MiniProject2

Group 20

March 7, 2022

1 Abstract

There are multiple widely used models to classify texts. In this project, we investigated Naive Bayes and logistic regression on two datasets: 20news group and sentiment140. We started with preprocessing the datasets, then implemented Naive Bayes and cross-validation methods. We also optimized these models by tuning the hyperparameters. Finally, we compared the performance of these models on both datasets using different training sets and test set splits. We concluded that Softmax Regression gave better accuracy in 20 news datasets, and Naive Bayes gave slightly better accuracy in the Sentiment140 dataset. But in general, the sentiment140 dataset has better accuracy than 20Newsgroups.

2 Introduction

Text classification plays an important role in the ML area due to its large number of applications in the real world industry, such as grouping messages and news. Therefore, many efficient machine learning algorithms are developed to help with classification. In this project, we aim to conduct classification using Navies Bayes and logistic regression along with cross-validation methods and compare their performance on two different datasets. Ultimately, we observe that the ideal alpha parameter for our NB is 0.007 for the 20Newsgroups dataset and 13 for the Sentiment140 dataset. We then compared the performance of NB with the logistic regression method on the two datasets. We also used different test-train set split to maximize accuracy for each model. We found that Sentiment140 has better accuracy in general, while the accuracy for each model is similar.

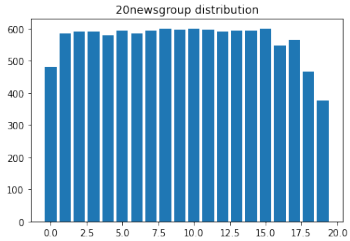
2.1 Related Work

The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. It is widely used by many researchers in machine learning and deep learning-related field. For example, the paper A simple approach to multilingual polarity classification in Twitter used this dataset to show the performance of multilingual polarity classification using sentiment analysis[1]. Sentiment140 dataset allows us to discover the sentiment of a brand, product, or topic on Twitter. It is also widely used for machine learning purposes. For example, in the paper Real-time sentiment analysis of tweets using Naive Bayes[2], the author used this dataset to improve classification for Naive Bayes.

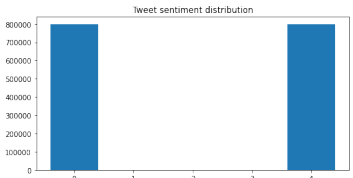
3 Datasets

The 20Newsgroups data set is a collection of newsgroup documents including 11314 training data and 7532 test data. By class distribution, we see that the training set is partitioned evenly across 20 topic groups. Some significant features like headers and names give an abundance of clues that distinguish newsgroups, which make the classifiers barely have to identify topics from the text at all. So we removed the blocks 'headers', 'footers', and 'quotes' to get a more realistic training and avoid over-fitting.
Sentiment140 contains tweets that are labeled as positive or negative. Our data set has 1600000 training data (half positive and half negative) and 359 test data. The data set has five features, but our focus is on classifying the sentiment of the text, hence we only need the tweet text.

To extract features from the text we first cleaned the text, including stop words removal, stemming, and regex filtering. Then we used Scikit-learn function CountVectorizer and TfidfTransformer to get frequencies of all the words appearing in one dataset. Note that the frequency we used is TF-IDF (Frequency Inverse Document Frequency), this helps avoid the potential problem of discrepancies compared to the bag of words. After this, we got a sparse matrix as our input.



(a) 20Newsgroups



(b) Sentiment140

Figure 1: Class distribution of datasets

4 Results

4.1 Comparison of model performance

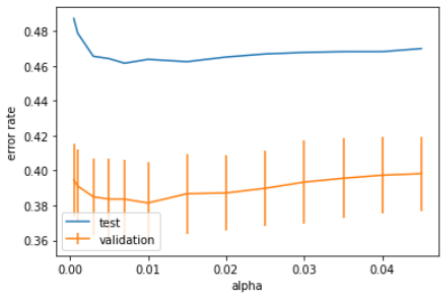
	20 news dataset	Sentiment140 dataset
Multinomial Naive Bayes	0.5384 with alpha=0.007	0.8273 with alpha=13
Logistic Regression	0.7010 with C = 10	0.8162 with C =10
winner	Logistic Regression	Multinomial Naive Bayes

(a) result table

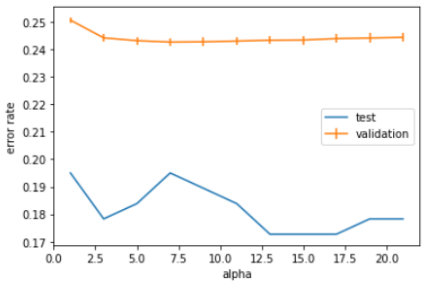
Figure 2: Result

From this table, we could see that Softmax Regression gave better accuracy in 20 news datasets, and Naive Bayes gave slightly better accuracy in the Sentiment140 dataset.

4.2 Impact of hyper-parameters



(a) 20Newsgroups



(b) Sentiment140

Figure 3: Accuracy for datasets by Naive Bayes

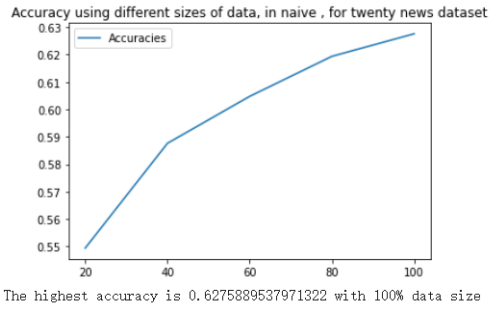
Figure 2: the left:Different alpha values and the performance of Naive Bayes, using 20 news group dataset;the right:Different alpha values and the performance of Naive Bayes, using sentiment140 dataset.

