DETAILED LECTURE NOTES

Lecture. 1

Campus: PCE Course: B.Tech Class/Section: ..... Date: .....
Name of Faculty: UPMA kumari Name of Subject: ML Code: .....

Introduction:- machine learning is a subset of artifical intelligence in the field of computer scienc that often uses statistical techqiues to give computers to ability to learn with data, without being explicity programmed. in the past decade machine learning has given us self-driving cars, practical speech recongition, effective web search and a improved understanding of a human genome.

machine learning task:-

machine learing task are typically clanified into two three part.

machine learning is an application of AI that provides systems the ability to automatically learn and improve from experience without being explicity programmed) ML focues on the development of computer programs that can access data and use it.

(1) Supervised learning :- the computer is presented with example input and desired output and the goal is to learn a general rule that maps inputs to outputs.

As special cases, the input signal can be only partially available, or restricted to special feedback.

(2) Semisupervised :- the computer is given only an incomplete training signal, a training set with some of the target output missing.

(3) Active learning :- the computer can only obtain training labels for limited set of instances (based on a budget) and also has to optimized it's choice to acquire label for. when used interactively these can be presented to the user for Labeling.

(4) Reinforcement learning :- training data is given only as feedback to the program action in a dynamic environment such as driving a vehicle, or playing a game against an opponent.

DETAILED LECTURE NOTES

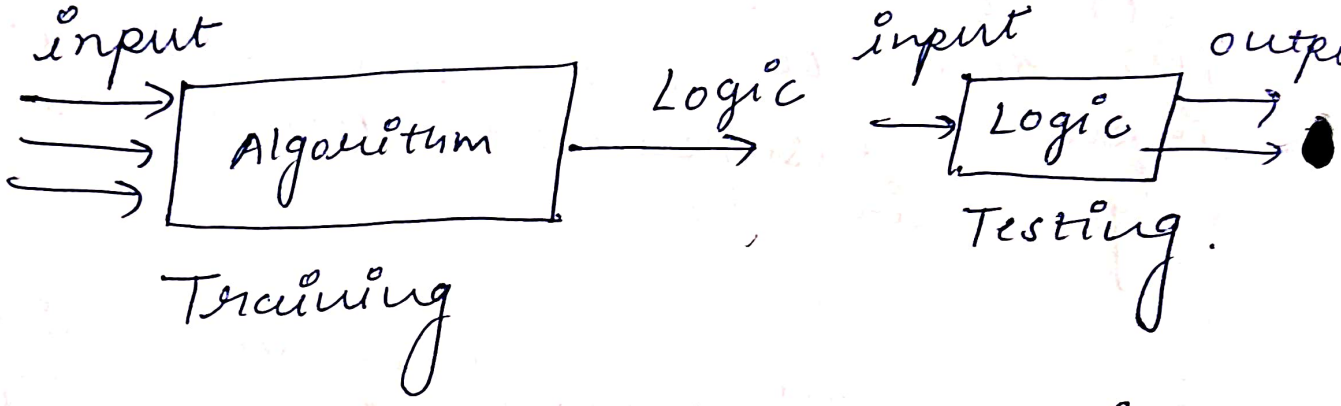
Campus: ..... Course: ..... Class/Section: ..... Date: .....
Name of Faculty: ..... Name of Subject: ..... Code: .....

(5) unsupervised learning :- No Label are given to the learning algo. learning it on it's own to find structure in it's input. unsupervised learning can be goal in itself.

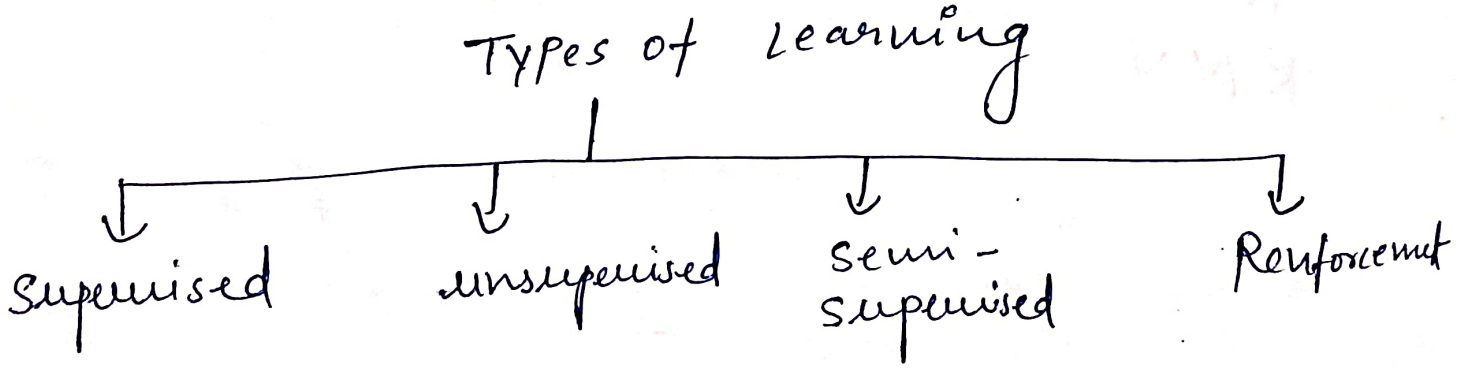
Supervised learning	unsupervised	instance based learning.
• find - s - algo.	EM algo.	Locally
• candidate elimination algo	k-mean algo.	weighted Regression algo
• Decision tree algo.		
• Back propagation algo.		
• Naive Bayes Algo.		
• KNN.		

definition of ML:- ML is one of the mo technologies that one would have come across and ability to learn ML is actually being used today

② supervised learning :-



what is learning for a machine:-
A machine is said to be learning from past experience (data feed in) with respect to some class to task. if it is perform in a given task improve with the experience.



