COMP551 - Assignment 2 Report

Abstract

In this project, we investigated the performance of two classification models, linear regression and logistic regression, and we measured the effects of their hyperparameters on two benchmark datasets. To test our models, we followed standard procedures to clean the data, and we implemented and verified the correctness of the models. From our experiments, we found interesting differences in how they assigned importance to features. These differences highlight the distinct ways linear regression and logistic regression manage classification, with linear regression minimizing squared error, and logistic regression optimizing decision boundaries. Ultimately, the logistic regression approach achieved better accuracy than multiple linear regression, though linear regression still produced relatively good results.

Introduction

Our first task was to load, store, and clean the data while following good experimental design practices in order to minimize the chances of a misinterpretation of the performance of our models due to abnormalities in the dataset. The two datasets were carefully selected to test our models over both binary and multi-class classification tasks. The binary dataset is the *Oxford Parkinson's Disease Detection Dataset*[1], a collection of biomedical voice measurements with 22 features to predict whether the patient had Parkinson's disease, and the multi-class dataset, *Wine Recognition Dataset*[2], is a chemical analysis of wines with 13 features, used to predict the type of wine (out of three possible types).

We implemented and compared two classification models: linear and logistic regression. For the binary dataset, we used simple multiple linear regression and logistic regression. For the multi-class dataset, we opted for multivariate multiple linear regression and multi-class logistic regression. Using these datasets, both models were trained, parameterized, and evaluated, through various experiments, in terms of accuracy achieved when predicting labels and Area Under the Receiver Operating Characteristic curve (AUROC). Results were then compared across various configurations to determine the optimal model with the highest predictive accuracy.

To determine the better performing model, the results of the experiments were compared across various configurations to determine the optimal model with the highest predictive accuracy. We concluded that while both models showed good predictive capabilities, logistic regression outperformed linear regression. Ultimately, this finding supports and is consistent with the fact that logistic regression is often better suited for classification tasks, whereas linear regression, though it can be adapted and used for classification, is preferred to predict continuous outputs.

Datasets

We worked with two UCI Machine Learning Repository datasets: the Parkinson's Disease dataset (binary classification) and the Wine dataset (multiclass classification). The Parkinson's dataset consists of various measurements from individual patients, intending to distinguish between those with and without Parkinson's disease. The Wine dataset contains the physicochemical properties of different wines, aiming to classify them into one of three wine cultivars.

To ensure the datasets were well-suited for modeling, we first performed standardization to normalize feature magnitudes. This was essential for logistic regression, which is sensitive to feature scaling, and for improving convergence during optimization. Additionally, we utilized simple linear regression to analyze feature importance, which guided our feature selection process.

Closed-form solutions of simple linear regressions, iteratively computed for each feature separately, served as the basis to determine each feature's correlation to the labels. In order to properly select relevant features, it was important to consider strong feature correlation in both the positive and negative directions. Because of this, we evaluated features on absolute simple linear regression coefficients and average ranking in feature importance ordering.

[1] Little, M. A., McSharry, P. E., Hunter, E. J., & Ramig, L. O. (2008). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 56(4), 1015-1022.
[2] Aeberhard, S., Coomans, D., & de Vel, O. (1992). Comparison of classifiers in high-dimensional settings. UCI Machine Learning Repository.

For the Parkinson's dataset, our analysis revealed that the most relevant features based on simple regression coefficients were spread1, PPE, spread2, MDVP:Shimmer, and MDVP:APQ, while features such as MDVP:Fhi, NHR, and DFA were among the least important. We used a feature selection threshold of 0.3 for standardized coefficients, eliminating low-relevance features to reduce dimensionality and improve model performance. This resulted in exactly the set of aforementioned highly relevant features kept after pruning. For the Wine dataset, who's ternary labels we translated to one-hot encodings, we identified key predictive features based on ranked importance between classes, including Flavanoids, Proanthocyanins, Alcalinity_of_ash, Hue, OD280/OD315_of_diluted_wines, and Proline, all of which were kept for model implementation with the wine dataset. Correlation analysis with class labels further confirmed the relevance of these features. Visualization experiments such as horizontal bar plots for regression coefficients and heatmaps of feature correlations served to further interpret the data structure, ensuring confidence in the model's implementation.

