Assignment_05

Friday, October 1, 2021 3:54 PM

Please see next page

Homework Assignment 5 [30 points]

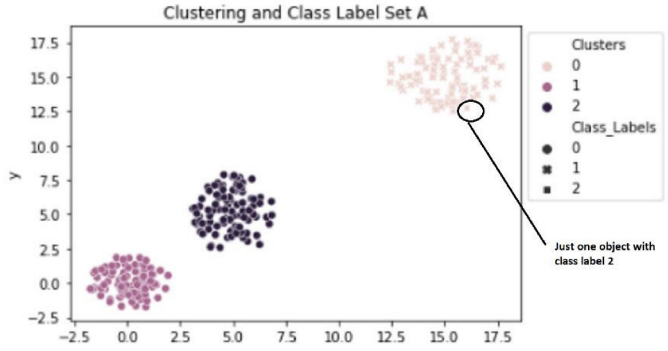
STAT430 Unsupervised Learning - Fall 2021

Due: Friday, October 1 on Compass at 11:59pm CST.

Questions #1-#5: Answer the questions in the jupyter notebook.

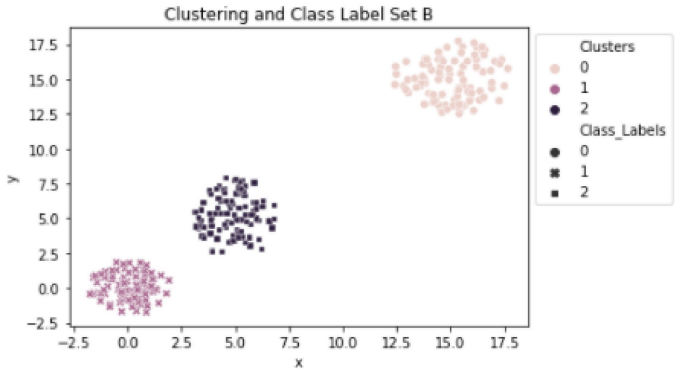
Question #6: [1.5 pts]

The three plots below display three sets of clustering labels (shown in the colors) and the class labels (shown by marker type) for the same dataset. For each of these sets of clustering labels and class labels we calculate the completeness score and the homogeneity score. Match the plots to the scores. No explanation required, but they may help if you are wrong.

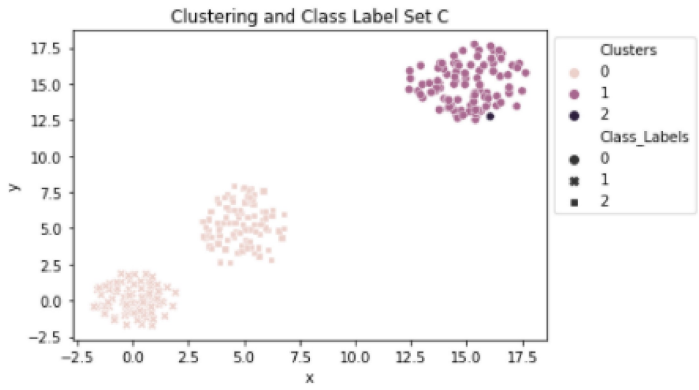


Score Set 1	Completeness Score = 0.97
	Homogeneity Score = 0.58
Score Set 2	Completeness Score=1
	Homogeneity Score = 1
Score Set 3	Completeness Score = 0.58
	Homogeneity Score = 0.97

Set A : Score Set 3



Set B : Score Set 2



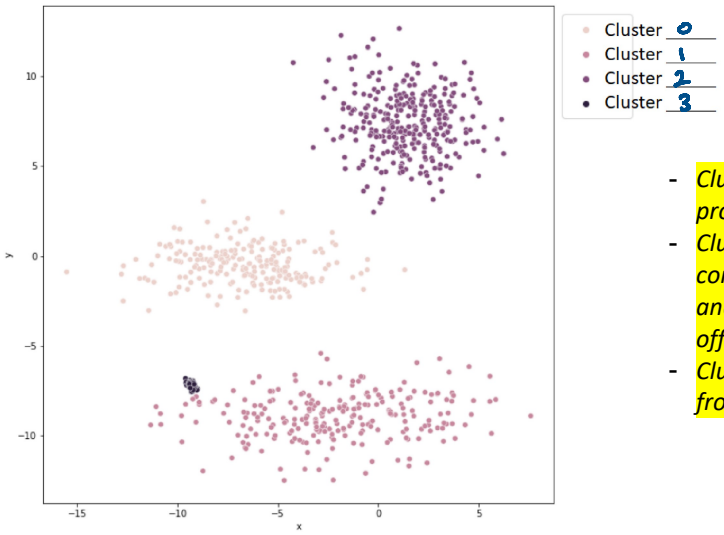
Set C : Score Set 1

Question #7: [2 pts]

