COMP551 Assignment 2

An Ni Xu, Jia Song, YiWen Zhang

1 Abstract

In this assignment, we will analyse two regression techniques seen in class: simple and multi-class logistic regressions on two textual datasets. We first run simple linear regression to select the most important features for each of the two datasets, then use one-hot-encoded data as input to each logistic regression models. To evaluate our data, we will be comparing the label classification of our models to the decision tree technique model from built-in Python library Sklearn. The goal is to compare the performance of simple and multi-class logistic regressions' classification with decision tree classification. We have concluded that simple logistic regression and multi-class logistic regression both perform better than decision tree model in classification.

2 Introduction

This assignment consists of fitting simple logistic regression on IMDB Review data and multiclass logistic regression on 20-new-groups data. Both datasets have training and testing subsets which will be used for training and evaluation correspondingly.

- IMDB Review dataset:
  - The labels "sentiment" consists of "pos" (positive or 1) and "neg" (negative or 0). Its training and testing subsets (2500 data points each), which are textual movie reviews.
  - A study has been done using this dataset: Learning Word Vectors for Sentiment Analysis, which has goal to capture semantic similarities between words by the use of a mix of unsupervised and supervised learning.
- 20-news-groups dataset:
  - It is generated by Sklearn module in Python, which comprises 18846 newsgroups posts as data points on 20 topics, where topics are the target labels.
  - A semi-supervised classification study has been used on this dataset. The results showed the F1-scores of the experiments for different training percentage and models.

To evaluate the performance, quality and accuracy of the regressions, we will compute ROC curves to find AUROC scores for binary classification, and classification accuracy for multi-class classification. For investigation purpose, we will compare our model's binary and multi-class classifications with Sklearn's decision tree classifications. In the end, we found that logistic and multi-class logistic regressions approach achieved better accuracy than Decision Tree approach.

3 Datasets

For both of our datasets, we use the ”train” subsets - with different percentages of random data selection : 20%, 40%, 60%, 80%, and 100% - for model training and perform performance evaluation on ”test” subsets. Filtering out irrelevant features is an important step to ensure quality regression fits and to avoid the runtime being too slow. Feature selection for IMDB dataset will be based on simple linear/logistic regression coefficients : those with highest absolute coefficients contributes more and thus should be selected. For 20-news-groups dataset we will use mutual information score from Sklearn to select the top 10 features/words per categorical class. Furthermore, we will only use the following five class labels out of the 20 : comp.windows.x, talk.politics.guns, rec.sport.baseball, sci.crypt, soc.religion.christian, talk.politics.guns, which are relatively different from one and another and is an easier combination to use for analysis and debugging.

