Assignment 1: Getting Started with Machine Learning

Anson Ng Howard Qin Aybuke Ekiz

February 26, 2024

Abstract

This study provides a comparative analysis of linear classification models, including Logistic regression, MultiClass regression, and decision trees, across two key datasets in natural language processing: the IMDB movie reviews and the "20 news group dataset." By employing linear regression for feature selection, our approach fine-tunes Logistic regression models to enhance predictive accuracy and efficiency. The investigation reveals that while decision trees offer the advantage of lower training times, they lag behind in performance compared to Logistic and MultiClass regression models. Furthermore, the study examines the impact of varied learning rates on the training effectiveness of models, particularly with the "20 news group dataset," highlighting how these adjustments can optimize model convergence and accuracy. The findings underscore the critical balance between training time and model performance, providing valuable insights for optimizing classification models in text analysis tasks.

Introduction

The objective of this project was to evaluate the effectiveness of various linear classification models in the context of natural language processing, specifically through sentiment analysis of the IMDB movie reviews dataset and text categorization of the "20 news group dataset." A key aspect of our methodology was the use of linear regression for feature selection in Logistic regression models, aiming to identify the most significant words for sentiment analysis, thereby streamlining the model to improve both efficiency and accuracy. This study compares the performance and training dynamics of Logistic regression, MultiClass regression, and Decision Trees. It was observed that decision trees, despite their quicker training times, underperformed in accuracy when compared to the Logistic and MultiClass regression models. Additionally, our research delved into the effects of varying learning rates on the models trained with the "20 news group dataset," aiming to discern the optimal conditions for model convergence and precision. Through this comprehensive analysis, we aimed to provide insights into the practical applications and limitations of these models, highlighting the importance of feature selection through linear regression and the strategic adjustment of learning rates in enhancing model performance for natural language processing tasks.

Datasets

Our project utilized two prominent datasets in the natural language processing domain: the IMDB movie reviews dataset for sentiment analysis and the "20 news group dataset" for text categorization. For the IMDB dataset, preprocessing involved reading a comprehensive vocabulary list and loading sentiment-labeled feature vectors. We then converted these vectors from a sparse to a dense matrix format, facilitating the transformation into a pandas DataFrame for further manipulation. A key step in our preprocessing was the filtering of words based on their occurrence frequency across reviews, eliminating those present in less than 1

In contrast, for the "20 news group dataset," we leveraged the CountVectorizer from the sklearn library, applying it with English stop words removal and a feature cap of 500 to prioritize the most significant terms. This was followed by an enhancement step using mutual information to identify and retain only the top-performing features, thereby optimizing the dataset for subsequent classification tasks. This feature selection process was critical in honing our models' focus on the most impactful elements within the text.

Results

Logistic Regression Experiment 1

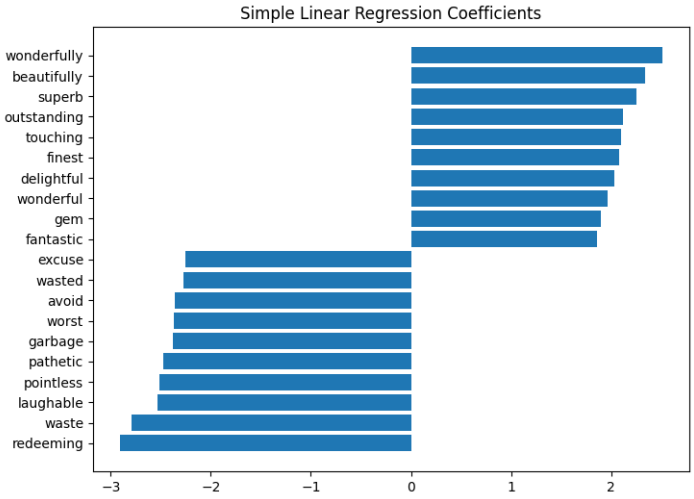

Figure 1: Top 20 words from linear regression

The provided bar chart, titled "Simple Linear Regression Coefficients," displays the top 20 features selected using a simple linear regression model from the IMDB dataset, characterized as the most influential in determining sentiment within movie reviews. The features are split into two categories based on their coefficients: the top 10 with positive coefficients and the top 10 with negative coefficients.

The magnitudes of these coefficients, indicated by the length of the bars, demonstrate the strength of the association between each word and the sentiment prediction. Longer bars for "wonderfully" and "redeeming" suggest a stronger predictive power for these words compared to others. This visualization provides clear insights into which specific words are strong predictors of sentiment in movie reviews, which is invaluable for refining feature selection and improving model performance in sentiment analysis tasks.

