Classification of Textual Data

COMP 551 - Assignment 2

Group 102
Sevag Baghdassarian (260980928)
Mike Zhu (261069547)
Rosa Wu (260901912)

Abstract

The assignment consisted of implementing the logistic regression and multiclass regression machine learning algorithms from scratch, applying them on two distinct textual benchmark data sets, analyzing their performance, and evaluating against KNN and BERT algorithm. We found that logistic regression and multiclass regression both performed better than their KNN counterparts, but were significantly slower to train. Further, the BERT model we implemented on the 20 news groups dataset significantly outperformed than the other algorithms implemented on the same dataset.

Introduction

We implemented and fine-tuned the logistic regression and multiclass regression algorithms based on two benchmark datasets concerning IMDB reviews and news groups. We preprocessed both datasets by removing stop words and rare words, generated word embeddings by employing bag-of-words and TF-IDF vectorization, and selected top features based on absolute z-score and Mutual Information (MI) [1, 2]. Moreover, we also compared the performance of our implemented algorithms with KNN, a non-parametric supervised learning method, and BERT, a bi-directional transformer-based architecture achieving state-of-the-art performance for a slew of Natural Language Processing (NLP) problems [3]. Based on our experiments, we found that multiclass regression and logistic regression obtained better accuracies and lower losses than KNN models, even when the hyperparameters were tuned for KNN, but logistic obtained better results than multiclass.

Data Sets

We utilised the IMDB reviews dataset to perform binary sentiment analysis using logistic regression while leveraging the 20 news groups dataset to train a multiclass prediction model and classify 4 news categories, including "comp.graphics", "rec.sport.hockey", "sci.med", and "soc.religion.christian", whose class distributions are demonstrated in Figure 1. We also removed stop words and rare words from both datasets. Besides, we created bag-of-words embedding for the IMDB dataset while using TF-IDF vectorization for the 20 news groups dataset. Further, we selected the top 100 features by the absolute z-score associated with continuous rating scores in the IMDB dataset while choosing the top 50 features based on Mutual Information (MI) in the 20 news groups dataset. Finally, we one-hot encoded our dataset, and further created train, validation, as well as test set.

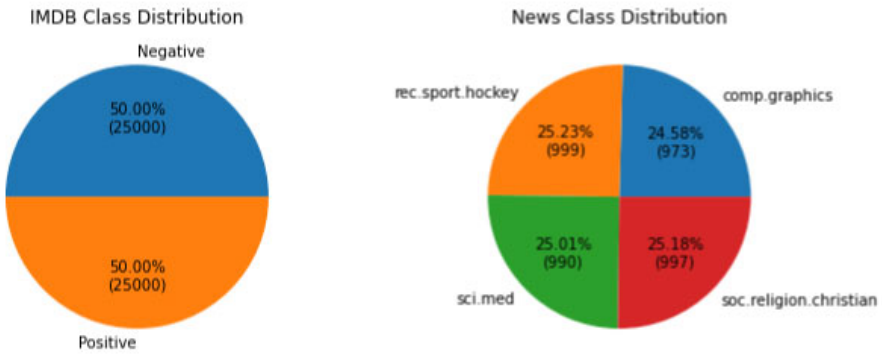

Figure 1: Class distributions of IMDB reviews dataset and 20 news groups dataset

Results

To start off, we evaluated the z-scores of the features of the IMDB dataset using simple linear regression. The features with the most positive/negative z-scores can be found in Figure 2.

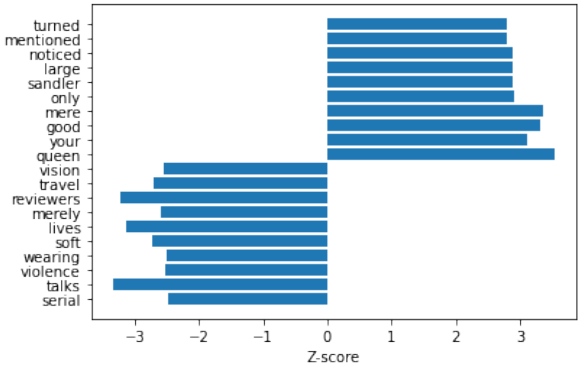

Figure 2: Z scores of the top 20 features from simple linear regression on the IMDB dataset

We implemented logistic regression with mini-batch update. Figure 3 shows a convergence plot of cross entropy as a function of iterations for logistic regression on the IMDB data set.

