**COSC 757 Data Mining Assignment 1**

**Spring 2015**

Due date: 02/25/2015 11:59 pm

Instructions: This is an individual assignment. Use Blackboard to submit your answers on the due date (no hard copies please). Late submissions will receive a zero grade.

1) **Exploratory data analysis with R:**  Choose a dataset from the UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml/>) to explore using R. You can use the functions provided in Lab 1 as a reference. Write a brief report showing your exploratory data analysis. You should at least show descriptive statistics for the data including visualizations of the distribution of the attributes, relationships between attributes, and distance between data objects.

2) **Algorithm for binary dissimilarity:** Download the Congressional Voting Records Dataset (<http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records>). Propose an algorithm in pseudo code or your favorite programming language (feel free to write an R script) that calculates the dissimilarity between the binary attributes in this dataset as shown in the text on page 71 in example 2.18. The output of the algorithm should be an nxn dissimilarity matrix. The algorithm should be void of already implemented distance functions (e.g. daisy). I want to see how you solve the problem.

3) **Data reduction and data normalization:** Download the Automobile dataset from (<http://archive.ics.uci.edu/ml/datasets/Automobile>). This dataset consists of 26 attributes including attributes related to the specification of an auto in terms of various characteristics, its assigned insurance risk rating, and its normalized losses in use as compared to other cars.

a) Select a data reduction technique to remove the number of dimensions in the dataset and show your results.

b) Some of the methods for data reduction require the data to be normalized (i.e. rescaling data measured in differing units. Use R to normalize the numerical attributes of this dataset (hint: the scale() function can be used to compute Z scores).

c) Select a continuous variable and experiment with two methods to bin the variable into discrete categories.