CMPE 442 Assignment #2

Due Monday, April 6.

Note: The assignment is to be done individually. You are going to submit the report and code separately. Please do not include codes in the report. Writing comments in the code is encouraged. Do not send me your codes to find the error!!! Ask conceptual questions.

In this assignment, you are going to implement Naïve Bayes document classifier and apply it to the 20 newsgroups dataset. In this dataset, each document is a posting that was made to one of 20 different newsgroups. Your goal is to write a program which can predict which newsgroup a given document was posted to. You are provided with the *TextClassifier.ipynb* file that contains some auxiliary stuff. You can improve the functions given in the file depending on your needs.

Data:

The dataset is already divided into training and testing set. You can fetch it from scikit datasets. Fetching and getting the testing and training set is given in the *TextClassifier.ipynb* file that is provided to you.

Vocabulary:

Before computing the parameters you have to construct the vocabulary (dictionary) from both train and test sets. Prior to vocabulary construction, we need to preprocess the data: remove numbers, white spaces, stop words, punctuations, etc. The function, *preProcessing(text)*, that does all of these preprocessing steps is given in the *TextClassifier.ipynb* file that is provided to you. This function takes the text as an input and returns tokens (individual words) back. Run and see the text before sending to the function and after to understand what is being done there.

In order to construct a dictionary send each text document that is in both test and train set to *preProcessing(text)* function and add the words sent back to the dictionary.

Model :

Once you construct the dictionary, estimate the parameters using **multinomial event modeling**. Note that the parameters are estimated using train set only. Do not use test set at learning stage! Use **Laplace smoothing** in order to avoid zero probabilities.

Q1 [20 pts]: Report the size of the dictionary that you constructed.

Q2 [10 pts]: What are the class priors?

Q3 [30 pts]: Report your overall testing accuracy (number of correctly classified documents in the test set over the total number of test documents), and print out the confusion matrix (the matrix C, where c_{ij} is the number of times a document with category j was classified as category i).

Q4 [20]: Are there any newsgroups that the algorithm confuses more often than others? Why do you think this is?

Q5 [10]: Report on 10 words for each category that are strong indicators of that category label.

Q6 [10]: How could you possibly increase the accuracy of your classifier? Suggest.