

Regresión Polinómica

Se desea predecir el peso de un pescado para venderlo en el mercado a partir de sus dimensiones y su especie. El archivo `Fish.csv` contiene una base de datos, donde se indica:

- **Species:** Variable categórica que representa la especie del pez. Incluye siete categorías: Bream, Roach, Whitefish, Parkki, Perch, Pike y Smelt.
- **Weight:** Variable numérica que representa el peso del pez en gramos. Esta es la variable objetivo en los modelos de predicción de peso.
- **Length1:** Variable numérica que indica la primera medición de longitud del pez (en centímetros). Representa la distancia desde el extremo frontal del pez (como la punta del hocico) hasta la base de la cabeza o el inicio de las aletas pectorales.
- **Length2:** Variable numérica para la segunda medición de longitud del pez (en centímetros). Corresponde a la longitud estándar, que abarca desde el extremo frontal del pez hasta la base de la aleta caudal (el inicio de la cola).
- **Length3:** Variable numérica para la tercera medición de longitud del pez (en centímetros). Esta es la longitud total, desde el extremo frontal del pez hasta la punta de la aleta caudal.
- **Height:** Variable numérica que representa la altura del pez, medida en el punto más alto de su cuerpo (en centímetros).
- **Width:** Variable numérica que representa el ancho o grosor del pez (en centímetros), medida en su punto más ancho.

(a) Exploración de datos:

1. Indicar la proporción de las variables categóricas representando probabilidades. 📖: Si no sabe qué tipo de variable aleatoria es la categórica, deberá buscar dicha información.
2. Para analizar las variables numéricas, utilice el comando `pairplot` (seaborn). Explique qué representan los gráficos.
3. Definir los conjuntos de entrenamiento, validación y testeo respetando los índices presentes en el archivo `indicesforfish.json`.

(b) Pre-procesamiento:

1. Utilizar el comando `OneHotEncoder` (sklearn) para codificar las variables categóricas como `one-hot`. 📖: En caso de no conocer el concepto, buscar información sobre `one-hot encoding`.
2. Utilice el comando `PolynomialFeatures` (sklearn) para crear un mapa polinómico de orden 2 sobre las variables numéricas (exceptuando el peso del pescado).
3. Combinar y posteriormente normalizar las salidas de las operaciones anteriores. 📖: Herramientas recomendadas son `ColumnTransformer` y `StandardScaler` (sklearn). Recordar que la normalización se define durante el entrenamiento.

(c) Regresión Lineal:

1. Utilizar `Pipeline` (sklearn) para combinar las operaciones anteriores con una regresión lineal. 📖: Se recomienda utilizar `RegressionLineal` (sklearn).
2. ¿Cuántos parámetros tiene el modelo? Explicar por qué.

3. Utilizar `set_config(display="diagram")` (sklearn) para hacer un diagrama del sistema definido.
4. Entrenar el sistema descrito anteriormente.
5. Reportar el error cuadrático medio de entrenamiento y validación.

(d) *Regularización:*

1. Utilizar `Ridge` (sklearn) para entrenar una regresión lineal regularizada para diferentes valores de $0 < \lambda < 25$.
2. Graficar el error cuadrático medio de entrenamiento y validación en función del hiperparámetro λ .
3. Reportar el λ que minimiza el error cuadrático medio de validación.
4. Reportar el error cuadrático medio de testeo para el λ obtenido anteriormente.