Assignment #5

Tuesday, November 7, 2023 9:38 AM

11813184

Data Mining

Assignment Five — **10 points**

Due: Thursday, November 16 @ 11:50 PM

Question 1 (3 points). Use the similarity matrix in the table below to perform complete link hierarchical clustering ***by hand (do not use Python)***. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.

	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

Question 2 (4 points). In this problem, you will perform K-means clustering ***by hand (do not use Python)***, with K = 2, on a small example with 6 observations and 2 features. The observations are as follows.

OBSERVATION	X1	X2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

(a) Plot the observations by hand.
(b) Arbitrarily partition the observations into 2 groups. Report the cluster label for each observation.
(c) Compute the centroid for each cluster.
(d) Assign each observation to the centroid to which it is closest, in terms of Euclidean distance. Report the cluster label for each observation.
(e) Repeat (c) and (d) until the answers obtained stop changing.
(f) In your plot from (a), color the observations according to the cluster labels obtained.

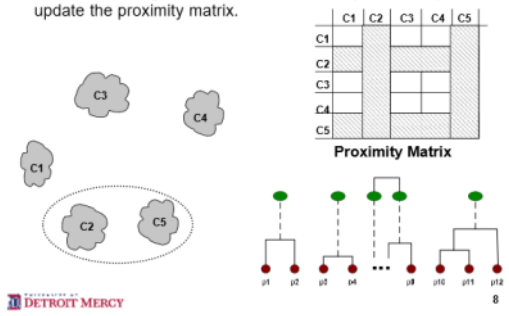
Question 3 (3 points). Apply K-means clustering on the Iris dataset using Python and visualize cluster assignments from K = 2 to K = 5 clusters. What value of K seems the best?
You can include the following statements to load the iris data:

```
from sklearn import datasets
iris = datasets.load_iris()
```

Good luck!

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

- We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.



Question 1

To complete link hierarchical clustering, we should consider each point in the similarity matrix as the initial centroids. Each stage where there are similar clusters (the closest), we merged them. We form the K clusters by assigning all the points to the closest centroid, after which, we would recalculate the centroid of each cluster, and keep doing this until the centroids do not change. We would dothis iteratively until all points in the matrix are accounted for fall into a single cluster.

1) For this question, the initial step would be to look at the proximity matrix and find the most correlated clusters with highest values such as 0.98 from p2 x p5. We would merge them as cluster "p2 union p5", which can just be written as "25".

2) After we do that, we calculate the similarity of each point respective to that "25" cluster. It depends on what techniquewe want to use for inter-cluster similarity, but for this example, we can use MIN, which calculates the minimum distance between any pair of data points, one from each of the two clusters.

These two steps would be repeated. We find the similarity of a point with respect to the cluster we created. See the matrix below.

	p25	p1	p3	p4
p25	1	min(0.10, 0.35) = 0.1	min(0.64, 0.85) = 0.64	min(0.47, 0.76) = 0.47
p1	min(0.10, 0.35) = 0.10	1	0.41	0.55
p3	**min(0.64, 0.85) = 0.64**	0.41	1	0.44
p4	min(0.47, 0.76) = 0.47	0.55	0.44	1

The max value we have in this matrix is 0.64. From that, we can say that p3 is clustered with original cluster "25" to now be"235". We repeat the process again with "235".

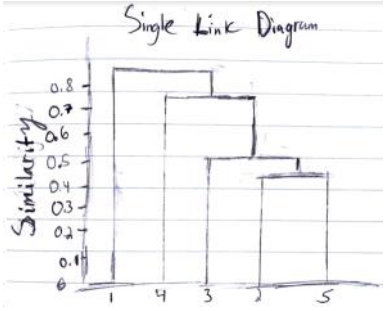
	p235	p1	p4
p235	1	min(0.41, 0.10) = 0.10	min(0.44, 0.47) = 0.44
p1	min(0.41, 0.10) = 0.10	1	**0.55**

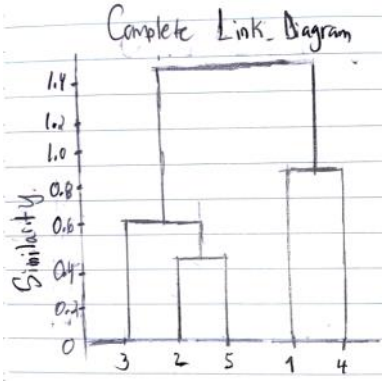
p4	min(0.44, 0.47) = 0.44	0.55	1

The max value we have in this matrix is 0.55 from p1 x p4. From that, we can say that they are clustered as "14".

	p235	p14
p235	1	0.10
p14	0.10	1

Correlation minimum is 0.10. See below for the single link and the complete link dendrogram.
* Note: Please ignore title heading mistake of the graphs, replacing "diagram" with "dendrogram".

