Assignment 1

Kubra Iqbal

Due Date: Friday*, September 2018, by midnight*
Total number of points: 30

This assignment covers the topics from Lectures #1 and #2.

Problem 1 (5 points): Differentiate between the following terms:
a. classification and clustering

Classification and clustering are both data mining functionalities under pattern discovery. Classification is a method under supervise learning where labels are always known and clustering is a method under unsupervised learning where all the labels are not known.

b. classification and prediction

Classification and prediction are both data mining functionalities under supervised learning. Classification models predict categorical class labels and prediction models predict continuous valued functions.

c. feature selection and feature extraction

Feature selection is used to remove unnecessary features from the data and on the other hand Feature selection is used to transform raw data into features suitable for modeling.

d. data mining and SQL

Data mining also known as (Knowledge discovery from data) is the process of extraction of interesting (Non-trivial, implicit, previously unknown and potentially useful) patters or knowledge from huge amount of data. Data mining also helps scientists in classifying and segmentation data. On the other hand SQL involves rederiving a subset of the existing data as specified by the user. The process is to specify the query and you get the known relevant data as output.

e. data warehouses and data marts

Data Warehouse can be viewed as a data repository for an organization that is set up to support strategic decision making. They store historical data of an organization in an integrated manner. The data in the Warehouse is never updated but only used to respond to queries from end users who are generally making decisions. Data warehouses also store billions of records because they are huge in size. Data marts on the other hand only hold one subject area and more summarized data. Data mart's are also built focused on a dimensional model using a schema.

Problem 2 (5 points):
Discuss whether or not each of the following activities is a data mining task.

(a) Monitoring the heart rate of a patient for abnormalities.

No, this is not a data mining task. Even though the data is interesting and potentially useful the extraction of this data can be done through advanced technology that is present these days. The heart monitors are able to detect abnormalities and simply tell the physician the outcomes.

(b) Computing the total number of courses offering by an university.

No, this is not a data mining task. Because the data for this is not hidden or a challenge to extract. The number of courses being offered by a university can be checked in the data base where all the records are saved.

(c) Sorting a student database based on student identification numbers.

No, this is not a data mining task. This information can be found in the data base – sorting the database would do the job to sort this data.

(d) Predicting the outcomes of tossing a (fair) pair of dice.

No, this is not a data mining task. The probability for this is already known. The examples of dice, cards and coins are usually used as examples to teach statistics and any other subject in that matter – the answers of these outcomes can be easily calculated.

(e) Monitoring seismic waves for earthquake activities.

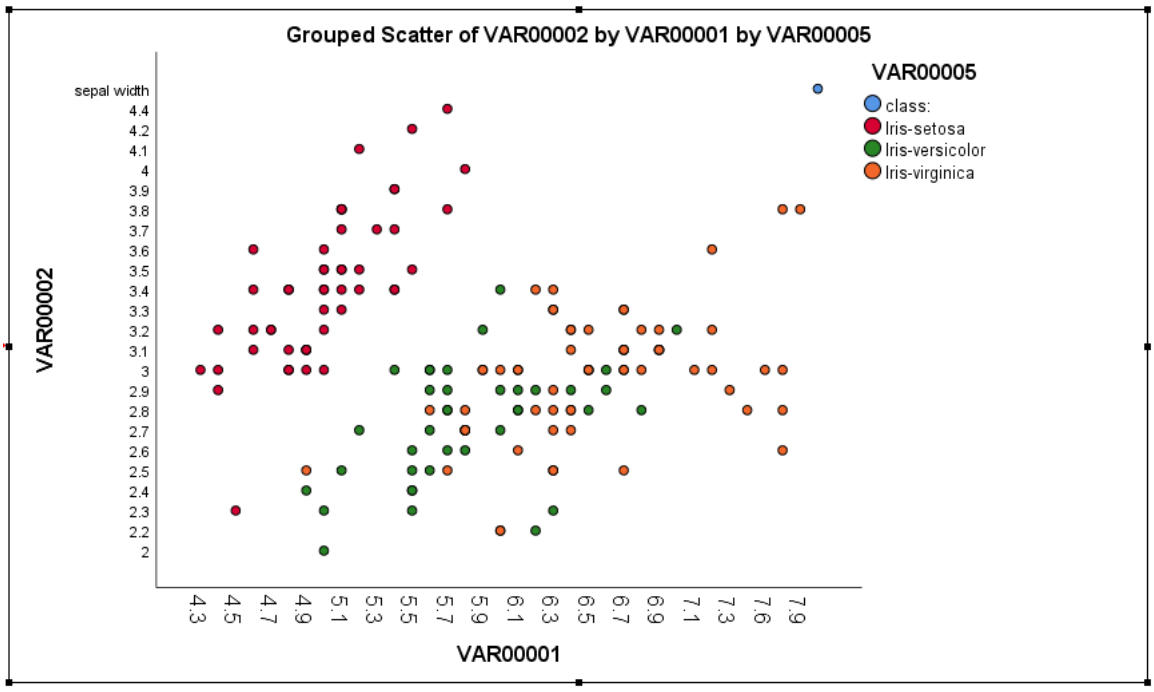
Yes, this is a data mining task because the data is previously unknown and useful. There can be massive data found which is raw and would need a lot pre-processing and taking out outliers to find out what is being looked for by the experts in this field.

