PAGE No.
DATE

Ensemble learning

Ensemble learning is a machine learning paradigm that combines the predictions of multiple model to improve overall performance & generalization.

Example :- Pappu pass ho gaya (Detail Explanation)

→ Why Ensemble learning ?

1). to overcome limitations of individual models by leaveraging their collective intellegence

2). Reduce overfitting & enhance robustness.

3). Improved performance on Imbalanced Datasets.

5). Enhanced Generalization
↳ By combining diverse model throgh various ensemble methods

Plus
12.26 18:33

model enables better generalization as model focuses on different aspects of data.

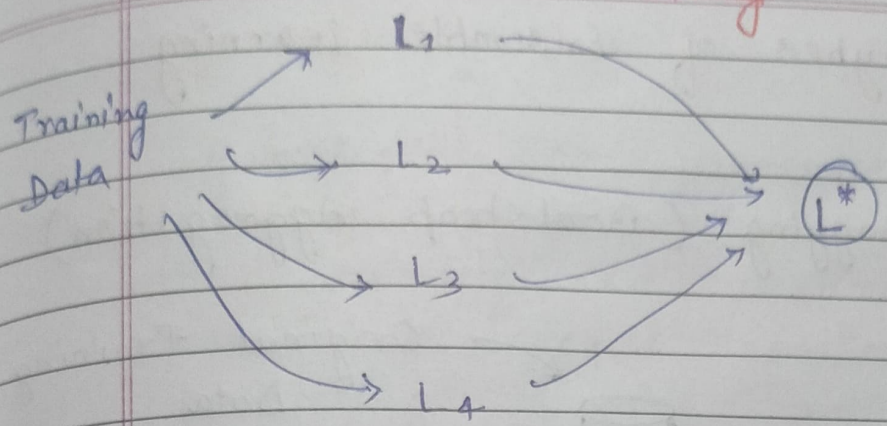
6). Increased model Accuracy

→ Base models: It constituent models that form the ensemble (Eg decesion tree, neural network etc).

→ Aggregation methods: Techniques for combining predictions (Eg voting, Averaging, stackingetc

In short, Ensemble learning is a powerful strategy to enhance model performance by Combining the strength of diverse base model & Understaning the trade-offs & selecting appropriate ensemble methods for effective predictions.

