Aprendizaje Automático I

Práctica Curso 2023/24

Práctica B

Proyecto sobre aprendizaje no supervisado

Esta práctica consta de una primera parte enfocada en el clustering jerárquico y una segunda parte enfocada en el clustering particional. La entrega consistirá en:

- Un cuaderno denominado 1_jerarquico.ipynb
- Un cuaderno denominado 2_particional.ipynb
- Un cuaderno denominado 3_densidad.ipynb
- Un cuaderno denominado 4_otros.ipynb
- Una memoria en formato PDF: memoria.pdf

Los 5 documentos se incluirán en un archivo comprimido .zip cuyo nombre seguirá el siguiente formato, donde XX será el grupo asignado en Moodle a la pareja de prácticas:

PRÁCTICA_NO_SUPERVISADO_23_24_GRUPO_XX.ZIP

Instrucciones

- **Se deberán utilizar celdas de markdown** para añadir las explicaciones necesarias sobre los pasos seguidos los resultados obtenidos
- El código deberá estar comentado.
- Los cuadernos **se entregarán con todas las salidas generadas y guardadas** para las celdas de código. Es decir, no será necesario ejecutarlo para ver lo que devolvió la ejecución.
- **Se deberá poder ejecutar el cuaderno hasta la última celda, sin errores,** en Google Colab. En caso de desarrollar la práctica en otro entorno, debéis aseguraros de que ejecuta correctamente en Google Colab.
- Se deberá hacer uso de gráficas para explicar los resultados y la inclusión de explicaciones detalladas, así como del trabajo realizado de cara a conseguir los mejores resultados posibles.

Preparación de la memoria

La memoria deberá describir en detalle el trabajo realizado **y no contendrá código.** Deberéis ayudaros del uso de gráficas y explicar en detalle todas las decisiones adoptadas y los resultados obtenidos.

Descripción del dataset: gasto turístico de viajeros en España

- Dataset: tourism_spain_2016_2023.zip

Para esta práctica se proporcionan datos del gasto turístico de viajeros extranjeros en España durante los años 2016 a 2023 (los meses ya disponibles). El dataset contiene una serie de columnas con la siguiente información para cada fila, que corresponden a un viajero concreto:

- mm_aaaa:
  - **Tipo**: string
  - **Definición:** mes y año de finalización del viaje
- A0:
  - **Tipo**: string
  - **Definición:** Encuesta de procedencia del dato
- A0_1:
  - **Tipo:** enumerado
  - **Definición:** Tipo de viajero
  - **Valores**:
    - 2 (turista no residente no en tránsito)
    - 6 (turista no residente en tránsito)
- A1:
  - **Tipo:** enumerado
  - **Definición:** vía de salida del país
  - **Valores**:
    - 1 - carretera
    - 2 - aeropuerto
    - 3 - puerto
    - 4 - tren
- País:
  - **Tipo**: string
  - **Definición:** País de residencia habitual
  - **Valores**:
    - 01: Alemania
    - 02: Bélgica
    - 03: Francia
    - 04: Irlanda
    - 05: Italia
    - 06: Países Bajos
    - 07: Portugal
    - 08: Reino Unido
    - 09: Suiza.
    - 10: Rusia
    - 11: Países Nórdicos (Dinamarca, Finlandia, Noruega, Suecia)
    - 12: Resto de Europa
    - 13: EEUU

        - 14: Resto de América
        - 15: Resto del mundo
- ccaa:
    - **Tipo**: string
    - **Definición:** comunidad autonóma de destino principal del viaje
    - **Valores:**
        - 01: Andalucía
        - 02: Aragón
        - 03: Principado de Asturias
        - 04: Illes Balears
        - 05: Canarias
        - 06: Cantabria
        - 07: Castilla y León
        - 08: Castilla-La Mancha
        - 09: Cataluña
        - 10: Comunitat Valenciana
        - 11: Extremadura
        - 12: Galicia
        - 13: Comunidad de Madrid
        - 14: Región de Murcia
        - 15: Comunidad Foral de Navarra
        - 16: País Vasco
        - 17: La Rioja
        - 18: Ceuta
        - 19: Melilla

