CS6350 Fall 2021
Big Data Management Analytics and Management

Assignment 4

Due Date: Nov 30th, 2021 (by 11:59 a.m.)

Part #1
Objective:

This assignment is for you to learn about clustering and recommendation system, particularly about different techniques of clustering.

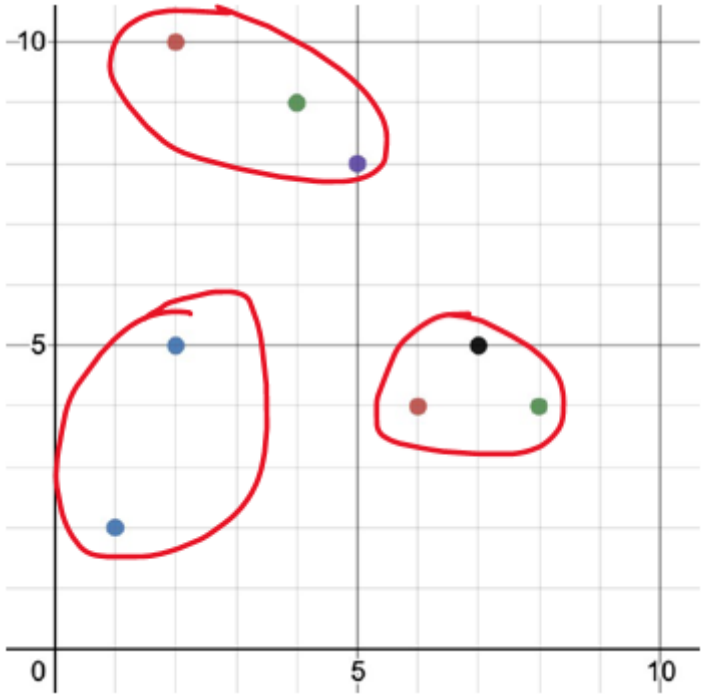
Please solve the following problems. No computer programming is required to solve the problems.

Problem Statement:

1. K-Means algorithm:

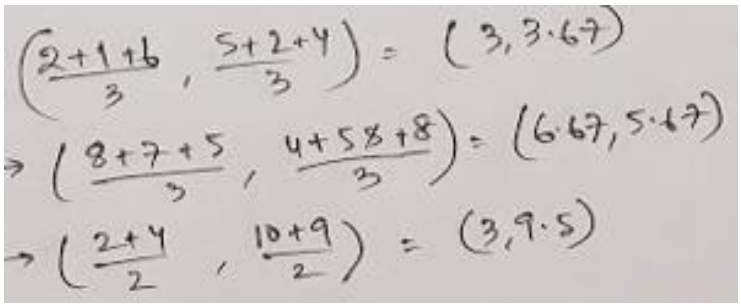
Consider the following eight points in a 2-dimensional space: {(2, 10); (2, 5); (8, 4); (5, 8); (7, 5); (6, 4); (1, 2); (4, 9)}. Use the Euclidean distance metric to measure the separation of the points.

a. Plot the data points and group them into appropriate clusters. How many clusters are required and what are contents of each of these clusters?



b. Now consider we want to divide the points into 3 initial clusters (C1, C2, C3) with centers defined as {(2, 5), (5, 8), (4, 9)} respectively.

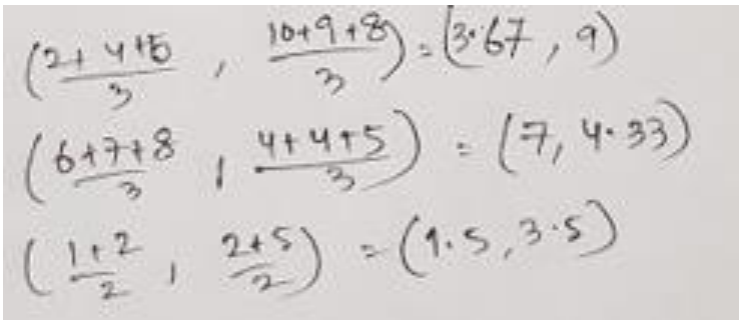
c. What's the center of the cluster after one iteration?