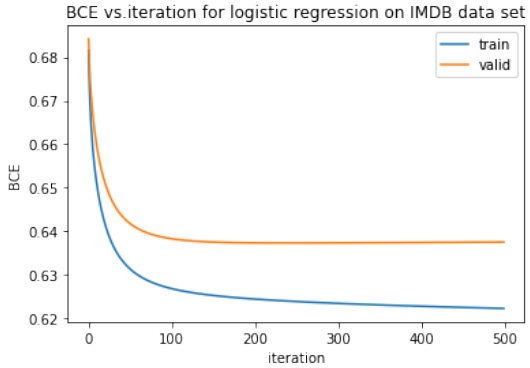

Figure 3: Cross entropy vs. iteration for logistic regression on the IMDB data set with batch size 30 and learning rate 0.01.

Figure 4 shows a convergence plot of cross entropy as a function of iterations for multiclass regression on the news data set. Cross entropy for the validation data started to increase after about 1000 iterations.

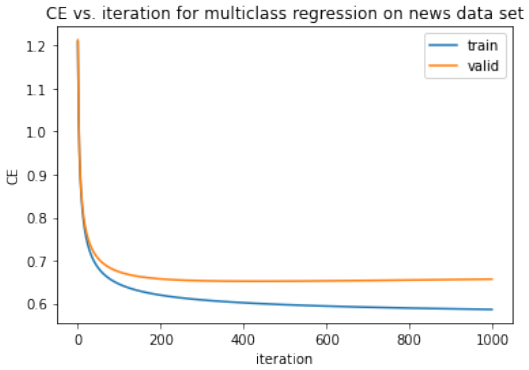

Figure 4: Cross entropy vs. iteration for multiclass regression on the news data set with a learning rate of 0.1

The KNN model for the IMDB test data achieved an AUROC of 0.5709, whereas logistic regression achieved 0.6854. Figure 5 shows a comparison of the ROC curves of both models.

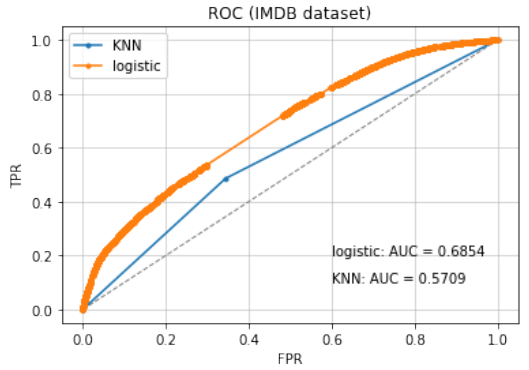

Figure 5: ROC curves for KNN and logistic regression on the IMDB data set.

Furthermore, for all the different percentages of the IMDB data sampled, logistic regression performed superior to KNN. Figure 6 shows a comparison of the AUROCs of logistic regression and KNN according to the different proportions of data sampled.

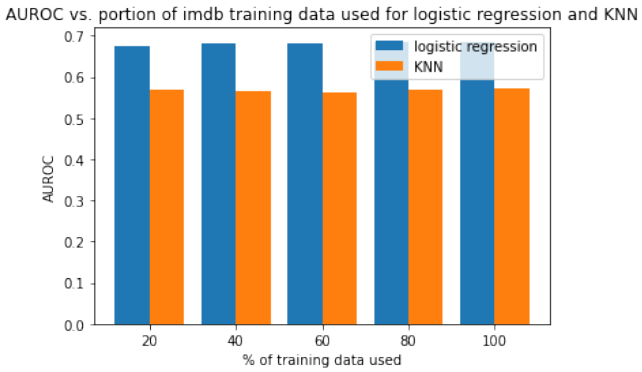

Figure 6: AUROC vs. % of news training data used for logistic regression and KNN on the IMDB data set

As for the KNN model for the news data, we implemented some hyperparameter tuning by validating different k values and different distance functions (namely, Euclidean and Manhattan). Euclidean distance and k=1 proved to result in the highest prediction accuracy. On the news data test set, the KNN model achieved a classification accuracy of 0.648 and a binary CE loss of 8.107, whereas our implementation of multiclass regression achieved an accuracy of 0.693 and a CE loss of 0.734. Figure 7 shows a comparison of the classification accuracies of multiclass regression and KNN according to the proportion of the training data used to train the models. In all cases,

multiclass regression obtained a slightly better classification accuracy than KNN, and the use of lower percentages of the data resulted in slightly lower accuracies for KNN.

