25/8/22

Module 5

CLUSTERING

★ K-Means Clustering –

Dataset –

{22, 9, 12, 15, 10, 27, 35, 18, 36, 11}

K = 3

C_1 = 22 {22, 27, 35, 18, 36} : Iteration 1

C_2 = 9 {9, 10}

C_3 = 12 {12, 15, 11}

Mean C_1 = 27.6

C_2 = 9.5

C_3 = 12.667

Iteration 2 :

C_1 = 27.6 {27, 35, 36, 22}

C_2 = 9.5 {9, 10, 11}

C_3 = 12.667 {12, 18, 15}

Mean C_1 = 30

C_2 = 10

C_3 = 15

Iteration 3 :

C_1 = 30 {35, 36, 27, }

C_2 = 10 {9, 10, 11, 12}

C_3 = 15 {18, 15, 22}

Mean C_1 = 32.667

C_2 = 10.5

C_3 = 18.334

25/8/22

Iteration 4:

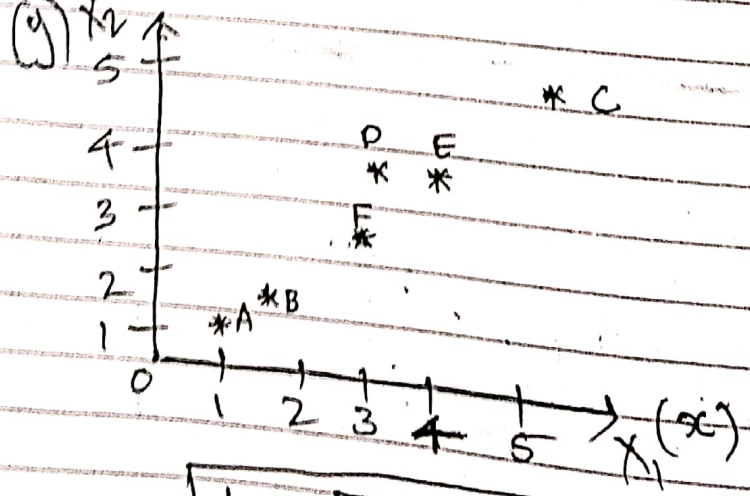
$C_1 = 32.67$ {35, 36, 27}
$C_2 = 10.5$ {9, 10, 11, 12}
$C_3 = 18.34$ {22, 18, 15,}

Mean $C_1 = 32.67$
$C_2 = 10.5$
$C_3 = 18.34$

✱ K-Means Clustering for two dimensional data

	$X_1(x)$	$X_2(y)$
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.3

K = 2



$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

25/8/22

Distance Matrix -

	A	B	C	D	E	F
A	0	0.707				
B	0.707	0				
C	5.6568	4.95	0			
D	3.605	2.91	2.2	0		
E	4.242	3.53	1.41	1	0	
F	3.047	2.34	2.97	0.7	1.2	0

Iteration 1 -

$C_1 = (1,1)$

$C_2 = (1.5, 1.5)$

	A	B	C	D	E	F
C_1	0	0.707	5.65	3.60	4.24	3.04
C_2	0.707	0	4.95	2.91	3.53	2.34

	A	B	C	D	E	F
C_1	1	0	0	0	0	0
C_2	0	1	1	1	1	1

(small ko 1 big ko 0)

