Beg. //

Cluster Analysis

It is a collection of data objects : simmilar within the same group , disimmilar objects in other groups.

Cluster analysis (Clustering)

The process of identifying these similarities and grouping similar data objects into clusters based on their charactaristics.

Clustering is a form of unsupervised learning because there are no predifined classes for the data points.

Typical Application

Insight : understanding data distribution.

Preprocessing : preparing data for other Algorithms.

k means → data into k clusters

DBSCAN → clusters based on d

Clustering as preprocessing Tool.

Summarization: Groups similar data ~~togth~~ together, making it easier to work with large datasets.

- helps regression, PCA, Classification by reveling patterns within groups.

Compression: Image processing → vector quantization. to reduce img size

K-nearest neighbors: ~~Als helps~~ clusters helps quickly finding. K-nearest neighbors by focusing nearby clusters.

Clustering can help identify outliers.

~~#~~Good clustering

A good clustering method creates clusters with high intra-class similarity and low inter similarity. The quality of clustering depends on the similarity measure and the implementation of Algorithm and its ability to discover hidden pattern in the data revealing relationships.

Measure Quality of Clustering

Quality of Clustering accessed using disimilarity and Similarity matric.
↳ measured through distance function. denoted (d.J)

* Distance function depends on the
  interval-Scaled → temp, Boolean → yes, no, Categorical → Colors, ordinal data → ranking;

* Weights depends on Application

The Quality of clustering is typically evaluated using Quality function. that measure how well clustering. achives the desired grouping.

Euclidean distance - $d(g_1, g_2) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

Manhattan distance - $d(g_1, g_2) = \sum_{i=1}^{n} |(x_i - y_i)|$

Minkowski " = $d(g_1, g_2) = \sqrt[m]{\sum_{i=1}^{n} (x_i - y_i)^m}$

Major Clustering Approach

Partitioning Approach: Creating various partitions of the Data and evaluating them using Criterion

methods → k means, k-medoids.

Hierarchical approach: Its creating tree like Structure by either merging Clusters or dividing them.

methods - Diana, Agnes

Density based on connectivity function.

methods — DBSACN

Grid based on multiple lvl granularity → Em, Som

model hypothesizes fits it to the data.

Link based on use relationship or links between objects → Linkclus.

Partitioning Algo

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} (p - c_i)^2$$

K-means → Each cluster represents by center of cluster

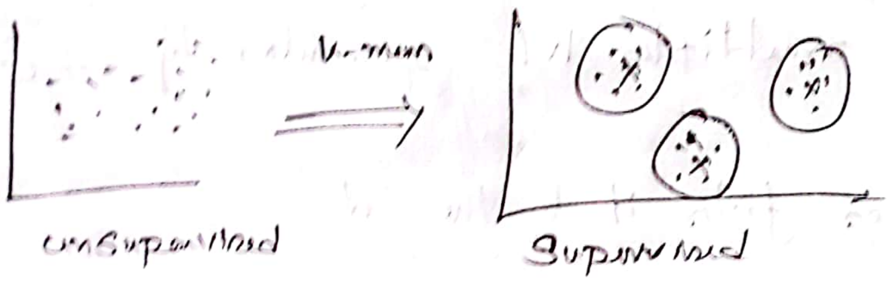
K-medoids → Each " " " one of the objects in the cluster

K-means (unsupervised)

↳ center of cluster.

K → no. of pre-defined cluster

K = 2 রিম 2 Clusters.



x - Centralized

* Elbow and Silhouette method used for K value detection

k means Example

$A_1(2,10)$ $A_2(2,5)$ $A_3(8,4)$ $B_1 = 5,8$ $B_2 = 7,5$ $B_3 = 6,4$

$C_1 = 1,2$, $C_2 = 4,9$

A_1 B_1 C_1 → Center [using Euclidean distance]

⇒ Euclidean distance $d(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$.

