Aprendizaje Automático I

Práctica Curso 2024/25

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: accidentes de Tráficos en Madrid

Se proporciona un dataset con información de accidentes de tráfico en Madrid.

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.

A continuación, se plantean una serie de tareas. El objetivo, en todas ellas, **es extraer información relevante sobre las condiciones en las que se producen accidentes en Madrid, tales como el efecto de la lluvia, determinadas localizaciones o determinados tipos de vehículos. Hacer un análisis no dirigido a este objetivo supondrá no alcanzar la nota mínima.**

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, las zonas con mayor índice de siniestralidad para cada tipo de vehículo. Ayúdate de diferentes modelos para hacer un análisis detallado de esta relación.

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.