COMP 551 Mini-Project 2

Yingjie Xu YJ.XU@MAIL.MCGILL.CA (260862481)
Hengxian Jiang HENGXIAN.JIANG@MAIL.MCGILL.CA (260830557)
Tianhao You TIANHAO.YOU@MAIL.MCGILL.CA (260830663)

Abstract

In this project, we investigated the performance of two machine learning methods — Naive Bayes and Logistics Regression. Each method is implemented and tested by two datasets separately to evaluate the accuracy of the outcomes. Normalization and data cleaning is adopted to improve the predicting model. After comparing the outcome accuracy, we found that the Logistics Regression method predicts more accurately for one dataset and they have similar performance on another dataset. The performance of the Linear Regression method is not good when doing 20-classes classification. However, it performs very well on binary classification. And we discovered that as we increase the number of training set data, the performance of both methods is increasing. We found that both methods predict more accurately for the second dataset than the first dataset and the reason for this may be that the second dataset has far fewer classes to be classified.

1. Introduction

Machine learning is a method used to make predictions or decisions by building mathematical models based on training data. To generate good predictions of data based on decision boundaries, Multinomial Naive Bayes(MNB) and Logistics Regression(LR) are investigated to implement the model.

The 20 newsgroups dataset comprises around 18000 newsgroups posts on 20 topics. The 20 newsgroups collection has become a popular dataset for experiments in text applications of machine learning techniques, such as text classification we did in this project.

The large movie review dataset contains a set of 25,000 highly polar movie reviews for training and 25,000 for testing. This dataset contains movie reviews along with their associated binary sentiment polarity labels.

By analyzing the performance of two models in the two datasets respectively, we found that logistics regression may be more suitable for text classification when there are only two classes to be classified because it performs far better than the MNB in the second dataset and it predicts slightly less accurately in the first dataset that contains 20 classes.

The noticeable discoveries when implementing these two model to the datasets are:

Model 1. Logistics Regression predicts more accurately for the second dataset and predicts slightly less accurate than the Multinomial Naive Bayes model in the first dataset. 2. Multinomial Naive Bayes model predicts slightly more accurately for the first dataset and performed not as well as LR in the second dataset.

Dataset 1. Text data contains a bunch of information and this may take quite a while to process and classify. 2. Some text data download from the internet are dirty and might require a lot of work to clean in order to achieve better performance.

The key takeaway in this project was that by adjusting and manipulating the hyperparameters for both models, we have a better understanding of the impact of changing those parameters would have on the model and the interactive relationship between different parameters. And we also learned that text classification required a lot of work in data cleaning and processing especially for data that directly crawling from the internet. The key to obtaining high accuracy is to process the data thoroughly and carefully.

2. Datasets

2.1. Summary

For this mini-project, we are given two datasets. The first one is the 20 newsgroups dataset. comprises around 18000 newsgroups posts on 20 topics. We use this text information to predict their belonging classes. The second dataset is the IMDB review dataset. This dataset includes the positive and negative reviews for the movies. In the train/test sets, a negative review has a score ≤ 4 out of 10 and a positive review has a score ≥ 7 out of 10 from the IMDB website. We use this information to predict whether that review is positive or negative.

2.2. Data processing

For both datasets, we loaded all the data and then perform the data cleaning. We removed the special characters and punctuation marks. We also delete anything after a '/n'. And we use CountVectorizer to convert our pure text data into a vector on the basis of the frequency of each word that occurs in the entire text. And then we use this generated sparse matrix as input to perform text classification.

2.3. Training, validation and testing sets

Training and testing sets are already given in the original dataset. For the IMDb dataset, the train and test sets contain a disjoint set of movies, so no significant performance is obtained by memorizing movie-unique terms and their associated with observed labels.

2.4. Class distribution

2.4.1. 20 newsgroups dataset:

Classes of newsgroups distributed almost evenly and there's no skew in the distribution.

2.4.2. IMDB movie review dataset:

Class distribution for this dataset is even. It is 1: 1 distributed in between positive and negative reviews.

3. Result

3.1. Dataset 1 Naive Bayes

By performing the Naive Bayes with Multinomial model, and controlling the alpha in range (0.1, 0.3, 0.5, 0.7, 0.9), the resulting graph of Cross-Validation is shown below. From the graph (Figure 1), we can directly find that the validation accuracy reached the highest at alpha = 0.1, and the accuracy is around 72%. Thus, we chose alpha = 0.1 as our hyperparameter for Naive Bayes Implementation, the final testing result is around 66%.

3.2. Dataset 1 Logistic Regression and comparison (Task 3.3)

For the logistic regression, we use directly the package from sklearn. We tune the hyperparameters for logistic regression by changing the tolerance and max number of iterations. By looking at the resulting trend, we

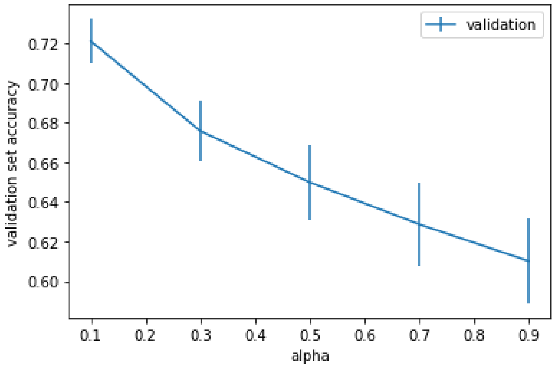

Figure 1: Cross validation with NB

found normally lower tolerance and a higher number of iterations would result in a better model.