3 each

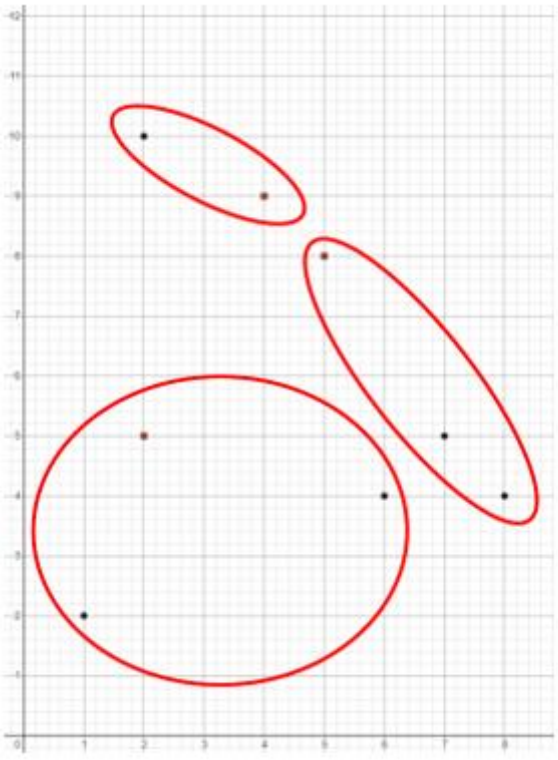
d. What's the center of the cluster after one 2nd iteration?
Centers are same as part (e)

e. What's the center of the cluster after one 3rd iteration?



f. Compare the results of each iteration with your answers in part (a).

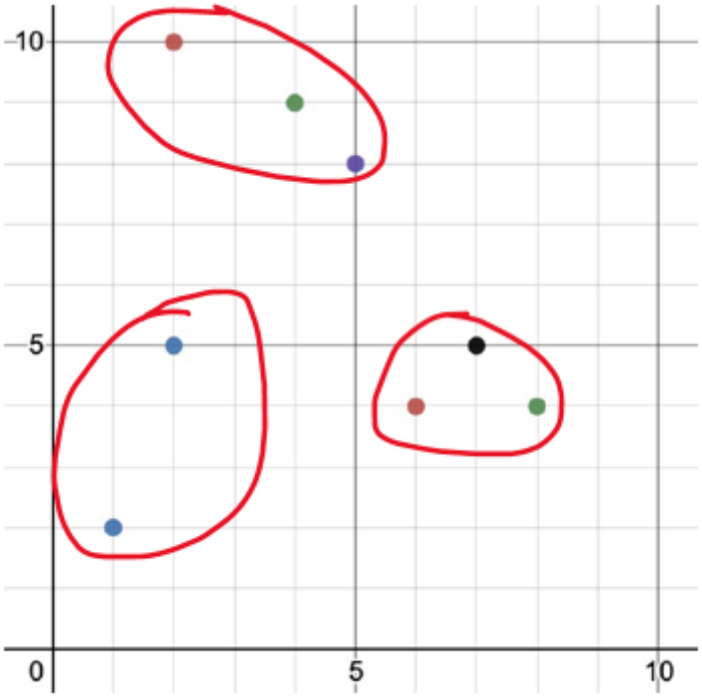
- 1st iteration:



- 2nd and 3rd iteration: the clusters are same as part (a)

g. How many iterations are required for the clusters to converge?
2/3

h. What are the resulting centers and resulting clusters (K=3)? Plot the final data points.

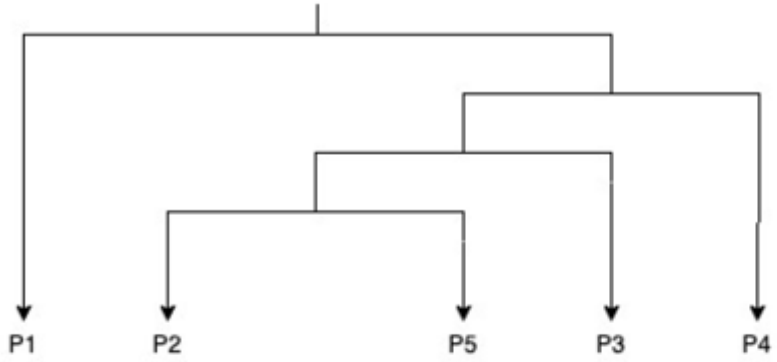


2. Hierarchical algorithm:

Use the similarity matrix in the table below to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged. (with enough explanation and calculation)

