STAT 380 Final Exam, Spring 2017
Instructor: M. Haran, Penn State University

NAME:

I have neither given nor received any assistance in the taking of this exam.

Signature:

Instructions:

1. This is a closed book exam.
2. *Turn off all electronic devices*! You will be asked to leave if your electronic device rings, vibrates, or makes any sound during the exam.
3. Please verify that your exam paper contains all 10 questions.
4. To earn credit, write clearly, and show your work. When in doubt, explain things clearly. If you need extra paper then please write on the back of your paper.

DO NOT WRITE BELOW THIS LINE

Question	Marks	Max
1		8
2		12
3		6
4		14
5		12
6		10
7		6
8		12
9		12
10		8
Total		100

Q1 Suppose you have 1000 emails, of which 600 are spam and 400 are non-spam ("ham"). The goal is to develop a classification method based on these emails for classifying spam versus ham emails. Your friend has an approach that fits a logistic regression based on several characteristics of these emails. Call this Method A. You want to use a naive Bayes approach (using "bag of words" as capturing the email characteristics) that fits and predicts spam versus ham. Call this Method B. Describe clearly *in around 5 sentences/steps* how you would use cross-validation to determine whether you prefer Method A or Method B. No code necessary, but if you are vague in your description you will lose points. [8 pts]

- Randomly split the data into 10 groups of 100 emails each.
- Using each method, train a model on the first 9 groups and test it on the tenth group. Record the number of incorrectly classified emails for each model, call them Err_1^A and Err_1^B.
- Repeat the previous step, testing on a different group each time, until you have $\mathrm{Err}_1^A, \ldots, \mathrm{Err}_{10}^A$ and $\mathrm{Err}_1^B, \ldots, \mathrm{Err}_{10}^B$.
- The cross-validated misclassification rate for each method is the average of these values. $CV_A = \frac{1}{10}\sum_{i=1}^{10} \mathrm{Err}_i^A$ and $CV_B = \frac{1}{10}\sum_{i=1}^{10} \mathrm{Err}_i^B$.
- Choose the method with the smaller cross-validated misclassification rate.

Q2 Consider the chips database discussed in lecture about CPU (central processing unit) development of PCs (personal computers) over time. The database has the following variables (column): processor, date, transistors, microns, clockspeed, width, mips. Write SQL queries to do the following:

(a) Calculate how many rows are in the chips table. [3 pts]

```
SELECT COUNT(*) FROM chips
```

(b) How many chips have attribute (variable) width *not* equal to 32? [3 pts]

```
SELECT COUNT(*) FROM chips WHERE NOT width = 32
```

(c) How many chips are in each width group? [3 pts]

```
SELECT width, COUNT(*) FROM chips GROUP BY width
```

(d) What is the average micron for each unique value of width? [3 pts]

```
SELECT width, AVG(micron) FROM chips GROUP BY width
```

Q3 Suppose you are using a naive Bayes approach to figuring out whether an email message is spam or non-spam ("ham"). Consider an email that has the phrase "Are your taxes too high?" (everything else in the email is discarded as junk by your text mining code). Now suppose you have the following information in your training data of 600 spam emails and 400 ham emails: the phrase "Are your taxes too high?" shows up in 7% of all the training emails. This phrase shows up in 10% of your (training) spam emails and it shows up in 2.5% of your (training) ham emails. Based on just this information, estimate the odds that the email is spam. That is, what is your estimate of the ratio between the probability that the email is spam to the probability that the email is ham. It is enough to write out your answer in terms of products or ratios (e.g. $\frac{0.8\times 0.3}{0.2\times 0.5}$); you do not need to calculate the final answer. Do you think you will classify this email is spam or ham? Why or why not? [6 pts]

Define the following events:

$$\begin{aligned} S &: \text{the email is spam} \\ \neg S &: \text{the email is ham (i.e. not spam)} \\ T &: \text{the email contains the phrase “Are your taxes too high?”} \end{aligned}$$

By Bayes' Theorem,

$$P(S|T) = \frac{P(S)P(T|S)}{P(S)P(T|S) + P(\neg S)P(T|\neg S)} \text{ and } P(\neg S|T) = \frac{P(\neg S)P(T|\neg S)}{P(S)P(T|S) + P(\neg S)P(T|\neg S)}.$$

When we take the ratio the denominator cancels out, so we have

$$\text{odds} = \frac{P(S|T)}{P(\neg S|T)} = \frac{P(S)P(T|S)}{P(\neg S)P(T|\neg S)}.$$

Let's base our prior probabilities on the frequency of spam and ham in the training data. Since 600 of the 1000 training set emails are spam, let $P(S) = 0.6$ and $P(\neg S) = 0.4$.

$$\text{odds} = \frac{(0.6)(0.1)}{(0.4)(0.025)}.$$

It's clear that the numerator will be larger than the denominator, which means the odds are greater than 1. This indicates that we will probably classify the email as spam.

