**CS6350 Spring 2021**

**Big Data Management Analytics and Management**

**Assignment 4**

**Due Date: May 3, 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.

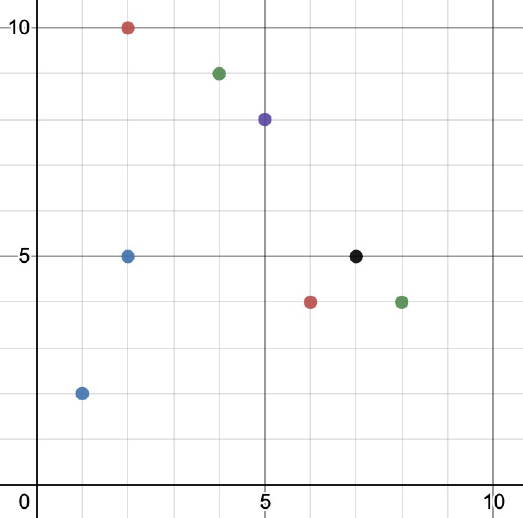
Please solve the following problems. No computer programming is required to solve the problems.

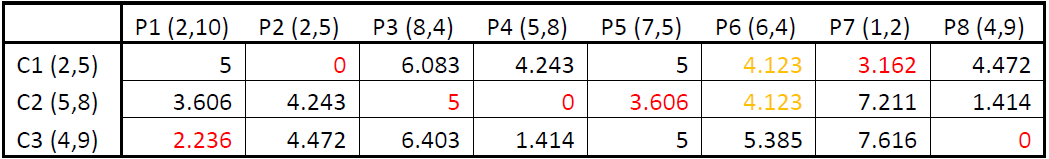
***Problem Statement:***

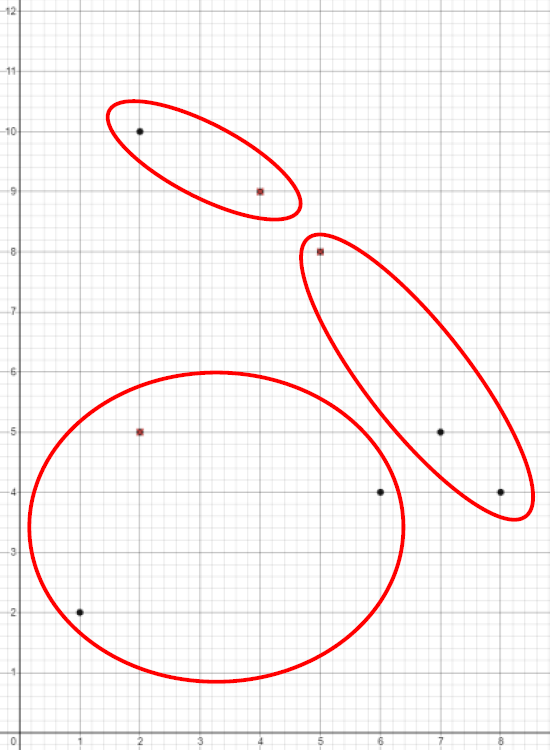
1. K-Means algorithm:

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)}. Suppose we plan to use the Euclidean distance metric, and we are interested in clustering these points into 3 clusters.

