Instituto Tecnológico y de Estudios Superiores de Monterrey

MÓDULO 2: Portafolio de Análisis

Inteligencia Artificial Avanzada para la ciencia de datos 101.

Ingrid Giselle Paz Ramírez | A00826973

Profesor

Iván Mauricio Amaya contreras

16 de septiembre de 2022

1. Análisis del desempeño del modelo

Se implementó un modelo de Regresión Logística utilizando la librería sci-kit learn.
Dataset utilizado: **Iris dataset**
Consiste en 150 muestras de 3 diferentes especies de Iris (Setosa, Versiocolour y Virginica), así como la longitud de pétalos y sépalos. Los nombres de las columnas encontradas en el dataset son: Sepal Length, Sepal Width, Petal Length y Petal Width.

1.1. Separación de datos y Evaluación del Modelo

Realizar el proceso de validación del modelo de Regresión Logística después de probarlo en un subonjunto de datos requiere de la creación de un conjunto de datos para la validación. El subconjunto de validación se crea a partir de los datos del conjunto de pruebas, por lo que se utiliza la función *train_test_split* nuevamente con el subconjunto de prueba:

```
#Dividir el dataset
x = df.drop(["target"], axis=1)
y = df['target']
xtrain, xtest, ytrain, ytest = train_test_split(x,y,test_size=0.15, random_state=42)
#Crear conjunto de validación
xtest, xval, ytest, yval = train_test_split(xtest, ytest,
test_size = 0.2, random_state = 42)
\caption{División del dataset}
```

Por lo que el 85 % de los datos se utilizó para entrenar el modelo, un 12 % se utilza para probar el modelo y el 3 % restante es utilizado para la validación.

Conjunto	Tamaño	Porcentaje
Entrenamiento	120	80 %
Prueba	24	16 %
Validación	6	4 %

Tabla 1: División del Dataset

Después de entrenar el modelo, utilizando las métricas de sci-kit learn, los resultados obtenidos han sido los siguientes

```
Coeficients and Interception:  [-0.11633479 -0.05977785  0.25491375  0.54759598] 0.25252758981814827
Accuracy:  1.0
R2 score:  0.9574973067781964
Mse:  0.033574176069306634
Rmse:  0.1832325737124997
```

Figura 1: Métricas del modelo sin ajustar

Los hiperparámetros a considerar dentro de la Regresión Logística han sido:

- Solver. Elegir el algoritmo que usa el modelo para la optimización, en donde las opciones son: newton-cg, lbfgs, liblinear, sag, saga
- Penalty .Se utiliza para regular el overfitting al introducir penalizaciones
- C. Un alto valor de C le indica al modelo darle más importancia a los datos de entrenamiento que a los de ajuste.

Se implementó un grid search de forma que se buscaran automaticamente los hiperparámetros que mejoraban el desempeño del modelo, en el que los resultados han sido los siguientes:

```
Best: 0.963889 using {'C': 10, 'penalty': 'l2', 'solver': 'lbfgs'}
0.958333 (0.051595) with: {'C': 100, 'penalty': 'l2', 'solver': 'newton-cg'}
0.955556 (0.055833) with: {'C': 100, 'penalty': 'l2', 'solver': 'lbfgs'}
0.961111 (0.046812) with: {'C': 100, 'penalty': 'l2', 'solver': 'liblinear'}
0.961111 (0.046812) with: {'C': 10, 'penalty': 'l2', 'solver': 'newton-cg'}
0.963889 (0.046564) with: {'C': 10, 'penalty': 'l2', 'solver': 'lbfgs'}
0.961111 (0.046812) with: {'C': 10, 'penalty': 'l2', 'solver': 'liblinear'}
0.958333 (0.059900) with: {'C': 1.0, 'penalty': 'l2', 'solver': 'newton-cg'}
0.958333 (0.059900) with: {'C': 1.0, 'penalty': 'l2', 'solver': 'lbfgs'}
0.955556 (0.059835) with: {'C': 1.0, 'penalty': 'l2', 'solver': 'liblinear'}
0.944444 (0.065734) with: {'C': 0.1, 'penalty': 'l2', 'solver': 'newton-cg'}
0.944444 (0.065734) with: {'C': 0.1, 'penalty': 'l2', 'solver': 'lbfgs'}
0.855556 (0.077380) with: {'C': 0.1, 'penalty': 'l2', 'solver': 'liblinear'}
0.861111 (0.092128) with: {'C': 0.01, 'penalty': 'l2', 'solver': 'newton-cg'}
0.861111 (0.092128) with: {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
0.658333 (0.025000) with: {'C': 0.01, 'penalty': 'l2', 'solver': 'liblinear'}
```

