Northeastern University, College of Computer and Information Science

CS 6220 Data Mining — Assignment 2

Exploring Data with Pandas

Prior to beginning your work on this assignment, download and run this notebook file(https://goo.gl/BFprVd), which will cover some basics on data exploration, loading data, extracting basic statistics from the various features, and generating visualizations.

Assignment Description:

This assignment will require that you implement and interpret some of the data understanding concepts that were introduced in class, such as summary statistics and data visualizations. 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 notebook files that were shared. 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 so long as all references and sources are properly cited. You are also encouraged to use code libraries, but be sure to acknowledge any source code that was not written by you by mentioning the original author(s) directly in your source code (comment or header).

Submission:
Submit your ipynb file through the Assignment Submission Portal as done in Assignments 1.

1 Iris Dataset [35 points]

Using your own module of choice (we recommend pandas), download the Iris flower dataset available HERE (http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data) into a DataFrame. For more details about the dataset and to obtain the feature names, check this link(http://archive.ics.uci.edu/ml/datasets/Iris). It is always recommended that you 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.

Below is a snippet of code that you can use to load that dataset into a pandas dataframe:

Listing 1: Loading Iris Dataset

```
import pandas as pd
fileURL = "http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
df = pd.read_csv(fileURL, names=["Sepal Length", "Sepal Width",
                                 "Petal Length", "Petal Width",
                                 "Name"], header=None)
```

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

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.

Box Plots: 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 [35 points]

Repeat the same process described in Part 1, but this time load THIS DATASET(http://archive.ics.uci.edu/ml/machine-learning-databases/pendigits/pendigits.tra). Note that the Digits Dataset is much larger than the Iris dataset, both with respect to the number of instances and the number of features. A description of this dataset can be found here(http://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits).

3 Conceptual Questions [30 points]

Answer the following questions about the analysis you just performed. Include the answers to this questions as text content (using markdown or text cells on Jupyter notebook) in the same notebook file used for parts 1 and 2.

3.1 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 length 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?

3.2 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?

3.3 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