Tiffany Wong A20442087
CS422 - HW 7 Exercises 11/07/2021

1. Exercises: 1.1 (on PDF file separate)

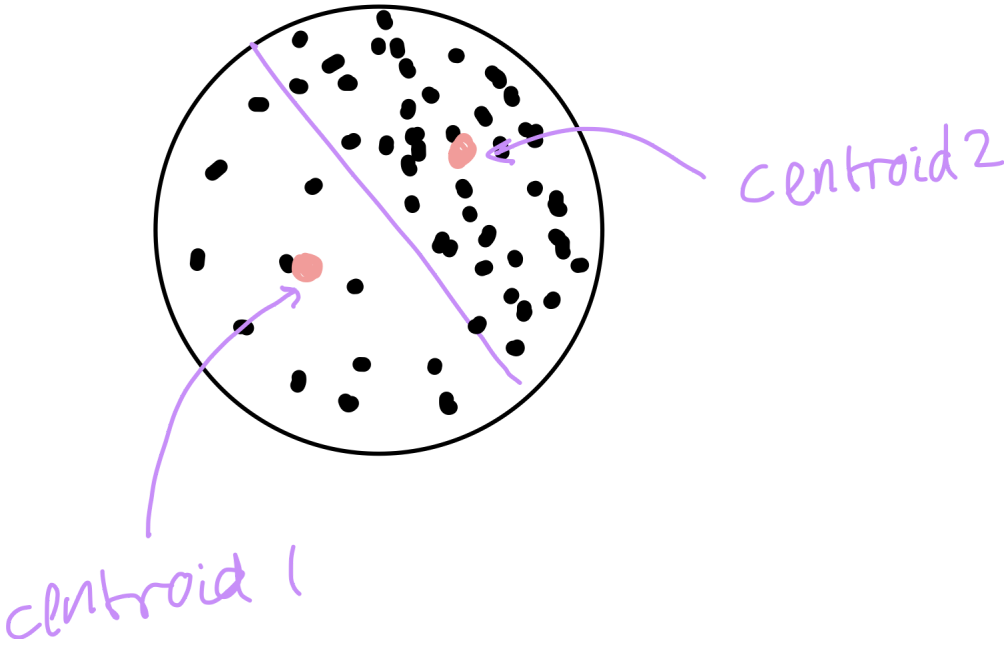
1.2 (2 points) **Tan, Ch 7, Ex. 2, 6, 11, 12.**

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

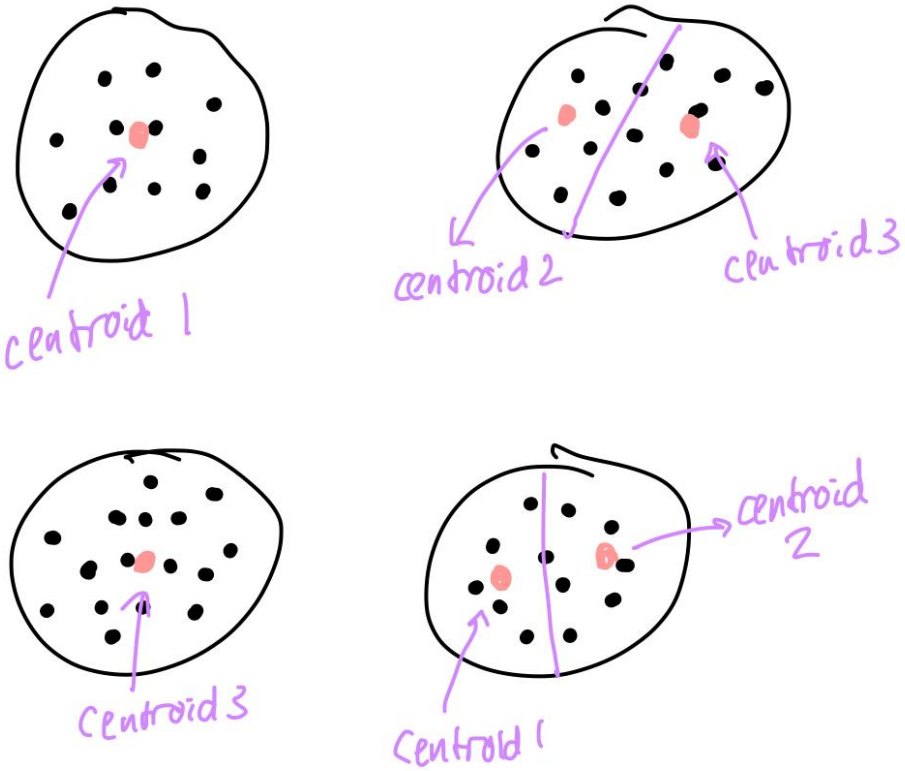
Figure 7.35.
Points for **Exercise 2**.

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).

