C S 488/508 Introduction to Data Mining
Homework 5: Clustering

Objective

In this homework, you will do exercises to understand basic *concepts* of several representative clustering algorithms.

Q1. (30 pts) K-means

Given a data set with the following points, *manually* run the K-means algorithm with two iterators (the initial iteration and one more iteration) and report the discovered clusters for K=2.

```
x1=(1, 1),  x2=(5, 5), x3 =(5, 6), x4=(7, 8), x5=(8, 6), x6=(8, 7)
```

Run the algorithm by utilizing `x1` and `x2` as the initial cluster centroids. At each iteration, show the initial centroids, the clusters formed at that iteration, the new centroids, and the SSE value. Use Euclidian distance to calculate distances.

(NOTE: you can utilize software (e.g., Excel, Python) to facilitate your calculation in each step.)

Q2. (40 points) Hierarchical

Use the **distance** matrix in the following table to do hierarchical clustering.

(a) (20 pts) Use MIN to calculate cluster distances.

(b) (20 pts) (**CS 508 only**) Use AVERAGE to calculate cluster distances.

At each step show the updated matrix with distances between clusters. Show your results by drawing the final dendrogram.

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

Q3. (30 pts) MST-based clustering method

Given a data set with the following points, *manually* run the Minimum Spanning Tree (MST) based clustering algorithm.

```
x1=(1, 1),  x2=(5, 5), x3 =(5, 6), x4=(7, 8), x5=(8, 6), x6=(8, 7)
```

(a) (15 pts) Draw the minimum spanning tree you get from the dataset (i.e., after finish running Step 1 of the algorithm). Use Euclidean distance for distance calculation. Succinctly explain your calculation.

(b) (15 pts) What are the clusters after you run two iterations of Step 3 of the algorithm? Please explain succinctly.

Submission instructions

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

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