IS467: Assignment 1

Jasmine Dumas

January 17, 2016

### Problem 1 (5 points): Differentiate between the following terms:

1. *classification and clustering*: Classification and clustering are both data mining functionalities under pattern discovery but classification is a method under *supervised learning* where the labels are *known* and clustering is a method under *unsupervised learning* where the labels are *unknown*.
2. *classification and prediction*: Classification and prediction are both data mining functionalities under *supervised learning* but known labels are of a different type/class; classification has labels that are categorical/discrete and prediction has labels that are continuous.
3. *data warehouse and database*: A database is a collection of interrelated data and a database warehouse is a repository of information collected from multiple sources, stored under a unified schema and usually resides at a single site. The difference being that the data warehouse is never updated and contains billions of records and the database can be modified, updated and smaller. databases are also an initial step before the data warehouse (ref pg 7 in textbook).
4. *data mining and OLAP*: In data warehouses, OLAP (online analytical processing) tools provide multiple views of the data and is stored in a special format called multi-dimensional hyper-box structure but OLAP *does not learn from data* and data mining tools obtain their answers *through learning* about the different relationships between attributes of data records.
5. *machine learning and statistics*: Machine learning and statistics are both disciplines that data mining adopts but statistics is more theory-based and focused on testing hypothesis and mathematical methods where as machine learning is more heuristic and focused on improvement performance of a learning agent.

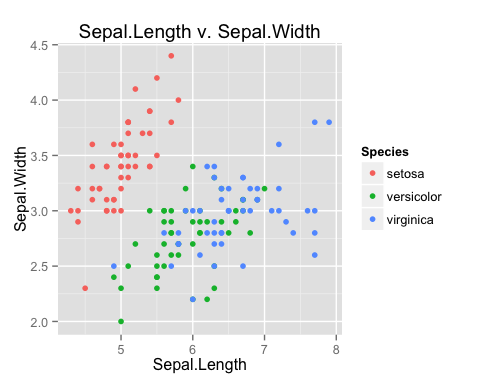
### Problem 2 (5 points): Discuss whether or not each of the following activities is a data mining task.

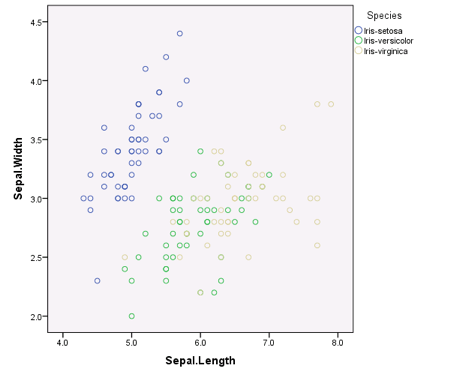
1. *Monitoring the heart rate of a patient for abnormalities*: No, this is not a data mining task because although the data is interesting and potentially useful the extraction and collection is trivial due to the advanced technology of heart monitors that are inherently able to detect abnormalities and simply output those values for physician evaluation.
2. *Computing the total sales of a company*: No, this is not a data mining task because the data for this is not hidden or challenging to extract and just a matter of aggregating the total sales of the company which can be completed by the company accountant or controller in excel using the SUM function.
3. *Sorting a student database based on student identification numbers*: No, this is not a data mining task because this is a trivial to complete, and readily available in the database. Sorting the database simply requires a SQL statement ie SELECT \* FROM studentdb ORDER BY studentID DESC;
4. *Predicting the outcomes of tossing a (fair) pair of dice*: No, this is not a data mining task because this is not interesting, already known probability, and trivial to obtain the data. The utilization of dice, coins and cards are frequent examples in statistics and probability so the answers about the outcomes of trials can be easily found or calculated out.
5. *Monitoring seismic waves for earthquake activities*: Yes, this is a data mining tasks because the data is interesting, non-trivial, implicit, previously unknown, and potentially useful. Massive seismic raw data would need pre-processing, filtering out of noise and needs interpretation by seismologist experts.

