Homework 2

Submitted by Sekharan Natarajan (200110398)

Q1.

**a) Why k-means clustering is sub-optimal?**

During every iteration, K means tries to minimize the SSE for a given set of centroids by assigning points to their nearest centroid. It recomputes the centroid to again minimize SSE. This only finds a local minimum with respect to the SSE and is influenced by the initial choice of centroid and not based on all possible choices. Hence for bad choices of initial centroids, it can produce suboptimal clusters (ex . Empty clusters)

**b) List 3 advantages and 3 disadvantages of k-means clustering.**

**Advantages of K means clustering algorithm:**

1. K means clustering algorithm is very simple and can be used with a variety of data types. (Its not only limited to euclidean space but can also be used with documents (doc term matrices)

2. It's simplicity leads to faster run times and hence is less expensive. (K means is usually faster than other clever algorithms)

3. K means can be used as a pre-clustering technique as it helps reduce the space into disjoint smaller subsets on which other algorithms can be applied.

**Disadvantages of K means clustering algorithm:**

1. K means cannot handle non-globular clusters properly

2. It is poor at detecting clusters with varying densities and sizes.

3. It has trouble clustering data that contains outliers and sometimes produces empty clusters.

4. It is only suitable for data where there is a notion of centroid/ center.

**c) As K means is sub-optimal, describe 3 ways to select initial centroids that may help finding good solution.**

1. Perform multiple trials : During each trial select different set of random initial centroids and then choose the trial which produces smallest SSE (Sum of Squared Error)

2. Perform Hierarchical clustering to a sample of points: In this approach K clusters are formed from the small sample of data and the centroids of these clusters are used for performing K means

3. After selecting a random first centroid, each successive centroid is chosen in such a way that they are farthest away from the initial centroid already selected.

4. Bisecting k means can solve the initialization problem of k means.

**d) Using the data in Table 1, show first 3 iterations of K-means with following initial centroids. (Show both calculations and scatter plot for each iteration)**

Iteration 1 : Centroids : m1 = (1,3), m2 = (2,3)

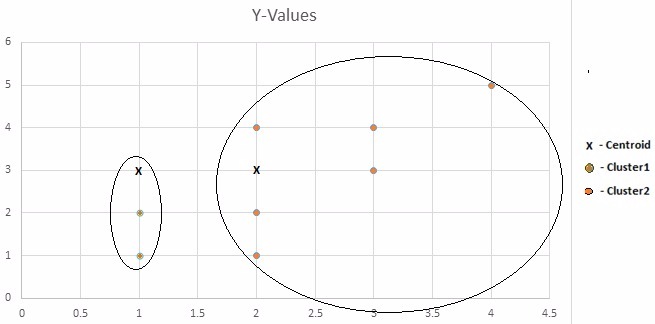
Distances are Euclidean distance with A1 and A2 belonging to X and Y axis respectively.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | A1 | A2 | Distance from m1 | Distance from m2 |
| 1 | 1 | 1 | 2.00 | 2.73 |
| 2 | 2 | 1 | 2.23 | 2.00 |
| 3 | 1 | 2 | 1.00 | 1.41 |
| 4 | 2 | 2 | 1.41 | 1.00 |
| 5 | 3 | 3 | 2.00 | 1.00 |
| 6 | 3 | 4 | 2.23 | 1.41 |
| 7 | 4 | 5 | 3.60 | 2.82 |
| 8 | 2 | 3 | 1.41 | 1.00 |

Cluster 1 has the following ids : 1,3

Cluster 2 has the following ids : 2,3,4,5,6,8

New centroid after iteration 1 : (1,1.5) and (2.67,3.17)



