The University of Texas at Austin
CE 395 R 5-Data Mining – Spring 2018

Homework 1 - Individual Assignment

DUE DATE: 02/08/18

Please submit the assignment at the beginning of the lecture.

GUIDELINES:

- Visit and browse the UCI Machine Learning Repository and KD Nuggets, available at the following URLs:
  - http://archive.ics.uci.edu/ml/
  - http://www.kdnuggets.com/datasets/index.html
- Select two data sets that have data quality problems. ***One of the data sets must have at least one numerical attribute***. Please note that the data set files are accessed through ftp sites and can be downloaded to your computer. Most data sets have a file with a ".names" extension. This file contains general information about the data set, such as number of instances, attribute names and types, among others. The files with a ".data" extension store the data, usually in the CSV format that can be opened using Notepad, Excel, and other software.
- ***Problem 1:*** For the two data sets selected, identify data quality problems (e.g.: missing values, noise, and inconsistent data) and propose a solution to these problems. The solution proposed for the first data set can be different than the solution proposed for the second. Create two new "clean" data set files.
- ***Problem 2:*** For ***one*** numerical attribute in ***one*** of the data sets, normalize the data using one of the methods discussed in class.

SUBMISSION:

The following files should be submitted: ***Make sure that you include the homework number, your name, the file number, and the course number in the file names to facilitate identification (e.g. "HW1_Joe__File 1_CE395R5.doc")***

- File 1 (Problem 1 Item a): one document (in Word or PDF format) presenting the data sets selected, the data quality problems identified, and the proposed solutions;
- Files 2 and 3 (Problem 1 Items b and c): the two "clean" data set files (in CSV or Excel format).
- File 4 (Problem 2): the new version of the data set file containing the normalized data (in CSV or Excel format).