DETAILED LECTURE NOTES

Lecture: 13

Campus: ..... Course: .....
Name of Faculty: .....
Class/Section: .....
Name of Subject: .....
Date: .....
Code: .....

Supervised learning:- supervised learning is when the model is getting trained on a labelled dataset. labelled dataset is one which have both input and output parameters. in this type of learning both training and validation dataset are labelled

Example:-

User id	Genders	Age	Salary	Purchased
156212	M.	Male	29	52000
—	—	—	—	—
156213		Female	21	—

→ classification

0
1
0

Fig. ①

Temperature.	pressure	Humity.	Label.
20°C	906	195.	mild direction.
13°C	120	195- --	8.2216
17°C			8.7216.

Regression.

both the above figure have label data set:-

- Figure A:- it is a dataset of a shopping store which is useful in predicting whether a customer will purchase a particular product under consideration or not based on his/her gender, age and salary.

input:- Gender, Age, salary.
output:- purchased i.e. 0 or 1 means yes the customer will purchase and 0 means customer won't purchase it.

- Figure B:- it is a meteorological dataset which serves the purpose of predicting wind speed based on different parameters

input:- Temperature, pressure, humidity, wind direction.
output:- wind speed.

** Training the system:- the training model data is usually split into (radio of 80:20, 80% as training data and 20% is Testing data. The model learns from training data only.

DETAILED LECTURE NOTES

Lecture:-4

Campus: ..... Course: ..... Class/Section: ..... Date: .....
Name of Faculty: ..... Name of Subject: ..... Code: .....

- Supervised → classification (defined labels)
  → Regression (no label defined)

Types of supervised learning:-

① classification:- it is a supervised learning task where output is having defined labels (discrete value). For example in Figure A. output:- purchased has defined labels 0 and 1, 1 means the customer will purchase and 0 means that customer won't purchase the goal here is to predict discrete value belonging to a particular class and evaluate on the basis of accuracy.

it can be either binary and multi clan classification, In binary classification Mo predict either 0 and 1. Yes or no. bu in case of multi clan classification, Mo predict more that one clan.

Example:- Gmail classifies mail in more than one classes like social, promotions, updates, forum.

(2) Regression:- it is a supervised learning task where output is having continous value

Example:- Figure B:- output:- wind speed is not having any discrete value but is continous in the particular range. the goal here is to predict a value as much closer to actual output value as our model can and then evaluation is done by calculating error value. the smaller the error the greater the accuracy of our regression model.

Example of supervised learning Algorithm:-

1. Linear Regression
2. Nearest Neighbor
3. Gaussian Naive Bayes.
4. Decision Trees
5. Support Vector Machine (SVM)
6. Random Forest.

DETAILED LECTURE NOTES

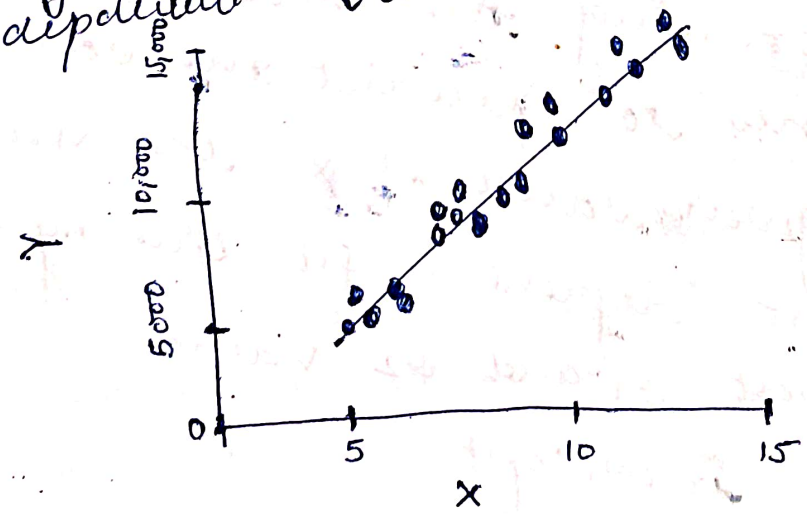
Lecture - 5

Campus: ..... Course: ..... Class/Section: ..... Date: .....
Name of Faculty: ..... Name of Subject: ..... Code: .....

Linear Regression:-

LR is a machine learning algorithm based on supervised learning. it performs a regression task. Regression model a target prediction value based on independent variable. it is mostly used for finding out the relationship between variable and forecasting.

Different regression model differ based on - the kind of relationship between dependent and independent variables. they are considering and the no. of independent variable being used.



linear Regression perform the task to pre... a dependent variable value (y) based on given independent variable (x). So this regression technique Find out a linear relationship between x (input) and y (output). Hence the name is linear Regression in the figure above X (input) is the work experience and Y (output) is the salary of a person. the regression line is the best fit line for our model.

equation of linear regression:-

$$Y = \theta_1 + \theta_2 . X$$

while training the model we are given:-
x:- input training data
y:- labels to data (supervised learning)
when training the model, it fits the best lines to predict the value of y for a given value of X. the model gets the best regression fit line by finding the best θ_1 and θ_2 values.

