Práctica 8. Modelos no paramétricos: Procesos Gaussianos

Objetivo

El objetivo de esta práctica es utilizar **Procesos Gaussianos** (*GPs*), un método no paramétrico de aproximación de funciones basado en kernels. Utilizaremos *GPs* 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) Optimiza ahora el kernel elegido utilizando la función `fit`. 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 Madrid. Concretamente, usaremos el dataset que os aportamos en la carpeta `historico`.
   Tu objetivo es utilizar procesos Gaussianos para predecir (interpolar) la calidad del aire en el área de la ciudad comprendida por las mas de veinte **estaciones** de medición que se encuentran en Madrid.
   b) Propón *kernels* para realizar las estimaciones con los contaminantes que elijas analizar. Justifica tus elecciones.
   c) Razona como podrías utilizar ell proceso Gausiano para colocar la siguiente estación. No puedes colocarla fuera del area ya cubierta por otras estaciones.

3. **Opcional**: Utilizaremos el notebook `P8_extra.ipynb`:
   a) En la carpeta `datosDiariosPorAnio` dispones de los ficheros con datos diarios desde el 2001 para todas las estaciones de Madrid.
   b) Utiliza lo que has aprendido en el apartado anterior para hacer una predicción de los datos a futuro 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. Selecciona casos interesantes observando diferentes estaciones y contaminantes.
   c) ¿Qué *kernels* has diseñado? Justifica tu respuesta y evalúa los resultados obtenidos.

> **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. Entregar en un único fichero comprimido.

Recuerda:

- Trae la práctica preparada para aprovechar la sesión de prácticas al máximo.
- Si te atascas, pregunta en la sesión o en tutorías.
- Debes citar correctamente todas las fuentes utilizadas.
- Tienes 6 días desde tu sesión para depositar la práctica en Moodle.
- Deberás defenderla en tu próxima sesión de prácticas.