Temas Selectos de Física Computacional III: Introducción a la ciencia de datos. Clase 3: 18 Febrero 2025

Dr. Leonid Serkin

1 Introducción

El propósito de esta clase es practicar con el proceso de preparación de datos tabulares, aprendiendo cómo manejar anomalías en los datos.

2 Conjunto de datos Global Land Temperature

El conjunto de datos de Temperatura Global de la Tierra (Global Land Temperature, GLT) es una gran colección de mediciones mantenidas activamente por Berkeley Earth (`https://berkeleyearth.org/data/`). Contiene los datos fuente en bruto medidos con estaciones alrededor del mundo, además de un formato intermedio y varios archivos de salida formateados. Los datos abarcan desde aproximadamente 1750 hasta días recientes con disponibilidad mensual y diaria. Las mediciones se proporcionan por hemisferios, estados, países, ciudades y más.

Trabajarás con una versión modificada, más pequeña pero más difícil, del conjunto de datos original de GLT, para destacar la importancia de la preprocesamiento de datos. Más específicamente, esta versión didáctica contiene los archivos de salida formateados de las principales ciudades del mundo con granularidad mensual. Por simplicidad, el análisis abarcará casi dos siglos (es decir, entre los años 1817 y 2012).

El conjunto de datos está compuesto por aproximadamente 200,000 filas correspondientes a las mediciones tomadas el primer día del mes en una ciudad dada. Cada medición se describe con 7 valores:

- Fecha, cuando se tomó la medición
- Temperatura Media ($AverageTemperature$)
- Incertidumbre de la Temperatura Media ($AverageTemperatureUncertainty$)
- Ciudad, desde donde se tomó la medición

- País
- Latitud
- Longitud

El conjunto de datos está disponible en formato CSV:

`GLT_filtrado.csv`

3 Preprocesamiento de datos en GLT

Empecemos[1]

1. Carga el conjunto de datos de Temperatura Global de la Tierra como una lista de listas.
2. Tómate un momento para inspeccionar mejor los atributos con los que vas a trabajar. Cuántos de ellos son continuos o discretos, valores o texto? Cuántas entradas hay? Cuántas ciudades distintas?
3. Analiza el atributo *AverageTemperature*, que contiene valores faltantes. Cuenta cuantos valores faltantes tienes en el conjunto. Crees que es un porcentaje grande?

Existen muchas formas de manejar valores faltantes. Se puede decidir eliminar una fila del conjunto de datos si contiene un valor faltante. Esta estrategia puede adoptarse cuando el conjunto de datos es grande y la pérdida de información no afecta la distribución general. Otra solución común es llenar cada valor faltante. Si los datos no tienen un orden específico, se pueden reemplazar con la media (o la mediana) del atributo involucrado. Los datos temporales, en cambio, permiten reemplazar los valores faltantes con los valores de las filas adyacentes, por ejemplo, promediándolos. Claramente, esta técnica es posible si el tipo de datos permite calcular la media.

1. Llena cualquier vacío con la media aritmética entre la medición antecedente más cercana y la medición sucesiva más cercana en el tiempo, tomada en la misma ciudad. Asume las siguientes reglas para casos extremos:
    - Puede suceder que un valor faltante no tenga una medición precedente (o sucesiva). Esto sucede cuando el valor faltante es el primero (o el último) valor del conjunto de datos. Si este es el caso, considera que el valor faltante está precedido (o seguido) por un 0, luego calcula la media en consecuencia.

[1] Eres libre de utilizar cualquier herramienta que prefieras, ya sea Python puro, NumPy, Pandas u otras bibliotecas para manejar y analizar los datos.

```
original = [ '', 5, 6, '' ]
paso_1 = [ 2.5, 5, 6, '' ] # (0 + 5) / 2
paso_2 = [ 2.5, 5, 6, 3 ] # (6 + 0) / 2
```

Aquí te doy un código que lo hace, pero puedes crear el tuyo:

```
import numpy as np

def llenar_faltantes(lista):
    # Convertir la lista a un array de NumPy, reemplazando '' con NaN
    datos = np.array([float(x) if x != '' else np.nan for x in lista])

    # Recorrer todos los elementos de la lista
    for i in range(len(datos)):
        if np.isnan(datos[i]):  # Si el valor es NaN (antes era '')
            # Buscar el valor anterior más cercano
            previo = 0
            for j in range(i-1, -1, -1):
                if not np.isnan(datos[j]):
                    previo = datos[j]
                    break

            # Buscar el valor siguiente más cercano
            siguiente = 0
            for j in range(i+1, len(datos)):
                if not np.isnan(datos[j]):
                    siguiente = datos[j]
                    break

            # Reemplazar NaN con la media de los vecinos más cercanos
            datos[i] = (previo + siguiente) / 2

    return datos.tolist()  # Convertir el array de NumPy de vuelta a lista
```

- Si hay valores faltantes consecutivos, simplemente cálculalos en orden temporal y utiliza los valores recién insertados para evaluar los siguientes. Aquí tienes un ejemplo con una lista simple donde se han aplicado ambas reglas:

```
original = [ '', '', 24, 28.9 ]
paso_1 = [ 12, '', 24, 28.9 ] # (0 + 24) / 2
paso_2 = [ 12, 18, 24, 28.9 ] # (12 + 24) / 2
```

Entonces, las modificaciones a la función previa deberían de ser:
a) en lugar de recorrer todos los datos de golpe, los valores faltantes se deben calcular uno por uno en el orden en que aparecen;
b) cuando hay valores consecutivos en blanco, el segundo debe usar el valor recién insertado del primero.

4 Cambio climático? Calentamineto global?

Ahora después de resolver el problema de los valores faltantes de nuestro conjunto de datos, vamos a estudiarlo:

1. Encuentra la temperatura más caliente y más fría en estas ciudades: Cairo, Moscow, Peking, Rome, Mexico.

2. Hay una anomalía en la distribución de datos. Con la ayuda de Matplotlib, grafica la distribución de las temperaturas medias terrestres para Roma y Bangkok.

3. Como puedes ver, Roma y Bangkok tienen distribuciones de temperatura muy diferentes. Sin embargo, lo que resulta extraño es la gran diferencia en la magnitud de sus temperaturas. Es posible que todos los sensores de las estaciones de Bangkok estuvieran defectuosos? Qué crees que pudo haber sucedido aquí? Hay otras ciudades que presentan el mismo problema? Usa tu creatividad para analizarlo y proponer una explicación.

4. Ahora sí, ya que lo encontraste, calcula la temperatura media anual para las siguientes ciudades: London, New York, Bangkok, Tokyo, Paris, Lima. Puedes agregar las ciudades que desees y que te interesen.

5. Identifica el top 5 de las ciudades donde el cambio climático ha sido más pronunciado desde el siglo XIX. Hay muchas maneras de abordar esta tarea. Empieza con graficar la temperatura promedio anual de cada ciudad y observar cuáles han tenido el mayor incremento, y después divide los datos en periodos de 5, 10, 25 años y observa cómo han cambiado las temperaturas medias.