Regresión Lineal Múltiple en Python

Por: Angel Francisco Hernández Gámez

Introducción

La regresión lineal es una técnica estadística ampliamente utilizada en el ámbito del análisis de datos y el aprendizaje automático. Su propósito principal es modelar la relación entre una variable dependiente y una o más variables independientes. En el caso más simple, la regresión lineal simple involucra una sola variable predictora, mientras que en la regresión lineal múltiple se consideran varias variables independientes para mejorar la capacidad predictiva del modelo.

Este enfoque resulta particularmente útil cuando se desea realizar predicciones cuantitativas a partir de datos numéricos. La fórmula general de la regresión lineal múltiple es:

$$Y = b + m_1X_1 + m_2X_2 + \cdots + m_nX_n$$

donde Y es la variable objetivo, $X1$ a Xn son las variables predictoras, m_1 a m_n son los coeficientes que indican la influencia de cada variable independiente, y b es el intercepto.

En este informe, se presenta el desarrollo de una actividad basada en la implementación de regresión lineal múltiple utilizando Python y la biblioteca scikit-learn, siguiendo los pasos propuestos en el libro "Aprende Machine Learning" de Juan Ignacio Bagnato.

Metodología

La actividad consistió en analizar un conjunto de datos sobre artículos relacionados con Machine Learning, con el fin de predecir cuántas veces sería compartido un artículo en función de sus características. Los pasos realizados fueron los siguientes:

Se cargó el conjunto de datos en formato CSV, el cual contenía columnas como el número de palabras, enlaces, comentarios, imágenes y número de veces compartido.

Se realizó una limpieza y filtrado de los datos para eliminar outliers. En concreto, se conservaron solo aquellos artículos con menos de 3000 palabras y menos de 80,000 compartidos.

Se creó una nueva variable llamada "suma", que representa la cantidad total de elementos interactivos del artículo (enlaces, comentarios e imágenes), calculada como:

$$suma = enlaces + comentarios + imágenes$$

Se construyó una matriz de entrada con dos variables predictoras: número de palabras y la variable "suma".

Se dividieron los datos en matrices de entrada (X) y salida (Y), y se entrenó un modelo de regresión lineal múltiple utilizando la clase $LinearRegression$ de scikit-learn.

Finalmente, se generaron predicciones sobre los mismos datos y se evaluó el modelo utilizando métricas como el error cuadrático medio y el coeficiente de determinación R^2.

Además, se hizo una predicción para un artículo hipotético con las siguientes características: 2000 palabras, 10 enlaces, 4 comentarios y 6 imágenes. La variable "suma" en este caso es:

$$suma = 10 + 4 + 6 = 20$$

Resultados

Los coeficientes obtenidos por el modelo fueron los siguientes:

Coeficiente para "Word count": 6.63

Coeficiente para "suma": -483.41

Esto sugiere que por cada palabra adicional en un artículo, se incrementa el número de compartidos en promedio por 6.63, mientras que un incremento en la variable "suma" se asocia con una disminución en el número de compartidos.

En cuanto a las métricas de evaluación, se obtuvo:

Error cuadrático medio (MSE): $352{,}122{,}816.48$

Coeficiente de determinación (R^2): 0.11

El valor de R^2 indica que el modelo explica solo el 11% de la variabilidad observada en los datos, lo cual sugiere un bajo poder predictivo. No obstante, representa una mejora frente al modelo de regresión simple.

La predicción para el artículo con 2000 palabras y una suma de 20 elementos interactivos se calculó como:

$$Y = 6.63 \times 2000 - 483.41 \times 20 = 13{,}260 - 9{,}668.2 = 3{,}591.8$$

(El modelo original predijo alrededor de 20,518; esta fórmula muestra cómo se puede calcular de manera manual.)

Conclusión

La implementación de un modelo de regresión lineal múltiple en Python permitió explorar cómo se pueden combinar múltiples variables predictoras para estimar una variable de salida continua. A pesar de que el rendimiento del modelo fue modesto, esta experiencia ofreció un acercamiento práctico a los fundamentos del aprendizaje automático supervisado.

El bajo valor de R^2 sugiere que el modelo podría mejorarse incorporando nuevas variables, utilizando técnicas de regularización o incluso probando algoritmos no lineales. Además, realizar una selección de características más exhaustiva o aplicar reducción de dimensionalidad podría ayudar a mejorar el ajuste.

En resumen, esta actividad permitió afianzar los conceptos teóricos de la regresión lineal y su aplicación práctica mediante el uso de herramientas modernas como pandas, numpy y scikit-learn. Representa un paso importante en el camino hacia la comprensión y dominio del análisis predictivo con Machine Learning.