Name: Amy Aumpansub
IS 467 Due: April 17, 2018

Assignment 1

Problem 1

a. What is the difference between classification and clustering (5%)?

Classification is a supervised learning technique which builds the model derived from the class-labeled (training) data set to predict categorical target (dependent) variable. The target value is known in the training set. On the other hand, clustering is an unsupervised technique, which does have a target value to predict, but it groups records which maximizes the intra-class similarity and minimizes the interclass similarity.

b. In a data table, what do the columns represent and what do the rows represent (5%)?

Columns represent the attribute, which represents a characteristic or feature of a data object such as customer name, address, telephone number. Rows represent data objects such as a customer object.

c. After loading new data into SPSS, describe two tasks you might do to clean your data (5%).

First task is to detect outliers by finding the normalized scores or z-scores of variables. In SPSS, we use the "Analyze" menu -> Descriptive Statistics->Descriptives-> click Save Standardized values as variables. The new variable shows the z-scores, then we sort descending to determine the potential outliers. The absolute value of z-scores greater than 2.5 (for a dataset with less than 80 samples) and z-scores greater than 3.0 (for a dataset with more than 80 samples) would be considered as outliers.

Second task is to fill in the missing values. In SPSS, we use the "Transform" menu -> Replacing missing values. Then, we select the methods such as Series Mean to replace the missing values. The new variable is created with completed values, which the mean and distribution of new variable are similar to those of the original variable, suggesting that the replacement is reasonable.

d. Explain which type of data mining algorithm (also called data mining functionality) would you use to answer each of these questions and why?

i. What are five groups of customers who buy similar things (5%)?

I would use Clustering because it can be used to identify homogeneous subpopulations of customers by grouping the customers who buy similar things based on the principle that maximizes the intraclass similarity and minimizes the interclass similarity. In this case, the things that are bought within the same customer group have a high similarity, but they are dissimilar to things in other clusters.

ii. I sell milk – can I predict if a user will buy that based on the other things they bought (5%)?

Association Analysis Functions would be implemented to find frequent patterns including a frequent itemset which shows the itemset that frequently appears together in the transactional data set. In this case, I will use the association rule to find which items are frequently bought together with milk within each transaction. The correlation analysis can also be additionally performed when the association rules do not satisfy the minimum support and confident threshold. We will use the correlation analysis to discover correlations between associated attribute–value pairs.

Name: Amy Aumpansub
IS 467 Due: April 17, 2018

Problem 2

Explain in few words whether or not each of the following activities is a data mining task and why.

a. Dividing the customers of a company according to their gender (5%).
No, this activity is not a data mining task because it can be done by using a simple database query to divide customers data by gender.

b. Computing the total sales of a company (5%).

No, this activity is not a data mining task because the total sales can be calculated by using simple accounting. It does not discover the knowledge or uncover pattern from the original data set.

c. Sorting a student database based on student identification numbers (5%).

No, this activity is not a data mining task because we can use a simple database query to sort student database by their id numbers.

d. Estimating the probability of the outcomes of tossing a (fair) pair of dice (5%).

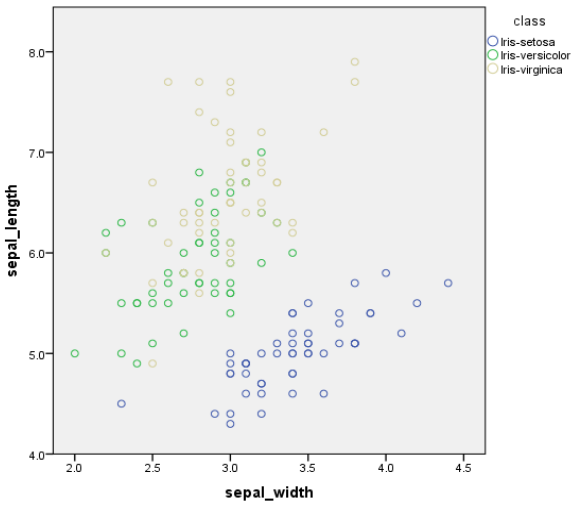
No, this activity is not a data mining task because we use the probability calculation to estimate the probability of the outcomes of tossing a fair pair of dice.

e. Predicting the future stock price of a company using historical records (5%)

Yes, this activity is a data mining task because it can be done by using the predictive mining task, which the historical records are mined to develop model to predict future stock price.

