COM S 573: Machine Learning

Homework #4

Abdurahman Mohammed

April 1, 2022

1

1.1 Single (MIN) Hierarchical clustering

Given the following similarity matrix, we will look for points with highest similarity. We can see that $P2$ and $P5$ are the ones with maximum similarity. Hence, will join them and update our similarity matrix.

Table 1: Similarity matrix.

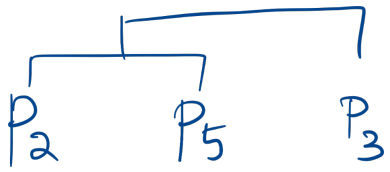
	p1	**p2**	**p3**	**p4**	**p5**
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	**0.98**
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	**0.98**	0.85	0.76	1.00

When constructing our new similarity matrix, we will be choosing the highest similarity measure or in other words the minimum distance two given points.

	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

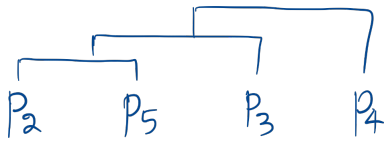
After building the new similarity matrix, we will look for points with highest similarity score. Then, we will find out that point $P3$ has highest similarity with the union in $P2$ and $P5$. So we will update our similarity matrix as well as connect $P3$ with the union of $P2$ and $P5$.

	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



After that, Looking at our similarity matrix, we can see that $P4$ has the highest similarity with our points $P2 \cup P5 \cup P3$. Hence, we will connect them in our dendogram and update our similarity matrix accordingly.

	p1	**p2 ∪ p5 ∪ p3 ∪ p4**
p1	1.00	**0.55**
p2 ∪ p5 ∪ p3 ∪ p4	**0.55**	1.00



The next step will be to create a connection between $P2 \cup P5 \cup P3 \cup P4$ and $P1$. And the final dendogram will look as follows.

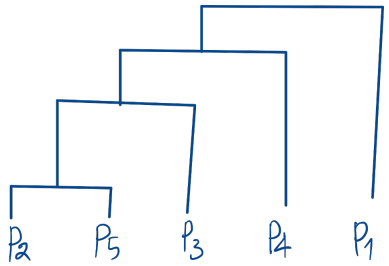
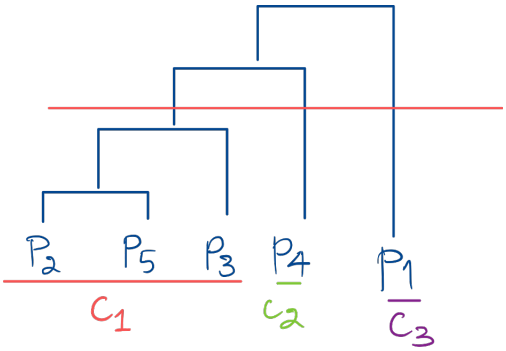


Figure 1: Principal component

If we consider making 3 clusters we can cut the dendogram as follows.



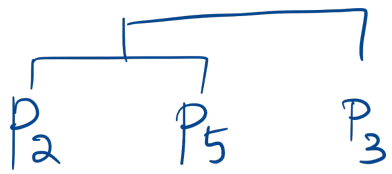
1.2 Complete(MAX) Hierarchical clustering

Given the following similarity matrix, we will look for points with highest similarity. We can see that $P2$ and $P5$ are the ones with maximum similarity of 0.98. Hence, will join them and update our similarity matrix. For the complete(MAX) hierarchical clustering case, When constructing our new similarity matrix, we will be choosing the lowest similarity measure or in other words the maximum distance two given points.

	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	**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

After building the new similarity matrix, we will look for points with highest similarity score. Then, we will find out that point $P3$ has highest similarity with the union in $P2$ and $P5$. So we will update our similarity matrix as well as connect $P3$ with the union of $P2$ and $P5$.

