Homework 2

Your Name Here, Your UNI ID

2019-10-19

Question 1

Perform 20 iterations of gradient descent to minimize the given cost function, starting from the initial β given. Follow these steps as a guideline:

1. Select your own learning rate, and be aware you may need to adjust this parameter to achieve a suitable descent.

2. Calculate the gradient ($\triangledown C$) of the Cost function in terms of β.

3. Iterate to adjust β as such: $\beta_{t+1} = \beta_t + (learningrate) * \triangledown C(\beta_t)$. It would be helpful to do this in a loop.
4. Finally, print a dataframe with a row for each iteration and the updated value of β. If β is a vector with multiple components, let your dataframe include a column for each component.

Additionally, comment on the following points:

1. What value of β would minimize the cost function, and can β be calculated explicitly?

2. Does your gradient descent method improve your guess of β correctly?

a) $C(\beta) = (\beta - 3)^2 + 5$, $\beta_0 = 5$

b) $C(\beta) = \frac{1}{2}(Y - X\beta)^2$, $\beta_0 = \begin{vmatrix} 0 \\ 0 \end{vmatrix}$, note: X and Y matrices are given below

Question 2

Using the randomly generated 1100 by 3 data set, you are going to practice different clustering methods to learn how each method works, to observe how each provides the output graphically, and how to make a decision based from the output. To make the practice effective, you are going to let the last column of dataset be an integer and be the cluster number but treat the column as unknown value. Then you are going to cluster the dataset using only first two columns and test your results by verifying the total number of clusters.

1. Generate a random 1100 row dataset containing three columns, the first column from -1.5 to 1.5, the second from -3.5 to 1.5, and the third column with integers from 1 to 4.

2. K-mean method - Using the basic function $kmeans()$, determine the appropriate number of clusters. You can tabulate the number of clusters from 1 to 40 and the total within-cluster variances. Then plot the scree plot to visually support your decisions on the cluster number.

3. DBSCAN method - install the package called fpc and compute DBSCAN using $fpc :: dbscan()$ and $dbsca :: dbscan()$. Determine the optimal eps parameter value and draw the k-distance plot. The cluster using $fp :: dbscan()$ function provides the same result with the result using $dbsca :: dbscan()$ function. Verify which function provides the better computation.

4. Hierarchal clustering - calculate the pairwise distance between observations. Create various dendrograms using complete and average linkage. Cut the dendrogram into groups of 5, 6 and 7. Discuss which is the most appropriate number of groups.

5. Summarize the results from 2, 3, and 4.

Question 3

In this problem, you are going to apply unsupervise learning techniques learned from Question 2. The goal in this problem is to apply clustering techniques and identify the correct number of clusters.

1. Download the file `HW2_Q5_1_1.csv` from the HW#2 folder.
2. Repeat the process you have done in Question 2 part 2.
3. Repeat the process you have done in Question 2 part 4.
4. Discuss how the cluster number from 2 and 3 are the same or different.

Question 4

This problem is an extension of Question 3 to practice the logistic regression. The actual cluster number will be given later in HW#3.

1. Split the data into train and test dataframe by 0.8 and 0.2 ratios.
2. Using the cluster number from Kmean in Question 3-2, predict the cluster number and calculate its accuracy using logistic regression.
3. Do same with the result obtained from the hierarchal clustering.

Question 5

Using the different dataset, "HW2_Q6_1_1.csv", you are planning to do what you have done in Q4.

1. Unfortunately, the kmean function is not installed in your R package and you have to write the code from the scratch. Write a function call kmean.alt function that calculates the within cluster variance, aggregates the data by the cluster number, and plots "total within cluster vs. number of cluster". The actual cluster number will be given later in HW#3.
2. Add the cluster number to the last column of dataset, predict the cluster number using logistic regression and calculate the accuracy of your model.
3. Using the pre-instaled Kmeans function, compare the performance of part 1. You can similar to part 2.