Logistic Regression : An Implementation and Analysis

Eamonn Lye

ABSTRACT

The goal of this project was to investigate the performance of logistic regression and multiclass regression models on two datasets. The goal of the logistic regression model was to classify movie reviews as either positive or negative based on their vocabulary. The logistic regression model had 87.9% classification accuracy on test data, and the K-nearest neighbors model had 69.7% accuracy. The multiclass regression model was classifying news articles into categories based on vocabulary. My model uses a cross entropy loss function, in combination with a gradient descent algorithm, to update its parameters. The K-nearest neighbors model correctly classified 54.5% of the newsgroups dataset. Multiclass regression testing accuracy was able to achieve a testing accuracy of 70.9% based on stopping criteria. Thus, the K-nearest neighbors model performed noticeably worse than each of the logistic and multiclass regression models that I implemented. Furthermore, the KNN model had a longer execution time compared to both regression models.

INTRODUCTION

My models were evaluated on two benchmark datasets. One dataset contained movie reviews written on IMDB and the score attributed by the reviewer on a scale from 1 to 10. The goal was to determine whether a given review is positive based on the vocabulary used in the review. The other data set contains 20 distinct types of news articles, of which I chose 4, and the goal is to determine to which news group a given article corresponds. I used a logistic regression model to analyze the first dataset and a multiclass regression model to analyze the second dataset. On the IMDB reviews dataset, the logistic regression model had 87.9% classification accuracy and the K-nearest neighbors model had 69.7% classification accuracy. On the news groups data, the multiclass logistic regression model had 70.9% classification accuracy and the K-nearest neighbors model had 54.5% classification accuracy. These results are similar to observations on the IMDB dataset made by Das, Kamalanathan, and Alphonse, who achieved 90.47% accuracy using a logistic regression model and 56.52% accuracy using a K-nearest neighbors model. Sharma and Dey also analyzed the IMDB dataset, and although a logistic regression model was not used, their K-nearest neighbors model underperformed the other models that they used. Similarly, existing research indicates that a regression model outperforms a K-nearest neighbors model on the newsgroups dataset. Mingyang Jiang et al. (2016) found that implementing a softmax regression model improved testing accuracy by over 20% compared to a KNN model.

DATASETS

One of the datasets I worked with contained IMDB reviews that were categorized as either positive or negative. To start cleaning the IMDB dataset, I first removed stopwords and rare words. Important features were selected by finding the words with the strongest association with numerical rating. The other dataset, 20-newsgroups, contained news articles, which were classified according to the subject of the article. I first chose a subset of these comments, choosing comments relating to computer graphics, baseball, space, and Christianity. First, the data was converted from text to numerical data using sklearn's tf-idf transformer. This allows for the data to be used in a multiclass regression model. Then, from the transformed data, important features were selected using mutual information. After both of the datasets were

cleaned, I determined the class distributions by simply counting the number of samples that belonged to each class. I found that both of the datasets were reasonably balanced.

RESULTS

The discrepancy in classification accuracy for each model and the K-nearest neighbors model indicates that the K-nearest neighbors model is not well suited to handle the datasets that I am analyzing. My logistic regression model achieved an accuracy of 87.9%, compared to only 69.7% by the K-nearest neighbors model. The AUROC curve also shows that my logistic regression model (AUROC = 0.87) outperforms the K-nearest neighbors model (AUROC = 0.7) (Figure 2). As mentioned in the introduction, this discrepancy coincides with observations made by other researchers.

Varying the size of the dataset provides the results that one would intuitively expect. The AUROC of the logistic regression model displays an increasing trend as the proportion of the dataset on which the model is trained increases (Figure 3). As for the multiclass regression model, two trends can be observed. First, the training accuracy decreases, and second, the validation and testing accuracies increase (Figure 4) as the proportion of the dataset used increases. This corresponds with general principles of modeling, where a small training dataset tends to result in an overfit model since it has not been trained on data that is truly representative of the population.

