SEMANA 11 (junio 10, 12 ,14)

AGRUPAMIENTO DE DATOS (CLUSTERING)

El **agrupamiento de datos o clustering** es una técnica que se utiliza para organizar elementos en grupos, de manera que los elementos dentro de un mismo grupo sean más similares entre sí que con los elementos de otros grupos. Es como organizar un conjunto de objetos en varias cajas, donde cada caja contiene objetos que se parecen mucho entre sí.

Explicación Sencilla:

Imagina que tienes una caja llena de diferentes tipos de dulces: caramelos, chocolates, gomitas y piruletas. El objetivo del clustering es agrupar estos dulces de manera que todos los caramelos estén juntos, todos los chocolates estén juntos, todas las gomitas estén juntas y todas las piruletas estén juntas. Así, cada grupo o "cluster" contiene dulces que son similares.

Pasos Básicos del Clustering:

1. **Recolección de Datos:**
    - Primero, obtienes información sobre los elementos que quieres agrupar. En nuestro ejemplo, esto sería la forma, el color y el sabor de los dulces.
2. **Elección de Características:**
    - Decides qué características usar para comparar los elementos. En los dulces, podrías usar la forma (redonda, cuadrada), el color (rojo, verde, amarillo) y el sabor (dulce, ácido).
3. **Medición de Similitud:**
    - Utilizas una métrica para medir qué tan similares son dos elementos. Por ejemplo, podrías decidir que los dulces del mismo color son más similares entre sí.
4. **Agrupamiento:**

- Usas un algoritmo de clustering para agrupar los elementos. Este algoritmo busca agrupar los elementos de tal forma que los más similares queden juntos en el mismo grupo.

Ejemplo Sencillo:

Imagina que tienes un grupo de estudiantes y quieres agruparlos según sus hábitos de estudio y calificaciones. Podrías usar clustering para agrupar a los estudiantes en:

- **Grupo 1:** Estudiantes que estudian mucho y tienen calificaciones altas.
- **Grupo 2:** Estudiantes que estudian poco y tienen calificaciones bajas.
- **Grupo 3:** Estudiantes que estudian moderadamente y tienen calificaciones promedio.

Aplicaciones en la Vida Real:

- **Marketing:** Agrupar clientes con hábitos de compra similares para ofrecerles promociones personalizadas.
- **Salud:** Agrupar pacientes con síntomas similares para un diagnóstico más preciso.
- **Redes Sociales:** Agrupar usuarios con intereses similares para recomendarles contenido relevante.

Resumen:

El clustering es una forma de encontrar patrones y relaciones en datos sin necesidad de saber de antemano a qué grupo pertenece cada elemento. Es una herramienta poderosa para organizar y entender grandes cantidades de información, facilitando la toma de decisiones y la personalización de servicios.

Vamos a generar un ejemplo práctico de agrupación de datos utilizando un conjunto de datos ficticio de clientes de una tienda en línea. Este conjunto de datos incluirá las siguientes características para cada cliente:

5. Edad
6. Ingresos anuales (en miles de dólares)
7. Gasto anual en la tienda (en miles de dólares)

Queremos agrupar a estos clientes en tres clusters para identificar distintos segmentos de mercado. Utilizaremos el algoritmo k-means para este propósito.

Paso 1: Crear el Conjunto de Datos Ficticio

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

# Crear un conjunto de datos ficticio
data = {
    'Edad': [25, 34, 45, 23, 35, 52, 46, 52, 27, 40, 22, 48, 55, 29, 31, 49, 36, 28, 44, 50],
    'Ingresos Anuales (k$)': [15, 20, 35, 12, 25, 45, 40, 55, 18, 30, 10, 38, 60, 22, 21, 50, 27, 17, 42, 53],
    'Gasto Anual (k$)': [2, 5, 20, 3, 10, 30, 25, 40, 7, 15, 2, 22, 50, 6, 8, 35, 11, 4, 27, 45]
}

# Convertir a DataFrame
df = pd.DataFrame(data)

# Visualizar el conjunto de datos
print(df)
```

Paso 2: Aplicar el Algoritmo de Clustering (k-means)

```
# Seleccionar las características para el clustering
X = df[['Edad', 'Ingresos Anuales (k$)', 'Gasto Anual (k$)']]

# Aplicar k-means con k=3
kmeans = KMeans(n_clusters=3, random_state=0).fit(X)

# Agregar las etiquetas de los clusters al DataFrame
df['Cluster'] = kmeans.labels_

# Visualizar los clusters asignados
print(df)
```

Paso 3: Visualización de los Resultados

```
# Visualizar los clusters en un gráfico 3D
fig = plt.figure(figsize=(10, 7))
ax = fig.add_subplot(111, projection='3d')

# Asignar colores a los clusters
colors = ['r', 'g', 'b']

for i in range(3):
    cluster_data = df[df['Cluster'] == i]
    ax.scatter(cluster_data['Edad'], cluster_data['Ingresos Anuales (k$)'], cluster_data['Gasto Anual (k$)'], c=colors[i], label=f'Cluster {i}')
```