θ_1 :- intercept
θ_2 :- co-efficient of x

Campus: ..... Course: ..... Class/Section: ..... Date: .....
Name of Faculty: ..... Name of Subject: ..... Code: .....

The fundamental Naive Bayes assumption is that each feature make an:-

- independent
- equal

- we assume that no. pair of features are dependent. For example, the temperature being 'Hot' has nothing to do with the humidity or the outlook being Rainy has no effect on the mind. Hence, the
- feature are assumed to be independent.

Bayes' theorm finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorm is stated mathematically as the following equation.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

where A and B are event and P(B) ? 0

- Basically we are trying to find probabil of event A, given the event B is tru Event B is also termed as evidence.

- P(A) is a priori of A (the prior probabili i.e probability of event before evidence is seen) the evidence is an attribu value of an unknown instance (here it is event B)

- P(A|B) is a Posteriori probability of B probability of event after evidence is seen.

$$P(y|X) = \frac{P(X|y)\,P(y)}{P(X)}$$

where y is class variable, and x is a dependent Feature vector (of size n) where:-

$$X = (x_1, x_2, x_3 \ldots x_n)$$

$$X = (\text{Rainy}, \text{Hot}, \text{high}, \text{false})$$

$$Y = \text{NO}$$

find the best θ_1 and θ_2 values, we get the best fit line. So when we are finally using our model for prediction. it will predict the value of $\hat{y}$ for the input value of x.

How to update θ_1 and θ_2 values to get the best fit line?

Cost function:- (J):- By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predict value and true value is minimum. So. it is very important to update the θ_1 and θ_2 values, to reach the best value that minimize the error between predicted y value. and true (y) value.

minimize $1/n \sum_{i=1}^{n} (Predi - yi)^2$

$$J = 1/n \sum_{i=1}^{n} (Predi - yi)^2$$

Cost function J of linear Regression is the Root mean sequined Error (RMSE) between Predicted y value (Pred) and true value (y).

Gradient Descent :- To update θ_1 and θ_2 valu in order to reduce cost function (minimizi RMSE value) and achieving the best fit Lin the model uses Gradiet Descent. the ide is to start with random θ_1 and θ_2 values a then iteratively updating the values, reachi minimum cost.

(2) K-Nearest Neighbours :- K-nearest Neighbors is one of the most basic yet essential classificatio algorithms in machine learning. it belongs to be supervised learing domain an finds intense application in pattern recagitio data mining, instrusion detection.

Neayest Neighbor:-

- **Naive Bayes:-** Naive Bayes classifier are a collection of classification algorithm Based on Bayes theorm. it is not a single algorithm but a family of algorithm where all of them share a common principle. every pair of feature being classified is indepdent of each other

- Example:- Consider a fictional dataset that decribe the weather conditions for playing a game of golf. Given the weather conditions each tiple clanifies the condition as fit(yes) or unfit (No) for playing golf.

Example:-

	Outlook	Tempature	Humidity	windy	Play Golf
0	Rainy	HOT	High	False	NO
1	Rainy	HOT	High	True	NO
2	Sunny	Mild	High	True	Yes
3	Sunny	Cool	Normal	False	Yes
4	Rainy	Cool	Normal	True	NO
5	Rainy	Mild	High	True	Yes

the dataset is divided into two parts namely Feature matrix and the response vector.

- Feature Matrix:- Feature matrix contain all the vector (rows) of dataset in which each vector consist of the value of dependent feature. in above dataset Feature are outlook, Temperature, Humidity and windy.

- Response Vector contain the value of class variable (prediction or output) for each row of feature matrix. in above dataset, the class variable name is play golf.

So basicaly P(x/y) here means the probability of "Not playing golf" given that the weather condition are Rainy outlook, i.e Temperature is "not high humidity" and no. wind".

Now, it's time to put a naive assumption to the Baye's theorm. which is independm among the feature. So now. we split evidence into the independent parts.

If any two event A and B are indipdent then.

$$P(A, B) = P(A), P(B)$$

machine learning AI का part है।
AI computer को सीखना है। human की तरह
behave करना।
machine learning learn own methods.
ML is a part of data mining.
data mining and ML are work on data.

ML works on:-
→ analysing
→ Response
→ feel.

AI and ML works on ~~machine learning~~ pattern recognition
in pattern recognition are:-
- digit
- charater
- food
- face recognition.

clustering are use.

Key Terminales:-

(1) data can be transfer by information:-

data
|
csv ⟶
image ⟶ information.
Text ⟶

(2) problem solving Tool:-

(3) combination of CS, engineering, & statistics
↓ (application) (mathmatic)
pratical, programing description of algo.

(4) Interperets (adapt the data) data & act on it.

(5) optimize perfomen criteria using past experien

DETAILED LECTURE NOTES

Campus: ..... Course: ..... Class/Section: ..... Date: .....
Name of Faculty: ..... Name of Subject: ..... Code: .....

Types of Learning :-

1. Supervised learning.
2. Unsupervised learning.
3. Semi-supervised learning.

Supervised learning are teacher oriented learning system ko pehle se pta rahta hai output kya aane wala hai.