Ensemble learning



Base learners / model

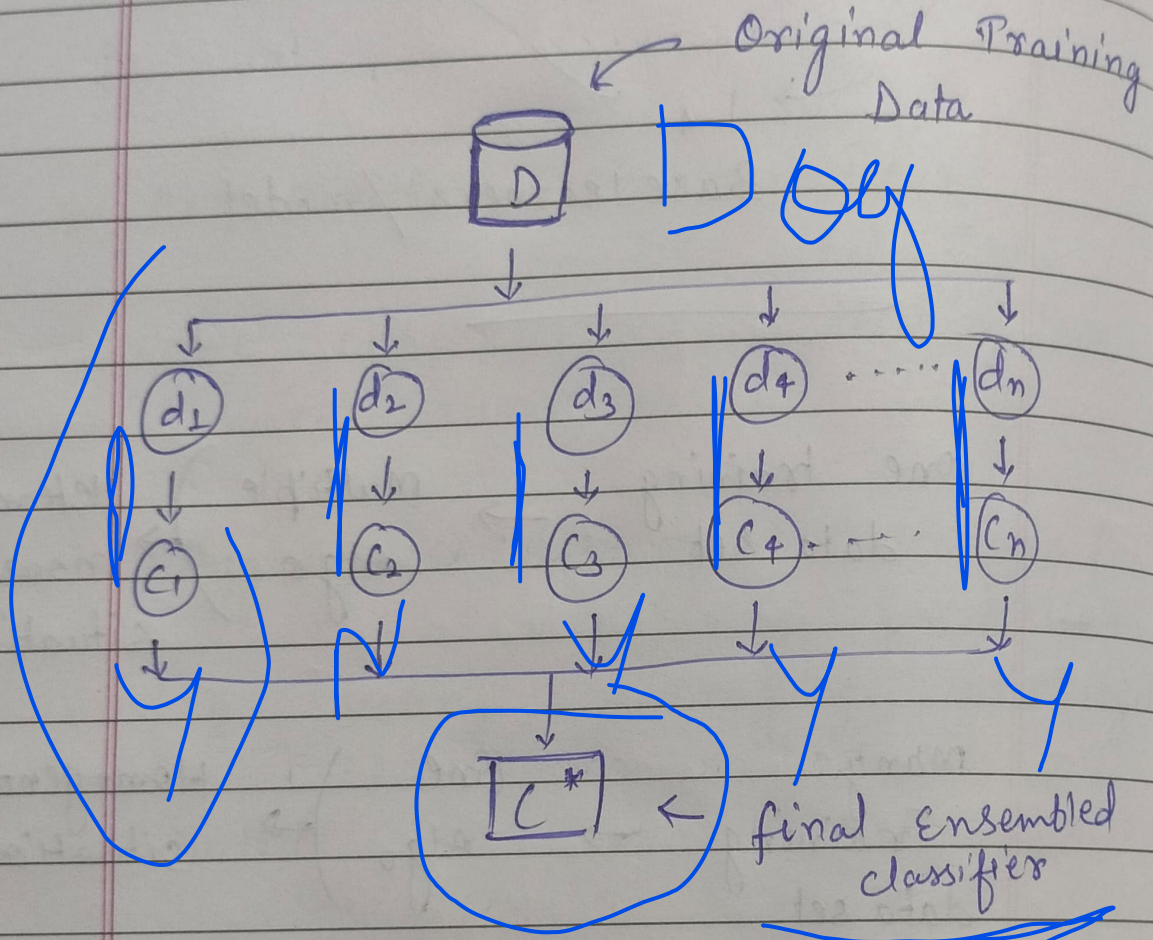
---

one training data set → Multiple Algo → Heterogenous situation

Many Training data set → One Algo → Homogenous situation

→ Types of Ensemble learning

<1>. Bagging (Bootstrap Aggregation)



Random sampling with replacement

Multiple dataset are also known as Bootstraped Sample

these samples are generated so that we can train the classifiers based on these samples.

By Defination →

Bagging is an Ensemble learning technique that aims to improve the stability and accuracy of machine learning algorithms by constructing multiple models in parallel on different random subsets of the training dataset

Process →

1). Bootstrap sampling :-

Randomly draw, with Replacment, multiple subsets (bags) from the original training dataset.
Same size subset, but it may contain repeated or missed data.

2). Model Traing :-

Train base model independently on each bootstrap sample.

3. Prediction Aggregation :-

Aggregate prediction from all model to make final prediction

E.g. Voting, Averaging etc.

→ Advantages

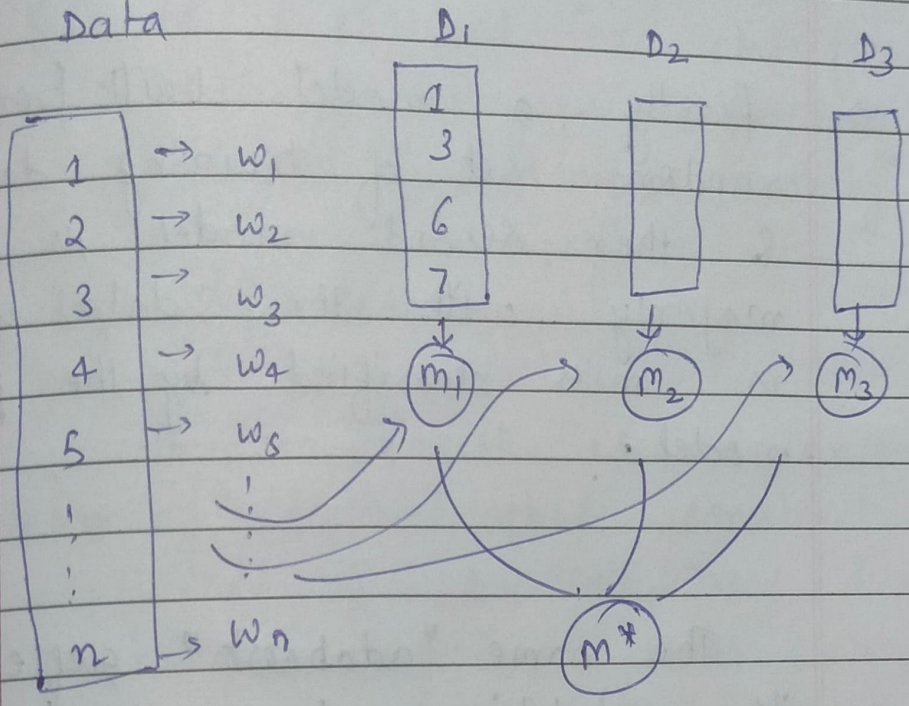
1. Reduce varience
2. Mitigate overfitting

This Bagging method is well suited for parallel processing, as base models can be trained independently on different subsets simultaneously.

