CSC 47-02: Machine Learning & Its Applications (SPRING 2016)

Project 1: Naïve Bayes

Kate Evans
Computer Science

Instructor
Dr. Vinayak Elangovan, Ph.D.
Department of Computer Science

Submitted On: 2/24/16

TABLE OF CONTENTS

1. Introduction

Naïve Bayes classifiers are commonly used for text classification. For this project, a Naïve Bayes classifier was implemented to classify SUV reviews as positive reviews or negative reviews. Customer written SUV reviews were found online and manually classified as positive or negative and split into various sizes of training and testing data. [4] Then a Java program was developed to take these reviews and train the classifier. The classifier then tests the remaining reviews and classifies them as positive or negative. This report details the design of this program and the results in produced for each size of training and testing data. The program also has the option to implement evidential learning. The design and results of this type of learning are also detailed in this report.

2. Theory/Background

A Naïve Bayes classifier is a basic type of probabilistic classifier commonly used for supervised learning. It is based on Bayes Theorem which states states that $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$. [1] The classifier uses this theorem to determine what class a given data set belongs to based on the probability of the data set's various characteristics belonging to a specific class. [2] A common application is text categorization where a "Bag of Words" is used to identify keywords in a text and the classifier calculates probabilities based on the frequency of these keywords in each class. [2]

There is another form of Bayes Theorem that can be used for Naïve Bayes classification. This equation is $P(W|C) = \frac{count(W,C)+1}{count(C)+V}$. [3] This equation uses the frequency of a word in a specific class, the number of words found in that class, and the total number of unique words to calculate the probability of that word belonging to that class. [3] It also adds one to the number of frequency of the word in the class as a smoothing factor. [3] This is the equation used in this project.

3. Design

Figure 1 below shows the relationship between the four major methods used to train the Naïve Bayes classifier and test it on SUV reviews. The main method controls the flow of the program. readTraining() is called to to read the training data from a text file and based on the user designated data split. For instance the instance the user can choose to train on fifty percent of the data and test on the other fifty percent or they can choose to train on eighty percent of the data and test on the remaining twenty percent. It also reads in the "Bag of Words" from a text file. For this classifier there are nineteen keywords found in the "Bag of Words" each of which commonly appear in SUV reviews. It returns the initial training table containing the keywords found in each SUV review and how it is classified. It also counts the total number of reviews it reads and how many of them are classified as positive or classified as negative. The table below displays what the training table would look like.

TRAINING TABLE 50/50:

DOC	Words	Class
1	love,regret,great,great,not,issues,not,not,	P
2	not,good,replaced,not,not,never,	N
3	not,never,not,never,not,	N
4	love,love,comfortable,	P
5	not,comfortable,good,problem,never,	N
6	great,comfortable,great,not,problems,	P
7	good,issues,	P
8	great,great,recommend,	N
9	problems,problems,never,problems,problems,	N
10	great,excellent,great,	P
11	good,comfort,recommend,	P
12	excellent,comfort,good,great,great,	P
13	great,replace,love,	P
14	unreliable,issues,not,happy,	N
15	not,happy,not,not,not,	N

Table 1 Displays the training data for a classifier trained on fifty percent of the training data.

Next, train() is called and passed the the training table to calculate prior probabilities and conditional independence for each word in the bag of words. After this point the test() method is called. However, this process does vary slightly based on whether or not the user selected to perform non-evidential learning or evidential learning.