$\sqrt{(2-2)^2 + (10-10)^2} = 0$ same zero

	Data (x_1)	(y_1)	Center A_1 (x_2 = 2, y_2 = 10)	B_1 (5, 8)	C_1 (1, 2)	Cluster ↳ 3 টার মধ্যে smaller যত নম্বর	New cluster
A_1	2	10	0	3.61	8.06	1	
A_2	2	5	5	4.24	3.16	3	
A_3	8	4	8.49	5	7.28	2	
B_1	5	8	3.61	0	7.21	2	
B_2	7	5	7.07	3.61	6.71	2	
B_3	6	4	7.21	4.12	5.39	2	
C_1	1	2	8.06	7.21	0.0	3	
C_2	4	9	2.24	1.41	7.62	2	

New cluster:

A_1 = 1 নং ক্লাস্টারে আছে 5টা

A_1 1 only

$x_1 = 2$, $y_1 = 10$

B_1 = 2 5টা

x_1 এর $\frac{8+5+7+4+6}{5} = 6$

এর for y_1 $\frac{30}{5} = 6$

$B_1 = (6, 6)$

$C_1 = (1.5, 3.5)$

Same way

নতুন center -

$A_1 = (2, 10)$

$B_1 = (6, 6)$

$C_1 = (1.5, 3.5)$

এভাবে same way তে যখন cluster গুলো same আসবে তখন আমরা বুঝবো যে আর change আসবেনা iteration বন্ধ করবো new center বের করে

Step 2

	A_1 (2,10)	A_2 (2,5)	A_3 (8,4)	B_1 (5,8)	B_2 (7,5)	B_3 (6,4)	C_1 (1,2)	C_2 (4,9)
$A \rightarrow$	0	5	8	3	7	7	8	2
$B \rightarrow$	5	4	2	2	1	2	6	3
$C \rightarrow$	6	1	6	5	5	9	1	6
	1	3	2	2	2	2	3	1

Center

$A_1 = 2, 10$

$B_1 = 6, 6$

$C_1 = 1.5, 3.5$

ⓢ এখানে যেটা হবে smaller সেটা নিবো

আগে নিবো

② বার, So change হয়ে গেছে আবার করা লাগবে

new Center

A_1 = 1 ও 8 এর মান = $\frac{2+4}{2} = 3$, $\frac{19}{2} = 9.5 \Rightarrow (3, 9.5)$

$B_1 = (6.5, 5.25)$

$C_1 = (1.5, 3.5)$

$$D_3 = \begin{bmatrix} 1.2 & 4.61 & 7.43 & 2.5 & 6. & 6 & 7 & 1 \\ 6 & 4 & 3 & 5 & 1 & 3 & 6 & 4 \\ 6 & 1 & 6 & 5 & 5 & 4 & 1 & 6 \end{bmatrix}$$

$$\begin{array}{cccccccc} 1 & 3 & 2 & 2 & 2 & 2 & 3 & 1 \end{array}$$

mil নাই Next iteration

New Centroids

$A_1 = 3.67, 9$

$B_1 = (7, 4.33)$

$C_1 = (1.5, 3.5)$

$$D_4 = \begin{array}{cccccccc} A_1 & A_2 & A_3 & B_1 & B_2 & B_3 & C_1 & C_2 \end{array}$$
$$\begin{bmatrix} 1 & 4 & 6 & 1 & 5 & 3 & 7 & 0 \\ 7 & 5 & 1 & 4 & 0 & 1 & 6 & 3 \\ 6 & 1 & 6 & 5 & 5 & 4 & 1 & 6 \end{bmatrix}$$