Figure 6 displays the words from the IMDB dataset with the greatest Z-scores and these words, by inspection, appear to be good predictors of whether a given review is positive.

Figure 7 also displays the top 20 words ranked based on their W-coefficients where the top 10 most positively and negatively weighted words are shown in order. Although different from the words shown by figure 6, the words do seem to be good indicators of whether the sample should be classified as positive or negative.

The heatmap for the most positively associated features with each of the news group classes demonstrates the differences between the words associated with the classes. The ‘Christian’ group is strongly associated with words that you would intuitively associate with the category, such as ‘bible’, ‘jesus’ and ‘god’. On the other hand, the most positive features for the ‘medicine’ category are ‘get’, ‘such’, ‘edu’, ‘too’ and ‘most’, which are all words that are not particularly associated with this category. The other categories, ‘computer graphics’ and ‘hockey’, have positive features more in line with the ‘Christian’ category. I believe this may be the result of the ‘medicine’ category being strongly associated with words that were filtered out along with the rest of the rare words.

As for the stopping criteria of the multiclass regression model I explored 3 different options, first with a maximum iteration of 5000 which resulted in an accuracy of 70.4%, then I used the stopping criteria of stopping at the optimal iteration, in which I stop at the iteration with the lowest cross entropy loss for the validation set at iteration 3706, which resulted in a testing accuracy of 70.9%, lastly, I also used a early stopping method which saved a window of the latest cross entropy loss for the validation set, and once the average of that window is lower than the newest cross entropy loss obtained from the validation set, the model is stopped, this method resulted in a accuracy of around 70.8% which is similar to using the optimal iteration stopping method as the two methods are very similar as they both try to stop around the iterations with the lowest entropy loss. However, one advantage the early stopping method has

over the optimal iteration method performance wise is that for the early stopping method the model does not need to be re-initialized with the optimal iteration conceived from the last time the model is ran resulting from the model only having to fit itself to the training data once instead of twice.

Different learning rates were also experimented with for the multiclass regression model which resulted in an optimal learning rate of 0.001 where higher values resulted in a lower accuracy and lower values resulted in the model not converging within a suitable amount of iterations while insignificantly improving the models accuracy.

Other models such as LASSO and ridge regression were also experimented with due to the open-ended nature of the assignment in which the sklearn implementation was used. As a result of their implementation, both LASSO and ridge models provide a performance either similar or better on both the IMDB and the newsgroups data sets, with the LASSO regression model receiving a test accuracy of 86.3% on the IMDB data set and 70.9% on the news data set, while the ridge regression model received a test accuracy of 86.3% on the IMDB dataset and 62% on the news dataset.

The convergence plot for the multiclass regression can also be seen in Figure 1a, and as the number of iterations increases two trends can be observed. First of all, both validation and training cross entropy decrease at a slowing rate. Secondly, Training cross entropy decreases faster than validation cross entropy. This aligns with the theory of gradient descent and cross entropy, and indicates that I used an appropriate learning rate. For multiclass regression, the correctness of the gradient was also verified through both the small perturbation method, which resulted in a value of 6.48e-07, and through the monitoring of loss of cross entropy at each iteration, and this can be seen in figure 1a. The correctness of the logistic regression was also verified through the monitoring of the loss of the cross entropy which can also be seen in figure 1b.

DISCUSSION

On both benchmark datasets, the K-nearest neighbors model performed worse than the logistic and multiclass regression models. This result reflects what was seen in class, with K-nearest neighbors models performing poorly on high-dimensional data. Future investigation into analysis of these kinds of datasets and models should be focused on the data itself, rather than the models, as the choice of appropriate key features is integral to getting high accuracy. However, this is more of a statistics problem than a machine learning problem. For the newsgroups dataset, despite using TF-IDF rather than pure counts or proportions, the class ‘medicine’ still contained unrelated words as key features. Furthermore, despite the other classes containing more relevant words, my model’s accuracy was negatively impacted by this class’ unrelated important features. Thus, obtaining the words that are actually the most indicative of a given class immediately provides a major boost to a regression model’s accuracy.