Problem 3: (15 points) Fisher's iris data (download the IRIS dataset from http://archive.ics.uci.edu/ml/datasets/Iris) consists of measurements on the sepal length, sepal width, petal length, and petal width of 150 iris specimens. There are 50 specimens from each of three species.

Use SPSS to answer the following questions:

a. Visualize and interpret the relationship between the two sepal variables, sepal length and sepal width. Provide the scatterplot that you created to visualize the data along with your interpretation. When you plot the data, you may want to use different colors/signs for representing the data points belonging to the different three class species. Do you think that a classification algorithm will be successful in classifying the data with respect to these two variables? Justify your answer.

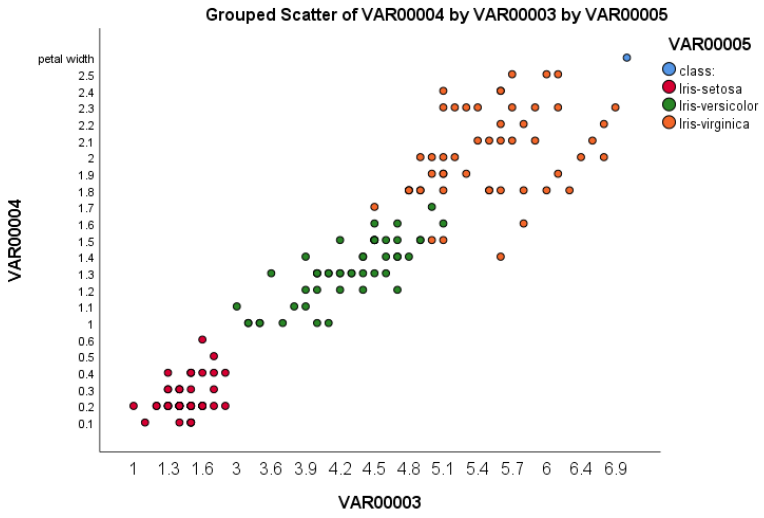
GGraph



The algorithm worked good enough because two classes which are: Iris versicolor and Iris Virgincia are pointing almost same values in some areas. Iris setosa class was separated. It might be hard to make predication from this classification algorithm for most cases.

b. Repeat part a. for the petal variables.

