COMP 551: Applied Machine Learning Report for Project 2

Zijian Pei, Xiaoxiao Shang, Yueng Zhang

March 8, 2022

Contents

1 Abstract

The purpose of this project is to investigate the performance of Naive Bayes and 5-fold cross validation with Softmax Regression from Scikit-learn package on 20 Newsgroups Dataset and Sentiment140 Dataset. Based on the experimental results, we found that the Logistic Regression has better performance on both datasets than Naive Bayes when the number of training data is fixed. We conducted experiments by setting different hyperparameters and recorded the one with the best average 5-fold result. For Softmax Regression, we implemented our models with different multi-classes, loss functions and maximum iterations; for Naive Bayes, we tested different smoothing parameters. Furthermore, we carried out extra experiments by adding stopwords and vocabulary list to identify potential factors that should affect the accuracy.

2 Introduction

[2]The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents; the task is to do multi-classification. The response variable has twenty classes, corresponding to twenty newsgroups. The features are in String format and can not be processed directly by the two algorithms. [4]The Sentiment140 Dataset is a collection of 1,600,000 data entries with five features and one response variable; the task is to do binary classification. The response variable has 2 classes, denoted as 0 and 4. Similarly, the features are in String format and can not be processed directly. For 20 Newsgroups Dataset, the best models are Naive Bayes with $\alpha = 0.01$ and Logistic Regression with maximum iteration 1986 and saga solver. For Sentiment140 Dataset, the best models are Naive Bayes with $\alpha = 0.98$ and Logistic Regression with maximum iteration 17032 and saga solver.

3 Datasets

In the process of converting text data, we followed the steps from [3]tutorial and transformed the training and test data in the String format to a count table with format of integers, a data type suitable for classification. For the Sentiment140 Dataset, we figured out that the data is arranged by the response variable. Therefore, shuffling is needed to guarantee all training dataset after cross validation have instances from both classes. After shuffling, we followed the steps described to transform the text data. In addition to these procedures, the number of training data from Sentiment140 Dataset is extremely large. Consequently, we decided to use only a portion of them as training data to speed up the computation. For both datasets, we use a cross_validation_split function to do 5-fold cross validation and selected the pair of training and validation set that guarantees the highest accuracy on the validation set. In particular, different values of hyperparameters are tested in determining the best Naive Bayes models.

4 Result

For the 20 Newsgroup Dataset, the best models are Naive Bayesian with $\alpha = 0.01$ and Logistic Regression conditioning on the best training and validation pairs under 5-fold cross validation. In Naive Bayes, the accuracy on the validation data is 0.8233, which is the highest among all hyper-parameter and training and validation pairs. The accuracy of this model on the test data is 0.8375. In Logistic Regression, the accuracy on the validation data is 0.7361, which is the highest among all training data pairs. The accuracy of this model on the test data is 0.7836.

Model	Accuracy (Validation)	Accuracy (Testing)
Naive Bayes	0.8233	0.8375
Logistic Regression	0.7361	0.7836

Table 1: 20 Newsgroups Dataset: Best Accuracy for Naive Bayes and Logistic Regression

For the Sentiment140 Dataset, we found that the best models are Naive Bayesian with $\alpha = 0.98$ and Logistic Regression conditioning on the best training and validation pairs under 5-fold cross validation. In Naive

Bayesian, the accuracy on the validation data is 0.716325, which is the highest among all hyper-parameter and training and validation pairs. The accuracy of this model on the test data is 0.7159. In Logistic Regression model, the accuracy on the validation data is 0.7242, which is the highest among all training data pairs. The accuracy of this model on the test data is 0.7242.

Model	Accuracy (Validation)	Accuracy (Testing)
Naive Bayes	0.716325	0.7159
Logistic Regression	0.7242	0.7242

Table 2: Sentiment140 Dataset: Best Accuracy for Naive Bayes and Logistic Regression

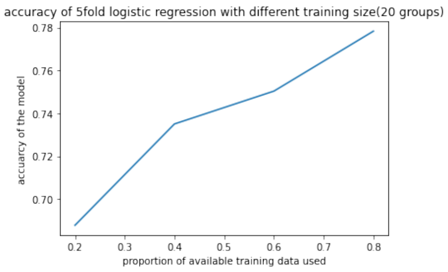
5 Extra Experiments

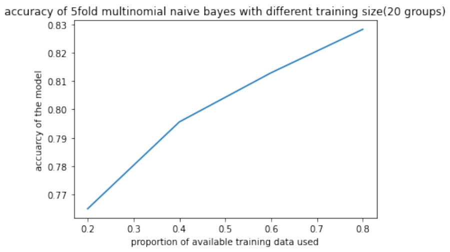
5.1 Experiment 1: Number of Training Data vs Accuracy

