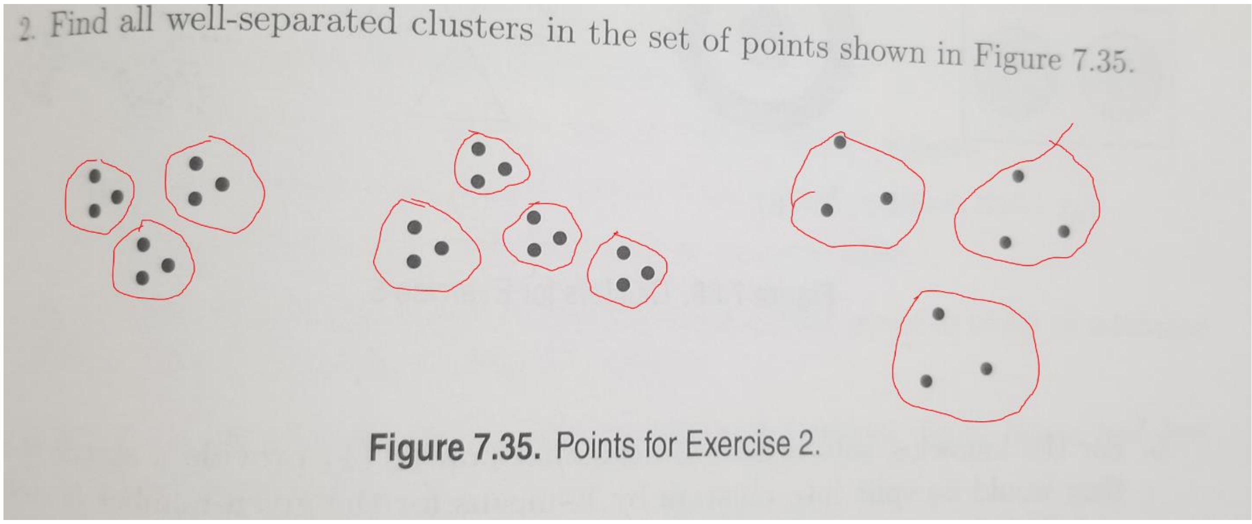
CS 422 HW 3

Matt Venter

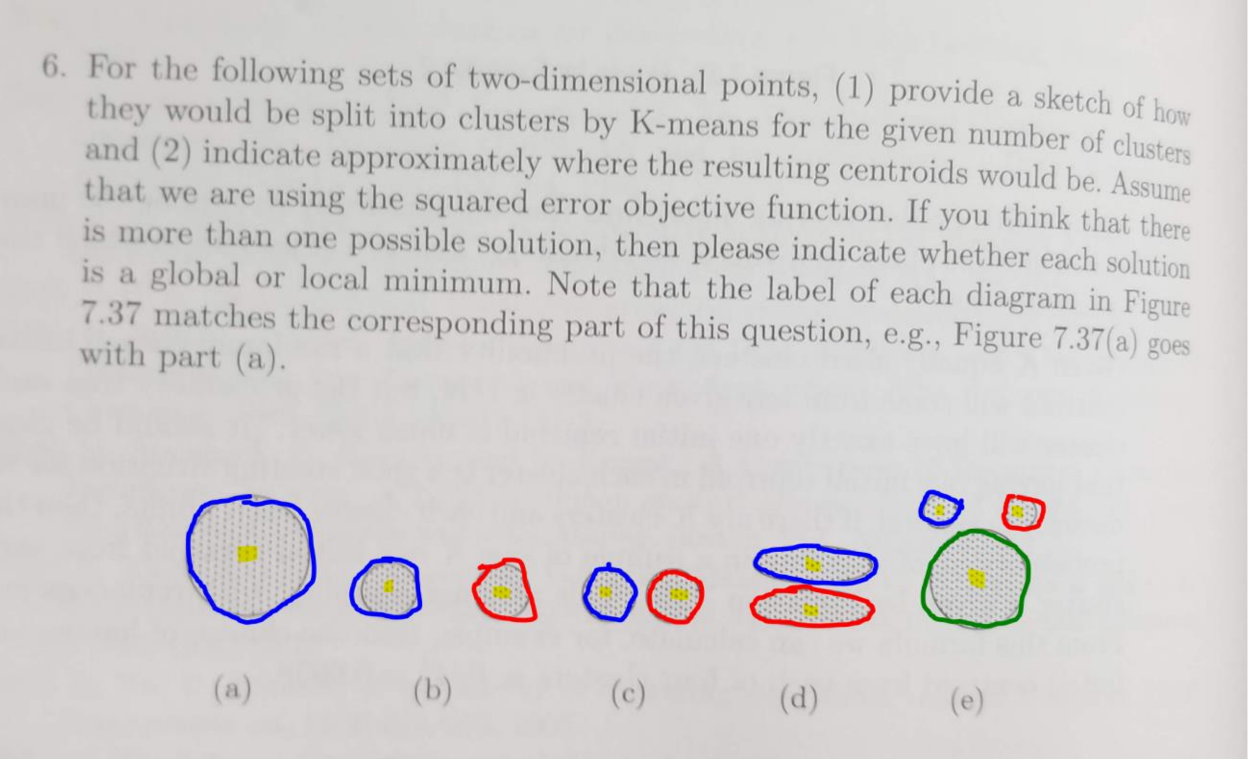
Question 2:



2. Find all well-separated clusters in the set of points shown in Figure 7.35.

Figure 7.35. Points for Exercise 2.

Question 6:

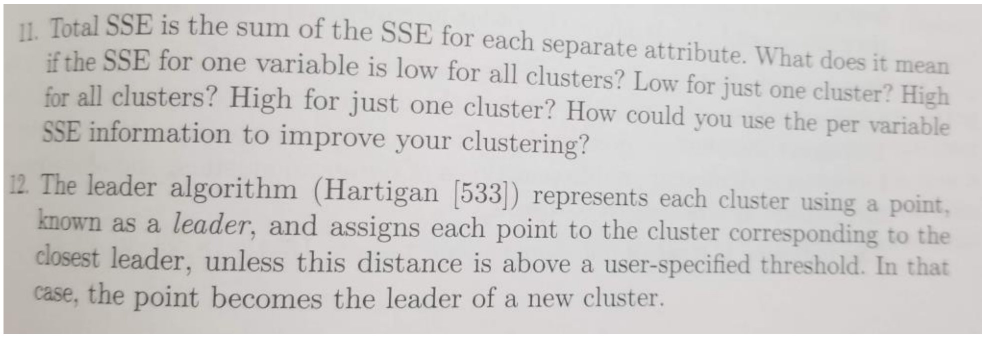


6. For the following sets of two-dimensional points, (1) provide a sketch of how they would be split into clusters by K-means for the given number of clusters and (2) indicate approximately where the resulting centroids would be. Assume that we are using the squared error objective function. If you think that there is more than one possible solution, then please indicate whether each solution is a global or local minimum. Note that the label of each diagram in Figure 7.37 matches the corresponding part of this question, e.g., Figure 7.37(a) goes with part (a).

Red/blue/green = different cluster

Yellow = centroid location for each cluster

Question 11/12:



11. Total SSE is the sum of the SSE for each separate attribute. What does it mean if the SSE for one variable is low for all clusters? Low for just one cluster? High for all clusters? High for just one cluster? How could you use the per variable SSE information to improve your clustering?

12. The leader algorithm (Hartigan [533]) represents each cluster using a point, known as a *leader*, and assigns each point to the cluster corresponding to the closest leader, unless this distance is above a user-specified threshold. In that case, the point becomes the leader of a new cluster.

11.

What does it mean if the SSE for one variable is low for all clusters?

It would mean the variable is fundamentally a constant and therefore is of little use when partitioning data into groups (can be excluded from cluster calculations)

Low for just one cluster?

That one attribute is very similer to all members of said cluster and helps define it.

High for all clusters?

The variable being mentioned is probably just noise

High for just one cluster?

Its funky and defines other clusters well so it just does not define this cluster specifically

How could you use the per variable SSE information to improve your clustering?

We start by excluding all high/low SSE attributes that affect all clusters. Then we try to home in on what defines certain clusters and work on those.

12.

a. What are the advantages and disadvantages for this algorithm as opposed to k-means clustering

Advantages include

- May be faster in some instances
- Clusters will be tighter

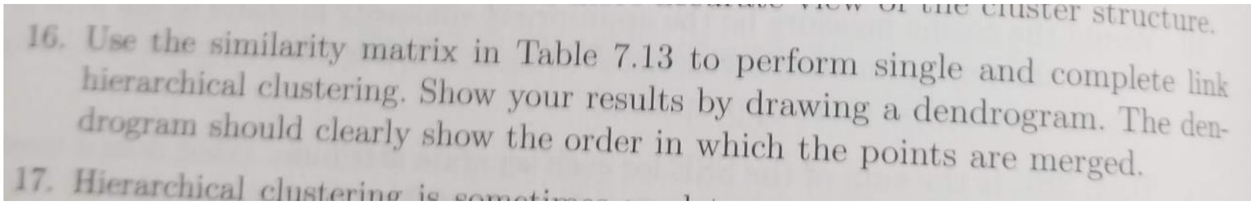
Disadvantages Include:

- Unpredictable k-values
- Will result in different clusters if it is started somewhere else
- Clusters may not be completely accurate

b. Suggest ways to improve the algorithm

Honestly, I hate the idea of this algorithm, but I would prefer it if it come closer to k-means and lets the leaders change over time if it is no longer optimal.

