Mehryar Mohri
Introduction to Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 1
Solution (written by Andres Muñoz)

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 absence (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?

   Solution: The following are the results for the case where the 10% most frequent words were taken off the vocabulary. The confusion table was capped because of matters of space. The vocabulary was extracted only from the training files. The error rate for the other vocabulary sizes were $.37$ and $.48$. The most confused categories were misc.forsale and comp.sys.mac.hardware. There were several pairs of categories that were never confused. Vocabulary Size: 108584, Error rate: 0.29.

	aa	cg	cosw	ibm	mac	win	mfs	aut	mot	rbb	rsh	cry	ele	med	sp	src
alt.atheism	153	1	0	1	1	0	0	1	0	1	2	3	1	1	6	13
comp.graphics	2	138	4	14	16	18	7	0	3	5	2	10	8	1	6	2
comp.os.ms-win	7	7	137	21	12	19	10	4	2	1	4	2	10	0	2	1
comp.sys.ibm.pc	2	8	9	162	18	4	16	1	2	0	0	6	7	4	4	2
comp.sys.mac	2	7	2	10	163	2	9	7	6	4	2	3	13	4	2	1
comp.windows.x	2	19	6	10	13	155	2	1	7	3	2	5	9	1	7	0
misc.forsale	6	6	4	18	22	3	150	3	1	2	3	1	11	2	5	2
rec.autos	2	5	1	4	4	1	8	191	7	1	1	2	1	5	4	0
rec.motorcycles	1	1	0	0	2	1	3	8	218	1	0	1	3	1	1	1
rec.sport.baseball	5	2	0	0	3	0	1	1	2	218	8	0	1	1	3	1
rec.sport.hockey	3	1	0	2	3	0	0	0	1	9	221	1	1	1	2	0
sci.crypt	3	4	0	2	4	4	3	3	4	3	0	189	3	2	8	2
sci.electronics	5	11	0	10	11	1	6	15	5	4	3	9	152	7	4	1
sci.med	2	9	1	1	5	1	6	2	7	3	2	1	6	182	6	4
sci.space	5	8	0	2	4	2	0	3	2	2	5	2	5	4	190	2
soc.religion.christian	5	0	0	0	2	1	0	0	1	2	1	1	1	3	1	215

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 $\Pr(A \mid HT) = \Pr(A \mid T)$).

Solution: Conditioning on the outcome of the first toss, $\Pr(A)$ can be rewriten as

$$\Pr(A) = \Pr(A \mid H)P(H) + \Pr(A \mid T)P(T) = \frac{1}{2}\Big[\Pr(A \mid H) + \Pr(A \mid T)\Big]. \quad (1)$$

Proceeding in the same way with $\Pr(A \mid H)$ yields:

$$\Pr(A \mid H) = \frac{1}{2}\Big[\Pr(A \mid HH) + \Pr(A \mid HT)\Big].$$

Since the coin tosses are independent, the probability of event A given that the outcome of the first toss is heads and that of the second tails is the same as the probability of A conditioned on the first outcome being tails. Thus, $\Pr(A \mid H)$ can be rewritten as

$$\Pr(A \mid H) = \frac{1}{2}\Big[\Pr(A \mid HH) + \Pr(A \mid T)\Big]. \tag{2}$$

Applying the same argument to $\Pr(A \mid HH)$ leads to

$$\Pr(A \mid HH) = \frac{1}{2}\Big[\Pr(A \mid HHH) + \Pr(A \mid HHT)\Big] = \frac{1}{2}\Big[1 + \Pr(A \mid T)\Big],$$

using the fact that $\Pr(A \mid HHH) = 1$: thus after three consecutive heads, the probability of A is the same as that of obtaining two consecutive tails, which is guaranteed to occur. Plugging in this expression in (2) gives

$$\Pr(A \mid H) = \frac{1}{4} + \frac{3}{4}\Pr(A \mid T). \tag{3}$$

Now, $P(A \mid T)$ can be analyzed in a similar way:

$$P(A \mid T) = \frac{1}{2}\Big[\Pr(A \mid TH) + \Pr(A \mid TT)\Big] = \frac{1}{2}\Pr(A \mid H),$$

using the fact that by definition of A, $\Pr(A \mid TT) = 0$. Thus (4) gives

$$2\Pr(A \mid T) = \frac{1}{4} + \frac{3}{4}\Pr(A \mid T), \tag{4}$$

that is $\Pr(A \mid T) = 1/5$ and $\Pr(A \mid H) = 2\Pr(A \mid T) = 2/5$. In view of these equalities, (1) gives

$$\Pr(A) = \frac{3}{10}.$$