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 1b 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

(a) **Single (Min)**

- **State 1:** Merge p2 and p5 (Highest similarity: 0.98).

	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

Similarity matrix: Before merging p2 and p5

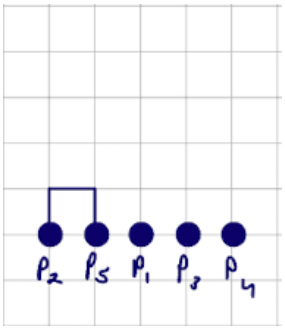

figure S_1: Dendrogram after merging p2 and p5

- **State 2:** Merge {p2, p5} and p3 (Current highest similarity: 0.85)

- Updated Similarities:
  - s({p2, p5}, p1) = max(0.10, 0.35) = 0.35
  - s({p2, p5}, p3) = max(0.64, 0.85) = **0.85**
  - s({p2, p5}, p4) = max(0.47, 0.76) = 0.76

	p1	**{p2}{p5}**	**p3**	**p4**
p1	1.00	0.35	0.41	0.55
{p2}{p5}	0.35	1.00	0.85	0.76
p3	0.41	0.85	1.00	0.44
p4	0.55	0.76	0.44	1.00

Similarity matrix: Before merging {p2, p5} and p3

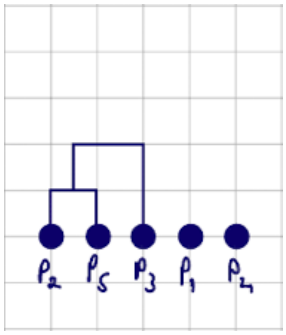

figure S_2: Dendrogram after merging {p2, p5} and p3

- **State 3:** Merge {p2, p3, p5} and p4 (Next highest similarity: 0.76)
  - Updated Similarities:
    - s({p2, p3, p5}, p1) = max(0.10, 0.41, 0.35) = 0.41
    - s({p2, p3, p5}, p4) = max(0.47, 0.44, 0.76) = **0.76**

	p1	**{p2}{p5}{p3}**	**p4**
p1	1.00	0.41	0.55
{p2}{p5}{p3}	0.41	1.00	0.76
p4	0.55	0.76	1.00

Similarity matrix: Before merging {p2, p3, p5} and p4

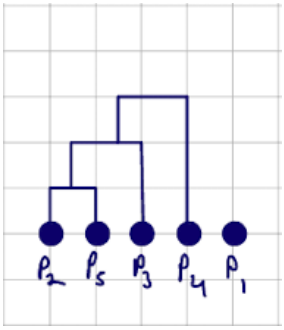

figure S_3: Dendrogram after merging {p2, p3, p5} and p4

- **State 4:** Merge p2, p3, p4, p5 and p1 (Similarity: 0.55)
  - Updated Similarities:
    - s({p2, p3, p4, p5}, p1) = max(0.10, 0.41, 0.55, 0.35) = **0.55**

	p1	{p2}{p5}{p3}{p4}
p1	1.00	0.55
{p2}{p5}{p3}{p4}	0.55	1.00

Similarity matrix: Before merging {p2, p3, p4, p5} and p1

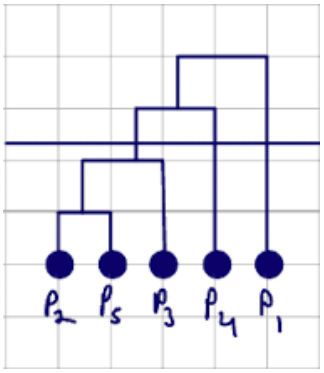

figure S_4: Dendrogram after merging {p2, p3, p4, p5} and p1

Final Clusters:

- **Cluster 1:** p1
- **Cluster 2:** p4
- **Cluster 3:** p2, p3, p5

(b) **Complete (Max)**

- **State 1:** Merge p2 and p5 (Highest similarity: 0.98).

	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

Similarity matrix: Before merging {p2} and {p5}

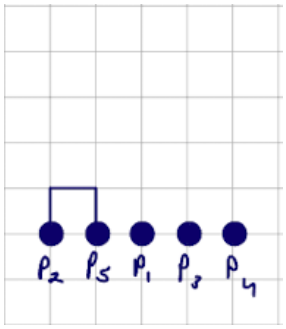

figure S_1: Dendrogram after merging {p2} and {p5}

- **State 2:** Merge {p2, p5} and p3 (Current highest similarity: 0.64)
  - Updated Similarities:
    - s({p2, p5}, p1) = min(0.10, 0.35) = 0.10
    - s({p2, p5}, p3) = min(0.64, 0.85) = **0.64**
    - s({p2, p5}, p4) = min(0.47, 0.76) = 0.47

	p1	{p2}{p5}	p3	p4
p1	1.00	0.10	0.41	0.55
{p2}{p5}	0.10	1.00	0.64	0.47
p3	0.41	0.64	1.00	0.44
p4	0.55	0.47	0.44	1.00

