Fall COMP 551 Assignment 2

October 2022
Group 3: Steve Wen, Chloe Si, Navid Hassan Zadeh

Abstract

In this project we implemented binary and multi-class logistic regression models and investigated their performance on two benchmark datasets: IMDB dataset, and 20newsgroup dataset. We found that the two logistic regression model achieved better results than KNN but they need significantly more time to train. What's more, we can conclude that the logistic regression models are more stable as the accuracy is barely affected by the training size while this is a key variant of the accuracy of KNN classifier.

Introduction

Text classification is a popular machine learning technique used commonly to classify a text into one or many predefined categories. In text classification, it is important to choose an appropriate model and the best representative data as features for the model. In this experiment we implemented binary and multi-class logistic regression and applied them on two datasets: IMDB reviews (binary) and 20newsgroup (multi-class). In preprocessing, we used z-score and mutual information score to select the most effective words as features for the IMDB and 20newsgroup datasets, respectively. In addition we analyzed how the changes in the size of the training data can affect the performance of the models and tried different model stopping criteria.

Dataset

IMDB Reviews Dataset

This dataset contains 50,000 movie reviews evenly split into 25k train and test sets with their positive and negative class labels. Only highly polarized reviews are included with a score ≥ 7 been classified as positive and a score ≤ 4 been classified as negative and no more than 30 reviews are included per movie. Using the provided tokenized bag of words, we removed stopwords, punctuations, rare words(occurrence<1%) and common words(occurrence>60%). The percentages are chosen to make the preprocessed vocabulary size resonable and the represented vector as dense as possible. We experimented with different thresholds, but eventually found out this pair (0.01, 0.6) is the best.

20 newsgroup

This dataset contains around 18,000 news articles partitioned almost evenly among 20 categories. For this experiment, we extracted four categories of distinct fields, namely, comp.windows.x, rec.motorcycles, sci.space, and talk.politics.misc. For preprocessing, similar to IMDB dataset, we firstly removed stopwords, punctuations, rare words (occurrence<2%) and common words(occurrence>50%) since this could make the vocabulary size smaller and thus quicker to tokenize. Then we tokenized the remaining words and further filtered out words that occurred almost evenly in each class(keep words that occurred >40% in only one class). This step is done because we want the representatives of each class to be the unique but not confounding ones.

Method

Logistic Regression and IMDB Reviews Dataset

Logistic regression classification algorithm is used to predict a binary outcome based on a series of independent variables. In this IMDB reviews dataset, we aimed to predict whether a review is negative or positive based on tokenized words, which is a useful application of logistic regression. In our implementation, after preprocessing the vocabulary by filtering out useless words(details in dataset section), we reduced the number of features to around 1600. Then we chose the top 500 representatives as selected features by the magnitude of their z-scores and further normalized the inputs. The model approximates the weight matrix by recursively updating it in each step using the gradient descent with an appropriate learning rate and stopping criteria(maximum iteration number and difference between gradients).

Multi-class Logistic Regression and 20Newsgroups dataset

Multi-class logistic regression classification algorithm is a generalized version of logistic regression allowing prediction of more than 2 classes. Given the 20newsgroups dataset of 20 categories, we chose 4 of them to perform multi-class regression classification. We preprocessed the data in a similar manner as in IMDB dataset(details in dataset section) and finished with around 300 useful words as features. Then we further selected 50 representatives of each class without repetition by calculating the mutual information score between every word and each class. We also normalized the input of the multi-class regression model, and the model recursively updates the weight matrix using gradient descent with an appropriate learning rate and stopping criteria(maximum iteration and difference between loss).

Results

In analyzing the results, we found that feature selection plays an important role in improving the accuracy of the models. In the IMDB dataset, we selected 500 features (words) with the highest magnitude of z-score while in the 20Newsgroups dataset, we only selected 200 features with top MI score, considering the relatively small sample size. We implemented the logistic regression and multi-class regression classifiers from scratch and compared their performance with KNN to see which classification model works better and under what conditions. The results indicate that the binary logistic regression models perform better than KNN with a difference of over 10% and multi-class logistic regression model outcompetes KNN with a difference in accuracy around 20%. This is in consistent with KNN's property that it's vulnerable when dealing with high-dimensional data. What's more, since KNN is a lazy-training model which simply fits in all the training samples, the size of the training set matters a lot in improving the accuracy while regression model is not sensitive to it. Additionally, one thing worth noting is that KNN performs much worse than the parametric methods that we implemented, we hypothesize that could be largely due to the high dimension feature space. Previous works have shown that KNN performs poorly on high dimensional data, which is the case in our implementation. We also computed gradient perturbation to ensure our gradient computation is correct. We got 9.46e-06 on IMDB dataset with Logistic Regression and 2.87e-10 on 20 Newsgroup with Multiclass Linear Regression. The results indicate our computation should be correct.