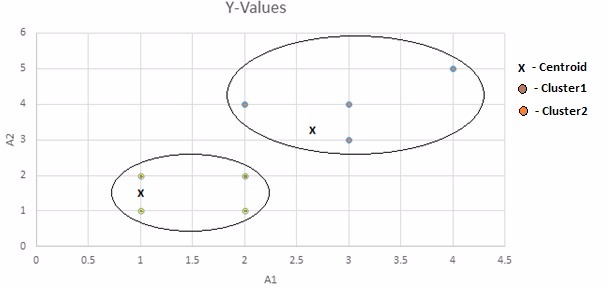
Iteration 2 : Centroids : m1 = (1,1.5), m2 = (2.67,3.17)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | A1 | A2 | Distance from m1 | Distance from m2 |
| 1 | 1 | 1 | 0.70 | 3.60 |
| 2 | 2 | 1 | 0.70 | 3.16 |
| 3 | 1 | 2 | 0.70 | 2.82 |
| 4 | 2 | 2 | 0.70 | 2.23 |
| 5 | 3 | 3 | 2.12 | 1.00 |
| 6 | 3 | 4 | 2.91 | 0.00 |
| 7 | 4 | 5 | 4.30 | 1.41 |
| 8 | 2 | 3 | 2.54 | 1.00 |

Cluster 1 has the following ids : 1,2,3,4

Cluster 2 has the following ids : 5,6,7,8

New centroid after iteration 2 : (1.5,1.5) and (3,4)



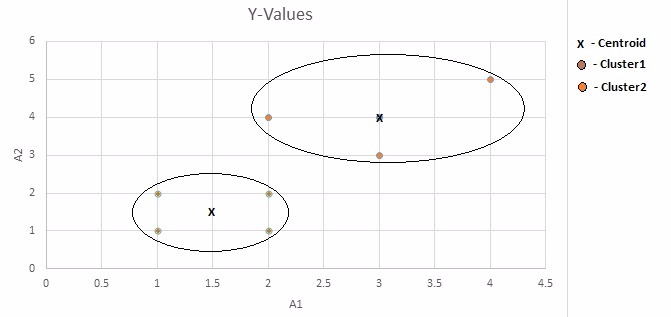
Iteration 3 : Centroids : m1 = (1.5,1.5), m2 = (3,4)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | A1 | A2 | Distance from m1 | Distance from m2 |
| 1 | 1 | 1 | 0.50 | 2.73 |
| 2 | 2 | 1 | 1.11 | 2.27 |
| 3 | 1 | 2 | 0.50 | 2.03 |
| 4 | 2 | 2 | 1.11 | 1.34 |
| 5 | 3 | 3 | 2.50 | 0.37 |
| 6 | 3 | 4 | 3.20 | 0.89 |
| 7 | 4 | 5 | 4.60 | 2.26 |
| 8 | 2 | 3 | 2.69 | 1.06 |

Cluster 1 has the following ids : 1,2,3,4

Cluster 2 has the following ids : 5,6,7,8

New centroid after iteration 2 : (1.5,1.5) and (3,4)



**e) Using the k-means cluster solution found in (d), predict label for points (3,1) and (2,3).**

|  |  |  |  |
| --- | --- | --- | --- |
| A1 | A2 | Distance from m1 (1.5,1.5) | Distance from m2 (3,4) |
| 3 | 1 | 1.58 | 3.00 |
| 2 | 3 | 1.58 | 1.41 |

It is clear form the above table that (3,1) is closer to m1 (1.5,1.5) and belongs to cluster 1

and (2,3) is closer to m2 (3,4) and belongs to cluster 2.

**Q2. a) Using the data shown in table 2, perform hierarchical clustering using min and max; show resulting distance matrix at each step.**

Given matrix:

|  |  |  |
| --- | --- | --- |
| ID | A1 | A2 |
| 1 | 1 | 1 |
| 2 | 2 | 1 |
| 3 | 2 | 3 |
| 4 | 3 | 4 |
| 5 | 5 | 3 |

Method : MIN

Step 1:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID/ID | 1 | 2 | 3 | 4 | 5 |
| 1 | 0.00 | 1.00 | 2.23 | 3.60 | 4.47 |
| 2 | 1.00 | 0.00 | 2.00 | 3.16 | 3.60 |
| 3 | 2.23 | 2.00 | 0.00 | 1.414 | 3.00 |
| 4 | 3.60 | 3.16 | 1.414 | 0.00 | 2.23 |
| 5 | 4.47 | 3.60 | 3.00 | 2.23 | 0.00 |

