Assignment 1 —IS467 1/12/2016

Sarah Cummings

Problem 1. Differentiate:

a. Classification v. Clustering: Classification is an example of supervised learning, where we use labeled training data to determine a function. Clustering is an example of unsupervised learning, a learning process in which the input examples are not labeled into classes.

b. Classification v. Prediction: Classification predicts categorical discrete labels, whereas prediction estimates quantitative continuous values. Prediction problems are generally solved with regression, and classification is solved with methods such as decision trees and k-nearest neighbors.

c. Data warehouse v. Database: A database is a collection of interrelated data and a set of software programs to manage and access the data. A data warehouse is much larger and generally stores history for a company of organization that will not be changed. It is a repository of multiple heterogeneous data sources organized under a unified schema, generally used to facilitate management decision making.

d. Data mining v. OLAP: Data mining is the extraction of interesting, intrinsic patterns and nuggets of information from large quantities of data. OLAP stands for online analytical processing, and these operations use background information to allow the presentation of data to be viewed at different levels of abstraction.

e. Machine learning v. statistics: Statistics studies collection, analysis, interpretation, and presentation of data. It is focused on testing hypotheses is is more theory based. Machine learning, on the other hand, investigates how computers learn and improve their performance based on data. Machine learning is more heuristic than theoretical, and often looks at real time learning.

Problem 2. Is it data mining? Discuss.

a. Monitoring the heart rate of patient for abnormalities is not a data mining task. In this task, we are collecting data (the patient's heart rate), but we are not discovering an intrinsic pattern through data mining explorative method.

b. Computing the total sales of a company is not a data mining task, as this information is easily totaled and is extrinsic in nature.

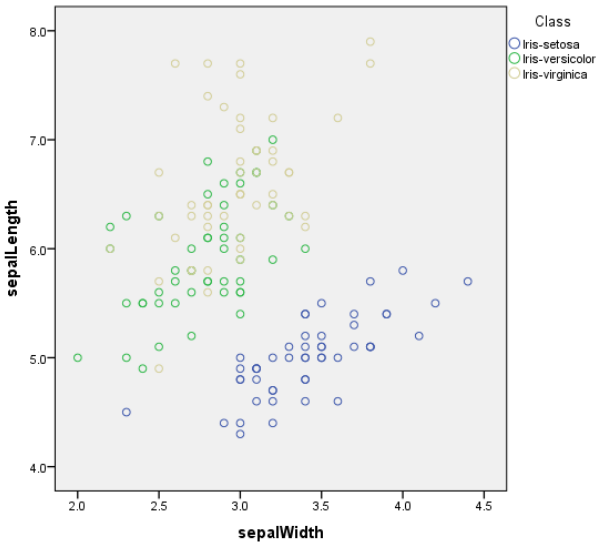
c. Sorting a student database based on identification numbers is not a data mining task, as we are not discovering any new or useful information in the process.

d. Predicting the outcomes of tossing a fair pair of dice is not a data mining task since there is not likely to be an interesting pattern or bit of information discovered.

e. Monitoring seismic waves for earthquake activities is not a data mining task, although we could later use the collected data for a data mining task if we collect a large amount of data and look for patterns.

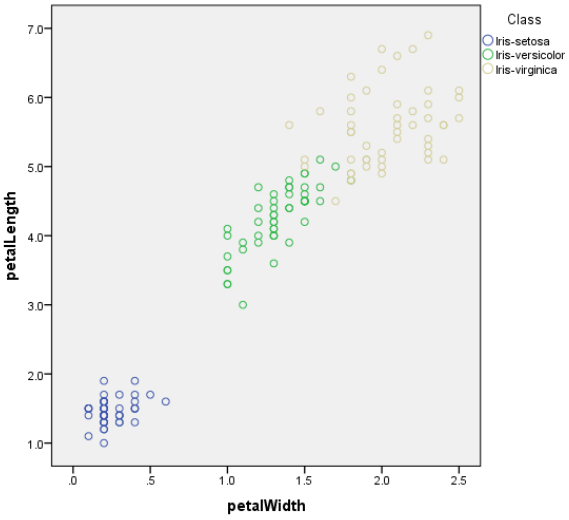
Problem 3. Fisher's iris data in SPSS

a. Visualization of sepal variables, and interpretation of their relationship



From the scatterplot provided by SPSS to the left, it appears as though sepal length has a positive relationship with sepal width. The slope of the corresponding relationship between sepal length is steeper for the Iris-versicolor and iris-Virginia than it is for the iris-sedosa class. In general, as sepal width increases, sepal length increases. However, sepal-length is shorter for a given width for the iris-sedosa class as it is for the other classes comparatively. Classification would likely be successful in classifying the data with these two variables as there is some distinctive grouping among the classes—though the overlap between iris-versicular and iris-virginica may pose a problem.

b. Repeat part a for Petal length and width.