16.



view of the cluster structure.

16. Use the similarity matrix in Table 7.13 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.

17. Hierarchical clustering is sometimes

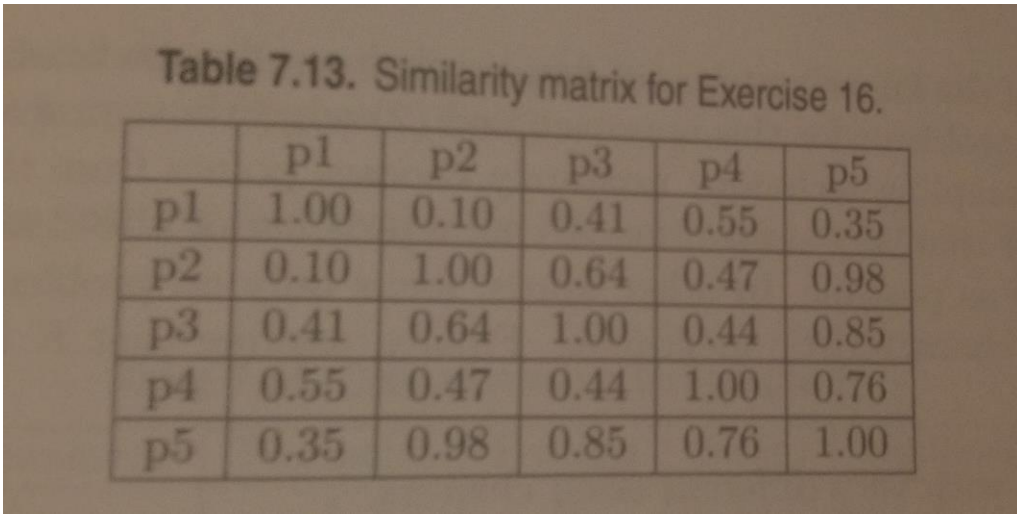
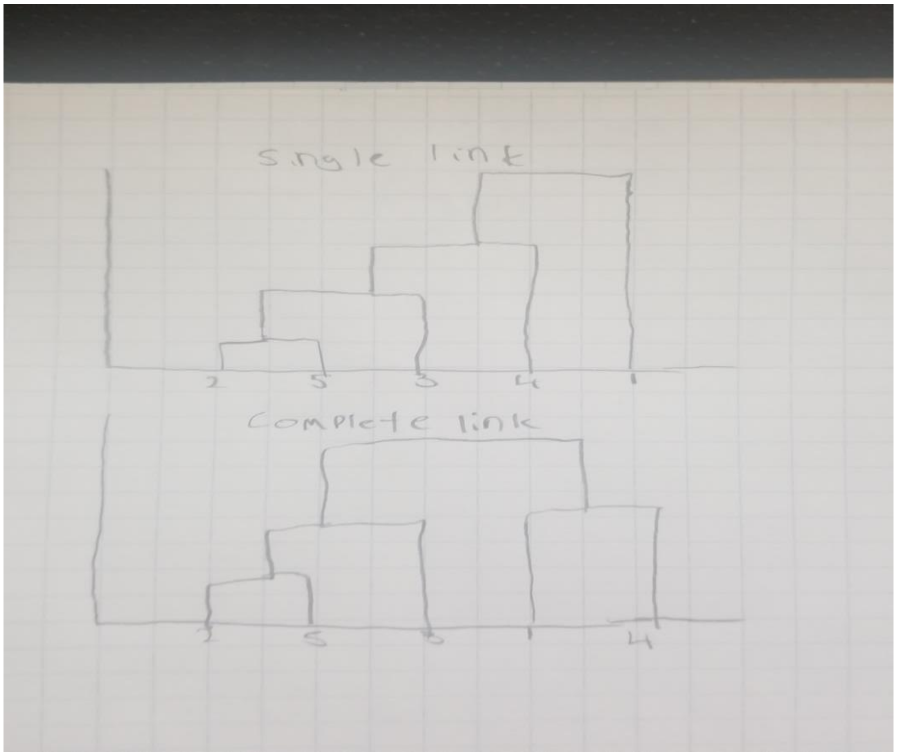


Table 7.13. Similarity matrix for Exercise 16.

	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
2
5
3
4
1
Complete link
2
5
3
1
4