Results

Experiment 1.

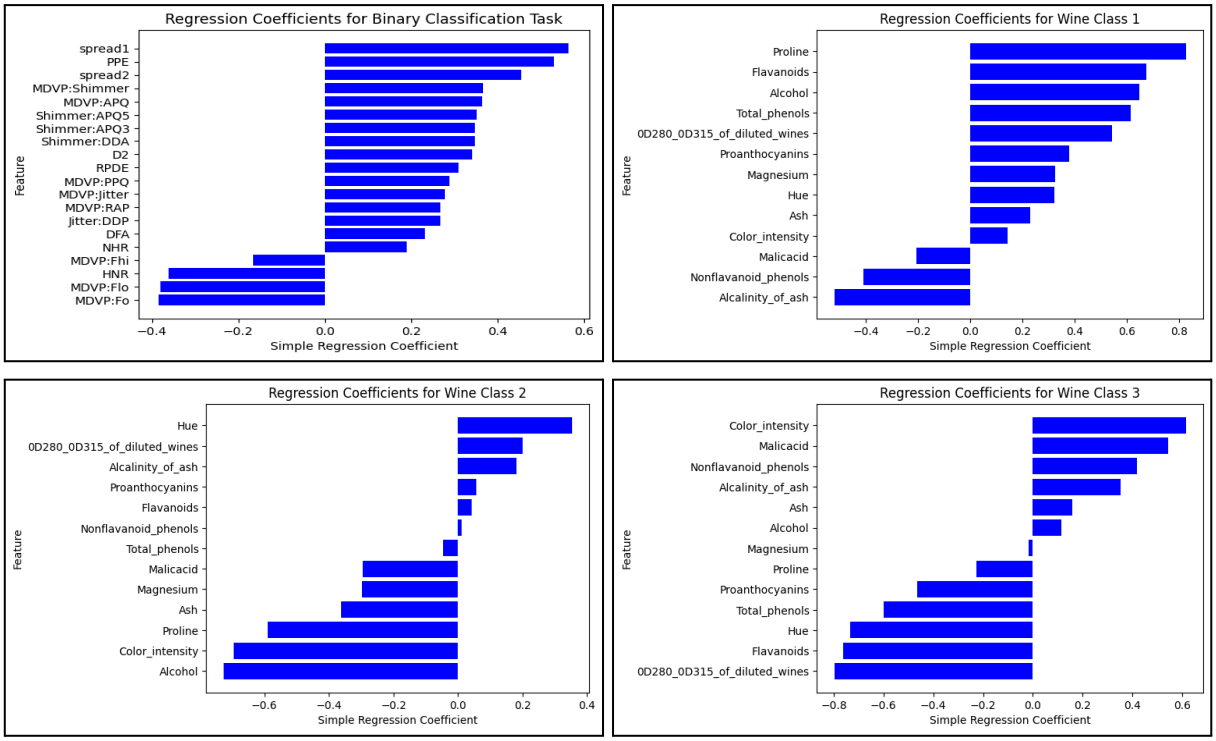

Figure 1-4: Features ranked by similarity to response variable for (from left to right, top to bottom) binary classification, wine class 1, wine class 2, wine class 3 (multiclass classification)

An immediate observation is that there is a healthy amount of variety between the three classes of wine in terms of feature similarity to the response; for example, some features that are most positively related to one class are suddenly most negatively related to the next.