→ Boosting

Boosting is a family of ensemble ML Algos where a series of weak learns are trained sequentially with each subsequent model focusing on correcting the errors made by previous one. final Prediction is weighed combination of predictions of all.

→ Boosting - Adaboost



Boosting is also a homogeneous weak learner's model same as bagging but it works differently from bagging.

In this model learner learn sequentially (instead of parallel as bagging) & adaptively to improve model prediction of a learning algorithm.

Here a strong classifier is achieved by using weak models in series.

⇒ firstly, a model built from random set of training dataset & then second model is built majorly with the data which is miss classified by the previous model.

The name "adaboost" reflects its adaptive nature; it adapts to the weaknesses of the previous models by giving more weight to miss classified data/instances.

→ Strength:

1). High accuracy due to sequential learning

2). versatility: applicable to various types of weak learners.

→ limitation :-

1). sensitive to noisy data

2) overfitting

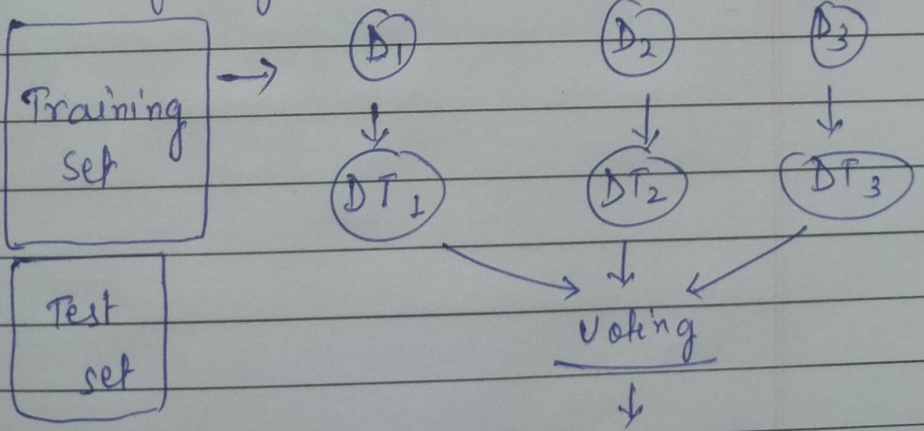
→ Random forest

It belongs to supervised learning & It can used for classification as well as Regression problems. It uses "Ensemble learning"

By Defination :-

Random forest is an ensemble learning method that constructs a multitude of decesion Trees during training & then takes average of all outputs to improve the predictive accuracy of that dataset.

1. The greater number of trees in the forest leads to higher accuracy & prevents the problem of over fitting.



PAGE No.
DATE

Clustering

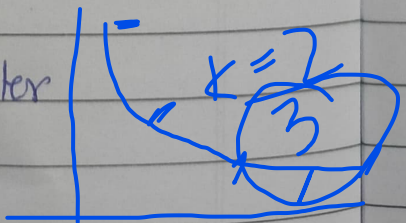
Clustering is a type of unsupervised learning in machine learning that involves grouping similar data points together based on certain criteria. The goals is to discover inherent pattern or structure within the data without prior knowledge of the class labels.

