Comparative Analysis of Linear Classification and Logistic Regression on Breast Cancer Dataset

Jan Vincent G. Elleazar
Department of Computer Science, College of Information and Computing Sciences, University of Santo Tomas
España Blvd, Sampaloc, Manila, 1008 Metro Manila, Philippines
janvincent.elleazar.cics@ust.edu.ph

I. MODEL BEHAVIOR

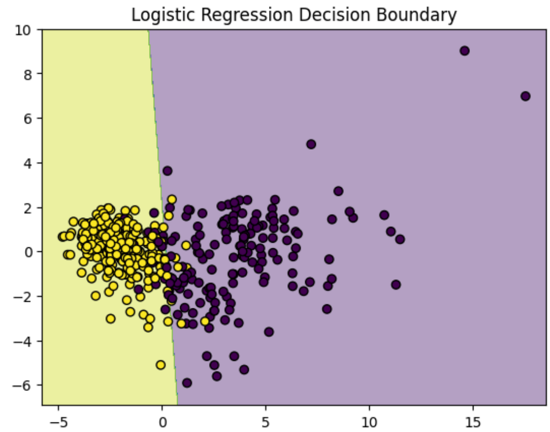

Fig. 1 Decision boundary of Logistic Regression

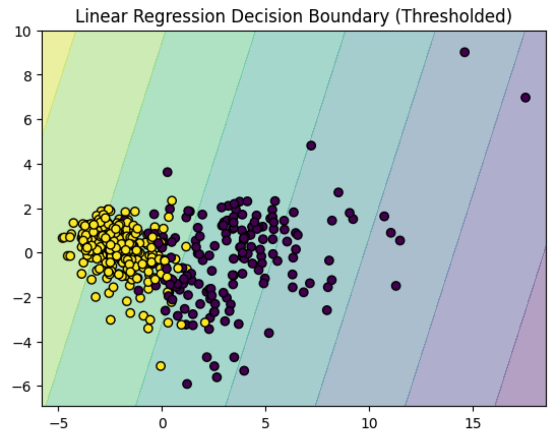

Fig. 2 Decision boundary of Thresholded Linear Regression

In high-dimensional data, Logistic Regression maintains a smooth, interpretable hyperplane as its decision boundary as seen on Figure 1, effectively classifying data points based on probabilities derived from its sigmoid function. This boundary remains relatively robust and less prone to overfitting. In contrast, Linear Regression, when adapted for classification through thresholding, creates a linear decision boundary that becomes increasingly complex and sensitive in high-dimensional spaces as seen on Figure 2. This complexity makes the boundary difficult to interpret and highly susceptible to overfitting, as the model attempts to fit the data too closely. Consequently, the performance of thresholded linear regression heavily relies on the chosen threshold, which can be challenging to optimize in high dimensions.

Logistic Regression handles class imbalance better than Linear Regression due to its probabilistic approach. By using the sigmoid function, it outputs probabilities for each class, allowing for classification even with dispersed data points. This probabilistic interpretation enables the model to learn from and classify minority classes effectively. Linear Regression, while adaptable for classification through thresholding, lacks this probabilistic framework, making it more susceptible to being biased by the majority class and less effective in handling imbalanced datasets.

II. FEATURE IMPORTANCE

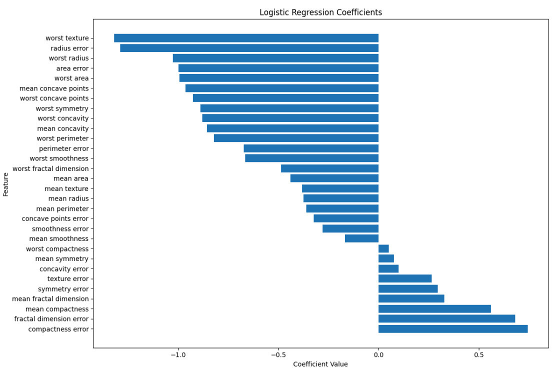

Fig. 3 Bar plot of feature coefficients in Logistic Regression

Feature importance in Logistic Regression is determined by the magnitude of the coefficients. 1 Larger absolute coefficient values indicate greater influence on the classification decision. Positive coefficients increase the log-odds of a positive outcome, while negative coefficients decrease them. As seen in Figure 3, nine features exhibit positive coefficients, suggesting they promote a malignant classification, with Compactness Error, Fractal Dimension Error, and Mean Compactness having the strongest positive impact. Conversely, 21 features have negative coefficients, indicating they favor a benign classification, with Worst Texture and Radius Error exhibiting the largest negative influence.

Irrelevant features introduce noise into both Linear and Logistic Regression models, potentially hindering performance. While L1 regularization in Logistic Regression can effectively shrink the coefficients of these irrelevant features towards zero, essentially performing feature selection, regularization techniques can also be applied in Linear Regression to mitigate the detrimental effects of noise from such features.

III. SCALABILITY

Logistic Regression generally performs better than Linear Regression on significantly larger datasets. Its ability to handle high-dimensional data, coupled with the possibility of hyperparameter tuning and solver optimization, makes it more scalable and efficient. While linear regression can utilize gradient descent, it often struggles to achieve comparable performance on such massive datasets due to its inherent limitations in handling high dimensionality and the computational demands of optimization.