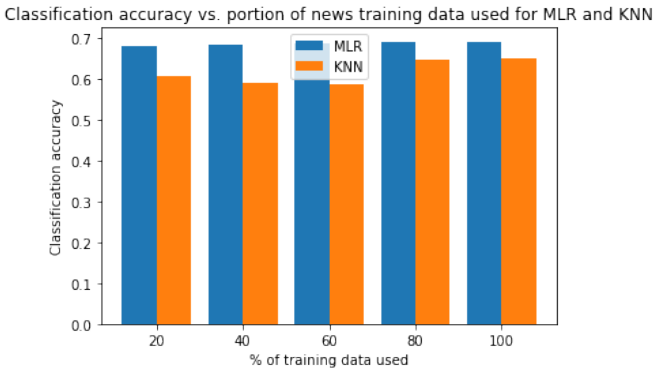

Figure 7: Classification accuracy vs. % of news training data used for multiclass regression and KNN

Figure 8 shows the 10 most positive and the 10 most negative features from the logistic regression on the IMDB data with the coefficient as the x-axis and the feature names as the y-axis

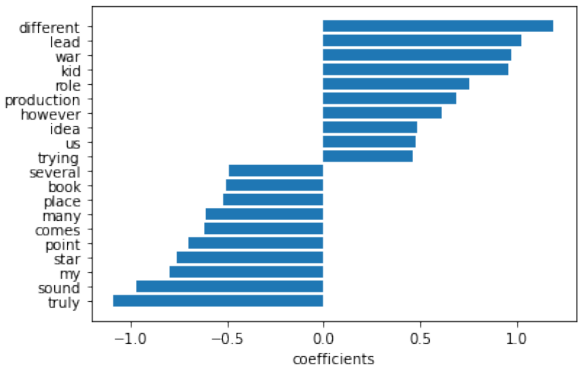

Figure 8: Coefficients of the top 20 features from logistic regression on the IMDB dataset

Figure 9 shows a feature importance heatmap for the news data set.

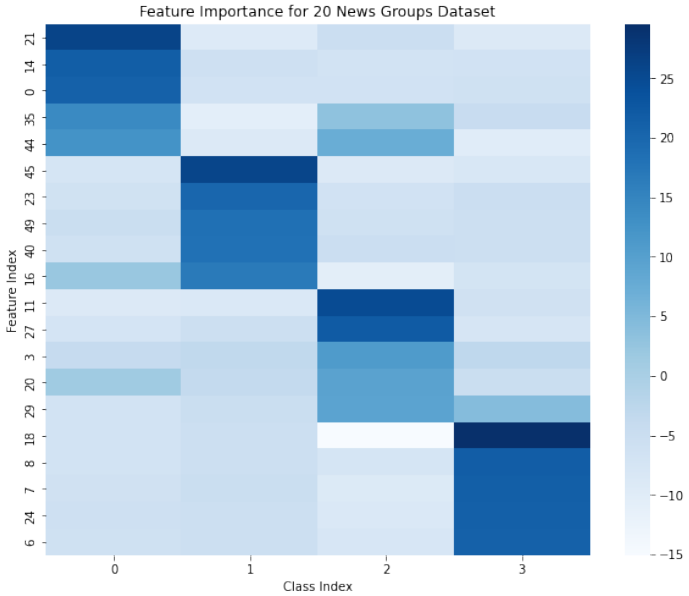

Figure 9: Feature importance heatmap for the news data set.

Last but not least, to evaluate the model's performance on the 20 news groups dataset with the state-of-the-art algorithms, we also utilised a pre-trained BERT model based on uncased English language using a masked language modelling (MLM) objective. We also extended our pre-trained BERT model with an additional 4-unit output layer to classify 4 distinct news categories. We further added a Dropout with a probability of 0.3 to mitigate the overfitting problems. Our BERT model trained on the 20 news groups dataset achieves a loss of 0.213 and an accuracy of 14.889 on the test set, which outperforms the other algorithms implemented on the same dataset.

Conclusion

In conclusion, logistic and multiclass regression performs better than their KNN counterparts. Further, our BERT model also outperforms the other algorithms implemented on the 20 news groups dataset.

Statement of Contributions

Sevag Baghdassarian: Task 2 (Multiclass Regression, KNN), Task 3 and report write-up
Mike Zhu: Task 1, Task 2 (Logistic Regression), creative points (Mini-batch update, BERT model, TF-IDF vectorization), and report write-up
Rosa Wu: Task 2 (Logistic Regression), Task 3 and report write-up

References

[1] C. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, 1948.

[2] J. Leskovec, A. Rajaraman, and J. Ullman, "Data mining," in *Mining of Massive Datasets*, 2003.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019.