CS 5751 – Spring 2019 – Homework 1
Assigned: 02/05/2020
Due: 02/19/2020 at 11:59 p.m.
Submit your file(s) to Canvas and a hardcopy in class on 02/21/2020
Maximum Grade: 100 pts.
Attendance is graded separately

Objectives: The objectives of this homework are the following:

- Setting up R and RStudio.
- Learn how to write functions in R.
- Learn how to use simple exploratory data analysis tools in R.
- Learn how to implement the gradient descent algorithm.
- Learn how to fit linear regression models in R.

Notes:

- This homework is to be done in groups of 2 people. There must be just 1 submission per group.
- Submit **a single PDF file** named: yourlastnames_lab1.pdf with the answers to your lab. To generate this PDF file, you **must use R markdown**. Check the video here for more information https://rmarkdown.rstudio.com/lesson-1.html.
- See the slides posted on canvas to learn how to install R and RStudio and to learn basic R commands.

Special Notes on the Dynamics of the Labs:

- Labs have only one submission.
- The TA will take attendance using canvas.

Activity 1: (Proof of Completion Due 02/05) (14 pts.) (Software Installation) Do the following tasks:

i. Download R from r-project.org
ii. Download RStudio from rstudio.com
iii. Install these packages: tidyverse, arules, caret, markdown.

For this exercise, there is nothing to submit. Just show the TA that everything is setup correctly.

Activity 2: (14 pts.) (Simple Exploratory Data Analysis) Using R, perform the following tasks:

a) **(Proof of Completion Due 02/05)** (2 pts.) Read the dataset contained in the file 'artificial_data.csv' into a data frame. This dataset contains only one column and this column is named *value*.
b) **(Proof of Completion Due 02/05)** (2 pts.) Plot a histogram of the attribute *value*. This histogram needs to have labels for the x and the y axes and needs to have a title. Do it with both base R graphics and with ggplot.

c) (7 pts.) Make a function named *central* which receives as input parameters: a vector *v* of real values, an integer *num_rounds*, an integer *sample_size* and that does the following:
   i. Draw *num_rounds* samples, each of size *sample_size,* with replacement from the vector *v.* Use the R function *sample* for this. For example, if *v = (7,8,4,5,10,11)*, then if we draw 2 samples each of size 3 we could obtain: {{7,10,4}, {11,11,8}, {5,5,7}},
   ii. For each sample *i* obtained in part (i), compute its sample mean $\bar{X}_i$ and its sample standard deviation s_i. In the example above, the sample means of each of the above 3 samples would be 7, 10, 5.66666 respectively. The sample standard deviation for each of the samples would be 3, 1.732, and 1.155, respectively. For computing the mean and standard deviation use the functions *mean* and *sd*.
   iii. Plot a histogram of the means $\bar{X}_i$ of all the samples. In the example above, you would make a histogram out of the three values {7, 10, 5.66666}. Use the function *hist* for this part.
   iv. Plot a histogram of the standard deviation σ_s of all the samples. In the example above, you would make a histogram out of {3, 1.732, 1.155}.
   v. The function must return a list containing just two real numbers: mean($\bar{X}_i$) and mean(s_i) over all the samples *i*. Continuing the example above, this means that you function returns the following list with only two elements: [(7+10 +5.66666) /3, (3 + 1.732 + 1.155) / 3] which is equal to [7.5555, 1.9623].

d) (3 pts.) Comment on what you observe in the histogram of part 2.c.iii and explain what is going on.

To help you with this exercise, check the slides on how to install R that I will post on canvas. Also check the slides containing an introduction to R that I will post on canvas.

Activity 3: (14 pts.) (Gradient Concept) Perform the following tasks related to the fundamentals of gradient descent.

a) Compute the gradient of the function $f(x, y) = 2x^5y^3 - \frac{3}{x}y^2$ by hand. Show all your work.
b) Find and plot the values of the gradient of the above function at the points $(1,0)$ and $(3,2)$. You can draw the plot by hand and scan it.

Activity 4: (14 pts.) (Fitting Linear Regression Models) Using R, do the following tasks:

a) Find the optimum value for x using the normal equations directly (do not use gradient descent).
b) Find the optimum value for x using the lm command in R (do not use gradient descent and do not use the normal equations).

Activity 5: (14 pts.) (Gradient Descent Basics) Run by hand and in a step-by-step fashion the gradient descent algorithm. Start from x_0 and compute up to and including x_2. Use $\lambda = 0.1$, x_0= [0.1, 0.1, 0.1], and H and z the same as the ones we used in class.

<u>Activity 6</u>: (30 pts.) (Gradient Descent Implementation) Using R, perform the following tasks related to the gradient descent method.

a) Implement the gradient descent algorithm in R from scratch. For this activity, you cannot use any package in R that already implements this method. You can start from the Matlab implementation that I showed you in class.
b) Run your algorithm on the H and z matrices that we used in class to find x. Choose 2 different vectors for x_0 and 3 different values for λ, then run your algorithm for all 6 combinations. Try to do your best so that the values you choose lead the algorithm to converge to the true (and only) solution.
c) Make a plot of $\log_{10}\|x_{i+1} - x_i\|_2$ as a function of the iteration number i, just as we did in class, for each combination of values for x_0 and λ. Your plot needs to have a proper title, and the axes need to be properly labelled. The title must include the value of λ and x_0 that you chose for that plot.
d) Comment on what you observe in your plots.