Assignment 1

Lavinia Wang #1473704

Due Date: Friday*, January 19th, 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

Answer:
Classification method belongs to the supervised learning category. In general, in classification you have a set of predefined labels and want to know which label a new object belongs to, eg. whether the customer is loyal to the brand, and the output is binary like yes or no. Algorithms applied in classification problems include Decision Tree, K-nearest neighbor, Neural Networks, Supportive Vector Machines, and Random Forest etc.
Clustering, aka unsupervised learning, tries to group a set of objects and find whether there is some relationship between the objects. There is no predefined label in clustering. Algorithms applied include K-means clustering and hierarchical clustering.

b. classification and prediction

Answer:
Classification problem deals with discrete or categorical labels/classes of dataset. For example, the label/class we want to classify is age group, which has 7 levels (level 1: 18-25, level 2: 26-34, level 4: 35- 44, level 5: 45-54, level 6: 55- 64, level 7 65+). The first observation belongs to level 1, meaning the person's age is in range 18-25.
Prediction problem deals with continuous or numerical labels/ classes and predicts missing values of dataset. For example, the label we want to discover is the mileage of one specific make in 2020.

c. dimensionality reduction and numerosity reduction

Answer:
Dimensionality reduction is the process of reducing the number of random attributes under consideration by obtaining a set of principal attributess. One example is principal component analysis(PCA).
Numerosity reduction is a technique of choosing smaller forms or data representation to reduce the volume of data. One example is sampling instead of using the whole population.

d. data mining and OLAP

Answer:

OLAP (on-line analytical processing) as the name suggests is a compilation of ways to query multi-dimensional databases. Typically, a data cube is used to display the output of an OLAP. The rows and columns are formed by the dimensions of the query. Aggregation on multiple tables is often used to obtain summaries. For example, it can be used to find out about the sales of this year in Wal-Mart compared to last year? OLAP does not learn from data nor do create new knowledge; they are simple special-purpose visualization tools.
Unlike OLAP, data mining deals with extraction or learning implicit/hidden information.

e. data mining and machine learning

Answer:
Data Mining is about using Statistics as well as other programming methods to find patterns hidden in the data so that you can explain some phenomenon.
Machine Learning focuses on writing algorithms in a way such that machines are able to learn on their own and use the learnings to tell about new dataset whenever it comes in.
In short, in terms of data mining, you essentially have the objective "knowledge discovery"; in machine learning, you have the objective "learn from training data and predict or estimate future".

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.

Answer: Yes. We would build a model of normal behavior of heart rate and compare with when unusual heart behavior occurred. This could be considered as a classification problem if we had examples of both normal and abnormal heart behavior.

(b) Computing the total sales of a company.

Answer: No. This is an accounting problem.

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

Answer: No. This is a database query.

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

Answer: No. Since the dice is fair, this is a probability calculation.

(e) Monitoring seismic waves for earthquake activities.
Answer: Yes. In this case, we would build a model of different types of seismic wave behavior associated with earthquake activities. When one of these different types of seismic activity was observed, we are able to predict whether earthquake is going to happen or what earthquake activity is going to take place. This is an example of classification problem.

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.