$$\begin{array}{cccccccc} 1 & 3 & 2 & 1 & 2 & 2 & 3 & 1 \end{array}$$

মিলে গেছে আর তাই থামা হল

① $\rightarrow A_1, C_2, B_1$

② $\rightarrow A_3, B_2, B_3$

③ $\rightarrow A_2, C_1$

Am

K medoids Clustering

Its identifies representive object in clusters

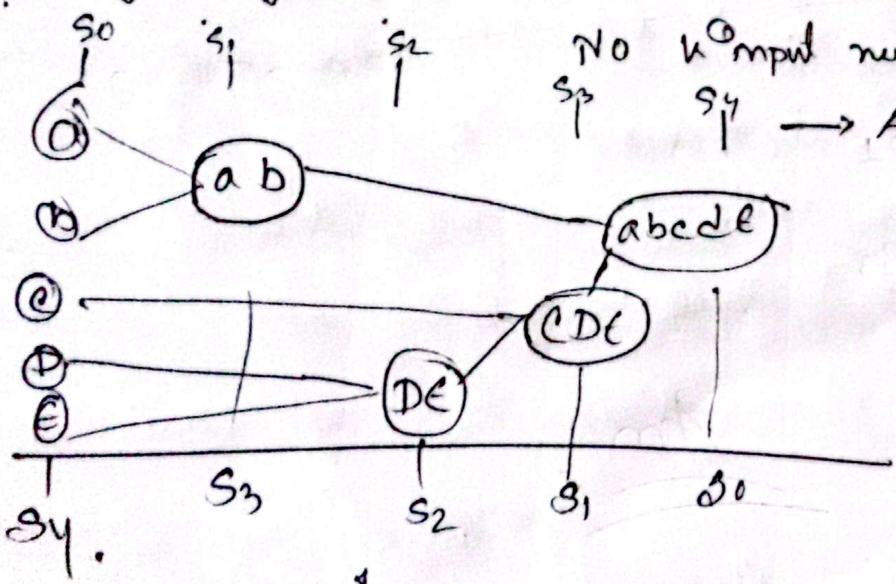
PAM - (partitioning around medoids) starts with initial medoids and iteratively replace one with a non-medoid if it reduces the total distance

CLARA → improved efficiency in large dataset
PAM samples

Hierarchical Clustering

Its builds a hierarchy of cluster by either iteratively merging smaller clusters into large one or spilitting larger clusters into smaller ones.

No k input need.



→ AGNES
single link approach
merge closest nodes

→ DIANA
inverse of AGNES
Iteratively spits all data in ...

Denodogram

(অনেক নতুন (অনেক data আছে) —

Density Clustering

Its focusing on grouping data points based on local density. its handle noise, one scan, discover clusters.

Key methods

DBSCAN - can identify noise an outliers.

Optics - extends DBSCAN

Density based spatial clustering application noise.

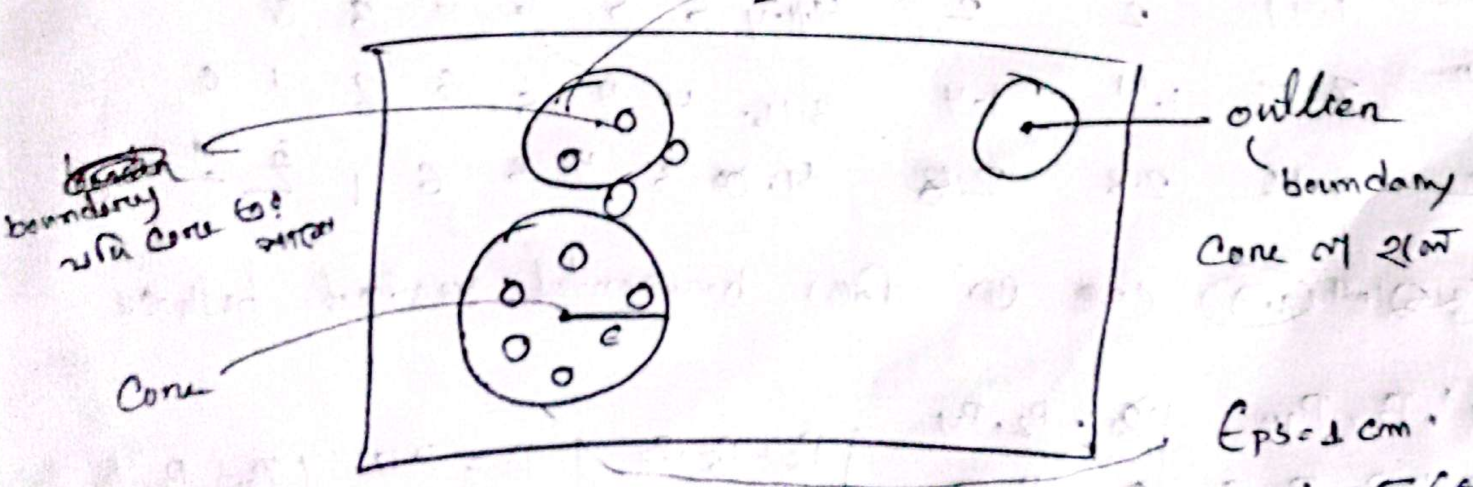
DBSCAN Parameters:

Eps (ε) = minimum ~~point~~ distance around a point to look for other points.

কিসের circle হবে তার মাপ

Minpts = Minimum no of points required within Eps distance for a point to be considered.

(minimum ৩টা circle থাকবে)



Eps = 1 cm

minpts = 5 (5 টা থাকবে)

DBSCAN Math

$\epsilon = 1.9$, minpts = 4.