Dataset	Model	Accuracy
IMDB Dataset	**Our implementation of Binary logistic regression**	**82.87%**
	KNN (k=30)	71.86%
20 Newsgroup	**Our implementation of Multiclass logistic regression**	**74.75%**
	KNN (k=5)	55.68%

Table 1: The results for comparing the accuracy of different models on both dataset. (including our implementation of logistic regression as well as sklearn's KNN)

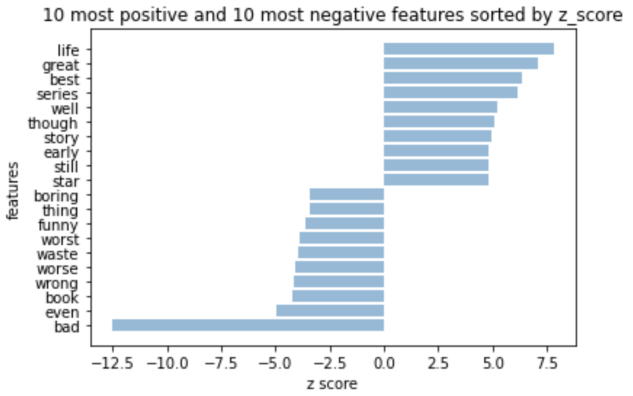
Logistic Regression and IMDB Dataset

Figure 1,2,3 demonstrate experiment results of logistic regression on the IMDB dataset.

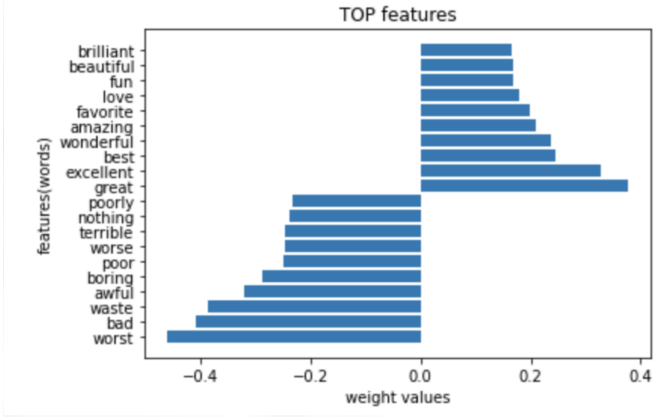
Figure 1(a) contains the 10 most positive and negative z-scores with their corresponding features(words). We can observe that some of these features make sense and properly fall inside the desired classes of whether a movie is good or bad, such as "worse", "bad", "worst", "great", "best", "well", etc. However there are also words that do not seem to have any connection with those classes. So using z-score as feature selection is somewhat reasonable, but we do need further training since there are still some words that are not good representatives of their classes.

Figure 1(b) contains the 10 most positive and negative weights after training with their corresponding features. These features are all very strong indicators in determining whether a movie review is positive or negative. Since the larger the absolute weight, the more important the feature in classification, the model indeed learned which features matter the most and how to classify reviews based on these representatives.

Figure 2(a) is the convergence plot with learning rate = 0.001. Since the logistic regression cross entropy loss is a convex function, the convergence plot should be a smooth decreasing curve and the result is as expected.

