CMSC 691 — Introduction to Data Science — Fall 2019

1)Suppose that the data mining task is to cluster points (with (x, y) representing location) into three clusters, where the points are A1(2,10),A2(2,5),A3(8,4),B1(5,8),B2(7,5),B3(6,4),C1(1,2),C2(4,9). The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to show only (a) The three cluster centers after the first round of execution. (b) The final three clusters

Lets calculate the Euclidean distance between all these points

Euclidean distance d(p,q)= $\sqrt{(q1-b1)^2+(q2-b2)^2}$

Distance of A1 from all the points:

A1=0
A2= $\sqrt{25}$
A3= $\sqrt{72}$
B1= $\sqrt{13}$
B2= $\sqrt{50}$
B3= $\sqrt{52}$
C1= $\sqrt{65}$
C2= $\sqrt{5}$

Distance of A2 from all the points:

A2=0
A3= $\sqrt{37}$
B1= $\sqrt{18}$
B2= $\sqrt{25}$
B3= $\sqrt{17}$
C1= $\sqrt{10}$
C2= $\sqrt{20}$

Distance of A3 from all the points:

A3= 0
B1= $\sqrt{25}$
B2= $\sqrt{2}$
B3= $\sqrt{2}$
C1= $\sqrt{53}$
C2= $\sqrt{41}$

Distance of B1 from all the points:

B1= 0
B2= $\sqrt{13}$
B3= $\sqrt{17}$
C1= $\sqrt{52}$
C2= $\sqrt{2}$

Distance of B2 from all the points:
B2= 0
B3= $\sqrt{2}$
C1= $\sqrt{45}$
C2= $\sqrt{25}$

Distance of B3 from all the points:
B3=0
C1= $\sqrt{29}$
C2= $\sqrt{29}$

Distance of C1 from all the points:
C1=0
C2= $\sqrt{58}$

Distance of C2 from all the points:
C2=0

Initially we assign A1(Center 1), B1(Center 2), and C1(Center 3) as the center of each cluster, respectively.

1st iteration:
A1
d(A1,Center 1)=0 as A1 is Center 1
d(A1,Center 2)= $\sqrt{13}$ >0
d(A1,Center 3)= $\sqrt{65}$ >0
A1 ε Cluster 1

A2
d(A2,Center 1)= $\sqrt{25}$ =5
d(A2,Center 2)= $\sqrt{18}$ =4.24
d(A2,Center 3)= $\sqrt{10}$ =3.16- Smallest of all

A2 ε Cluster 3

A3:
d(A3,Center 1)= $\sqrt{36}$ = 6
d(A3,Center 2)= $\sqrt{25}$ = 5-Smallest of all
d(A3,Center 3)= $\sqrt{53}$ = 7.28

A3 ε Cluster 2

B1:
d(B1,Center 1)= $\sqrt{13}$ = 6
d(B1,Center 2)= 0 -Smallest of all
d(B1,Center 3)= $\sqrt{52}$

B1 ε Cluster 2

B2:
d(B2,Center 1)= $\sqrt{50}$ = 7.07
d(B2,Center 2)= $\sqrt{13}$ =3.60- Smallest of All
d(B2,Center 3)= $\sqrt{45}$ =6.70

B2 ε Cluster 2

B3:
d(B3,Center 1)= $\sqrt{52}$ = 7.21
d(B3,Center 2)= $\sqrt{17}$ =4.12- Smallest of All
d(B3,Center 3)= $\sqrt{29}$ =5.38

B3 ε Cluster 2

C1:
d(C1,Center 1)= $\sqrt{65}$
d(C1,Center 2)= $\sqrt{52}$
d(C1,Center 3)= 0 Smallest of All

C1 ε Cluster 3

C2:
d(C2,Center 1)= $\sqrt{5}$

d(C2,Center 2)= $\sqrt{2}$ =Smallest of All
d(C2,Center 3)= $\sqrt{58}$

C2 ε Cluster 2

New clusters: Cluster 1:A1, Cluster 2: A3,B1,B2,B3,C2, Cluster 3: A2, C1

The three cluster centers after the first round of execution.
Center 1: (2,10)

