticamente la estructura implícita en una colección de documentos, identificando los temas más frecuentes dentro de la colección y distribuyendo los documentos en varios grupos (clusters).



Método de las K-Medias (K-Means)

- Algoritmo de agrupamiento k-Means fue propuesto por J. Hartigan y M. A. Wong en 1979.
- Dado un conjunto de n objetos distintos, el algoritmo de agrupación en clústeres k-Means divide los objetos en k número de clústeres, de modo que la similitud entre objetos es alta pero la similitud entre clústeres es baja.

1. La idea principal es definir K centroides de manera aleatoria.
2. Tomar cada punto del conjunto de datos y situarlo en la clase de su centroide más cercano.
3. Recalcular el centroide de cada grupo y volver a distribuir todos los objetos según el centroide más cercano.
4. El proceso se repite hasta ya no haber cambios en los grupos.



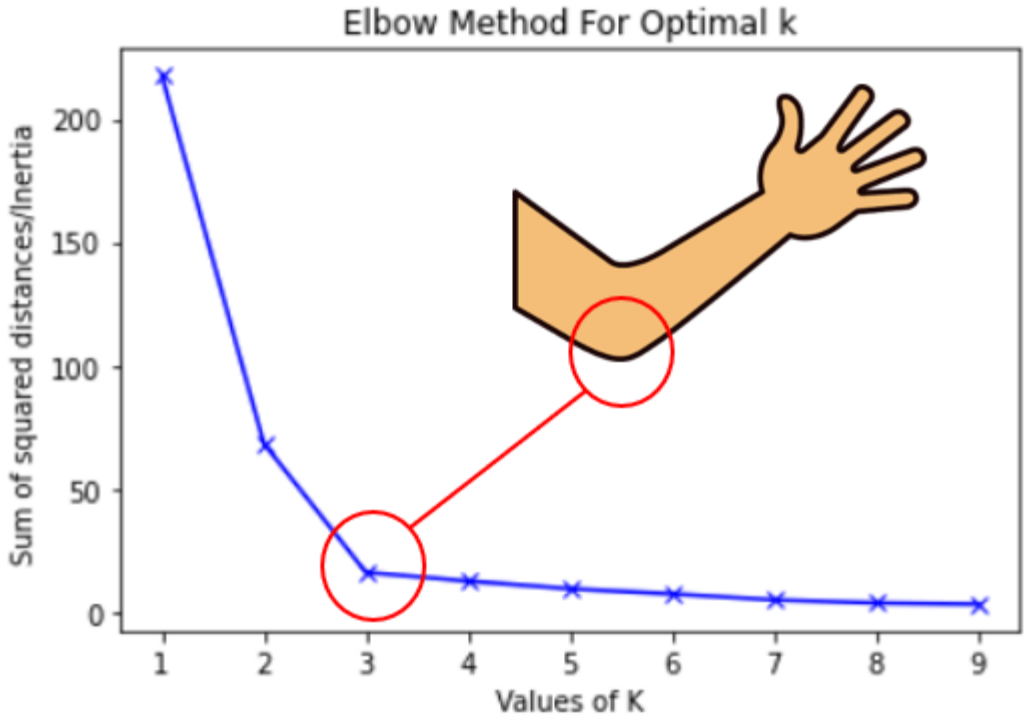
Número óptimo de clústeres

Las técnicas más populares para seleccionar el número óptimo de clústeres:

- Estadística de brechas
- Método del codo
- Coeficiente de silueta
- Índice Calinski-Harabasz
- Índice Davies-Bouldin
- Dendrograma
- Criterio de información bayesiana (BIC)

Método del codo

- La suma de los cuadrados en cada número de clústeres se calcula y se gráfica.
- En la gráfica se busca un cambio de pendiente, de empinada a poco profunda (un codo) para determinar el número óptimo de clústeres.



¡Gracias!

Autor: Erick Quintana Martínez