CS559: Assignment 3
Due 4/14/2020 Tuesday 11:59 PM

Homework #3 is an individual work: each student must submit their own work. Use of partial or entire solutions obtained from others or online is strictly prohibited. Electronic submission on Canvas is mandatory.

Question 1.: 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.

a) 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.
b) K-mean method - 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.
c) 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.
d) Summarize the results from b) and c).

Question 2-1.: In this problem, you are going to apply unsupervised learning techniques learned from Question 1. The goal in this problem is to apply clustering techniques and identify the correct number of clusters.

a) Download the file **HW3_Q2_1_1.csv** from the Assignment 3 folder.
b) Repeat the process you have done in Question 1 part b).
c) Repeat the process you have done in Question 1 part c).
d) Discuss how the cluster number from 2 and 3 are the same or different.

Question 2-2: This problem is an extension of Question 2-1 to practice the random forest classification. The actual cluster number will be given later in Question 3.

a) Split the data into train and test dataframe by 0.8 and 0.2 ratios.
b) Using the cluster number from Kmean in Question 2-1-a), predict the cluster number and calculate its accuracy using random forest classification.
c) Do same with the result obtained from the hierarchal clustering result.

Question 3: Recall Questions 2-1 and 2-2, you have applied various unsupervised learning techniques to cluster the given dataset, **HW3_Q2_1_1.csv**, whose cluster ids (or classes) were unknown. In this practice, you are going to evaluate the work that you did on the previous homework. Download the training data set which is the same data from Questions 2-2 and 2-2 but with the true class label. The goal of Question 3 is to tune the hyper parameters of techniques to get exact cluster numbers.

a) Report hyper parameter values you used for K-means and the number of clusters you have found in Question 2-2.
b) Load the file **HW3_Q3_1.csv** (hereafter the new dataset in Question 3). Determine the number of classes (the cluster number in our case) on the dataset. Evaluate and comment on the quality and accuracy of your initial work in Question 2-2.
c) Using k-mean technique, tune the hyper parameters until you can cluster the **new dataset** having the exact cluster number you answered in number 2-2. The new class id integers may not be exact with the true class integer but the number of clusters should be the same.
    (a) Load the **new dataset** and cluster. Make a scatter plot of **new dataset** by the true class. Use the

first column on the *x-axis* and the second column on the *y-axis*. Calculate the centroids in clusters.
(b) Eliminate the third column (the class column).
(c) Cluster the new dataset using K-mean and tune the hyperparameters until you can obtain the
exact cluster numbers or number of classes. Report the new hyper parameters.
(d) Make the same plot by new cluster id obtained by the tuned hyper parameters.
(e) Calculate and tabulate the new centroids in clusters.
(f) Comment of the similarities and differences between two plots.

Question 4-1.: Using the different dataset, **HW3_Q4_01_1.csv**, you are planning to do what you have done in Question 2-2.

a) Unfortunately, the Kmean function is not available in your python 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 Question 4-2.
b) Add the cluster number to the last column of dataset, predict the cluster number using **logistic regression** and calculate the accuracy of your model.
c) Using the pre-installed Kmean function, compare the performance of part 1. You can similar to part 2.

Question 4-2.: Recall in Question 4-1, you have clustered the **HW3_Q4_01_1.csv** dataset and tested the classification accuracy using the obtained cluster id as classes with the **logistic regression** (hereafter the old model). In this practice, you will improve the classification of the model you made in Question 4-1 on the datasets **HW3_Q4_1.csv** and **HW3_Q4_2.csv** (hereafter ***new train*** and ***new test data*** in Question 4-2, respectively) by using the logistic regression model (hereafter the new model). Then you will improve the classification model further with different supervised learning techniques (for the simplicity, new learnings). Load the old model used in Question 4-1. If you have not done, you need to go back to Question 4-1 and save the model as "old model".

a) Using the ***new test data***, **HW3_Q4_2.csv** which contains the true class in the last column of the dataset, test the performance of the "old model". Comment on the results - number of identified classes and the accuracy of the model.
b) Using the ***new train data***, **HW3_Q4_1.csv**, run the **logistic regression** and determine the accuracy of the "new model" with the ***new test data***.
c) Use the **random forest** technique to classify the data. Just like a tuning the **logistic regression** practice above, you will use **HW3_Q4_1.csv** as the train and **HW3_Q4_2.csv** as the test dataset again.
    (a) Load the train data and use **random forest** for classification.
    (b) Calculate the accuracy of each learning using the test data
    (c) Assemble the learnings and determine the classification accuracy.
    (d) Make a table to present the accuracy of each learning.
d) Use the **gradient boosting** technique to classify the data. Just like a tuning the logistic regression practice above, you will use **HW3_Q4_01.csv** as the train and **HW3_Q4_02.csv** as the test dataset again.
    (a) Load the train data and use **gradient boosting** for classification.
    (b) Calculate the accuracy of each learning using the test data
    (c) Assemble the learnings and determine the classification accuracy.
    (d) Make a table to present the accuracy of each learning.
e) Write short paragraphs of your results - which learning has the highest accuracy, the lowest, advantages and disadvantages of learnings you tested.

CS559: Assignment 3
Due 4/14/2020 Tuesday 11:59 PM

Video Presentation (Due by 4/6th Tuesday 11:59 PM)

Jaka Jazbec - Decision Tree
Daniel Kadyrov - Random Forest
Manoj Menon - Boosting