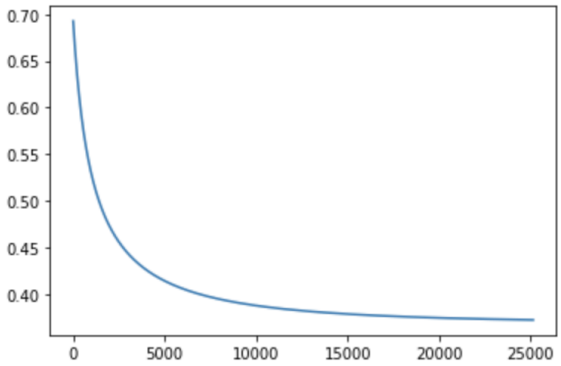


(a) TOP 10 negative and positive features by z-score

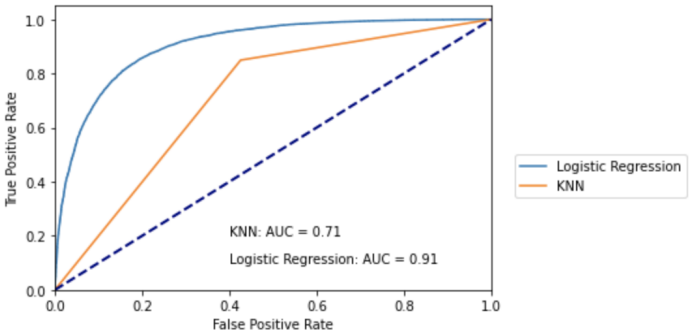


(b) TOP 10 negative and positive weight

Figure 1



(a) Convergence of CE



(b) ROC curve comparison of models

Figure 2

Figure 2(b) is the ROC curve of logistic regression and KNN. The logistic regression curve is closer to top-left corner, which indicates a better performance than KNN.

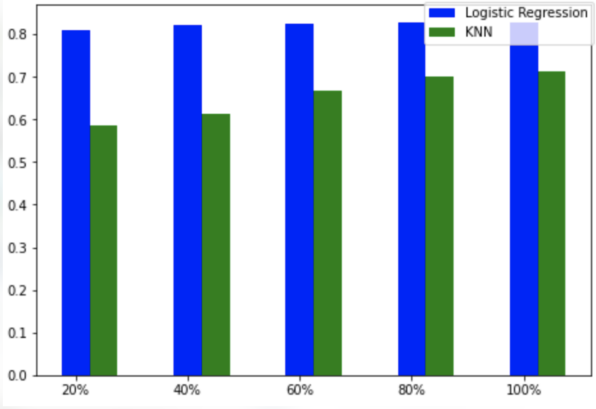
Figure 3(a) and 3(b) demonstrate the accuracy and AUROC of logistic regression and KNN models on 20%, 40%, 60%, 80% and 100% of training data respectively. The plots show that changes in the size of the training data barely affects the accuracy of the logistic regression, while the impact on KNN's accuracy is much more tangible. We know training in KNN classifier consists of merely fitting in all the training samples. Thus with a higher size of training data, there will be more closely-positioned neighbouring data points, which results in a higher accuracy for KNN. However, in logistic regression we seek for representative features and they should be constant no matter how large the training size.

Multi-class Logistic Regression and 20Newsgroup Dataset

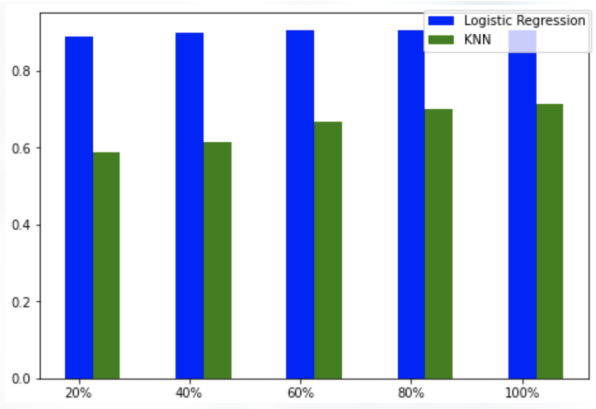
We see in Figure 4(a) that after changing the size of the training data the accuracy of the multi-class logistic regression model are barely affected. However KNN's accuracy changes more significantly.

In Figure 4(b), we see that the training error has an overall smooth and decreasing trend with learning rate= 0.001. This is consistent with expected result as softmax loss is a convex function.

Figure 5 shows the heatmap of the multi-class logistic regression for the 20newsgroup dataset. We have taken the selected features corresponding to the top 5 weights of each class and put them in the heatmap as rows. Each block of the heatmap is the weight of the feature(row) in one of the four class(column) and every row sums to 1. Thus, it's obvious that the top 5 features of each class are indeed representatives of their own class(the weight is much larger than the other 3) and there is barely confounding ones. The words are also clearly a good indicator in determining which class they belong to. The heatmap indicates that our model has a good performance after training.