Eg: K- means clustering
DB Scan
Hierarchical clustering

→ K- means clustering

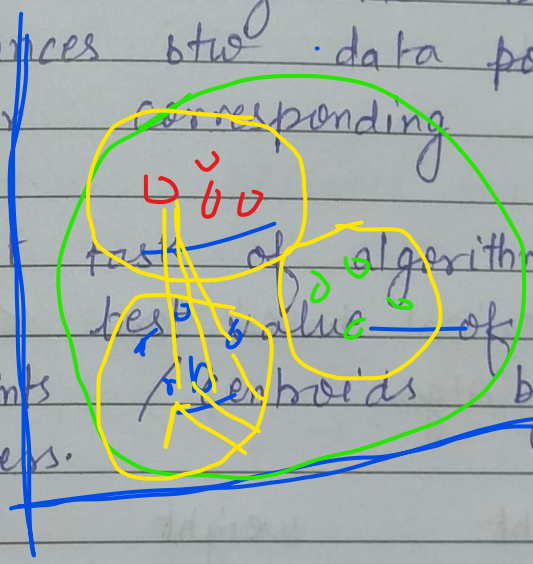
K- means clustering is an unsupervised machine learning algorithm which groups the unlabeled dataset into different clusters.

K- Predefined cluster



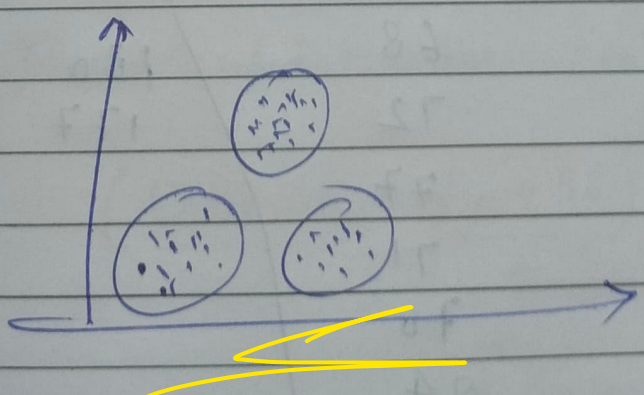
* It is an iterative algorithm that divides the unlabeled dataset into K clusters in such way that each instance of dataset belongs only one group that has similar properties.

→ It a centroid based algorithm the main goal of this algorithm is to minimize the sum of distances btw data point & their corresponding cluster.



→ first task of algorithm is —: to find the value of k center points / centroids by iterative process.

→ Assign each datapoint to its closet cluster.



→ steps involved in K-means clustering

1). Decide n clusters.

2). initilize centroids

3). assign cluster

4). move centroids

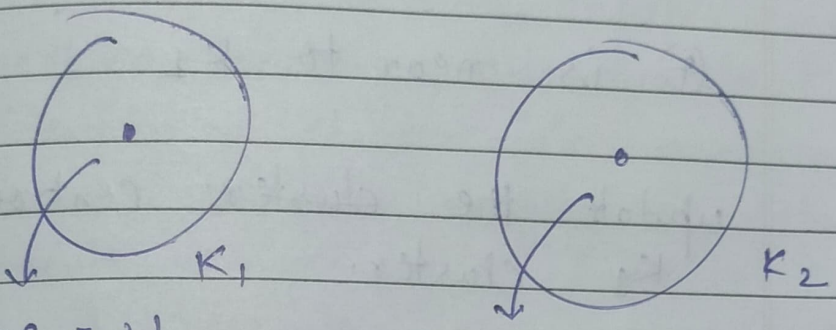
finish

→ Numerical based on K-means clustering Algo

Height	weight
185	72
170	56
168	60
179	68
182	72
188	77
180	71
180	70
183	84
180	88
180	67
177	76

Euclidian Distance

$$\sqrt{(x_0 - x_c)^2 + (y_0 - y_c)^2}$$



(185, 72)

(170, 56)

Euclidian Distance for (3)

ED to 3 K_1 → 20.00

↳ K_2 → 4.48

(3) near to K_2
it will go in K_2 cluster.

Now update the centroid of K_2

$$\left(\frac{170+168}{2}, \frac{60+56}{2}\right) = (169, 58)$$

↑
New centroid for K_2 cluster

PAGE No.
DATE

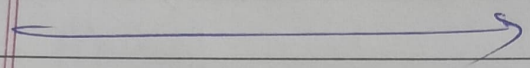
similarly for ④

$K_1 \rightarrow 6.32$

$K_2 \rightarrow 14.14$

④ is near to K_1

update the cluster centroid of K_1 cluster.



