University of Notre Dame, Computer Science and Engineering

# CSE 40647/60647 Data Mining — Assignment 1
# Due Date: February 10th, 2014 at 11:59pm ET

## Data Understanding

February 1, 2014

This assignment will require you to implement and interpret some of the data understanding concepts that were introduced in class, such as *summary statistics* and *data visualization*. Further, you will be working with real-world data retrieved from an online repository, and while you will be asked to utilize a variety of modules and functions, these have all been covered in the lecture demos. Keep in mind that the main objective of this assignment is to highlight the insights that we can derive from the data understanding process—the coding aspect is secondary. Accordingly, you are welcome to consult any online documentation and/or code that has been posted to the course website, so long as all references and sources are properly cited. You are also encouraged to use code libraries, so long as you acknowledge any source code that was not written by you by mentioning the original author(s) directly in your source code (comment or header).

**You are expected to submit a single IPython Notebook file following the same instructions and naming convention described in Assignment 0. Answers to the conceptual questions can be embedded to the notebook file as *markdown* cells, and you may use *heading* cells to further organize your document.**

## 1 Iris Dataset (20 points)

Using your own module of choice (we recommend pandas), download the Iris flower dataset available **here**[1] into a DataFrame. For more details about the dataset and to obtain the feature names, check **this link**. It is always recommended to familiarize yourself with the data you intend to use for data mining purposes. The Iris dataset in particular has a rich history, having been introduced in 1936 by Sir Ronald Fisher, often considered one of the fathers of modern statistical theory.

[1] See the code we provided for Data Transformation here for an example of how this can be done.

### 1.1 Summary statistics

Print the first 5 elements of your DataFrame using the command `head()`. How many features are there and what are their types (e.g., numeric, nominal)?

Compute and display summary statistics for each feature available in the dataset. These must include the minimum value, maximum value, mean, range, standard deviation, variance, count, and 25:50:75% percentiles.

### 1.2 Data Visualization

#### 1.2.1 Histograms

To illustrate the feature distributions, create a histogram for each feature in the dataset. You may plot each histogram individually or combine them all into a single plot. When generating histograms for this assignment, use the default number of bins. Recall that a histogram provides a graphical representation of the distribution of the data.

#### 1.2.2 Boxplots

To further assess the data, create a boxplot for each feature in the dataset. All of the boxplots will be combined into a single plot. Recall that a boxplot provides a graphical representation of the location and variation of the data through their quartiles; they are especially useful for comparing distributions and identifying outliers.

## 2 Pen-Based Handwritten Digits Dataset (20 points)

Repeat the same process described in Section 1 but this time load **this dataset**, which we will refer to as Digits. Note that Digits is a much larger dataset than Iris, both with respect to the number of instances and the number of features. A description of the dataset can be found **here**.

## 3 Conceptual Questions (20 points)

Consider the histograms you generated for the Iris dataset. How do the shapes of the histograms for petal length and petal width differ from those for sepal length and sepal width? Now consider just the petal width histogram. Is there a particular value of petal length (which ranges from 1.0 to 6.9) where the distribution of petal lengths (as illustrated by the histogram) could be best segmented into two parts?

Now consider the boxplots you generated for the Iris dataset. There should be four boxplots, one for each feature. Based upon these boxplots, is there a pair of features that appear to have significantly different medians? Recall that the degree of overlap between variabilities is an important initial indicator of the likelihood that differences in means or medians are meaningful. Also, based solely upon the box plots, which feature appears to explain the greatest amount of the data?

Lastly, consider the boxplots you generated for the Digits dataset. Do you observe any outliers? If so, for what features? Now consider the corresponding histograms. What sort of

distribution do the second and forth features display? With that in mind, explain the outliers, or lack thereof, in terms of what you observe from the histograms.

## 4 Graduate Student Portion (20 points / +10 points for undergraduates)

Use random sampling with replacement to generate 5 new datasets that are each the same size (same number of instances) as the original Iris dataset. Now compute and display the summary statistics for each of these new datasets. Are they similar? Use at least one form of visualization to contrast these datasets.