CS 422: Data Mining

Solution to Spring 2018: Homework 3

Saptarshi Chatterjee
CWID: A20413922

April 8, 2018

1 Tan, Chapter 8 , Exercise 2, 6, 11, 12, 16

Q.2 2. Find all well-separated clusters in the set of points shown in Figure 8.35.?

Answer -

2. Find all well-separated clusters in the set of points shown in Figure 8.35.

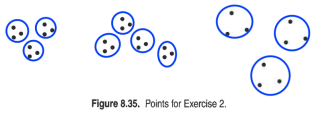
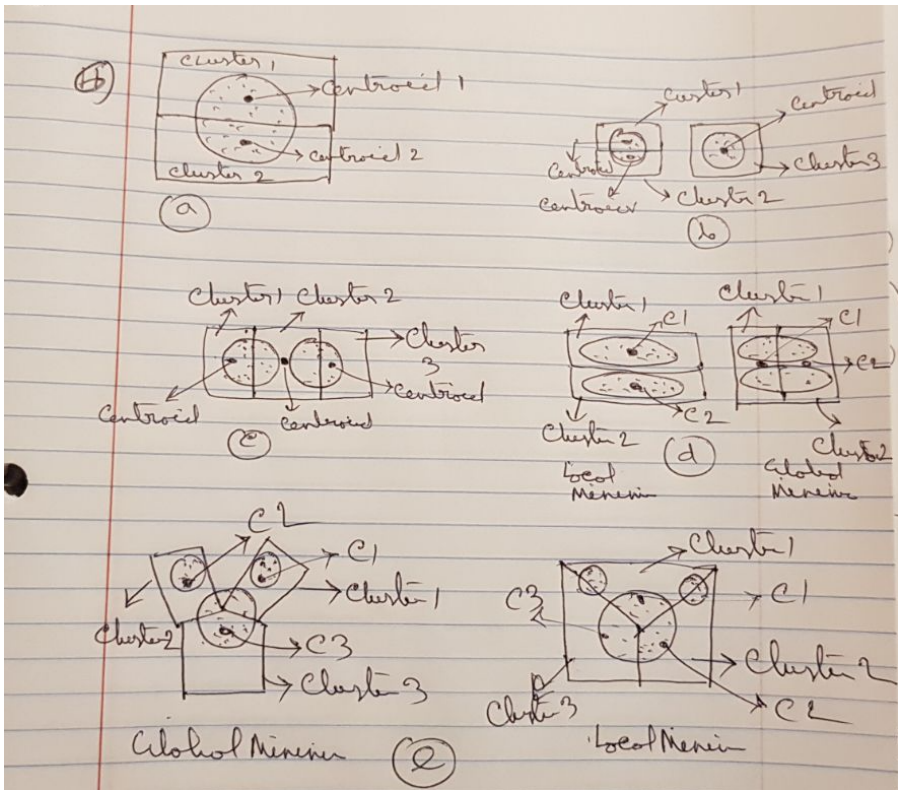


Figure 8.35. Points for Exercise 2.

Q.6 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

Answer -



6.a) There can be infinite ways to partition the points into two clusters. We can draw a diameter through the circle and the two halves of the circle will represent 2 clusters. As we can draw an infinite number of diameters just by changing the angle to the x-axis, so there is an Infinite number of ways to partition. And all the partition will have almost same global minimum error.

If we draw a perpendicular bisector to the diameter then the midpoint between the center and the circumference on both halves of the circle will be the position of the centroids of the cluster

6.b) Make any circle a cluster. And Make each equal halves of the other circle a cluster. Thus we get 3 clusters. Centroids will be center of the big cluster and midpoint between the center and the circumference on both halves of the 2nd circle will be the position of the 2nd and 3rd centroids. And all the partition will have almost same global minimum error.

6.c) Take the maximum distance between 2 points on the circumference of the circles. Equally, divide the line into 3 parts. These points will be centroids. And 3 rectangles that have these points as the center will be our three clusters.

6.d) There are 2 solution the left 1 produces Local minimum error and the right 1 will produce Global minimum error.

6.e) There are 2 solution the left 1 produces Global minimum error and the right 1 will produce Local minimum error.

Q.11 Total SSE is the sum of the SSE ...

Answer -
11.i) A particular attribute is constant, if it's SSE is constant for all the clusters. And it has minimal impact on clustering result.

11.ii) If a particular variable has low SSE for just one cluster then that variable dominates in defining the cluster

11.iii) If A particular attribute has high SSE for all the clusters then there is a high probability that this attribute is noise.

11.iv) If A particular attribute has high SSE for only one cluster then this variable doesn't help in defining the cluster, and the attribute that has low SSE for that cluster dominated this cluster.

11.v) Per variable SSE information helps us eliminate attribute that has little impact is defining the clusters. attributes that have low SSE for all the clusters are effectively constant and has a low impact on defining the cluster. attributes that have high SSE for all the clusters are essentially noise, and they impact overall SSE.

Q.12 The leader algorithm (Hartigan 1394]) represents each ...

Answer -
12 a) Leader algorithm as defined in https://onlinecourses.science.psu.edu/stat505/book/export/html/148 has following Pros and Cons

Pros -

1) We can create the entire cluster in O(n) time complexity where n is number of element in cluster
2) It will Always return the same result if the order of the input element are same.

Cons -

1) We cant have a predefined number of clusters like K means . So even if we know the value of K we can't control the Algo to make exact number of cluster
2) This simplistic Algo does not take SSE into consideration so sum of error is high . K means will almost always have a better result.

12 b) (i) Run the Algo in the 1st pass with threshold as mean of all the consecutive distances. (ii) Now Tweak the threshold value to get the predefined k clusters. (ii) Also after each pass we can calculate total SSE and by modifying threshold values we can check which configuration giving minimum SSE.

Q.16 Use the similarity matrix in Table 8.1 to perform single and complete link hierarchical clustering ...

Answer -