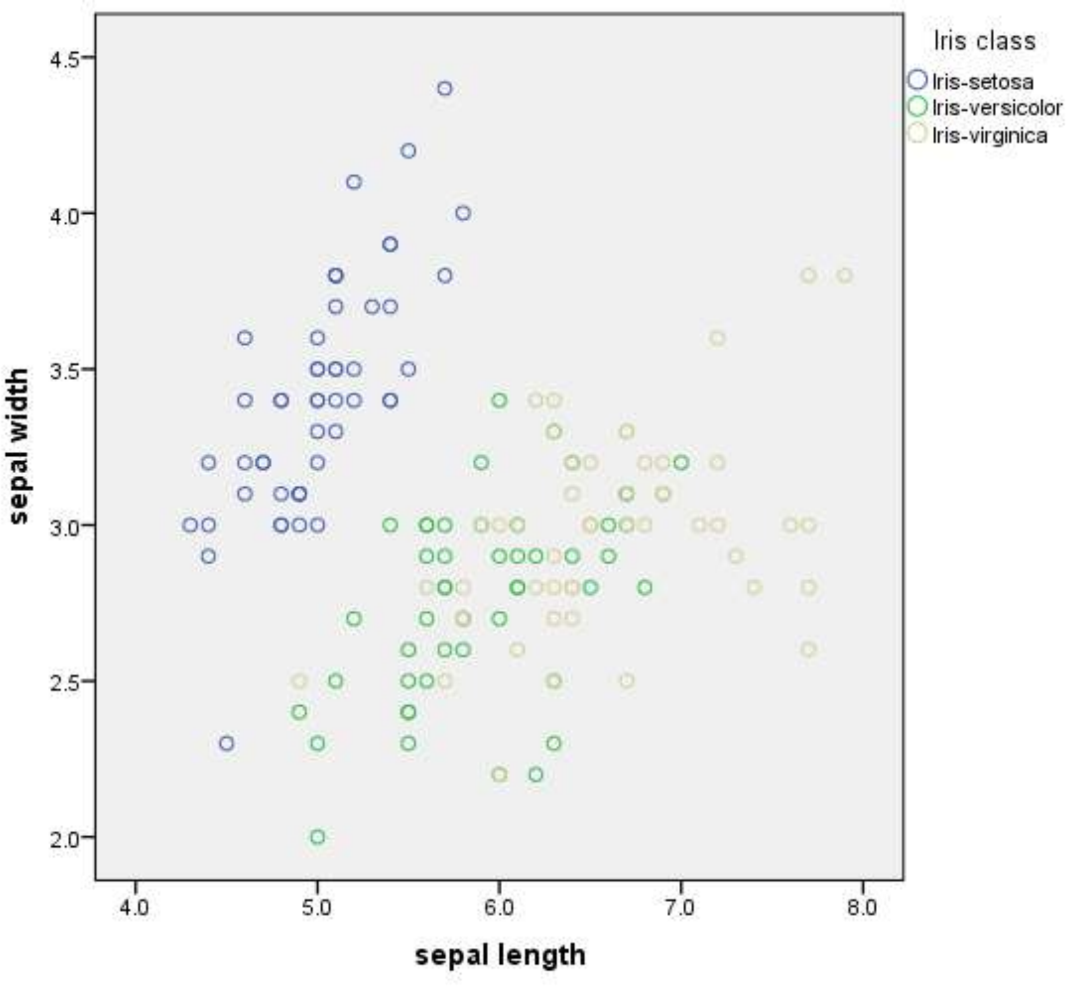


Figure1: Scatterplot of sepal length vs sepal width

From figure 1, we can see that iris-setosa clearly distinguishes from other two species and is easy to classify. But the other two, iris-versicolor and iris-virginica are not clearly distinguishable as setosa, because we can find overlaps located in the area with length 5.0 - 7.0 and width 2.2 - 3.4. Therefore, it is not possible to classify these two species and, the classification algorithm will not be successful in classifying the data with respect to these two variables.

b. Repeat part a. for the petal variables.