### 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/R 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.

library(ggplot2)  
plot1 <- ggplot(iris, aes(x=Sepal.Length, y=Sepal.Width, colour=Species)) +  
 geom\_point() + ggtitle("Sepal.Length v. Sepal.Width")  
plot1

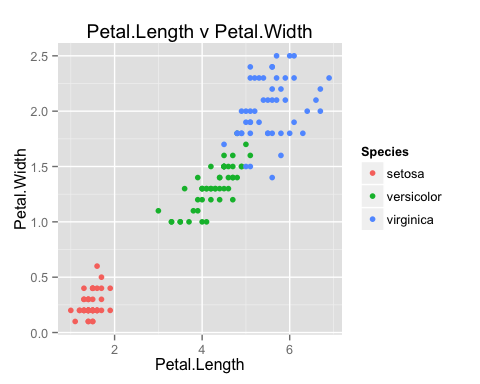


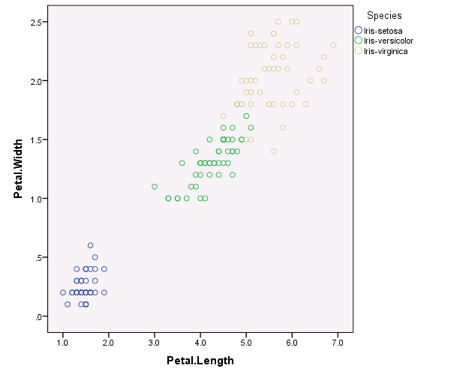
SPSS Scatterplot for Sepal Variables:  


There appears to be a positive linear correlation between Sepal.Length and Sepal.Width. By only relying on these two variables I don't believe that a classification algorithm will correctly classify the data seeing as the setosa species has a distinct grouping but the versicolor and virginica species is somewhat mixed.

1. Repeat part a. for the petal variables.

plot2 <- ggplot(iris, aes(x=Petal.Length, y=Petal.Width, colour=Species)) +  
 geom\_point() + ggtitle("Petal.Length v Petal.Width")  
plot2



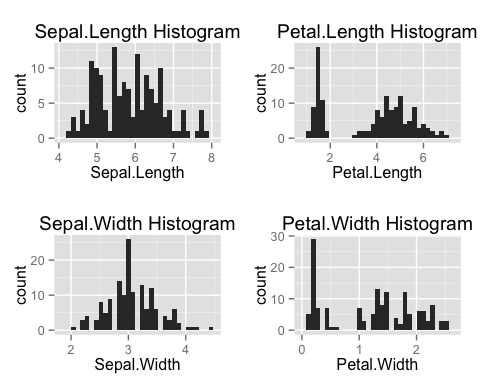
SPSS Scatterplot for Petal Variables:  


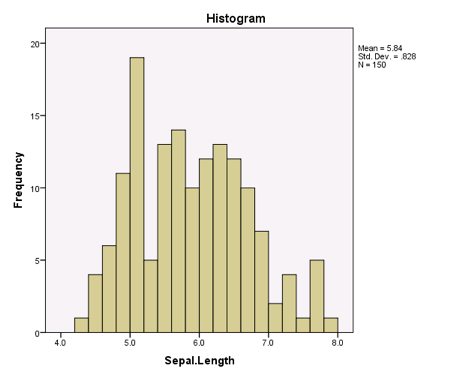
There appears to be a positive linear correlation between Petal.Width and Petal.Length. A classification algorithm should be able to successfully classify the species based on these two variables as the grouping among the species is distinct.

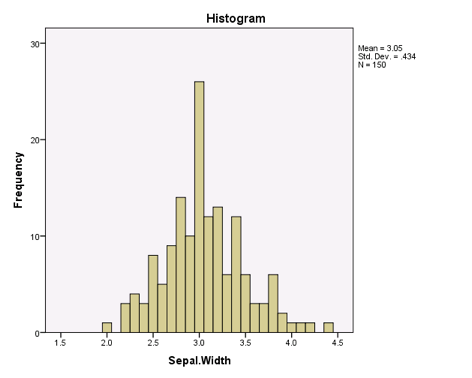
1. Draw the histograms of the four variables and interpret the distributions of each one of the four variables.

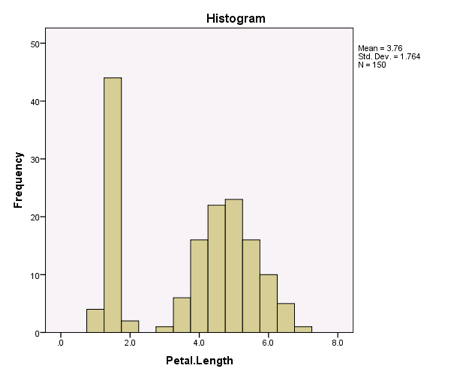
source("/Users/jasminedumas/Desktop/depaul/CSC423/multiplot.R")  
plot3 <- ggplot(iris) +  
 geom\_histogram(aes(x=Sepal.Length)) + ggtitle("Sepal.Length Histogram")  
  
plot4 <- ggplot(iris) +  
 geom\_histogram(aes(x=Sepal.Width)) + ggtitle("Sepal.Width Histogram")  
  
plot5 <- ggplot(iris) +  
 geom\_histogram(aes(x=Petal.Length)) + ggtitle("Petal.Length Histogram")  
  
plot6 <- ggplot(iris) +  
 geom\_histogram(aes(x=Petal.Width)) + ggtitle("Petal.Width Histogram")  
  
multiplot(plot3, plot4, plot5, plot6, cols=2)