1. Plot the data points to see what appropriate clusters might be;







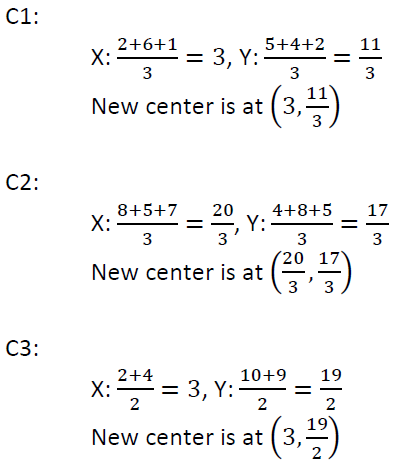
C1: P2, P6, P7

C2: P3, P4, P5

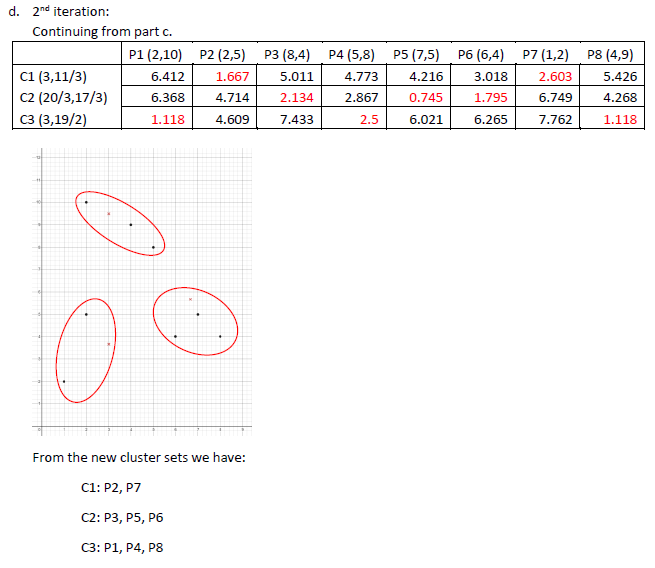
C3: P1, P8

Note that P6 has equal distance to C1 or C2 so we randomly chose it to be included in C1.

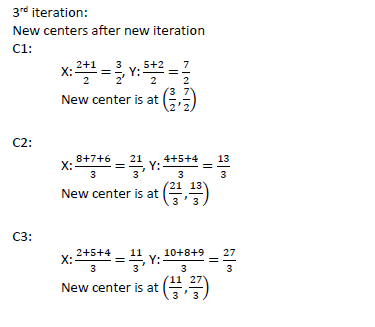
1. Form three initial clusters with points {(2, 5), (5, 8), (4, 9)} as initial cluster centers;

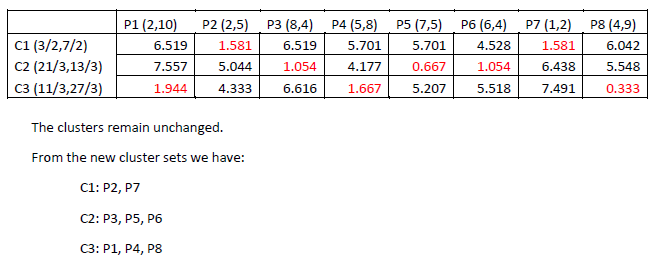


e-



f-

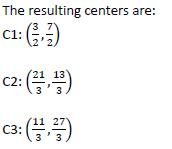




How many iterations are required for the clusters to converge?



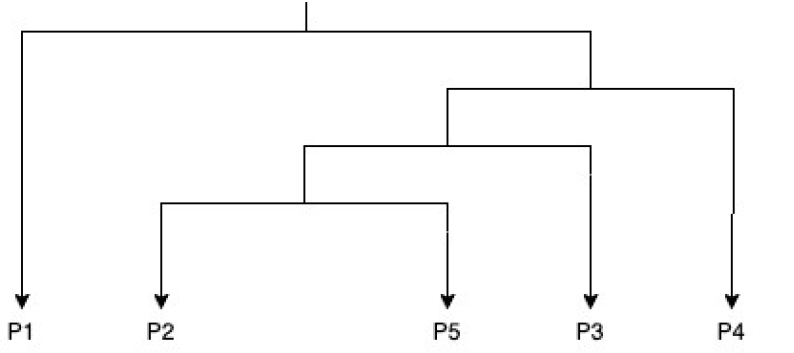
1. What are the resulting centers and resulting clusters? (K=3)

