COM S 5730 Homework 4

1. Please put required code files and report into a compressed file "HW4_FirstName_LastName.zip"
2. Unlimited number of submissions are allowed on Canvas and the latest one will be graded.
3. Due: **Tuesday Nov. 05, 2024 at 11:59pm.**
4. No later submission is accepted.
5. Please read and follow submission instructions. No exception will be made to accommodate incorrectly submitted files/reports.
6. All students are required to typeset their reports using latex. Overleaf (`https://www.overleaf.com/learn/latex/Tutorials`) can be a good start.

1. (40 points) **Hierarchical clustering**

   Use the similarity matrix in Table 1 to perform (1) single (MIN) and (2) complete (MAX) link hierarchical clustering. Show each step with dendrogram and the corresponding similarity matrix update. The dendrogram should clearly show the order in which the points are merged. Suppose we choose to use 3 clusters, Show the cut in each final dendrogram.

   Table 1: Similarity matrix.

	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

2. (30 points) **K-Medians Clustering**

   The K-means algorithm can be summarized as below:

   ```
   (a) Select K points as the initial centroids.
   (b) repeat
   (c)     Form K clusters by assigning all points to the closest centroid.
   (d)     Recompute the centroid of each cluster.
   (e) until The centroids don't change.
   ```

   K-medians clustering is a variation of k-means clustering where it calculates the median for each cluster to determine its center instead of using the mean. Also, K-medians makes use of the Manhattan distance for points assignment.

   (a) (8 points) Please show the algorithm of K-medians in the above format.

   (b) (6 points) Please explain how you will compute the median for each cluster.

   (c) (6 points) Does K-medians help to avoid the outlier problem? Justify your answer.

3. (30 points) **Principal Components Analysis**

   Given three data points: $(-1, -1), (0, 0), (1, 1)$.

   (a) (10 points) Show the first Principal Component (actual vector) without using Eigendecomposition. Justify your answer.

   (b) (10 points) If use the 1^{st} principle component to transform the data into 1-d space. What are the new data?