COSC 5340 Data Science and Big Data Analysis

Assignment 1

Dissimilarity/Distance Matrix

Choose one dataset from the UCI (<https://archive.ics.uci.edu/ml/datasets.php>) as your input data to analysis the similarity/dissimilarity of the objects in the dataset.

Download the dataset.

Use Python or any other programming language that you are familiar with to write a program to:

1: analysis and pre-process the dataset using preprocessing techniques needed for your dataset we covered in our course, e.g. handling missing.

2: Chose the proper distance/similarity function for each attribute in your dataset, your program needs to compute the distance/similarity for each attribute.

3: Combining distance/similarity of all attributes in the dataset, compute the overall distance/similarity between any pair of objects in the dataset. Write your distance/similarity matrix to an output file.

Upload your complete projects includes: your program, your input dataset, output file of the distance/similarity matrix, a read me file for the grader to compile your program on his/her machine and briefly explain any techniques that you applied in your project.

Bonus (50): use the same technique that you applied on a second dataset to compute the distance/similarity function.