ml-clustering-proj

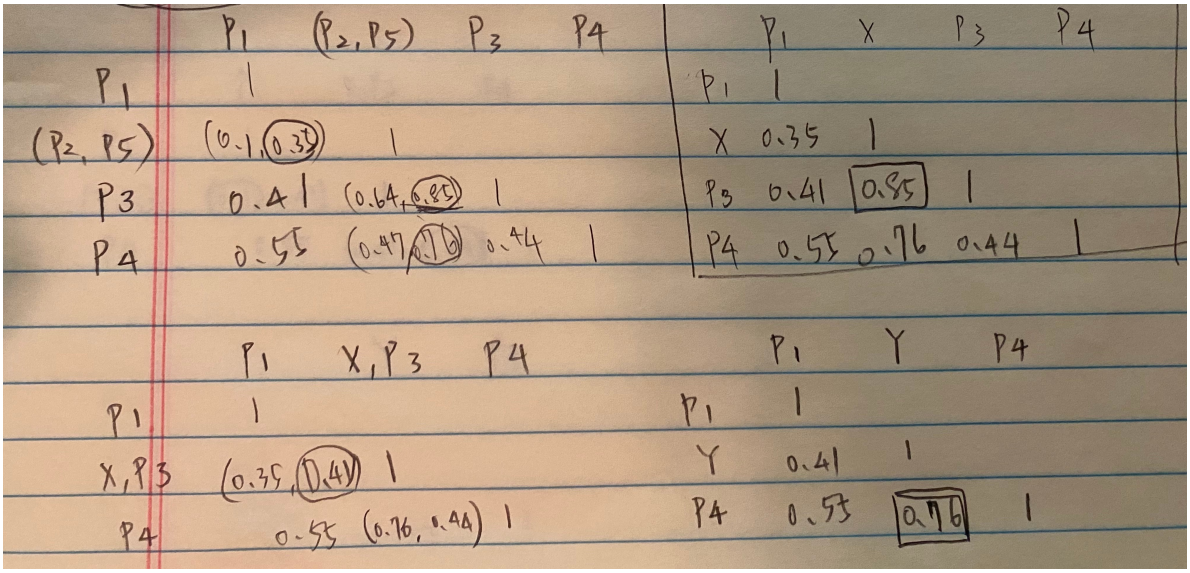
---

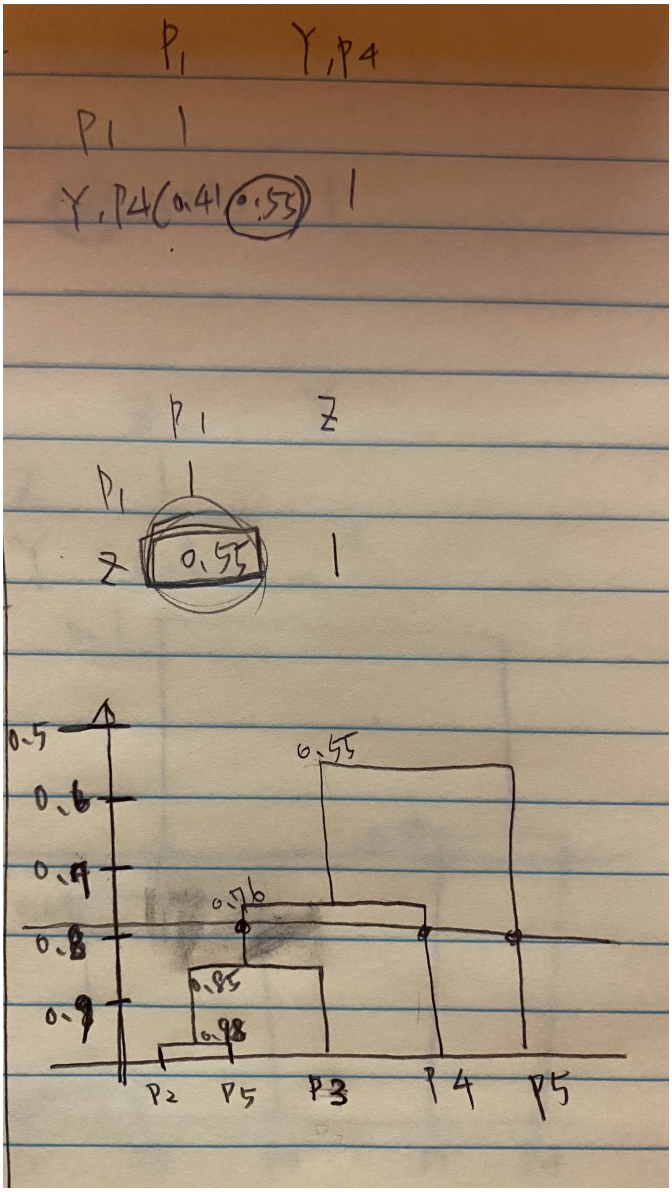
1. **Hierarchical clustering**

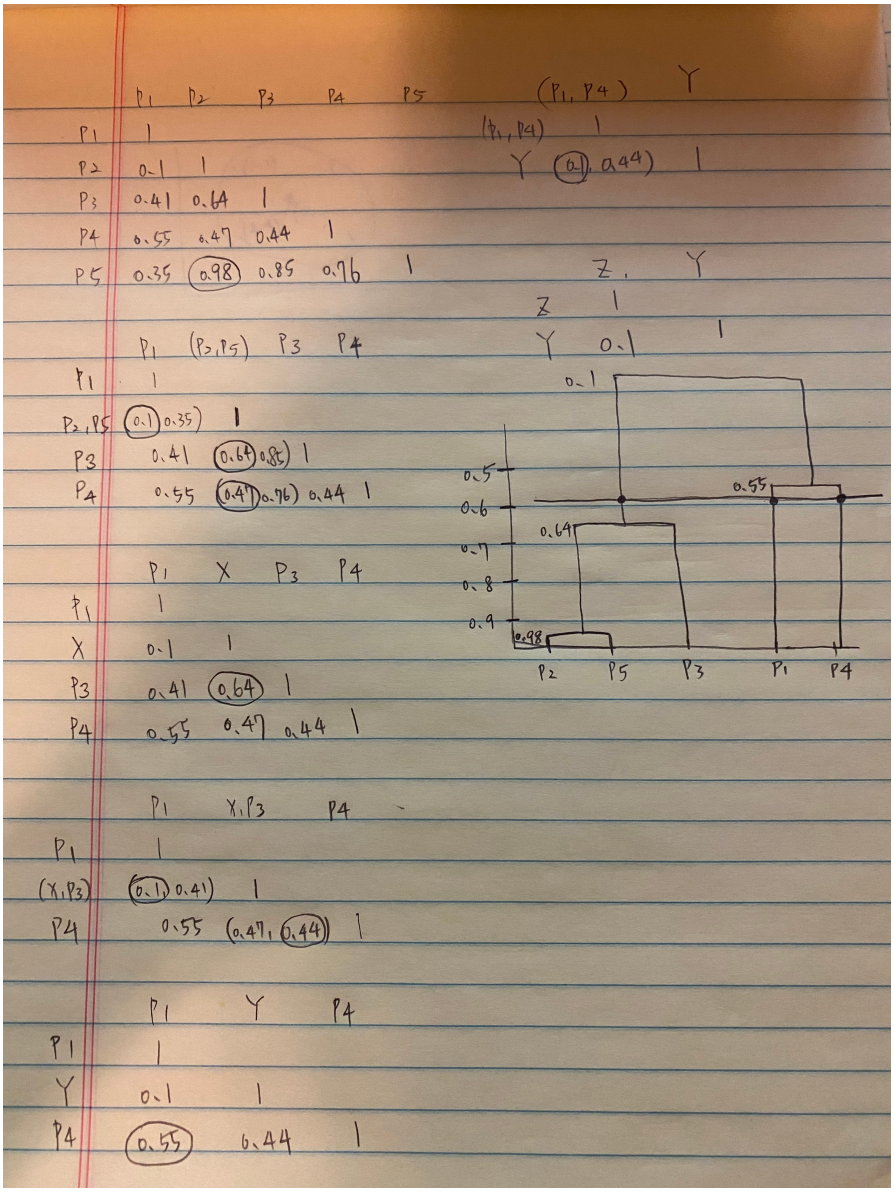
   Use the similarity matrix in Table 1 to perform (1) single (MIN) and (2) complete (MAX) link hierarchical clustering. Show each step with dendrogram and the corresponding similarity matrix update. The dendrogram should clearly show the order in which the points are merged. Suppose we choose to use 3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




P1 Y,P4
P1 1
Y,P4(0.41 0.55) 1
P1 Z
P1
Z 0.55 1
0.5
0.6
0.7
0.8
0.9
0.55
0.76
0.85
0.98
P2 P5 P3 P4 P5



2. **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) Please show the algorithm of K-medians in the above format.