Experiment 2.

```
{'spread1': 1.1803571616014489e-22,
 'PPE': 1.8579645581439362e-23,
 'spread2': 2.387059863735771e-22,
 'MDVP:Flo': 3.586760874125967e-21,
 'MDVP:Fo': 2.339409183845726e-23}
```

```
array([[1.80221026e-17, 7.17667276e-18, 4.96297455e-18],
       [7.93113124e-20, 1.03503754e-17, 1.00746353e-17],
       [4.46639707e-19, 2.24678664e-18, 1.23787650e-16],
       [3.15599311e-20, 2.14092257e-19, 6.57281705e-19],
       [1.01388580e-17, 2.68603399e-18, 1.87844980e-16],
       [7.80509935e-19, 1.13582231e-21, 1.15831148e-20]])
```

Figure 5 & 6: Checking of gradient implementation via a small perturbation. Shown are the differences in analytical and numerical gradients for each feature per class (left: binary, right: multiclass)

To verify our implementation of the gradient calculation for both logistic regression models, we made use of the perturbation method. While the goal stated in the assignment was `10e-6`, our differences greatly exceeded this expectation, with each difference being no larger than `10e-16`, indicating that our gradient calculation was correct and accurate for both logistic regression models.

Experiment 3.

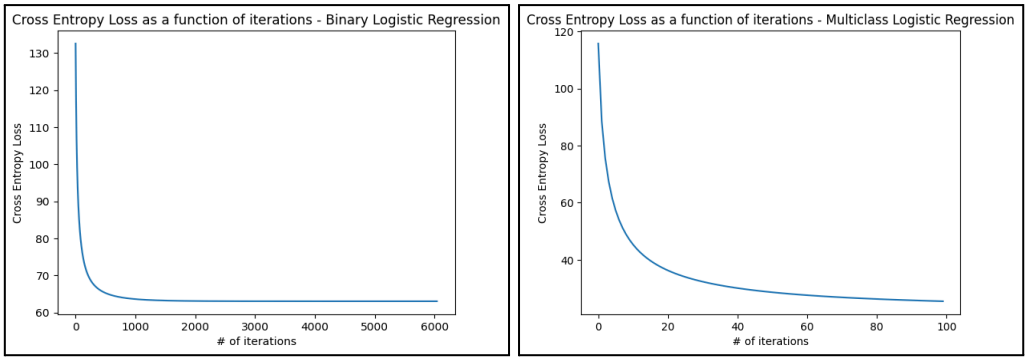

Figure 7 & 8: Convergence plots of binary and multiclass regression over iterations, with learning rates $\alpha = 0.05$ and $\alpha = 0.005$, respectively

We observe that the loss for binary logistic regression converges due to a stopping criterion (norm of gradient attained a smaller value than the parameter epsilon used when initializing the classifier), and takes ~6000 iterations to converge, while the loss for multiclass logistic regression converges due to the maximum number of iterations (100) being reached.

Experiment 4.

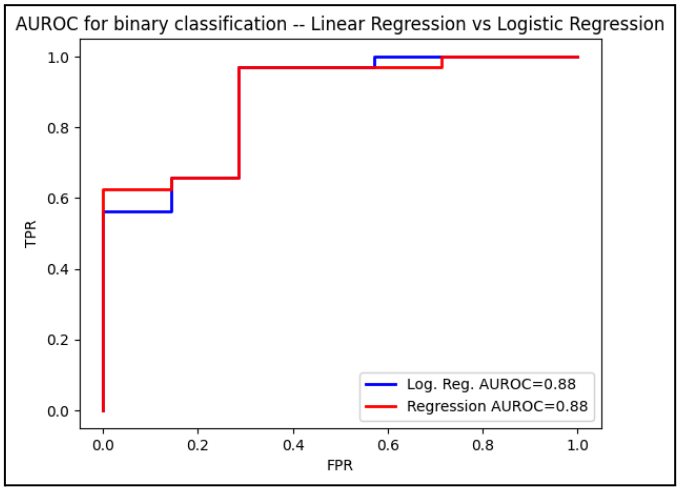

Figure 9: Comparison of AUROC for linear vs logistic regression on binary classification task

We observe that both linear and logistic regression have AUROC scores of 0.88. What's interesting is that when comparing the classification accuracies on the binary task's test set, logistic regression has an accuracy of only 74.36%, whereas linear regression had an accuracy of 82.05%. So although both models could distinguish between positive and negative with the same amount of effectiveness, there is a noticeable difference in accuracy between the two models. This might suggest that at a particular threshold(s), one model struggles more than the other in correctly predicting the response variable, but overall, the two models are comparable in terms of AUROC.