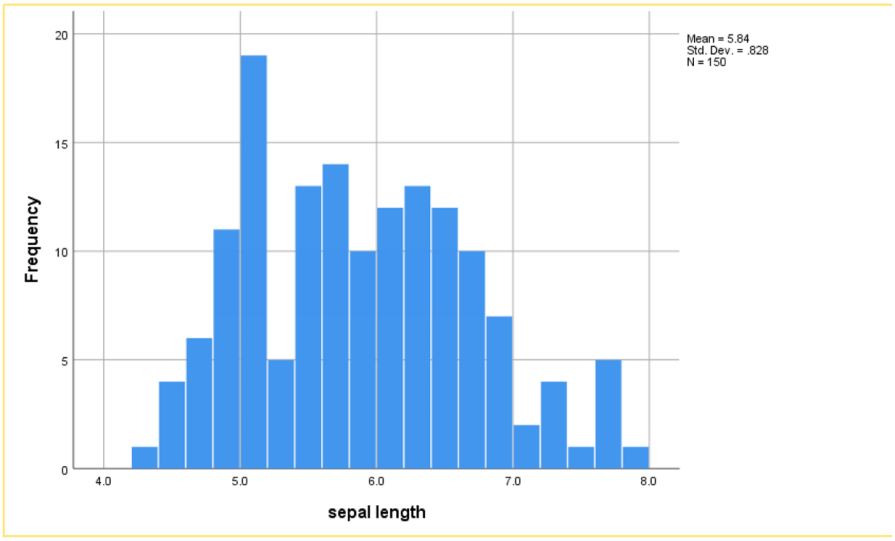
GGraph

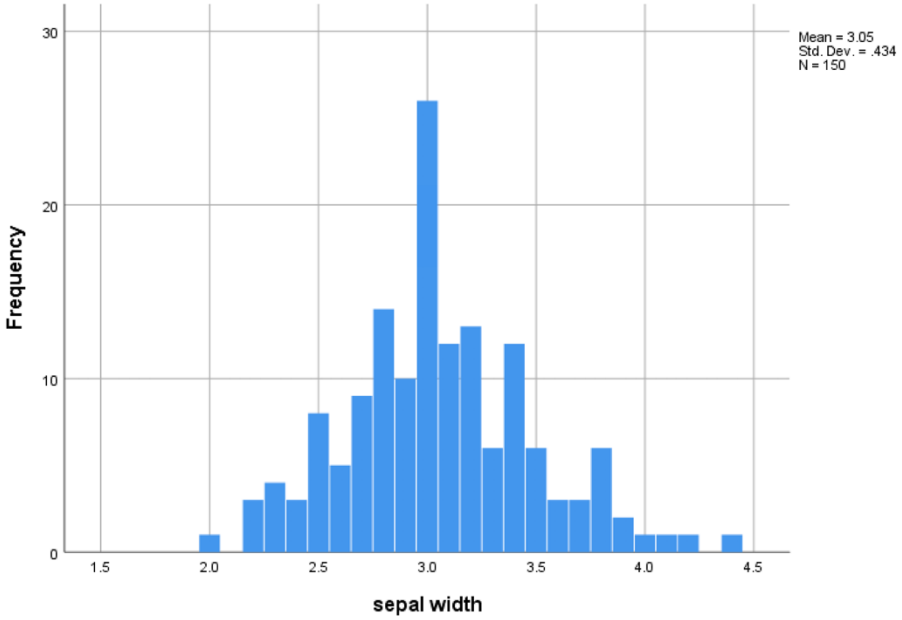


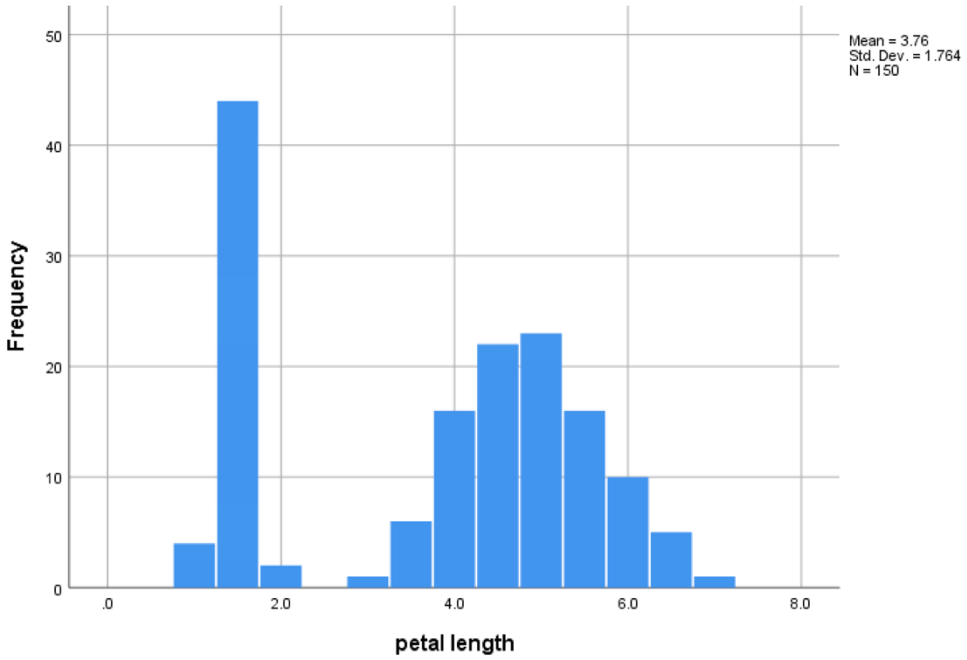
Classifications looks successful for petal variables. This is because it will help to make easy predications base on the results of the scatter plot.