Q4 For each part of this problem, you are given a regular expression and a single string. Draw a box around *all* characters in the string that are matched by the regular expression. For example, if the regular expression is `ee` and the string is `beekeeper`, then you would answer b$\boxed{\texttt{ee}}$k$\boxed{\texttt{ee}}$per.

(a) Regular expression: `the` [2 pts]

The quick brown fox jumped over $\boxed{\texttt{the}}$ lazy dog in $\boxed{\texttt{the}}$ park.

(b) Regular expression: `ce.` [3 pts]

This is a senten$\boxed{\texttt{ce.}}$ This is $\boxed{\texttt{cer}}$tainly another senten$\boxed{\texttt{ce.}}$

(c) Regular expression: `ce\\.` [3 pts]

This is a senten$\boxed{\texttt{ce.}}$ This is certainly another senten$\boxed{\texttt{ce.}}$

(d) Regular expression: `^This` [3 pts]

$\boxed{\texttt{This}}$ is a sentence. This is certainly another sentence.

(e) Regular expression: `\\<.{3}\\>` [3 pts]

$\boxed{\texttt{How}}$ $\boxed{\texttt{now}}$ brown $\boxed{\texttt{cow}}$?

Q5 (a) Write a regular expression that will match any `td` HTML tags (both opening and closing tags, that is `<td>` or `</td>`). [6 pts]

```
</?td>
```

or

```
<td>|</td>
```

(b) Write a regular expression that will match any text that begins with `<td>` and ends with `</td>`. That is, it has to match paired `td` HTML tags as well as all characters in between the tags. (Hint: This should be similar to what you did in your homework on processing email messages.) [6 pts]

```
<td>.*?</td>
```

Note the `?` to make the star quantifier lazy instead of greedy. If we did `<td>.*</td>`, this would match everything from the very first `<td>` to the very last `</td>`, not each pair separately.

Q6 The lottery number `lotto` consists of a sequence of 6 numbers, where each of the 6 numbers is drawn at random from the set $\{X, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ (all numbers and one letter, X, which is the "wild card"). Suppose it costs 10 dollars to play and that you win if all 6 of the numbers in `lotto` are either even (2,4,6, or 8) or X. Your payouts are: $100 \times \text{Xcount}^2$ where Xcount is the number of Xs in `lotto`. Write a program in `R` to approximate the expected amout you win when you play 1 game of `lotto`. Hint: You will need to use Monte Carlo. [10 pts]

I used 0 instead of X in my simulation to avoid the hassle of dealing with both numeric and character types. Plus 0 is even, so checking if "all numbers are either even or X" simply becomes checking if all numbers are even.

```
simulate_draw <- function(cost) {
  sequence <- sample(0:9, size = 6, replace = TRUE)
  if (sum(sequence %% 2) != 0) return(-cost)
  xcount <- length(which(sequence == 0))
  100 * xcount^2 - cost
}

trials <- replicate(100000, simulate_draw(10))
mean(trials)
```

Q7 Write a function called yLim which takes as input the vectors x and y. This function returns a vector of length 2, which contains the values in y that occur in the positions of the minimum and maximum values of x. For example,

```
> x = c(100, 13, 1, 20)
> y = c(5, 7, 9, 0)
> yLim(x, y)
[1]  9  5
```

Since the minimum of x is in the 3rd element of x and the maximum of x occurs in its first position, the return value is a vector containing the 3rd and first values in y (which are 9 and 5, respectively). Additionally, the input x is required, and y has a default value of x. If x is shorter than y, then a message should be issued but the computation is carried out. If y is shorter than x, then the function is terminated and a message is issued. You may assume that there is a unique minimum and maximum in x and there are no NAs in either x or y. [6 pts]

```
yLim <- function(x, y = x) {
  if (length(x) > length(y)) stop(’y must be as least as long as x’)
  if (length(x) < length(y)) warning(’y is longer than x’)
  y[c(which.min(x), which.max(x))]
}
```

Q8 (a) Write a simple expression (1 line of code) in `R` to create a vector of 500 0s followed by 40 3s and 10 repeats of the sequence 1 through 5. To clarify, a vector of 5 0s followed by 4 3s and 3 repeats of the sequence 1 through 5 would look like this: 0 0 0 0 0 3 3 3 3 1 2 3 4 5 1 2 3 4 5 1 2 3 4 5 [4 pts]

```
c(rep(0, 500), rep(3, 40), rep(1:5, 10))
```