Experiment 5.

Our multiclass logistic regression and multivariate linear regression classifiers scored accuracies of 98.15% and 100% on the multiclass wine dataset, respectively. This is certainly a step up from the accuracies achieved by the logistic regression and linear regression models on the binary classification task. One could speculate that this is due to how the features were selected; for the multiclass case, features were selected based on their average "ranking" of importance for the three classes. On the other hand, for the binary classification dataset, features were selected based on their raw similarity, and perhaps the selected threshold led to potentially informative/relevant features being dropped unknowingly.

Experiment 6.

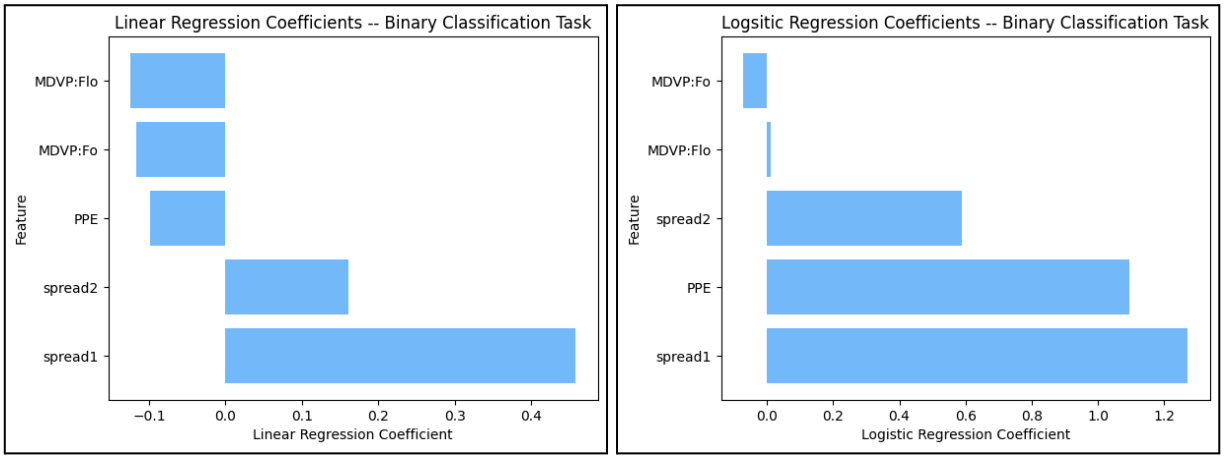

Figure 10: Comparison of coefficient values after fitting on wine dataset. Linear regression (left) vs multiclass logistic regression (right)

From the two horizontal barplots shown above, we can see that although the values of the coefficients that the multiclass logistic regression model arrived at are slightly different to those that the closed form solution of linear regression computes, the relative importance of the features is for the most part, preserved (the only swap being between spread2 and PPE). One point of interest is the fact that PPE receives a negative coefficient in the linear regression model, but a positive coefficient in the multiclass logistic regression model. As to what could have caused this point of difference, it's possible that modifying the stopping criterion of the multiclass logistic regression model to account for something like norm of the gradient or training loss reaching a certain threshold could have helped fix this discrepancy, as the classifier implemented in our notebook only relies on maximum iterations as a stopping criterion.

Experiment 7.

