

Programando Regresión Logística en Python

March 24, 2025

1 Introducción

En este informe se documenta el desarrollo de un modelo de Regresión Logística multiclase implementado en Python con la biblioteca `scikit-learn`. El objetivo es predecir el sistema operativo de los usuarios de un sitio web, basándose en datos recolectados de comportamiento en línea, como duración de la sesión, número de páginas vistas, acciones realizadas y el valor de dichas acciones.

Este tipo de modelo es útil para problemas de clasificación donde la variable objetivo tiene más de dos clases. En este caso, las clases representan tres sistemas operativos: Windows, Macintosh y Linux.

2 Metodología

Se trabajó con un conjunto de datos de 170 registros. Cada registro contiene las siguientes variables:

- **duración:** duración de la visita (en segundos)
- **páginas:** cantidad de páginas vistas
- **acciones:** interacciones del usuario (clicks, scrolls, etc.)
- **valor:** puntuación total de las acciones realizadas
- **clase:** sistema operativo del usuario (0=Windows, 1=Macintosh, 2=Linux)

El modelo fue entrenado con el 80% de los datos y validado con el 20% restante. También se aplicó validación cruzada para evaluar la estabilidad del modelo.

3 Resultados

Los resultados obtenidos muestran un buen rendimiento general del modelo.

Resumen numérico

- **Precisión del modelo (entrenamiento):** 0.776
- **Precisión en validación (hold-out):** 0.852
- **Predicción de usuario ficticio:** 2 (Linux)

Matriz de Confusión

```
[[16  0  2]
 [ 3  3  0]
 [ 0  0 10]]
```

Reporte de Clasificación

	precision	recall	f1-score	support
0	0.84	0.89	0.86	18
1	1.00	0.50	0.67	6
2	0.83	1.00	0.91	10
accuracy			0.85	34
macro avg	0.89	0.80	0.81	34
weighted avg	0.87	0.85	0.84	34

4 Conclusión

El modelo de regresión logística demostró una precisión del 85% al predecir el sistema operativo de usuarios a partir de sus patrones de navegación. Se observó que la clase con mejor desempeño fue la correspondiente a usuarios de Windows, mientras que los usuarios de Macintosh presentaron menor precisión, posiblemente por su menor representación en los datos.

Se concluye que la regresión logística multiclase es efectiva en este contexto y que, con un conjunto de datos más amplio y balanceado, los resultados podrían mejorar aún más. Esta técnica puede aplicarse a muchos otros escenarios de clasificación de usuarios o clientes en entornos reales.

Anexo

El código fuente del modelo fue desarrollado en Python 3.9