Controlling Spam

Comparing Statistical and non-statistical methods

1 Introduction

Steve Raber, of Ciphertrust has said: "Spam is the cholesterol of the Internet, it is just clogging up everything."

Many companies are investing heavily in anti-spam research and development and are looking at innovative ways that technology can contribute to helping solve the spam problem for users worldwide.

Many methods to combat spam in e-mail have been invented over time. To introduce yourself to the issues involved in spam and the methods that are being used to stop it, go to

tt http://paulgraham.com/antispam.html and click on "Stopping Spam." There you will see that there are statistical based and non-statistical based methods to filter spam.

Question 1: Summarize in a small table the advantages and disadvantages of each of the methods described there. Make the table short. Do not just copy and paste what is in the web page.

Question 2: Now go back to the main page and choose "A Plan for Spam" and read it. After you are done reading it go back and then choose "Better Bayesian Filtering." Summarize briefly how the Bayesian method does the filtering.

2 Main Activity

In this activity, we are going to simulate how the Bayesian filters work.

Question 3: Choose among your and your friend's mails several e-mails at random days and random times (chosen as we said in class a simple random sample should be chosen), some spam and some not spam (5 of each should suffice). Having the mails in a file will help a lot in what comes next. So ask your friends to forward you some of that e-mail. List all the words that appear in each type of email. Find the frequency distribution of words in the spam mail and the frequency table of words in the non-spam mail. Select from each frequency table the words that do not appear in the other frequency table. This will give you the frequency distribution of words unique to the

spam mail and the frequency distribution of words unique to non-spam mail. Order each table with the most frequent words and their frequencies appearing first in decreasing order of frequency. Compute also the relative frequencies of each word.

Question 4: Now select from the frequency tables that you prepared in Question 3, the words that appear in both types of mails, and compute their frequencies and relative frequencies in the spam and non-spam mail. Again, order the words in decreasing order of frequency.

Question 5: For the tables in Question 4, do a Chi-square test to see if the frequency of use of those common words is statistically the same in the two groups.

Question 6: This is the testing part: For the tables in Question 3, I will give you in two weeks several pieces of email that I received without telling you what kind of mail it is. Use the results in questions 3, 4 and 5 to determine whether the mail is junk mail or not. To do this, first do for each message:

- Put in one column the words in the new message that only appeared in spam in the training set.
- Put in another column the words in the new message that only appeared in non-spam in the training set.
- Put in another separate column the words in the new message that appeared in the "common" words in the training set.

Question 7: Determine now whether the new e-mail messages (the test set) are spam or not.

To catalog each of the new messages as spam or non-spam, use the following rules:

- If the frequency of spam-only words is larger than the frequency of non-spam-only words and the frequency of common-but-higher-in-spam words is larger than the frequency of common-but-higher-in-nonspam, then the message is spam.
- If the frequency of spam-only words is smaller than the frequency of non-spam-only words and the frequency of common-but-higher-in-spam words is smaller than the frequency of common-but-higher-in-nonspam, then the message is non-spam.
- If the frequency of spam-only words is larger than the frequency of non-spam-only words and the frequency of common-but-higher-in-spam words is smaller than the frequency of common-but-higher-in-nonspam, then the message is spam.
- Everything else is 50-50, so we don't know.

Question 8: Based on the outcome of part 7, what is the probability that your method has of catching spam with these messages? This would be the number of new messages that you catalogued correctly, divided by the total number of messages.

Now choose any of the other filtering methods (other than the bayesian) that you read about in Question 1, and determine whether the mail I am giving you is junk or not using that method. What is the probability that this other method has of catching spam with these messages?

Question 9: Based on the above, update the method you developed above, to incorporate what you have learned with the testing.

Basically, is what you have done from question 3 to 8 something (roughly speaking) similar to what the Bayesian method does? Of course, you are welcome to write an algorithm in any language that will do all the steps you do above. Feel free to improve anything.