Mehryar Mohri
Introduction to Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 1
September 19, 2011
Due: September 30, 2011

A. Naive Bayes

The objective of this exercise is to apply the Naive Bayes algorithm presented in class to a particular document classification problem. Download the following data set which contains 20 different newsgroups, each with $1{,}000$ messages.

`http://www.cs.nyu.edu/~mohri/mlu/newsgroup.tar.gz`

The learning problem consists of assigning a newsgroup class to an email message. There is a directory for each newsgroup and the file name for each message in a directory is a number. For each directory, sort the files in order of increasing numbers and use the first 750 files for training, the rest for testing.

1. Extract the total vocabulary of all the messages in all the directories, training plus test messages. Form the following three vocabularies V_1, V_2, V_3, obtained by disregarding the $n\%$ most frequent words, with $n = 10, 20, 40$. Report the sizes of all three vocabularies.

2. Use as features the presence (1) or absense (0) of a word of the vocabulary in the message. Use additive smoothing to estimate conditional probabilities and apply the Naive Bayes algorithm to the training and test set. Report the classification accuracy obtained for V_1, V_2, and V_3 on both sets.

   To do this, you can use simple Unix scripts, for example:

   ```
   cat * | LC_CTYPE=C tr -cs a-zA-Z '\n' | \
   sort | uniq -c | sort -nr
   ```

   to count the number of occurrences of each word in the files of the current directory, or

   ```
   (for i in `ls` ; do cat $i/*; done) | \
   LC_CTYPE=C tr -cs a-zA-Z '\n' |
   sort | uniq -c | sort -nr
   ```

to count them all from the root directory.

Alernatively, you can write special-purpose programs in the language of your choice or use a scripting language of your choice.

3. Report the *confusion table* on the test set: for each pair of newsgroups (a, b) count the number of times that label b is predicted when the correct label is a. Which group has the highest accuracy? Which two groups are most likely to be confused? Which two groups are the least likely to be confused?

B. Probability (bonus question)

Suppose we toss a fair coin multiple times. What is the probability of obtaining three consecutive heads (HHH) any time before two consecutive tails (TT)? (*hint*: you can express the probability of this event A as follows: $\Pr(A) = \Pr(A \mid H)\Pr(H) + \Pr(A \mid T)\Pr(T)$ and proceed similarly for $\Pr(A \mid H)$ and $\Pr(A \mid T)$; observe that $\Pr(A \mid HT) = \Pr(A \mid T)$).