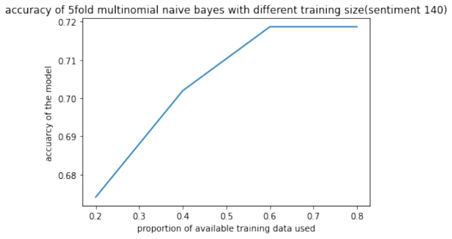
After considering factors that can potentially affected the models' accuracy, we decided to test how different amounts of training data would affect the results.

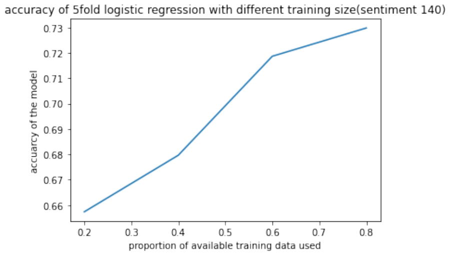
In the following figures, we can tell that for both datasets, the prediction accuracy of both Naive Bayes and Logistic Regression increases as the number of training data increases. The accuracy of Naive Bayes reaches 0.83 and 0.73 respectively for 20 Newgroups Dataset and Sentiment140 Dataset when maximum number of training data is applied under 5-fold cross validation; the accuracy of Logistic Regression reaches 0.78 for 20 Newgroups Dataset when maximum number of training data is applied under the same setting.

However, for the Sentiment140 Dataset, the prediction accuracy of the Naive Bayes keeps rising at the first, and stays steady when the proportion of training data is greater than 0.6.









5.2 Experiment 2: Stop Words and Other Vocabulary Set

When doing extra experiments on Sentiment140 Dataset, we found that words like "am", "are", "is", "me", "you", "he", "she", etc. are irrelevant to different labels. These common words can appear anywhere, which means their appearance should not give any useful information on the label. Consequently, we applied one parameter of CountVectorizer, stopwords, when reading data to eliminate the effects of these words. We also found a [1]vocabulary text file relevant to the dataset. We believed it is worthwhile studying how the words from this file would affect the accuracy of the model by cutting down the appearance of words irrelevant to classification.Then we compared the refined models' accuracy with the original result from Naive Bayes.

Model	Accuracy
Logistic Regression	0.7242
Logistic Regression with stop words	0.7577
Logistic Regression with own vocabulary file	0.7270

Table 3: Sentiment140 Dataset: Accuracy for different means

6 Discussion and Conclusion

First, we ran cross validation on both algorithms on both datasets, and we summarized the performance of the best models. For both data sets, the accuracy is higher when implementing Logistic Regression model. The maximum iteration for two datasets are different but they all achieved highest accuracy when applying l_2 loss and multinomial. Besides, models' performance on 20 Newsgroups Dataset is better than the Sentiment140 Dataset in terms of accuracy.
Then we tested how the amount of training data should affect the models' accuracy, and found that the accuracy of Logistic Regression and Naive Bayes is always positively related to the number of training data.
In the extra experiments , we tried to increase the accuracy of Logistic Regression by applying stopwords and the vocabulary list found on Internet. From Table 3, we can tell that the accuracy increases from 0.7242 to 0.7577 and 0.7270 when applying stopwords and vocabulary list respectively.
For future investigation, we consider using different models and techniques. For example, we can set up different learning rates or different stopping criteria for the gradient descent method.

7 Contributions

Zijian Pei: Cross Validation Implementation, Data Cleaning, Data analysis, Report Writing
Xiaoxiao Shang: Naive Bayes and Logistic Regression Implementation, Data Analysis, Report Writing
Yuteng Zhang: Naive Bayes and Logistic Regression Implementation, Data Analysis, Report Writing
Code for building Naive Bayes and Logistic Regression references from tutorial

References

[1] DWYL. *english-words*. Website. https://github.com/dwyl/english-words/blob/master/words.txt. 2020.

[2] Scikit Learn. *The 20 newsgroups text dataset*. Website. https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html. 2019.

[3] Scikit Learn. *Working With Text Data*. Website. https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html. 2019.

[4] *Sentiment140*. Website. http://help.sentiment140.com/for-students. 2019.