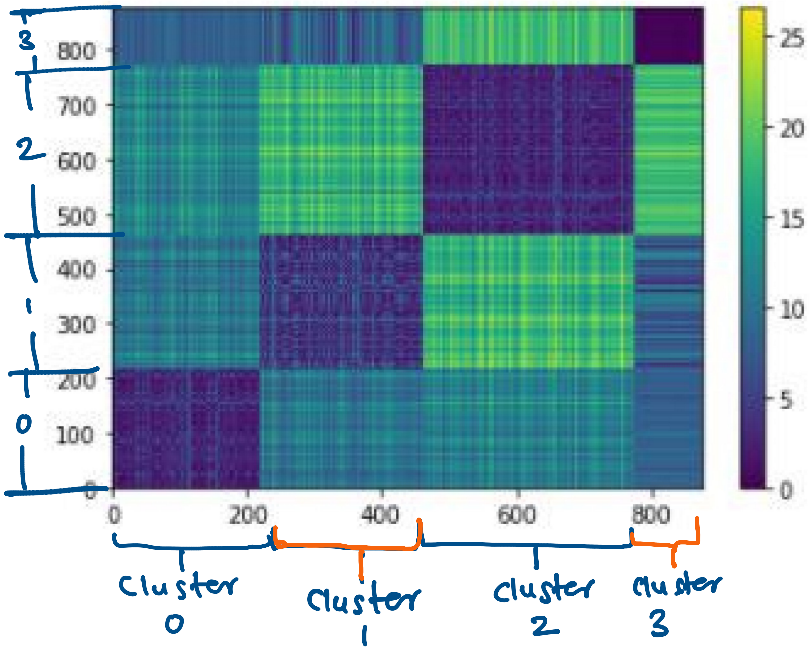
The dataset represented in the scatterplot below has been clustered into four clusters (cluster 0, cluster 1, cluster 2, and cluster 3). These respective cluster labels should be given in the legend. The cluster sorted pairwise similarity matrix for this dataset and clustering is also given below. No explanation required, but they may help if you are wrong.

- Cluster 0 contains objects [0, 218],
- Cluster 1 contains objects [219, 462]
- Cluster 2 contains objects [463, 744], and
- Cluster 3 contains objects [745,875]

Use this information to fill in the blanks in the legend in the scatterplot. Explain your answers.



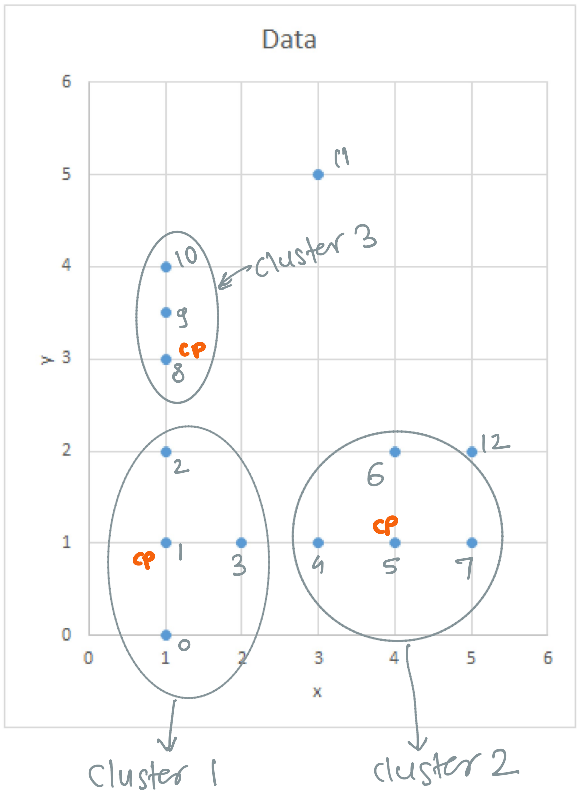
- *Cluster 3 is dense and closely packed hence low proximity*
- *Cluster 3 and cluster 0 are very close to each other compared to others, hence their separation is low and can be seen by darker color when we see their off diagonal intersection*
- *Cluster 2 and cluster 3 are very far away, can be seen from their off diagonal intersection*



Question #8: [2.5 pts]

A dataset with 13 objects is shown in the table and the scatterplot below. Select the values of ϵ and $minpts$ in the DBSCAN algorithm that will yield the following desired cluster and noise point assignments shown in the table below. No explanation required, but they may help if you are wrong. *Hint: In the presence of border point ties, the border point can be assigned arbitrarily to either core point.*

	Data		
	x	**y**	**Desired Assignment**
Object 0	1	0	Cluster 1
Object 1	1	1	Cluster 1
Object 2	1	2	Cluster 1
Object 3	2	1	Cluster 1
Object 4	3	1	Cluster 2
Object 5	4	1	Cluster 2
Object 6	4	2	Cluster 2
Object 7	5	1	Cluster 2
Object 8	1	3	Cluster 3
Object 9	1	3.5	Cluster 3
Object 10	1	4	Cluster 3
Object 11	3	5	Noise
Object 12	5	2	Noise