final output

$K_1 = \{1, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$

$K_2 = \{2, 3\}$

DB scan cl
(Dens

DB Sca
machine
identifies
based
points.
Unlik
not he
number
& can
arbitra

DB
on thi
& no

each
neigh
con t
numb

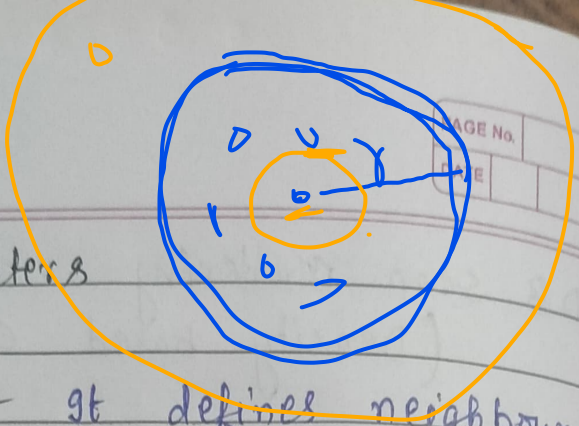
DB Scan Clustering
(Density based clustering)*

DB Scan is an unsupervised machine learning algorithm that identifies clusters in a dataset based on the density of data points.

Unlike K-means, DB Scan does not require specifying the number of clusters beforehead & can discover clusters of arbitrary shaps.

DB Scan algorithm is based on these intutive notion of clusters & noise.

The key idea is that for each point of a cluster, the neighbourhood radius has to contain at least a minium number of points.



* Parameters

(1). eps :- It defines neighbourhood around a data point that is if the distance btw two datapoints is <= eps than they are considered neighbours.

Small eps → Outliers ↑

large eps ↑ → merging of clusters. ↓

thats why to find eps we use K- distance graph.

(2). minpts :- Minimum number of neighbours (data points) with in eps radius.

large Dataset ↑ → choose large minpts ↑

in general

minpts >= D+1

where D is dimensions of Dataset

⇒ 3 types of Data point

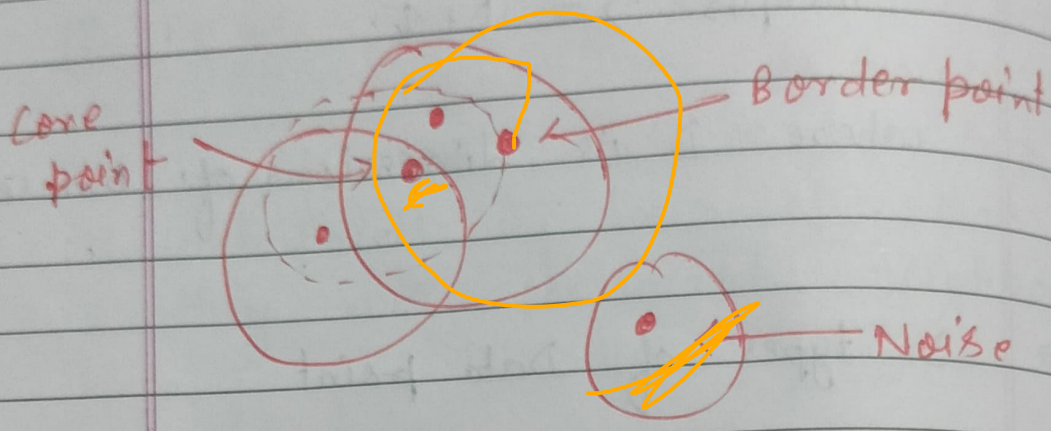
1. Core point : A point is corepoint if it has more than minpts within eps.

2. Border point :- A point which has fewer than minpts within eps but it is neighbour of core point

3. Outlier / Noise : Not a core point nor border point.

PAGE No.
DATE

Numerical Based on DB Scan



is a point a from any
Directly Density Reachable ? point

Sol": 2 Conditions (take two point p & q

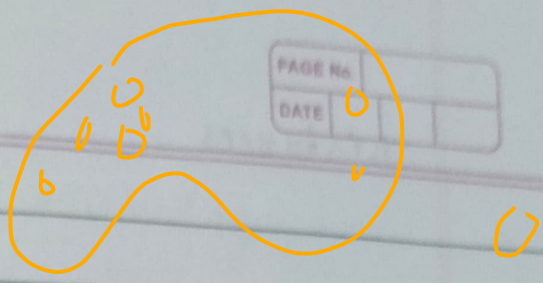
1.) q should neaghbour of p

2). p must be a core point

Advantages of DB Scan

(1). Robust to outlier

(2). flexible in cluster shape



Limitations

1. Difficulty with varying Density
2. sensitive to parameter selection (choice of eps & minpts)

→ Numerical

Data Points

Ques:

$P_1 = (3,7)$ $P_2 (4,6)$ $P_3 = (5,5)$
$P_4 = (6,4)$ $P_5 (7,3)$ $P_6 = (6,2)$
$P_7 = (7,2)$ $P_8 (0,4)$ $P_9 = (3,3)$
$P_{10} = (2,6)$ $P_{11} (3,5)$ $P_{12} = (2,4)$

minPts = 4

& epsilon (ε) / eps = 1.9

Distances

	P_1	P_2	P_3	P_4	P_5	P_6	P_7 ...
P_1	0	—	—	—			
P_2	1.41	0	—	—			
P_3	2.83	1.41	0	—			
P_4	4.24	2.83		0			
P_5	5.66						
P_6	5.83						
.....	!	!					
P_{12}	3.16						

go like this

Step 1: Calculate Distance of each datapoint with other Datapoint

Step 2: for each point, figure out all the distances with datapoints which are <= eps.

Step 3: Now Categorize all points as core / border / Noise accordingly

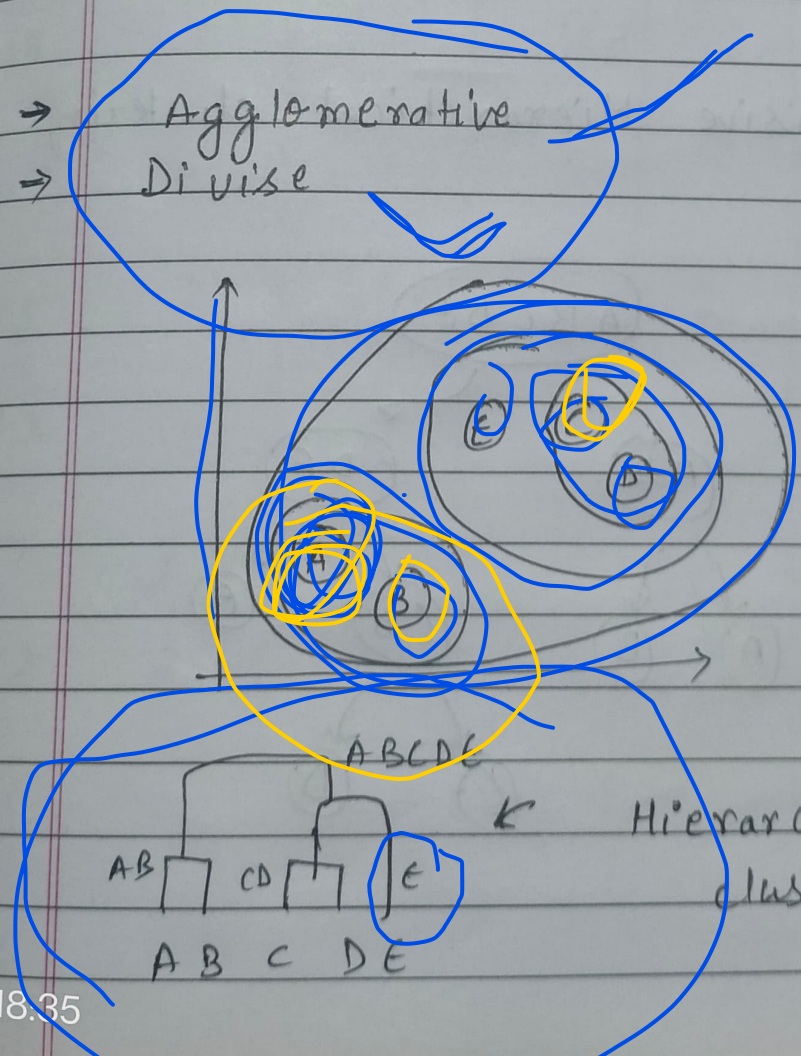
Get detailed Explanation in video

Hierarchical clustering

Hierarchical clustering is an unsupervised clustering algorithm that organizes data into ~~hierarchical~~ hierarchy of nested clusters.

