Machine Learning Assignment 1
Naive Bayes

Group Members

Ritesh Kumar Singh	2019H1030154H
Simran Batra	2019H1030024H
Parth Bhope	2019H1030023H

Dataset Description

The dataset contains 1000 reviews and their sentiment i.e. 0 or 1 indicating the class label. The dataset contained 500 reviews that were labelled 0 and 500 that were labelled as 1.
The reviews contain symbols , numbers, and english words.

Approach

We have built the model with two approaches and compared the results.

1. Build the model without any text preprocessing.
2. Text preprocessing followed by building the model

Text Preprocessing

The text consisted of some uppercase and mostly lowercase words so all the alphabets were converted to lowercase using string function lower() in python as the lowercase and uppercase version of the same word have the same meaning and importance in classification.

The reviews contain many stopwords like a an the is which do not help in finding the context of the reviews.We took a list of english stopwords of 191 words from the internet and put it in a set() and checked if a word was present in this set then that stopword was discarded. This gave us all the important words in the reviews.

The documents (reviews) contain symbols like @#%!,’’ which do not contribute to the classification process hence these symbols were removed.Now if some words like Ah!!

Were present in the review which wouldn't be removed in the stopwords removal step we repeated the stopword removal step to eliminate any remaining stopwords.

5 Fold split

The dataset was split into 5 parts, the tuples were randomly selected using numpy's sample function, in each iteration 1,2,3,4,5 the ith fold was used as a testing set and the remaining 4 were used to build the model. The size of the vocabulary when preprocessing was not done on an average was 2000, and when preprocessed was 1500.

Building the model

We maintained dictionaries (map of words and their frequencies in each class) to store the frequencies of each unique word in the vocabulary made using the training example. Prior probabilities of each class were also calculated as the count of class 0 examples divided by the total examples.

Laplacian smoothing

The frequency of each word is added with a pseudo count of 1, so for words that are not present in the class have a count of 1 and not 0 which prevents the probability of the testing example becoming 0.

Testing

For each testing review the classifier assumes that each word is independent of the other and hence the product of class conditional probability of each word is multiplied with prior probability of that class.

The class conditional probability of a word

$$= P(\text{word}_i / \text{class}=0) = \frac{\text{count}(\text{word}_i \text{ in class0 reviews}) + \text{pseudo count}(1)}{(\#\text{unique words in}(\text{class0}) + \text{ total vocab size})}$$

The denominator is the sum of total vocabulary size and class vocabulary as we are using Laplacian smoothing

Handling out of vocabulary words

The model has built the vocabulary only using the training data. The testing examples that have words not present in the corpus were not neglected as the pseudo count will make the numerator as 1 in the class conditional probability.
For the words in the vocabulary we have stored the counts , using which we calculate the class conditional probability,

The probability that a review belongs to class 0 is compared to that of class 1, the review is assigned a label 0 if the probability of it belonging to class 0 is more than class 1 and vice versa.

Results

Approach 1: No text preprocessing applied
Average Accuracy of 0.747 +- 0.0331 was obtained over the 5 folds.
Average F Score of 0.7228 +- 0.0464

Approach 2 : Text preprocessing applied
Average Accuracy of 0.768 +- 0.0256 was obtained over the 5 folds.
Average F Score of 0.7298 +- 0.0354

Conclusion

In the first scenario when the text was not preprocessed the accuracy was slightly lower than the second scenario as due to the cleaning processes we are left with only the important words that make up the context of the review which leads to a better accuracy.