~~small~~ $G_1 = \{A\}$

$C_2 = \{B, C, D, E, F\}$

Calculate mean

$C_1 = (1,1)$

$C_2 = \left(\frac{1.5+5+3+4+3}{5}, \frac{1.5+5+4+4+3.5}{5}\right)$

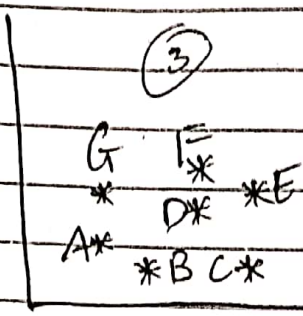
$= (3.3, 3.56)$

$= (3.3, 3.6)$

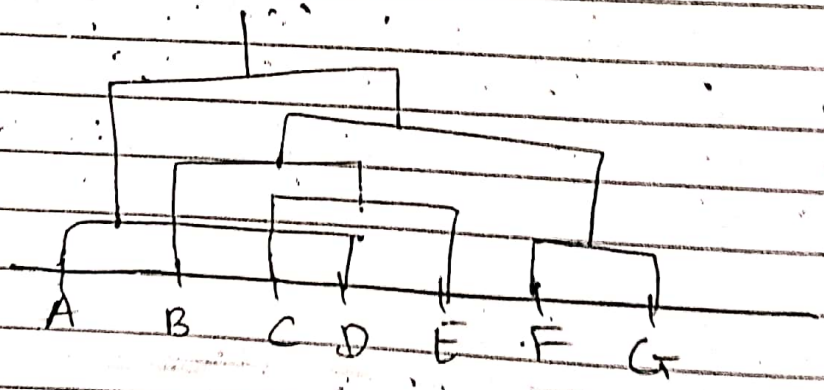
27/9/22

☆ Hierarchical Clustering :-

- Agglomnitive
- Divisive



Dendogram



There are 2 approaches of Hierarchical clustering
Agglomnitive (Bottom up) & Divisive (Top Down) Approaches.
Following is comparison between two approaches

1/1/22

Agglomnitive	Divisive
1) It starts with a single data point as a cluster	1) It starts with single big cluster which consists of all the data samples
2) Recursively add two or more appropriate clusters	2) Recursively divides into smaller clusters
3) The algorithm stops when k-number of clusters are achieved if k is specified otherwise it will stop after forming single large cluster	3) It will stop when k no of clusters are achieved if k is specified otherwise it will form the cluster of individual data points

Hierarchical clustering uses distance matrix as a clustering criterior. The distance matrix or adjacency matrix can be constructed using 3 approaches which needs to 3 different techniques of forming the cluster.

1) Single linkage → $D(C_1, C_2) = \min D(i,j)$ $i \in C_1$ & $j \in C_2$

2) Complete linkage → $D(C_1, C_2) = \max D(i,j)$ $i \in C_1$ & $i \in C_2$

3) Average linkage → $D(C_1, C_2) = \frac{T C_1 C_2}{N C_1 \times N C_2}$

7/9/22

where TC_1C_2 is sum of all pairwise clusters between C_1 & C_2. NC_1 is no of datapoints in C_1 & NC_2 is no of datapoints in C_2

Consider the following distance matrix given & apply hierarchical clustering using single, complete & average linkages dendograms.

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	4	5	9	0	
5	5	10	2	8	0

* Single linkage →

		C_1	C_2	C_3	C_4	C_5
		1	2	3	4	5
C_1	1	0				
C_2	2	9	0			
C_3	3	3	7	0		
C_4	4	4	5	9	0	
C_5	5	5	10	[2]	8	0

Iteration 1 -

The distance between datapoint 3 & datapoint 5 is minimum.

So we can group 3 & 5 in one cluster

	3,5	1	2	4
3,5	0			
1	3	0		
2	7	9	0	
4	8	4	5	0

Iteration 2 -

The distance between cluster of 3, 5 & 1 is smallest. So combine 3, 5 & 1 as a single cluster.