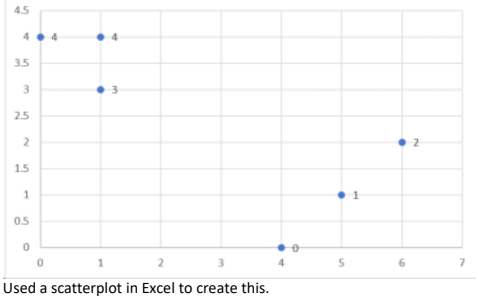




Question 2
Consider K = 2 initial clusters, see below.

Part A:

Observation	X1	X2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0



Used a scatterplot in Excel to create this.

Part B:
Let's arbitrarily assign Cluster 1: {1, 2, 3} and Cluster 2: {4, 5, 6}.

Part C:

Cluster	Centroid	Centroid
	x1	x2
1	((1+1+0)/3) = 0.6667	((4+3+4)/3) = 3.6667
2	((5+6+4)/3) = 5	((1+2+0)/3) = 1

Cluster 1 centroid: (x1, x2)	Cluster 2 centroid: (x1, x2).
Cluster 1 centroid: (0.67, 3.67)	Cluster 2 centroid: (5, 1).

Part D:

Observation	Cluster 1	Cluster 2
1	SQRT((3.67-4)^2+(0.67-1)^2) = 0.4666	SQRT((1-4)^2+(5-1)^2) = 5
2	SQRT((3.67-3)^2+(0.67-1)^2) = 0.7468	SQRT((1-3)^2+(5-1)^2) = 4.4721
3	SQRT((3.67-4)^2+(0.67-0)^2) = 0.7468	SQRT((1-4)^2+(5-0)^2) = 5.8309

4	SQRT((3.67-1)^2+(0.67-5)^2) = 5.0870	SQRT((1-1)^2+(5-5)^2) = 0
5	SQRT((3.67-2)^2+(0.67-6)^2) = 5.5854	SQRT((1-2)^2+(5-6)^2) = 1.4142
6	SQRT((3.67-0)^2+(0.67-4)^2) = 4.9555	SQRT((1-0)^2+(5-4)^2) = 1.4142

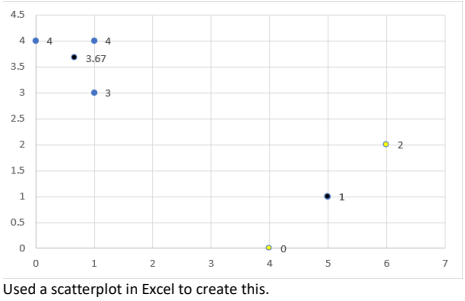
Observation 1 is closer to Cluster 1.
Observation 2 is closer to Cluster 1.
Observation 3 is closer to Cluster 1.
Observation 4 is closer to Cluster 2.
Observation 5 is closer to Cluster 2.
Observation 6 is closer to Cluster 2.

No need to update the clusters.

Part E:
Because each observation is mapped to the appropriate cluster, and that there are no additional observations for any of theseclusters to reallocate to, the answers stop changing. Fortunately, my selection of the correct observations was inferred upon through theaid of the visual scatterplot.

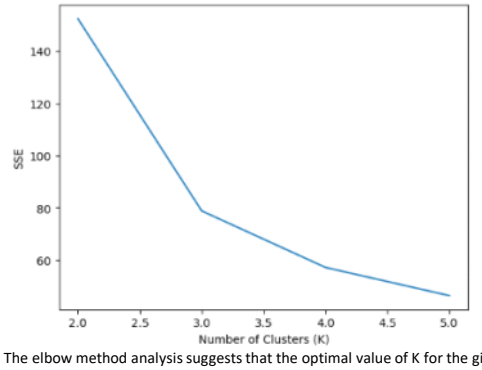
Part F:

Observation	X1	X2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0
7	0.67	3.67
8	5	1



Used a scatterplot in Excel to create this.
* Note: Not able to see Observation 4 due to centroid being right on top of it.

Question 3



The elbow method analysis suggests that the optimal value of K for the given dataset is K=3. The "elbow point" in the sum ofsquared distances (SSE) plot indicates that adding more clusters beyond K=3 does not significantly reduce the SSE. Therefore, K=3 provides a reasonable balance between capturing the underlying structure in the data and avoiding overcomplexity, making it the best choice in this context.