(b) Consider the following list, aList:

```
aList
  $x
  [1] "a" "b" "c" "d" "e"
  $mat
       [,1] [,2]
  [1,]    8    5
  [2,]    7    4
  [3,]    6    3
$zz
$zz$x
[1] 1 2 3
$zz$y
[1] 7 8 9 10 11
$zz$z
[1]  TRUE  TRUE FALSE FALSE  TRUE
$one [1] 100
```

Write down what will appear at the console when R evaluates each of the following expressions (note: some expressions may result in error messages. Also: pay close attention to parentheses!): [2 pts each]

i. `aList$x[2:4]`

```
[1] "b" "c" "d"
```

ii. `aList$mat + aList$one + 3`

```
     [,1] [,2]
[1,]  111  108
[2,]  110  107
[3,]  109  106
```

iii. `sapply(aList$zz, sum)`

```
x  y  z
6 45  3
```

iv. `length(aList[["zz"]][[1]])`

```
[1] 3
```

Q9 Recall the temperatures data sets that you have worked on/studied in homework and in lecture.

(a) Assume the first line of the file with url `http://www.stat.psu.edu/Temp.dat` contains the names of the columns of the data. Write the `R` command you will use to read in this data set into a data frame called tempSCE. [2 pts]

```
tempSCE <- read.csv('http://www.stat.psu.edu/Temp.dat', header = T)
```

(b) The first column is date, columns 2-4 are Tmax, Tmin, Tdat (maximum, minimum and average daily temperature respectively), and column 5 is ErrorFlag (1=error, 0=no error). Write a command to delete all rows that correspond to an ErrorFlag value of 1. [2 pts]

```
tempSCE <- tempSCE[which(tempSCE$ErrorFlag != 1), ]
```

(c) If any one of Tmax, Tmin or Tdat is -9999, remove the entire row of data. [2 pts]

```
tempSCE <- tempSCE[which(tempSCE$Tmax != -9999 &
                         tempSCE$Tmin != -9999 &
                         tempSCE$Tdat != -9999), ]
```

(d) Suppose date is in the form YearMonthDate, for e.g. 20030923 corresponds to September 23, 2003. Write `R` code that finds the minimum temperature for the year 2008. That is, your code should return the smallest minimum temperature value (smallest Tmin) for the year 2008. [6 pts]

```
min(tempSCE$Tmin[which(substr(tempSCE$date, 1, 4) == '2008')])
```

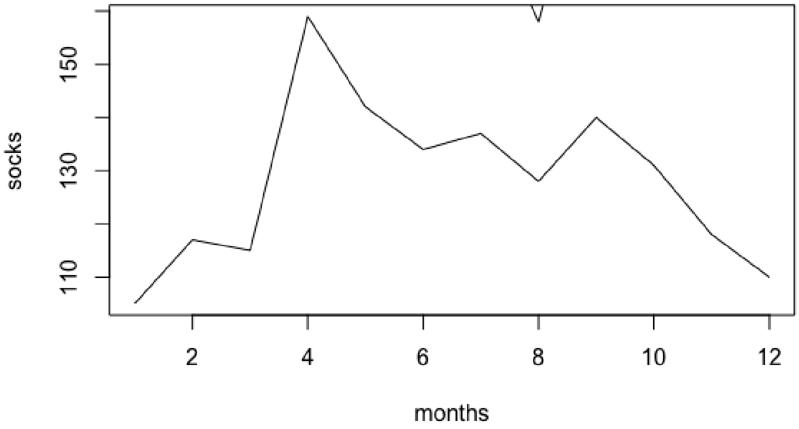
Q10 Suppose you work for a company that sells two products: socks and computers. You are asked to make a plot comparing the number of products sold per month over the last year. You have the following data:

month	1	2	3	4	5	6	7	8	9	10	11	12
computers	241	235	229	210	220	194	182	158	199	214	246	239
socks	105	117	115	159	142	134	137	128	140	131	118	110

You make a plot in R with this code:

```
plot(month, socks, type = 'l')
lines(month, computers)
```

This is the output you see:



(a) You expected to see a line for socks and a line for computers. Why is this not the case? What can you do to make both lines visible? [4 pts]

The line for computers is drawn, but it is almost entirely outside the range of the y axis. To fix it we need to use the `ylim` argument to `plot()` to include the computer data.

(b) List some other ways the plot could be improved. What needs to be done before showing it to your boss? [4 pts]

At minimum, students should say that they would add some way to differentiate between the two lines (line type or color), and a legend to tell which is which. I hope they also say make better labels (y should be “quantity”, x should have the month names instead of numbers), a descriptive title, etc.