	1,3,5	2	4
1,3,5	0		
2	7	0	
4	4	5	0

Iteration 3 -

Distance between cluster 1, 3, 5 & 4 is smallest. So combine 1, 3, 5 & 4 as a single cluster

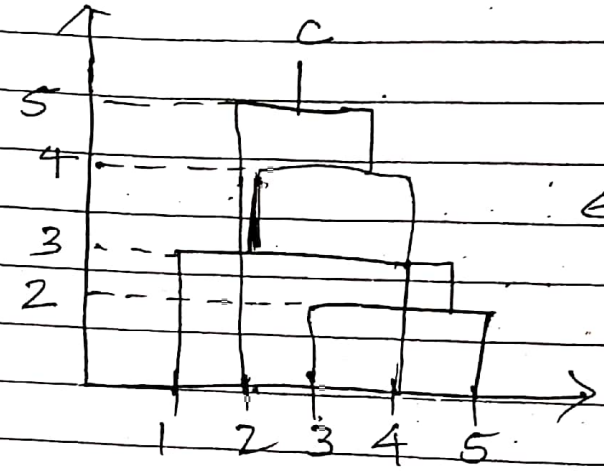
	1,3,4,5	2
1,3,4,5	0	
2	5	0

Iteration 4 -

Distance between cluster 1, 3, 4, 5 & 2 is smallest. So we can club into 1, 2, 3, 4, 5 cluster

: 7/9/22

	1, 2, 3, 4 5
~~1~~ 2, 3, 4, 5	○

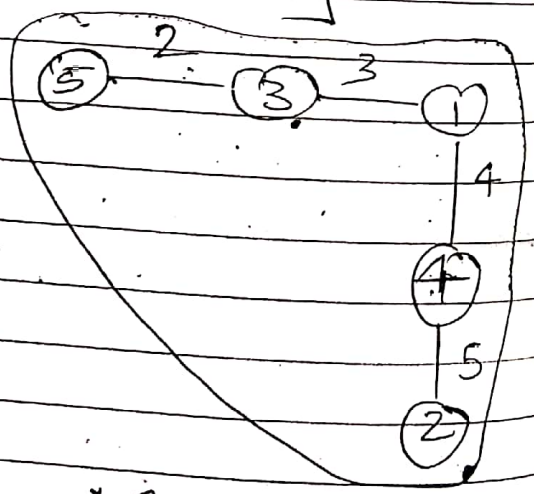


← Dendogram (Top Down)

* Complete linkage -

Edge	Cost
5, 3	2
1, 3	3
1, 4	4
4, 2	5
4, 5	5
3, 2	7
4, 5	8
4, 3	9
1, 2	9
2, 5	10

Initially



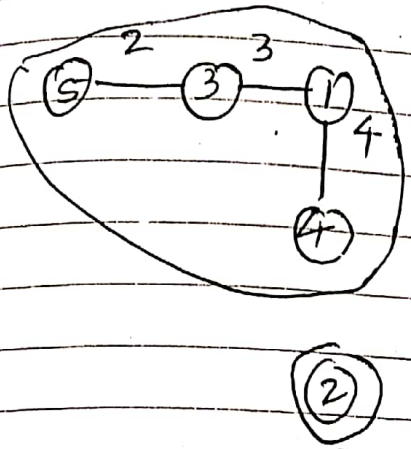
MST

↓

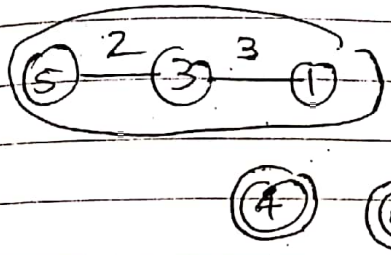
Minimum Spanning Tree

1/22.

