**IE 6318 Data Mining and Analytics**

**Homework 1**

Data Exploration

1. Download the “Iris.data” data set from <http://archive.ics.uci.edu/ml/>. The dataset has 4 attributes; the 5th column is the class label. Do a raw data processing to replace the names of the three classes by numerical numbers 1, 2 and 3, save the data into a file that can be loadable by your program. For example, in Matlab you can load and save the data as .mat file for convenient use.

2. Explore the Iris dataset and report the following:

1) 2D scatter plots of the four attributes. (In Matlab use function ‘plotmatrix’ or ‘gplotmatrix’)

2) 3D scatter plot of three attributes (sepal length, sepal width, petal width). (In Matlab use function ‘scatter3’)

3) Visualization of the feature matrix (column 1-4). (In Matlab use function ‘imagesc’)

4) Histogram of the four attributes for the three classes. (In Matlab use function ‘hist’)

5) Boxplots of the four attributes for the three classes. (In Matlab use function ‘boxplot’)

6) Calculate the correlation matrix of the four attributes and visualize the correlation matrix.

7) Parallel coordinates plot of the four attributes.

3. Practice Data Distance Measures

1) Make a function for Minkowski Distance. (3 function inputs: vector A, vector B, and order r)

2) Make a function for T-statistics Distance. (3 function inputs: time series A, time series B)

3) Make a function for Mahalanobis Distance. (3 function inputs: vector A, vector B, and covariance matrix M)

4. Assume a new iris sample S has a feature vector of [5.0000, 3.5000, 1.4600, 0.2540]. Calculate the distances between the new sample and the 150 samples in the iris dataset using the distance functions you made.

1) Calculate Minkowski distances with r = 1, 2, 100, respectively, and plot the obtained distances.

2) Calculate Mahalanobis distances and plot the obtained distances.

5. We provide a dataset with two time series in HW1\_DataMining.txt file. Perform the following analysis:

1) Plot the generated two time series in one plot

2) Calculate the T-statistics distance between the two time series.

3) Calculate the correlation of the two time series

4) Normalize the feature matrix of the IRIS dataset such that after normalization each feature has a mean of 0 and a standard deviation of 1.