Unsupervised :- is mein no. idea of output

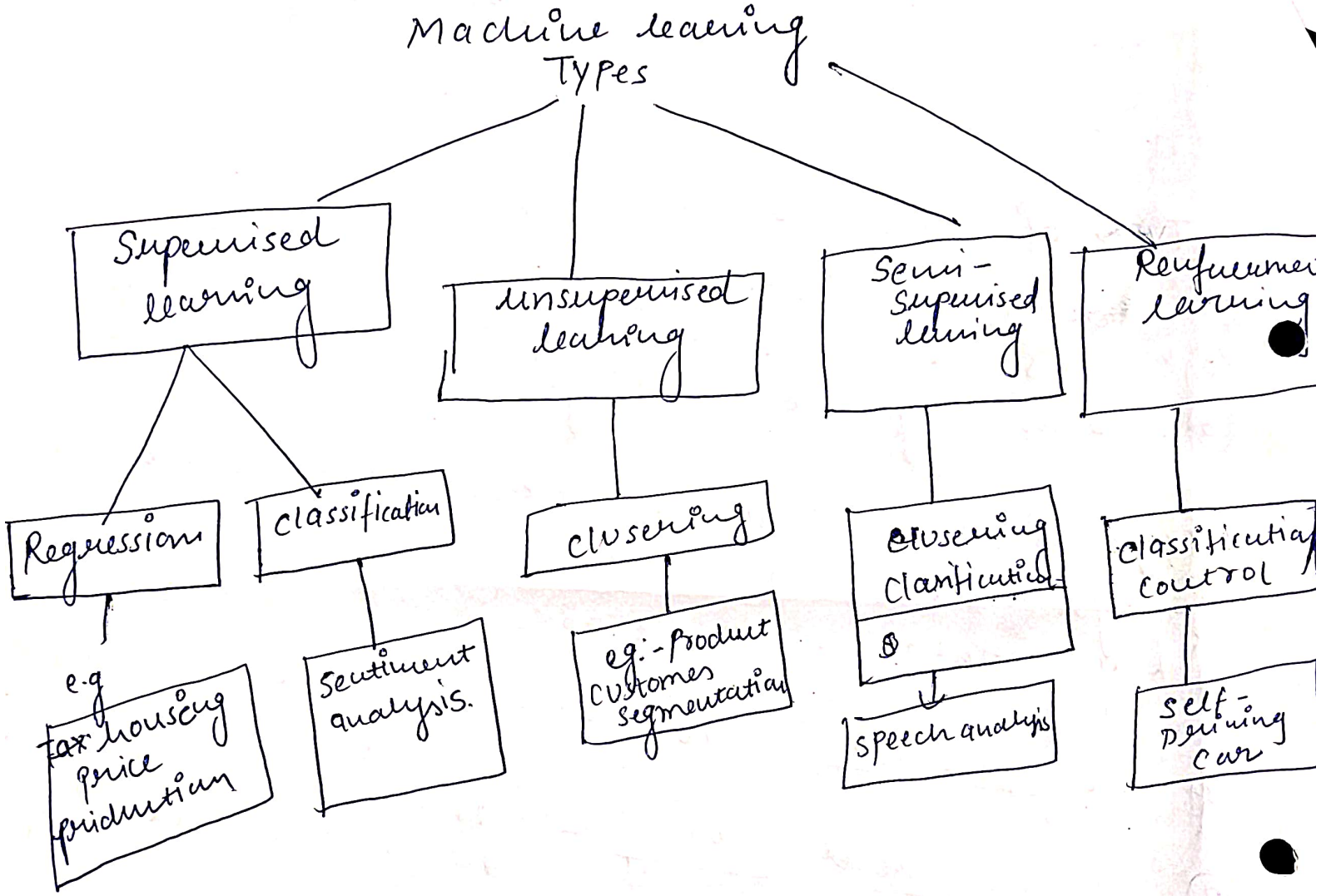
3. Reforcement learning :- : Reward based
   : feedback oriented.

Steps of ML :-

1. collect the data
2. prepare the input data (order example.csv)
3. analyzing the data
   → pattern-check
   → bounaries.
4. Train the algorithm (actual ML work)
5. Algo. Testing
6. use it in Real word application.

Key Terminology:-

① Expert system – Domain specific



DETAILED LECTURE NOTES

Campus: ..... Course: .....
Name of Faculty: .....
Class/Section: .....
Name of Subject: .....
Date: .....
Code: .....

① Supervised learning:- in this learning the computer will get the labelled input and the desired output

② Unsupervised:- in this learning computer will get the input without the desired output. The main aim of this model is to find the structure in the input.

③ Semi-supervised learning:- in this learning is a class of Machine learning task and Techniques that also make use of unlabelled data for training. typically a small amount of label data with a large amount of unlabled data.

Semi supervised learing Falls between unsupervised learning and Supervised learning.

Machine learning applications and goals.

(1) To improve the image classification in apps like Google photos, deep learning is required.

(2) To improve detection of malicious software. machine-learning is used in anti-virus software.

(3) Anti-spam softwares are trained better by machine learning. identify email message as spam or non-spam and if it happens to be spam then penalize them.

(4) Tagging people in photoes like facebook does it. it shows whether a face is present or not in a photograph.

(5) To identify patients suffering from a particular disease or Not.

(6) For speech recongition.

(7) To predict a set of customers who will respond to a particular type of advertisement or promotion like Amazon, recommend products especially for you.

(8) To categorize news articles to different categories of politics, sports, etc.

