1 Abstract

In this assignment, we were tasked with implementing both logistic regression and multi-class regression machine learning models and test their binary classification and multi-class classification performance against K-Nearest Neighbors respectively. To test the performance of binary classification of logistic regression vs. KNN, the models were tasked with performing sentiment analysis on IMBD movie reviews to determine if the reviews were either positive or negative. To test the performance of multi-class classification of multi-class regression vs. KNN, the models were tasked with classifying specific newsgroup posts into 4 specific categories. In both tests, we found that logisitic regression and multi-class regression significantly outperform KNN.

2 Introduction

The task we were given for this assignment was to (i) acquire, clean, and preprocess data from both the IMDB dataset and the 20-news group dataset, (ii) implement both logistic regression and multiclass regression machine learning models from scratch, (iii) conduct binary classification on IMDB reviews and multi-class classification on 20-news group datasets, (iv) compare AUROC and classification accuracy with K-Nearest Neighbors model from *sklearn*, (v) report the top features found from in datasets, (vi) and compare the accuracy of the two models as a function of the size of the training data. We found that logistic regression significantly outperformed KNN on binary classification of positive/negative IMDB movie reviews with an AUROC of 0.8779 compared to KNN's AUROC of 0.6343. Logistic regression outperformed KNN with controlling the training size for 20%, 40%, 60%, and 80% of the data as well. We also found that multi-class regression outperformed KNN on multi-class classification on the 20-news group dataset with a classification accuracy of 74.80% compared to KNN's best classification accuracy being 46.12% (with k=1, KNN performed $\sim 10\%$ worse using the sklearn default of k=5). The IMDB dataset was initially introduced by researchers at Stanford in their paper *Learning Word Vectors for Sentiment Analysis* in 2011, and has since been used as a benchmark dataset in several papers of testing sentiment analysis performance of specific machine learning models. The 20-news-group dataset is also used alongside the IMDB dataset for testing text classification tasks on a variety of machine learning models. For example, in *Semi-supervised Sequence Learning*, two Google researchers used both 20-news group and IMDB datasets to test their recurrent neural network (RNN) model's text classification performance by modeling sequential data.

3 Datasets

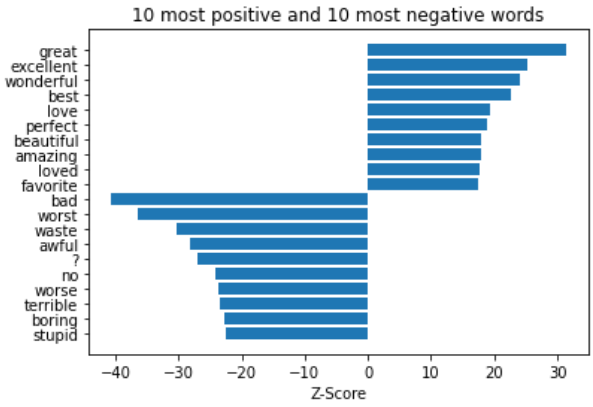
The IMDB datasets contains 50 000 movies reviews (25,000 test samples and 25,000 training samples) of balanced distribution (25,000 positive reviews and 25,000 negative reviews). We use logistic regression to perform binary classification on this dataset. We first were tasked with removing the rare words (which are words that appear very infrequently, in less than 1% of documents) as wells as stop words, words that appear in more than half of the datasets. Second, we selected the top 100 features according to their absolute z-scores associations with the continuous rating scores. The z-scores were obtained by using Simple Linear Regression Hypothesis testing. By doing so we use the continuous rating score to obtain the necessary features.

$$z = \frac{\hat{w}}{\sqrt{Var[\hat{w}]}} = \frac{1/Nx^t y}{\sqrt{1/N}} = \frac{x^t y}{N} \sim \mathcal{N}(0,1).$$

This was the formula used to calculate the z-scores
The news dataset from scikit learn contains 18,000 newsgroup posts on twenty topics that are split into training and testing subsets. Same as in the IMDB dataset, we filter out rare words and stop words. Further, we chose *rec.motorcycles*, *soc.religion.christian*, *sci.med* and *comp.windows.x* for our four categories, from the twenty available. For the feature selection, we used one-hot encoding to convert our categorical data variables into binary vectors. This is done in order for our models to process the data.

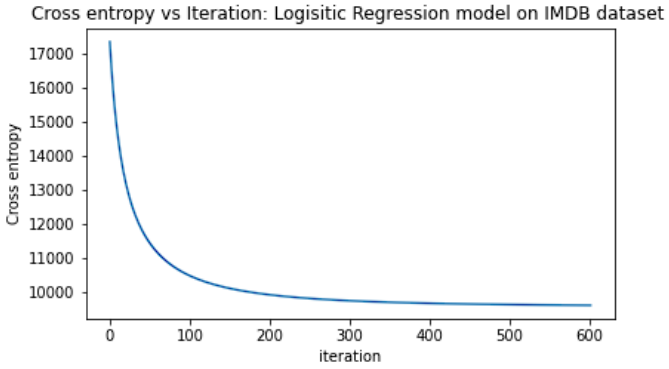
4 Results

(1) After using the simple linear regression hypothesis testing, we found the following z-scores for the ten highest and lowest words.