Similarity matrix: Before merging {p2, p5} and {p3}

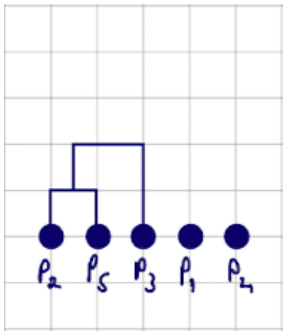

figure S_2: Dendrogram after merging {p2, p5} and {p3}

- **State 3:** Merge p1 and p4 (Next highest similarity: 0.55)
  - Updated Similarities:
    - s({p2, p3, p5}, p1) = min(0.10, 0.41, 0.35) = 0.10
    - s({p2, p3, p5}, p4) = min(0.47, 0.44, 0.76) = 0.44

	p1	{p2}{p5}{p3}	p4
p1	1.00	0.10	0.55
{p2}{p5}{p3}	0.10	1.00	0.44
p4	0.55	0.44	1.00

Similarity matrix: Before merging {p1} and {p4}

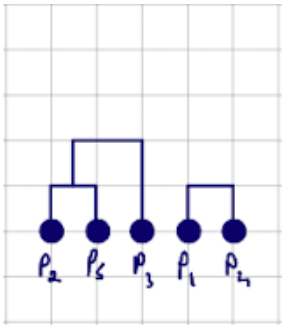

figure S_1: Dendrogram after merging {p1} and {p4}

- **State 4:** Merge p1, p4 and p2, p3, p5 (Similarity: 0.10)
  - Updated Similarities:
    - s({p1, p4}, {p2, p3, p5}) = min(0.10, 0.41, 0.35, 0.47, 0.44, 0.76) = 0.10

	{p1}{p4}	{p2}{p5}{p3}
{p1}{p4}	1.00	0.10
{p2}{p5}{p3}	0.10	1.00

Similarity matrix: Before merging {p1, p4}, {p2, p3, p5}

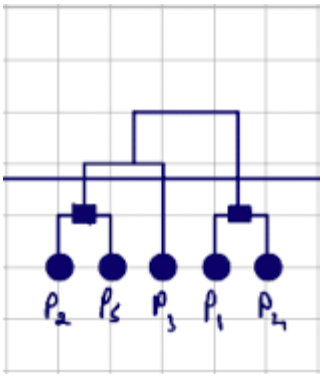

figure S_1: Dendrogram after merging {p1, p4}, {p2, p3, p5}

Final Clusters:

- **Cluster 1:** p2, p5
- **Cluster 2:** p3
- **Cluster 3:** p1, p4

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.

The K-medians algorithm can be summarized as below:

i. Select K points as the initial centroids.
ii. **repeat**
iii. Form K clusters by assigning all points to the closest centroid using Manhattan distance as the distance metric.
iv. Recompute the centroid of each cluster by calculating the median of the data points in that cluster.
v. **until** The centroids don't change.

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

Given a cluster K with X data points $X = [x_1, x_2, ..., x_n]$ where $x_i \in \mathbb{R}^d$, the centroid recomputation for K will be based on the median of each dimension.

For each dimension d, sort the values of X along d. The median, $Med(X^d)$, is calculated as:

$$Med(X^d) = \begin{cases} X[\frac{n+1}{2}] \text{ if n is odd} \\ \\ \frac{X[\frac{n}{2}]+X[\frac{n}{2}+1]}{2}] \text{ if n is even} \end{cases}$$

Compute the median independently for each dimension d, resulting in a new centroid $(Med(X^1), Med(X^2), ..., Med(X^d))$ for the cluster.

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

Yes, K-medians helps to avoid the influence of outliers because the median is not as affected by outliers as the mean. In K-medians clustering, using the median to determine the cluster centroid makes it more robust to extreme values, whereas K-means, which relies on the mean, can have its centroids skewed by outliers.

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.

- **Step 1: Calculate the mean**

$$(\frac{-1+0+1}{3}, \frac{-1+0+1}{3}) = (0, 0)$$

- **Step 2: Subtract the mean out of each point**

(-1, -1): $(-1, -1) - (0, 0) = (-1, -1)$

(0,0): $(0, 0) - (0, 0) = (0, 0)$

(1, 1): $(1, 1) - (0, 0) = (1, 1)$

- **Step 3: Find vector along the maximum distance in the positive direction**

(1, 1) vector along maximum variance.

- **Step 4: Normalize vector by dividing it by its magnitude to have a unit vector for direction.**

(1, 1) Magnitude: $\sqrt{1^2 + 1^2} = \sqrt{2}$

Normalized Vector: $(1, 1) \div \sqrt{2} = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]$

- The 1^{st} principle component is $[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]^T$.

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

- **Step 5: Use the 1^{st} principle component to transform the data into 1-d space**

$$\textbf{(-1, -1):}\ (-1,-1)\cdot[\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}]^T=\frac{-2}{\sqrt{2}}=-\sqrt{2}$$

$$\textbf{(0,0):}\ (0,0)\cdot[\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}]^T=0$$

$$\textbf{(1, 1):}\ (1,1)\cdot[\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}]^T=\frac{2}{\sqrt{2}}=\sqrt{2}$$

The new data is: $-\sqrt{2}, 0, \sqrt{2}$