Actividad: Regresión Lineal con Términos Cuadráticos

Profesor de Introducción al Deep Learning

Contexto

Eres parte de un equipo de científicos de datos en una startup ambiental que busca predecir el **índice de calidad del aire (AQI)** en ciudades de Europa. Los datos simulados contienen 5 variables predictoras: - `NO2` (dióxido de nitrógeno) - `PM10` (partículas en suspensión) - `SO2` (dióxido de azufre) - `CO` (monóxido de carbono) - `O3` (ozono)

Tu proposito es **modelar el AQI** a partir de estas variables para predecir la calidad del aire en tiempo real. Para esto debes implementar un modelo de regresión lineal que mejor se ajuste a los datos y evaluar su desempeño con validación cruzada, (el modelo puede tener términos cuadráticos, cúbicos, de una o todas sus variables).

Recuerde que

Error Cuadrático Medio (MSE)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Coeficiente de Determinación (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Preguntas

1. Análisis Exploratorio y Preprocesamiento

- **a)** Realice un análisis exploratorio de los datos, en este caso, todas las variables son importantes para el modelo, para este análisis se recomienda visualizar la correlación de las variables entre ellas y con el AQI.

2. Ecuación Normal

- **a)** Deriva matemáticamente la ecuación normal para θ incluyendo términos cuadráticos, cúbicos, etc.:

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

- **b)** Si $\mathbf{X}^\top \mathbf{X}$, como es la condición de la matriz ¿tiene algún problema este número de condición?
- **c)** ¿Qué ventajas tiene la ecuación normal sobre el gradiente descendente?
- **d)** Implemente un algoritmo, para esto divida los datos en 80% para entrenamiento y un 20% para el test, unicamente usando "numpy", que calcule los coeficientes $\hat{\beta}$ de la regresión lineal, incluyendo términos cuadráticos, cúbicos, etc, para mejorar el ajuste del modelo. Escoja el modelo que mejor se ajuste a los datos test.

3. Gradiente Descendente

- **a)** Explica por qué se normalizan las características antes de aplicar gradiente descendente
- **b)** Implementa un algoritmo de gradiente descendente para regresión lineal, unicamente usando "numpy", incluyendo términos cuadráticos, cúbicos, etc. para mejorar el ajuste del modelo. para mejorar el ajuste del modelo. Escoja el modelo que mejor se ajuste a los datos test.
- **c)** si los modelos de gradiente descendente y ecuación normal son diferentes, ¿cuál es el motivo?, y si son iguales, ¿cual tiene mejor desempeño?
- **c)** ¿Qué ventajas tiene el gradiente descendente sobre la ecuación normal?
- **d)** Si el MSE no disminuye después de 200 épocas, ¿qué hiperparámetros ajustarías?

4. Cross Validation

- **a)** Implemente un algoritmo de validación cruzada para evaluar el desempeño de los modelos de regresión lineal implementados en los puntos 2 y 3. ¿Qué modelo tiene mejor desempeño? Con 5 folds.
- **b)** ¿Qué ventajas tiene la validación cruzada sobre la división de los datos en entrenamiento y test?

5. Evaluación del Modelo

- **a)** Evalúe el desempeño de los modelos implementados en los puntos 2 y 3, usando el error cuadrático medio (MSE) y el coeficiente de determinación (R^2). ¿Qué modelo tiene mejor desempeño?