Constructing a Naive Bayes Spam Filter

This homework aims at developing and testing a program that allows to classify e-mails into one of two classes: spam or legitimate. We will encode the e-mail by defining an event $S$, which corresponds to the fact that the e-mail is spam. Its complement will be $S^c$, namely the event that the e-mail is instead legitimate. It is in general assumed that the prior probability of these events is equal

$$P(S) = P(S^c) = \frac{1}{2}.$$  

(Notice that in reality, there are more spam e-mails circulating out there, than legitimate ones.)

An e-mail is classified on the basis of the occurrence or not of a certain list of words $W_1, W_2, \ldots W_m$. Here $W_i$ is the event that word $i$ (say, ‘hello’) appears in the e-mail. Its complement $W_i^c$ is the event that the word does not appear. These occurrences are assumed to be conditionally independent given $S$ or $S^c$. Recall that this means

$$P(\tilde{W}_1 \cap \tilde{W}_2 \cap \cdots \cap \tilde{W}_m | S) = P(\tilde{W}_1 | S) P(\tilde{W}_2 | S) \cdots P(\tilde{W}_m | S),$$  

$$P(\tilde{W}_1 \cap \tilde{W}_2 \cap \cdots \cap \tilde{W}_m | S^c) = P(\tilde{W}_1 | S^c) P(\tilde{W}_2 | S^c) \cdots P(\tilde{W}_m | S^c),$$

where, for each $i$, $\tilde{W}_i = W_i$ or $\tilde{W}_i = W_i^c$ and this identity holds for all such choices.

Use Bayes formula to express the posterior expectation that an e-mail is spam, given which of the words in the list actually occur.

An e-mail is classified to spam if the posterior expectation of $S$ is 0.5 or larger.

You are required to implement this method and test it on a publicly available dataset, maintained at UC Irvine at

http://archive.ics.uci.edu/ml/datasets/Spambase

The data are contained in the file spambase.data which is in csv (comma separated variable) format. One advantage of this dataset is that the curators did the text processing work for us, and all we are left with is the classification problem.

Here is a partial description of the dataset (we will not use all of the data available):
The file `spambase.data` contains 4601 rows. Each row corresponds one e-mail, and has 58 entries. Entries 49-57 will not be used for the homework.

The last entry of each row is a 1 if the e-mail was spam, and 0 otherwise.

The first 48 entries of each row correspond to frequencies of 48 words in the e-mail. In particular, the $i$-th entry is a 0 if the $i$-th word does not appear in the e-mail, and it is strictly positive if it does appear. Hence the event $W_i$ in our model above corresponds to the $i$-th entry being strictly positive, for $i \in \{1, \ldots, m\}$, $m = 48$.

In order to implement and test the Bayesian classification method discussed above you are required to:

1. ‘Put aside’ the last 200 instances of legitimate e-mails, and the last 200 instances of spam. These will be used for testing your classification algorithm.
2. Use the remaining 4201 e-mails to estimate the probabilities $P(W_i|S)$ and $P(W_i|S^c)$ for $i \in \{1, \ldots, m\}$, $m = 48$.
   A simple way to do this is to use the fraction of spam e-mails that contain word $i$ as an estimate for $P(W_i|S)$, and the fraction of non-spam e-mails that contain it as an estimate for $P(W_i|S^c)$. It is advisable to add to these fractions a small number (e.g. $1/4201$) to avoid the appearance of zeros among these probabilities.
3. Use these probability estimates and Bayes formula to compute the posterior probability that each of the 400 test e-mail is spam.
4. Classify it as spam if the posterior probability is at least 0.5, and as legitimate otherwise.
5. Report the number of ‘false positives’ (legitimate e-mails that were classified as spam) and ‘false negatives’ (spam e-mails that were classified as legitimate).
6. Sometimes a better estimator for the probabilities is the following (here the hat denotes the fact that we are estimating this probability)
   $$\hat{P}(W_i|S) = \frac{n(i; S) + k}{n(S) + k},$$
   where $n(i; S)$ denotes the number of spam training e-mails that contain word $i$, and $n(S)$ the total number of spam training e-mails. An analogous formula can be written for $\hat{P}(W_i|S^c)$. Further $k > 0$ (typically bigger than 1) is a parameter.
   Try to repeat the above with a few distinct values of $k$ of your choice and report the false negatives and false positives in each case.