$P_1 = (3,7)$ $P_3 = (5,5)$ $P_5 = (7,3)$ $P_7 = (7,2)$ $P_9 = (3,3)$ $P_{11} = (3,5)$

$P_2 = (4,6)$ $(P_4) = 6,4$ $P_6 = (6,2)$ $P_8 = (8,4)$ $P_{10} = (2,6)$ $P_{12} = (2,4)$

use equladian : $\left(\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}\right)$

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}
P_1	0											
P_2	1.41	0										
P_3	2.83	1.41	0									
P_4	4.24	2.83	1.4	0								
P_5	5.66	4.24	2.8	1.4	0							
P_6	5.83	4.47	3.6	2	1	0						
P_6	6.40	5	3.1	1.4	1	1	0					
P_8	5.83	4.47	2.8	2	1	2	2	0				
P_9	4.00	3.16	3.1	3.9	4	3	4	3	0			
P_{10}	1.41	2	2	4.4	5	5	6	6	3	6		
P_{11}	2.00	1.4	1.4	3.16	4	4	5	5	2	1	0	
P_{12}	3.16	2.8	2.8	4.00	5	4	5	6	1	2	1	0

[illegible] (1.9) কম হবে নিবো। horizontal ও vertical both ck.

P_1 : P_2, P_{10}
P_2 : P_1, P_3, P_{11}
P_3 = P_2, P_4
P_4 = P_3, P_5
P_6 = P_5, P_7
P_7 = P_5, P_6
P_9 = P_{12}
P_{10} = P_1, P_{11}
P_{11} = P_2, P_{10}, P_{12}
P_{12} = P_9, P_{11}

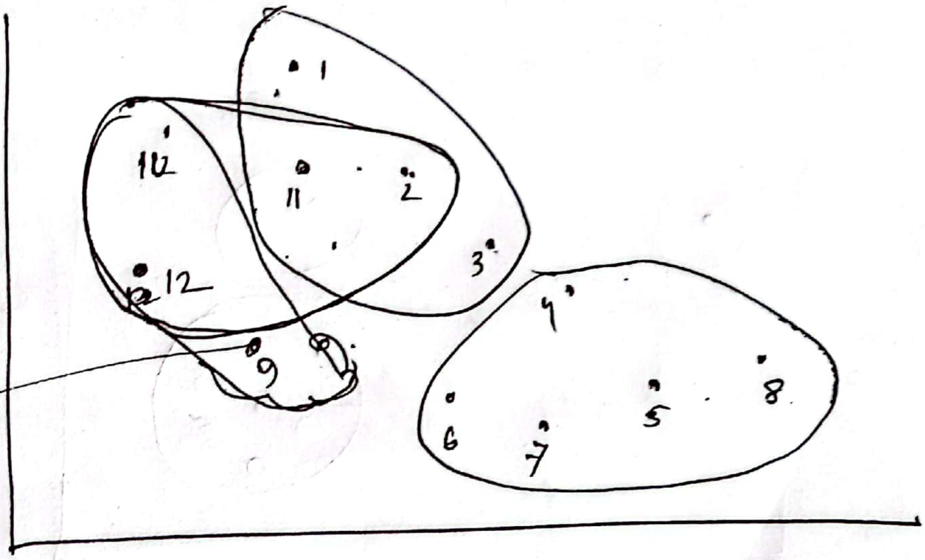
points		status
P_1	border	P_1, P_2, P_{10} (3) টি Noise
P_2		P_2 P_1 P_3 P_{11} (4) same Core
P_3	border	Noise
P_4	border	Noise
P_5		core
P_6	border	Noise
P_7	border	Noise
P_8	border	"
P_9		"
P_{10}	border	"
P_{11}		core (4 টি আছে)
P_{12}	border	Noise

* যদি midpoint এর data হয় then noise অর্থাৎ core

④ যেগুলো Noise তার কোন data core থাকলে তাহলে Noise ও border হবে

P_9 ও P_{12} আছে যা core point না

৫ টা আছে



only noise

Clustering Application

Biology - classifies organisms into hierarchical categories

Information Retrieval - Group document by topic for improve search and org.

Land use : identifies areas with land use patterns

Marketing : Segment customers based on behavior

Earthquake Studies : Loc of earthquake plate bound. areas where most are happen

Climate Studies : find patterns of weather and understand climate change