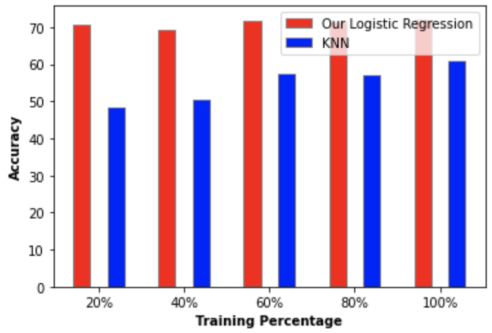


(a) Accuracy of logistic regression and KNN with various training size

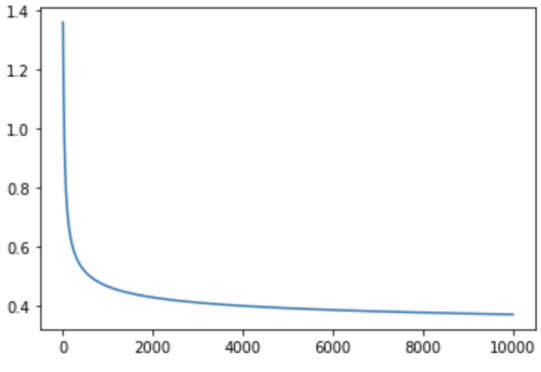


(b) AUROC of logistic regression and KNN with various training size

Figure 3



(a) Accuracy of multi-class regression and KNN



(b) The convergence of CE

Figure 4

Discussion and Conclusion

We implemented from scratch the binary and multiclass logistic regression and evaluated their performance by applying them on two datasets, namely the IMDB and 20newsgroup. We performed a through feature extraction procedure using word count and word frequency in both datasets and calculated z-score and mutual information score to rank the features' importance. This was a necessary and important step because the performance of the model was strongly affected by the choice of selected features. We excluded too common words and too rare words, also those ones that confound the results.

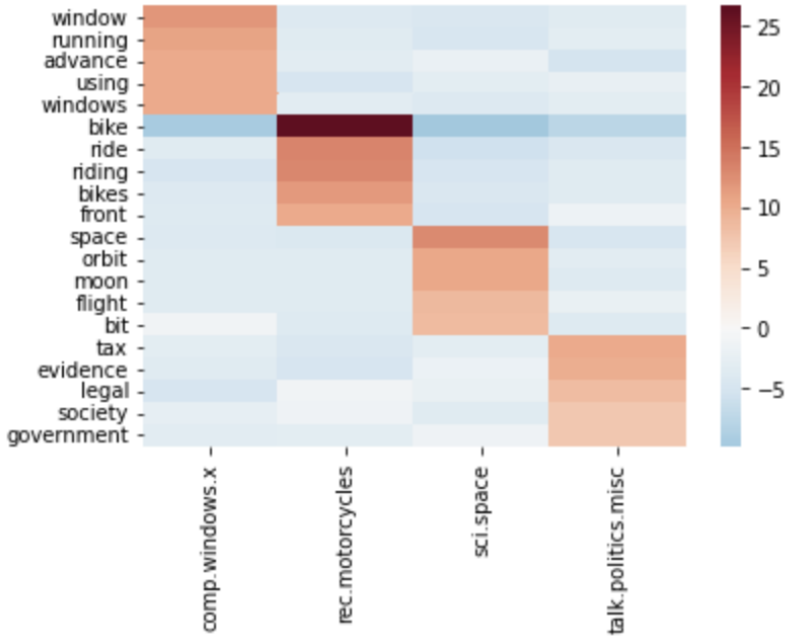
We concluded that the binary and multi-class logistic regression classifiers are more accurate and perform significantly better than KNN classifier in both datasets with a difference in accuracy over 10% and 20% respectively. The latter larger difference could be attributed to KNN's worse performance on higher dimensional data.

What's more, the size of the training dataset has significant different effects on the 2 models. Accuracy of (multi-class) logistic regression almost has nothing to do with the training size as the model seeks for potential representatives, while accuracy of KNN affected a lot by the size due to its simple fitting training process.

For further investigations, we would like to find other ways to select those features and to improve the training speed. Also, the accuracy of the 20newsgroup is not very high, so we may try regularization on this dataset to see if it can be improved as the size of samples and features relatively close compared to other datasets.

Statement of Contributions

Navid Hassen Zadeh: Multiclass regression Preprocessing, Report Writing
Chloe Si: Feature selection, Model Construction, Report Writing
Steve Wen: Model Construction, Model Debugging, Feature Selection, Report Writing



(a)

Figure 5: heat map of multi-class logistic regression for the 20newsgroup dataset

References

- Large Movie Review Dataset, Maas, Andrew L *et al*, 2011
  `https://ai.stanford.edu/~amaas/data/sentiment/`
- Sklearn 20 newsgroups dataset (classification)
  `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_20newsgroups.html`
  item Probabilistic Machine Learning: An introduction, Kevin P. Murphy 2022
  `https://probml.github.io/pml-book/book1.html`