POORNIMA
COLLEGE OF ENGINEERING

DETAILED LECTURE NOTES

Campus: ..... Course: .....
Name of Faculty: .....
Class/Section: .....
Name of Subject: .....
Date: .....
Code: .....

Supervised versus unsupervised Learning

Basic For Comparision	Supervised learning	Unsupervised learning.
Basic	Deals with Labelled data	Handles unlabeled data.
Computional Complexity	High	low
Analyzation	offline	Real-time.
Accuracy	produce accurate results	Produces moderate results.
sub-domain	Classification and regression.	Clustering and Association rule mining.

machine learning application:-

① **virtual personal Assistants:-** Siri, Alexa, google are some of the popular example of virtual personal Assistant. As the name suggest, they assist in finding information. ML is an important part of these personal assistant.

virtual assistant are integrated to a variety of platforms:-

- Smart speakers:- Google Home.
- Smart phones
- mobile Apps.

② **predicitions while commuting:-**

Traffic predictions:- we all have been using gps navigation. while we do that, our current locations and velocities are being saved at a central server for managing traffic.

online Transportation:- when booking a cab, the app estimates the price of the ride. ML to define price shown by predicting the rides demand. in the entire cycle of servics, ML is playing a major role

Random forest advantage & disadvantage.

Advantage:- Random forest is based on many trees on the subset of the data and combines the output of all the trees therefore improve the accuracy.

(2) Random forest can be used to slove both classification as well as regrian.

(3) Random forest can automatically handle missing value.

(4) Random forest is usually robust to outlier and can handle them automatically

(5) Random forest is very stable.

Disadvantage:-

(1) complexity

Random forest a lot of trees and combine their output.

(2) longer training period.

⇒ Random forest require much more time to train as compared to decision trees to generate a lot of trees and output.

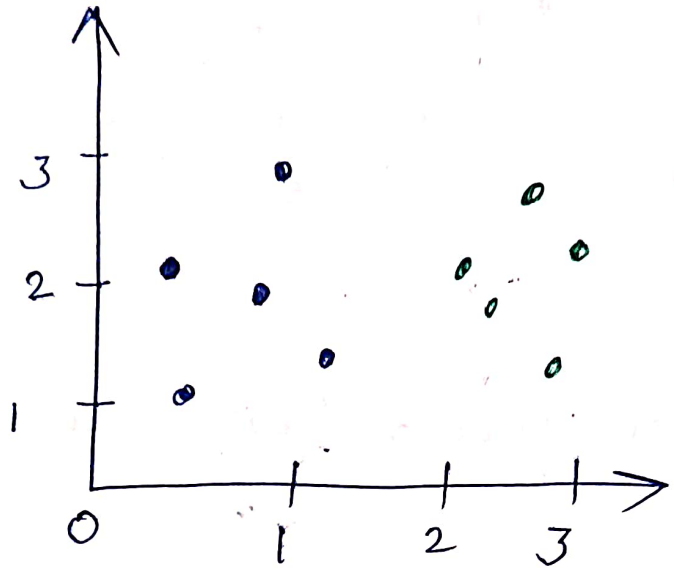
Gini impurity :- the Gini impurity of a node is the probability that a random chosen sample in a node would be incorrectly labeled if it was labeled by the distribution of sample in the node.

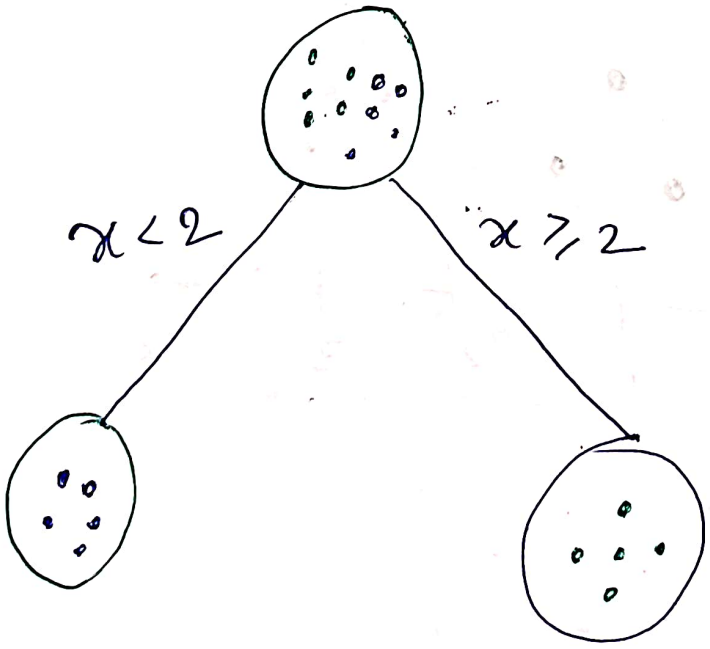
$$I_G(n) = 1 - \sum_{i=1}^{j} (P_i)^2$$

the Gini impurity of a node n is 1 minus the sum over all the classes j (for a binary classification task this is 2) of the fraction of examples in each class P_i squared.