4 Results

4.1 Regression Convergence Plots

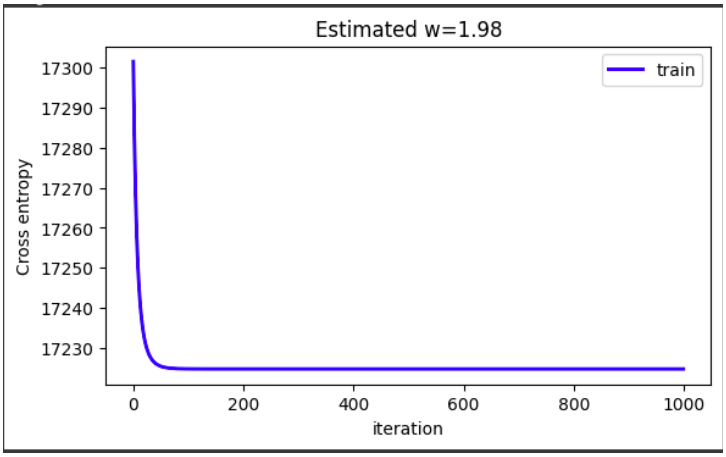

Figure 1: Simple Logistic Regression Convergence

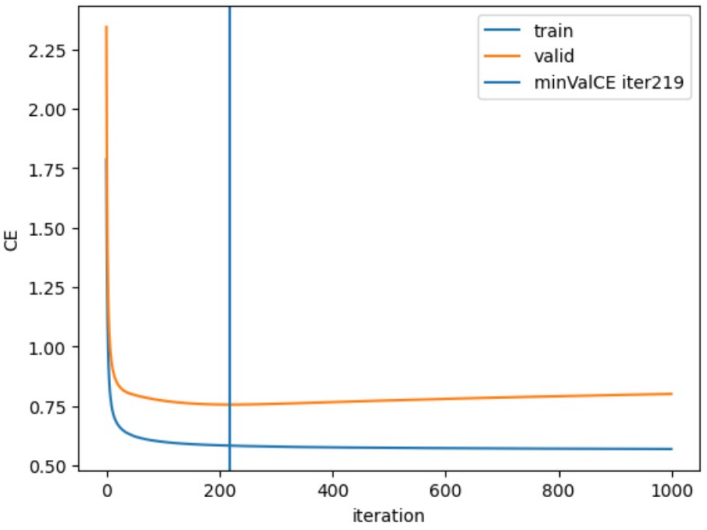

Figure 2: Multi-Class Logistic Regression Convergence

In both simple logistic and multi-class regression errors convergence plots, the training error(cross entropy) curves continues to converge, which implies that gradient has been verified. Additionally, we could verify gradient by checking small perturbation, which should be approximately 10e-8 or even smaller. Otherwise, gradient calculation and/or loss function is incorrect. Furthermore, we can also verify for sign of overfitting by checking for potential increase at a specific point in validation error.

4.2 IMDB dataset analysis

4.2.1 Top 20 IMDB Features

From the bar plots below, we can see that the top features are quite different from the two types of logistic regressions, and when there are common features, their coefficients are also different. However, the features/words that have very positive and very negative coefficients all make sense. In other words, features/words that are used to describe good and bad movies are the suitable ones to use.

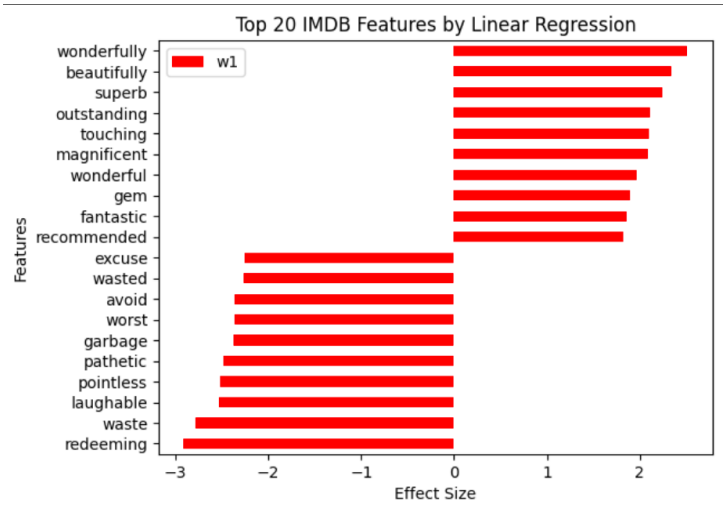

Figure 3: Top 20 IMDB Features by Simple Linear Regression

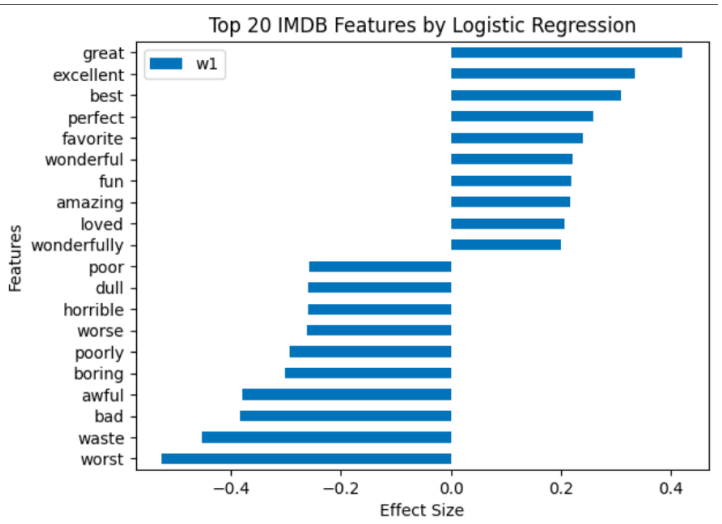

Figure 4: Top 20 IMDB Features by Logistic Regression

This phenomenon is due to the fact that logistic regression uses logistic function to transform class label into classification probabilities to perform regression, while linear regression directly uses target class labels for model fitting and prediction.