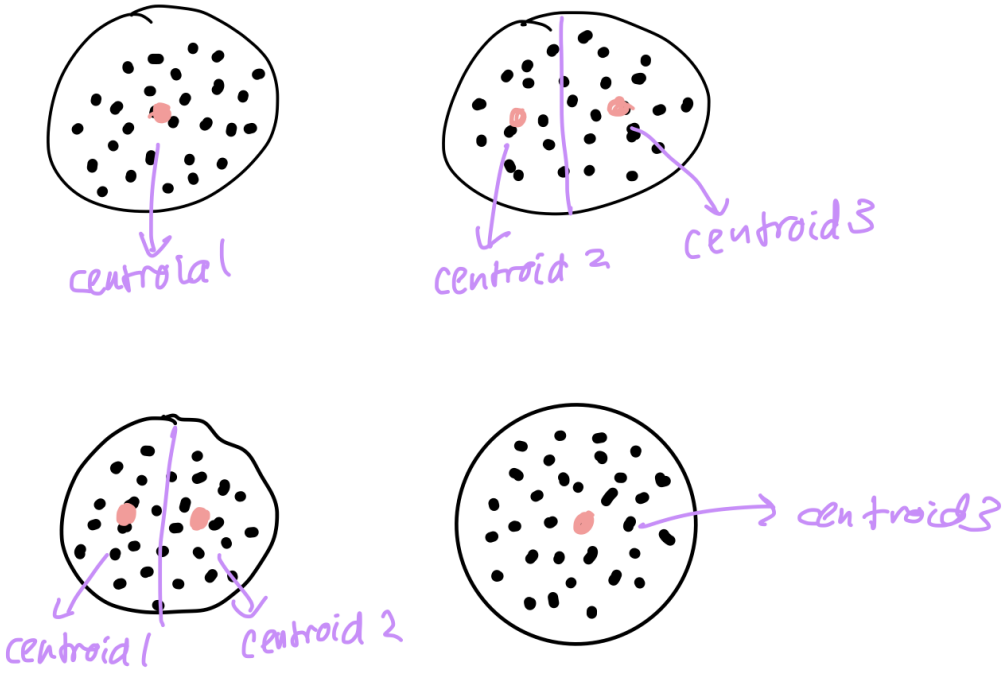
a. *K=2. Assuming that the points are uniformly distributed in the circle, how many possible ways are there (in theory) to partition the points into two clusters? What can you say about the positions of the two centroids? (Again, you don't need to provide exact centroid locations, just a qualitative description.)*



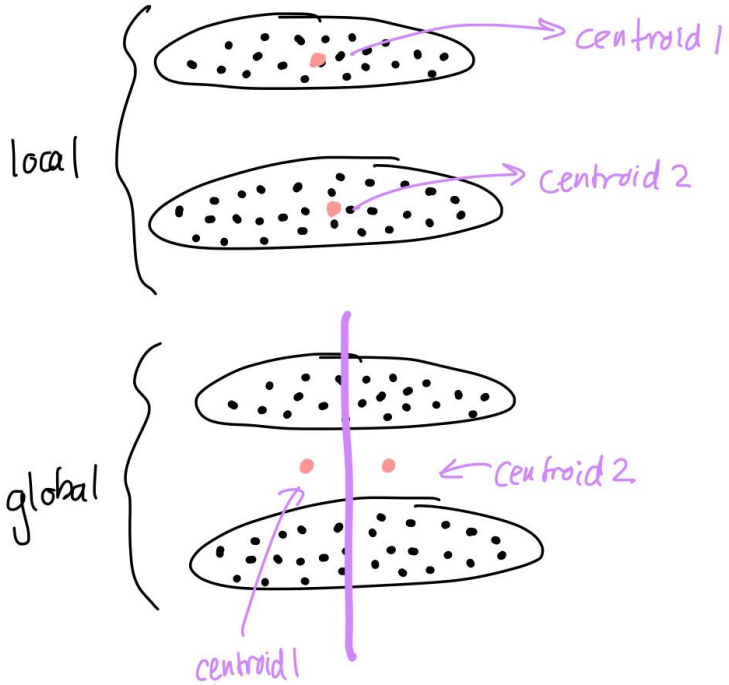
b. *K=3. The distance between the edges of the circles is slightly greater than the radii of the circles.*