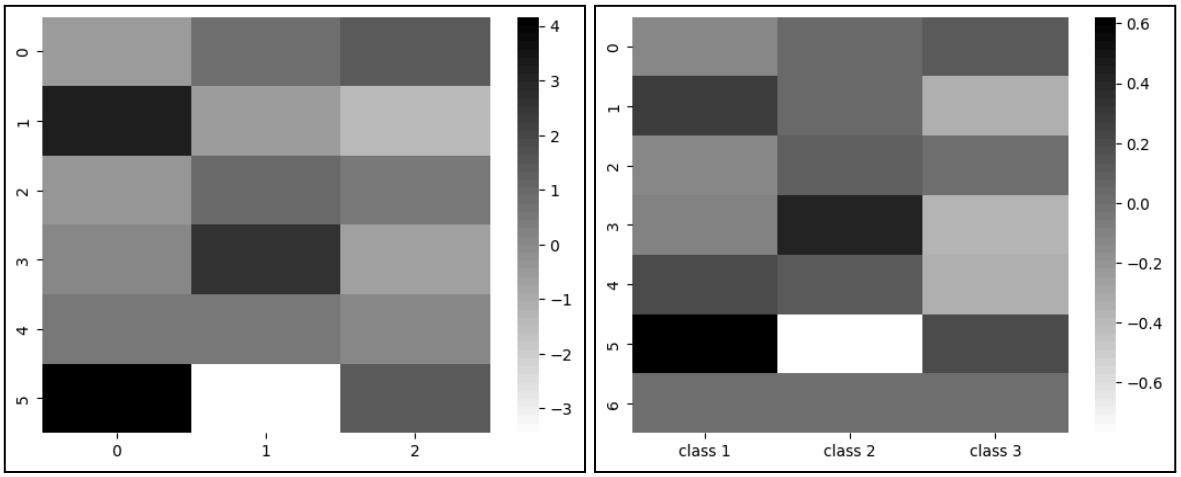

Figure 11 & 12: Side-by-side heatmap comparison of linear regression/true coefficients (left) vs multiclass logistic coefficients (right)

Lastly, we investigated the difference between linear regression and logistic regression in terms of what features both models prioritized as important when computing weights. Although the heatmaps of the two models have different scales, as is shown by the number line on the right-hand side of each respective graph, we can see that the models nonetheless have very similar prioritization of features when observing the colours in each of the respective slots (this is also corroborated by the discussion from Experiment 6). For instance, both models prioritize feature 5 heavily for class 1, effectively disregard it for class 2, and prioritize it a moderate amount of class 3. We can observe a similar pattern for the rest of the heatmap, with only slight deviations in priority (indicated by changes in shade, from white to gray to black) occurring at certain locations within the heatmap.

Discussion & Conclusion

Our experiments demonstrated how different parameter choices influenced the performance of both linear and logistic regression models. We explored the impact of learning rates, gradient convergence, and feature selection across binary and multiclass classification tasks to identify key strengths and weaknesses in each approach. However, logistic regression remains the more efficient model for classification because its loss function, cross-entropy loss, is designed to optimize for categorical target variables, whereas linear regression minimizes squared error, which does not align well with classification tasks. Logistic regression was also found to be more stable in optimization, as seen in the convergence plots. A comparison of regression coefficients between linear and logistic regression revealed discrepancies in feature ranking. In the Parkinson's dataset, logistic regression correctly identified spread1, PPE, and spread2 as the most significant predictors of Parkinson's disease. In current Parkinsons research, this aligns with established medical research on vocal deterioration in Parkinson's patients[1]. Linear regression, on the other hand, assigned greater importance to MDVP:Flo, which has weaker diagnostic significance. This difference arises because linear regression assigns coefficients based on minimizing squared error, whereas logistic regression optimizes decision boundaries in classification tasks. Similarly, in the Wine dataset, logistic regression emphasized Proline, Flavanoids, and OD280/OD315_of_diluted_wines as top predictors, aligning with domain knowledge that these chemical compounds are strong differentiators among wine classes. Linear regression, however, placed higher importance on features such as Alcohol and Malic Acid, which may not be as strong discriminators across all classes[2]. From our study, we conclude that while both models can be leveraged for classification tasks, logistic regression is the more reliable approach due to its probabilistic framework and well-defined loss function (cross-entropy). However, fine-tuning hyperparameters—such as learning rate, iteration limits, and feature selection thresholds—can significantly impact performance.

[1] Little, M. A., McSharry, P. E., Hunter, E. J., & Ramig, L. O. (2008). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 56(4), 1015-1022.
[2] Li, H., Liang, Y., Xu, Q., & Cao, D. (2008). Identification of red wine categories based on physicochemical properties and chemometrics. *Journal of Chemometrics*, 22(5), 399-406.