APPENDIX

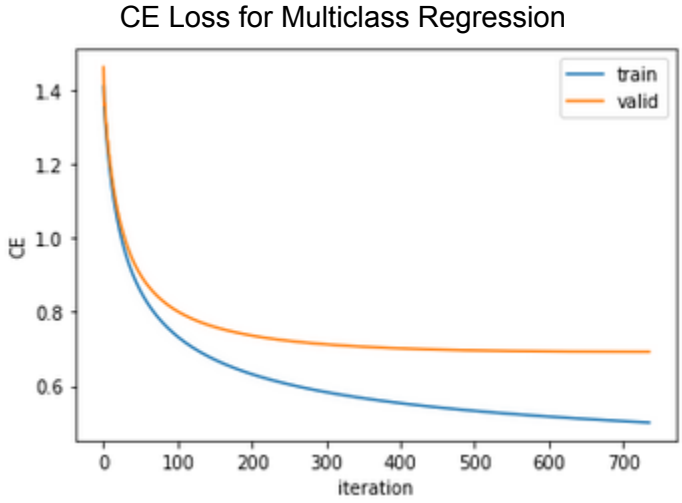

Figure 1a Cross entropy loss for multiclass regression

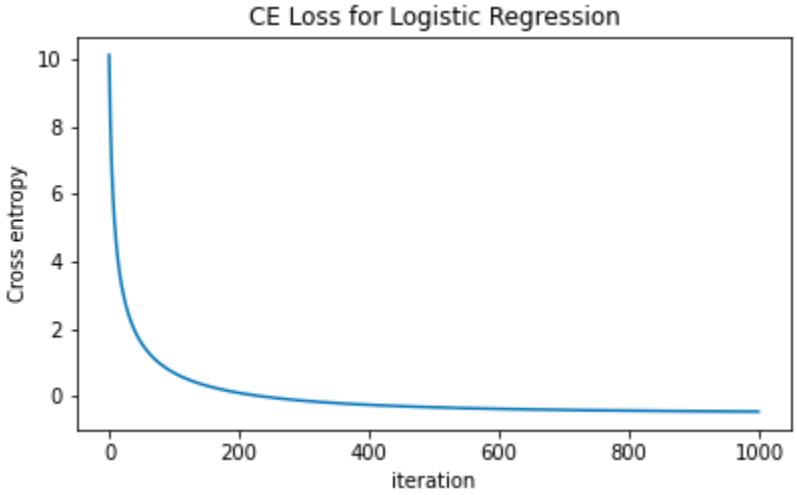

Figure 1b Cross entropy loss for logistic regression

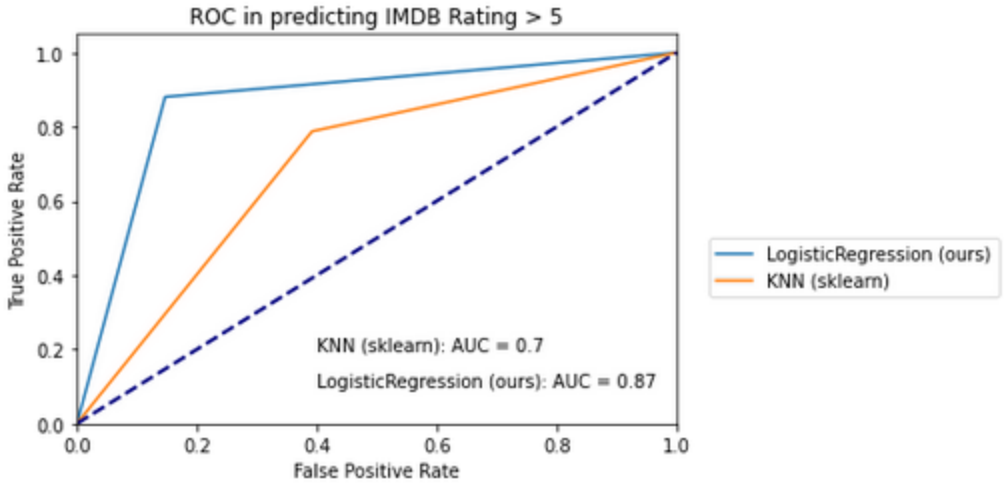

Figure 2 AUROC for KNN and logistic regression models

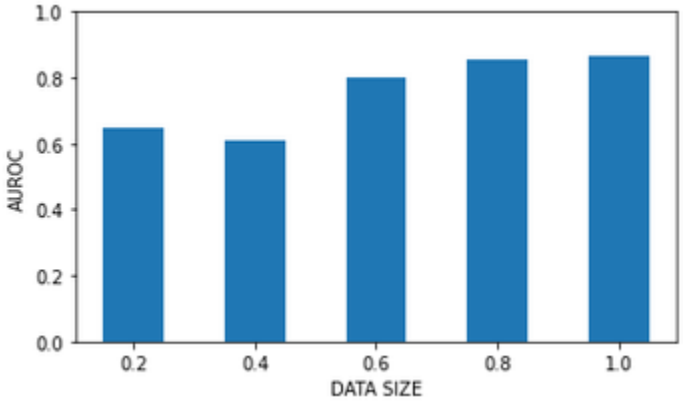

Figure 3 AUROC of logistic regression as a function of proportion of dataset used