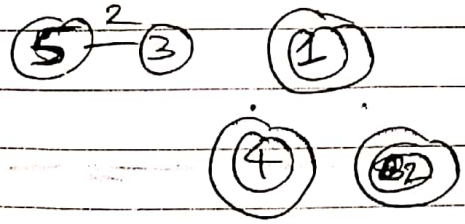
Iteration 1 -



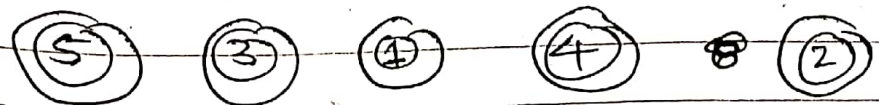
Iteration 2 -

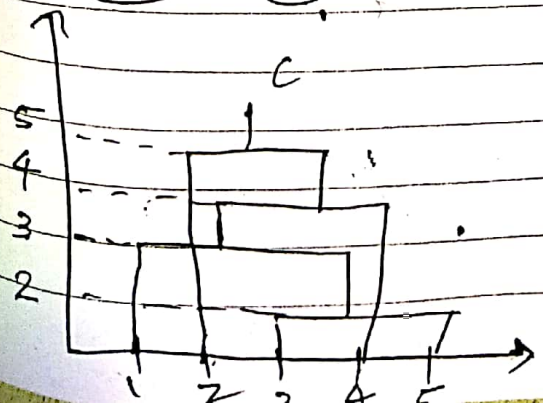


Iteration 3 -



Iteration 4





← Dendogram (Bottom Up)

7/9/22

* Average Linkage -

	C1	C2	C3	C4	C5
	1	2	3	4	5
C1 1	0				
C2 2	9	0			
C3 3	3	7	0		
C4 4	4	5	9	0	
C5 5	5	10	[2]	8	0

Iteration 1 -

Distance between 3 & 5 is minimum so we can club together

	3,5	1	2	4
3,5	0			
1	3+5 = 8/2 = [4]	0		
2	7+10 = 17/2 = 8.5	9	0	
4	9+8 = 17/2 = 8.5	[4]	5	0

Iteration 2 -

Distance between 1 & 4 is minimum so we combine 1 & 4

	3,5	1,4	2
3,5	0		
1,4	3+9+5+8 = 25/4 = [6.25]	0	
2	8.5	9+5 = 14/2 = 7	0

Iteration 3 -

Distance between 3,5 & 1,4 cluster is minimum so we can club together

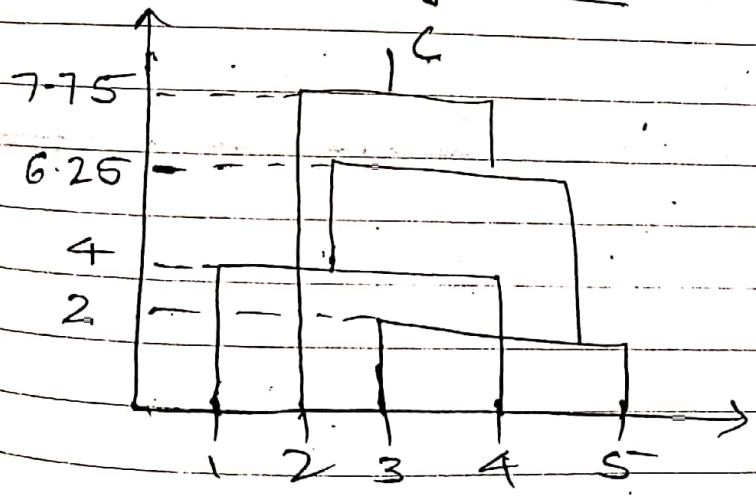
	1,4,3,5	2
1,4,3,5	○	
2	8.5+7=15.5=7.75	○

Iteration 4 :-

Distance between 1,4,3,5 & 2 is minimum so we can club together

	1,2,3,4,5
1,2,3,4,5	○

Dendogram (Top-Down)



Consider the following datasamples with 2 features & construct the dendogram with single, complete & average linkages

7/9/22

	x_1	x_2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

A) Distance Matrix –

	A	B	C	D	E	F
A	0					
B	0.707	0				
C	5.65	4.95	0			
D	3.605	2.91	2.23	0		
E	4.24	3.53	1.41	1	0	
F	3.05	2.34	2.97	0.7	1.2	0