Problem 3

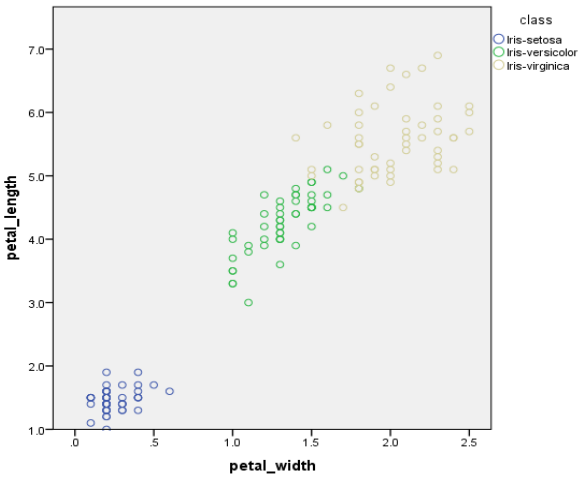
a. Visualize the relationship between the two sepal variables, sepal length and sepal width, using a scatter plot. Use different colors or symbols per class so you can see how the classes are related to this pair of variables. We talked in class about how classifiers work, broadly-speaking. Do you think that a classification algorithm using these two variables will be successful in classifying data with respect to the class labels we have? Explain why or why not and include the plot image with your answer (10%).



Sepal_length and Sepal_width

The scatterplot shows a very weak linear correlation, suggesting that these two variables are not linearly correlated. The classification algorithm using these two variables are unclear for Iris-versicolor and Iris-virginica, thus it unsuccessfully classifies data with respect to the class labels. The plot shows a mix result of Iris-versicolor and Iris-virginica, in which some samples from Iris-virginica have longer sepal length with shorter width, others have longer length with longer width. The Iris-setosa shows a weak linear trend that the sepal_width slightly increases as the sepal_length increases.

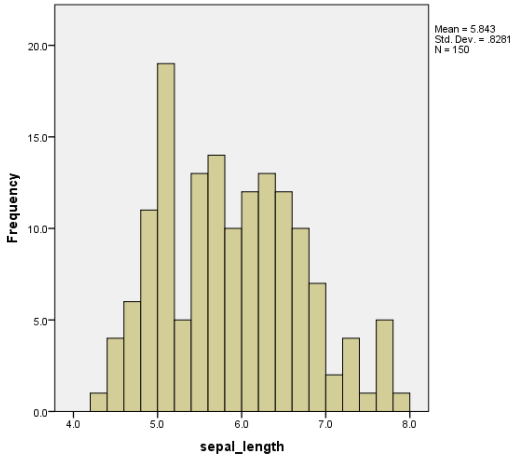
b. Repeat part (a) for the petal variables (10%).



Petal_length and Petal_width

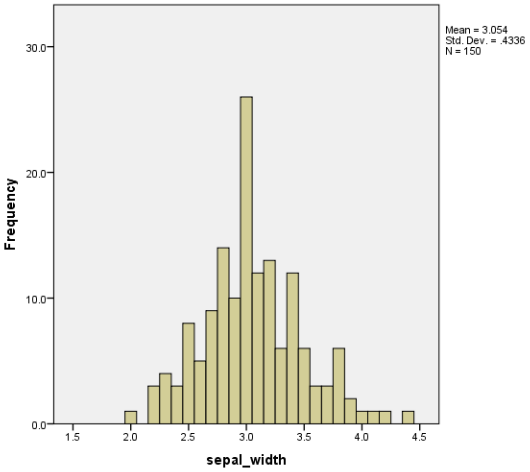
The scatterplot shows a positively strong correlation, suggesting that these two variables are positively and linearly correlated. The classification algorithm using these two variables shows a clear linear trend and a high correlation with class, thus it successfully classifies data with respect to the class labels. The plot indicates that compared to other classes, the "Iris-setosa" has the shortest petal width and petal length. The "Iris-versicolor" has a moderate petal width and moderate petal length. Among three classes, the "Iris-virginica" has the longest petal length and the longest petal width (cm).

c. Create a histogram for each of the four variables. Histograms in SPSS are just a different graph type from scatterplots. Describe what you can tell about the distribution of each variable (10%).



