Comparison Between Naïve Bayes and Logistic Regression

Runze Yin **Ze Meng** **Yuquan Xu**
Graduate school of Arts & Sciences
{ry180, zm196, yx241}@georgetown.edu

Abstract

There are mainly two methods for Supervised learning, Generative model, and Discriminative model, both are commonly used in NLP. Due to the difference in algorithms of different approaches, there will always exits the difference in their performances on different datasets and that raises a problem: whether one of the models is generally better than the other across multiple datasets, or is there any preference for the certain types of dataset for different methods? In this paper, we will compare the performance of Naïve Bayes and Logistic Regression on three different datasets. We will also compare the accuracy of different type of the datasets on the same method.

1 Introduction/Objective

Supervised learning is crucial in natural language processing and there are mainly two types of methods in it, which are generative model and discriminative model (Afshin, et al., 2019). Both methods are commonly and wildly used in NLP. However, it seems that discriminative model is more prevailing than generative model. Just like Andrew claimed in his article, “Indeed, leaving aside computational issues and matters such as handling missing data, the prevailing consensus seems to be that discriminative classifiers are almost always to be preferred to generative ones.” (Andrew et al., 2002). Indeed, generative model has some shortcoming like the complexity in computing, but it is not a sufficient reason to explain that why generative model is not as preferred as discriminative model, and it also has its own advantages like only need a small amount of training data in the classification process (Merry Anggraeni et al., 2019). So, it will be a very meaningful to discuss why generative model are more preferred in NLP even though they both have pros and cons. However, it is a very general topic if we just compare the generative model and discriminative model since each model has many different methods and it is impossible to compare all of them each by each. To be more realistic, we will only compare Naïve Bayes as one of the generative models and Logistic Regression as one of the discriminative models to see their performance and the preference on different types of data.

We mainly have two objectives in this paper. First, compare the performance of Naïve Bayes, Logistic Regression, and the Baseline on the same dataset. Since we are applying models on several datasets, if one of the models is way outperformed than the other, we may conclude that one method is generally better than the other. If they perform differently on different dataset, we will try to analyze why this happens.

Secondly, we will compare the accuracy with same method on different datasets. Three datasets that we will analyze has three totally different feature to discriminate, attitudes, authorships, and the authenticity. It will be highly likely that different datasets have different preference in methods, and we will try to find out whether it is true base on the result we have.

2 Dataset

For this research, we are going to use three different datasets to compare the performance of the Naïve Baye and the Logistic Regression on different types of data. The first data we use is Fake and real news dataset, which originally were two datasets, one contains a list of articles considered as "real" news and one contains a list of articles

considered as "fake" news. We firstly combine two datasets and create a column called true_or_fake. If the article is considered as "real" news, the value on column true_or_fake will be "true", otherwise it will be "fake". After data cleaning, the new dataset has 44916 rows and 2 columns, which are texts and labels.

The second dataset is Tweet data. It has two columns, name of the person who tweeted and tweet content. Since there are more than two authors in this dataset, we remove all rows expect tweets written by Barack Obama or Neil deGrasse Tyson. After data cleaning, the dataset contains 5110 rows and 2 columns.

The last dataset we will use is Amazon Fine Food Reviews dataset. It consists of reviews of fine foods from Amazon. The data span a period of more than 10 years, including about 500,000 reviews up to October 2012. Reviews include product and user information, ratings, and a plain text review. It also includes reviews from all other Amazon categories. We classify ratings as two groups: reviews that have 5 stars rating and reviews do not have 5 stars rating. Finally, we save them in a new column called FiveStar. If the review has a 5-star rating, it is labeled as "Yes". If it doesn't, it is labeled as "No". Because the origin dataset is too large, we decide to only extract first 10,000 rows of it as our sample data. Then, we drop all unimportant columns and only leave column Text and FiveStar.