Logistic Regression Experiment 2

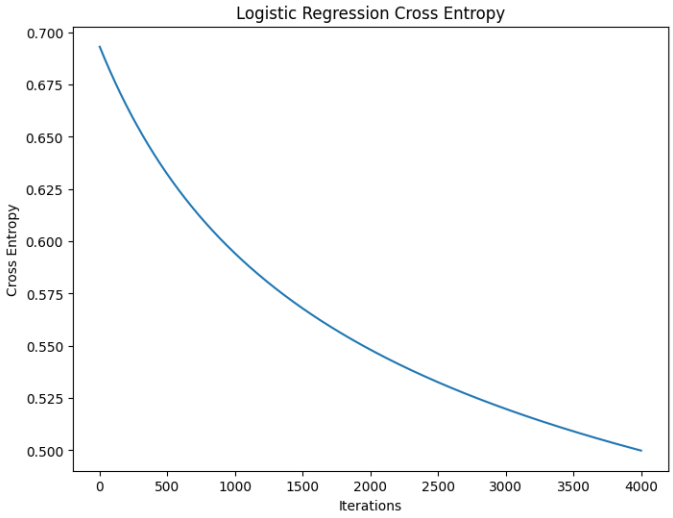

Figure 2: Convergence plot on how the Logistic regression converge

This decreasing trend suggests that the model's parameters are being refined towards an optimal set that minimizes the loss function. However, it is important to note that the curve is still on a downward trajectory at the last iteration presented, which implies that the model has not yet fully converged to its potential minimum loss. The choice to limit the number of iterations to around 4,000 was a trade-off between computational efficiency

and model performance. While a higher number of iterations could potentially lead to a lower loss and thus a better model.

Logistic Regression Experiment 3

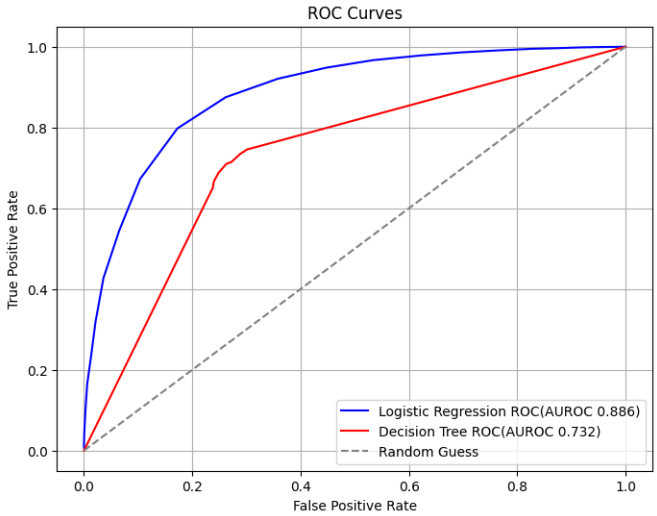

Figure 3: ROC curve comparing Logistic regression and decision tree

The graph displays the Receiver Operating Characteristic (ROC) curves for Logistic regression and decision tree models. The Logistic regression model demonstrates superior predictive performance over the decision tree model for the dataset under analysis, as reflected in the ROC curves and corresponding AUC values.

Logistic Regression Experiment 4

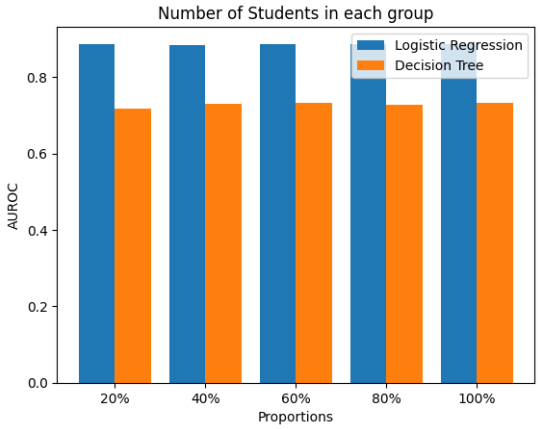

Figure 4: AUROC of Logistic regression and DT on the test data as a function of 5 proportions of training data

The bar plot presents a comparative analysis of the performance, measured by the Area Under the Receiver Operating Characteristic Curve (AUROC), of Logistic Regression and Decision Tree models across different training set sizes. Logistic Regression consistently outperforms the Decision Tree at every level of training data proportion, from 20% to 100%. However, the incremental gains in AUROC for both models seem to plateau beyond the 60% data mark, suggesting that additional data does not significantly enhance the predictive accuracy. The stability of the AUROC across training sizes indicates that both models are relatively robust, but Logistic Regression generally provides a better discriminative power for the classification task at hand.

