Aprendizaje Automático I

Ejercicio propuesto: City Bike NYC

DSLab

octubre, 2023

Introducción

El sistema de uso compartido de bicicletas en la ciudad de Nueva York (EE.UU.) publica diariamente gran cantidad de datos de actividad sobre su uso.

Estos datos han dado lugar, como no, a algunos análisis sobre la evolución de este servicio y posibles factores que puedan influenciar su uso. En esta práctica vamos a proponer el análisis de datos resumen diarios sobre la utilización de este servicio entre julio de 2013 y noviembre de 2015.

La filosofía de esta práctica es fomentar que consultéis la documentación en línea tanto de Pandas como de Seaborn, para así familiarizaros más con los diferentes métodos disponibles para resolver los ejercicios propuestos. En cada pregunta, se ofrecen consejos sobre partes relevantes de esta documentación relacionadas con las tareas que se piden.

Descripción de variables

El archivo de datos que vamos a utilizar puede obtenerse de esta url. Se trata de un fichero en formato CSV, que se ha creado mezclando datos del City Bike System con datos de la National Oceanic and Atmospheric Administration (NOAA), sobre NYC. El fichero cuenta con las siguientes columnas:

- `date`: fecha del dato, en formato YYYY-MM-DD.
- `trips`: entero positivo, número total de viajes acumulados ese día.

- `precipitation`: entero positivo, cantidad de lluvia total registrada ese día (pulgadas).
- `snow_depth`: entero positivo, altura de nieve (pulgadas).
- `snowfall`: entero positivo, registro de precipitación en forma de nieve (pulgadas).
- `max_temperature`: entero, temperatura máxima registrada (ºF).
- `min_temperature`: entero, temperatura mínima registrada (ºF).
- `average_wind_speed`: entero, velocidad promedio del viento (MPH, millas por hora).
- `dow`: [0, 7]; código de día de la semana, 0 corresponde al domingo.
- `year`: Año del registro.
- `month`: Mes del registro.
- `holiday`: Valor lógico, indica si esa fecha es festivo (TRUE) o no (FALSE).
- `stations_in_service`: Número de estaciones para tomar o dejar bicicletas que estaban en servicio ese día.
- `weekday`: Valor lógico, indica si esa fecha corresponde a un día entre semana (de lunes a viernes, ambos inclusive).
- `weekday_non_holiday`: Valor lógico, indica si la fecha corresponde a un día entre semana festivo.

Los datos están tomados con frecuencia diaria (filas del archivo).

Ejercicio 1

Genera una tabla con valores estadísticos resumen para las variables cuantitativas de este conjunto de datos.

Ejercicio 2

Crea un gráfico que represente la evolución del número total de viajes en bicicleta registrados en el sistema cada mes.

A continuación, genera otro gráfico con la evolución de la media mensual de temperaturas máximas y mínimas.

¿Se pueden observar patrones estacionales o algún tipo de relación entre ambas variables?

Ejercicio 3

Representa un gráfico con dos paneles, en el que cada panel muestre el histograma y función de densidad de probabilidad del número total de viajes diarios realizados. El panel izquierdo mostrará la distribución del total de viajes diarios en dias no festivos y el panel derecho mostrará la misma distribución pero para días festivos.

Ejercicio 4

Calcula cual es, en promedio el día de la semana en el que más viajes en bicicleta se realizan y el día que menos viajes registra, usando toda la serie de valores. Si es posible, intenta visualizar estos datos por paneles para mostrar tus conclusiones.