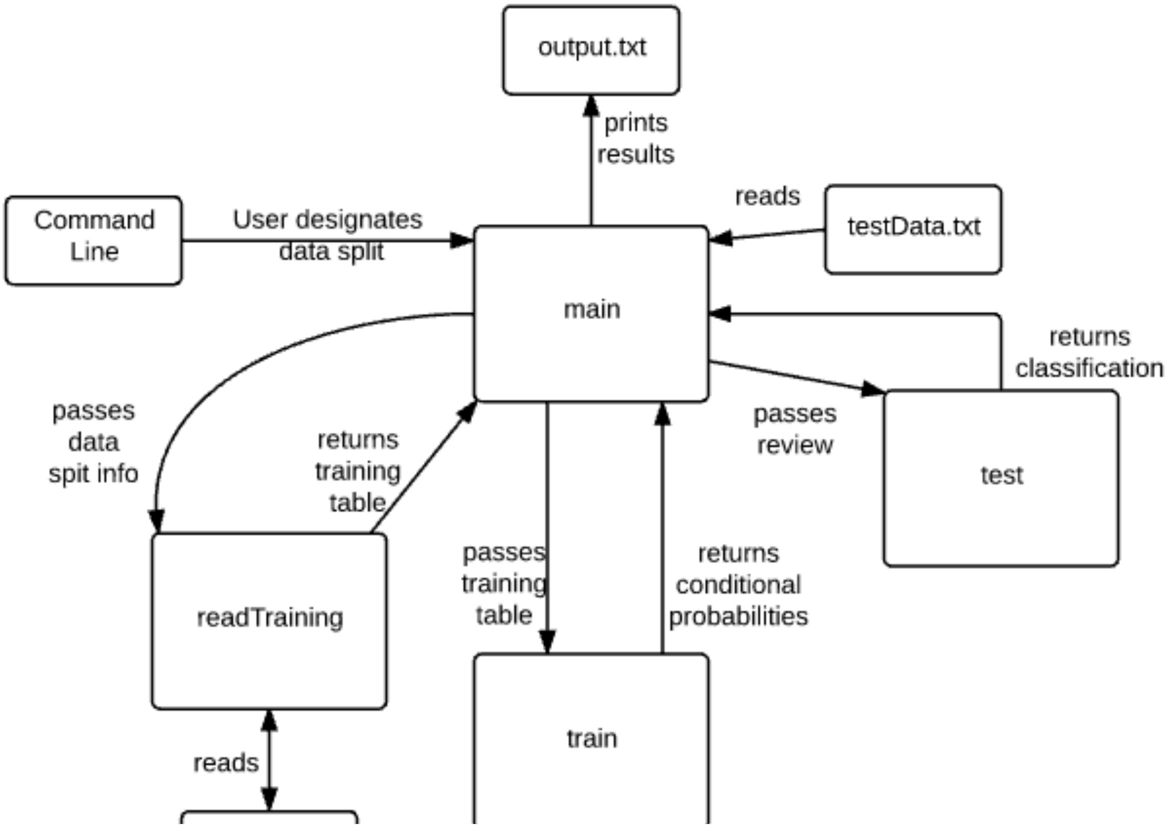


Figure 1 Naïve Bayes Classification Design Diagram

3.1 Non-Evidential Learning

For non-evidential learning the process for testing the data is quite simple. The main method reads in the test data file from a text file and passes each review to test() as a parameter. test() then returns the keywords it extracted and the classification for the review. The result is then recorded to be output to the command line and a text file at the end of the program.

3.2 Evidential Learning

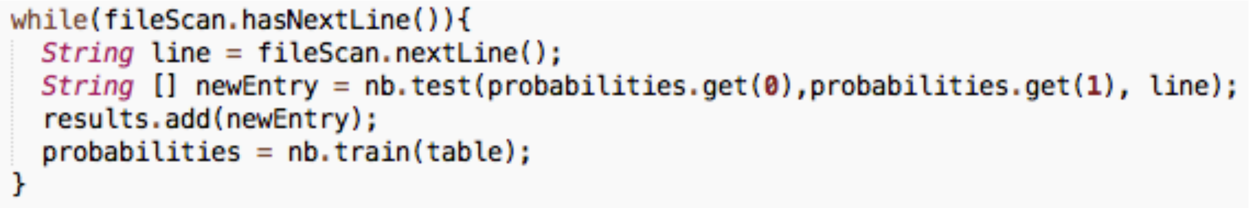
For evidential learning the classification for each tested review is added to the training table and the classifier retrained. For this program this type of learning can only be selected for initially training on eighty percent of the data and tested using the remaining twenty percent. Table 2 shows the initial training table for this form of learning.