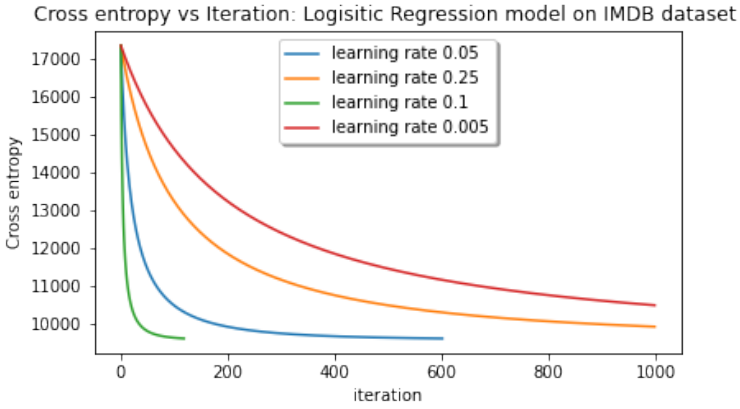


From the above graph it can be observed that the z-scores accurately depict whether a movie is good or bad.

(2) The learning curve of logistic regression. We can see that it converges given our chosen learning rate of 0.05. We made sure to first hold out half of the training data as a validation set to find the optimal max-iteration number (602 iterations) to retrain the model on to avoid overfitting.

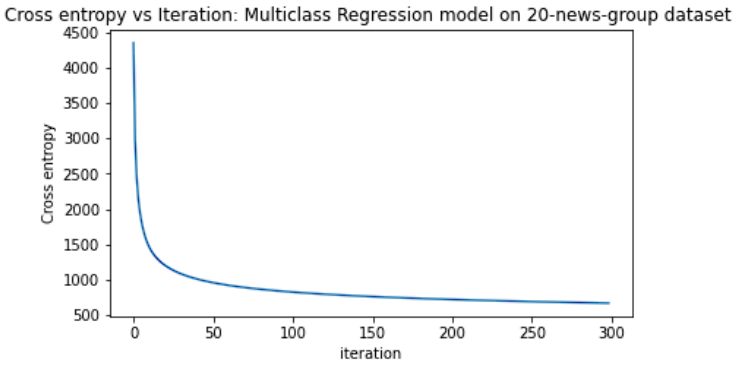


The learning curve of logistic regression for different learning rates. The different learning rates used were 0.05(the original learning rate used to train the model), 0.1, 0.25, 0.005. The models were trained twice for each different learning rate used. First with a validation set to obtain optimal max-iteration number(602,999,121,999) respectively. Then the models were retrained using the latter numbers to avoid over fitting.



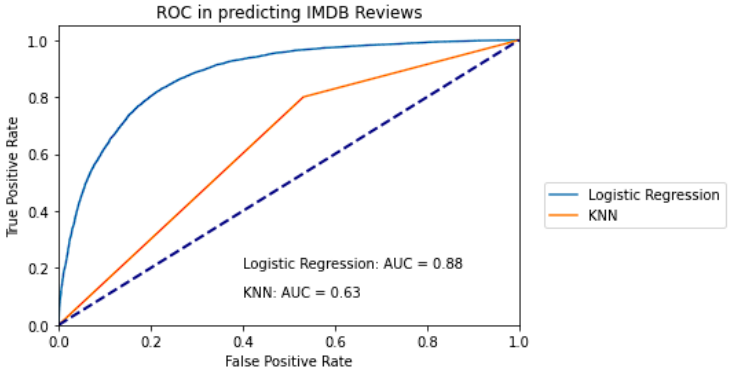
Similarly for multi-class regression, we see that the cross entropy converges for our learning rate of 0.0005. Once again we made sure to train the model twice, once using a validation set to find the

optimal max-iteration number (299 iterations) and retrained the entire model using that number to avoid overfitting.



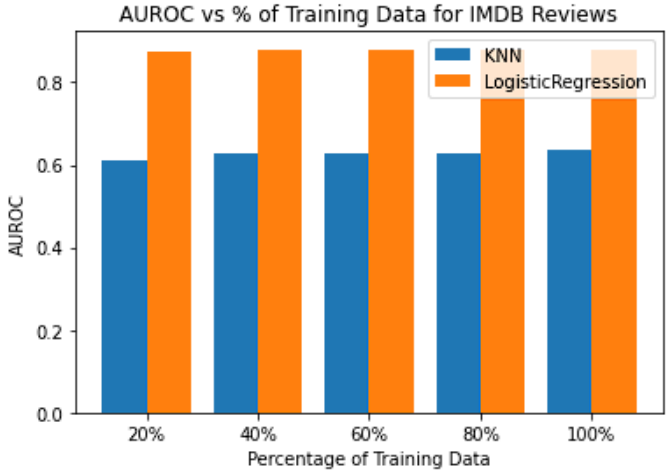
We checked a the gradient of both models using small perturbation and both our models yielded very small results (1.08e-11 for logistic regression and 1.52e-09 for multi-class regression).