3 Background

The comparison of the Naïve Bayes and the Logistic Regression fulfills different goals. It can make people who are not experts understand the details of how a prediction was made (Biran and McKeown, 2014). It is also used to see which model has better performance. Pranckevicius and Marcinkevicius (2017) compared these two models to reveal the relationship between classification accuracy, the size of training set and n-gram properties by analyzing short texts for product review data form Amazon. On the other hand, Tsangaratos and Ilia (2016) used it to prove the hypothesis. They state that the Naive Bayes has better generalization ability against Logistic Regression and Naïve Bayes is more reliable for constructing landslide susceptibility maps. They validated their thoughts by comparing the Naïve Bayes and the Logistic Regression. Based on the result of Ng and Jordan's study (2001), although as discriminative learning, the Logistic Regression has lower error rate on prediction, the Naïve Bayes, a generative learning, may also approach its asymptotic error much faster. In addition, people also try to improve the performance of models before the comparison. Lasso and Ridge Logistic Regression can be applied for comparison and a form of Tf-idf term weighting with cosine normalization can be used for text representation (Genkin et al., 2007). On the other hand, Guo (2010) made some modifications for the Naïve Bayes to make it have better performance on the text data and improve the accuracy of output.

Unlike methods mentioned above, we investigate the performance of the Naïve Bayes and the Logistic Regression model on tweets, Amazon food reviews, and news data to see which model people should use according to the data they analyze. In this study, like Genkin (2017), we use Tf-idf to extract features.

4 Methodology

Different evaluation metrics such as accuracy, precision and recall, F1 score, sensitivity and specificity, and ROC curve and AUC will be used to evaluate the classification models. And these metrics will be compared between different models and within the same model. In other words, these metrics will be compared among Naïve Bayes, Logistic Regression, and the Baseline when they are trained and validated on the same dataset. Also, this paper will compare these metrics on the three different datasets using the same model to see if the domain or length or certain characteristics of a particular dataset will have a great effect on the model performance.

4.1 Naïve Bayes

Naïve Bayes is a generative model that is widely used in NLP. It is mainly based on the Bayes Theorem to calculate the probabilities and make the prediction. Naïve Bayes assumes all the features are all independent from each other. Even though the theorem behind the algorithm is simple, it needs quite a lot computing the calculate the probability. However, this method requires relatively less amount of training data to determine all the parameters that calculation needs (Merry Anggraeni et al 2019).

Methods	accuracy	precision	recall	F1_score	ROC AUC
Multinomial NB	94.0	91.7	96.7	94.1	NA
Logistic Regression	95.9	96.4	95.5	95.9	98.9
Zero-rule	50.1	NA	NA	NA	NA

Table 1: Tweet

Methods	accuracy	precision	recall	F1_score	ROC AUC
Multinomial NB	93.8	937	93.3	93.5	NA
Logistic Regression	98.8	98.7	98.7	98.7	99.9
Zero-rule	52.3	NA	NA	NA	NA

Table 2: News

Methods	accuracy	precision	recall	F1_score	ROC AUC
Multinomial NB	75.3	78.6	45.1	57.3	NA
Logistic Regression	78.2	75.2	60.7	67.1	84.2
Zero-rule	63.3	NA	NA	NA	NA

Table 3: Amazon Reviews

4.2 Logistic Regression

As a discriminative classifier, Logistic Regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative distribution function of logistic distribution. Logistic Regression is commonly applied to all sorts of NLP tasks, and any property of the input can be a feature. In this paper, Logistic Regression is used to train and validate on the three text datasets from different domains: Tweets, News, and Amazon Reviews.

4.3 Comparison

Different evaluation metrics such as accuracy, precision and recall, F1 score, sensitivity and specificity, and ROC curve and AUC will be used to evaluate the classification models. And these metrics will be compared between different models and within the same model. In other words, these metrics will be compared among Naïve Bayes, Logistic Regression, and the Baseline when they are trained and validated on the same dataset. Also, this paper will compare these metrics on the three different datasets using the same model to see if the domain or length or certain characteristics of a particular dataset will have a great effect on the model performance.

