2 Project 2 (Naive Bayes)

Due date: March 3

In this project the Naive Bayes Spam filtering is implemented. The data are all in the given zip file. All the preprocessing was done for this dataset. Please do not redistribute the dataset outside this class.

- The files train-features, train-features-50,...,train-features-400 contain pre-processed versions of messages used for training. First collumn is the message number, second column is the word's counter in the dictionary, third column is the number of occurrences of the word in the current email.

- The files train-labels, train-labels-50,...,train-labels-400 are labels of the messages in the above files. 1 corresponds to spam, and 0 corresponds to non-spam messages.

- test-features is a set of messages to test your classifier.

- test-labels are labeling the messages in test-features into spam/non-spam (1/0)

For the final program you should use the files train-features and test-features, however while developing your project you can see what happens if you use smaller training sets.

To understand the structure of the files you can find attached training spam message number 1 (which is itself pre-processed by removing all the numbers, punctuation, common words, etc). This message is message number 351 according to training-labels (350 are spam and 350 are ham). After ordering the message alphabetically, the first word "addres" is labeled 2 in our dictionary and it appears ones in the message. The word "app" labeled 6 appears once, the word "application" labeled 7 appears twice, "area", labeled 9 appears 3 times, etc...

For implementation, based on the training data you can build a dictionary with all the words in the messages together with number of occurences in spam and ham messages. You do not necessarily have to use a dictionary, you can just store the number of occurences in a matrix. Based on this file, using the Naive Bayes algorithm you can calculate the probability whether a certain message is a spam or ham. The file test-features contains the test messages. The file test-labels contains the labels of those messages as spam or ham. For each message in the file test-features use the Naive Bayes algorithm to decide whether it is a spam or a ham message.