Lab 12: Advance Data Analysis and Visualization

By: *Prof. N. Hemachandra & R. Deval* Date 25-Oct-2023

Objective: In this lab, students will use some basic Machine Learning techniques to cluster a given data set and some advanced data analysis for visualization.

Instructions: Below are some instructions. Please go through them carefully:

- Use Lab12_Practice_ClevelandHearData and Lab12_Practice_Clustering as practice files to get familiarized to various techniques to be used in this particular lab.
- Also, explicitly mention the assumptions used throughout your modeling technique.
- Use Matplotlib.pyplot, Pandas, Numpy, sklearn and other library documentation if you need help.
- Your task is to **submit the questions** asked below in **.ipynb file itself**.
- Use the traditional approach to name your files for submission:
  - **ROLLNUMBER_IE507_Lab12_Q*i*.ipynb;** for i^{th} question

Question 1: Advance Data Analysis: Clustering

Clustering using K-Means

Given data set $D = \{(x^i\}_{i=1}^m$ of points x^i, we wish to cluster these points into groups. The clustering should be done so that the points belonging to the same group (or) cluster should be similar compared to those from other groups (or) clusters.

Given the number of clusters K, the clustering is achieved by finding a partition $(C_1, C_2, \ldots, C_K) \subseteq D^K$, such that the following objective is optimized:

$$\min_{(C_1,C_2,\ldots,C_K)\subseteq D^K} \sum_{k=1}^{K} \sum_{x\in C_k} \|x - \frac{1}{|C_k|} \sum_{u\in C_k} u\| \tag{1}$$

This optimization algorithm can be equivalently written as:

$$\min_{(C_1,C_2,\ldots,C_K)\subseteq D^K} \sum_{k=1}^{K} \sum_{x\in C_k} \|x - \mu^k\| \tag{2}$$

where $\mu^k = \frac{1}{|C_k|} \sum_{u\in C_k} u$.

In general, solving this optimization problem is NP-hard. However an iterative technique called K-Means algorithm has been developed which can effectively cluster the points by finding a local optimum of the above optimization problem.

The idea of the algorithm is as follows:

- Input: Data set $D = \{(x^i)\}_{i=1}^m$.
- Start with a random initialization of means $(\mu^1, \mu^2, \ldots, \mu^K)$ (usually identified from the data set itself).
- Repeat:
  1. Construct partitions $(C_1, C_2, \ldots, C_K)$ such that C_j contains points from D which are closer to $\mu^j, \forall j \in \{1, 2 \ldots, K\}$.
  2. Recompute means $\mu^j = \frac{1}{|C_j|} \sum_{x\in C_j} x, \forall j \in \{1, 2, \ldots, K\}$.
     Until partitions $(C_1, C_2, \ldots, C_K)$ do not change.

Question 1.1: Problem based on S1.txt file

Link for S1.txt
Link for test.txt

1. Try to understand the K-**Means++** algorithm implemented in scikit-learn package.
2. For the data in S1.txt, vary the number of clusters K by choosing from the set $\{6, 7, 8, 9, 10, 11, 12, 13\}$ and use the data in the KMeans function of scikit learn package.
3. For each value of K, prepare the scatter plots depicting the clusters using different colors along with the cluster centers depicted in the same plot with a color different from those used for clusters.
4. Explain your observations about the clustering results you obtained when the value of K is increased from 6 till 13.
5. Consider the `test.txt` file given above and find the predictions for the points in `test.txt` for clustering obtained for each value of K. Plot the points in the scatter plot and indicate the predicted cluster labels.
6. Explain your observations about the predictions obtained for different values of K.
7. Can you suggest your procedure that can be used to find the best choice for the number of clusters?
8. Implement your procedure for the data from S1.txt and report the best choice for the number of clusters.
9. Explain how you can modify the data in S1.txt so that the mean of each column is 0 and the variance is 1. This procedure is called column normalization.
10. Write the appropriate code to do the column normalization.
11. On the new data set thus obtained where the columns have mean 0 and variance 1, repeat the clustering for K in $\{5, 6, 7, 8, 9, 10, 11, 12, 13\}$.
12. For each value of K, prepare the scatter plots depicting the clusters using different colors along with the cluster centers depicted in the same plot with a color different from those used for clusters.
13. Explain your observations about the clustering results you obtained when the value of K is increased from 5 till 13.
14. Explain how you will modify the test data given in the `test.txt` file so that it can be used for prediction. Using your idea, convert the data in `test.txt` for prediction and report the predicted labels. Prepare scatter plots where you plot the transformed data from `test.txt`.
15. Explain your observations about the predictions obtained for different values of K.
16. Using your procedure to find the best choice of the number of clusters, report the best choice for the number of clusters for the column normalized data.
17. Did you observe any differences when the data from S1.txt was used without normalization and with normalization? Explain.
18. Did you observe any differences during prediction when the data from S1.txt was used for clustering without normalization and with normalization? Explain.
19. Explain a situation where normalizing the data might help.