4.2.2 ROC curves (with Decision Tree's) on IMDB Test Data

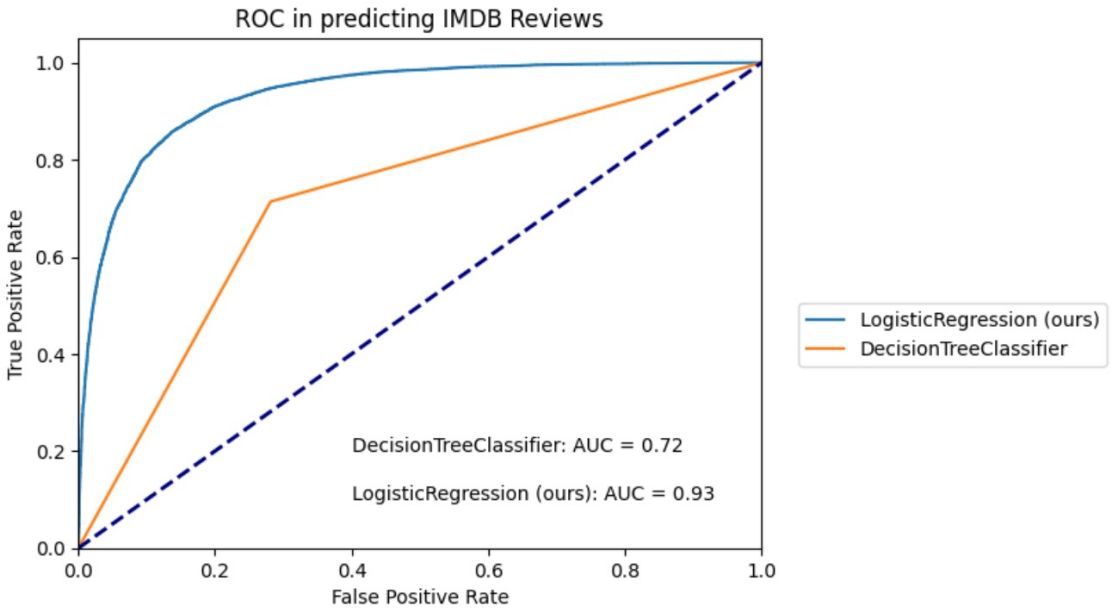

Figure 5: ROC curve

Based on the AUROCs, our simple logistic model does a better job than decision tree model from Sklearn in binary classification as its AUROCs score is higher in value. AUROC score of decision trees is 0.72, which is smaller than the AUROC score of logistic regression that is 0.93. Furthermore, simple logistic regression has a nearly perfect AUROC score, which means it is a highly efficient model for binary classification prediction on IMDB data.

