Page No.
Date

Week 6

Density Based Clustering

DBSCAN: Density based Spatial Clustering of Application and Noise.

⇓ data points ⇓ Outliers.

Concave

In k means / FCM / GMM

⇓

No way to Classify Outlier.

i A j
are density-connected

→ every data is defined in following way.

1) Core
2) border
3) outlier.

2 parameter 1) Mp: points in neighbour pts
2) ε : neighbour radius.

# of pts in radius ≥ Mp

Classify it as core points.

border

If no. of points ≤ Mp - Then is border

⇓

Necessary cond^n

Page No.
Date

border — Density.

Sufficient cond^n: reachable from any core points and # pts $< M_p$.

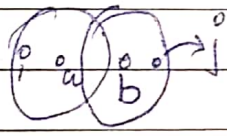
~~Outlier~~

Outlier: pts which are not core or border are outlier

for parameter choosing

ϵ should as small as possible
M_p should be as large as possible.
for high Density.

Density connected



i & j are not density reachable to each other

but i, j are density reachable to some core pts so i, j are density connected

Density connected is transitive property

$i \to j \quad j \to k$.

So $i \to k$ are density connected

Clustering Process.

→ Randomly choose a data pts which is not part of any cluster or outlier

→ Find out if data pts is core pt.

→ ~~At Start~~

if pt is core pt:

~~for~~ start new cluster

check for neighbour of ~~the~~ new cluster.

Keep on doing until u don't have any pt in neighbour that is not classified as core or border (part of cluster)

~~Hierarchical~~

Extra

ADVANTAGES

→ Doesn't require a-priori specification of number of cluster.

⇒

→ Able to identify noise data while clustering

⇒ find arbitrarily size and arbitrarily shaped cluster

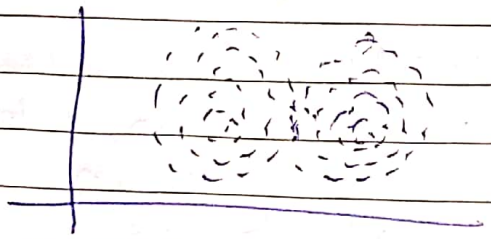
~~Disadvantage~~

Page No.
Date

DBSCAN Disadvantage.

1) fails in case of varying density clusters

2) Fails in case of neck type of Data.



DBSCAN fails to identify 2 cluster.

3) Doesn't work well in case of high Dimensional data.

Nearest-

SNN (Shared Neighbour)

SNN Tried clip ~~the~~ on the disadvantages to DBSCAN

⇒ For higher dimension SNN introduced some new notation for distance & density

Major points of SNN.

Distance

→ euclidian distance is not good choice of higher dimension

→ use different Similarity measure in terms of KNN.

→ Define Similarities in terms of new distance.

Jarvis-Patrick algorithm

Step 1 : SNN Sparsification.
→ Construct an SSN Graph from data matrix
→ if p & q have each other in KNN list
→ then create link between them

Step 2 : Weighting

Weight the links with $sim(p,q) = |NN(p) \cup NN(q)|$ where NN(p) & NN(q) are k neighbour of p & q resp

Step 3 : Filtering
→ then filter the edges
→ remove all edges with weight less than some threshold

Step 4 : clusters.
Lets all connected component be cluster.

Density

→ In euclidian space, density is no. of points per unit volume.

→ but as dimension increases volume decrease rapidly → density tends 0 for same N_p.

Page No.
Date

New Concept

if KNN is close, then region is most likely to be high density

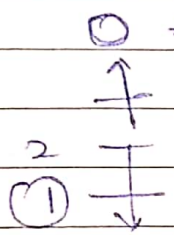
→ Su distance to Kth neighbour gives a measure of density of apt

~~Some advanced measures of Density~~

Hierarchical Clustering

1) Bottom up / Agglomerative
2) top-down / Divisive.

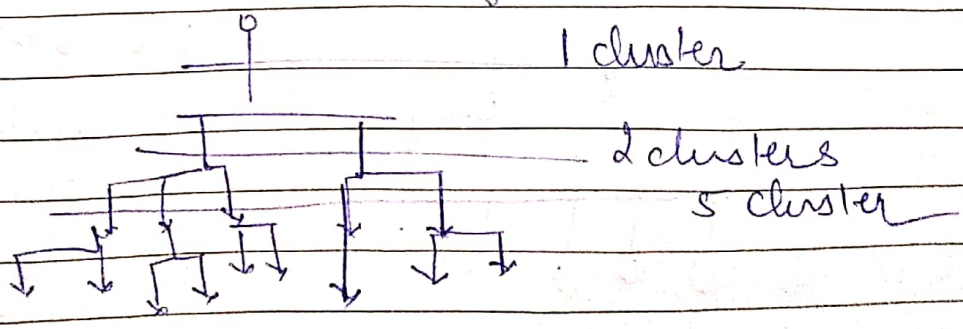
Levels are defined



. . . . . All pts are different cluster

x_1 x_2 - - - x_n

Tree structure / Dendogram.



