

Tarea final del curso Herramientas Computacionales 2021-1

D. Sierra-Porta

October 15, 2021

1 Descripción de las actividades

En el repositorio de Github (<https://github.com/sierraporta/Herramientas-Computacionales-Basicas>) que hemos creado con antelación para el curso y en donde hemos puesto varios códigos y ejercicios y clases he colocado una carpeta que se llama “Tarea_final” y allí he colocado cuatro conjuntos de datos:

1. “Global_active_power.csv”
2. “Global_reactive_power.csv”
3. “Global_intensity.csv”
4. “Voltaje.csv”

Estos datos (que yo he modificado de los originales), contiene 475024 mediciones recogidas en una casa situada en Sceaux (7 km de París, Francia) entre enero y noviembre de 2010 (11 meses).

Notas:

1. $(\text{Global_active_power} \times 1000 / 60)$ representa la energía activa consumida cada minuto (en vatios-hora) en el hogar por los equipos eléctricos.
2. $(\text{Global_reactive_power} \times 1000 / 60)$ representa la energía reactiva consumida cada minuto (en vatios-hora) en el hogar por los equipos eléctricos.
3. Todas las marcas de tiempo del calendario están presentes en el conjunto de datos.

Para esta tarea final sólo usará PANDAS y quizás algunas funciones de NUMPY y por su puesto MATPLOTLIB.

Deberá entregar un JUPYTER NOTEBOOK en grupos de tres personas como máximo con un código sencillo y coherente, además el JUPYTER NOTEBOOK deberá estar documentado, usando MARKDOWN para comentarios y secciones. El JUPYTER NOTEBOOK deberá estar en un repositorio GITHUB y debe contener la información de los integrantes del grupo con código e identificación. Usted deberá proveer el enlace correspondiente en el que el profesor pueda verlo y evaluarlo. Esta tarea final tiene como límite máximo de entrega para el día Martes 19 de octubre de 2021 hasta las 23:59 horas. No se evaluarán tareas luego de este día. El enlace deben enviarlo por el canal que se ha creado para dicho envío.

Se evaluará para esta tarea los siguientes criterios: orden y organización (1.5ptos), metodología (2ptos), documentación (1.5ptos) para un total de 5ptos.

A continuación de detallan las actividades.

2 Actividades

1. Debe leer los cuatro archivos. Todos los archivos tienen la misma marca de tiempo.
2. Deberá juntar/mezclar/concatenar los cuatro archivos para obtener un sólo archivo.
3. Una vez que tenga un sólo archivo con la información de las cuatro columnas “Global_active_power”, “Global_reactive_power”, “Global_intensity” y “Voltaje”, con la correspondiente columna de “Datetime”, verá que los datos tienen escalas de minutos. Seguidamente queremos hacer un *resample* de los datos para tenerlos ahora en escalas de días (con promedios diarios, con lo que obtendrá un nuevo conjunto

de datos que tendrá no más de 365 líneas), y también otro *resample* de los datos para tenerlos ahora en escalas de meses (con promedios mensuales, con lo que obtendrá un nuevo conjunto de datos que tendrá no más de 12 líneas). **Sugerencia:** use una librería de pandas llamada “pandas.Series.resample”, <https://pandas.pydata.org/docs/reference/api/pandas.Series.resample.html>.

4. Por último, queremos evaluar un modelo sencillo de regresión para tener una relación entre “Global_active_power” y “Global_intensity” y/o “Voltaje”, es decir, algo como

$$\text{Relación 1: } \text{Global_active_power} = A_1(\text{Global_intensity}) + B_1, \quad (1)$$

$$\text{Relación 2: } \text{Global_active_power} = A_2(\text{Voltaje}) + B_2, \quad (2)$$

$$\text{Relación 3: } \text{Global_active_power} = A_{31}(\text{Global_intensity}) + A_{32}(\text{Voltaje}) + B_3. \quad (3)$$

No olvide que al hacer la regresión debe especificar los valores de los parámetros A_i y B_i y sus errores correspondientes como hicimos en clase. Recuerde que en el repositorio hay códigos ya hechos para esto que los usamos para practicar. Esta parte debe hacerla usando la librería “curve_fit” del módulo “scipy.optimize” que vimos en clases anteriores.

5. Diga cual de las relaciones fue la mejor y cual la peor. Sería ideal hacer las regresiones con todos los dataset que construyó (es decir, los de frecuencia por minutos, por mes o por día) para hacer comparaciones de que es apropiado o que no. Haga gráficos de los datos y las líneas de regresión como lo hicimos en clase. Para evaluar el mejor modelo calcule los errores de la regresión con los datos.

Eso es todo. Buena suerte y recuerde ser ordenado y usar las buenas prácticas de programación que les han enseñado!