TRAINING TABLE 80/20:

DOC	Words	Prediction
1	issues,replace,love,	N
2	comfortable,good,	P
3	love,love,never,	P
4	great,recommend,	P
5	great,love,	P
6	happy,	P
7	not,not,recommend,not,not,happy,terrible,	N
8	bad,never,love,	P
9	love,comfortable,happy,	P
10	good,	N
11	love,good,bad,great,recommend,	P
12	terrible,	N
13	love,	P
14	bad,bad,great,recommend,bad,problems,	N
15	replaced,not,unreliable,never,problems,regret,not,	N
16	love,regret,great,great,not,issues,not,not,	P
17	not,good,replaced,not,not,never,	N
18	not,never,not,never,not,	N
19	love,love,comfortable,	P
20	not,comfortable,good,problem,never,	N
21	great,comfortable,great,not,problems,	P
22	good,issues,	P
23	great,great,recommend,	N
24	problems,problems,never,problems,problems,	N

Table 2 Displays the training data for a classifier initially trained on eighty percent of the training data.

Like in the non-evidential learning portion of this program the main method reads in the testing data from a text file and sends each review to test() to be classified. test() then returns the extracted keywords and the classification. The result is then added to the training table and train() is called again to recalculate prior probabilities and conditional independence. The training table is expanded and ends up looking like Table 3 below. The next review is then read and processed. The results are also recorded to be output at the end of the program. The code section below shows how this occurs.



```
while(fileScan.hasNextLine()){
  String line = fileScan.nextLine();
  String [] newEntry = nb.test(probabilities.get(0),probabilities.get(1), line);
  results.add(newEntry);
  probabilities = nb.train(table);
}
```

FINAL TRAINING TABLE 80/20 EL:

DOC	Words	Class
1	issues,replace,love,	N
2	comfortable,good,	P
3	love,love,never,	P
4	great,recommend,	P
5	great,love,	P
6	happy,	P
7	not,not,recommend,not,not,happy,terrible,	N
8	bad,never,love,	P
9	love,comfortable,happy,	P
10	good,	N
11	love,good,bad,great,recommend,	P
12	terrible,	N
13	love,	P
14	bad,bad,great,recommend,bad,problems,	N
15	replaced,not,unreliable,never,problems,regret,not,	N
16	love,regret,great,great,not,issues,not,not,	P
17	not,good,replaced,not,not,never,	N
18	not,never,not,never,not,	N
19	love,love,comfortable,	P
20	not,comfortable,good,problem,never,	N
21	great,comfortable,great,not,problems,	P
22	good,issues,	P

23	great,great,recommend,	N
24	problems,problems,never,problems,problems,	N
25	great,excellent,great,	P
26	good,comfort,recommend,	P
27	excellent,comfort,good,great,great,	P
28	great,replace,love,	P
29	unreliable,issues,not,happy,	N
30	not,happy,not,not,not,	N

Table 3 Displays the training data for a classifier having undergone evidential learning.

4. Results

4.1 50% Training Data, 50% Testing Data

The table below displays the results from testing fifty percent of the data after training the classifier on the other fifty percent of the data. Fifteen SUV reviews were tested. The classifier produced a correct classification for ten of the reviews and an incorrect classification for the remaining five. The classifier incorrectly classified four negative reviews as positive and one positive review as negative.

RESULTS TABLE 50/50:

DOC	Words	Prediction	Answer
1	issues,replace,love	P	N
2	comfortable,good	P	P
3	love,love,never	P	P
4	great,recommend	P	P
5	great,love	P	P
6	happy	N	P
7	not,not,recommend,not,not,happy,terrible	N	N
8	bad,never,love	P	P
9	love,comfortable,happy	P	P
10	good	P	N
11	love,good,bad,great,recommend	P	P
12	terrible	P	N
13	love	P	P
14	bad,bad,great,recommend,bad,problems	P	N
15	replaced,not,unreliable,never,problems,regret,not	N	N

Table 4.1 This table displays the results of the Naïve Bayes classifier trained on 50% of the data. It shows the words found in each review, if the classifier classified it as a negative or positive review, and what class the review actually belonged to.