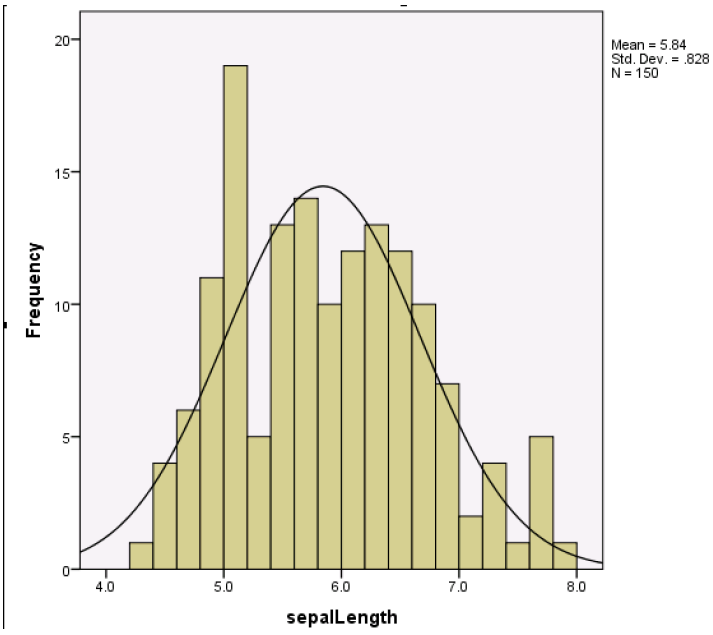


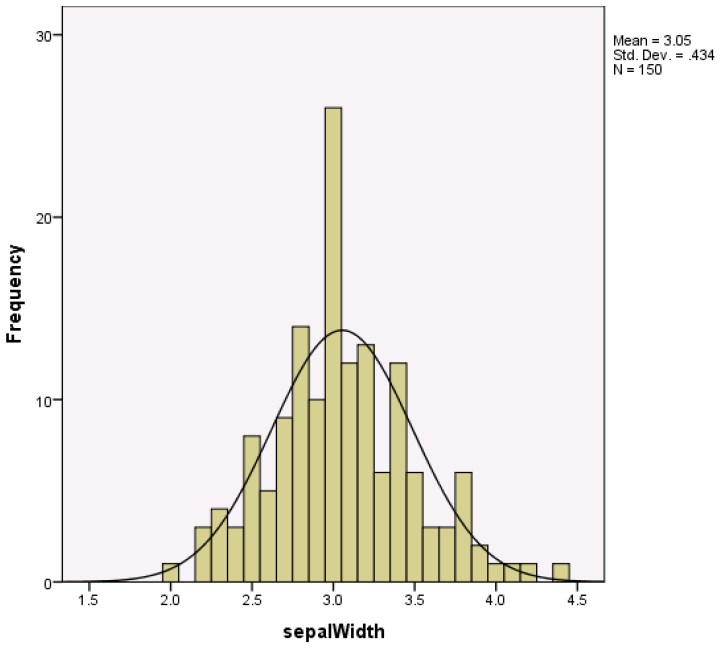
As seen in the SPSS scatterplot to the left, there is a positive relationship between petal length and petal width. We can also conclude that the petals for the Iris-sedosa class are much smaller than the iris-versicolor and iris-virginica classes, since the petal width and length measurements are smaller. The iris-versicolor petals are the next largest, and the iris-virginca petals are the biggest. Classification would be successful in classifying the data with these two variables as the three classes form three distinct groups with respect to petal width and petal length.

c. Histogram of each of the variables and interpretation of their distribution.