4.2.3 AUROCs (with Decision Tree's) on IMDB Test Data (Training Data at 20%, 40%, 60%, 80%, and 100%)

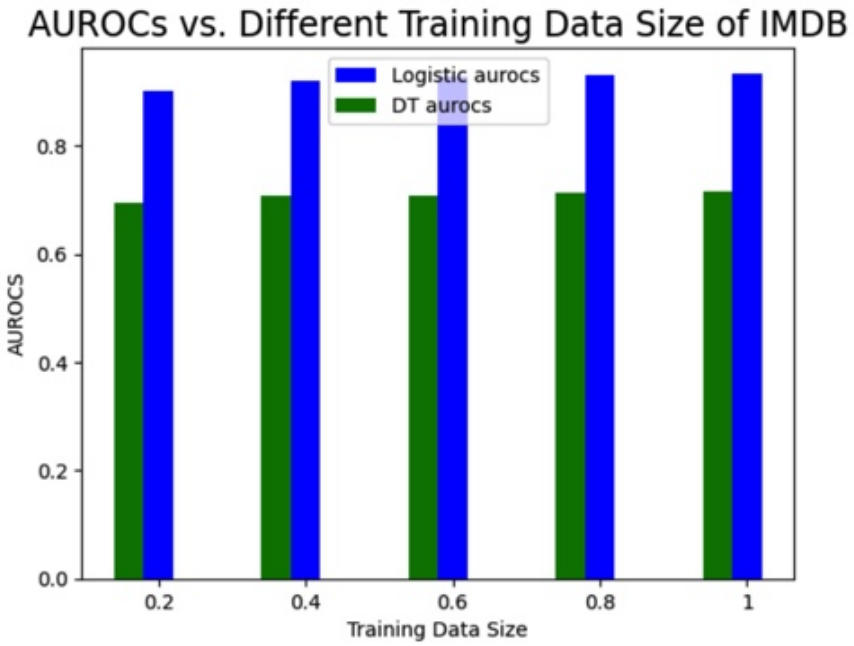

Figure 6: Decision Tree and Simple Logistic AUROC

Even though it isn't very obvious by eyes, but we can see a light increase in the AUROCs as the percentage of data used for training increases. The difference in AUROCs of logistic regression and decision trees is relatively large for all percentages of training data, where for all cases, our simple logistic regression's AUROC scores are higher than decision tree's AUROC scores. This plot results indicate that simple logistic regression approach performs way better than decision trees approach for binary classification experiment.

4.3 20-news-groups dataset analysis

4.3.1 Classification Accuracy (with Decision Tree's) of Test Data (Training Data at 20%, 40%, 60%, 80%, and 100%)

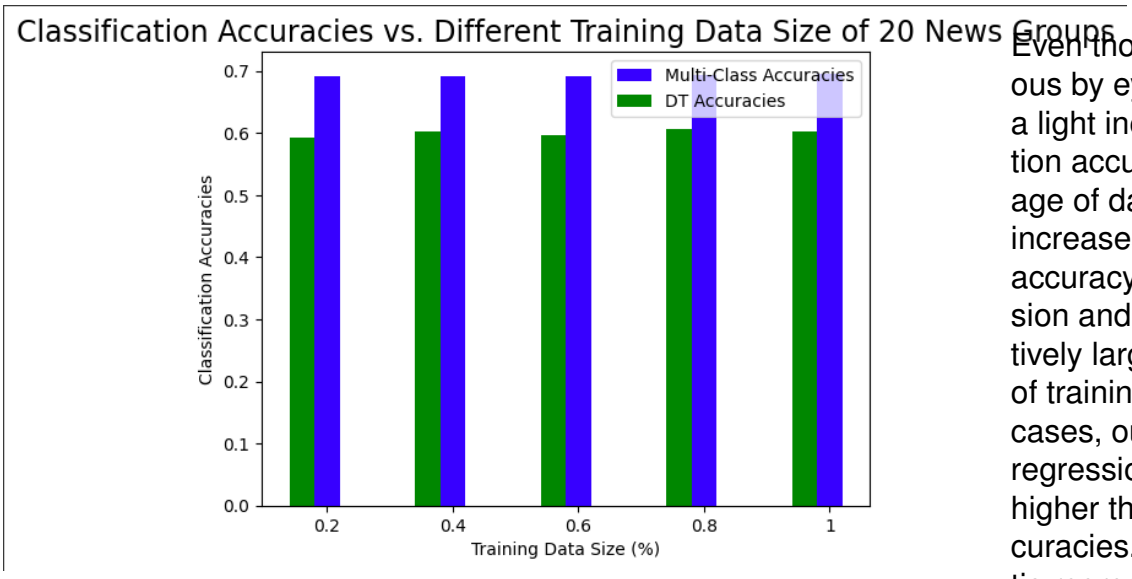

Figure 7: Decision Tree and Multi-Class Logistic classification accuracy on 20-news-groups dataset

Even though it isn't very obvious by eyes, but we can see a light increase in classification accuracy as the percentage of data used for training increases. The difference in accuracy of multi-class regression and decision trees is relatively large for all percentages of training data, where for all cases, our multi-class logistic regression's accuracies are higher than decision tree's accuracies. So, multi-class logistic regression performs better than decision trees approach.

