Jennifer Baltazar Rico

Reporte — Algoritmo de Agrupación k-means

1. Justificación del algoritmo

K-Means es un algoritmo de agrupamiento no supervisado que busca dividir un conjunto de datos en grupos (clusters) de manera que los elementos dentro de cada grupo sean lo más similares posible y los grupos entre sí sean lo más distintos posible. La similitud se mide generalmente a través de la distancia euclidiana entre los puntos.

Se eligió K-Means para este proyecto por varias razones:

- **Simplicidad y eficiencia:** Es rápido y fácil de implementar, especialmente en datasets numéricas de tamaño moderado.
- **Interpretabilidad:** Los clusters generados se pueden interpretar fácilmente y comparar con variables originales.
- **Flexibilidad:** Permite explorar distintos números de clusters mediante técnicas como el método del codo o Silhouette score.

Este algoritmo es apropiado para el conjunto de datos `samsung.csv`, ya que nos interesa identificar patrones o segmentaciones en los datos sin necesidad de tener etiquetas previas.

2. Descripción del diseño del modelo

El diseño del análisis se desarrolló siguiendo los pasos:

1. **Carga de datos y revisión inicial**
    - Se cargó el CSV `samsung.csv` y se inspeccionaron sus dimensiones y tipos de variables.
    - Se filtraron únicamente las columnas numéricas, ya que K-Means requiere datos cuantitativos para calcular distancias.
2. **Preprocesamiento y escalado**
    - Se aplicó **StandardScaler** para normalizar las variables. Esto evita que las variables con rangos mayores dominen la medida de distancia.
    - Se revisaron valores nulos y se completaron con estrategias simples si fue necesario.

3. **Determinación del número de clusters óptimo**
    - Se aplicó el **método del codo**, calculando la inercia para kkk desde 2 hasta 10.
    - Se seleccionó k=3k=3k=3 en este ejemplo, ya que se observó un punto de inflexión que indicaba que más clusters no reducían significativamente la inercia.
4. **Entrenamiento del modelo K-Means**
    - Se entrenó K-Means con `n_clusters=3` y `n_init=10` para asegurar estabilidad en los resultados.
    - Cada observación se asignó a un cluster, generando un vector de etiquetas.

5. **Reducción de dimensionalidad para visualización**
    - Se aplicó **PCA (Análisis de Componentes Principales)** para reducir los datos a 2 dimensiones, facilitando la visualización de los clusters en un plano.
    - Se graficaron los clusters coloreados por su etiqueta y los centroides del modelo.
6. **Guardado del modelo**
    - Se guardó el modelo K-Means entrenado con **joblib** para uso futuro.

3. Evaluación y métricas

Aunque K-Means es un algoritmo no supervisado, se pueden evaluar los clusters con métricas inernas como:

- **Inercia:** Suma de las distancias cuadradas de cada punto a su centroide. Valores menores indican clusters más compactos.
- **Silhouette Score (opcional):** Evalúa qué tan cerca está cada punto de su propio cluster comparado con otros clusters. Valores cercanos a 1 indican clusters bien definidos.

En nuestro ejemplo:

- La gráfica del método del codo mostró que k=3k=3k=3 proporciona un buen balance entre simplicidad y reducción de inercia.
- La proyección en 2D mediante PCA muestra que los clusters son diferenciables, aunque pueden solaparse parcialmente, lo cual es normal en datos complejos.

4. Gráfica personalizada e interpretación de resultados

La visualización PCA de los clusters muestra:

- Cada punto representa una observación del dataset `samsung.csv`.
- Los colores indican el cluster asignado por K-Means.
- Los centroides (marcados con “X” negra) representan el punto medio de cada cluster.

Interpretación:

- Cada cluster agrupa observaciones con características similares según las variables numéricas.
- Esto permite segmentar el dataset en grupos que podrían representar diferentes perfiles de mediciones, usuarios o productos, dependiendo de la naturaleza de los datos.

6. Conclusiones y posibles aplicaciones

- K-Means permitió identificar patrones ocultos en los datos sin necesidad de etiquetas.

Jennifer Baltazar Rico

- Los clusters obtenidos pueden servir para segmentación de usuarios, agrupación de productos o análisis de comportamiento en sensores, dependiendo del contexto del dataset.
- La combinación de K-Means con PCA facilita la interpretación visual de los resultados.
- Este análisis puede ser la base para análisis más avanzados, como clustering jerárquico o modelado predictivo posterior usando las etiquetas de cluster como features.