Descriptive Statistics

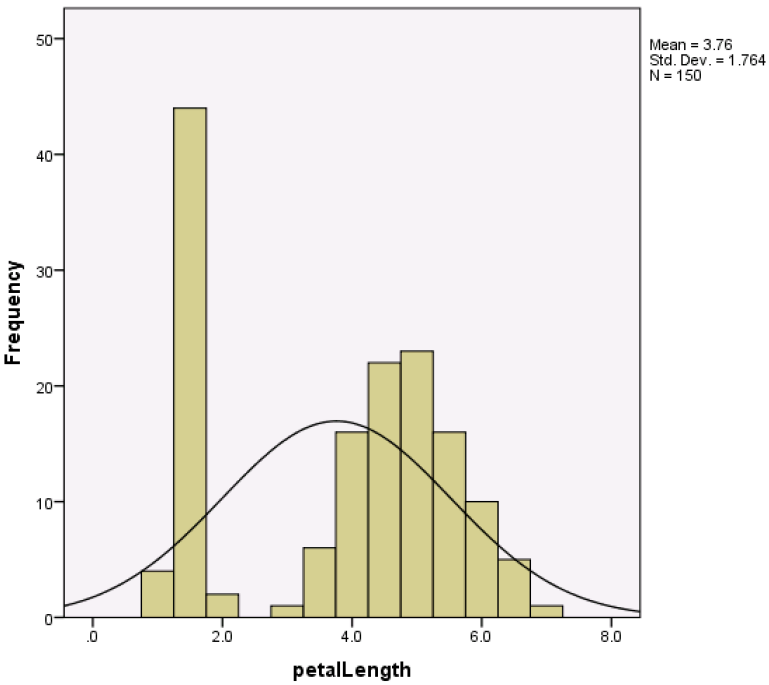
	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
sepalLength	150	4.3	7.9	5.843	.8281	.315	.198	-.552	.394
sepalWidth	150	2.0	4.4	3.054	.4336	.334	.198	.291	.394
petalLength	150	1.0	6.9	3.759	1.7644	-.274	.198	-1.402	.394
petalWidth	150	.1	2.5	1.199	.7632	-.105	.198	-1.340	.394
Valid N (listwise)	150								

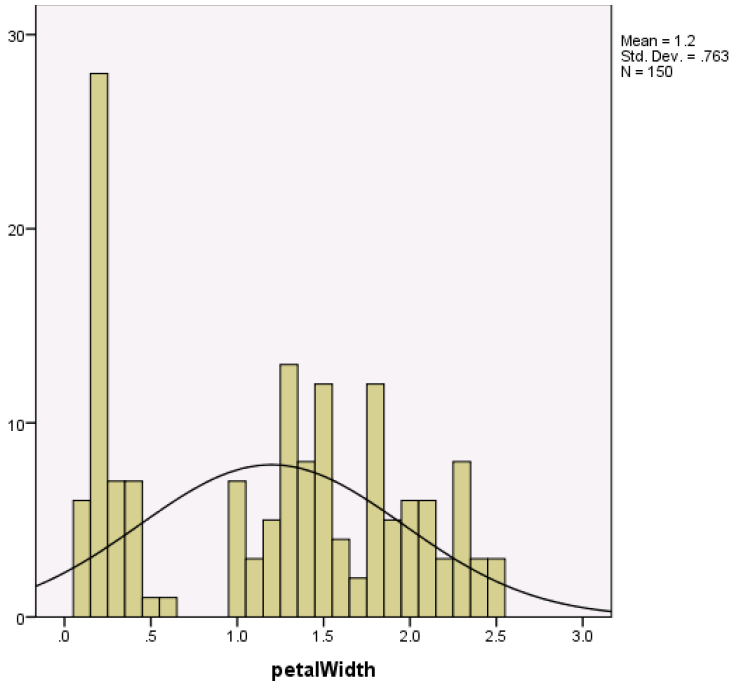




As seen in the histogram above, it appears as though the sepal length distribution is slightly skewed to the right. This is confirmed by a skewness statistic of 0.315 in the descriptive stats output. If the frequency of sepal lengths of 5.0 cm wasn't so high, this distribution would be close to normal. The mean for this distribution is 5.84 cm and the standard deviation is 0.828 cm. The mode is approximately 5.0 cm.

The histogram of sepal width shows it is the closest to normal distribution of all of our four variables. Though the frequency of the mean, 3.05 cm, is quite high, the distribution otherwise fits the normal curve fairly well. The standard deviation for this distribution is 0.434 cm.