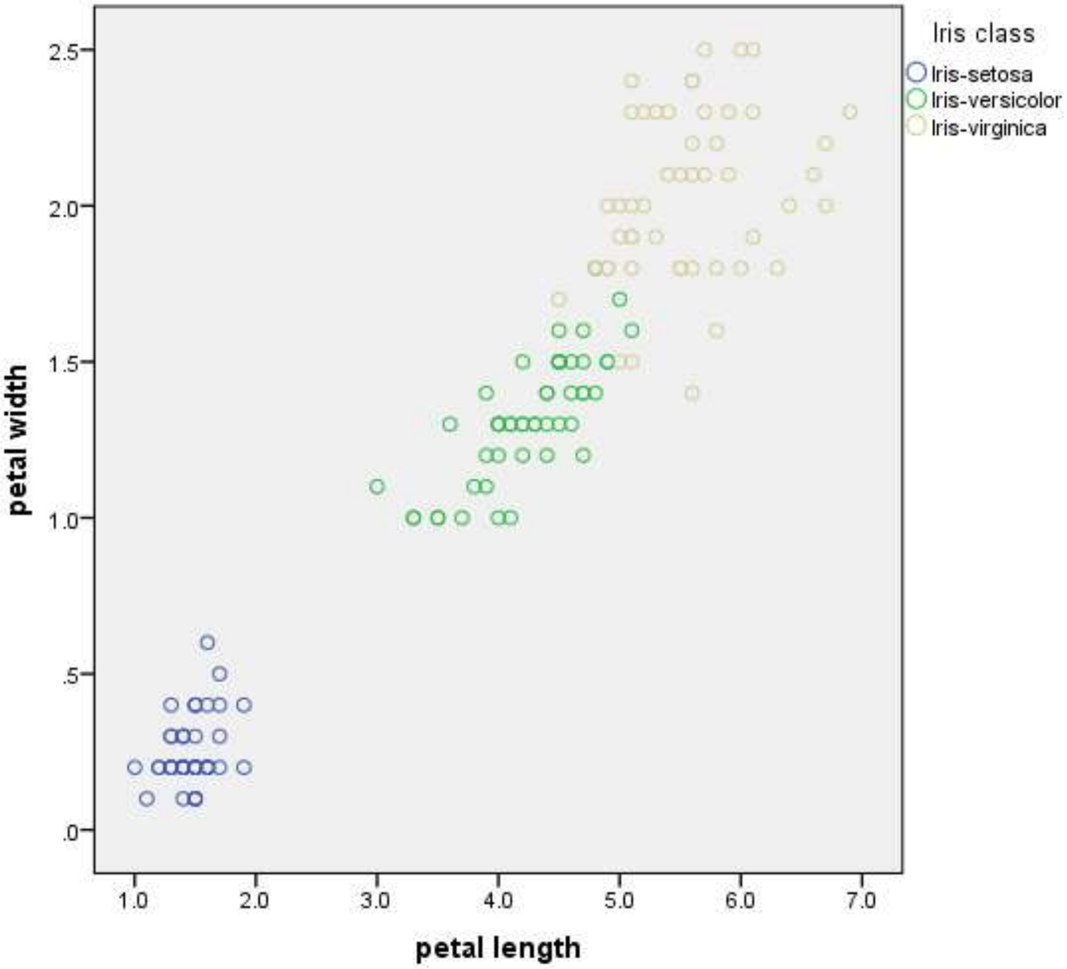
