PRML Assignment 3

Shrung D N
ME19B168

The model training and validation is done using Naive Bayes algorithm using two datasets, with a total of 11,300 mails.

The procedure followed for training is as follows:

- Download all dependencies required for data parsing, to export model etc
- Download and read data into a data frame
- Edit the data frame so that it contains only “message” and the associated “label” and change this “label” to only 0 and 1 (from “ham” and “spam” respectively)
- Split the data into training and validation data in the ratio 80:20
- Train the NB (Naive Bayes) model with training data:
    - Messages are processed before being fed for training:
    - Unwanted information such as brackets, numbers, punctuation marks, repeating words and other symbols are removed from the messages.
    - Words in the messages which have the number of characters less than a particular threshold (this is an hyperparameter. The value of 4 seemed to give the best results and hence the threshold is taken as 4 or more letters per word in this case) are removed. This is done as shorter words very rarely have important information stored in them. Doing this reduces the total number of words in the dictionary (dictionary is the collection of all the important words (i.e, features) that has appeared during the training process. The model makes use of these features for classification.
    - The words in the message is then tokenized, which is a process of converting a string of the entire message into a list of strings of words
    - The tokenized message is then parsed, where stop words (common words such as “a”, “an”, “the”) are removed and is then stemmed (a process where words are converted into its base form (for example “did” would be converted to “do”). This is again done to reduce the total number of words that would appear in the dictionary, as past tense of words more or less carry the same information as the present tense in most cases.
    - These processed messages are fed into the model, which then calculates the MLE (Maximum Likelihood Estimate) of the various parameters (probabilities) of the generative model. The result of MLE is made to undergo Laplace Smoothing so that the classification is done better.
    - The model is then exported, and is used to classify whether a mail is spam (+1) or not spam/ham (0) using the estimates of various parameters.

Some characteristics of the trained model:

- Number of features extracted (i.e number of words in the dictionary) = 84215
  (Note: Not all these features are important, as most of these would have come from a single mail or so, unlike the important features, which would have come from multiple mails. This is automatically taken care of during MLE, that is the probability associated with a less important feature would be very less as well)
- Validation accuracy = 93.6 %
- Approximate time taken to train = 6 minutes

Some details of the files present in the submitted folder:

- Train_model.pdf - pdf of the ipynb file used to train and export the model
- NB.py - Contains the class of the model, which is used to understand the class data when the model is imported
- NBmodel.pkl - is the exported model using the pickle module
- Get_predictions.py - is the file to be run to get the predictions of the mails contained in the folder "test" - the file prints a dictionary with the predicted value associated with each text file.

Other algorithms tried:

- SVM: SVM performed worse than NB algorithm. This could be due to the fact that the emails aren't linearly separable (due to the large number of features, it is very unlikely that a single hyperplane would divide the feature-space into two classes
- KNN: This algorithm takes a long time to predict (due to the large dataset) and it also requires the original data set to make predictions (for example, NBmodel.pkl occupied about 12 MB of memory if it contained the original data, and occupied about 4 MB when the original data was deleted from it)

Hence among the three, Naive Bayes algorithm gave the best result and hence it is used.