Informe de Clasificación del Dataset Iris utilizando Regresión Lineal

Leny Santiago López Cruz

Introducción

El presente informe tiene como objetivo mostrar la aplicación de un modelo de regresión lineal en la tarea de clasificación utilizando el conocido dataset Iris. Aunque la regresión lineal es un algoritmo diseñado originalmente para problemas de predicción de variables continuas, en este caso se empleó una técnica de adaptación que permite transformar sus salidas en decisiones de clasificación. El propósito de este ejercicio es explorar los alcances y limitaciones de esta aproximación no convencional en comparación con modelos más adecuados, como la regresión logística o los clasificadores lineales.

Solución

El dataset Iris contiene 150 muestras de flores clasificadas en tres especies: Setosa, Versicolor y Virginica. Cada muestra incluye cuatro características: longitud y ancho del sépalo, y longitud y ancho del pétalo.

Para abordar el problema, se aplicó la técnica de One-vs-Rest (OvR). Este método consiste en entrenar un modelo de regresión lineal para cada clase, de modo que cada uno predice la pertenencia a una especie particular frente al resto. En términos prácticos, se construyeron tres modelos independientes: uno para Setosa, otro para Versicolor y otro para Virginica. Durante la predicción, cada muestra se evalúa en los tres modelos, y se asigna a la clase que produzca el valor de salida más alto. De esta forma, las salidas continuas de la regresión lineal se convierten en etiquetas discretas correspondientes a las especies.

La evaluación se realizó dividiendo el dataset en un 70% para entrenamiento y un 30% para prueba, empleando métricas como *accuracy*, *precision*, *recall*, *F1-score*, además de la matriz de confusión para analizar el rendimiento en detalle.

```python
# 1. Cargar dataset
iris = load_iris()
X = iris.data
y = iris.target
class_names = iris.target_names

# 2. Dividir en entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# 3. Entrenar modelos (One-vs-Rest con regresión lineal)
models = []
for clase in np.unique(y):
    y_train_bin = (y_train == clase).astype(int)
    model = LinearRegression()
    model.fit(X_train, y_train_bin)
    models.append(model)

# 4. Predicción multiclase
def predict_multiclass(X):
    preds = np.column_stack([m.predict(X) for m in models])
    return np.argmax(preds, axis=1)
```

Resultados

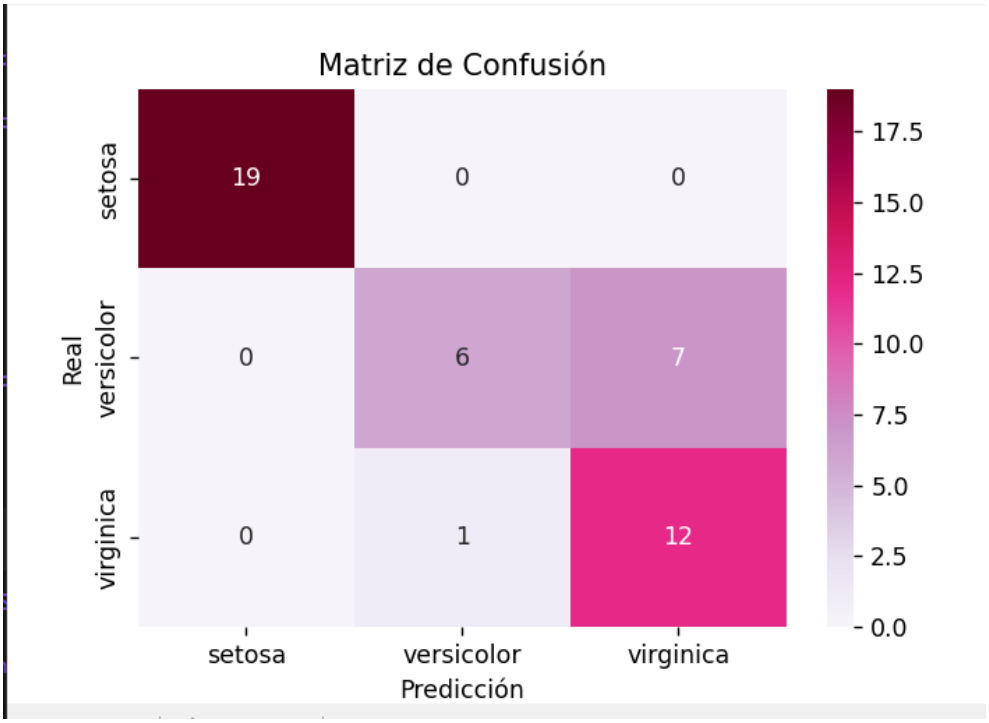
El modelo alcanzó un accuracy global de 82,2% sobre el conjunto de prueba. El reporte de clasificación mostró un desempeño desigual entre las clases:

- La clase Setosa fue identificada correctamente en todos los casos, con valores perfectos de precisión, recall y F1-score de 1.00.
- La clase Versicolor presentó mayores dificultades, alcanzando una precisión de 0.86 pero con un recall de 0.46, lo que indica que varias muestras de esta clase fueron clasificadas erróneamente como Virginica.
- La clase Virginica obtuvo una precisión de 0.63 pero un recall alto de 0.92, reflejando que el modelo tiende a predecir Virginica con frecuencia, aun cuando en algunos casos no corresponde.

```
Accuracy: 0.8222222222222222

Reporte de Clasificación:
              precision    recall  f1-score   support

      setosa       1.00      1.00      1.00        19
  versicolor       0.86      0.46      0.60        13
   virginica       0.63      0.92      0.75        13

    accuracy                           0.82        45
   macro avg       0.83      0.79      0.78        45
weighted avg       0.85      0.82      0.81        45
```

La matriz de confusión confirma este comportamiento. Mientras que todas las muestras de Setosa fueron correctamente clasificadas, se observó confusión significativa entre Versicolor y Virginica: 7 de las 13 muestras reales de Versicolor fueron clasificadas como Virginica, y 1 muestra de Virginica fue clasificada como Versicolor.

En términos globales, el promedio ponderado de las métricas alcanzó valores de precisión = 0.85, recall = 0.82 y F1-score = 0.81, lo que indica un desempeño razonable considerando las limitaciones del enfoque.



Conclusiones

El uso de la regresión lineal para clasificación, apoyado en la técnica **One-vs-Rest con argmax**, permitió obtener un desempeño aceptable en la clasificación del dataset Iris, con un accuracy superior al 80%. Sin embargo, los resultados evidencian limitaciones claras. El modelo es incapaz de separar con precisión suficiente las clases Versicolor y Virginica, lo que se refleja en un número elevado de falsos positivos y falsos negativos entre estas especies.

Este comportamiento se debe a que la regresión lineal no está diseñada para problemas de clasificación, y las salidas continuas no corresponden directamente a probabilidades bien calibradas.

En síntesis, el modelo funciona adecuadamente para distinguir Setosa, pero no resulta confiable para separar Versicolor y Virginica. Esto resalta la importancia de emplear algoritmos adecuados al problema, como la regresión logística, la cual ofrecerían un mejor desempeño en este contexto.

Repositorio: https://github.com/ELPIR0B0/Clasificaci-n-del-Dataset-Iris-utilizando-Regresi-n-Lineal