★ Single linkage –

Distance between F & D is smallest so combine together

Iteration1 –

	F,D	A	B	C	E
F,D	0				
A	3.05	0			
B	2.34	0.707	0		
C	2.23	5.65	4.95	0	
E	1	4.24	3.53	1.41	0

Iteration2 –

Distance between A & B is smallest so combine together

	F,D	A,B	C	E
F,D	0			
A,B	2.34	0		
C	2.23	4.95	0	
E	1.41	3.53	1.41	0

Iteration 3-

Distance between F,D & E is smallest so combine

	F,D,E	A,B	C
F,D,E	0		
A,B	2.34	0	
C	1.41	4.95	0

Iteration 4 -

Distance between F,D,E & C is smallest so combine

	F,D,E,C	A,B
F,D,E,C	0	
A,B	2.34	0

Iteration 5 -

Distance between F,D,E,F & AB smallest so combine

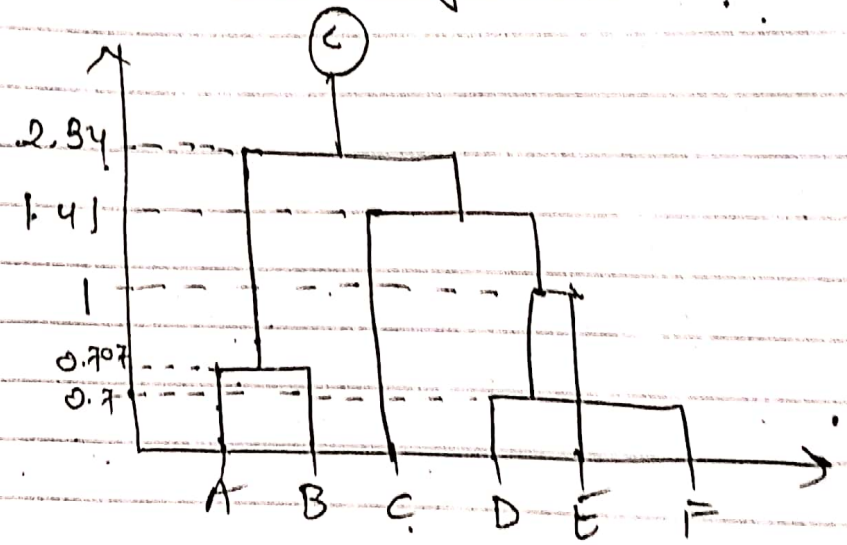
	A,B,C,D,E,F
A,B,C,D,E,F	0

7/9/22

✱ Average Linkage

Iteration 5 -

Dendogram



✱ Average linkage -

	A	B	C	D	E	F
A	0					
B	0.707	0				
C	5.65	4.95	0			
D	3.605	2.91	2.23	0		
E	4.24	3.53	1.41	1	0	
F	3.05	2.34	2.97	0.7	1.2	0