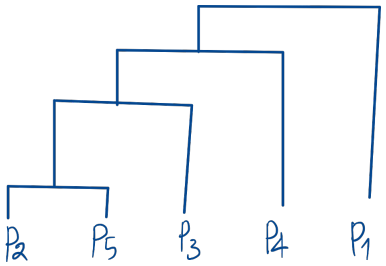


	p1	**p2 ∪ p5 ∪ p3**	**p4**
p1	1.00	0.10	**0.55**
p2 ∪ p5 ∪ p3	0.10	1.00	0.44
p4	**0.55**	0.44	1.00

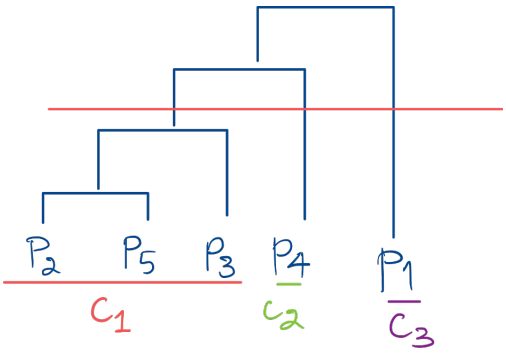
After that, Looking at our similarity matrix, we can see that $P4$ has the highest similarity with our points $P2 \cup P5 \cup P3$. Hence, we will connect them in our dendogram and update our similarity matrix accordingly.

	p1	**p2 ∪ p5 ∪ p3 ∪ p4**
p1	1.00	**0.10**
p2 ∪ p5 ∪ p3 ∪ p4	**0.10**	1.00

The next step will be to create a connection between $P2 \cup P5 \cup P3 \cup P4$ and $P1$. And the final dendogram will look as follows.



If we consider making 3 clusters we can cut the dendogram as follows.



2

2.1

Below is the algorithm for the K-Median clustering. The K-Median algorithm will look like

1. Select k points as initial centroids
2. Repeat
3. Assign every point to the nearest medians
4. Recompute the median using the median of each individual feature.
5. Repeat until medians stop to change.

2.2

Calculating the medians will be the same as the method of calculating the centroids in K-Means. To find the x-coordinates of the medians, we will find the medians of x-coordinates of the data. In the same way, we will us the y-coordinates of the data to find the y-coordinates of the median.

2.3

Using K-Medians can help us fix the problem of outliers in the K-Means algorithm. From a statistical point of view, we know that Median can be thought of as the middle point of data. It is not affected by outliers as much as the mean of a data is. Therefore, using K-Medians algorithm will helps us ensure our algorithm is not affected by outliers.

3

First let's plot the data points. After normalizing this principal component to be a unit length,we get

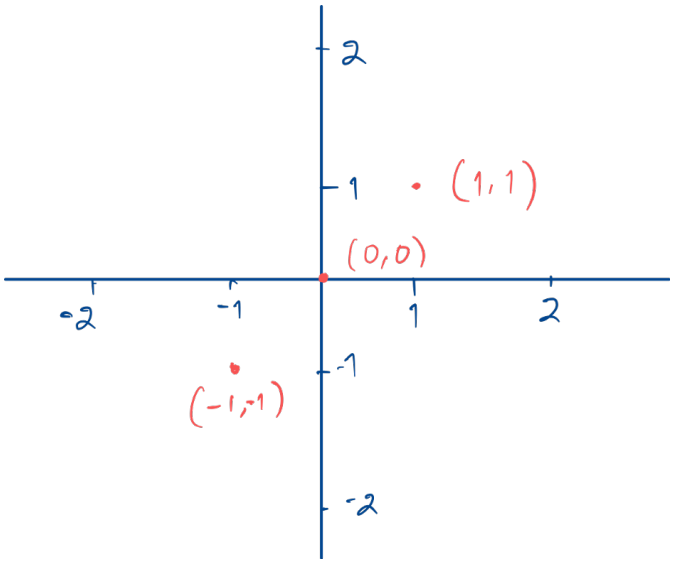


Figure 2: Data points

the following principal component.

$$\left[\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right]^T$$

The first principal component will be as follows.