```
ax.set_xlabel('Edad')
ax.set_ylabel('Ingresos Anuales (k$)')
ax.set_zlabel('Gasto Anual (k$)')
ax.legend()
plt.show()
```

Interpretación de los Clusters

Cluster 0:

- **Descripción:** Clientes más jóvenes con ingresos y gastos anuales relativamente bajos.
- **Estrategia de Marketing:** Ofrecer descuentos y promociones para atraer más compras.

Cluster 1:

- **Descripción:** Clientes de edad media con ingresos y gastos anuales moderados.
- **Estrategia de Marketing:** Ofrecer productos de gama media y promociones basadas en fidelidad.

Cluster 2:

- **Descripción:** Clientes mayores con ingresos y gastos anuales altos.
- **Estrategia de Marketing:** Enfocar en productos premium y servicios personalizados.

Este ejemplo demuestra cómo se puede utilizar clustering para segmentar un conjunto de datos ficticio de clientes y adaptar estrategias de marketing basadas en los resultados del análisis.

DEFINICIONES FORMALES

El **agrupamiento de datos (clustering)** es una técnica de análisis de datos que busca dividir un conjunto de datos en grupos o "clusters", de manera que los datos dentro de cada grupo sean más similares entre sí que con los datos de otros grupos. Este proceso es fundamental en el aprendizaje no supervisado, donde no se tiene información previa sobre las categorías de los datos. A continuación, se detallan los aspectos clave del clustering:

Objetivos del Clustering

- **Descubrimiento de estructuras ocultas:** Identificar patrones o estructuras en los datos que no son inmediatamente obvios.
- **Reducción de la dimensionalidad:** Simplificar el conjunto de datos agrupándolos en categorías más manejables.
- **Segmentación:** Dividir los datos en segmentos significativos para un análisis más detallado.

Tipos de Clustering

8. Clustering Particional:

- **Descripción:** Divide los datos en k clusters predefinidos.
- **Ejemplo: k-means, k-medoids.**
- **Aplicación: Agrupación de clientes en segmentos de mercado.**

9. Clustering Jerárquico:

- **Descripción:** Construye una jerarquía de clusters mediante un enfoque ascendente (agglomerative) o descendente (divisive).
- **Ejemplo:** Algoritmo de enlace completo (complete linkage), enlace simple (single linkage).
- **Aplicación:** Análisis de taxonomías biológicas.

10. Clustering Basado en Densidad:

- **Descripción:** Identifica clusters basados en la densidad de puntos de datos en el espacio.
- **Ejemplo:** DBSCAN (Density-Based Spatial Clustering of Applications with Noise).
- **Aplicación:** Detección de agrupaciones en datos espaciales y eliminación de ruido.

11. Clustering Basado en Modelos:

- **Descripción:** Asume que los datos son generados por una mezcla de distribuciones probabilísticas y trata de encontrar la distribución que mejor se ajuste.
- **Ejemplo:** Gaussian Mixture Models (GMM).
- **Aplicación:** Clasificación de documentos.

Algoritmos de Clustering

12. k-means:

- **Proceso:** Asigna cada punto de datos al cluster cuyo centroide es el más cercano, recalculando los centroides repetidamente hasta que las asignaciones no cambien.
- **Ventajas:** Simple y fácil de implementar.
- **Desventajas:** Requiere definir el número de clusters (k) previamente y puede converger a óptimos locales.

13. DBSCAN:

- **Proceso:** Agrupa puntos que están densamente conectados y marca como ruido los puntos que están aislados.
- **Ventajas:** No necesita especificar el número de clusters y puede detectar formas arbitrarias.
- **Desventajas:** Sensible a los parámetros de densidad y radio.

14. Agglomerative Hierarchical Clustering:

- **Proceso:** Comienza con cada punto como un cluster individual y fusiona los clusters más cercanos iterativamente.
- **Ventajas:** No necesita especificar el número de clusters.
- **Desventajas:** Alta complejidad computacional para grandes conjuntos de datos.

Aplicaciones del Clustering

15. Segmentación de clientes:

- **Descripción:** Agrupar clientes en segmentos basados en comportamientos o características similares.
- **Ejemplo:** Marketing personalizado y estrategias de ventas.

16. Detección de anomalías:

- **Descripción:** Identificar datos que no pertenecen a ningún cluster o que pertenecen a clusters muy pequeños.
- **Ejemplo:** Detección de fraudes en transacciones financieras.

17. Análisis de imágenes:

- **Descripción:** Agrupar píxeles o segmentos de imágenes para identificar objetos o regiones.
- **Ejemplo:** Segmentación de imágenes médicas para detección de tumores.

18. Bioinformática:

- **Descripción:** Agrupar secuencias de genes o proteínas para identificar funciones similares.
- **Ejemplo:** Clasificación de especies o identificación de enfermedades genéticas.

Desafíos del Clustering

- **Determinación del número de clusters:** Algunos métodos requieren especificar el número de clusters de antemano.
- **Escalabilidad:** Manejar grandes volúmenes de datos puede ser computacionalmente costoso.
- **Elección de la métrica de distancia:** La efectividad del clustering puede depender de la elección de la métrica de distancia adecuada.
- **Interpretación de resultados:** Los clusters obtenidos deben tener sentido en el contexto del problema específico.

El clustering es una herramienta poderosa y versátil en la minería de datos y el aprendizaje automático, permitiendo el descubrimiento de estructuras subyacentes y patrones ocultos en los datos sin necesidad de etiquetas previas.