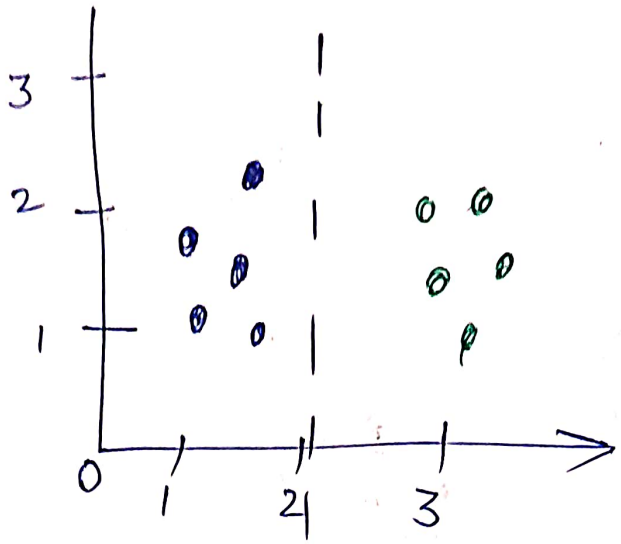
$$I_{root} = 1 - \left(\left(\frac{2}{6}\right)^2 + \left(\frac{4}{6}\right)^2 =\right.$$