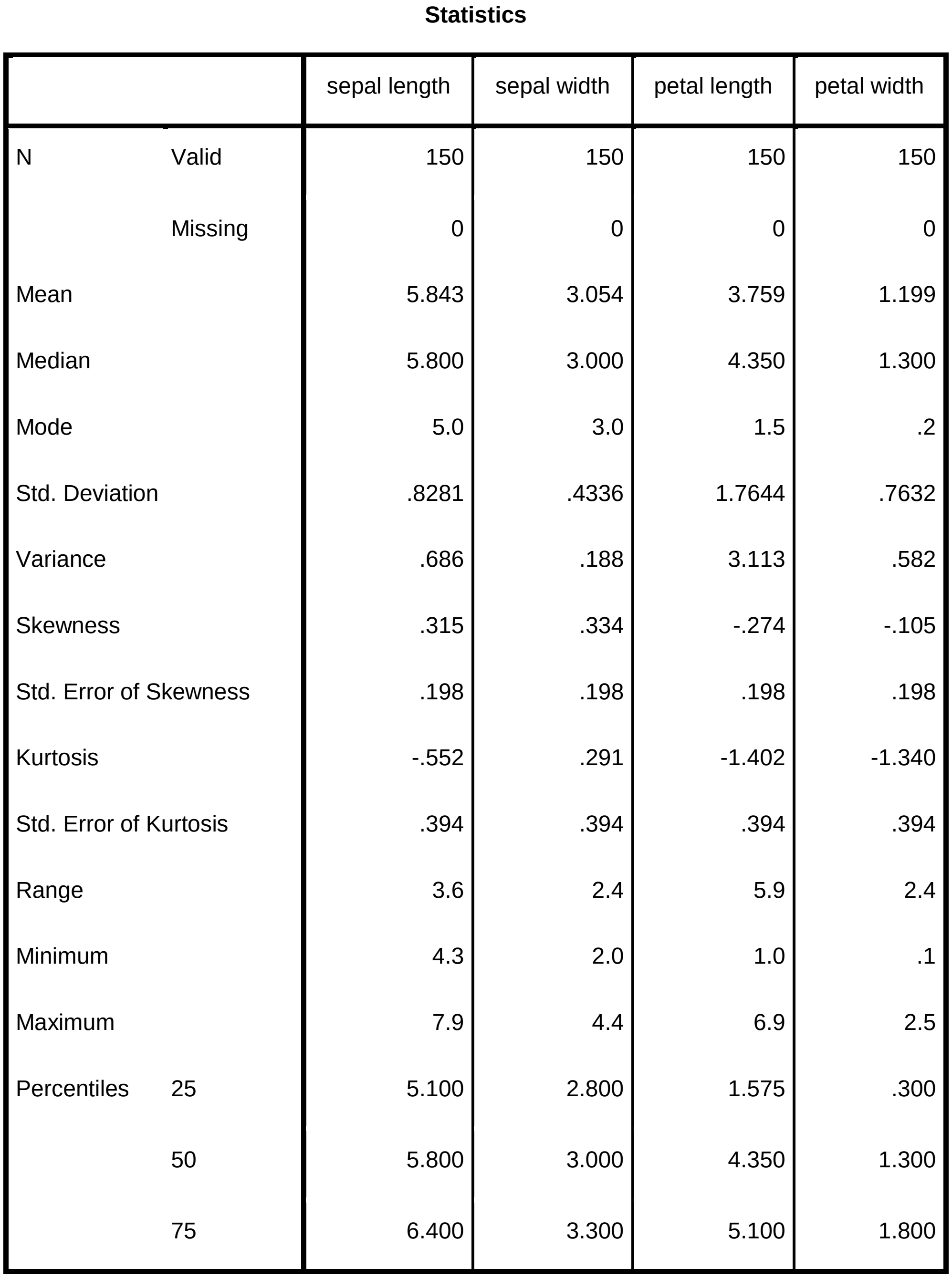


