C S 488/508 Introduction to Data Mining
Homework 1: Proximity and Pre-processing

1 Objective

This homework requires you to calculate the proximity among data points and conduct data preprocessing. This is an *individual* homework.

2 Questions

Q1. Data proximity

You are given the below four data points.

```
p1: 1   2   4   5
p2: 4   3   2   1
p3: 1   1   1   0   0
p4: 1   0   0   0   1
```

Please answer the following questions.

(1) (15 points) Write program to calculate the (1) Euclidean distance, (2) Manhattan distance, and (3) Cosine similarity between p1 and p2. You are NOT allowed to use functions that directly calculate these distances in exiting packages. Output the calculation results.

(2) (15 points) Write program to calculate the (1) SMC, (2) Jaccard coefficient, and (3) Hamming distance between p3 and p4. You are NOT allowed to use functions that directly calculate these distances in exiting packages. Output the calculation results.

(3) (15 points) Write program to calculate the DTW distance between p1 and p3. You can use functions in existing packages or write your own implementation. Output the distance and the warping path. From the warping path, draw a figure to show how these two data points are aligned. The figure can be put in the PDF file.

(4) (10 points, Non-programming question) Give a **concrete example** to show that Cosine similarity can better capture the similarity than SMC.

Q2. Data preprocessing

Use the dataset **BreastCancerCoimbra_imbalanced.csv**, which can be downloaded from Canvas. The dataset description about the attribute information can be found from `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra#`.
Please complete the following tasks.

(1) (15 points) Write program to create a sample of size 10 which is randomly selected (without replacement) from the original data.

(2) (15 points) Write program to create a stratified sample of size 10 (without replacement) from the original data.

(3) (15 points, Non-programming question, **CS 508 only**) In the random sample and the stratified sample that you generate above, are there any major difference regarding the number of instances for each class label? Why?

General requirements

For all the programming questions, put the code for all these question to one file. Please properly organize the code to make grading easy.

For the non-programming questions, put the answers to one `hw_non-program_yourlastname.PDF`. The answers need to be concise and informative. Further requirements are as follows. (a) Use A4 page size. (b) Margin at each of the top, bottom, left, and right sides is 1.0 inch. (c) The font size is 10pt; font type is Times New Roman. (d) Single line space.

3 Submission instructions

A zipped file `hw-lastname.zip` consisting of all the code and the PDF file.

4 Grading criteria

(1) CS 508 students need to answer all the questions.

(2) CS 488 students do not need to answer questions marked with **(CS 508 only)** although you have the freedom to work on them. Your scores will be scaled to 100. If CS 488 students answer the questions marked with **(CS 508 only)**, you will not have any points deducted if your answers are wrong; you will not get any extra points either if your answers are correct.

(3) The score allocation has been put beside the questions.

(4) Please make sure that you test your code **thoroughly**.

(5) FIVE points will be deducted if files are not submitted in the required format.