PRML ASSIGNMENT 3 REPORT

NAME- BIBHUTI MAJHI ROLLNO- CS22M031

Naive Bayes Spam Classifier:

Data reference:

1) Dataset downloaded from https://www.kaggle.com/datasets/wanderfj/enron-spam?resource=download

Preprocessing:

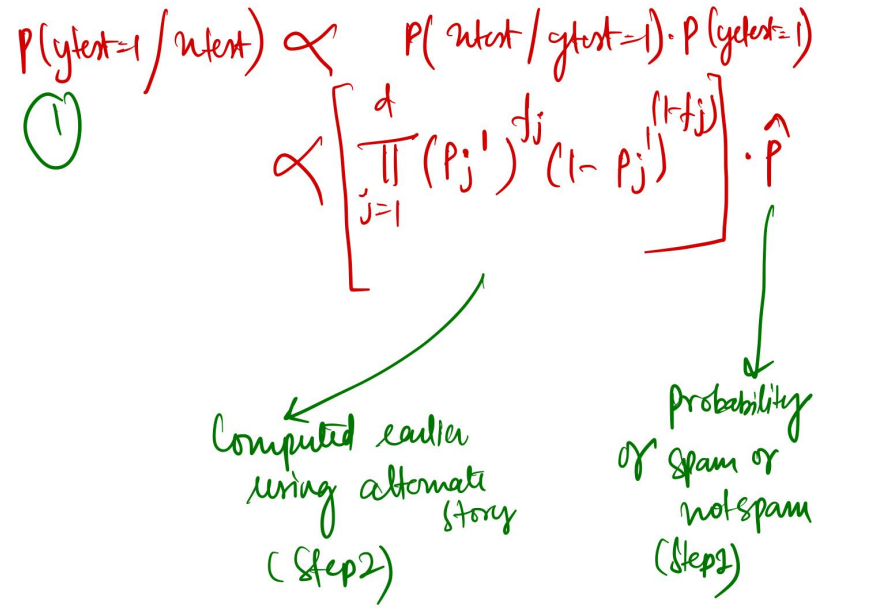
1) In the given dataset we are given two folders, one has all the spam emails and another one has all the non-spam emails.
2) I had to make a dictionary of words for training purposes, but the emails have many unnecessary characters and symbols so in preprocessing I removed all these unnecessary symbols and divided the email into a list of words all in lower characters.
3) I traversed through all the emails both spam and non-spam and preprocessed them one by one and after preprocessing maintained a dictionary that stores unique words found in the emails after preprocessing.
4) After traversing through all the emails we now have a dictionary of 50499 words.

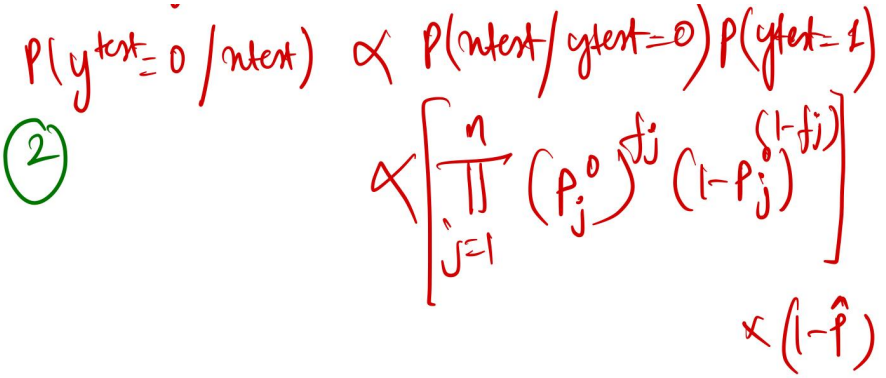
Training the model:

1) Now similar to the Dictionary that we made I maintained two more dictionaries that store the probability of that word being in spam or ham.
2) To do this I had to traverse all the words of spam emails again and stored the {word, frequency} key-value pair in the dictionary spamwords.
3) Same thing I did for the Ham emails and created a dictionary hamwords that stores the {word,frequency} key-value pair.
4) For handling the case that a particular word may not appear in the set of spam or ham emails, I added a mail with all the words that are in the dictionary, meaning added 1 to all the values of the dictionary spamwords if there is any word having count of occurrence = 0 in spamwords.
5) Same thing I did for hamwords dictionary.
6) For each {word,frequency} pair I divided the frequency by the no of words in the dictionary to get the probability.
7) Now we have both dictionaries storing the key-value pair { word, probability of occurrence of word in spam or ham emails}
8) Phat is the prior that says the probability of an email to be spam or non-spam. Calculated Phat as the ratio of no of spam email / total no of emails.
9) I got Phat = 0.2901

Prediction for test email:

1) For a given test email I preprocessed it and got the list of words present in that mail.
2) Then created a feature vector of size equal to the size of the dictionary. Each feature[i] can take value 1 or 0 denoting that if that word[i] is present in the email or not.
3) Calculated probham = Probability(test email is spam / feature vector) and Probspam = Probability(test email is non-spam/ feature vector) by the following formula for naive bayes.





4) If (probspam > probham) then return 1
   Else return 0
5) I took another set of emails that are spam which was not used in the training purpose and checked how many emails does my algorithm classifies as spam and found the accuracy by dividing with no of emails in the folder.
6) Training Accuracy = no of spam classified by algorithm / no of emails = 87.47%

SVM Classifier:

Preprocessing:

1) For each email from spam and nonspam folders I made their feature vector and stored that in a list called Xtrain and their corresponding label in Ytrain. Label 1 for spam and label 0 for nonspam.
2) From sklearn imported svm and then made and instance of SVM classifier using svm.SVC(C = 0.1,kernel = ‘linear’)
3) I trained the model by using svc.fit method and passing the Xtrain and Ytrain as the parameters.
4) Selected Xtrain as the testing data and checked the Training accuracy which turned out to be 99.85%