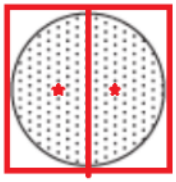
Mark Gameng

HW9

2.

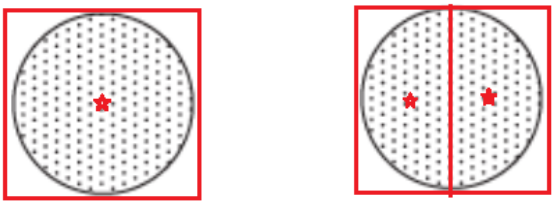
6.

a.



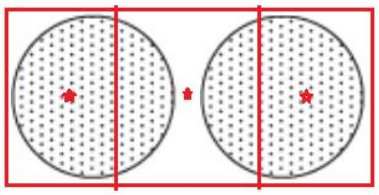
The split can be anywhere (any angle) as long as it equally splits the circle, so infinite ways to do so. The centroids will be symmetrical and close to the places shown in the diagram as stars.

b.



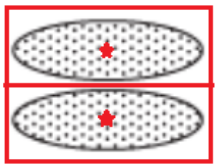
The two circles are farther away than the radius so separate. Two similar solutions – the first circle being the one bisected – but also infinite ways to do so because the bisection can be of any angle. The centroid is as shown, with the first circle being in the exact middle, and the second circle the two centroids are symmetrical in terms of the line between them.

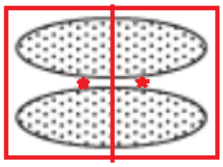
c.



Since distance between the circle is short and K = 3, will result in these kinds of sections. The centroids will be close to the places shown in the diagram.

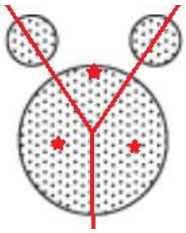
d.





Two different ways, with the first one separating the two circles, and the second one splitting both circles in half as shown in the diagram. The centroids for the first one will be in the center of the circles, while for the second one, the centroids will be symmetrical and in the middle of the two circles, as shown above. I think the first one would be a global minimum and the second would be a local minimum

e.



With K=3 the result will be similar to the diagram shown above. Not sure if there is any other solution.

11. If the SSE for one variable is low for all clusters, then that variable is quite useless in terms of defining the clusters. If the SSE for one variable is low for one cluster then that variable is useful as it is helpful in forming that specific cluster. If the SSE for one variable is high for all clusters then that variable is quite useless, and probably just noise. If the SSE for one variable is high for one cluster then for that specific cluster, it doesn't really help, but that variable is still useful for other clusters. With the per variable SSE information, you can ignore cases that are useless or even negatively affects the clustering such as variables with low or high SSE for all clusters.
12.
    a. One of the benefits is that it is more efficient than Kmeans(each point is only compared to all leaders once) and because of it, it may be better to use for huge data sets. Some of the disadvantages are that the clusters are dependent on the order of the points in the dataset due to how the algorithm works and thus, the number of clusters as well. Unlike the K-means where the user can specify the number of clusters, they cannot in the leader algorithm. Also, the user will have to specify a threshold for the leader algorithm to work.
    b. Rather than a user-specified threshold, the leader algorithm may be improved by using a sample or the whole data set to find an important information that can be used to better set the threshold for better results (for example the variation of the points or the mean).