Naive Bayes Spam Classification

(from Prof. Berg, COMP 590-133)

In this assignment you will be building a Naive Bayes spam classification system for predicting whether an email is spam or ham.

Data: Preprocessed spam and ham email data: emails.tar.gz.

To perform classification, we split the data into 4 parts: spam training documents (listed in spamtraining.txt), ham training documents (listed in hamtraining.txt), spam testing documents (listed in spamtesting.txt), and ham testing documents (listed in hamtesting.txt). The spam/ham training documents will form your labeled training set (the documents whose label you know) and the spam/ham testing documents will form your testing set (the documents whose label you need to predict).

Computing Features

As features for spam/ham classification we will use a lexicon of words. Compute a lexicon consisting of the set of unique words occurring more than *k* times in the entire document collection (*k* is a number you should set experimentally by examining the lexicon and classification accuracy produced for a few values of *k*, e.g. 1,2,10,...).

Training

The goal of the training stage is to estimate the parameters of the model from the training set. Parameters include the **likelihoods** P(Word | class) for every word (w) in your lexicon for each class (spam and ham). The likelihood estimate is defined as:

$$P(Word = w \mid class) = \frac{\text{\# of times word w occurs in training examples from this class}}{\text{total \# of words in training examples from this class}}$$

In addition, you should smooth the likelihoods to reduce the effect of small probabilities. *Laplace smoothing* is a very simple method that increases the observation count of every value ‘w’ by some constant *m*. This corresponds to adding *m* to the numerator above, and *m***V* to the denominator (where *V* is the number of words in your lexicon). The higher the value of *m*, the stronger the smoothing. Experiment with different integer values of *m* (sample a few values between 1 and 50) and find the one that gives the highest classification accuracy.

You should also estimate the parameters for the priors P(class) as the empirical frequencies of the classes in the training set (i.e. percentage of spam and ham documents).

Testing

You will perform maximum a posteriori (MAP) classification of spam or ham according to your learned Naive Bayes model. Suppose a test document contains words w1 , w2 , ... , w200 . According to this model, the posterior probability (up to scale) of each class given spam or ham is:

$$P(\text{class}) \cdot P(\text{w1} \mid \text{class}) \cdot P(\text{w2} \mid \text{class}) \cdot \ldots \cdot P(\text{w200} \mid \text{class})$$

To avoid underflow, you should compute the log of the above quantity:

$$\log P(\text{class}) + \log P(\text{w1} \mid \text{class}) + \log P(\text{w2} \mid \text{class}) + \ldots + \log P(\text{w200} \mid \text{class})$$

For each test document, compute the above probabilities for the spam and ham classes. Classify the test document as the most probable class according to your computation. Also include in your report a discussion of whether it would make any difference to use maximum likelihood (ML) classification for the provided data.

Measuring Performance

1. Compute the overall accuracy -- what percentage of documents in the testing set were classified correctly?
2. Compute the spam accuracy -- what percentage of documents in the spam testing set were classified as spam?
3. Compute the ham accuracy -- what percentage of documents in the ham testing set were classified as ham?