Documentación para el Código de Clustering

Departamento de Innovación

Alvaro Infante

Breve explicación del código:

Este código es un análisis de clustering que se aplica a un conjunto de datos proporcionado por el usuario. Comienza importando las bibliotecas necesarias y definiendo una serie de funciones que se utilizan en el análisis. En primer lugar, se cargan los datos, se rellenan los valores faltantes y se normalizan los datos numéricos. Luego se determina el número óptimo de clusters utilizando el método del codo y se realiza el análisis de clustering KMeans. Finalmente, se identifican los clientes que se comportan de manera extraña en función de su distancia al centroide de su cluster.

Diccionario de Variables:

- `**df**`: DataFrame de los datos cargados y procesados.
- `**n_clusters**`: Número de clusters determinados por el método del codo.
- `**kmeans**`: Modelo de KMeans entrenado en los datos.
- `**wcss**`: Lista de la suma de cuadrados intra-cluster para diferentes números de clusters.
- `**new_client**`: DataFrame del nuevo cliente ingresado por el usuario.
- `**new_client_cluster**`: Cluster al que pertenece el nuevo cliente.
- `**cluster_sizes**`: Tamaños de cada cluster.
- `**weird_clients**`: Lista de clientes que se comportan de manera extraña.

Resumen de la Matemática:

Determinación del número de clusters:

El código determina el número óptimo de clusters utilizando el método del codo. Esto implica entrenar el modelo KMeans en los datos para un rango de números de clusters (en este caso, de 1 a 10) y calcular la suma de cuadrados intra-cluster (WCSS) para cada número de clusters. El WCSS para K clusters se calcula como:

$$WCSS = \Sigma\Sigma ||x_i - c_j||^2$$

donde:

- x_i son los puntos de datos en el cluster
- c_j es el centroide del cluster
- $||x_i - c_j||^2$ es el cuadrado de la distancia euclidiana entre el punto de datos y el centroide del cluster

El número óptimo de clusters es aquel para el cual el WCSS comienza a disminuir más lentamente, que se puede visualizar como el "codo" en un gráfico de WCSS frente al número de clusters.

Creación de los clusters:

Una vez determinado el número óptimo de clusters, el código crea los clusters utilizando el análisis de clustering KMeans. KMeans es un algoritmo de clustering que divide los datos en K clusters distintos, donde K es un número especificado por el usuario (en este caso, determinado por el método del codo). Cada cluster se define por su centroide, que es el promedio de todos los puntos en ese cluster. La posición del centroide c_j se calcula como:

c_j = (1 / |S_j|) Σ x_i para todo x_i en S_j

donde S_j es el conjunto de puntos de datos en el cluster.

Identificación de clientes que se comportan de manera extraña:

El código identifica los clientes que se comportan de manera extraña en función de su distancia al centroide de su cluster. Primero, calcula la distancia euclidiana de cada punto a su centroide. La distancia euclidiana entre un punto x_i y un centroide c_j se calcula como:

d(x_i, c_j) = sqrt(Σ (x_ik - c_jk)^2)

donde x_ik y c_jk son las coordenadas del punto y del centroide respectivamente.

Luego, identifica como "extraños" a aquellos puntos cuya distancia al centroide es mayor que la media más la desviación estándar de todas las distancias. Este criterio asume que las distancias siguen una distribución normal y que los puntos que están a más de una desviación estándar de la media son inusualmente lejos del centroide.