Points 1 and 2 are combined

Step 2:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID/ID | 1,2 | 3 | 4 | 5 |
| 1,2 | 0.00 | 2.00 | 3.16 | 3.60 |
| 3 | 2.00 | 0.00 | 1.41 | 3.00 |
| 4 | 3.16 | 1.41 | 0.00 | 2.23 |
| 5 | 3.60 | 3.00 | 2.23 | 0.00 |

Points 3 and 4 are merged

Step 3:

|  |  |  |  |
| --- | --- | --- | --- |
| ID/ID | 1,2 | 3,4 | 5 |
| 1,2 | 0.00 | 2.00 | 3.60 |
| 3,4 | 2.00 | 0.00 | 2.23 |
| 5 | 3.60 | 2.23 | 0.00 |

Clusters (1,2) and (3,4) are merged

Step 4:

|  |  |  |
| --- | --- | --- |
| ID/ID | 1,2,3,4 | 5 |
| 1,2,3,4 | 0.00 | 2.23 |
| 5 | 2.23 | 0.00 |

All the points now belong to one big cluster.

Step 5:

|  |  |
| --- | --- |
| ID/ID | 1,2,3,4,5 |
| 1,2,3,4,5 | 0.00 |

Method : MAX

Step 1:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID/ID | 1 | 2 | 3 | 4 | 5 |
| 1 | 0.00 | 1.00 | 2.23 | 3.60 | 4.47 |
| 2 | 1.00 | 0.00 | 2.00 | 3.16 | 3.60 |
| 3 | 2.23 | 2.00 | 0.00 | 1.414 | 3.00 |
| 4 | 3.60 | 3.16 | 1.414 | 0.00 | 2.23 |
| 5 | 4.47 | 3.60 | 3.00 | 2.23 | 0.00 |

Points 1 and 2 are merged.

Step 2:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID/ID | 1,2 | 3 | 4 | 5 |
| 1,2 | 0.00 | 2.23 | 3.60 | 4.47 |
| 3 | 2.23 | 0.00 | 1.41 | 3.00 |
| 4 | 3.60 | 1.41 | 0.00 | 2.23 |
| 5 | 4.47 | 3.00 | 2.23 | 0.00 |

Points 3 and 4 are merged.

Step 3:

|  |  |  |  |
| --- | --- | --- | --- |
| ID/ID | 1,2 | 3,4 | 5 |
| 1,2 | 0.00 | 3.60 | 4.47 |
| 3,4 | 3.60 | 0.00 | 3.00 |
| 5 | 4.47 | 3.00 | 0.00 |

Point 5 is added to cluster (3,4).

Step 4:

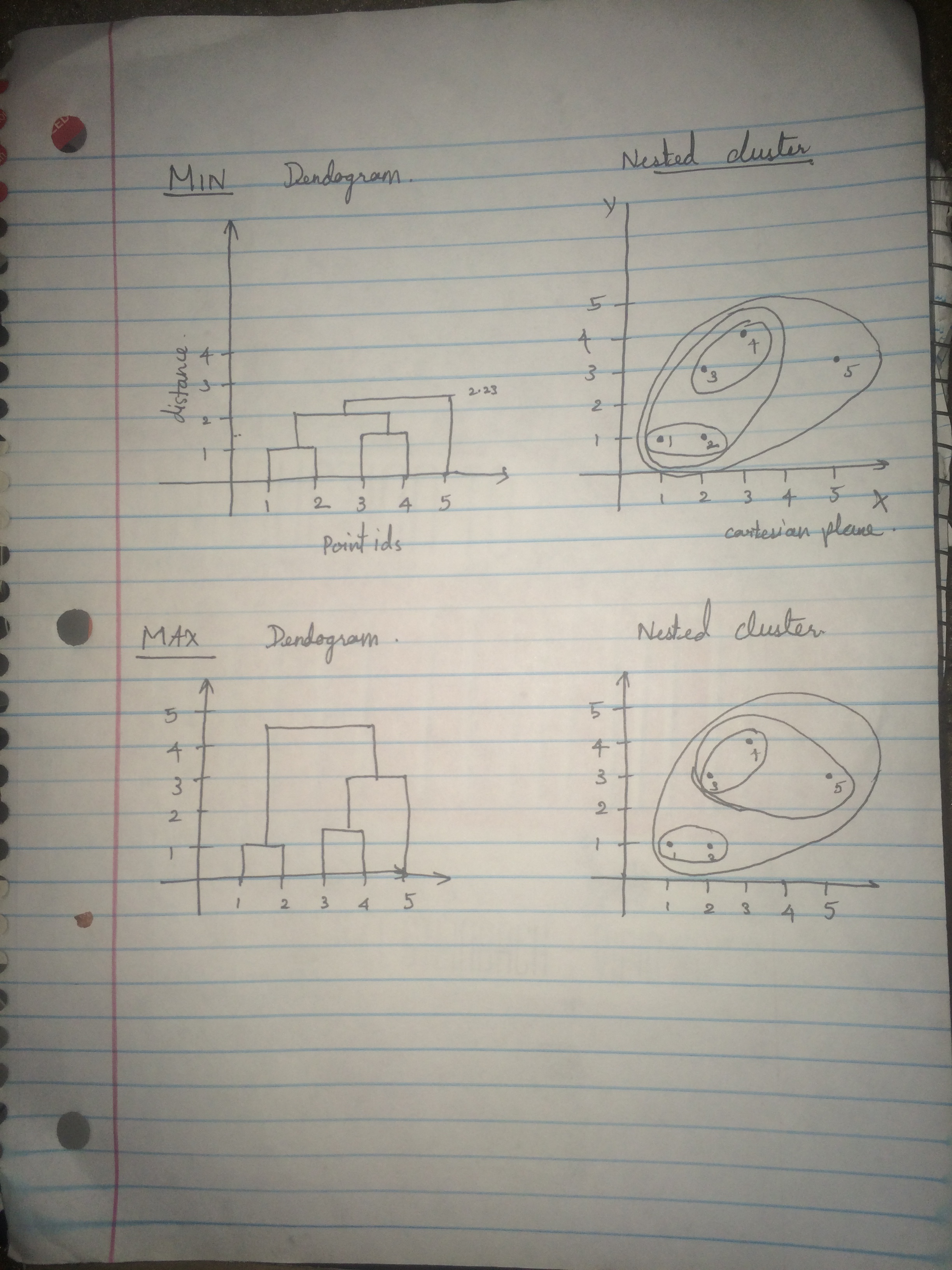
|  |  |  |
| --- | --- | --- |
| ID/ID | 1,2 | 3.4.5 |
| 1,2 | 0.00 | 4.47 |
| 3,4,5 | 4.47 | 0.00 |

All the points now belong to one big cluster.

Step 5:

|  |  |
| --- | --- |
| ID/ID | 1,2,3,4,5 |
| 1,2,3,4,5 | 0.00 |

**(b) Draw nested cluster and dendrogram for each of final clustering.**



**Q3.**

**a) Define the core,border and noise points:**

Core points : These points are present inside a density based cluster and have at least “MinPoints” number of points within its neighborhood. The neighborhood is defined by a user defined parameter called as epsilon distance (eps) and “MinPoints” is also a user defined parameter.

Border points : A border point is not a core point but is present within the boundary/ neighborhood of another core point/points .

Noise points : Noise points are those points that are neither a core point nor a border point. They are usually outliers and present sparsely far away from dense regions.

