Name : Dhaarani Shanmugam

Subject : Data Mining and discovery

Topic : Linear Regression Vs Classification

Github : Link

Linear Regression vs Classification

Introduction:

The performance of two distinct methods Linear Regression and Logistic Regression applied to the Breast Cancer Wisconsin (Original) dataset and compared and contrasted in this study. The dataset includes peculiarities associated with the structure of cell nuclei in breast cancer biopsies. The goal is to forecast each sample's class, which may indicate either benign (0) or malignant (1).By thresholding its continuous predictions to the binary classes, the Linear Regression model is handled as a classifier in this report. The second method is a binary classification algorithm called logistic regression. To guarantee the quality of incoming data, fundamental data preprocessing procedures are also carried out, such as addressing missing values, feature scaling, and label encoding.

Data Preprocessing:

1. **Missing Data Handling:** '?' was used to indicate missing values in the Bare_nuclei column. To ensure that all numeric values were filled, these were swapped out for NaN and imputed using the mean technique.

2. **Converting Data Types:** At first, every column was of type object. To ensure that the models handled each feature correctly, pd.to_numeric was used to convert each feature to a numeric representation.

3. **Feature Scaling:** StandardScaler was used to scale the features variance to unity and mean to zero, which is necessary for models that depend on distance-based metrics, such as Logistic Regression.

4. **Label Encoding:** LabelEncoder was used to encode the target variable, Class, into numerical labels, transforming the two categories (benign and malignant) into binary values (0 and 1).

5. **Train-Test Split:** It was used to split the dataset to a training data set which is 70% and a test dataset which is 30%.

Logistic Regression:

Logistic Regression was used as a classifier with 5000 iterations to make sure convergence is present. After the model is trained, the performance on the test set was evaluated using the following metrics:

Accuracy: 96.19%

Precision, Recall, F1-Score: The model exhibited excellent precision , recall for both classes, especially for class 0 (benign) with a recall of 0.99. The precision , recall for class 1 (malignant) were .97 and .91, correspondingly.With only a few errors, the model will be able to correctly categorize the vast majority of the data,, which can be seen in the logistic regression confusion matrix.

Linear Regression (as a Classifier):

Linear Regression was applied as a classifier by predicting continuous values and then thresholding the predictions at 0.5 to classify the instances. The model was trained and evaluated using the following metrics:

Accuracy: 96.19%Mean Squared Error (MSE): 0.038

Classification Metrics: Similar to Logistic Regression, Linear Regression showed a precision of .97 for class one and .96 for class Zero, with recall values of .91 and .99, corresponingly.The confusion matrix revealed that Linear Regression performed similarly to Logistic Regression in terms of classification, although it is not typically used for classification tasks.

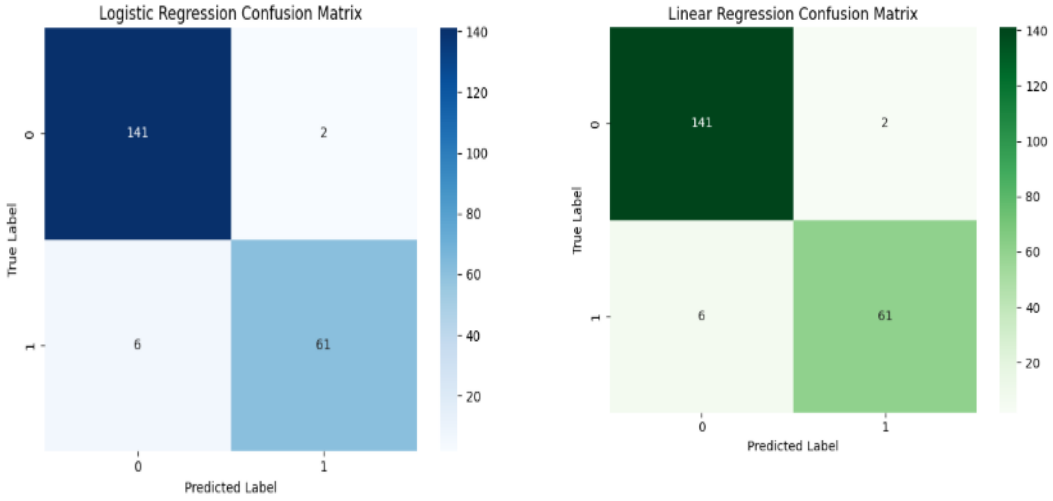
Results Comparison: Both models achieved the same accuracy of 96.19%, which highlights the potential for Linear Regression to be used as a classifier despite it being a regression-based method.

Model	Accuracy	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1-Score (Class 0)	F1-Score (Class 1)
Logistic Regression	96.19%	0.96	0.97	0.99	0.91	0.97	0.94
Linear Regression	96.19%	0.96	0.97	0.99	0.91	0.97	0.94

However, Logistic Regression is generally more suitable for classification tasks due to its probabilistic nature and the output being constrained to a [0,1] range, which aligns better with binary classification.

Confusion Matrices and Model Performance Visualization:

- Logistic Regression: Predicted class 0 (benign) with high accuracy but had a slightly lower recall for class 1 (malignant), which could be improved by adjusting the decision threshold.
- Linear Regression: The confusion matrix for Linear Regression was similar, reinforcing its viability as a classifier, but it generally lacks the inherent classification characteristics of Logistic Regression.



Conclusion: In the case of applying the Breast Cancer dataset to, Logistic Regression and Linear Regression (as a classifier) yielded results that were almost equal. However, because of its probabilistic basis, Logistic Regression is a better option for binary classification applications. Although it can be modified for classification, linear regression is primarily used for regression and does not inherently supply the probabilities required for a binary classification problem.

References:

- James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R.* 2nd ed. New York: Springer.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A. and Cournapeau, D. (2011). *Scikit-learn: Machine Learning in Python*. *Journal of Machine Learning Research*, 12, pp. 2825–2830.

- Bishop, C.M. (2006). *Pattern Recognition and Machine Learning.* New York: Springer.

- Han, J., Kamber, M. and Pei, J. (2012). *Data Mining: Concepts and Techniques.* 3rd ed. San Francisco: Morgan Kaufmann.

- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), pp. 301–320.