

Práctica 8. Modelos no paramétricos: Procesos Gaussianos

Objetivo

El objetivo de esta práctica es utilizar **Procesos Gaussianos** (*GP*s), un método no paramétrico de aproximación de funciones basado en kernels. Utilizaremos *GP*s para resolver un problema de regresión no lineal y explorar el papel que juega el kernel y sus hiperparámetros en el ajuste de la función.

Estudio previo

- Repasa las transparencias de clase y estudia las funciones auxiliares proporcionadas para esta práctica.
- Familiarízate con la documentación de la clase `GaussianProcessRegressor` de `scikit-learn`, los distintos tipos de *kernels* que se pueden utilizar y su interpretación (estos enlaces: [1] y [2], pueden ayudarte a esto último).

Desarrollo de la práctica

1. **Regresión simple.** Utilizaremos el notebook `P8_1D.ipynb`. El notebook te guiará en los distintos pasos:
 - a) Emplea la clase `GaussianProcessRegressor` para ajustar el dataset `weightdata_clean.mat`. Utiliza primero un kernel `RBF` y ajusta sus parámetros manualmente. Explica cómo cambian los resultados.
 - b) Ahora, explora el comportamiento de otros *kernels*. Puedes ver los distintos tipos de *kernels* soportados por `scikit-learn`. Justifica tu elección y los resultados que obtienes con los parámetros elegidos manualmente. Para impedir la optimización de los hiperparámetros, usa `optimizer=None` al instanciar `GaussianProcessRegressor`.
 - c) Ahora, optimiza el kernel elegido, realizando para ello la instanciación de `GaussianProcessRegressor` sin especificar `optimizer=None`. Compara los resultados con tu ajuste manual.
2. **Regresión 2D.** Utilizaremos el notebook `P8_2D.ipynb`:
 - a) Revisa el notebook proporcionado. En este apartado, trabajaremos con los datos, accesibles públicamente en el [portal del Ayuntamiento](#), de calidad del aire en Zaragoza. Concretamente, usaremos el dataset que os aportamos `calidad_aire_2023.csv`.

Tu objetivo es utilizar procesos Gaussianos para predecir (interpolan) la calidad del aire en el área de la ciudad comprendida por las [ocho estaciones](#) de medición que se encuentran en Zaragoza.
 - b) Propón *kernels* para realizar las estimaciones con los contaminantes que elijas analizar. Justifica tus elecciones.
 - c) ¿Dónde colocarías tu la siguiente estación? ¿Por qué?

3. **Opcional:** Utilizaremos el notebook `P8_extra.ipynb` :

- a) Dispones de varios datasets con los datos de calidad de aire de la estación “Centro”, ubicada en Plaza Aragón. Cada fichero tiene los datos de un indicador: `03ZgzCentro.csv` , `N2ZgzCentro.csv` , `COZgzCentro.csv` y `noxZgzCentro.csv` .
- b) Utiliza lo que has aprendido en el apartado anterior para hacer una predicción de los datos a partir del histórico o, si existen, para rellenar huecos en él. Para ello, observa los datos, diseña un kernel adecuado, optimiza los hiperparámetros y analiza los resultados.
- c) ¿Qué *kernels* has diseñado? Justifica tu respuesta y evalúa los resultados obtenidos.

i Info

La nota máxima sin el apartado opcional es 8.

A entregar en Moodle

Los notebooks `P8_1D.ipynb` , `P8_2D.ipynb` y `P8_extra.ipynb` con el código de cada apartado, los resultados, su interpretación y las conclusiones que hayas obtenido.