4.3.2 Heatmap of Top Five Features for Each of the 5 Classes

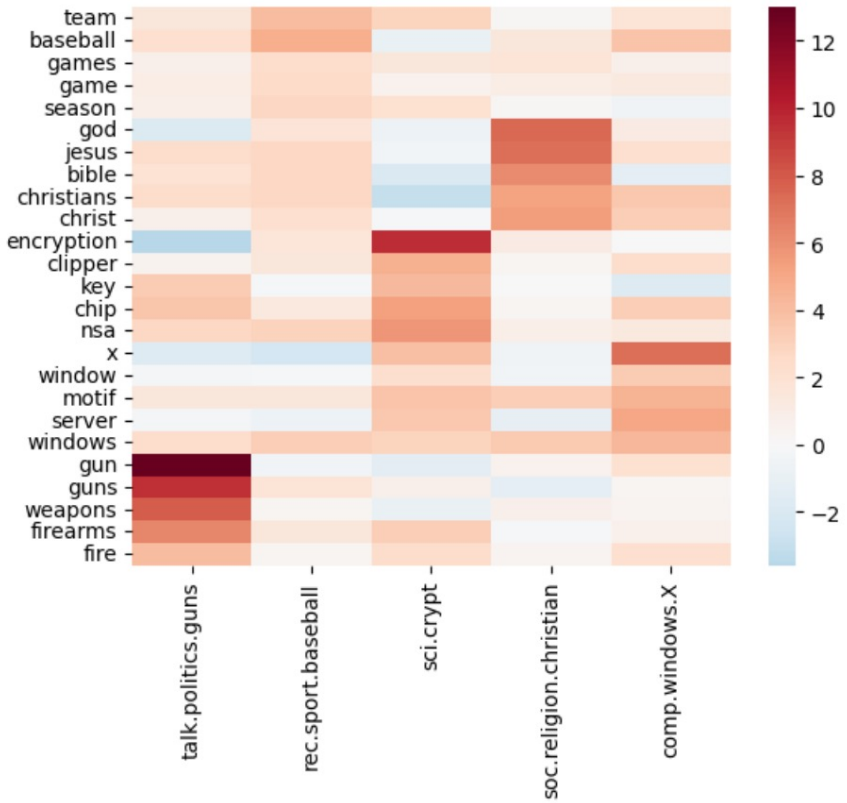

Figure 8: Heatmap for 20-news-groups' Five Chosen Classes

From the legend, we know that the warmer the color, the higher the feature's coefficient, and thus its importance to a categorical class. For each class, the features' contributions are very different since our classes are relatively distinct from each other: we can see blocks of darker colors, from which we can clearly see the presence of 5 distinct classes even without the names on the x-axis. For instance, in the first column, the most important features are: "gun", "guns", "weapons", "firearms", "weapon", from which we can deduce that the posts in this class must be about weapons, specifically guns - the actual class is "talk.politics.guns", similar to our educated guess.

5 Discussion and Conclusion

In this assignment, we have compared decision tree's AUROCs with simple logistic regression's for binary classification on the IMDB data, as well as decision tree's classification accuracy with multi-class logistic regression's for multi-class classification on the 20-news-groups data. The goal of this assignment is in fact to compare the performance of decision trees and simple/multi-class logistic regression. As a conclusion, we found out that in general logistic regression approach performs better than decision trees approach since a noticeable difference could be observed in the results obtained.