Figure 2: Scatterplot of petal length vs petal width

From figure 2, we can see that three species, iris-setosa, iris-versicolor and iris-virginica clearly distinguishes from one another and is easy to classify by either length or width, for instance, from length 1.0 to 2.0, the iris class is setosa; from 3.0 to 5.0, the class is versicolor; from 5.0 to 7.0, the class is virginica. A classification algorithm will be successful in classifying the data with respect to these two variables. Note that at the border of length 5.0 and width 1.5, versicolor and virginica cluttered. This may cause mismatch in classification.

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



Statistics

		sepal length	sepal width	petal length	petal width
N	Valid	150	150	150	150
	Missing	0	0	0	0
Mean		5.843	3.054	3.759	1.199
Median		5.800	3.000	4.350	1.300
Mode		5.0	3.0	1.5	.2
Std. Deviation		.8281	.4336	1.7644	.7632
Variance		.686	.188	3.113	.582
Skewness		.315	.334	-.274	-.105
Std. Error of Skewness		.198	.198	.198	.198
Kurtosis		-.552	.291	-1.402	-1.340
Std. Error of Kurtosis		.394	.394	.394	.394
Range		3.6	2.4	5.9	2.4
Minimum		4.3	2.0	1.0	.1
Maximum		7.9	4.4	6.9	2.5
Percentiles	25	5.100	2.800	1.575	.300
	50	5.800	3.000	4.350	1.300
	75	6.400	3.300	5.100	1.800

Table 1: Summary Statistics of four variables

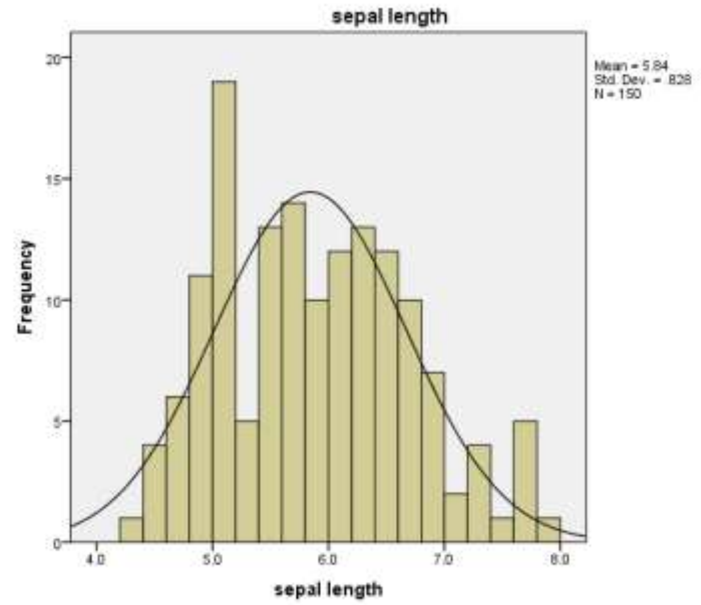


Figure 3: Histogram of sepal length

From figure 3 and table 1, the distribution has a mean of 5.84 and a standard deviation of 0.828. It is symmetric with skewness slightly equals 0 and thin tail as kurtosis is less than 3. It has a min of 4.3 and max of 7.9, giving a range of 3.6. The most common length is in 5.0. There is no apparent outlier. The distribution of sepal length is approximately normal.

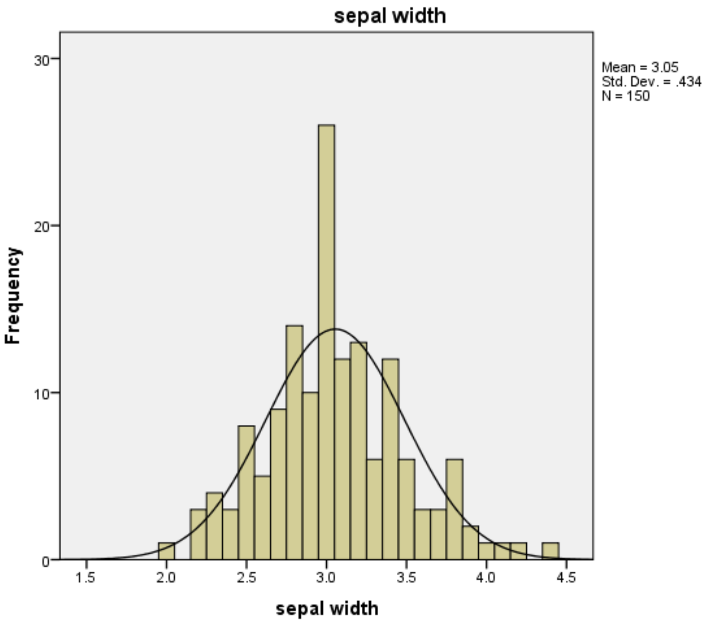


Figure 4: Histogram of sepal width

From figure 4 and table 1, the distribution has a mean of 3.05 and a standard deviation of 0.434. It is symmetric with skewness slightly equals 0 and thin tail as kurtosis is less than 3. It has a min of

2.0 and max of 4.4, giving a range of 2.4. The most common length is in 3.0. There is no apparent outlier. The distribution of sepal width is approximately normal.