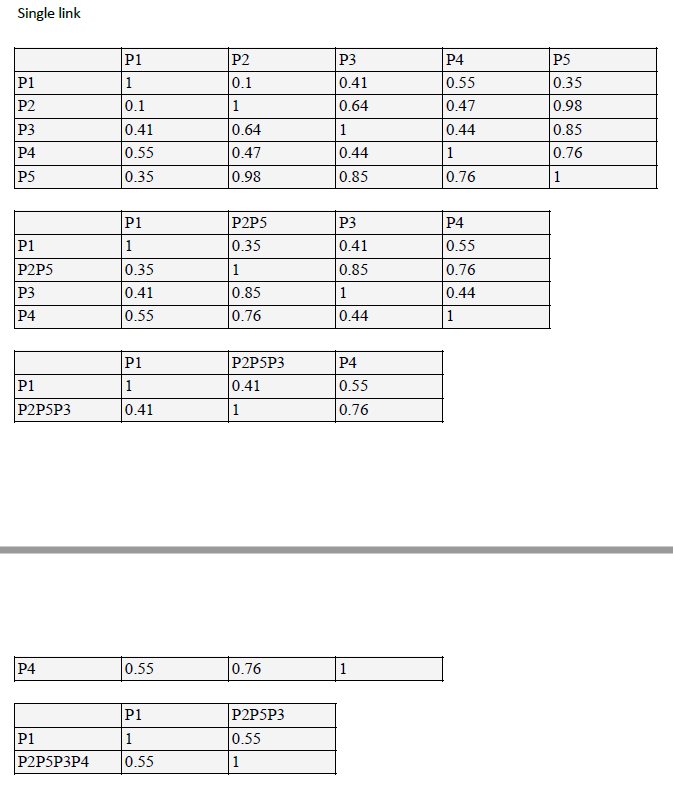


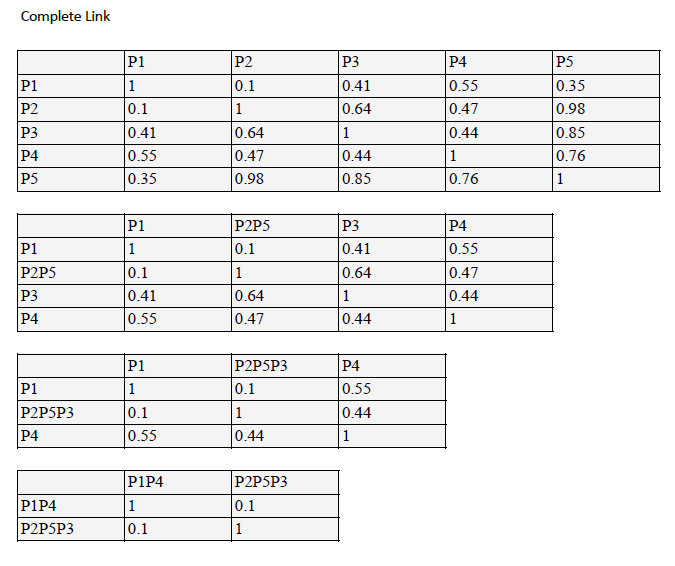
2. Hierarchical algorithm:

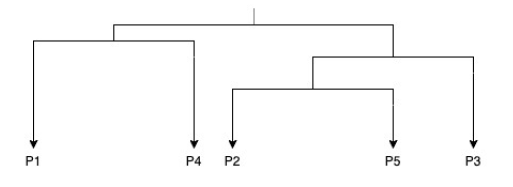
Use the similarity matrix in Table 1 to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 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 |





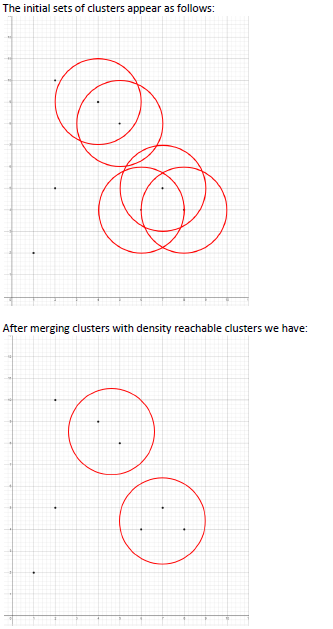




3. DBSCAN algorithm

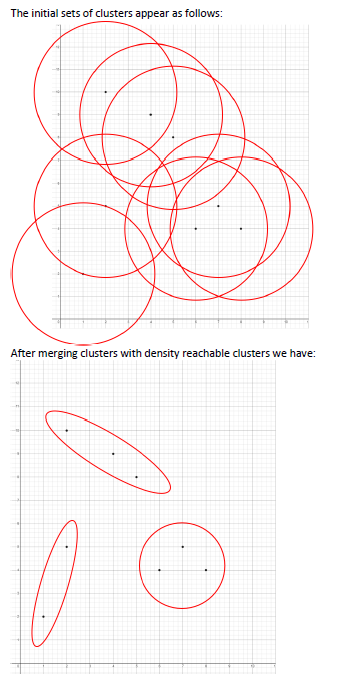
Consider the following eight point in a 2-dimensional space: {(2, 10); (2, 5); (8, 4); (5, 8); (7, 5); (6, 4); (1, 2); (4, 9)}. Suppose we use the Euclidean distance metric.

1. If Epsilon is 2 and min\_samples is 2, what are the clusters that DBSCAN would discover. Plot the discovered clusters.



We can circle the noises too, but usually will remove.

1. What if Epsilon is increased to ?



**Part #2: BigTabe and Cassandra**

Q1. Compare BigTable with Cassandra.

Q2. What is Cassandra?

Q3. Explain the concept of tunable consistency in Cassandra.

Q4. Define memtable.

Q5. What is SSTable? How is it different from other relational table.

Q6. Explain CAP theorem.

Q7. Describe difference between Tablet Server and Tablets.

**Find from Lectures**

**Part 3: Programming**

**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.

**The code depends on each programmer.**