Page No.
Date

Eg 500 cities name of nation

O → All belong to one nation

→ National capital

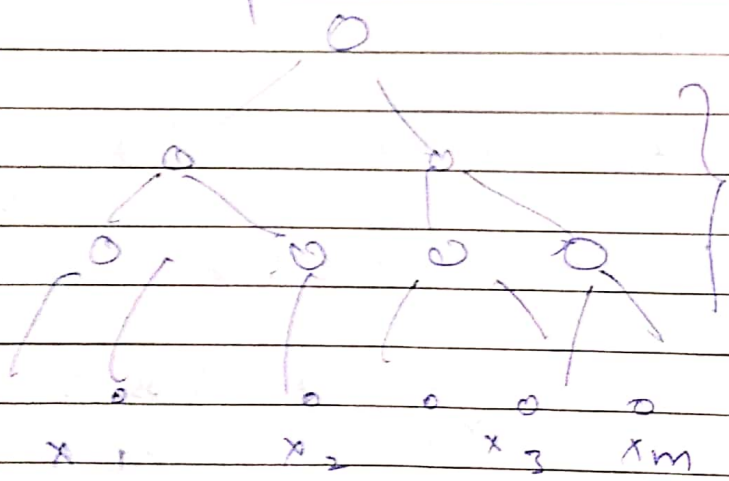
→ State capital

→ districts

→ 500 city name

~~Speech data having multiple speaker.~~
~~also eg of tie.~~

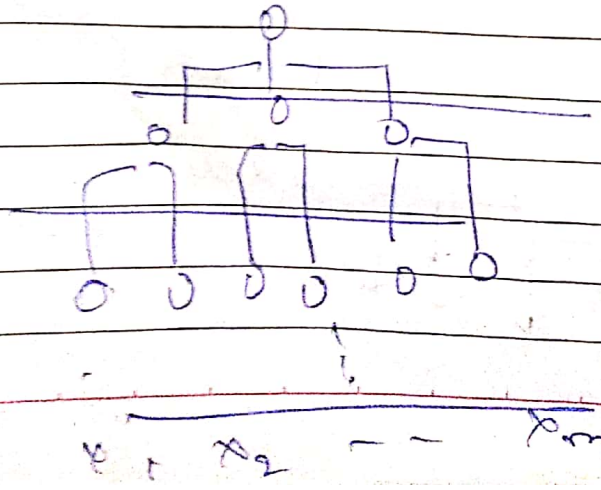
Bottom up.



we keep merging until we get one.

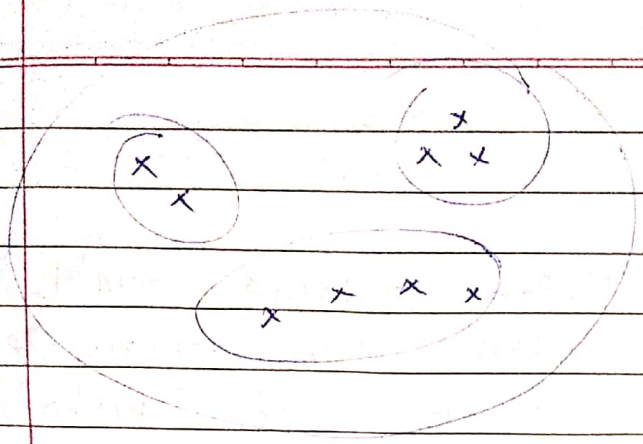
We all feature use clusters.

top-down



We start with on & keep on divid until we have no. of cluster = data set [of

Scanned with CamScanner

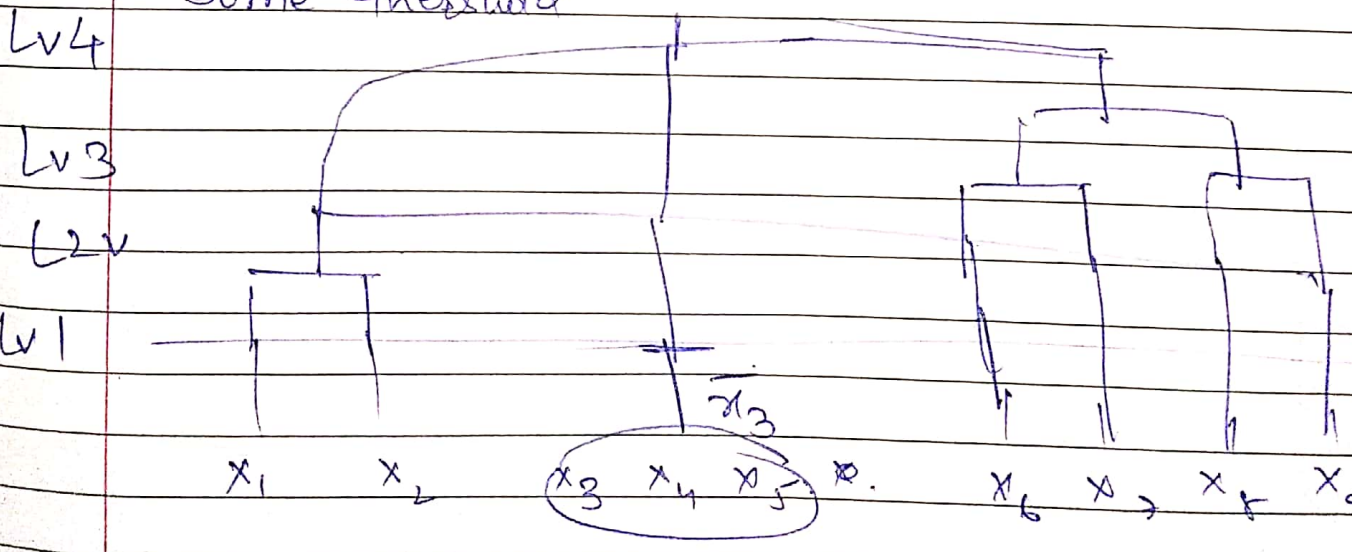


which feature vector are closest of each other

	x_1	x_2	x_3	---	x_9
x_1	0				
x_2		0			
x_3			0		
⋮					
x_9					0

We have this distance matrix

Now we will find which are closest & less than some threshold



Now we find New distance matrix with $\bar{x}_3$

by euclidian distance the clustering is done

for top-down

we say thy don't belong to same cluster if at some level if distance b/w mean vector are greater than that level threshold

we can say it is reverse of bottom-up