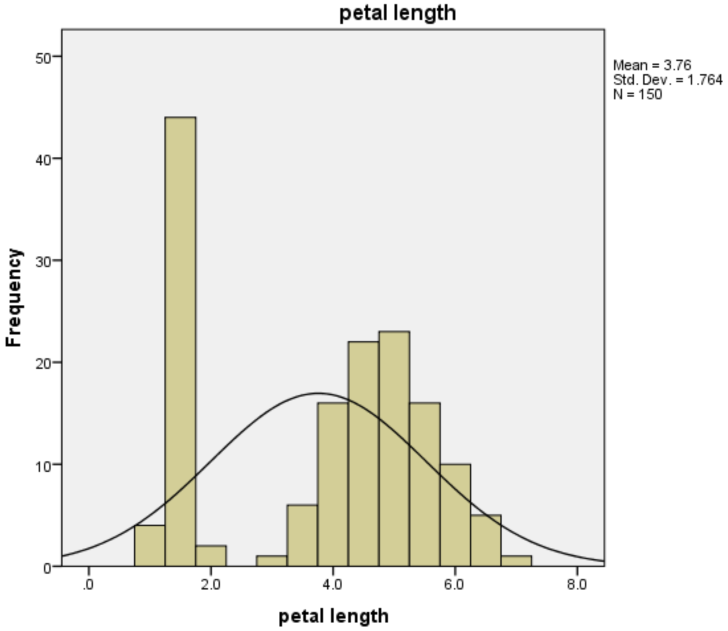


Figure 5: Histogram of petal length

From figure 5 and table 1, the distribution has a mean of 3.76 and a standard deviation of 1.764, which makes the distribution look short and wide. The median appears to be larger than the mean, which is more typical of left-skewed distributions. Kurtosis is much less than 3 meaning it has very thin tail. It has a min of 1.0 and max of 6.9, giving a range of 5.9. The most common length is in 1.5. There is no apparent outlier. The distribution of petal length is not normal

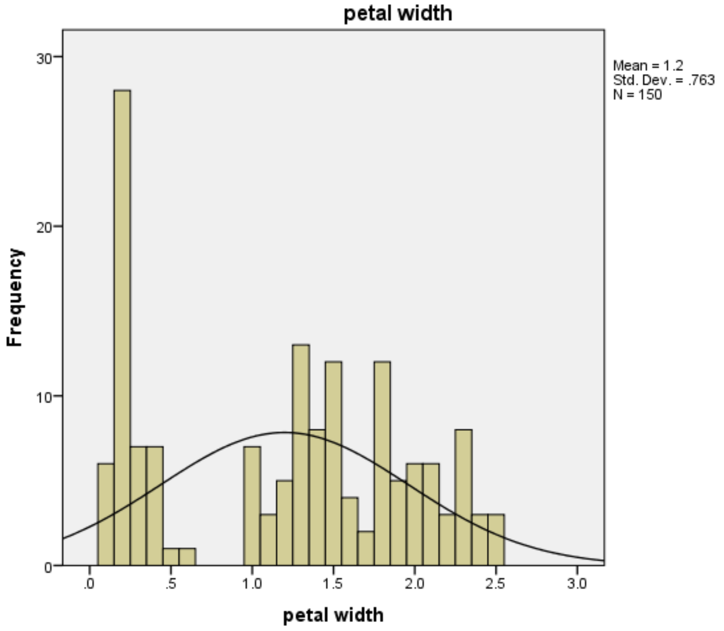


Figure 6: Histogram of petal width

From figure 6 and table 1, the distribution has a mean of 1.2 and a standard deviation of 0.763. It is slightly skewed to the left with very thin tail as kurtosis is much less than 3. It has a min of 0.1 and max of 2.5, giving a range of 2.4. The most common length is in 0.2. There is no apparent outlier. The distribution of petal width is not norm