Center 2: Taking average of the x and y coordinates of A3,B1,B2,B3,C2 respectively= ((8+5+7+6+4)/5, (4+8+5+4+9)/5)= (6,6)

Center 3:Taking average of the x and y coordinates of A2, C1 respectively= ((2+1)/2,(5+2)/2)=(1.5,3.5)

After performing 2nd iteration the new Centers and Clusters will be:
New clusters: Cluster 1:A1,C2 Cluster 2: A3,B1,B2,B3 Cluster 3: A2, C1

The three cluster centers after the second round of execution.
Center 1: (3,9.5)

Center 2: (6.5,5.25)

Center 3: (1.5,3.5)

After performing 3rd iteration the new Centers and Clusters will be:
New clusters: Cluster 1:A1,B1,C2 Cluster 2: A3,B2,B3 Cluster 3: A2, C1

The three cluster centers after the third round of execution.
Center 1: (3.66,9)

Center 2: (7,4.33)

Center 3: (1.5,3.5)
The final iteration will be the third iteration as after performing the 4th iteration the Clusters don't change.

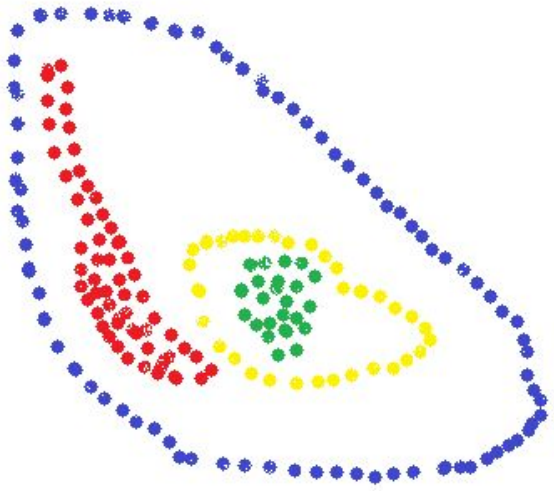
2)Why is it that BIRCH encounters difficulties in finding clusters of arbitrary shape but OPTICS does not? Propose modifications to BIRCH to help it find clusters of arbitrary shape.

Balanced Iterative Reducing and Clustering Using Hierarchies (BIRCH) begins by partitioning objects hierarchically using tree structures, and then applies other clustering algorithms to refine the clusters while **Ordering points to identify the clustering structure** (**OPTICS**) is an algorithm for finding density-based clusters in spatial data. Since BIRCH is a hierarchical based clustering approach, it makes use of the Intercluster proximity and Euclidean distance which tends to form only the spherical based structures otherwise it won't be able to achieve a high quality clustering effect and hence it encounters difficulty in finding clusters of arbitrary shape. OPTICS uses a density based clustering approach which tends to form dense regions on the basis of the points connected in a defined radius, which helps in forming clusters of arbitrary shape.

In order for BIRCH to find clusters of arbitrary shapes, we should use density based distance and connectivity based distance just like OPTICS does, instead of using the Euclidean distance which might lead to forming spherical shapes again. We should apply this technique at low levels of the CF Tree formation method, which might be useful in creating the arbitrary shapes.

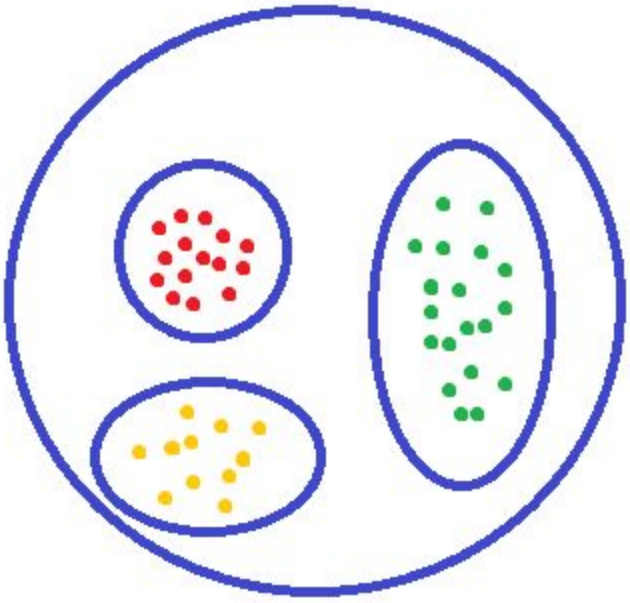
3)Present conditions under which density-based clustering is more suitable than partitioning-based clustering and hierarchical clustering. Give application examples to support your argument.

