CS276 Programming Exercise 2

NagaChaitanya Vellanki
vellanki@stanford.edu

May 30, 2011

1 Introduction

In this project, I have implemented Multivariate Naive Bayes(MVB), Multinomial Naive Bayes(MNB), Complement Naive Bayes(CNB), Weight-normalized Complement Naive Bayes(WCNB), Transformed Weight-normalized Naive Bayes(TWCNB). I have also implemented Chi-Square feature selection method and used k-fold cross validation to the test the accuracy of the classifiers. The implementation of the above classifiers is done considering efficiency, modular design.

2 Multivariate Naive Bayes, Multinomial Naive Bayes Classifiers

The implementation of MVB, MNB classifiers was trained and tested on 18828 messages. Logrithms of probabilities, add-one or Laplace smoothing were used to prevent underflow, over-fitting issues. MNB performed better than MVB since it takes the term frequencies into account. The accuracy of these classifiers based on the 18828 messages is shown in Table 1.

Table 1: MVB, MNB

Classifier	Correct Predictions	Accuracy
MVB	16001	84.98%
MNB	17893	95.03%

3 Feature selection using Chi-Square

Chi-Square feature selection method was used to select the Topk words from each newsgroup. The TopK words from each newsgroup were combined into a set for a given value of Topk and the MNB, MVB were retrained using only the topk words from Chi-Square feature selection. The accuracy of MNB decreased initially since it takes into account term frequencies and the accuracy of MVB improved since small number of features were considered. The accuracy of MNB improved as more words were selected from each newsgroup.

Table 2: MVB Chi-Square, MNB Chi-Sqaure

TopK Words	Correct Predictions MVB	Correct Predictions MNB
300	16258	14291
400	16357	15390
500	16370	16008
600	16354	16407
700	16351	16672
800	16377	16892
900	16412	17040
1000	16409	17156
1100	16401	17266
1200	16398	17334
1300	16414	17368
1400	16407	17405
1500	16396	17446

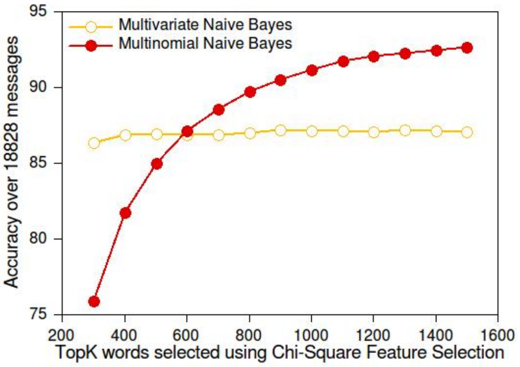

Figure 1: Chi-Square Feature Selection

4 K-fold Cross Validation

For K-fold cross validation, I have divided the messages in each newsgroup into K subsets. In iteration 1, the first subset from each newsgroup was considered to be part of the test data set and the rest of the subsets were used for training. In iteration 2, the second subset from each newsgroup was considered to be part of the test data set and the rest of the subsets were used for training. For

K =10, the training set size would be 16952 and the test set size would be 1876 for a total set of 18828 messages. The above process was repeated for K times. In each iteration, training and testing data sets were used with MVB, MVB Chi-Square, MNB, MNB Chi-Square, CNB, WCNB and TWCNB. The validation was performed for K= 10, 15 and 20. Top 300 words from each newsgroup were selected using Chi-Square during this process for MVB Chi-Sqaure and MNBChi-Sqaure.

Table 3: K-fold cross validation, Avg. number of Correct Predictions on Test set for value of K

K	Training Set Size	Test Set Size	MVB	MVB Chi-Square	MNB	MNB Chi-Square
10	16952	1876	1456	1557	1654	1356
15	17582	1246	972	1104	1100	900
20	17896	932	727	775	823	675

Table 4: K-fold cross validation, Avg. number of Correct Predictions on Test set for value of K

K	Training Set Size	Test Set Size	CNB	WCNB	TWCNB
10	16952	1876	1690	1678	1672
15	17582	1246	1121	1115	1113
20	17896	932	839	834	832

5 Improving Multinomial Naive Bayes Classifier

I have improved the MNB classifier by implementing various suggestions in the paper [1]. The following versions of MNB were implemented in order

1. Complement Naive Bayes(CNB)
2. Weight-normalized Complement Naive Bayes(WCNB)
3. Transformed Weight-normalized Complement Naive Bayes(TWCNB) with
   (a) Term Frequency(TF) transform
   (b) Inverse Document Frequency(IDF) transform
   (c) Length Normalization(LN)

The training data has different number of messages in each newsgroup, this causes MNB to choose one the larger newsgroup. CNB avoids this kind of skew in the training data. WCNB normalizes the newsgroup weights since some newsgroups can have more dependencies between words and can violate the independence assumption in Naive Bayes. The TF transform will lower the counts of terms which have large counts. The IDF transform will down weigh the common words and increase the weight for rare terms. LN is done since long messages can have more terms and can contribute to large term counts. In my observation TWCNB performed better than WCNB and CNB. The improvements achieved more accuracy than the MNB, The accuracy of the improvements compared to MNB are shown in the Table 3.

Table 5: Accuracy of CNB, WCNB, TWCNB compared to MNB on 18828 messages

Classifier	Correct Predictions	Accuracy
MNB	17893	95.03%
CNB	18128	96.28%
WCNB	18092	96.09%
TWCNB(TF)	18104	96.15%
TWCNB(TF, IDF)	18360	97.51%
TWCNB(TF, IDF, LN)	18372	97.57%

6 Experimenting with different techniques

The words in the corpus are lowercased and stemmed while reading in by the MessageFeatures class. I disabled stemming, lowercasing on the MessageFeatures class by commenting the code out and I observed the accuracy of TWCNB increased to 98.65% from 97.57%.

Table 6: Accuracy of classifiers on 18828 messages after stemming and lowercasing are disabled

Classifier	Correct Predictions	Accuracy
MVB	15755	83.67%
MVB Chi-Square(top 300)	16154	85.79%
MNB	18157	96.43%
MNB Chi-Square(top 300)	11483	60.98%
CNB	18375	97.59%
WCNB	18343	97.42%
TWCNB(TF, IDF, LN)	18575	98.65%

Also, removing email address , numbers and hyperlinks improved the accuracy of MNB Chi-Sqaure from 75.90% to 79.68%.

7 References

[1] Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, and David R. Karger. Tackling the poor assumptions of naive bayes text classifiers. In In Proceedings of the

[1] "Tackling the poor assumptions of Naive Bayes Text Classifier"

Twentieth International Conference on Machine Learning, pages 616623, 2003.
[2] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. 2008. Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA.