Actividad 1

Actividades evaluables

La actividad evaluable asociada al boque 1 de la asignatura va a constituir en un ejercicio de clasificación empleando los tres métodos descritos en el tema 3 y la aplicación de varias de las técnicas estudiadas en el tema 2. Para ello habrá que crear un notebook con celdas de tipo code y de tipo markdown, en las de este último tipo se incluirán comentarios de lo que se esté haciendo en las de tipo code o sobre los resultados.

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

Datos

Los datos provienen de la web InsideAirBnB, dedicada al estudio de los alquileres vacacionales ofrecidos en la plataforma AirBnB. Es necesario descargar el fichero airbnb.csv de la carpeta datos de los documentos de Aprendizaje Automático I. Este fichero es una versión editada, a fin de facilitar la tarea, del listado original de información sobre las ofertas existentes, para la ciudad de Madrid, en abril de 2017. Contiene 13321 registros con 11 campos cada uno, correspondientes a diferentes características de cada oferta de alojamiento.

El fichero se puede descargar a mano y colocarlo en un directorio desde donde lo cargue el notebook.

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.

3.7.2. Estudio estadístico y limpieza de datos

Se realizará un breve estudio estadístico de los datos numéricos y de la variable room_type se contarán los valores de cada clase. Si las clases no estuvieran balanceadas habrá que usar los mecanismos que puedan tener los algoritmos de clasificación para tratar con este caso. Para Naive Bayes ya se habla de esto en la actividades autoevaluables, para los otros métodos consultar en la documentación de sus APIs los parámetros de los constructores para ver si hay mecanismos para balancear.

En cuanto a la limpieza de datos estudiar si hay datos faltantes, transformar datos categóricos y escalar datos numéricos haciendo uso de pipelines cuando sea posible.

3.7.3. Entrenamiento y validación

1. Dividir los datos en conjunto de entrenamiento y test de manera que el conjunto de test sea un 20 % del total.
2. Evaluar los modelos midiendo la exactitud (accuracy) usando validación cruzada para los tres métodos estudiados utilizando los parámetros por defecto de los tres métodos, salvo lo comentado para el balance, para cross_val_score usar cv=10. Comparar los resultados de los tres modelos.
3. Realizar la afinación de hiperparámetros para KNN y Árboles de decisión empleando Grid- SearchCV, Naive Bayes no tiene hiperparámetros que afinar.

a) Para KNN buscando el valor óptimo de K.

b) Para Árboles de decisión: variando max_leaf_nodes entre 2 y 50 ambos inclusive1, min_samples_split entre 2 y 6 ambos inclusive y max_depth entre 1 y 20 ambos inclusive[1]. Ver la documentación de la API.

3.7.4. Valoración

Se valorará la creación del código que realice los requisitos enumerados, la presencia de comentarios 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.

[1] range(a, b) incluye a pero llega hasta b-1