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

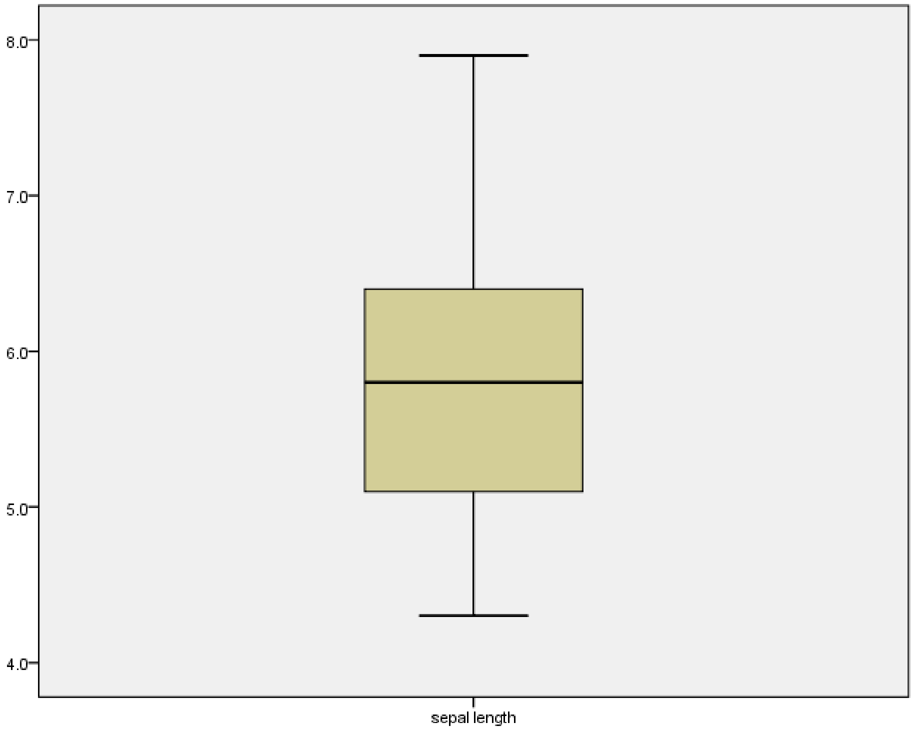


Figure 7: Boxplot of sepal length

Since sepal length is normally distributed, we use mean (aka μ) and standard deviation (aka σ) to determine if there is any outlier. From the above histogram, we find that μ is 5.84 and σ is 0.828. The range will be between 5.84 + (3*0.828) i.e. 8.324 and 5.84 - (3*0.828) i.e. 3.356. There is no value outside this range. So there are no outliers in the data with respect to the sepal length. We can verify this by looking at the boxplot figure 7.

e. Repeat d. for the petal length.

