Modelo Estadístico

Grupo MAR - Universidad de Murcia

El modelo estadístico aporta a SinQlair una nueva forma de pronosticar los índices de calidad y la concentración de contaminantes, con mayor capacidad y menor coste computacional que el modelo WRF+CHIMERE o MM5+CHIMERE. Esto permite realizar pronósticos más largos en el tiempo, de hecho, se utilizan 10 días aunque este valor puede cambiarse internamente en el sistema.

La principal característica del modelo estadístico es que encuentra posibles relaciones empíricas partiendo de una gran base de datos, tanto meteorológicos como de contaminación. Más concretamente, se fundamenta en explicar la concentración de contaminantes a partir de la situación meteorológica que rija en el momento.

Los datos utilizados son de dos tipos. Por un lado están los predictores, que van a estar compuestos por datos meteorológicos (históricos del modelo GFS) de distintas variables como la precipitación, la temperatura, el viento, la radiación solar y la altura de la capa límite entre otros. Para estas variables, dada la gran cantidad de datos disponible y con el fin de reducirlos, se opta por realizar un análisis en componentes principales (PCA) para construir las climatologías. Por otro lado, están los predictandos, que van a ser las concentraciones de los distintos contaminantes del sistema, esto es, NO_2, SO_2, O_3, PM_{10} y $PM_{2,5}$. Estos últimos son aportados por la Red de Vigilancia de la CARM. Con estos datos, se construye la base de datos sobre la que el modelo encuentra las relaciones antes comentadas.

Para encontrar las relaciones, cada día se descargan datos meteorológicos que deben ser preprocesados para analizar la situación, y para ello se realiza la regresión (pronóstico de un predictando en función de los predictores). Para tal fin, exiten una gran diversidad de modelos dentro del campo del *Machine Learning*, entre los que destacan las **Redes Neuronales Densamente Conectadas** y los **Modelos Random Forest**, ambos utilizados en el caso de SinQlair.

Cuando la regresión a través de estos modelos es aplicada sobre unos datos de entrada, se produce una salida que debe ser corregida convenientemente en lo que se conoce como *calibración del modelo*. Además, para asegurar que el modelo se comporta adecuadamente, es necesario validarlo una vez calibrado. Esto se consigue por medio de la técnica de la crossvalidación.

Todas estas técnicas de *Machine Learning* vienen incorporadas en librerías específicas de *Python* como *Scikit-Learn*, que es una de las que se han utilizado.

A modo de resumen:

1. Se descargan los datos del GFS cada día.
2. Se preprocesan dichos datos.
3. Se llama al script que hace la regresión, utilizando los modelos entrenados.
4. Se postprocesan las concentraciones obtenidas y se calculan los ICA.