**b) Mark core, border and noise points for DBSCAN data given in Figure 1. Assume unit squares. MinPoints = 2 and radius (EpsilonDist) = 1.5 units. Please note all data points (\*) are at the intersection of grid lines.**

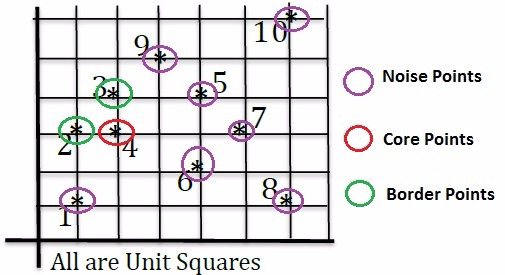
For this problem, Manhattan distance is used to check if a particular point is within it's neighbor hood or not.

Core points : 4,

Border points : 2,3

Noise points: 1,5,6,7,8,9,10

Explanation : Only points 2 and 3 are within epsilonDist from point 4. Hence point 4 becomes a core point and points 2 and 3 become border points. Points 5,6 and 7 are at a distance of two from each other which is greater that epsilonDist and are not close to 4 and hence are classified as Noise points. Points 1,8,9,10 are also don't have 2 points in their neighborhood and are far away from the only core point . Hence they are also classified as Noise points.



**c) Explain the conditions (situations) where DSCAN does not work well.**

DBSCAN doesn't work so well when the clusters have widely varying densities. (I.e if there is large differences in densities)

DBSCAN doesn't work well with higher dimensional data because it's not clear how to define densities for those type of data (Curse of dimensionality)

**d) Is DBSCAN deterministic? If not, explain how can you make it deterministic.**

DBSCAN is not completely deterministic. Sometimes border points can belong to more than one cluster and can be parter of either cluster. The result depends on the order of processing the data (Note : This doesn't usually occur and even if it occurs, it won't radically change the cluster properties).

DB Scan can be made deterministic by using an alternative called DBSCAN\* which considers the border points which belong to multiple clusters as noise and hence doesn't assign them to any cluster thus producing a fully deterministic cluster.

References :

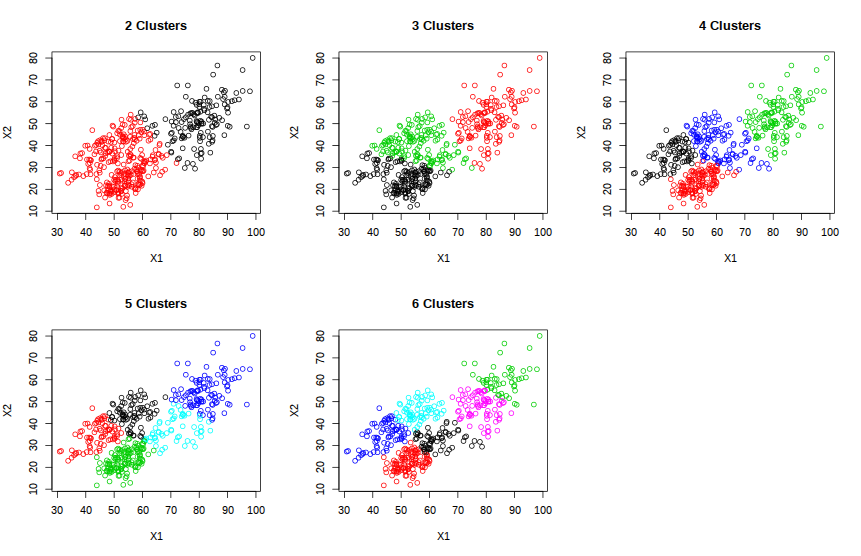
http://stats.stackexchange.com/questions/58855/why-do-we-use-k-means-instead-of-other-algorithms

<https://en.wikipedia.org/wiki/DBSCAN>

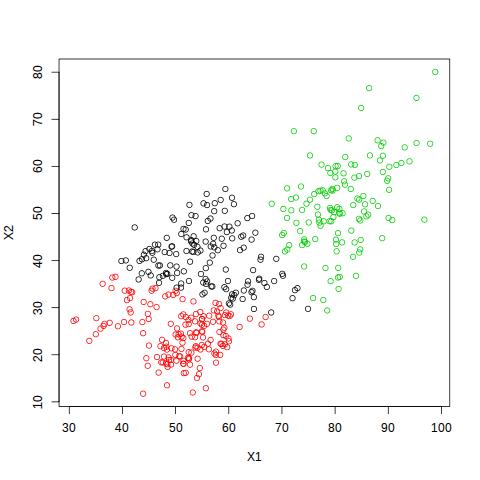
Introduction to Data mining by Pang-Ning Tan et al