Question 1.2: Problem based on multiple data set

We provide multiple dataset in file b4.txt, e3.txt and u1.txt. Try to answer following questions.
Link for b4.txt
Link for e3.txt
Link for u1.txt

1. Vary the number of clusters K by choosing from the set $\{3, 5, 7, 9, 11, 13, 15, 17, 19\}$ and use the data in the K-Means function of scikit-learn package.
2. For each value of K, prepare the scatter plots depicting the clusters using different colors along with the cluster centers depicted in the same plot with a color different from those used for clusters.
3. Explain your observations about the clustering results you obtained when the value of K is increased from 3 till 19.
4. Consider the `test.txt` given above and find the predictions for the points in `test.txt` for clustering obtained for each value of K. Plot the points in the scatter plot and indicate the predicted cluster labels.
5. Explain your observations about the predictions obtained for different values of K.
6. Using your procedure to find the best choice of number of clusters derived in Problem 2.1, report the best choice for the number of clusters for the column normalized data.
7. Normalize the columns of the data to be of mean 0 and variance 1.
8. On the new data set thus obtained where the columns have mean 0 and variance 1, repeat the clustering for K in $\{3, 5, 7, 9, 11, 13, 15, 17, 19\}$.
9. For each value of K, prepare the scatter plots depicting the clusters using different colors along with the cluster centers depicted in the same plot with a color different from those used for clusters.
10. Explain your observations about the clustering results you obtained when the value of K is increased from 3 till 19.
11. Modify the test data given in `test.txt` file so that it can be used for prediction and report the predicted labels. Prepare scatter plots where you plot the transformed data from `test.txt`.
12. Explain your observations about the predictions obtained for different values of K.
13. Using your procedure to find the best choice of number of clusters, report the best choice for the number of clusters for the column normalized data.
14. Did you observe any differences when the data was used without normalization and with normalization? Explain.
15. Did you observe any differences during prediction when the data was used for clustering without normalization and with normalization? Explain.

Question 2: Introduction to Data Visualization: A Case Study from Cleveland Heart Data

Link for cleveland_heart_attr.csv
Link for cleveland_heart_attr_description.txt

1. After loading the data into the pandas dataframe **df**, write code to identify the number of rows and columns that **df** has, and print them.

2. Why are the num_major_vessels_fluroscopy and thal columns considered object types? Write the reason.

3. From the histogram on age attribute, identify the number of bins and bin size. Report these quantities.

4. Plot the histogram on age attribute for 50 bins and report the bin size and your observations.

5. What is the KDE option useful for in histplot()? Explain the details.

6. Plot pandas based histogram and seaborn based histogram for serum_cholesterol at tribute. Use bin sizes from {default, 20, 50, 100, 200, 500}. For seaborn, use KDE. Report the observations.

7. In the plot depicting the histogram of serum_cholesterol attribute containing mean and median, add also the vertical lines to represent the 25 percentile and 75 percentile values in the serum_cholesterol attribute. Use different colors and appropriate legend.

8. Change the order in the bar plots for gender vs serum_cholesterol from male, female to female, male and replot.

9. Explain the difference between the bar plot obtained using the median estimator for gender vs serum_cholesterol and the bar plot obtained before.

10. Explain the observations from the bar plot containing gender vs serum_cholesterol grouped according to chest pain type.

11. Note that the chest_pain_type attribute is numerical and hence is of less value in the bar plot obtained for gender vs serum_cholesterol grouped according to chest_pain_type. To make the plot more meaningful, insert a new column to the dataframe which contains the description according to the corresponding chest_pain_type code. Name this column as chest_pain_type description. To fill the values in this chest_pain_type description column, take the description for chest_pain_type from description file. Construct the bar plot for gender vs serum_cholesterol grouped according to chest pain type description. Add an appropriate legend and display the legend in a position where the bar graphs are clearly visible.

12. Add an appropriate annotation indicating the value of the upper boundary values of the bar plots in the gender vs serum_cholesterol grouped according to chest pain type.

13. Add an appropriate annotation with pointed arrows and with textual description in bar plot of gender vs serum_cholesterol grouped according to chest pain type. Color the arrow with a color other than red.

14. Explain your observations from the scatter plot obtained for age vs serum_cholesterol.

15. What does light-colored bands and the dark central line indicate for age vs serum_cholesterol plot?

16. What do the upper and lower boundaries of the box of chest_pain_type and serum_cholesterol indicate? What does the line inside the box indicate? What are the points marked beyond the error bars? Explain.

17. Discuss the observations made from the box plot for chest_pain_type and serum_cholesterol grouped according to gender.

18. Use violin plot to plot the relationship between chest_pain_type and serum_cholesterol and discuss the observations. Group the violinplots based on gender information and discuss the observations.