Binary Logistic Regression Report

Learning Rate and Accuracy

Learning Rate (lr) [Max_iter = 500]	Validation Accuracy
0.1	81.82%
0.01	75.76%
0.001	74.24%
0.0001	72.73%

Chosen Learning Rate: 0.1

Test Accuracy: 77.61% (lr = 0.1)

Problem Understanding

After going through the Binary Logistic Regression using Iris dataset, I applied the same knowledge to the diabetes dataset. The dataset consisted of 442 patients records with 10 features like age, sex, BMI, and blood pressure.

Dataset Analysis

The dataset was analyzed to understand how features and labels are related. It was necessary to process the data properly for training and testing the model. Approach to the Implementation

Data Preprocessing

Following tasks were performed:

→Loading Data: The dataset was loaded using pandas.

→Shuffling: Data was shuffled to avoid bias.

→Splitting Data: The dataset was split into three parts:

→Training set: 75%

→Validation set: 15%

→Test set: 15%

→Features and labels were extracted.

→Feature values were normalized for the same scaling.

→Target variables were binarized.

→Bias term was added in the dataset.

Sigmoid Function

The sigmoid function was used to scale the output between 0 and 1, thus making it suitable for binary classification. The equation of the sigmoid function is as follows:Training the Model

$$\sigma(x) = \frac{1}{1 + e^{-x}}.$$

The steps involved in training are as follows:

→Initialization: Initialize random values for theta parameters, which represent feature importance.

→Weighted Sum and Sigmoid: Calculate the weighted sum of features and apply the sigmoid function to obtain probabilities.

→Cost Function: Calculate the difference between predictions and actual values.

->Gradient Descent: Theta values updated iteratively to minimize cost.

Model Validation

The validation was done by:

Computing the weighted sum of features and applying trained theta values. Using the sigmoid function to produce probabilities. Converting probabilities into binary predictions, threshold = 0.5. Calculating accuracy by comparing predictions to actual labels.

Results Analysis

Different learning rates were tried to see the best results:

->Learning rate 0.1: Best validation accuracy of 81.82%.

-> Learning rate 0.01: Slightly lower accuracy of 75.76%.

-> Learning rate 0.001: The accuracy went down to 74.24%.

-> Learning rate 0.0001: Lowest accuracy of 72.73%.

With a learning rate of 0.1, the model showed a test accuracy of 77.61%. This, though not high, is decent enough considering the noise associated with medical data.

Conclusion

The model did a good job in predicting with logistic regression. A learning rate of 0.1 provided the right balance between stability and accuracy. In spite of the approach shows the capability of logistic regression in medical predictions, given the complexity of the dataset from diabetes.