5.1 Comparing Simple Logistic Regression & DT on the IMDB Reviews dataset (Binary Classification)

We have computed ROC curves and AUROC score of simple logistic regression and decision trees to evaluate and compare models for the binary classification on IMDB dataset. By definition, the higher the AUROC scores, the better the classification since there are more true positives and less false positives. From our resulting plots, we notice that simple logistic regression approach is better than decision trees approach since simple logistic regression has AUROC score of 0.93, which is way better than AUROC score of decision trees which is 0.72. Furthermore, for different training data size, AUROC scores of simple logistic regression are always greater than AUROC scores of decision trees. There is a significant difference between the two models' AUROC scores, thus we conclude that for binary classification, simple logistic regression approach should be the preferred model.

5.2 Comparing Multi-class Regression & DT on the 20-news-groups dataset (Multi-class Classification)

We have computed classification accuracies of multi-class logistic regression and decision trees to evaluate and compare models for the multi-class classification on 20-news-groups dataset. By definition, the higher the accuracy, the better the classification since there are more correctly classified data points and less misclassified data points. From our resulting plots, we notice that multi-class logistic regression approach is better than decision trees approach since it always has classification accuracies of approximately 70% for different training data size, which are higher than decision trees accuracies which only are approximately 60%. There is a noticeable difference between the two models' classification accuracies, thus we conclude that for multi-class classification, multi-class logistic regression approach should be the preferred model.

5.3 Conclusion

From the experiment, we can conclude that both simple and multi-class logistic regression approaches perform better than decision trees for classifying datasets with multiple features and classes(simple logistic regression for binary classification and multi-class logistic regression for multi-class classification). Clearly, there are noticeable difference in the AUROCs scores for binary classification and noticeable difference in the classification accuracies for multi-class classificiation. Thus, the results obtained indicate that for classification type of experiment, simple and multi-class logistic regression models should be the preferred models used.

5.4 Further Studies

5.4.1 Overfitting

As possible future studies, we can investigate on overfitting and classification accuracy by trying different numbers of features and/or numbers of classes and/or ways of splitting training, validation and test data points. For example, in our experiment, we see overfitting in the multi-class regression as the validation error curve starts to increase when a specific number of iterations is reached. So, if one wishes to more accurately study a given dataset, preventing overfitting by stopping at the appropriate iteration could be consider ideal.

5.4.2 Prevention of Possible Errors by Built-in Library Methods

Even though results obtained during this experiment are accurate, sometimes using methods of built-in library would cause potential negligible errors. For example, in our experiment the use of CountVectorizer would sometimes cause minor words count errors. But luckily, these errors are negligible so the outcomes of using the method to pre-process data would still be highly accurate. However, if we were to apply it to a highly large dataset that have 10 times more data than the one we have in this experiment, negligible errors would accumulate and possibly leading to many enormous and huge errors that could highly affect the outcomes of such experiment. To solve this situation, one could employ other methods to pre-process dataset using more robust technique or dividing the dataset into smaller subset for running experiment.

5.4.3 Comparison with Other Models

We can also compare our regression models with the ones in the built-in Sklearn library and/or other classification/regression methods such as K-nearest neighbors approach for further model perfor-

mance evaluation. Comparing with other models available would give us insights into choosing more appropriate and suitable models for future experiment that requires classification experiment. Very often in machine learning experiments, resources are limited. Therefore, knowing which types of models are suitable for a specific type of experiment would largely prevent much erroneous and misleading outcomes, thus giving more accurate results and conclusion for analysis.

6 Statement of Contributions

- Jia: Code writer and editor, report editor
- An Ni: Main Code writer , report editor
- Yi Wen: Main Report writer, code editor

7 References

Maas, Andrew. “Large Movie Review Dataset.” Sentiment Analysis, ai.stanford.edu/ amaas/data/sentiment/.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).

“Sklearn.Datasets.Fetch_20newsgroups.” Scikit, scikit-learn.org/stable/modules/generated sklearn.datasets.fe 20newsgroups.htmlsklearn.datasets.fetch _20newsgroups.html.

“Semi-Supervised Classification on a Text Dataset.” Scikit, scikit-learn.org/stable/auto_examples/semi_-supervised plot_semi_supervised_newsgroups.html#sphx-glr-auto-examples-semi-supervised-plot-semi-supervised-newsgroups-py.