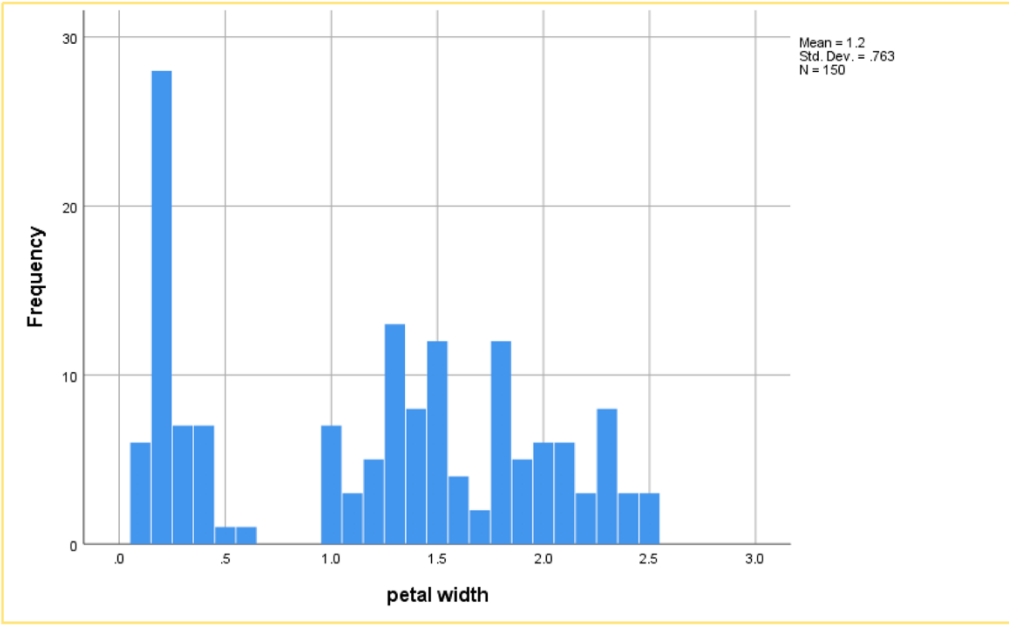
c. Draw the histograms of the four variables and interpret the distributions of each one of the four variables.

[DataSet1]


