IE 5300 Data Mining and Analytics
Homework 1

Data Exploration and Classification Using KNN and Decision Tree

1. Download the "Iris.data" data set from http://archive.ics.uci.edu/ml/, and convert the raw dataset into Matlab data format (.mat). The dataset has 4 feature attributes; the last column is the class label. Replace the names of the three classes by numerical numbers 1, 2 and 3. Give a brief description of the dataset in the first part of your report.

2. Explore the Iris dataset , and report the following:

1) 2D scatter plots of the four attributes use Matlab function 'plotmatrix' or 'gplotmatrix'.

2) 3D scatter plot of three attributes (sepal length, sepal width, petal width) using the Matlab function 'scatter3'.

3) Visualization of the feature matrix (column 1-4), use Matlab function 'imagesc'.

4) Histogram of the four attributes for the three classes, use Matlab function 'hist'.

5) Boxplots of the four attributes for the three classes, use Matlab 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 Matlab function for Minkowski Distance. (Three function inputs: sample A, sample B, and distance order r)

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

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

Assume a new iris sample S has a feature vector of [5.0000, 3.5000, 1.4600, 0.2540]. Calculate the distances of the new sample to the 150 samples in the iris dataset

4) using Minkowski distance with r = 1, 2, 100, respectively. Plot the obtained distances.

5) using Mahalanobis distance. Plot the obtained distances.

Generate two time series data by the code: X = mvnrnd([0;0],[1 .3;.3 1],100);

6) Plot the generated two time series in one plot

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

8) Calculate the correlation of the two time series

9) 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.