Figura 2: Hiperparámetros para la regresión

En donde se observa que la mejor opción para el modelo es usando C: 100, penalty: l2 y solver: lbfgs.

Las métricas para el nuevo modelo quedan de la siguiente manera:

```
Accuracy:  1.0
R2 score:  1.0
Mse:  0.0
Rmse:  0.0
Coeficients and Interception:  [[-0.37993204  2.06173382 -4.0770181  -1.86887276]
 [ 0.94560501  0.14854084 -0.4086544  -2.00605747]
 [-0.56567296 -2.21027466  4.48567251  3.87493022]] [ 13.04360167   2.30270533 -15.34630701]
```

Figura 3: Métricas para el nuevo modelo

Es importante mencionar que estas métricas fueron calculadas con el conjunto de validación.

A través de la comparación de los valores de las métricas, se observa como los resultados mejoran al ajustar los hiperparámetros. Sin embargo, el nivel de error descrito en el MSE se hace cero al incluir las mejoras.

1.2. Sesgo

El sesgo de predicción es obtenido con la diferencia entre el promedio de las predicciones y el promedio de todas las muestras. Al realizar esta operación con el modelo ajustado con las nuevas correcciones, el sesgo ha resultado de **0.3333**. Se puede graficar la frecuencia de las clases del dataset para analizar el valor del sesgo obtenido.

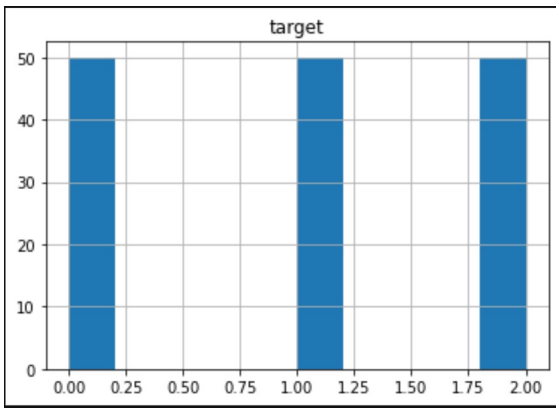

Figura 4: Frecuencias para la clase de Iris

Como el dataset es pequeño, además de que contiene la misma cantidad de datos para cada especie de Iris diferente, el resultado del sesgo obtenido puede ser interpretado como bajo debido a que solo existen tres clases diferentes y sobretodo a que el dataset es pequeño. La significancia del sesgo se observa en el porcentaje de error mse obtenido con el modelo, puesto

que este es de 0.04 antes de mejorar el modelo y 0 después de ajustar los hiperparámetros. Los coeficientes al aplicar el Cross validation Score también han mostrado un sesgo bajo.

1.3. Varianza

La varianza es uno de los indicadores tomados en cuanta para la selección de hiperparámetros. La regresión logística asume homocedasticidad, por lo que se busca que la varianza sea constante. Una forma de analizar la varianza es con el coeficiente de determinación R^2. En el modelo ajustado, el R^2 score ha sido de **0.9638**, esta métrica muestra la relación entre la varianza de las predicciones con la de las muestras contenidas en el dataset. Al tener un coeficiente alto, la diferencia de varianzas es pequeña, por lo que el modelo presenta varianza baja.

1.4. Ajuste

Para determinar el ajuste del modelo, se observan los valores de Accuracy para las predicciones hechas con los datos de prueba y para las predicciones hechas con el conjunto de validación; estos mostrados previamente en las figuras 1 y 3, la diferencia entre ambos valores ha sido de **0.0362** por lo que el ajuste del modelo se considera adecuado.