

Control de versiones

Profesores: Walter Gómez & Nicolás Caro

Considere el archivo de datos correspondiente a su grupo adjunto a la tarea en la plataforma. El archivo contiene 4 columnas: Año, Mes, Fecha (como variable continua) y CO2 (concentración). El objetivo final de la tarea es ajustar una función $F(t)$ que describa aproximadamente la dependencia de la concentración respecto a la fecha. Para ello se deben seguir los siguientes pasos, utilizando la librería `scipy.optimize` para optimizar.

Pasos

1. El dato de CO2 debe ser un número real positivo. Identifique el intervalo de tiempo continuo (meses consecutivos) más extenso que cumple tres condiciones:

- El intervalo termina con la última observación de Septiembre 2020.
- El intervalo comienza con una observación con un dato válido de CO2.
- En todo el intervalo más del 95% de las observaciones tienen un dato de CO2 válido.

Extraiga ese intervalo como la base de datos a ser utilizada en esta tarea.

2. Cree una nueva columna '*CO2_completa*' que rellene los datos de CO2 faltantes de su base de datos utilizando la siguiente metodología:

- Identificar cada intervalo que comienza y termina con una observación válida, pero que tiene entremedio solo observaciones faltantes de CO2, por ejemplo como el siguiente (mostrando solo fecha y CO2):

$\{ \dots (1958.7068, 315.33)(1958.7890, -99.99)(1958.8740, -99.99)(1959.9562, -99.99)(1959.0411, 315.09) \dots \},$

- Rellenar los valores intermedios del intervalo utilizando la recta que une los extremos del intervalo. En el ejemplo anterior sería la recta $C(t) = \frac{1}{0.3343}[(t - 1958.7068) \cdot 315.09 - (t - 1959.0411) \cdot 315.33]$, que genera la siguiente secuencia rellenada

$\{ \dots (1958.7068, 315.33)(1958.7890, 315.271)(1958.8740, 315.210)(1959.9562, 315.151)(1959.0411, 315.09) \dots \}$

3. Ajuste una función polinomial de grado n de la forma

$$F_{\theta}(t) = \theta_0 + \theta_1(t - T) + \theta_2(t - T)^2 + \dots + \theta_n(t - T)^n$$

que minimice el error cuadrático medio (donde T es la primera fecha de sus datos). O sea, para un valor del grado n dado, resuelva el problema:

$$\min_{\theta} \frac{1}{N} \sum_{j=1}^N (x_j - F_{\theta}(t_j))^2$$

donde N es el número total de datos, t_j la variable fecha y x_j la variable *CO2_completa*. Pruebe con los valores $n = 0, 1, 2, 3, 4$ y muestre los resultados obtenidos indicando en cada caso:

- Los valores optimales del vector θ .
 - El valor optimal de la función de error cuadrático medio.
 - Tiempo requerido en el proceso de optimización.
 - Un gráfico que contenga el ploteo de dos curvas, una con los puntos originales y otra con la función $F_\theta(t)$ obtenida.
4. Seleccione una de las funciones ajustadas que a su criterio mejor aproxima los datos y genere una nueva columna *CO2_centrado* que contenga para cada fecha la resta de *CO2_completa* menos el valor de la función $F_\theta(t)$ seleccionada. Indique la media y la desviación standard de *CO2_completa*.
 5. Ajuste ahora con una componente periódica de la forma

$$F_\beta(t) = \beta_0 \cdot \sin(\beta_1 \cdot (t - T) + \beta_2) \cdot e^{\beta_3 t}$$

para capturar parcialmente la dependencia de *CO2_centrado* respecto al tiempo. O sea, resuelva el problema:

$$\min_{\beta} \frac{1}{N} \sum_{j=1}^N (z_j - F_\beta(t_j))^2$$

donde N es el número total de datos, t_j la variable fecha y z_j la variable *CO2_centrado*. Indique el parámetro β optimal, el tiempo requerido para la optimización y genere un gráfico que plotee simultaneamente *CO2_completa* y $F_\beta(t)$ (respecto al tiempo).

6. Finalmente defina su aproximación como $F(t) = F_\theta(t) + F_\beta(t)$ con las funciones calculadas y entregue la siguiente información
 - Error cuadrático medio total entre $F(t)$ y *CO2_completa*.
 - Gráfico que plotea simultaneamente *CO2_completa* y $F(t)$ (respecto al tiempo).
 - Predicción que entrega $F(t)$ para todos los meses entre Octubre 2020 y Abril 2022.

El primer objetivo de esta evaluación es medir conocimientos sobre el manejo general de Python. De manera simultanea, se evaluará el uso del sistema Git y el control de versiones. Por tal motivo, para este ejercicio se exige:

- Generar un repositorio de control de versiones Git donde se trabajará con este archivo.
- Por lo menos un commit por cada integrante del grupo.
- Por lo menos un merge.
- El formato de entrega de esta evaluación es un archivo .zip con el repositorio correspondiente a este trabajo.

Cada grupo debe subir su solución antes del día **Miércoles 08 junio a las 23:30 hrs**. El nombre del archivo zip debe seguir el siguiente formato: *Numero_grupo_IMA602.zip*.