Logistic Regression Experiment 5

The horizontal bar plot displays the top 20 features extracted from a Logistic regression model trained on the IMDB dataset. The features are individual words whose coefficients indicate their impact on the model's predictions. The ten most positive words, such as "great," "best," and "excellent," are associated with positive

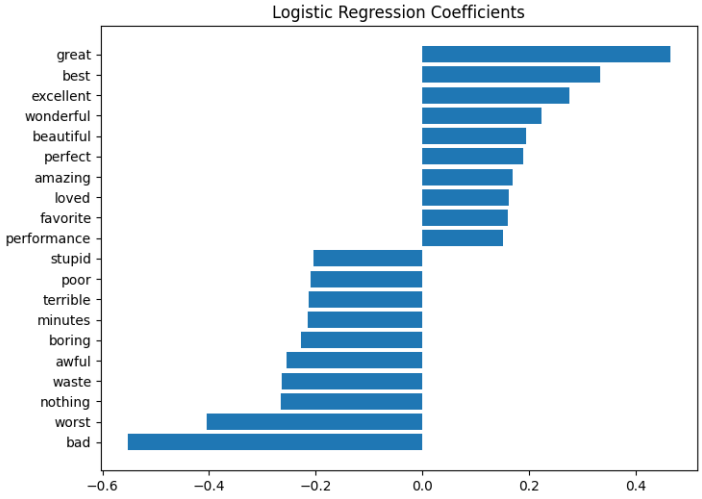

Figure 5: Top 20 words from MultiClass regression

sentiment, having positive coefficients that suggest a strong correlation with positive reviews. Conversely, the ten most negative words, including "bad," "worst," and "nothing," have negative coefficients, indicating a strong association with negative reviews. The magnitude of these coefficients reflects the relative weight or importance these words have in the Logistic regression model for classifying the sentiment of the reviews.

MultiClass regression Experiment 1

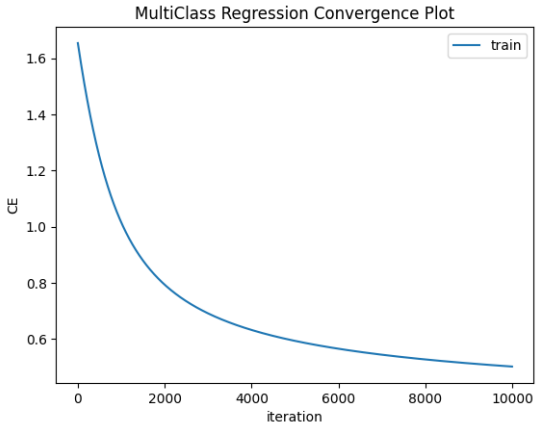

Figure 6: Convergence plot on how the MultiClass regression converge

The convergence plot illustrates the performance of a MultiClass regression model during the training phase. As training progresses through iterations, the Cross Entropy(CE) loss sharply decreases initially, indicating rapid learning, and then gradually levels off, suggesting the model is converging to a minimum loss. The flattening curve as iterations approach 10,000 implies that additional training provides diminishing improvements, which is typical behavior as a model approaches its optimal state. This plot suggests that the model's parameters are being effectively optimized over time.

MultiClass regression Experiment 2

The heatmap visualizes the top 5 most influential features for each of four selected classes from the 20-newsgroups dataset. Each row represents a feature (word or term), and each column represents a class (category). The color intensity indicates the strength of the association between features and their respective classes, with red indicating a stronger positive association and blue indicating a lesser one.