Density Based Clustering such as DBSCAN, OPTICS, DENCLUE can identify noise in the data while clustering and can find arbitrary size and shape clusters. It forms clusters on the basis of density, that means points in a cluster will have high density as compared with the points outside the cluster.



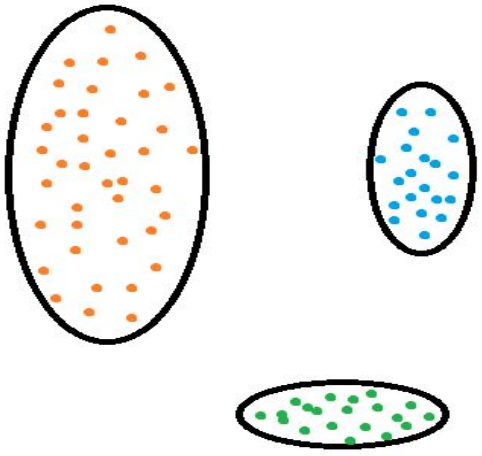
In case of Hierarchical Clustering:
It is not possible to undo the previous step: once the instances have been assigned to a cluster, they can no longer be moved around.

• Time complexity: Not suitable for large datasets
• Initial seeds have a strong impact on the final results
• The order of the data has an impact on the final results
• Very sensitive to outliers



In case of Partitioning Clustering:

- Forms Clusters only of Spherical Shape and encounters difficulty in making clusters of arbitrary size.
- Difficulty in finding clusters of different density and diameter.



4)Suppose that you are to allocate a number of automatic teller machines (ATMs) in a given region so as to satisfy a number of constraints. Households or workplaces may be clustered so that typically one ATM is assigned per cluster. The clustering, however, may

be constrained by two factors: (1) obstacle objects (i.e., there are bridges, rivers, and highways that can affect ATM accessibility), and (2) additional user-specified constraints such as that each ATM should serve at least 10,000 households. How can a clustering algorithm such as k-means be modified for quality clustering under both constraints?

In order to make any modifications for quality clustering we need to make sure that we don't disturb the actual functionality of the algorithm and also stay within the defined constraints which is, having some obstacle objects in the way of the ATM's and the households and each ATM should serve at least 10,000 households.

We can use the actual distance instead of calculating the Euclidean distance between the household and the ATMs, when we encounter obstacles such as bridges, rivers etc. To maintain the boundary condition which is each ATM serving at least 10000 households and also handles the obstacles in between as well, we need to calculate the actual central point from the ATM to the household which in this case will be an obstacle and not the average of all the objects in the cluster so that even if there are some outliers such as a household which is very far away then it won't affect the value of the centroid. This methodology is also known as PAM (**Partitioning Around Medoids**). In this manner if we keep the objects in small clusters, it will lead to a lot of micro clusters instead of creating big clusters and then people in that region will be able to reach the ATM's easily even in the presence of obstacles. So basically we aim towards keeping maximum obstacles that we can in the microclusters and avoid keeping obstacles along with the ATM in one cluster so that it can be more easily accessible from the other households.

As per my proposed modifications, I know that it won't be easy to select the exact obstacle which will act as a central point as there will be many obstacles in the way between the household and the ATM and choosing the right one won't be easy. While computing the most accurate obstacle to choose the central point the complexity of the algorithm will be impacted in a bad way as choosing the wrong obstacle will lead to the generation of wrong clusters or mapping the households with the wrong ATMs. While applying this clustering algorithm, we need to make sure that the choice of an obstacle (central point) gives us the desired results keeping in mind the defined constraints so that we can perform a quality clustering.