Actividad 3

La actividad evaluable asociada a este tema y al bloque 3 de la asignatura va a constituir en un ejercicio de clasificación y otro de regresión usando los datos de los alquileres vacacionales ofrecidos en la plataforma AirBnB que ya se usaron en la actividad del tema 3, ver sección 3.7. Para ello habrá que crear un notebook con celdas de tipo code y de tipo markdwon, en las de este último tipo se incluirán comentarios de lo que se esté haciendo en las de tipo code o sobre los resultados.

Para el estudio estadístico y limpieza de datos, ver subsección 3.7.2, se pueden tomar lo que ya se hiciera en la actividad del tema 3.

La fecha de entrega de la actividad aparece en la plataforma.

Clasificación

La tarea de clasificación consistirá en clasificar los datos según el tipo de alojamiento, definido en el campo room_type, a partir del resto de características. Es decir, en room_type estarán codificadas las clases y en el resto de campos los atributos.

Para el estudio estadístico y limpieza de datos, ver subsección 3.7.2, se puede tomar lo que ya se hiciera en la actividad del tema 3.

La clasificación se realizará usando LinearSVC y SVC con sus parámetros por defecto, para el entrenamiento y validación se seguirá lo descrito en los dos primeros puntos de la subsección 3.7.3, teniendo en cuenta que ahora son dos modelos. En cuanto al punto 3, afinación de hiperparámetros, se realizará para SVC, usando el kernel RBF, buscando los mejores valores para γ y C usando GridSearchCV. En este caso no se van a dar sugerencias del rango de valores a probar, por lo que hay que realizar búsquedas de manera que cada una refine la anterior. Proponemos seguir las indicaciones de la sección 3.2 de A Practical Guide to Support Vector Classification, bastará con dos búsquedas. Para crear las listas de parámetros se puede hacer a mano o usando numpy.logspace. Las listas están equiespaciadas en escala logarítmica, si se realizara una tercera búsqueda se podría usar una escala lineal. La idea es escala logarítmica para rangos elevados y lineal para rangos no elevados.

Regresión

La tarea de regresión consistirá en considerar el precio por noche como variable dependiente, y el resto de campos como variables independientes. Es decir, se tratará de predecir los valores del campo price a partir de todos los demás. Convendrá repasar el capítulo 2 de Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd Edition.

Puntos a tener en cuenta:

Convendrá estudiar estadísticamente las variables price, minimun_nights y calculated_host_listings_count para ver si hay outliers que posiblemente sea conveniente eliminar.

Probar con LinearSVR y SVR con parámetros por defecto, realizando el entrenamiento para cada caso usando todas las variables independientes y todas las independientes excepto neighbourhood, ya que esta variable es categórica y tiene muchos valores, lo que podría afectar a los resultados. Si los resultados aconsejan eliminar neighbourhood el siguiente punto se realizará sin tener en cuenta dicha variable.

Para el algoritmo que mejor resultado haya dado en el punto anterior realizar una afinación de hiperparámetros como se comentó en el apartado de clasificación:

- C si fuera LinearSVR, realizando 3 búsquedas.
- γ y C si fuera SVR, realizando 2 búsquedas.

Valoración

Se valorará la creación del código que realice los requisitos enumerados, la presencia de comen- tarios de lo que hace el código y sus resultados y las comparaciones entre los resultados obtenidos por los distintos algoritmos, así como algún gráfico que muestre los resultados. Para terminar incluir un apartado de conclusiones. Tanto los comentarios como las conclusiones no es necesario que sea extensos, sino que describan de forma concisa.