While performing Logistic Regression on dataset 1, we find that there is a small difference in the accuracy between the Naive Bayes and Logistic Regression. Naive Bayes has slightly higher accuracy than Logistic Regression. We changed the percentage of training data to be used to train the model in the plot below (Figure 2). It could also be concluded from the plot that a higher number of training data would usually result in a better performance for the model.

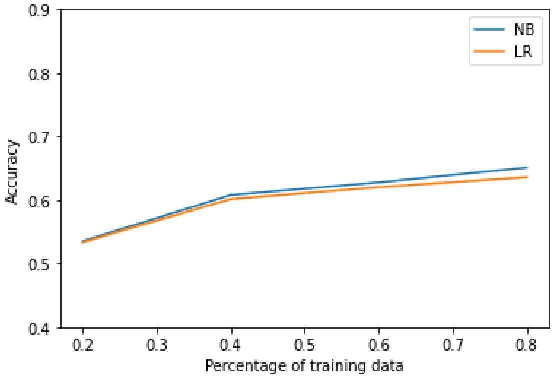

Figure 2: Comparison between NB and LR

3.3. Dataset 2 Naive Bayes

By performing the Naive Bayes, and controlling the alpha in range (0.1-1), the resulting graph of Cross-Validation is shown below (Figure 3). The accuracy and alpha are negatively correlated, accuracy is decreasing as alpha increasing. Hence, when alpha = 0.1, we got the highest validation accuracy around 80%. And the result accuracy from the testing set of Naive Bayes is 81%.

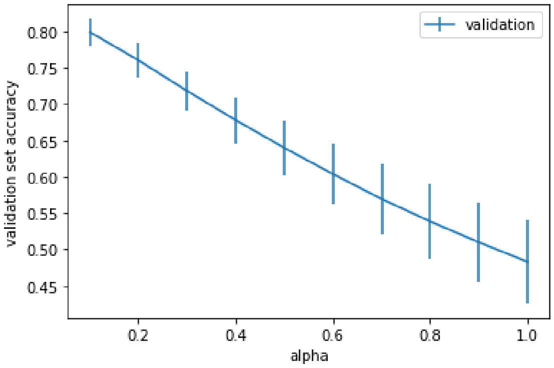

Figure 3: Cross validation with NB

3.4. Dataset 2 Logistic Regression and comparison (Task 3.3)

We repeated the same thing we did in 3.2. The trend of the accuracy of Logistic Regression is increasing as the percentage of data incresing. In another word, it could be concluded that a higher number of training data would usually result in a better performance for the model.

From the graph (Figure 4), we can easily find that the accuracy of Logistic Regression is slighly higher than the accuracy of Naive Bayes with same amount of data. Where accuracy of LR is around 86% and accuracy of NB is around 76%.

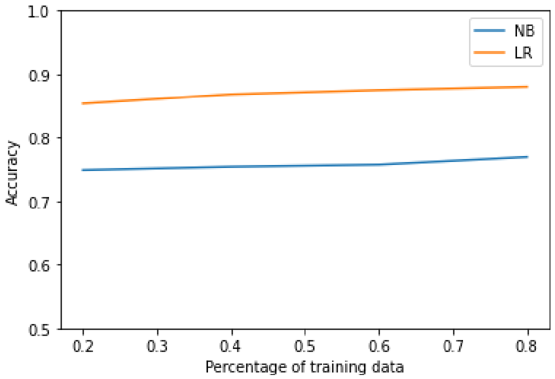

Figure 4: Comparison between NB and LR

3.5. Task 3.1

For the first dataset, which is the 20 newsgroup dataset, we built a target set as y_test which contains (0 - 19) represent the 20 newsgroups for multiclass classification.

For the second dataset, the IMDB dataset, there are two types of data. It either be positive or negative, so we identify them with a particular number, 0 represents the negative and 1 for the positive.

3.6. Task 3.2

3.6.1. Dataset 1:

By looking at the table, we could see that for dataset one Naive Bayes has higher accuracy than Logistic Regression. Naive Bayes is better for doing multi-class classification.

Table 1: D1 Performance Comparision

Model	Result
Naive Bayes	0.66078
Logistic Regression	0.64631

3.6.2. Dataset 2:

For the dataset two, Logistic Regression has higher accuracy than Naive Bayes. It could be concluded that logistic regression is better for doing binary class classification.

Table 2: D2 Performance Comparision

Model	Result
Naive Bayes	0.81062
Logistic Regression	0.88187

4. Discussion and conclusion

4.1. Key takeaways

1. By adjusting and manipulating the parameters for logistics regression model, we have a better understanding of the impact of changing those parameters would have on the model and the interactive relationship between different parameters.

2. Text classification required a lot of work in data cleaning and processing especially for data that directly crawling from the internet. The key to obtain the high accuracy is to process the data throughly and carefully.

3. Logistic Regression model performs better in the datasets that contains only two classes while it does not performed as good as MNB in the first dataset that has 20 classes to be classified.

4.2. Future investigation

1. Due to the large amount of data for both datasets, it takes quite a long time for data processing and model fitting and prediction. We may need to find more effective model to classify the data.

2. Selecting a model to predict a dataset is not easy, we may need to analyze the advantages and disadvantages of the models before applying them.

5. Statement of Contributions

Hengxian was responsible for implementing the Naive Bayes for the first dataset, wrote the abstract, introduction and discussion and conclusion section in the report. Yingjie was responsible for data processing, implemented Logistics Regression and k-fold cross validation for the first and the second dataset respectively, and wrote the dataset section in the report. Tianhao was responsible for implementing Naive Bayes for the second datasets Logistics Regression and k-fold cross validation for the first dataset, and wrote the result section in the report.