Mean = 5.84
Std. Dev. = .828
N = 150
Frequency
sepal length


Mean = 3.05
Std. Dev. = .434
N = 150
Frequency
sepal width

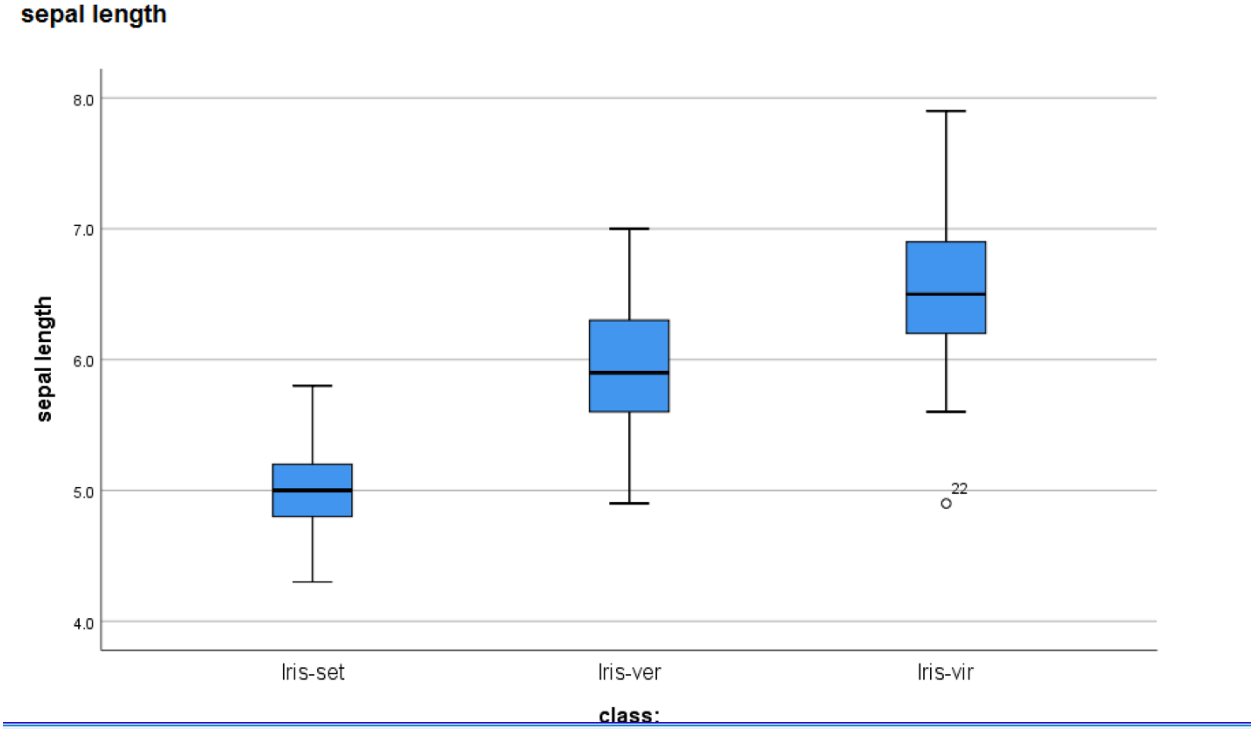




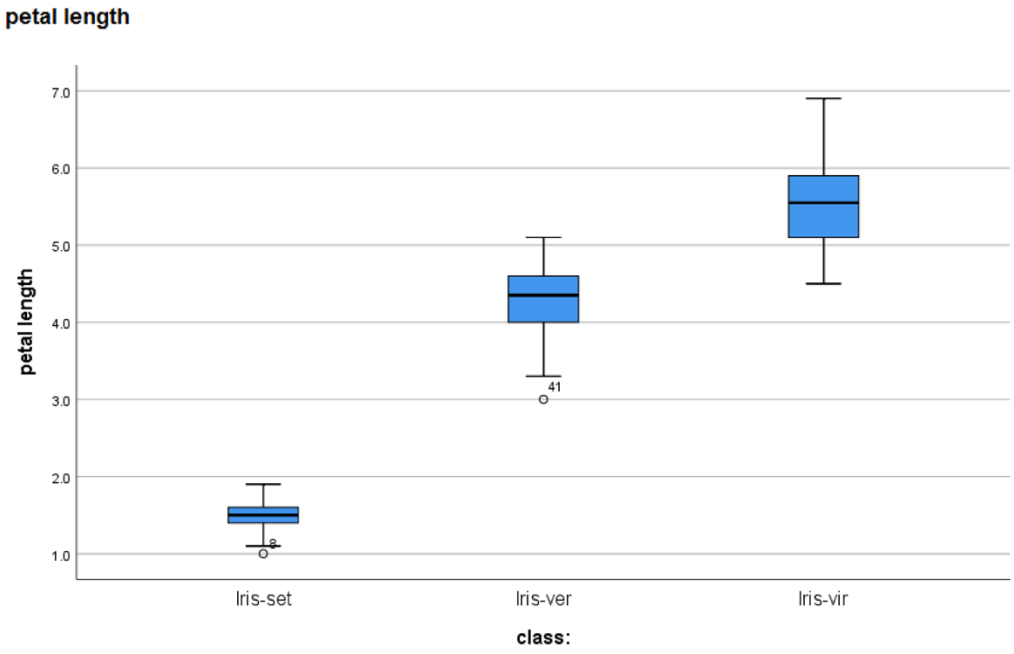
- The Histogram for the Sepal Length Histogram appears to have normal distribution. There seem to be distinct peaks from the distribution between the different species.
- The histogram for Sepal Width appear to have a normal distribution and the central peak is clear.
- The histogram for Petal Length appears to have two distinct groups. One of them displaying a normal distribution possibly from the distribution between the different species. To describe this, if the different species are not to be considered the histogram can be described as skewed right as the right-hand side has a peak with a higher count.