Min. points = 4

$\epsilon = 1$

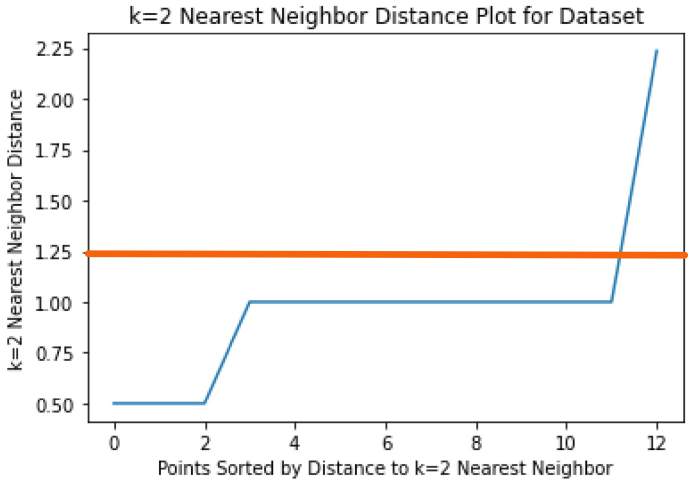
Object 2 is arbitrarily assigned to cluster 3. Here, it could have belonged to cluster 1 as well.

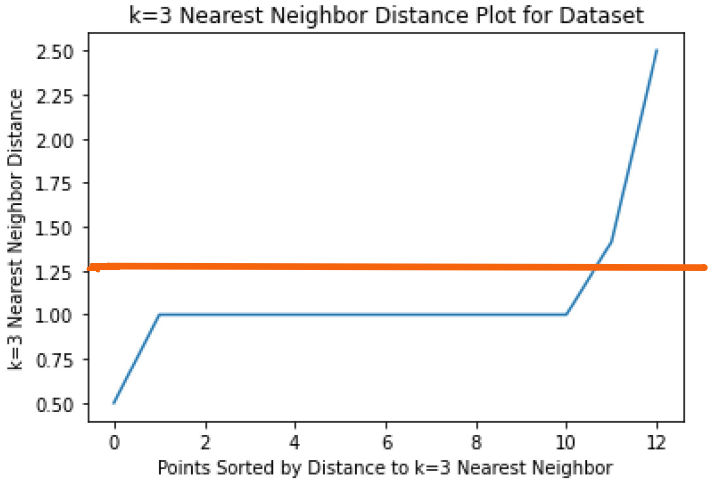
Question #9: [2.5 pts]

For another given dataset with 13 objects, we create two plots.

- The first plot represents the distance each object is to it's nearest neighbor (k=2) in the dataset. These distances have been sorted from lowest to highest.
- The second plot represents the distance each object is to it's 2nd nearest neighbor(k=3) in the dataset. These distances have been sorted from lowest to highest.

Suppose we were to cluster this dataset with DBSCAN and *minpts=3* and $\epsilon = 1.25$.





1. How many core points would this DBSCAN clustering have? Explain.

   Hard to tell, since any points with k = 3 nearest neighbor values less than or equal to epsilon could be either noise point, border point, or a core point.

2. Will this clustering have any noise points? Explain.

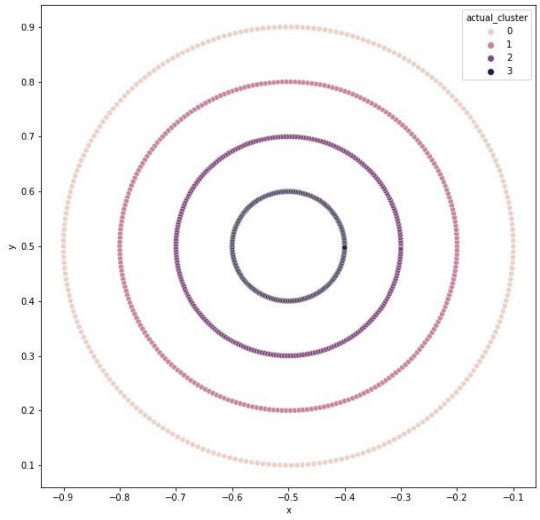
   Yes, This clustering will have two noise points, since object no. 11 and 12's nearest point when k = 3 is more than the epsilon value which is 1.25 here.

Question #10 [2 pt]: True or False

Circle whether the following statements below are true or false.

1. **True or False**: A silhouette plot and the average silhouette score (using the Euclidean distance) would be a useful plot and metric for evaluating the cohesion and separation of clusters shown in the plot below.

2. **True or False**: The cluster-sorted similarity matrix of the clustering above (using the Euclidean distance) would be a useful visualization for evaluating the cohesion and separation of clusters shown in the plot below.



3. **True or False**: We can use a t-SNE plot to determine the shapes of clusters in a high dimensional dataset.

   We can approximate cluster shapes

4. **True or False**: We can use a t-SNE plot to determine the presence of outliers in a high dimensional dataset.