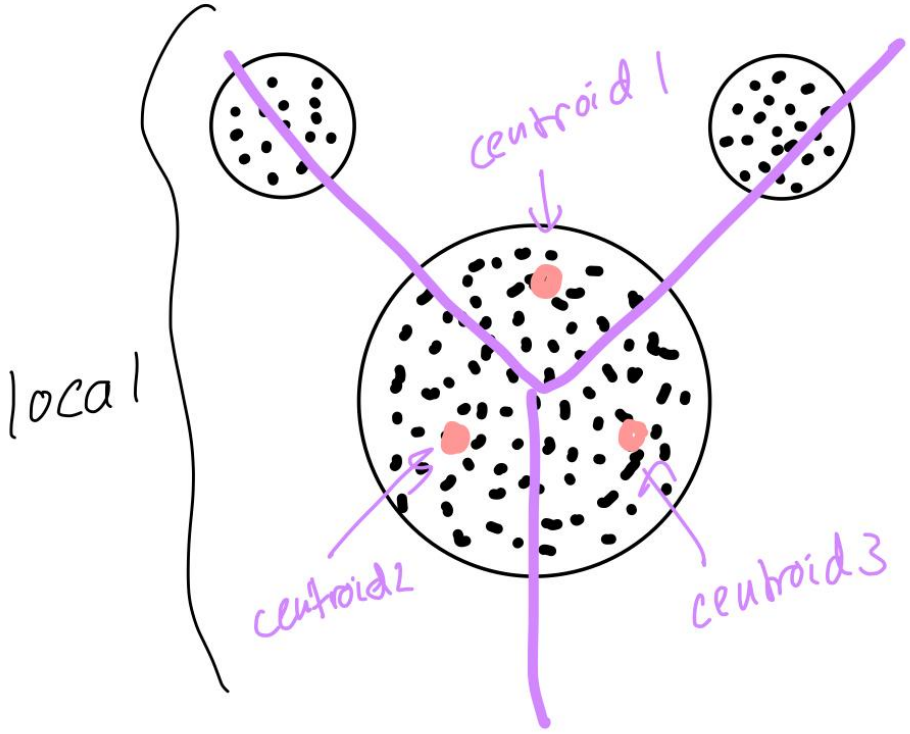
c. *K=3. The distance between the edges of the circles is much less than the radii of the circles.*

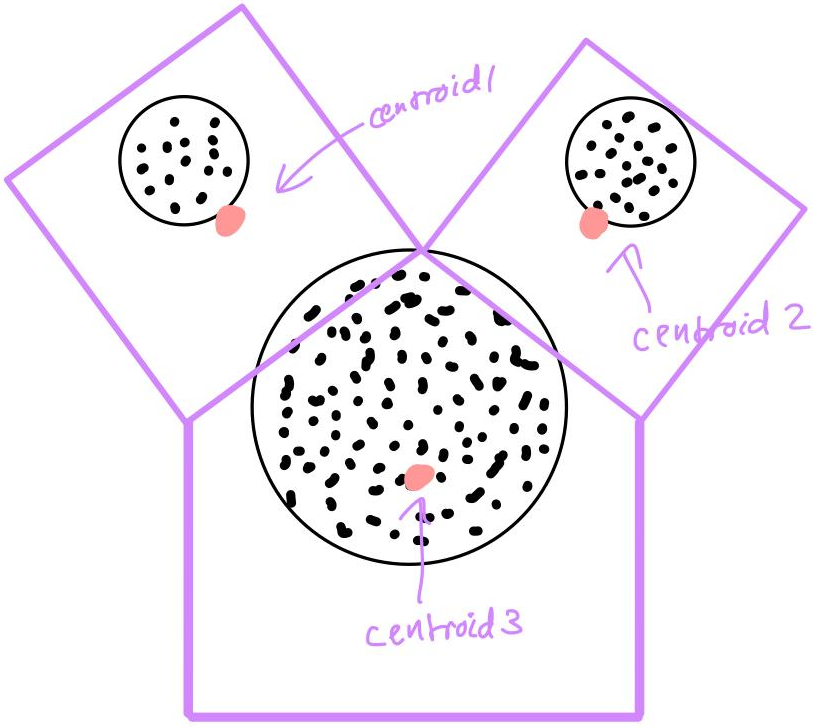


d. *K=2.*



e. *K=3. Hint: Use the symmetry of the situation and remember that we are looking for a rough sketch of what the result would be.*





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?*

It means that the variable is a constant and it provides no help in splitting the data into clusters.

- *Low for just one cluster?*

It means that that attribute will help in defining the cluster.

- *High for all clusters?*

It means that that variable then that means that the variable is noise.

- *High for just one cluster?*

It means that that variable does not help in defining the cluster.

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

You can use the per SSE information in order to help get rid of variables that provide no use in defining clusters.

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.

a. *What are the advantages and disadvantages of the leader algorithm as compared to K-means?*

The advantages to the leader algorithm when compared to the k-means algorithm is that it is easier to predict the leader value than the k-value. Also, it works better for global clusters and it works better with clusters of different sizes and densities.
The disadvantage of the leader algorithm is that the leader algorithm does not produce tight clusters unlike k-means, and it is computationally slower than k-means for large variables.

b. *Suggest ways in which the leader algorithm might be improved.*

One way to improve it is to introduce the concepts of uniqueness and agreement to the algorithm.
The concept of uniqueness will allow for only one processor to consider itself the leader for a cluster while the concept of agreement makes it so that the other processors know who the leader of the cluster is.