5 Results

5.1 Comparison between datasets

As the result shown in the tables above, if we compare the accuracy through different dataset, we quite surprisingly found that Tweet dataset are generally has the higher accuracy and Amazon Reviews has the lowest accuracy. For the Naïve Bayes, both Tweet and News dataset has an extremely high accuracy closing to 94%. However, it drops significantly when dealing with the Amazon Reviews dataset. As for the Logistic Regression model, the result is almost the same. Both News and Tweet dataset has extremely higher accuracy in prediction, which are 98.8% and 95.9%. As for Amazon Reviews, the accuracy is still quite low which equals to 78.2%. The result shows that, at least for Naïve Bayes and Logistic Regression model, the datasets that contains the authenticity or the authorship are easier to predict, and the prediction has generally higher accuracy. As for the datasets which need to predict the attitude, the accuracy is generally low for both Naïve Bayes and the logistic Regression model.

5.2 Comparison between classification models

As is shown in table1, in tweets data, the logistic Regression model has higher accuracy, precision and F1 score while the Naive Bayes model has higher recall, which means that the Naive Bayes model returns most of the relevant results.
As is shown in table 2, in news data, the logistic Regression model has higher accuracy, precision, recall and F1 score than the Naive Bayes model.
As is shown in table 3, in Amazon Reviews data, the logistic Regression model has higher accuracy, recall and F1 score, while the Naive Bayes model has higher precision, which means that it returns more relevant results than irrelevant ones.
The results show that the Logistic Regression classification method for has achieved the highest classification accuracy in comparison with Naïve Bayes (Pranckevicius and Marcinkevičius,2017) and it performs better on the three datasets.

6 Discussion

In this paper, the first finding is that the models using the Logistic Regression method perform better compared to those using the Naive Bayes method. This can be due to the fact that the Logistic Regression algorithm doesn't make as many assumptions as that of the Naive Bayes algorithm. The assumptions that the Naive Bayes algorithm mainly make are:

1. The Bag of Words assumption, which assumes that the position of the words in the document does not matter.

2. Conditional Independence assumption, which assumes that the feature probabilities are independent given the class.

However, in the real world, the order of words matters, and they are not independent. A phrase like "this movie was incredibly awful" or "what a great movie to make me fall asleep" shows an example of how both of these assumptions don't hold up.

The second finding is that Amazon Reviews sentiment classification has the worst results compared to authorship attribution for tweets and fake news detection. This may be caused by the fact that people might use sarcasm to express their negative sentiments using positive words and the presence of word ambiguity.

Last but not least, Naive Bayes and Logistic Regression have relative strength and perform better on aspects. As is shown in the results, the Naive Bayes model has higher recall than the logistic Regression model using tweets data and has higher precision than the logistic Regression model using Amazon Reviews data. It is not advisable to say which model is better than the other without a comprehensive comparison. In theory, Naive Bayes has a higher bias, but lower variance compared to Logistic Regression. If the data set follows the bias, then Naive Bayes will be a better classifier. So, with the small training data, model estimates may overfit the data using Logistic Regression. And as is mentioned before, Logistic Regression, which is discriminative, makes a prediction for the probability using a direct functional form whereas Naive Bayes, which is generative, figures out how the data was generated given the results.

References

Rahimi, A., Li, Y., & Cohn, T. (2019). Massively multilingual transfer for NER. *arXiv preprint arXiv:1902.00193*.

Merry Anggraeni et al 2019Literation Hearing Impairment (I-Chat Bot): Natural Language Processing (NLP) and Naïve Bayes Method

Ng A Y, Jordan M I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes[C]//Advances in neural information processing systems. 2002: 841-848.

Or Biran, Kathleen McKeown. Justification Narratives for Individual Classifications. 2014. Department of Computer Science, Columbia University.

Tomas Pranckevicius, Virginijus Marcinkevicius. 2017. Comparison of Naïve Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers

for Text Reviews Classification. Vilnius University, Institute of Mathematics and Informatics.

Tsangaratos P, Ilia I. 2016. Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size[J]. Catena, 145: 164-179.

Andrew Y. Ng, Michael I. Jordan. 2001. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. NIPS'01: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, pages 841–848.

Genkin, D. Lewis, and D. Madigan. 2007. Large-scale Bayesian logistic regression for text categorization. Technometrics, 49(3):291–304.

Qiang Guo. 2010. An Effective Algorithm for Improving the Performance of Naïve Bayes for Text Classification. Second International Conference on Computer Research and Development. 699 – 701.

Comparison of Naïve Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification by Tomas PRANCKEVIČIUS, Virginijus MARCINKEVIČIUS