(3) The AUROC curves of both the logistic regression and KNN on the IMDB dataset



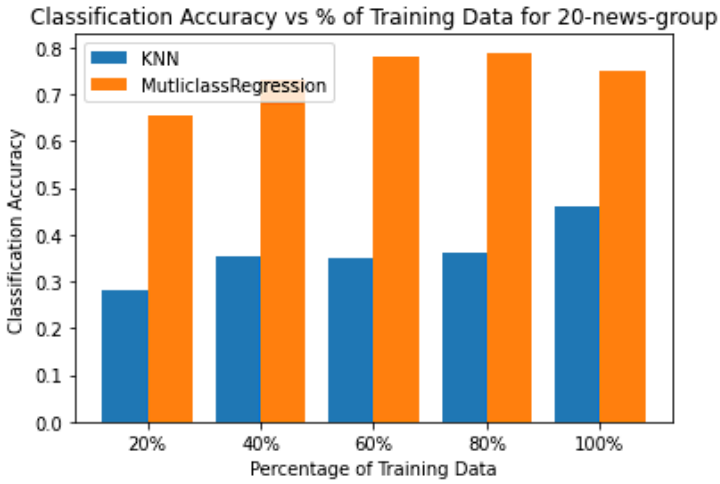
From the graph, we see that logistic regression performs better than KNN on the IMDB dataset, with a score of 0.88 vs 0.63.

(4) The AUROC of logistic regression and KNN as a function of the percent of training data used



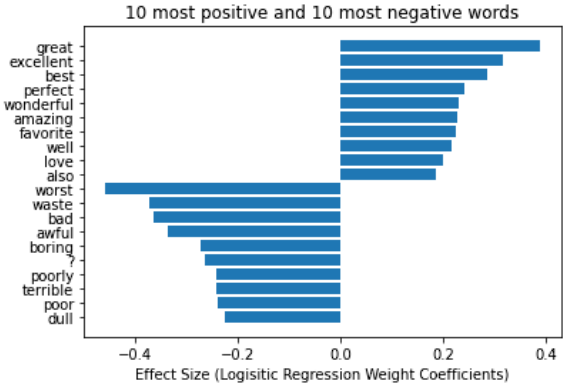
In the above graph. It can be observed that the logistic regression model is better at distinguishing between positive and negative classes then the KNN model, possible due to the limited amount of training data.

(5) The AUROC of multi-class regression and KNN as a function of the percent of training data used



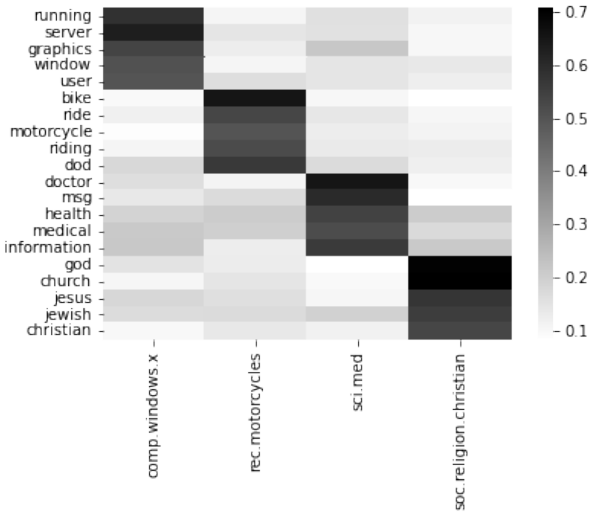
We can see that in the news data set, KNN performs very poorly when there is a limited amount of training data, whereas the performance of multi-class regression stays somewhat consistent throughout.

(6) After applying logistic regression, we found the top 20 features.



In the above graph we observe that the features obtained from logistic regression weight coefficients are similar to the ones obtained from z-scores. Thus it is accurately selecting features that demonstrate whether a movie is good or bad.

(7) A heatmap of the top five most positive features for every of the four categories



5 Discussion and Conclusion

In the case of logistic and multi class regression, the performance of both methods exceeds that of a simple classifier, such as KNN, especially when the number of features is very high. Consequently, the complexity requirement of the model must match. During the experiments, the prediction rate of the regression method greatly exceeds that of KNN, especially when the size of the training set is low. For future investigation, it would be interesting to apply some regularization method such as LASSO, ridge, or elastic net regression and see if the performance of our models would increase due to reduction in overfitting. It could also be interesting to introduce Adam (Adagrad + Momentum + RMSprop) to our gradient descent approach and see how that changes the performance as well.

6 Statement of Contribution

- William Huang (260972252) - Multi-class regression, report
- Dgebe Nicoals(260867207) - report,part of expirements
- Willie Habimana (260987793) - Extracting and preprocessing datasets, logistic regression, experiments, report