The primary goal is to create a tree like structure (known as dendrogram) that visually represents the relationships btw data points & clusters.

→ Agglomerative
→ Divise



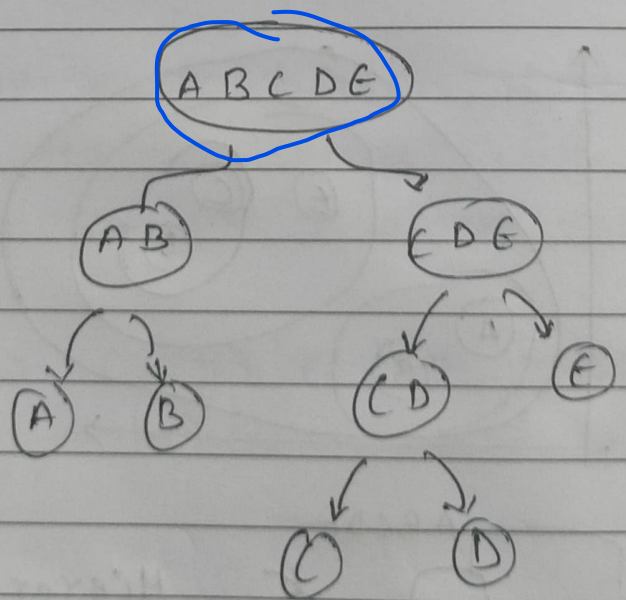
← Hierarchy of clusters

- Agglomerative Hierarchical clustering

It is bottom up approach in which the algorithm starts with taking all data points as single cluster & then merge them until they form one single cluster.

Dendrogram

- Divisive Hierarchical clustering



Divisive Algorithm is just opposite of agglomerative Hierarchical clustering.

It is top-down approach.

Single Linkage

Complete Linkage 23,45 34,34

Numerical

18, 22 , 35 ,(42,43),27

18, 22, 25, 42, 27, 43

step 1:

	18	22	25	42	27	43
18	0	4	7	19	24	25
22	4	0	3	5	20	21
25	7	3	0	2	17	18
42	19	5	2	0	15	16
27	24	20	17	15	0	1
43	25	21	18	16	1	0

Create cluster of 42 & 43

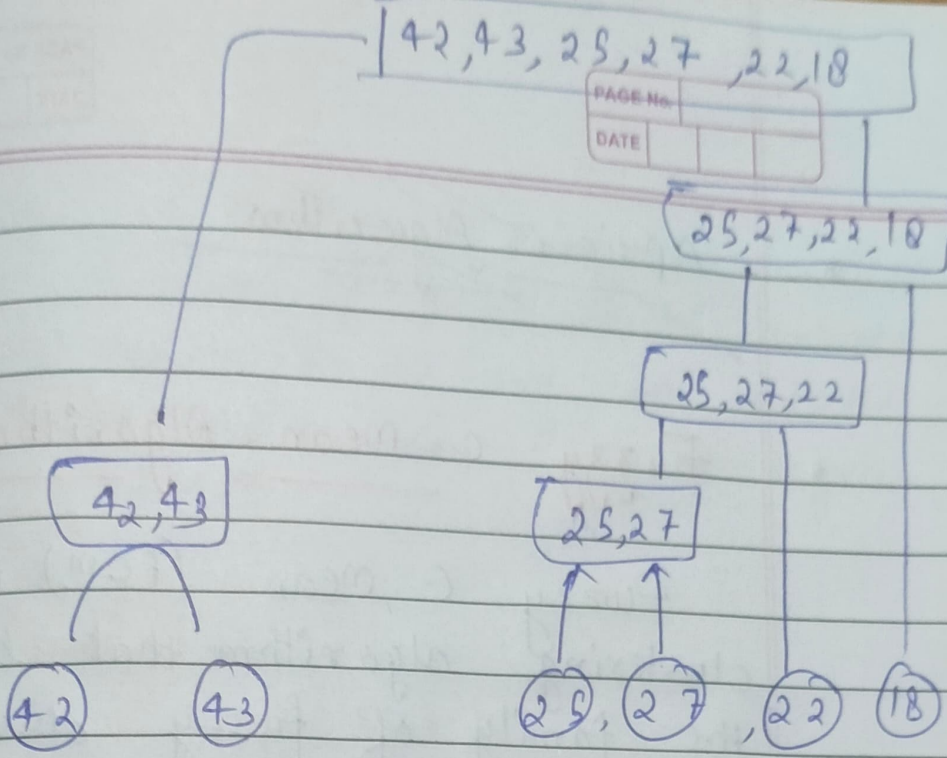
remove the 43 row
& now will have
10, 22, 25, 27, (42, 43)

