

Act11

Daniel Rojas

March 2025

1 Introducción

La regresión logística es un modelo del aprendizaje automatizado utilizado para resolver tareas de clasificación de datos. Es una variante de la regresión lineal específicamente usada para obtener predicciones discretas.

2 Metodología

Para realizar la actividad, se usó como referencia el capítulo de regresión logística del libro Aprende Machine Learning, adjuntado en las referencias del presente artículo.

Primero hay que preprocesar los datasets y separarlos en sets de entrenamiento y validación. Para ello podemos usar las siguientes líneas de código:

```
1 #Librerias a utilizar
2
3 import pandas as pd
4 import numpy as np
5 from sklearn import linear_model
6 from sklearn import model_selection
7 from sklearn.metrics import classification_report
8 from sklearn.metrics import confusion_matrix
9 from sklearn.metrics import accuracy_score
10 from sklearn.preprocessing import MinMaxScaler

1 #Cargar el dataset
2 dataframe = pd.read_csv("usuarios_win_mac_lin.csv")

1 #Creación del dataset de validación y de entrenamiento
2
3 X = np.array(dataframe.drop(['clase'],axis = 1))
4 y = np.array(dataframe['clase'])
5
6 validation_size = 0.20
7 seed = 7
8 X_train, X_validation, Y_train, Y_validation = model_selection.
    ↳ train_test_split(X, y, test_size=validation_size,
    ↳ random_state=seed)
```

Ahora solo queda entrenar el modelo y evaluar los resultados en el dataset de validación:

```
1 model.fit(X_train, Y_train)

1 #Observar las predicciones
2 predictions = model.predict(X_validation)
3 print(accuracy_score(Y_validation, predictions))
```

Se obtiene una precisión del 82.35%.

También se puede observar la matriz de confusión y el reporte de clasificación para conocer más información sobre la evaluación:

```
1 print(confusion_matrix(Y_validation, predictions))
2 print(classification_report(Y_validation, predictions))
```

De esta forma se obtiene la siguiente matriz:

$$\begin{pmatrix} 16 & 0 & 2 \\ 4 & 2 & 0 \\ 0 & 0 & 10 \end{pmatrix} \quad (1)$$

y el siguiente reporte:

	precision	recall	f1-score	support
0	0.80	0.89	0.84	18
1	1.00	0.33	0.50	6
2	0.83	1.00	0.91	10
accuracy			0.82	34
macro avg	0.88	0.74	0.75	34
weighted avg	0.85	0.82	0.80	34

Figure 1: Reporte de clasificación

3 Resultados

Observando la matriz de confusión, no se puede concluir si realmente el modelo es efectivo o no, ya que en general se manejan muy pocos datos tanto en el set de validación como en el de entrenamiento. Cabe destacar que por lo general se tiene más problemas para clasificar a la etiqueta 1, la cual corresponde a los usuarios de MAC

Por otro lado, el reporte de clasificación muestra un f1-score de 0.80, lo cual es bueno si no consideramos los pocos datos utilizados.

4 Conclusión

La regresión logística es un modelo fácil de usar y de bajos recursos bueno para resolver tareas de clasificación, aunque en este artículo solo se pudo apreciar su fácil implementación.

Por último, la matriz de confusión y el reporte de clasificación son muy buenas métricas para observar que tan bien se desempeñó nuestro modelo, ya que nos permiten observar de forma sencilla las áreas de oportunidad del modelo entrenado, como al mostrar el porcentaje de aciertos de cada categoría.

5 Referencias

Bagnato, J. I. (2020). Aprende Machine Learning.