MBD – Estadística – Práctica Ia (ML)

Descripción

Los sistemas de alquiler de bicicletas existentes en las grandes ciudades disponen de un sistema automatizado de recogida y retorno del vehículo a través de una red de estaciones distribuidas por toda la metrópolis. Con el uso de estos sistemas, las personas pueden alquilar una bici en una ubicación y retornarla en otra distinta en función de sus necesidades. Los datos generados por estos sistemas son atractivos para los investigadores debido a variables como la duración del viaje, los puntos de salida y destino y el tiempo de trayecto. Por tanto, los sistemas de intercambio de bicicletas funcionan como una red de sensores que son útiles para los estudios de movilidad. Con el objetivo de mejorar la gestión, una de estas empresas necesita anticiparse a la demanda que habrá en un determinado rango de tiempo en función de factores como la franja horaria, el tipo de día (laborable o festivo), la climatología, etc..

El objetivo de esta práctica es predecir la demanda en una serie de franjas horarias concretas, usando el conjunto de datos histórico como base para construir un modelo lineal.

Datos

Se entregarán dos conjuntos de datos que contendrán el número de bicicletas alquiladas en distintas franjas de una hora:

1. Datos de entrenamiento. Contendrán la variable respuesta (número de bicicletas alquiladas en esa franja)
2. Datos test. No contendrán la variable respuesta y ésta deberá predecirse basándose en los datos históricos del conjunto de entrenamiento.

Las variables presentes en los 2 conjuntos de datos son:
id: identificador de la franja horaria (no guarda relación con el orden temporal)
year: año (2011 o 2012)
hour: hora del día (0 a 23)
season: 1 = invierno, 2 =primavera, 3 = verano, 4 = otoño
holiday: si el día era festivo
workingday: si el día era laborable (ni festivo ni fin de semana)
weather: cuatro categorías (1 a 4) que van de mejor a peor tiempo
temp: temperatura en grados celsius
atemp: sensación de temperatura en grados celsius
humidity: humedad relativa
windspeed: velocidad del viento (km/h)
count (sólo en el conjunto de entrenamiento): número total de alquileres en esa franja

Evaluación

Se basará en el error cuadrático medio logarítmico entre la predicción y el valor real:

$$\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \left((\log(p_i + 1) - \log(a_i + 1) \right)^2}$$

Donde:
n es el número de observaciones (horas) en el conjunto de datos test
p_i es el número de bicicletas alquiladas predicho por el modelo
a_i es el contaje real
log(x) es el logaritmo natural

Se obtendrá una puntuación más alta cuanto menor sea este estadístico. Adicionalmente se valorará:

1. La parsimonia del modelo
2. El cumplimiento de las premisas
3. La presentación de los resultados

Entrega

La fecha límite para realizar la entrega será el día 22/11/2020 a través del campus.

La documentación a entregar será:

1. Una descripción del trabajo realizado. El documento debe incluir (como mínimo) la expresión del modelo final, el proceso de selección de variables y el análisis de las premisas. Además se debe especificar si se creó alguna variable derivada (calculada a partir de otras) o si se realizó alguna transformación (logaritmo, box-cox…) sobre las ya existentes. Extensión: 3 o 4 páginas. Formato: *.docx*
2. Fichero de texto con el contaje predicho en dos columnas separadas por una coma: la primera debe contener el identificador de la franja horaria y la segunda el número de alquileres predicho. Formato: *.txt.* Es muy importante que este fichero se llame exactamente "p1a.txt"
3. El código utilizado para realizar esta parte con comentarios explicativos (#). Formato: *.R*

Los 3 ficheros se subirán al campus en una carpeta comprimida cuyo nombre contenga el nombre y primer apellido de cada integrante del grupo. Por ejemplo "Jordi Gracia y Maria Andres.zip"