- The histogram for Petal Width appears to have two distinct groups and a singular peak. The distribution appears to be skewed right.

d. Determine if there are any outliers in the data with respect to the sepal length.



e. Repeat d. for the petal length.



Problem 4 (5 points): The following paper presented at the *ACM KDD 2017 Workshop on Machine Learning Meets Fashion* showcases an interesting application of data science to fashion and social media: "Identifying Fashion Accounts in Social Networks" by Doris Jung-Lin Lee, Jinda Han, Dana Chambourova, and Ranjitha Kumar:
https://kddfashion2017.mybluemix.net/final_submissions/ML4Fashion_paper_21.pdf

Read the paper and briefly answer the following questions:

1. What was the data used for the study? Include descriptions on the type of data and the size of the data.

The data collected was of 10k Twitter accounts using a content-based Snowball sampling approach and crowd sourced ground-truth labels for these accounts on Amazon Mechanical Turk. After the feature set is figured out then it will be possible to compute the feature set over the fashion accounts. Once the training set is obtained, then the classifier can be trained.

2. Was the data preprocessed or cleaned before applying any modeling techniques?

A Crowdsourced Dataset collection on Amazon Mechanical Turk was used and workers were asked to classify whether an given account is fashion, non-fashion or inaccessible. An inaccessible account is one that is either a deleted or private account. For each task the worker is shown a list of 10 Twitter accounts and they are asked to classify whether the account is fashion related or not. A account is deemed a fashion account if at least two out of three crowd workers classify it as a fashion account.

3. Did the authors solve a classification, a prediction, or a clustering problem as part of the pattern discovery stage? Justify your answer.

Classification. Since to classify the Twitter accounts as fashion or non-fashion – the features based on the data collected from the crawler which consisted of all the recent tweets posted by the account. Since the twitter word limit is 160 – the data was classified with the word "fashion" in their tweets or some word from the vocabulary that was listed out. There were two separate machine learning algorithms used for account classification – Naïve Bayes and Support Vector Machines.

4. For the problem identified, which algorithm(s) the authors use to solve that problem?

One of the issues that arrived after this experiment was that due to the 140-word limit on Twitter, a diverse set of fashion users use media attachments as a way to further their expression. This makes a very big challenge when the data is being sorted by the text-based approach. Since the media content is often self-exclamatory, tweets with media content are often associated with many descriptive tweet text related to fashion. To figure this issue, a post-analysis was conducted to understand how many accounts from the misclassified cases are media heavy accounts by examining 100 tweets from 100 account for both the false positive and false negative cases.

Another issue was that it was fairly common for a fashion user to be identified as fashion due to their occupational description on their profile descriptor. Some of these people are just using Twitter to discuss things related to their personal life unrelated to fashion. To make this search clearer, a more fine grained classification for whether someone is fashion-related by occupation or by tweet content.

Lastly, there are many bloggers, magazines or special interest pages on Twitter that post tweets related to multiple topics. These fashion accounts contributed largely to false positive rates of the data.

Submission Instructions

1. Answer the problems and write your answers in a Word document.
2. Submit your file online at the website at http://d2l.depaul.edu and check your submission
3. Keep a copy of all your submissions!
4. If you have questions about the homework, email me BEFORE the deadline.
5. Late submissions are allowed with a 5%, 10%, and 15% penalty for a one day, two days, and three days, respectively.
6. No late work will be accepted after three days since the assignment was due.