Act 10: Programando Regresión Lineal Múltiple en Python

Patricio Ricardí

March 2025

¿Qué es la regresión lineal múltiple?

Es un método estadístico que permite predecir un valor numérico (como el número de veces que se comparte un artículo) usando varias variables relacionadas. La fórmula básica es:

$$Predicción = Intercepto + (m_1 \times Variable_1) + (m_2 \times Variable_2) + \dots$$

Donde:

- $m_1, m_2 ...$ = Coeficientes (indican la importancia de cada variable)
- Intercepto = Valor base cuando todas las variables son cero

Pasos realizados

1. **Cargamos datos**: Usamos un archivo CSV con información de artículos sobre Machine Learning, incluyendo: - Cantidad de palabras - Número de enlaces, comentarios e imágenes - Veces que fue compartido

2. **Limpiamos datos**: - Quitamos artículos muy largos (> 3000 palabras) - Eliminamos casos extremos ($> 80,000$ compartidos)

3. **Creamos variable "interacciones"**:

$$interacciones = enlaces + comentarios + imágenes$$

4. **Preparamos modelo**: - Variables predictoras: `Word count` (palabras) y `interacciones` - Variable a predecir: `# Shares` (compartidos)

5. **Entrenamos modelo**: Usamos la librería `scikit-learn` de Python.

Código simplificado

[language=Python] Filtrar datos $datos_filtrados = dataset[(dataset["Wordcount"] < 3000)(dataset["Shares"] < 80000)]$

Crear variable "interacciones" interacciones = $\text{datos}_filtrados[\text{"}ofLinks\text{"}] + datos_filtrados[\text{"}ofcomments\text{"}].fillna(0) + datos_filtrados[\text{"}Imagesvideo\text{"}]$

Entrenar modelo modelo = LinearRegression() modelo.fit(X=$\text{variables}_predictoras, y = objetivo$)

Evaluar print("Error cuadrático medio:", $\text{mean}_squared_error(y_true, y_pred)$)

Resultados clave

- **Coeficientes**:
  - Por cada palabra: +6.63 compartidos
  - Por cada interacción: -483.41 compartidos
- **Precisión**:
  - Error medio: 352,122,816.48
  - $R^2 = 0.11$ (explica 11% de la variabilidad)
- **Ejemplo**:
  - Artículo con 2000 palabras y 20 interacciones:

$$(2000 \times 6.63) + (20 \times -483.41) = 3,591.8 compartidos$$

Conclusiones

- El modelo es básico pero útil como introducción.
- Para mejorarlo podríamos:
  - Usar más variables (ej: tema del artículo)
  - Probar modelos más complejos (ej: redes neuronales)
  - Mejorar la limpieza de datos
- Muestra cómo pequeños cambios en las variables pueden afectar las predicciones.