4.2 80% Training Data, 20% Testing Data

The table below displays the results from testing twenty percent of the data after training the classifier on the other eighty percent of the data. Six reviews were tested. The classifier correctly classified all six of the reviews.

RESULTS TABLE 80/20:

DOC	Words	Prediction	Answer
1	great,excellent,great,	P	P
2	good,comfort,recommend,	P	P
3	excellent,comfort,good,great,great,	P	P
4	great,replace,love,	P	P
5	unreliable,issues,not,happy,	N	N
6	not,happy,not,not,not,	N	N

Table 4.2 This table displays the results of the Naïve Bayes classifier trained on 80% of the data. It shows the words found in each review, if the classifier classified it as a negative or positive review, and what class the review actually belonged to.

4.3 80% Training Data, 20% Testing Data with Evidential Learning

The table below displays the results from testing twenty percent of the data. The classifier was initially trained on the other eighty percent of the data. Then after each test review was classified it was added to the training data and the classifier included it in its calculations. Six reviews were tested and added to the training data. The classifier correctly classified all six of the reviews.

RESULTS TABLE 80/20 EVIDENTIAL LEARNING:

DOC	Words	Prediction	Answer
1	great,excellent,great,	P	P
2	good,comfort,recommend,	P	P
3	excellent,comfort,good,great,great,	P	P
4	great,replace,love,	P	P
5	unreliable,issues,not,happy,	N	N
6	not,happy,not,not,not,	N	N

Table 4.3 This table displays the results of the Naïve Bayes classifier initially trained on 80% of the data and employed evidential learning. It shows the words found in each review, if the classifier classified it as a negative or positive review, and what class the review actually belonged to.

5. Conclusion

Overall, the Naïve Bayes classifier developed for this project performed fairly well. The accuracy of its results did vary based on how much training data was used with fifty percent training data performing less accurately than eighty percent training data. However, this was not surprising as more training data allows the classifier to calculate stronger probabilities.

5.1 50% Training Data, 50% Testing Data

As shown in the results section of this report the classifier trained on fifty percent of the data produced a correct classification for ten of the reviews and an incorrect classification for the remaining five. The classifier incorrectly classified four negative reviews as positive and one positive review as negative. This shows that the classifier was accurate 66.667% of the time which shows it is not a very strong or accurate classifier. This is probably due to the lack of training data. Out of its incorrect classifications four of them were Type I errors as they incorrectly classified as negative review as a positive review. Only one review was a Type II error as it was a positive review incorrectly classified as a negative review.

5.2 80% Training Data, 20% Testing Data

The classifier trained on eighty percent of the data correctly classified 100% of the test data. This accuracy was most likely due to the larger amount of training data compared to the classifier trained on fifty percent of the reviews. Due to this high level of accuracy, there was no difference between the non-evidential learning and the evidential learning classifiers implemented. There may have been more errors in either of the cases if more data had been collected. This project used thirty SUV reviews. If more had been collected there would have been more to train the classifier with but also more for the classifier to test and possibly incorrectly classify.

6. References

1. Elangovan, Vinayak, Dr. "Naïve Bayes." The College of New Jersey. 2 Feb. 2016. Lecture.
2. "Naive Bayes Classifier." *Wikipedia*. N.p., n.d. Web. 24 Feb. 2016.
<https://en.wikipedia.org/wiki/Naive_Bayes_classifier>.
3. "Text Classification and Naïve Bayes." Standford University. Lecture.
<https://web.stanford.edu/class/cs124/lec/naivebayes.pdf >
4. "Consumer Car Reviews." *cars.com*. N.p., n.d. Web. 17 Feb. 2016.
<http://www.cars.com/go/crp/consrevwidget.jsp>.