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

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

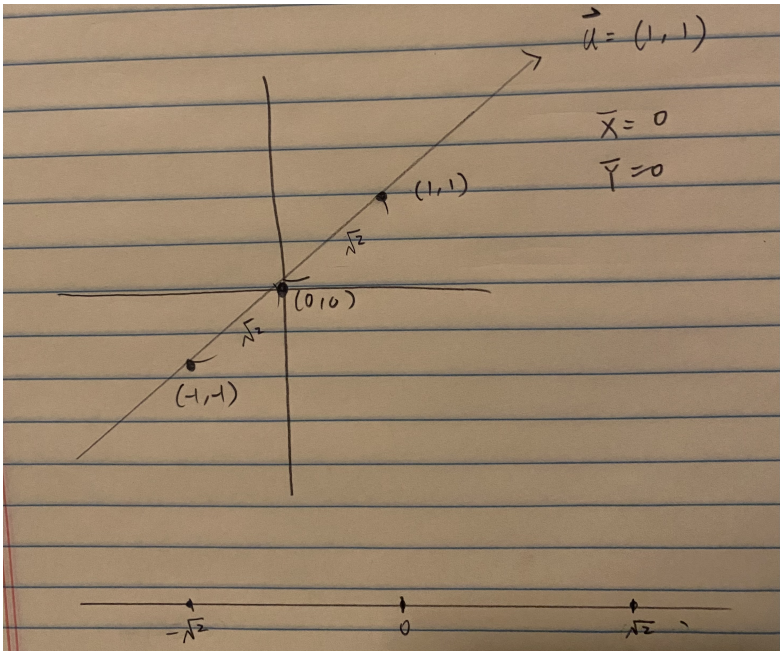
2. (1) (a) Select K points as the initial centroids. (b) repeat (c) Form K clusters by assigning all points to the closest centroid. (d) Recompute the centroid: Change the cluster center by to the median of the assigned points (e) until The centroids don't change. (2) Arrange the value of data points in each cluster in numerical order. Then find the median value. (3) Yes, because we arrange the data by its numerical order. No matter what the values of outliner is, the median would not change. If se use K-mean, the mean will be affected by outliner.

3. **Principal Components Analysis**

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

(a) Show the first Principal Component (actual vector) without using Eigen-decomposition. Justify your answer.

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



(1) vector = (1,1)

(2) $\sqrt{2}, 0, -\sqrt{2}$

4. **Principal Component Analysis:**

In this homework, you will apply the principal component analysis to a collection of handwritten digit images from the USPS dataset. The USPS dataset is in the "data" folder: USPS.mat. The starting code is in the "code" folder. The whole data has already been loaded into the matrix A. The matrix A has shape 3000×256 and contains all the images. Each row in A corresponds to a handwritten digit image (between 0 and 9) with size 16×16. You are expected to implement your solution based on the given codes. The only file you need to modify is the "solution.py" file. You can test your solution by running the "main.py" file.

(a) (15 points) In PCA, we obtain a projection matrix or reduce matrix $\boldsymbol{U} \in \mathbb{R}^{d \times p}$. Based on $\boldsymbol{U}$, we project the original centered data $\bar{\boldsymbol{X}} \in \mathbb{R}^{d \times n}$ into reduced data $\boldsymbol{Z} \in \mathbb{R}^{p \times n}$. Complete the ***_do_pca()*** method. You only need to center the data instead of applying mean normalization. Your code will be tested on $p = 10$, 50, 100, 200, total four different numbers of the principal components.

(b) Based on the projection matrix $\boldsymbol{U}$ and reduce data $\boldsymbol{Z}$, we can reconstruct the original data $\boldsymbol{X'}$ by $\boldsymbol{UZ}$ and adding back the original means. Here you need to Complete the ***reconstruction()*** method to reconstruct the reduced data.

(c) Based on the reconstructed data $\bar{\boldsymbol{X}}'$, we can compute measure the reconstruction error by $||\boldsymbol{X} - \boldsymbol{X'}||_F^2$. Complete the ***reconstruct_error()*** function to measuring the reconstruction error.

(d) Run "main.py" to see the reconstruction results and summarize your observations from the results into a short report. When you run the "main.py" file, a subset (the first two) of the reconstructed images based on p = 10, 50, 100, 200 principal components will be automatically saved on the "code" folder. Please attach these images into your report also.