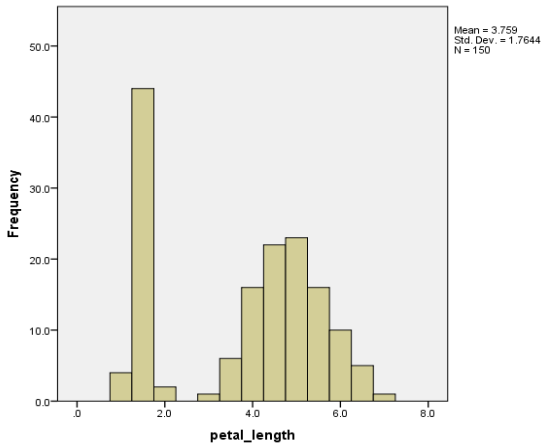
Sepal length

The histogram shows the distribution of sepal_length variable is close to normal. The mean of 5.843 is close to the mode of 5 and the median of 5.8. The statistics indicates the skewness value of 0.315 which confirms the distribution is relatively normal. The standard deviation of 0.8281 suggests the data has a small spread from the mean. The histogram also identifies that there are no potential outliers for the sepal_length variable.



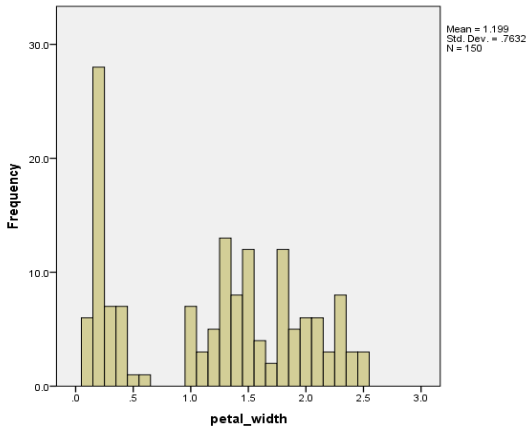
Sepal width

The histogram shows the distribution of sepal_width variable is close to normal. The mean of 3.054 is close to the mode of 3 and the median of 3. The standard deviation of 0.4336 suggests the data has small spread from the mean. The histogram also identifies that there are two potential outliers including the sepal_width values of 2 and 4.4. The z-score confirms that there is one outlier (the sepal_width value of 4.4), which has a z-score of 3.14 (see the statistic table on the last page).



Petal length

The histogram shows the distribution of petal_length variable is not normal and not symmetric. The distribution seems to be bimodal as it has two distinct modes. The mean of 3.759 is significantly higher and different from to the first mode of 1.5 and the median of 4.350. The standard deviation of 1.76644 suggests the data has a large spread from the mean. The histogram also identifies that there may be some potential outliers for the petal_length variable, but the z-scores of all values are less than 3, confirming that there are no outliers (see the statistic table below).



Petal width

The histogram shows the distribution of petal_width variable is not normal and seems to be bimodal as it has two distinct modes. The mean of 1.199 is significantly higher and different from to the first mode of 0.2 and the median of 1.3. The standard deviation of 0.7632 suggests the data has a large spread from the mean of 1.199. The histogram also identifies that there are no potential outliers for the petal_width variable, and the z-scores of all values of petal width variable are less than 3, confirming that there are no outliers.

d. Determine if there are any outliers in the data with respect to the sepal length (10%).

There are no potential outliers for the sepal length. The z-score variable has been created using the "Analyze" menu and "Descriptives" function in SPSS. The new column, "Zsepal_length" shows that all absolute values of z-score are less than 3 (the dataset contains 150 cases). The z-scores range from -1.86378 to 2.4837. The maximum absolute value of z-score for the sepal length is 2.4837, which is still less than 3. Thus, it confirms that there are no potential outliers in the data with respect to the sepal length.

e. Repeat d. for the petal length (10%).

There are no potential outliers for the petal length. The z-score variable has been created using the "Analyze" menu and "Descriptives" function in SPSS. The new column, "Zpetal_length" shows that all absolute values of z-score are less than 3. The z-scores range from -1.56350 to 1.78038. The maximum absolute value of z-score for the petal length is 1.78038, which is still less than 3. Thus, it confirms that there are no potential outliers in the data with respect to the sepal length.

Statistics

		sepal_length	sepal_width	petal_length	petal_width	Zscore (sepal_length)	Zscore (sepal_width)	Zscore (petal_length)	Zscore (petal_width)
N	Valid	150	150	150	150	150	150	150	150
	Missing	0	0	0	0	0	0	0	0
Minimum		4.3	2.0	1.0	.1	-1.86378	-2.43084	-1.56350	-1.43963
Maximum		7.9	4.4	6.9	2.5	2.48370	3.10428	1.78038	1.70519