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 fromthe 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 [60 Points]

Using your own module of choice (we recommend pandas), download the Iris flower dataset available http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.dataHERE (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 familiar- ize 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.

1.1 Summary Statistics [10 Points]

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 [25 Points]

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.

Pairwise Plot: To understand the relationship between the features, create scatter plot for each pair of the features. If there are are n features in the dataset, there should be $nC2$ plots.

Class-wise Visualization: Create histograms for each feature in a similar way for each of the different classes present in the data.

1.3 Conceptual Questions [25 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 visualization.

1. How many features are there? What are the types of the features (e.g., numeric, nominal, discrete, continuous)?

2. From the histograms of the whole data, how do the shapes of the histograms for petal length and petal width differ from those for sepal length and sepal width? 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. 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?

4. From the pairwise plots of the features, which features are most correlated from the plots? Mention at least three pairs.

5. Compare the histograms of each class to the histograms of the whole dataset. What differences do you see in the shapes?

2 Air Quality Dataset [40 Points]

Download the Air Quality dataset from here (http://archive.ics.uci.edu/ml/machine-learning-data 00360/AirQualityUCI.zip). Note that this 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/Air+Quality). Download the dataset in your machine, and then unzip. Use the *AirQualityUCI.xlsx* file for the data. You can use `pandas.read_excel('AirQualityUCI.xlsx')` to read the file in DataFrame.

2.1 Summary Statistics [5 Points]

Similarly as in Section 1, 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.

2.2 Data Visualization [15 Points]

Similarly as in Section 1, create histograms and boxplots for the dataset. Now, create boxplots without the outliers. You can use `showfliers=False` to remove outliers from the boxplots.

2.3 Conceptual Questions [20 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 visualization.

1. From the histograms, what abnormality can you see?
2. What abnormality can you see from the summary statistics?
3. How can you remove the abnormality from the data?
4. Show how does the histograms looks like after removing the abnormalities from the data?