$$1 - 5/9 = 0.444$$

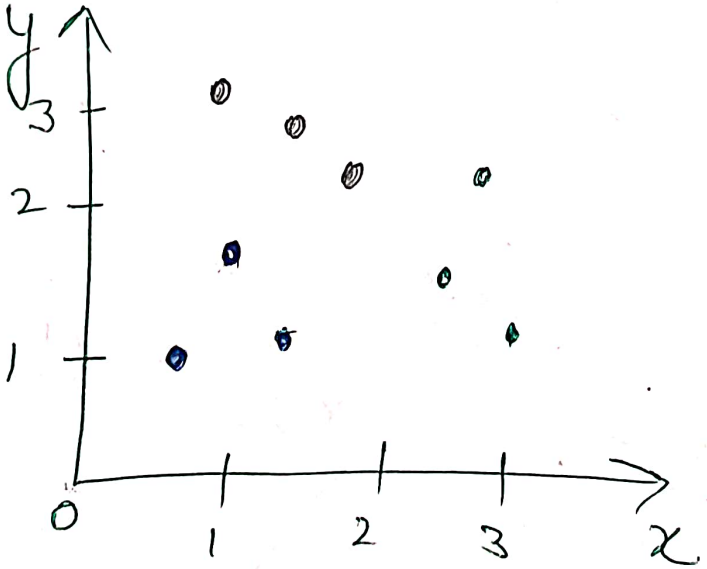


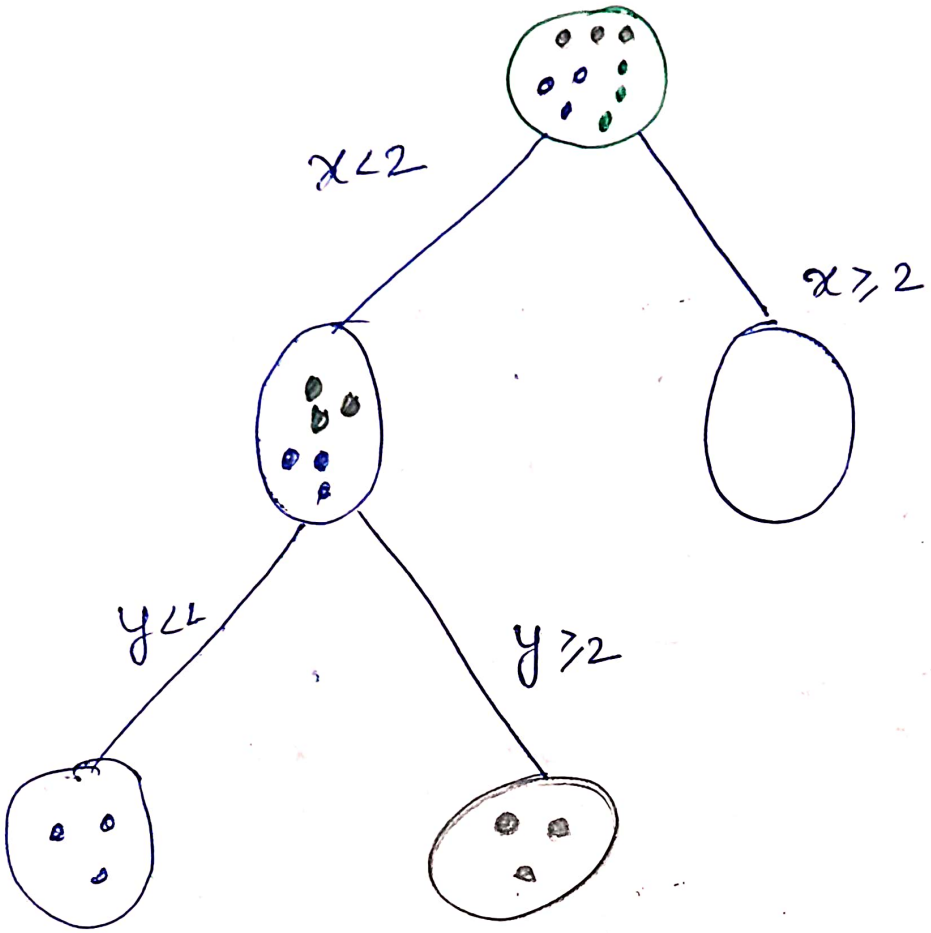


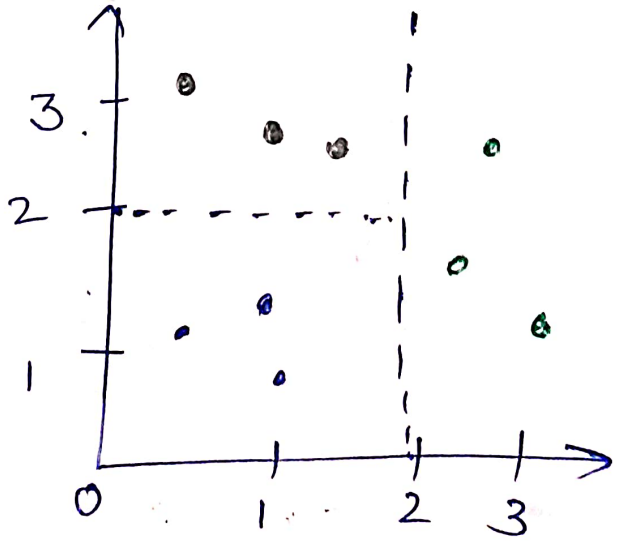
the simple decision is that $x < 2$ the left branch. $x \geq 2$ is right branch



dataset that has 3 clan now instead to 2.




x<2
x≥2
y<2
y≥2


3
2
1
0
1
2
3

$$G_I = \sum_{i=1}^{3} P_i * (1 - P(i))$$

$$3 * \left(\frac{1}{3} * \frac{2}{3}\right)$$

$$= \frac{2}{3}$$

$$G_{left} = 0*1 + 1*0 + 0*1 = \boxed{0}$$

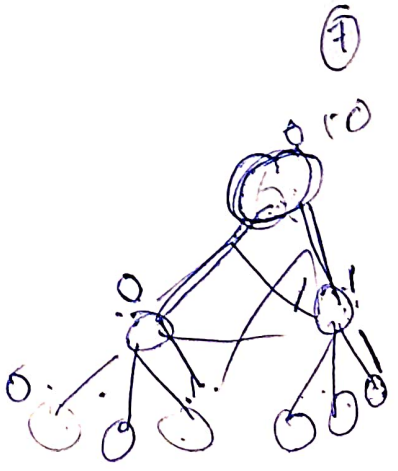
$$G_{right} = 3/8 * 5/8 + \frac{2}{8} * \frac{6}{8} + \frac{3}{8} * \frac{5}{8}$$

$$= \frac{21}{32}$$

$$G_{ain} = G_{initial} - 1/9\, G_{left} - \frac{8}{9} G_{right}$$

$$= \frac{2}{3} - 1/9 * 0 - \frac{8}{9} * \frac{21}{32}$$

$$= 0.083$$



DETAILED LECTURE NOTES

Support vector machine:- in machine learning sum or support vector machine are supervised learning model with analyze data used for classification and regression analysis.

Sum model is a representation of the examples as points in space, mapped so that the example of the separate| divided by a clear gap

it can solve linear or non-linear problems and work well for many partical problems

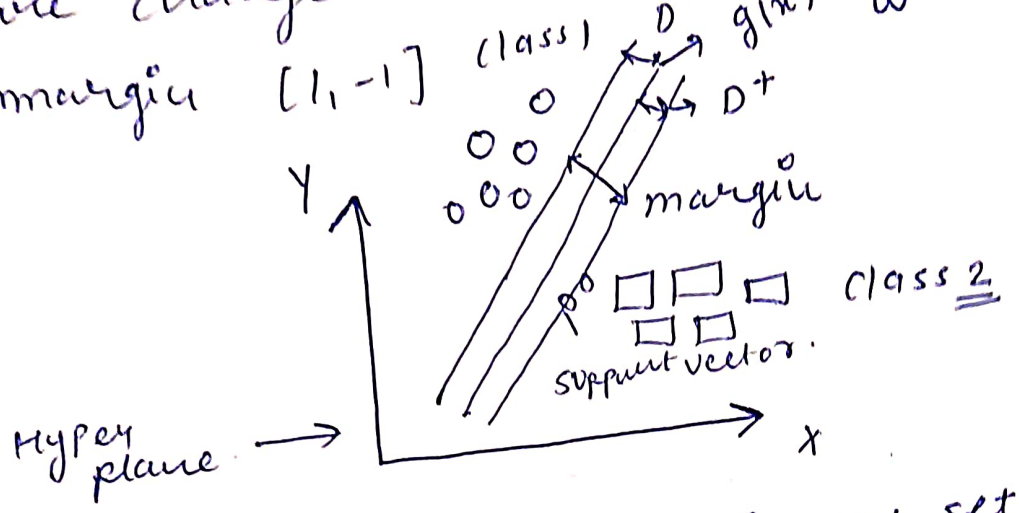
The idea of sum is simple

The algo. creates a line or a hyperplane which separates the data into clan.

Large margin intuition:-

In logistic regression, we take the output of the linear function and squash the value within the range of [0,1] using the sigmaid function. if the squashed value is greater than a threshold value we assign it a label 1. else we assign it a label 0. output of the linear function and if that output

is greater than t, we
class and if the output is -1.
is with another class. the threshold
are changed to 1 to -1 in SVM. wh
margin [1, -1]



SVM for linearly separt binary sets the goal is design a hyperplane that classifies all training vectors in two classes.

⇒ The best choice will be hyperplane that le
leaves the maximum margin from Both classes.