- A13:
    - **Tipo**: integer
    - **Definición:** número de pernoctaciones

- Aloja:
    - **Tipo:** enumerado
    - **Definición:** tipo de alojamiento
    - **Valores**:
        - 1: Hoteles y similares
        - 2: Resto de mercado
        - 3: Alojamiento no de mercado
- Motivo:
    - **Tipo:** enumerado
    - **Definición:** motivo principal del viaje
    - **Valores**:
        - 1: Ocio/vacaciones
        - 2: Negocios
        - 3: Resto
- A16:
    - **Tipo:** enumerado
    - **Definición**: Si el viajero usó un paquete turístico

    - **Valores**:
        - 1: Si
        - 6: No
- Gastototal:
    - **Tipo**: Decimal
    - **Definición:** Gasto total del viaje/excursión
- Factoregatur:
    - **Tipo**: Decimal
    - **Definición:** Factor de elevación de Egatur

Para esta práctica, las variables A0, A0_1, A0_7 y factoregatur pueden descartarse.

Enunciado

Se pide resolver las siguientes cuestiones:

0. Descripción del dataset

(Resolver en 1_jerarquico.ipynb)

Tarea 1: Realiza un análisis descriptivo del dataset. Analiza la distribución de los datos por cada una de las columnas, realiza los pasos de pre-procesamiento necesarios, justificando adecuadamente las acciones tomadas. Se deberá hacer uso de gráficas para entender los datos y las decisiones adoptadas.

1. Aplicación de algoritmos de clustering jerárquico

(Resolver en 1_jerarquico.ipynb)

Tarea 2.1: Aplica al menos 2 algoritmos de clustering jerárquico sobre el dataset proporcionado, probando y evaluando los efectos de la distancia utilizada (euclídea, coseno…).

Tarea 2.2: Analiza a determinadas profundidades la distribución de los ejemplos en el dendograma. ¿Es uniforme la distribución independientemente de la profundidad?

Tarea 2.3: ¿Cómo afectan las diferentes métricas de distancia a la estructura del dendrograma?

Tarea 2.4 Utiliza por lo menos dos índices de calidad de clustering y y analiza sus resultados.

Tarea 2.5 ¿Cuál es el número óptimo de clusters? ¿por qué?.

Tarea 2.6: Queremos conocer, con ayuda de métodos de clustering, el perfil de gasto del viajero (columna *perfil_gasto_viajero*) en función de tipo de alojamiento, número de noches, país de residencia y CCAA de destino. Implementa un método que reciba como argumentos estos 4 valores y devuelva el perfil de gasto de viajero más probable. Para ello, el método utilizará un modelo de clustering entrenado en los datos, justificando el modelo utilizado.

2. Aplicación de algoritmos de clustering particional

(Resolver en 2_particional.ipynb)

Tarea 3.1: Realiza el pre-procesamiento necesario para poder aplicar algoritmos de clustering particional.

Tarea 3.2: Establece el número más adecuado de clusters para el dataset proporcionado. Ayúdate de los métodos vistos (al menos 2) en la asignatura, así como de gráficas para justificar la decisión. Compara los resultados que obtienes con cada método.

Tarea 3.3: ¿Cómo varía la calidad del clustering con diferentes valores de 'K'?

Tarea 3.4: Con el número más adecuado de clusters, ayúdate de estadísticas para analizar a los viajeros incluidos en cada cluster.

Tarea 3.5: Compara los resultados obtenidos con K-means y el clustering aglomerativo/jerárquico. Discute las ventajas y desventajas de cada método en diferentes tipos de datos.

3. Aplicación de algoritmos de densidad

(Resolver en 3_densidad.ipynb)

Tarea 4.1: Realiza el pre-procesamiento necesario para poder aplicar algoritmos de densidad.

Tarea 4.2: Establece el radio (eps) y número de puntos mínimo número más adecuado de clusters para el dataset proporcionado.

Tarea 4.3: ¿Cómo varía la calidad del clustering con diferentes valores de 'eps' y de minpoints?

Tarea 4.4 Utiliza por lo menos dos índices de calidad de clustering y analiza sus resultados.

Tarea 4.5 ¿Cuál es el número óptimo de clusters? ¿por qué?

4. OPCIONAL: Aplicación de otros algoritmos

(Resolver en 5_otros.ipynb)

Tarea 5.1: Emplea otros algoritmos como HDBScan y compara con otros algoritmos su rendimiento.

Tarea 5.2: Emplea otros algoritmos como K-modes y compara con otros algoritmos su rendimiento.