IV. ROBUSTNESS

Logistic Regression is relatively less sensitive to outliers thanks to its sigmoid function, which helps to dampen the influence of extreme values. However, it can still be affected by noise, potentially leading to overfitting. Linear Regression, in contrast, is highly sensitive to outliers, as they can significantly distort the fitted hyperplane. Like Logistic Regression, Linear Regression is also susceptible to noise, which can also lead to overfitting.

Both Linear and Logistic Regression are sensitive to missing or incompatible features, requiring careful handling during data preprocessing. While removing rows or columns containing missing data is an option, it can lead to substantial data loss. Alternatively, techniques like imputation, such as winsorization, can estimate and replace missing values, allowing the models to utilize more of the available data.

V. INTERPRETABILITY

Logistic Regression offers better interpretability for classification prediction compared to Linear Regression. Its probabilistic approach allows for a clearer understanding of the relationship between features and the predicted class, such as the type of breast cancer in the breast cancer dataset. Linear Regression provides less interpretable results in classification prediction due to its lack of probabilistic framework as it doesn't offer the same level of insight into the likelihood of a particular outcome.

With complex datasets, a trade-off emerges between model interpretability and predictive power. Highly complex models, while often achieving greater predictive accuracy due to their ability to capture intricate patterns, become increasingly difficult to interpret. Their complexity obscures the relationship between features and predictions, making it harder to understand how the model arrives at its results. On the other hand, simpler, more interpretable models may sacrifice some predictive power in exchange for the clarity they provide.

VI. METRICS

TABLE I
EVALUATION METRICS OF LOGISTIC AND LINEAR REGRESSION

Metric	**Logistic Regression**	**Linear Regression (Thresholded)**
Accuracy	98%	95%
Precision	97%	92%
Recall	100%	100%
F1 Score	99%	96%
AUC Score	100%	100%

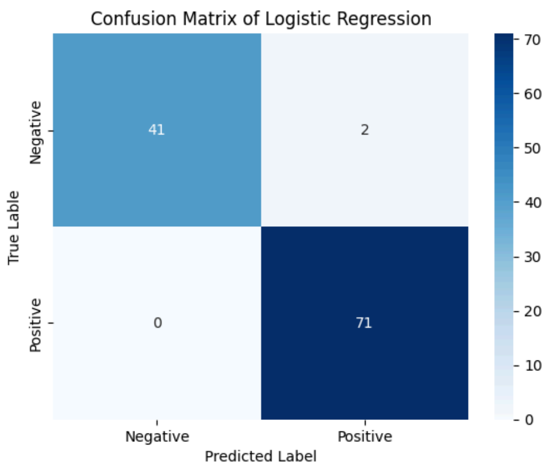

Fig. 4 Confusion matrix of Logistic Regression

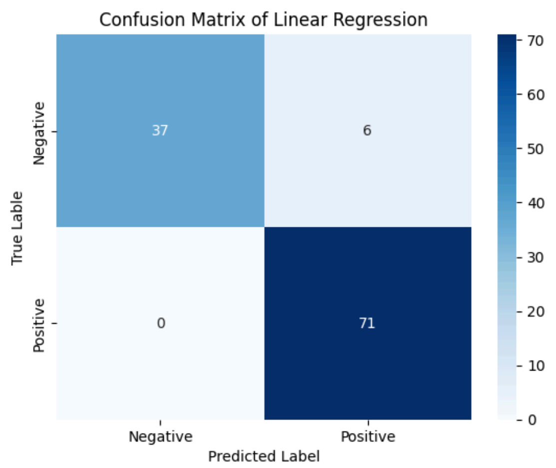

Fig. 5 Confusion Matrix of Thresholded Linear Regression

Table 1 shows the evaluation metrics of Logistic Regression and Thresholded Linear Regression. Logistic Regression has better accuracy, precision, and F1 score compared to Linear Regression. However, both models performed exceptionally well having 100% recall and AUC score.

Figure 4 and 5 shows the confusion matrix of Logistic Regression and Thresholded Linear Regression respectively. Logistic Regression has more True Negatives compared to Linear Regression. Linear Regression has more False Positives compared to Logistic Regression. Lastly, both models do not have False Negatives and have the same number of True Positives.

In terms of accuracy, it is easily understandable that Logistic Regression performs better than Linear Regression due to the classification nature of the dataset. Probabilistic output of Logistic Regression also contributes to its better accuracy.

In terms of precision and recall score, Logistic Regression has better precision but it has the same recall score with Linear Regression. This means that Logistic Regression has fewer False Positives than Linear Regression but both models predicted the same number of True Positives as shown in Figures 4 and 5. It shows that Logistic Regression is only predicting positive outcomes when it's more confident which leads to fewer false predictions. On the other hand, Linear Regression predicts True Positives and False Positives which leads to less accuracy.

Both models possess the same AUC score of 100%. This suggests that both models perfectly discriminate between classes. However, there is a potential issue with both models having the perfect AUC score since this means that there is no misclassification but as seen on Figures 4 and 5, there are still identified False Positives in the prediction.

Lastly, the Confusion Matrix determines the number of True Positives, True Negatives, False Positives, and False Negatives in the predictions of both models. As stated before, both models predicted the same number of True Positives, while Logistic Regression has less False Positives and more True Negatives compared to Linear Regression. This shows that Logistic Regression can predict better with True Negatives in high-dimensional data compared to Linear Regression..