Your Name: _____

(This is an INDIVIDUAL assignment)

CSC 7442: Data Mining and Knowledge Discovery from Databases

Professor E. Triantaphyllou
Louisiana State University
Department of Computer Science
Fall 2016

Today's date: **Monday, October 31, 2016**
Due dates: **Part 1 (on Similarities): Monday, November 14, 2016**
Part 2 (on Clustering): Monday, November 28, 2016

For each deadline submit your files by the closing of the day (i.e., 5:00 PM). Send them to me via email to trianta@csc.lsu.edu

MAIN GOALS:
For Part 1: The main goal is to gain some experience with data processing issues as related to similarity assessment.
For Part 2: The main goal is to gain some experience with clustering of data. Explore algorithmic and search issues in both subject areas.

1. Description of the Available Data

Attached to this announcement please find the following files:

1) PIMA_Diabetes_2015 in ARFF format
2) Training_Examples_TEXT_File1 in text format
3) Available_Distances_of_Pairs_of_Records1 in text format

The first file was retrieved from UCI's machine Learning Data Set Repository. It has all the data records available on that site. The second file is a random subset of the first data set (approximately of size 2/3 of the original size). This data set (i.e., the second file) was used to form a number of random pairs of records (150 out of all possibilities). Some may be duplicates.

It is assumed that a panel of experts has examined each such pair and came up with a value that assesses in their collective opinion how close the two records are of each other. The closer they are, the more similar the corresponding records are assumed to be. For this study such assessments have been **simulated** as follows. A Euclidean distance was assumed to exist and then it was used on pairs of vectors to come up with these distance values. This Euclidean distance expresses a weighted sum of the squared differences of the attribute values of pairs of records.

This data set is defined on 9 attributes. The last one is the class attribute. For the purpose of this study please ignore the class attribute (which assumes binary value). Thus, we have 8 attributes (all of which are numerical and ordinal). These values have been normalized by dividing the values of each attribute by the maximum value for that

attribute. The values normalized as described above were next used in the "hidden" weighted Euclidean distance formula.

2. Tasks for Part 1 (on Similarities)

Part 1 focuses on determining similarity values. For that you will have to first estimate the form of the Euclidean formula used to derive the values presented in the third file. In other words, you will have to use some type of search to estimate he weight values used in that formula. You may want to use a search method based on **steepest descent** as described in class. Once the "hidden" weighted Euclidean formula has been estimated, use it to determine similarity values between any pair of records from the second file.

In your report for Part 1 you will have to present what the Euclidean formula you derived (along with the weights used) and the values of the estimated distances of the pairs described in the input file. Determine the total difference (say, expressed as the sum of all squared differences divided by the number (= 150) of pairs).

3. Tasks for Part 2 (on Clustering)

Part 2 focuses on clustering. Thus, once you have estimated the "hidden" weighted Euclidean formula used in Part 1 to determine the distance values, use it to cluster the data described in the second file. The number of clusters of the proposed clustering scheme is up to you to decide.

In your report for Part 2 you will have to describe what objective function you tried to optimize in your clustering scheme. You may have to present and analyze a number of alternative clustering schemes and then describe and explain which one is the best one. For clustering data you can use any one of the methods we discussed in class.

Attach this form on front of your answers