Q4. a)

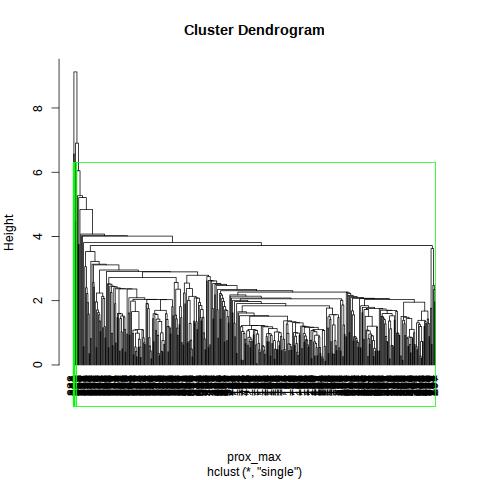
All the plots. For k = 2 to 6

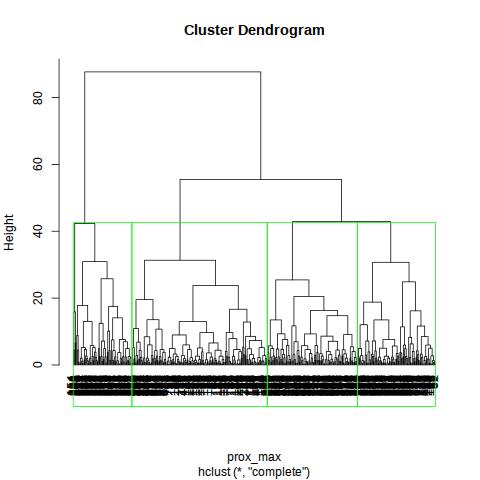


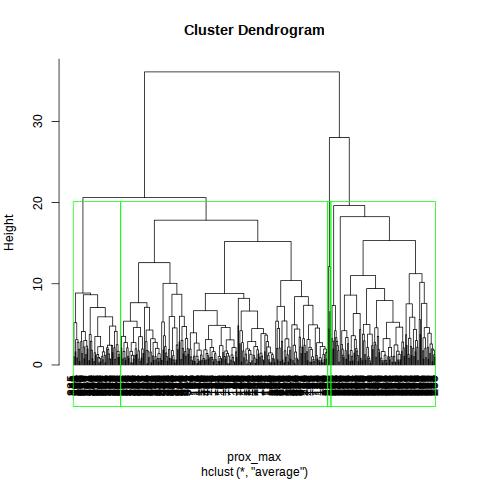
The best plot is for K=3 . While plotting the elbow plot for SSE Vs K, the elbow occurs at K=3 (Although the change is very subtle). Also These clusters seem to be more dense

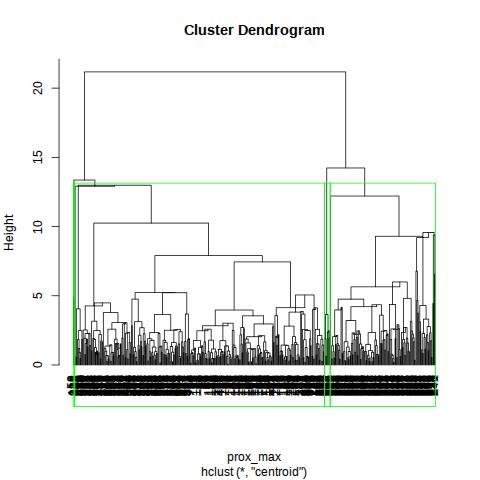


b)









For individual Cuts Please refer to the ZIP.

Average : a1 – a4

Single : s1 – s4

Centroid : ce1 – c4

Complete : c1-c4