20

	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



Single link

	P1	P2	P3	P4	P5
P1	1	0.1	0.41	0.55	0.35
P2	0.1	1	0.64	0.47	0.98
P3	0.41	0.64	1	0.44	0.85
P4	0.55	0.47	0.44	1	0.76
P5	0.35	0.98	0.85	0.76	1

	P1	P2P5	P3	P4
P1	1	0.35	0.41	0.55
P2P5	0.35	1	0.85	0.76
P3	0.41	0.85	1	0.44
P4	0.55	0.76	0.44	1

	P1	P2P5P3	P4
P1	1	0.41	0.55
P2P5P3	0.41	1	0.76

P4	0.55	0.76	1

	P1	P2P5P3
P1	1	0.55
P2P5P3P4	0.55	1

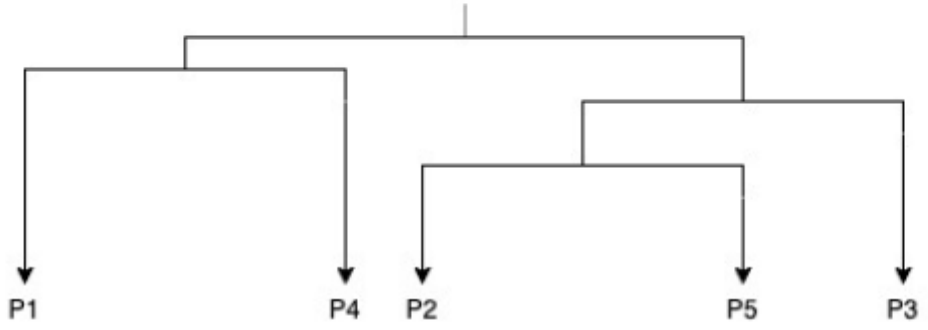
Complete Link

	P1	P2	P3	P4	P5
P1	1	0.1	0.41	0.55	0.35
P2	0.1	1	0.64	0.47	0.98
P3	0.41	0.64	1	0.44	0.85
P4	0.55	0.47	0.44	1	0.76
P5	0.35	0.98	0.85	0.76	1

	P1	P2P5	P3	P4
P1	1	0.1	0.41	0.55
P2P5	0.1	1	0.64	0.47
P3	0.41	0.64	1	0.44
P4	0.55	0.47	0.44	1

	P1	P2P5P3	P4
P1	1	0.1	0.55
P2P5P3	0.1	1	0.44
P4	0.55	0.44	1

	P1P4	P2P5P3
P1P4	1	0.1
P2P5P3	0.1	1



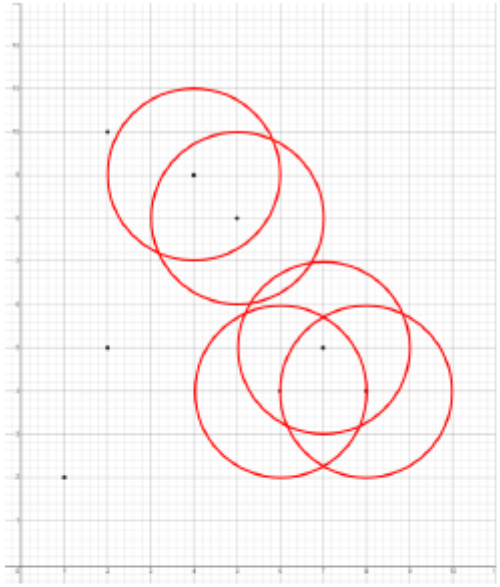
3. DBSCAN algorithm

Consider the following eight point in a 2-dimensional space: {(2, 10); (2, 5); (8, 4); (5, 8); (7, 5); (6, 4); (1, 2); (4, 9)}. Suppose we use the Euclidean distance metric.