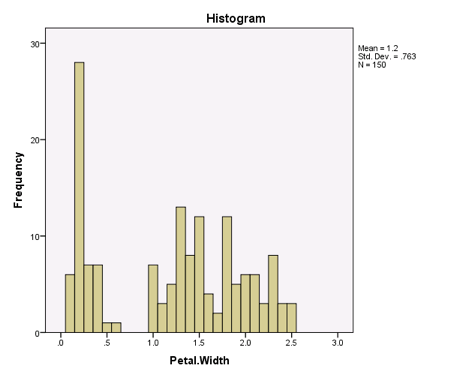
## stat\_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.  
## stat\_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.  
## stat\_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.  
## stat\_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.



SPSS Histograms for the four variables:  




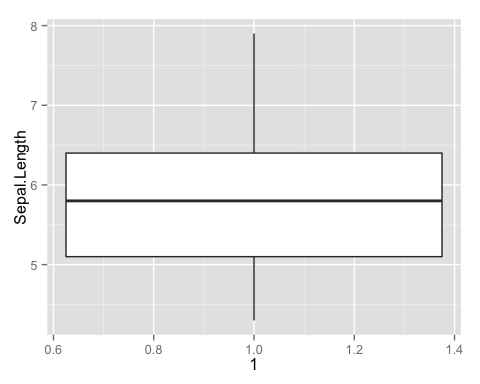




* The histogram for the Sepal.Length Histogram appears to have a normal distribution but there seems to be distinct peaks possibly from the distribution between the different species.
* The histogram for Sepal.Width appears to have a normal distribution with a clear central peak.
* The histogram for Petal.Length appears to have two distinct groups each displaying a normal distribution possibly from the distribution between the different species. If not considering the different species the histogram can also be described as being skewed right as the right hand-side has a peak with a higher count/kurtosis.
* The histogram for Petal.Width appears to have two distinct groups and a singular peak. The distribution appears to be skewed right.

1. Determine if there are any outliers in the data with respect to the sepal length: *There appears to be no outliers for Sepal.Length.*

# Graphically depicted in a boxplot  
plot7 <- ggplot(iris, aes(x=1, y=Sepal.Length)) +  
 geom\_boxplot(outlier.colour = "green")  
plot7

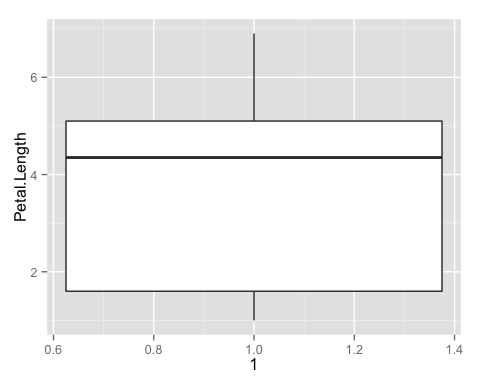


# http://stackoverflow.com/a/33525389/4143444  
is\_outlier <- function(x) {  
 return(x < quantile(x, 0.25) - 1.5 \* IQR(x) | x > quantile(x, 0.75) + 1.5 \* IQR(x))  
}  
is\_outlier(iris$Sepal.Length) # no outliers detected

## [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [34] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [45] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [56] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [67] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [78] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [89] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [100] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [111] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [122] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [144] FALSE FALSE FALSE FALSE FALSE FALSE FALSE

1. Repeat d. for the petal length.: *There appears to be no outliers for Petal.Length*

# Graphically depicted in a boxplot  
plot8 <- ggplot(iris, aes(x=1, y=Petal.Length)) +  
 geom\_boxplot(outlier.colour = "green")  
plot8



# http://stackoverflow.com/a/33525389/4143444  
is\_outlier <- function(x) {  
 return(x < quantile(x, 0.25) - 1.5 \* IQR(x) | x > quantile(x, 0.75) + 1.5 \* IQR(x))  
}  
is\_outlier(iris$Petal.Length) # no outliers detected

## [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [34] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [45] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [56] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [67] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [78] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [89] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [100] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [111] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [122] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [144] FALSE FALSE FALSE FALSE FALSE FALSE FALSE