	18	22	25	27	42,43
18	0	4	7	9	24
22	4	0	3	5	20
25	7	3	0	2	17
27	9	5	2	0	15
42,43	24	20	17	15	0

Create new cluster of 25, 27

full Explanation in video

((42,43), (((25,27), 22), 18))



why Hierarchical clustering?

complete linkage formula

Single linkage

minimum Distance

max Distance

PAGE No.
DATE

* Apriori Algorithm

• Fuzzy C-Mean Algorithm

Fuzzy C-Mean (FCM) is a clustering algorithm that belongs to the family of fuzzy clustering methods.

It is an extension of traditional K-means clustering, allowing data points to belong to multiple clusters with varying degree of membership rather than strictly assigning them to a single cluster.

Fuzzy clustering is particularly very useful when there is uncertainty or ambiguity in the assignment of point to cluster.

Use cases!-

1). Widely used in pattern recognition, image segmentation & data mining

→ One of the key parameter in FCM is fuzziness (m) which controls the degree to belong to multiple clusters with more Equal degrees.

⇒ Numerical

video Link

clustering

Hard clustering

- K-mean clustering

Soft clustering

- fuzzy C-mean

Numerical steps

(1). Given data points, based on number of clusters required initialize the membership table with random value

Note: But after all the Σ membership probability of each datapoint must be

PAGE No.
DATE

* Associate Rule learning

Associate Rule learning is a machine learning technique used to discover interesing relationship or associations among a set of variable ing large dataset.

The relationships are often expressed in the form of rules where one set of items implies another set of items some important concepts & algorithms related to association Rule are

1. Support
2. Confidence
3. Lift
4. Conviction

Rules

Assosiation Rules

1. Shoes - > socks

2. socks -> shoes

Shoes - > socks

=> support : Support is the proportion of transection in the dataset that contain a particular set of item.

Confidence is the measure of reliability of a rule.

Given: Min support = 50%.
Threshold confidence = 70%.

eg.

Tid	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Itemset	Support
1	2/4 → 50%
2	3/4 → 75%
3	3/4 → 75%
4	1/4 → 25%
5	3/4 → 75%

Remove all the items from itemset having support less than min support

So new itemset is

{1, 2, 3, 5}

(ii)

	Itemset	Support
		1/4 → 25%
X	(1,2)	2/4 → 50%
	(1,3)	1/4 → 25%
X	(1,5)	2/4 → 50%
	(2,3)	3/4 → 75%
	(2,5)	2/4 → 50%
	(3,5)	

again remove all item having support less than min support

(1,2 3,5) biplet

X	(1,2,3)	1/4 — 25%
	(2,3 5)	2/4 — 50%
X	(1,2,5)	1/4 = 25%

{2,3,5}

Now building Association Rules

Rule	Support	Confi.
(2 ^ 3) → 5	2	100%
(2 ^ 5) → 3	2	66%
(3 ^ 5) → 2	2	100%

$2 \rightarrow (3 \wedge 5)$	2	66%
$5 \rightarrow (2 \wedge 3)$	2	66%
$3 \rightarrow (2 \wedge 5)$	2	66%

$$\text{confidence} = \frac{S(A \cup B)}{S(A)}$$

Eg: $\underset{A}{(2 \wedge 3)} \rightarrow \underset{B}{5}$

$$= \frac{S(\overset{A}{2 \wedge 3} \cup \overset{B}{5})}{S(\underset{2 \wedge 3}{A})} = \frac{2}{2} = 100\%$$

Compare these Rules confidence with threshold confidence

those Rules, with confidence >= threshold confidence are association rules.

hence $(2 \wedge 3) \rightarrow 5$

$(3 \wedge 5) \rightarrow 2$