Nicholas Saveas
CS 422-02

HW 3 Recitation Problems

Exercise 1.1

Chp 8 # 11

If the within SSE for one variable is low for all clusters then the points in each cluster are mostly placed correctly. If it is low for only one cluster then the points in that cluster are mostly placed correctly.

If the within SSE for one variable is high for all clusters then the points in each cluster are mostly placed incorrectly. If it is high for only one cluster then the points in that cluster are mostly placed incorrectly.

Per variable if the SSE is high for any number of clusters then the points are probably placed incorrectly. One way to fix this is to change the number of clusters or to change the clustering algorithm.

Chp 8 # 12

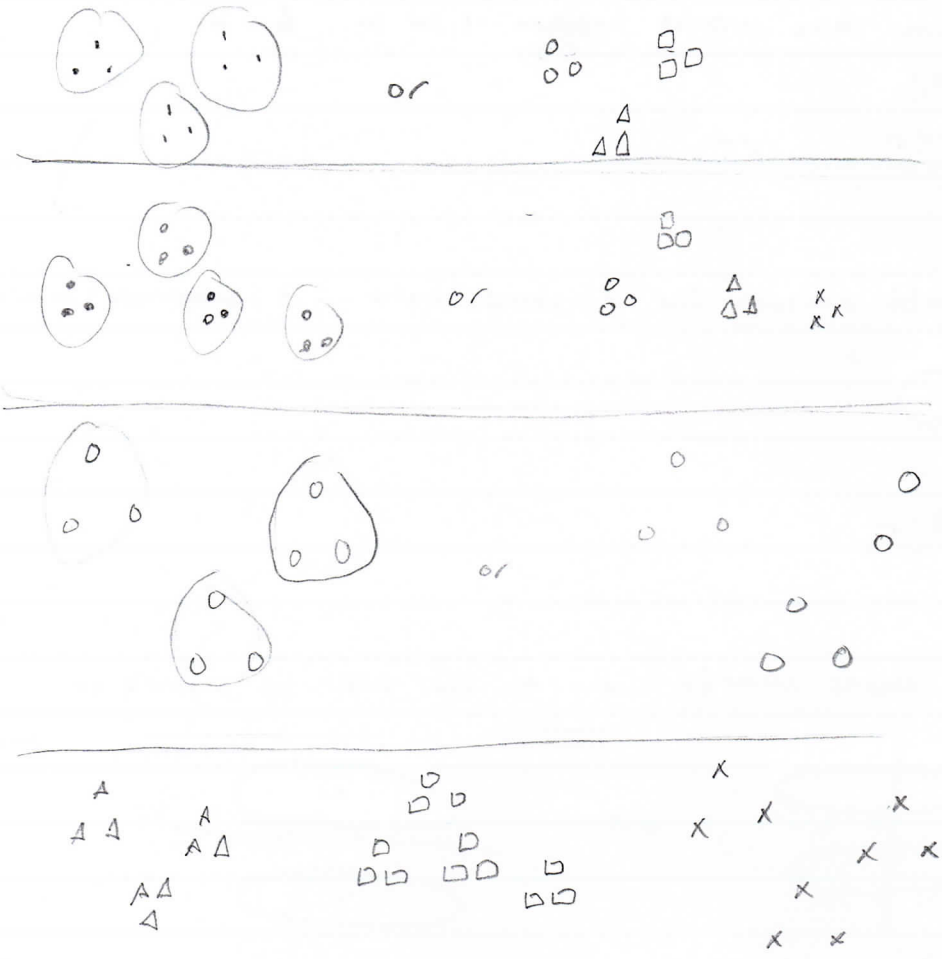
A) One of the advantages of using the leader algorithm is that it doesn't take as much time as the K-means algorithm - there is only one iteration and the algorithm doesn't look at the points very many times. One disadvantage is that the leader algorithm will likely place the points into incorrect clusters. If two leaders are on the edge of a circle the circle will be divided into two clusters because the leaders to not change locations. In the K-means algorithm one of the points in the circle will eventually push the other point out so that the circle becomes on cluster.

B) One way to improve the leader algorithm is to check now new leaders that are approx. two times the user-specified threshold because then it will create more defined, same-sized clusters. Under the current method of choosing new leaders, the algorithm will choose points on the edge of the radius defined by the threshold, but since points are chosen by the closest leader, the new leaders will take points away from the other leaders. This change may improve the performance.

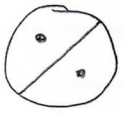
Nicholas Saveas

CS 422 HW 3 Chp 8

2.

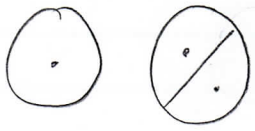


6. a) If the points are uniformly distributed then the circle would be divided into two halves. If one centroid is in one quadrant the other centroid is in the opposite centroid, and both centroids would be in the center

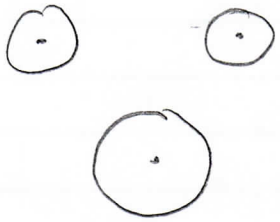


b) One centroid would be in the center of one circle and the other two centroids would be split like in 6 - a

c) I think the same thing would happen like in 6-b



e) k-means would divide the clusters into 3 seperate circles



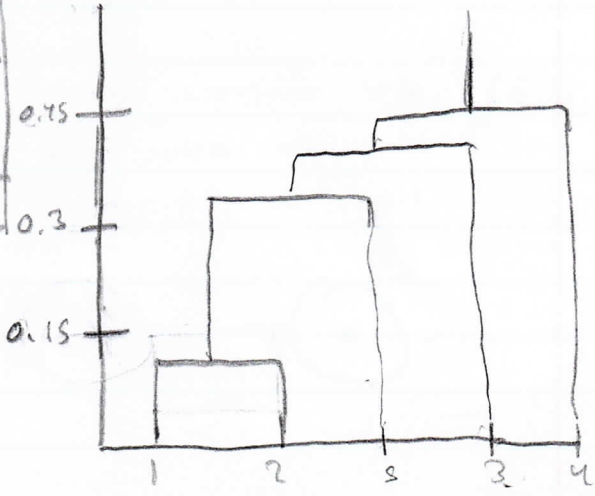
d) I think k means would divide the set of points in half.

16.

simple	P1 + P2	P3	P4	P5
P1+P2	X	X	X	X
P3	0.41	X	X	X
P4	0.47	0.44	X	X
P5	0.35	0.85	0.76	X

simple	P1+P2+P5	P3	P4
P1+P2+P5	X	X	X
P3	0.41	X	X
P4	0.47	0.44	X

simple	P1+P2+P5+P3	P4
P1+P2+P5+P3	X	X
P4	0.44	X



CS 422 HW Chp 8.

16.

complete	P1+P2	P3	P4	P5
P1+P2	x	x	x	x
P3	0.64	x	x	x
P4	0.55	0.44	x	x
P5	0.98	0.85	0.76	x

complete	P1+P2	P3+P4	P5
P1+P2	x	x	x
P3+P4	0.64	x	x
P5	0.98	0.85	x

complete	P1+P2+P3+P4	P5
~~→	x	x
P5	0.98	x