Iteration 1 -

Distance between D & F smallest so combine

D, F	A	B	C	E
O				
3.05	O			
3.3275	0.707	O		
2.6	5.65	4.95	O	
1.1	4.24	3.53	1.41	O

Iteration 2 -

Distance between A&B smallest so combine

D, F	A, B	C	E
O			
3.1887	O		
2.6	5.3	O	
1.1	3.885	1.41	O

Iteration 3 -

Distance between D, F & E smallest so combine

D, F, E	A, B	C
O		
3.5368	O	
2.005	5.3	O

Iteration 4 -

Distance between D, F, E & C smallest so combine

D, F, E, C	A, B
O	
4.4184	O

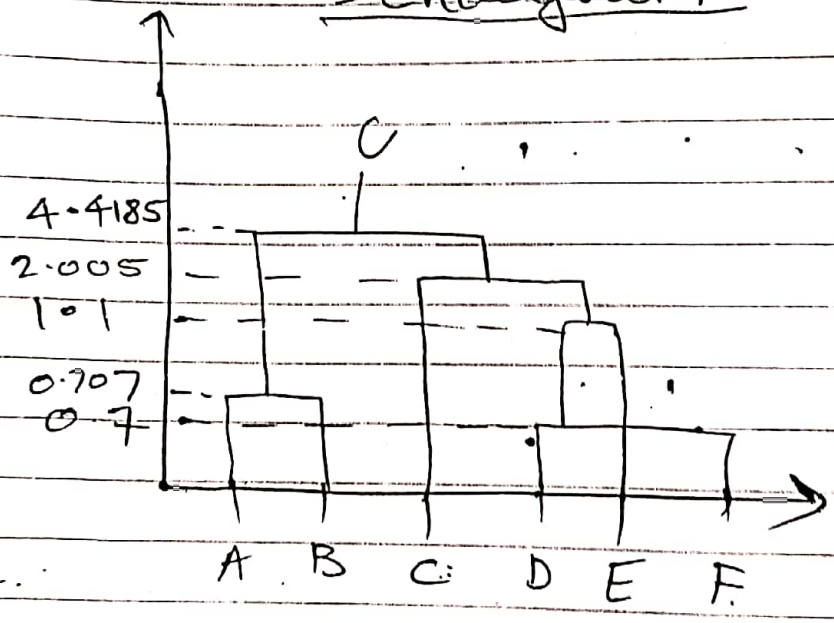
7/9/22

Iteration 5

Distance between D, F, E, C & A, B smallest so combine

	A, B, C, D, E, F
A, B, C, D, E, F	0

Dendogram



21/9/22

★ Density Based Clustering –

- Partioning the hierarchical methods are designed to find spherical clusters. They have difficulty in finding clusters of arbitary shape
- To find clusters of arbitary shape, we can model clusters as dense regions in the object space separated by sparse regions. This is the main idea behind density based clustering

22) DBSCAN (Density based Spatial clustering of Application with noise) -

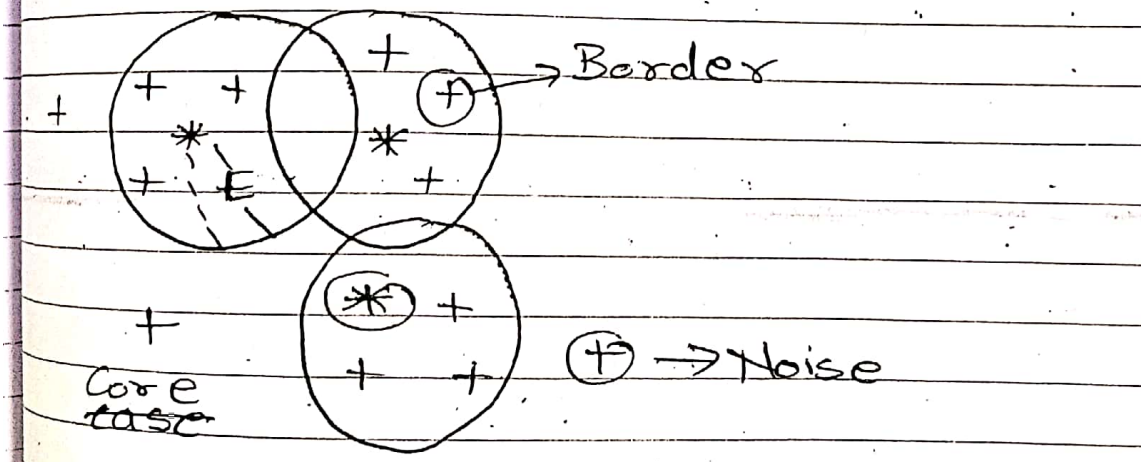
The density of an object O can be measured by the number of objects closed to O

The algorithm finds case objects, the objects whose neighbourhoods are connected to form dense regions as clusters.

The algo has 2 user defined parameter

E - neighbourhood - to specify neighbourhood an object O with raidus absolone (E) centered at O.