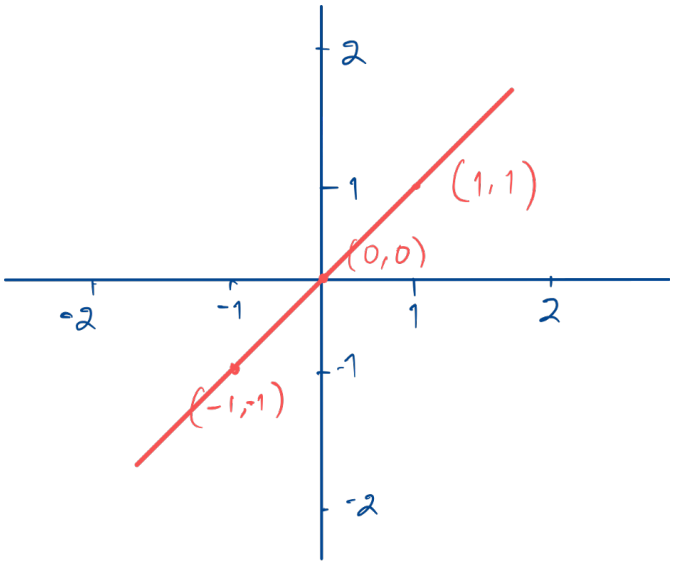


Figure 3: Principal component

If we use the first principal component to transform the data into 1-d space , the new data is $(\sqrt{2}), (0), (\sqrt{-2})$. And when we plot the new transformed points, our plot will look like this.

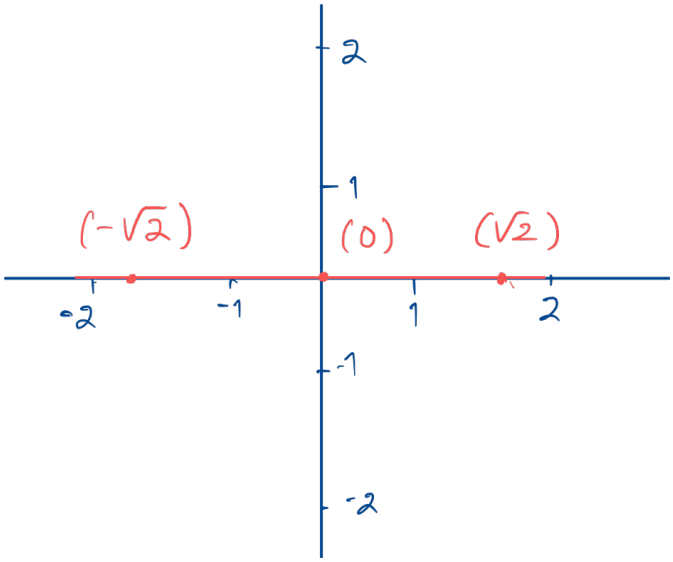
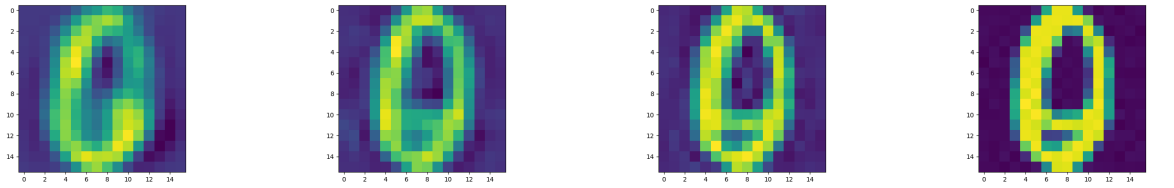


Figure 4: Transformed data

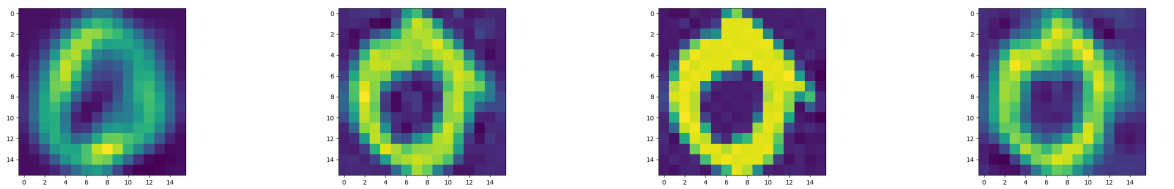
4

```
Reconstruction error for p = 10 is 394.1465
Reconstruction error for p = 50 is 202.5460
Reconstruction error for p = 100 is 119.5233
Reconstruction error for p = 200 is 37.0326
```

The resulting images are presented as follows.



(a) Image 1 10 PCs (b) Image 1 50 PCs (c) Image 1 100 PCs (d) Image 1 200 PCs



(e) Image 2 10 PCs (f) Image 2 50 PCs (g) Image 2 100 PCs (h) Image 2 200 PCs

Figure 5: Reconstructed images.

According to the results obtained, as the number of principal components used increases, the quality of the images also increases. On the other hand, when the number of principal components increase, the

error between the original image and the reconstructed image decreases. The highest construction error is obtained when the lowest number of principal components is used. This shows that the reconstruction error decreases as the number of principal components increase.