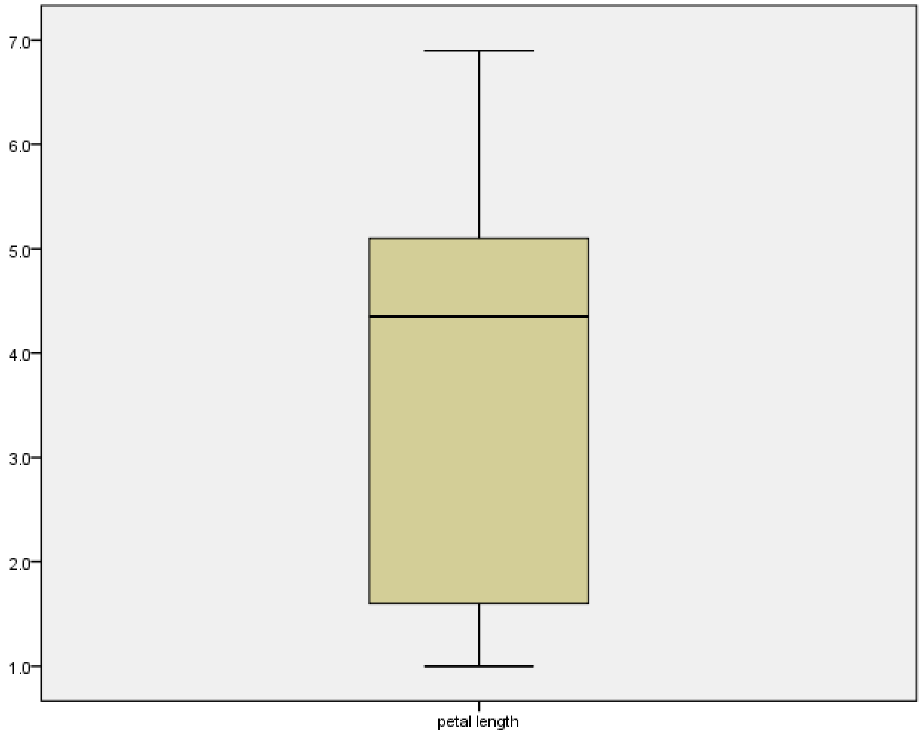


Figure 8: Boxplot of petal length

As is known petal length is not normal distribution, we will use 1.5*IQR criteria to detect outliers. The upper bound and lower bound is 5.1+1.5*(5.1 - 1.575), i.e. 10.388 and 1.575-1.5*(5.1 - 1.575), i.e. -3.713 respectively. There is no value outside this range. So there are no outliers in the data with respect to the petal length. We can verify this by looking at the boxplot figure 8.

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.

Answer:
“We decided to take a content-based, snowball sampling technique to collect our initial raw dataset. ”“In order to capture whether an account is fashion-related or not, we use two criterion: 1) number of fashion words from all the tweets posted by a user exceeds a threshold and 2) whether the word ‘fashion’ is contained in their profile description.” From these descriptions, we can tell that the dataset used for study is record data from Twitter.
The attributes of such dataset are fashion word. Each observation is the count of how many times the fashion word occurred. The last variable should be the label fashion-related account

with row contents indicating yes (the word ‘fashion’ is contained in their profile description) or no (the word ‘fashion’ is not contained in their profile description).
“Out of the 10230 unique labeled accounts, 26.72 %(2734) of the dataset is labeled as fashion accounts and the rest labeled as non-fashion.” The final dataset has 10230 observations.

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

Answer:
Data cleaning and data integration are indicated because the authors started with total 30510 responses and finalized with 10230 unique labelled observations. These numbers imply that redundant data, noisy data must be removed or smoothed before modeling.
Data transformation is applied, especially normalization. The following words describe the computation of normalization. “From the account information, we define fashion counts divided by total number of words in all tweets as normalized fashion counts, and we use normalized fashion counts and number of tweets and ... Another feature in used for classification is the normalized fashion counts, which is computed as the total number of fashion words divided by the total number of words over all tweets.”

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

Answer:
The authors solved a classification problem. The problem they tried to solve in this paper was to identify whether a Twitter account is fashion-related. The possible outcome is either the user is fashion-related or non-fashion related. If it were a prediction problem, the expected outcome would be a continuous label, like sales of 2018 is predicted 12345678. Then the variables or labels of the dataset should be totally different. If it were a clustering problem, the possible outcome would be certain number of clusters, representing groups of the dataset with distinguishable attributes.

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

Answer:
They used support-vector machine and Naive Bayes to conduct the classification.