For the 'comp.graphics' class, features such as 'graphics', 'image', and 'program' are most indicative. In the 'misc.forsale' class, terms like 'sale', 'offer', and 'price' are prominent. For 'rec.sport.baseball', words like 'baseball', 'team', and 'game' are most associated. Lastly, the 'sci.med' class is strongly associated with terms

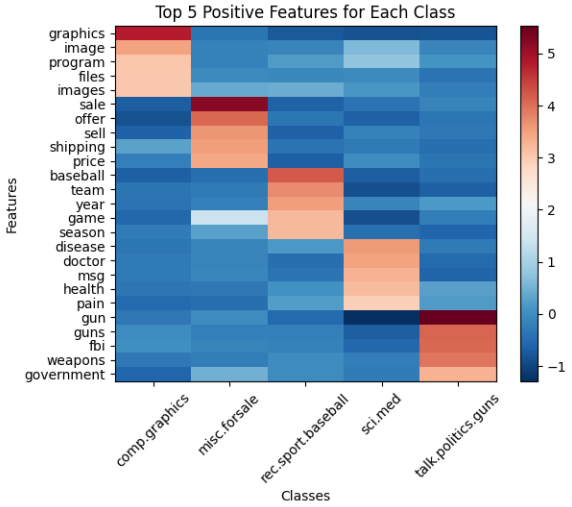

Figure 7: A heatmap showing the top 5 most positive features for each class

such as 'disease', 'doctor', and 'health'. The 'talk.politics.guns' class has strong associations with terms like 'guns', 'fbi', and 'government'.

This heatmap aids in understanding which terms are most informative for distinguishing between the different classes, which is crucial for the feature selection in text classification tasks.

MultiClass regression Experiment 3

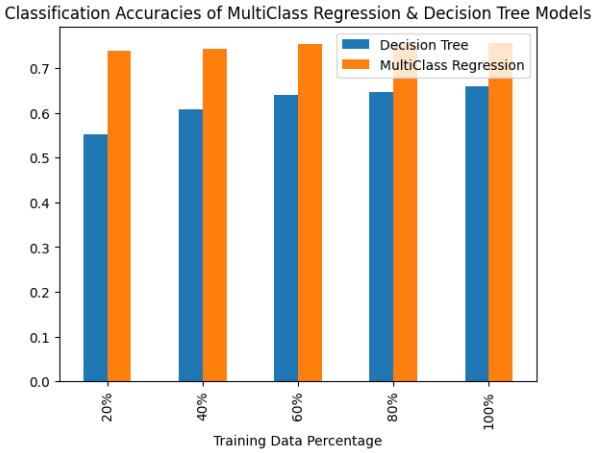

Figure 8: AUROC of MultiClass regression and DT on the test data as a function of 5 proportions of training data

The bar plot compares the classification accuracies of MultiClass Regression and Decision Tree models as a function of the training data percentage used, ranging from 20% to 100%. The accuracies for both models improve as more training data is utilized, with MultiClass Regression consistently outperforming the Decision Tree across all data proportions. The performance increase appears to plateau for both models beyond 60% training data, suggesting that additional data beyond this threshold offers diminishing returns on accuracy improvement. This visualization underscores the effectiveness of MultiClass Regression for this dataset and the importance of the quantity of training data in model performance.

Discussion and Conclusion

The key takeaways from this project are centered around the performance of the MultiClass Regression model across various learning rates and iterations. The experiment demonstrates that there is a nuanced relationship between the learning rate, the number of iterations, and the model's ability to converge to a solution that generalizes

well to unseen data.

The observed association between the key features and their respective classes provides a clear indication that the modela are capturing meaningful patterns within the dataset. For instance, the strong correlation of terms such as 'disease', 'doctor', and 'health' with the 'sci.med' class is intuitively satisfying, as these terms are central to the medical discourse. Similarly, the prevalence of words like 'awful', 'waste', and 'worst' in the negative review are to be expected given the subject matter.

Regarding the learning rate experiment, the results indicate that as the learning rate decreases, the classification accuracy generally increases. This is expected up to a point, as smaller learning rates can lead to more precise convergence on the loss surface. However, the final test accuracy drops slightly when the learning rate becomes excessively small, possibly due to the model's inability to converge within the set number of iterations. Notably, by increasing the number of iterations significantly, the model with the smallest learning rate does eventually converge, achieving a comparable accuracy to the higher learning rates. This suggests that a smaller learning rate can be effective if the model is allowed sufficient iterations to learn.

For future investigation, it would be worthwhile to explore a more dynamic approach to learning rate adjustment, such as learning rate annealing or the use of adaptive learning rate methods like AdaGrad, RMSProp, or Adam. These methods can potentially improve the rate of convergence and the final model performance. Additionally, incorporating more sophisticated regularization techniques might prevent overfitting and further improve model generalizability. Lastly, a thorough analysis of feature importance and selection could provide deeper insights and potentially enhance model accuracy.

Statement of Contribution

Anson: Datasets prepossessing and writing report
Howard: Datasets prepossessing, Logistic model implementation and running experiments
Aybuke: Logistic model implementation and running experiments