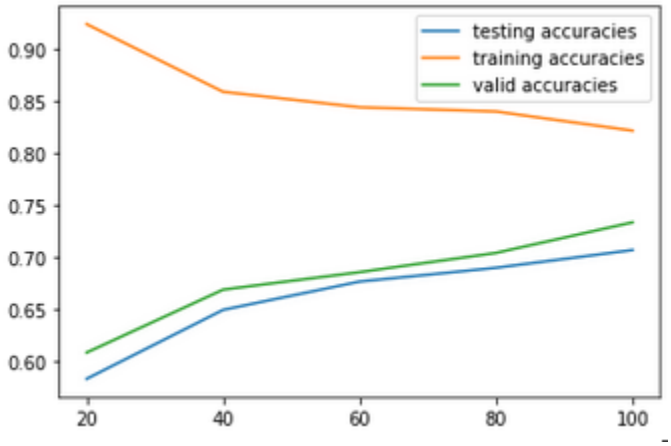

Figure 4 Multiclass regression as a function of proportion of dataset used

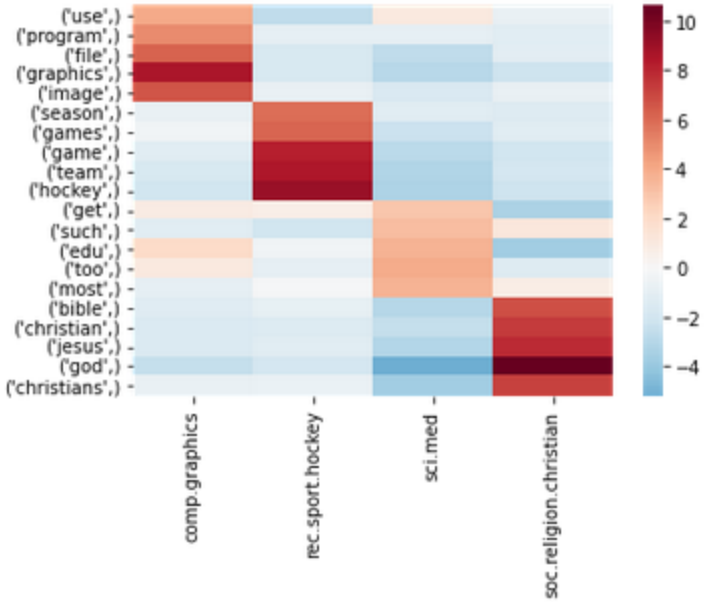

Figure 5 Heatmap of 5 most positive features for each news group

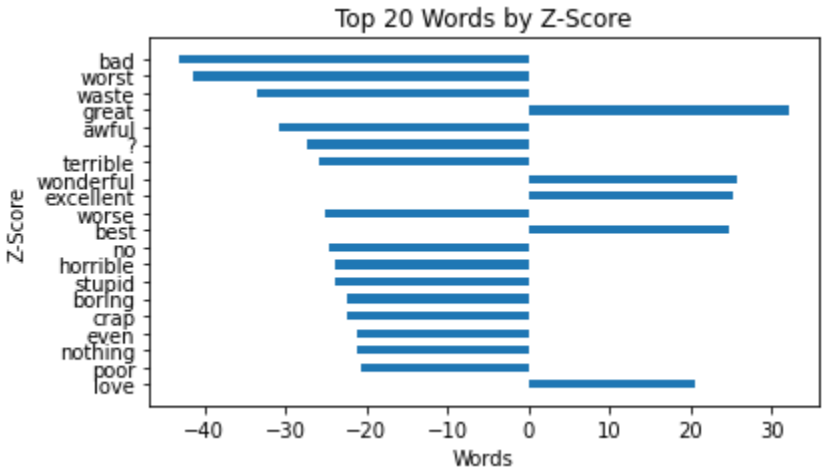

Figure 6 Top 20 Words by Z-score in the IMDB dataset

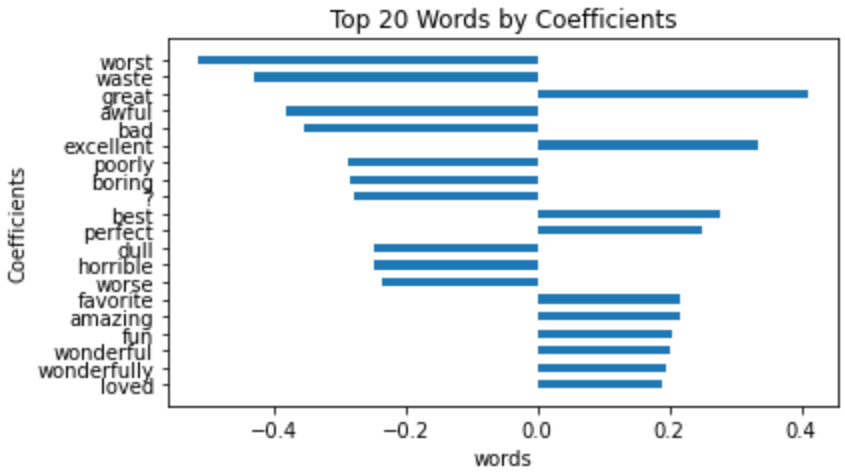

Figure 7 Top 20 Words by W-Coefficients in the IMDB dataset

REFERENCES

Das, Mamata et al. “A Comparative Study on TF-IDF Feature Weighting Method and Its Analysis Using Unstructured Dataset.” *COLINS* (2021).

Jiang, M., Liang, Y., Feng, X. *et al.* Text classification based on deep belief network and softmax regression. *Neural Computing & Applications* **29**, 61–70 (2018). https://doi.org/10.1007/s00521-016-2401-x

Lemoine, Julien, Benhadda, Hamid, Ah-Pine, Julien. Classification non supervisée de documents hétérogènes: Application au corpus ” 20 Newsgroups ”. 11th Information Processing and Management of Uncertainty Conference (IPMU 2006), Jul 2006, Paris, France. Hal-01504419

Sharma, Anuj, and Shubhamoy Dey. "A comparative study of feature selection and machine learning techniques for sentiment analysis." *Proceedings of the 2012 ACM research in applied computation symposium*. 2012.