Mean point - To decide the density threshold as dense region. In this concept, we come across 3 different data points



Core - This is a point that has at least 'm' points within distance n from itself

Border - This is a point that has at least one core point at distance n

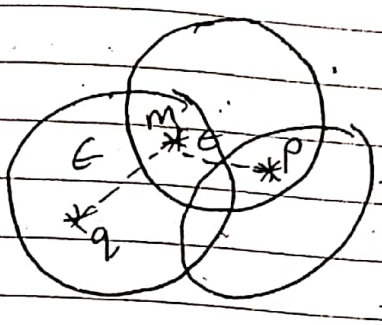
Noise - This is a point that is neither a core nor a border & it has less than 'm' points within distance n from itself

21/9/22

★ Concept of DBSCAN -

- The Algo proceeds by arbitarily picking up a point in the dataset (untill all points have been visited)
- If there are atleast mean points, points within raidus of ϵ to the point then we consider all these points to be part of the same clusters
- The clusters are then expanded by recurssively repeating the neighbourhood calculation for each neighbourhood
- Given a Set D of objects, the algo first identifies the core objects using ϵ and mean points, these core objects and their neighbourhoods are used to form a denser region
- For a core object q & an object p it is said that p is directly density reachable from q if p is within ϵ neighbourhood of q
- The object p is density reachable from a core obj q, if there is chain of obj p_1, p_2, p_n such that $p_1 = q$ & so on upto $p_n = p$ & p_{i+1} is directly density reachable from p_i where $1 \leq i \leq n$ wrt ϵ & mean points

21/9/22



- M is directly density reachable form q & p is density directly reachable form M. So P is ~~directly~~ density reachable form q
- If p & q are core, then they both are density reachable form each other if q is a core obj & p is not then p is density reachable form q but q is not from p
- Two points p_1 p_2 are density connected if there is an obj q such that both p_1 & p_2 are density reachable form q wrt ϵ & mean p
- A subset C of Dataset D is a dense region & so a cluster if -

i) for any 2 obj O_1 & O_2 $\in$ C, O_1 & O_2 are density connected

ii) There exists no obj $O \in C$ & another obj $O' \in (D-C)$ such that O & O' are density connected

★ The Algo -

Step 1- Initially all the objects in the given set are unvisited $D = \{ - - - \}$

21/9/22

Step 2- An Unvisited Object p is selected randomly $D = \{A, B, C\}$

3 - p is marked as visited

4 - If p is not a core obj then it is marked as noise point

5 - If p is a core obj then a new cluster C is created for p & all the objects in ϵ neighbourhood of p are added to a candidate set N, $N = \{ \ldots \}$

6 - For each object p' in N which is unvisited, the algo labels it visited & checks if it is a case obj.

A. $N = \{D, F, G\}$

B. $\rightarrow$

C. p in N

D.

E F G

7 - If p' is a core obj, then all the obj in ϵ neighbourhood of p' are added to N

8 - Algo adds the objects to cluster from C the objects of N that are not already added to clusters & such objects are removed from N

9 - The process is repeated until N is empty & the cluster C is completed

10 - To find the next cluster, algo randomly selects an unvisited obj from the remaining objects (D - C)

21/9/22

11 - The clustering process is continued untill all the points are visited.

28/9/22

Module 6

* EM Algorithm -

It is an approach for maximum likelihood estimation (MLE) in the presence of latent variables. The most discussed application of EM Algorithm is for clustering with mixture model. The EM Algorithm is an iterative approach that cycles between 2 steps i.e estimation step(E) and maximoization step(M)

E: Classify the data using current theory

M: Generate the best theory using the current classification of data.

Step E generates the expected classification for each example and Step M generates the most likely theory given the classified data.

Eg) Imagine there are two coins A and B. One is more likely to get heads, the other more likely to get tails. You pick one at random and toss it and then find out which one was it. Suppose we do it five times i.e