Práctica 1. Aprendizaje Automático

Fecha de entrega: 8 de marzo de 2020

Esta práctica tiene como objetivo aplicar a distintos conjuntos de datos algunos de los algoritmos de aprendizaje automático disponibles en el entorno Scikit-Learn de Python. Se elaborará una memoria mediante notebooks de jupyter que incluirán los apartados debidamente señalizados, con su código, la salida del mismo cuando corresponda, y las respuestas a las preguntas planteadas.

IMPORTANTE: Usa la opción **random_state** en las funciones que generan las particiones de datos y que implementan los algoritmos de aprendizaje automático para que los resultados del notebook sean reproducibles en todo momento.

Los conjunto de datos

Por cada conjunto de datos que utilices deberás incluir una breve descripción del mismo.

- Nombre del conjunto de datos
- Breve descripción del problema que describe
- Tabla con el nombre y tipo de las variables
- Tabla de estadísticos descriptivos de cada variable

Parte 1: Agrupamiento o clustering

Usa el conjunto de datos de causas de arrestos en los diferentes estados de Estados Unidos en 1973 que puedes descargar del campus virtual. El conjunto contiene las cifras de arrestos por asalto, asesinato y violación por 100.000 residentes. Además, contiene la variable con el porcentaje de población que vive en áreas urbanas, la cual no usaremos para el clustering.

El objetivo es realizar agrupamientos de los estados que presenten un perfil de arrestos similar e identificar cuál ese ese perfil de arrestos y qué estados pertenecen a él.

1) Describe el conjunto de datos tal y como se indica más arriba y extrae algunas conclusiones de las variables, su distribución y su correlación.
2) Considera si debes re-escalar las variables antes y el tipo de escalado que usas. Razona tu elección.
3) Aplica un algoritmo de clustering de los que hemos visto en clase con una parametrización (el valor de *k* en el algoritmo de *k*-medias, o la forma en la que se agrupan clusters en el caso jerárquico).
   Determina el número de clusters que consideras adecuado para el conjunto de datos y justifica tu elección.
4) Da un sentido a cada uno de los clusters que has obtenido en el contexto del problema que representa el conjunto de datos.
   Si obtienes un número mayor de 4 clusters, comenta solamente los dos los dos más numerosos y los dos menos numerosos.
   Para analizar los clusters:
   - Usa estadísticos descriptivos (número de individuos, media, desviación típica, mediana, cuartiles) para describir los clusters.

- Usa una matriz de gráficos de dispersión que pinte los clusters usando un color diferente para ver la separación de los clusters en función de cada par de variables de entrada. ¿Qué clusters se separan mejor y en función de qué variables? ¿y cuáles se cofunden más?
    - Para ello, usa la función **`seaborn.pair_plot`** de la librería de representación gráfica seaborn, como puedes ver en este ejemplo

Te recomendamos que uses las variables en su escala original y no en la transformada, ya que en la escala original puedes relacionar los valores de las variables con lo que representan en la vida real.

Documenta todo el proceso en un notebook de jupyter con comentarios, texto explicando las soluciones y toda la información que consideres necesaria.