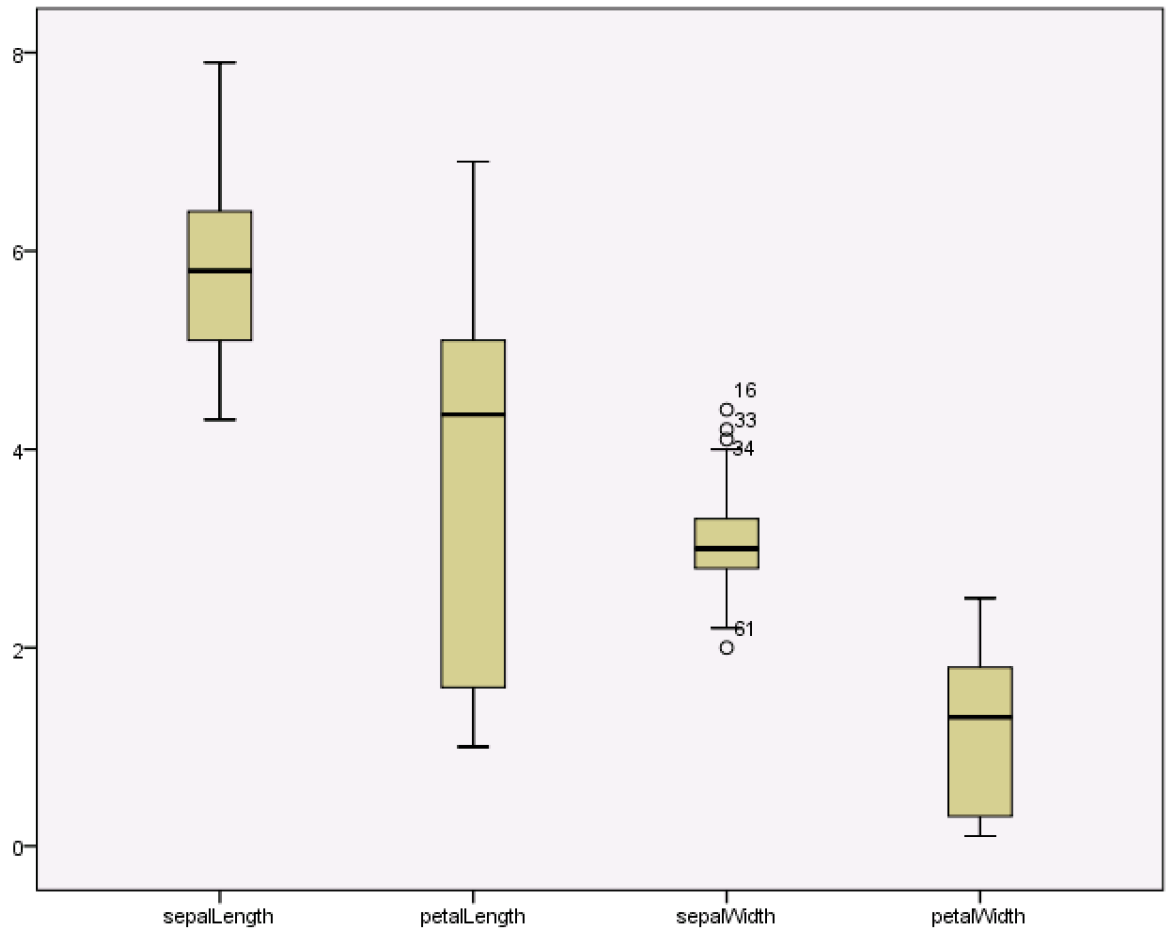




As seen in the histogram above, the petal length distribution appears to be skewed to the right. The mean is at the lower end of the spectrum of measurements, and the frequencies of higher values tapers off from there. The distribution could also be described as bi-modal, with some points around a mean of 1.76 cm, and another curve centering around the 5.0 cm mark. The mean for this distribution is 3.76 cm and the standard deviation is 1.764 cm.

The petal width distribution has a mean of 1.2 cm and a standard deviation of 0.763 cm. The skewness statistic is -1.340, which means its is negatively skewed or skewed left.

d/e. Outliers for sepal length and petal length



As seen in the box plots above, no outliers are indicated for either the sepal length or petal length variables. There are outliers for the sepal width variables, though, which is interesting.