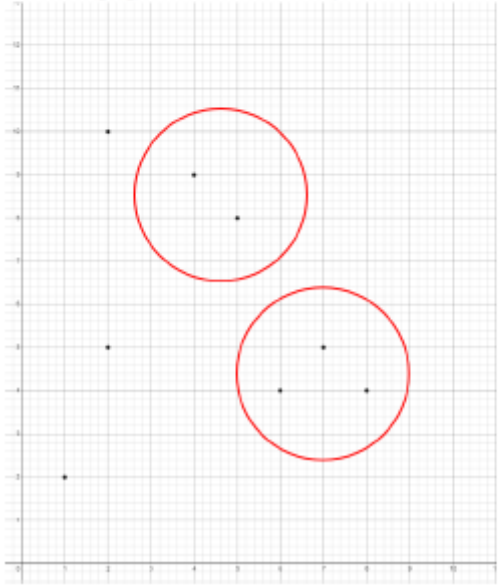
12

a. If Epsilon is 2 and min_samples is 2, what are the clusters that DBSCAN would discover. Plot the discovered clusters.

The initial sets of clusters appear as follows:

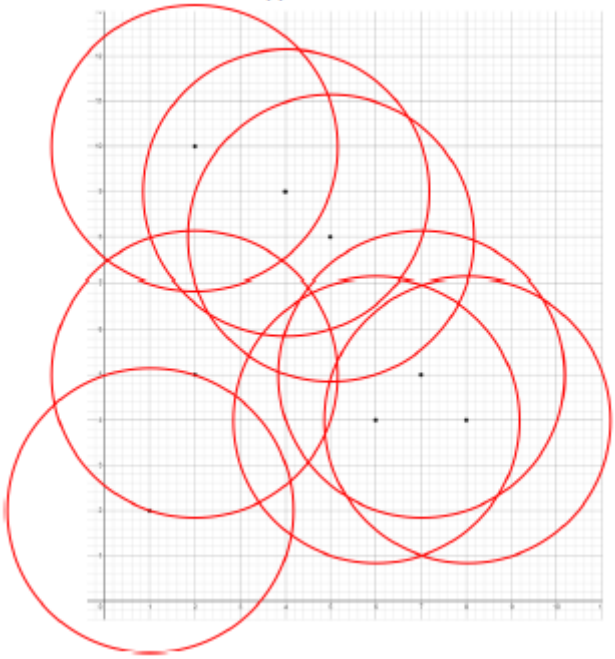


After merging clusters with density reachable clusters we have:

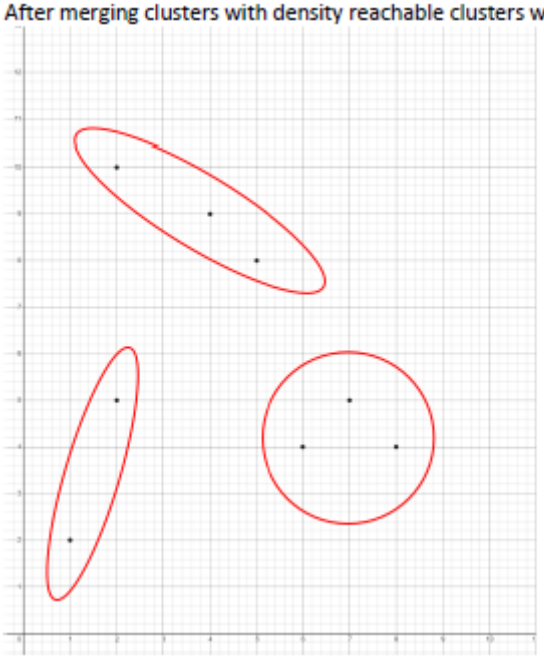


b. What if Epsilon is increased to $\sqrt{10}$?

The initial sets of clusters appear as follows:



After merging clusters with density reachable clusters we have:



3

4. Explain the shortcomings of BFR algorithm and describe how CURE algorithm overcomes the shortcomings.

Part #2: BigTable and Cassandra

Q1. Compare BigTable with Cassandra.
Q3. Explain the concept of tunable consistency in Cassandra.
Q4. Define memtable.
Q5. What is SSTable? How is it different from other relational tables?
Q6. Explain CAP theorem.
Q7. Describe difference between Tablet Server and Tablets.

4 each

Part #3: Recommendation Systems

Use Collaborative filtering to find the accuracy of ALS model. Use ratings.dat file. It contains: User id :: movie id :: ratings :: timestamp.
Your program should report the accuracy of the model. For details follow the link: https://spark.apache.org/docs/latest/mllib-collaborative-filtering.html. Please use 60% of the data for training and 40% for testing and report the MSE of the model. Submit the code along with the output of your code.

20