As the first part of our experiment, we test which hyper-parameter values could help our Naive Bayes model achieve their best performance, And we implement a 5-fold cross-validation method, which enables us to estimate the better accuracy of the NB model. On the 20news group dataset, we found that the best accuracy is 0.5384 with alpha =0.007. And on the Sentiment140 dataset, the highest accuracy is 0.8273 with alpha=13. From the figure, we could conclude a trend that the accuracy tends to decrease with the increasing alpha values.For Logistic Regression model, we tested the different C values(100,10,1,0.1,0.01).And we discovered that the best accuracy for both datasets is given when the C value is equal to 10. Due to the data problem, we could not draw a proper graph to show its trend. But from the results we have, we could get a conclusion, when C values are between the range from 10 to 0.1, it usually gave better performance

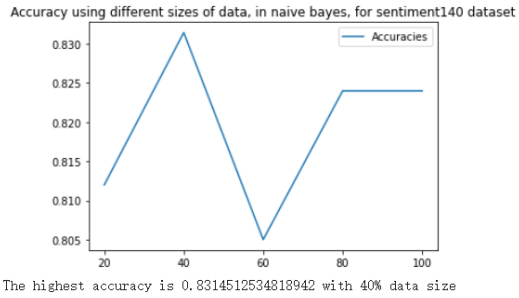
4.3 Impact of different algorithm

We also import the provided Multinomial Naive Bayes method from Sci-kit learn Library. And we did the same experiment as our Naive Bayes to MNB. So we found the highest accuracy for the 20news dataset increased to 61% with the default alpha values. But the accuracy for the Sentiment140 dataset is very close for both Naive methods. This may result from the different fit functions that used a different technique like Laplace smoothing, different feature selection, etc. For Logistic Regression, we tested three solvers('newton-cg', 'lbfgs', sag) which are all L2 penalty. The results of the experiment show that 'newton-cg' and 'lbfgs' gave the almost same result for both datasets. On the experiments of Softmax regression of Sentiment140 datasets, 'newton-cg' and 'lbfgs' maybe not converge in limiting iteration, but using sag algorithm could solve the problem. The 'newton-cg' and 'lbfgs' algorithms did not work well in the large-size dataset. We also import linear regression to fit both datasets and calculated the mean square error. The MSE for the 20news group is relatively large, about 17 and that for the Sentiment140 dataset is approximately 3. So, Large MSE means the error of our prediction is large, therefore supporting the results we get in the previous experiment.

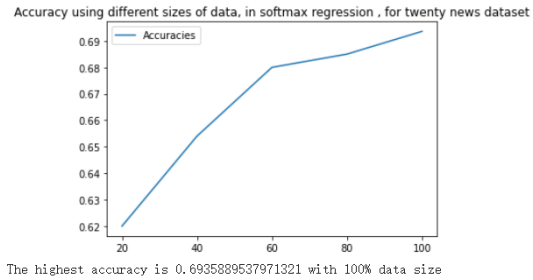
4.4 Impact of different data size



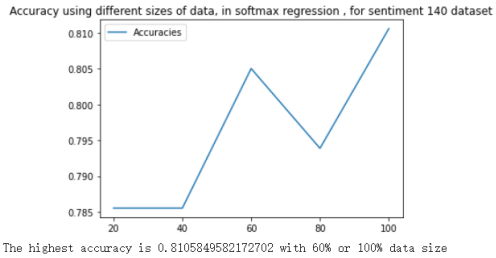
(a) 20Newsgroups using Naive Bayes



(b) Sentiment140 using Naive Bayes



(c) 20Newsgroups using Softmax Regression



(d) Sentiment140 using Softmax Regression

Figure 4: Accuracy for both datasets with different data size

Here we vary the size of the train set of our models and test how each model performs under these conditions. For both datasets, the overall trend is accuracy increasing with the data size increasing.

But we could see that the Sentiment140 dataset has the highest accuracy when the data size is 40%. This indicates we may overfit the Naive Bayes model.

5 Discussion and Conclusion

One conclusion we could draw from our experiment is that data pre-processing could play an important role when it comes to accuracy improvement. The different feature selection and text processing methods would lead to a huge difference in accuracy. Choosing a proper feature selection method could improve a lot the accuracy of our models. Some logistic regression methods may not be suitable to fit the large size dataset and get an accurate prediction. In addition, sometimes, a larger data size does not mean better accuracy. It also depends on our method implementation. We could get good prediction results from a small-size dataset if we could process the data properly and set features of the dataset correctly.

6 Statement of Contributions

Everyone contributes to writing the write-up. Tianyu Cao contributes to data processing. Wen Cui contributes to implementing Naive Bayes and cross-validation. Zhengxin Chen contributes to task3.

References

[1] E. S. Tellez, S. Miranda-Jiménez, M. Graff, D. Moctezuma, R. R. Suárez, and O. S. Siordia, "A simple approach to multilingual polarity classification in twitter," *Pattern Recognition Letters*, vol. 94, pp. 68–74, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865517301721

[2] A. Goel, J. Gautam, and S. Kumar, "Real time sentiment analysis of tweets using naive bayes," pp. 257–261, 2016.