$g(\bar{x}) \geq 1 \quad \forall \bar{x} \in$ class 1

$g(\bar{x}) \leq -1 \quad \forall \bar{x} \in$ class 2

total margin E is computed

$$\frac{1}{|\bar{w}|} + \frac{1}{|\vec{w}|} = \frac{2}{|\vec{w}|}$$

minimizing this term will maximize the separability

Polynomial kernal:-

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^d$$

② Gaussian kernal:-

$$K(x, y) = \exp\left(\frac{-\|x - y\|^2}{2\sigma^2}\right)$$

③ Gaussian radial basis function:-

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2$$

K mean clustering is a partion of n object into k clusters in which each object belongs to the cluster with the nearest mean. the method produce exactly k different cluster of greatest possible distinction. the best no. of cluster of k- leading to the greatest separation is not known as a priori and must be computed from the data.

the objective of k-mean clustering is to minimize total intra-cluster variance or the sequred error function.

objective function = $J = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_i^{(j)} - c_j \|^2$

($x_i^{(j)}$: case i, c_j: centroid of cluster)

K = no. of clusters

Distance function

n = no. of cases.

what is clustering :- (group of similar data point is called clusters)
clustering is a basically a technique that groups similar data points such that the points in the same group are more similar to each other than the points in the other groups.

Algorithm:- cluster the data into k where k is predefined.

(2) select k points at random as cluster center.

(3) Assign object to their closest cluster center according to the euclidean distance function.

(4) calculate the centroid or mean of all object in each cluster.

(5) Repeat steps 1, 2, 3, 4 until the same points are assigned to each cluster.

Centroid:- the middle of cluster. A centroid is a vector that contains one no. for each variable where each no. is the mean of a variable for the observation in the cluster.

advantage

(1) Easy to implement with a large no. of variables.

(2) k-means may be computationally faster than hierarchical clustering (if k is small)

disadvantage: (1) difficult to predict the no. of cluster.

(2) the order of data has an impact on the final result.

DETAILED LECTURE NOTES

K mean clustering two type cluster:-

① Central ② Hierarchy.

- Step ① Take mean value.

Step.2 Find nearest no. of mean and put in cluster.

Step 3:- Repeat one and two until we get Some mean.

- $K = \{2, 3, 4, 10, 11, 12, 20, 25, 30\}$

$K = 2$

$m_1 = 4$ $\quad$ $m_2 = 12$

$K_1 = \{2, 3, 4\}$ $\quad$ $K_2 = \{10, 11, 12, 20, 25, 30\}$

$m_1 = \frac{2+3+4}{3}$ $\quad$ $m_2 = \frac{108}{6} = 18$

$m_1 = 3$

$m_1 = 3,$ $m_2 = 18$

$K_1 = \{2, 3, 4, 10\}$ $K_2 = \{11, 12, 20, 25, 30\}$

$m_1 = 4.75_{(5)}$ $m_2 = 19.6_{(20)}$

$K_1 = \{2, 3, 4, 10, 11, 12\}$ ✓ $K_2 = \{20, 25, 30\}$

$m_1 = 7$ $m_2 = 25$

$K_1 = \{2, 3, 4, 10, 11, 12\}$ ✓ $K_2 = \{20, 25, 30\}$

$m_1 = 7$ $m_2 = 25$

Thus we are getting same mean we have to stop.

new cluster will be

$K_1 = \{2, 3, 4, 10, 11, 12\}$

$K_2 = \{20, 25, 30\}$

Unit-II ML

Grouping unlabelled items using k-mean clustering :-

k-mean clustering :- clustering is one of the most common exploratory data analysis techniques used to get an intuition about the structure of the data. it can be defined as the task of identifying subgroups in the data such that data points in the same subgroups (cluster) are very similar while data points in different cluster are very different.

clustering analysis can be done on the basis of features where we try to find subgroups of samples based on features or on the basis of samples where we try to find subgroups of features based on samples.

clustering is used in market segmentation etc.

K mean clustering :- K mean algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non overlapping subgroups [clusters] where each data point belongs to only one group. it tries to make the inter-cluster datapoint as similar as possible while also keeping the clusters are different as possible.

The way K-mean algo. works is as

(1) Specify no. of cluster K.

(2) initialize centroid by first shuffling the dataset and then randomly selecting K data points for the centroid without replacement.

(3) keep iterating until there is no. change to the centroid. i.e. assignment of data points to clusters is not changing.

- Compute the sum of the squared distance between data points and all centroids.
- Assign each data point to the closest cluster (centroid)
- compute the centroid for the cluster by taking the average of the all data points the belongs to each cluster.

the objective function is:-

$$J = \sum_{i=1}^{M} \sum_{k=1}^{K} w_{ik} \|x_i - v_k\|^2 \quad \text{— (1)}$$

where $w_{ik} = 1$ for data point x_i if it belongs to cluster k, otherwise $w_{ik} = 0$

Also v_k is the centroid of x_i cluster.

$$\frac{\partial J}{\partial w_{ik}} = \sum_{i=1}^{M} \sum_{i=1}^{K} \|x_i - v_k\|^2$$

$$\Rightarrow w_{ik} = \begin{cases} 1 & \text{if } k = \text{argmin} \|x_i - v_j\|^2 \\ 0 & \text{otherwise} \end{cases}$$

in other words, assign the data point to the closest cluster judged by it's sum of squared distance from cluster's centroi

$$\frac{\partial J}{\partial v_k} = 2 \sum_{i=1}^{M} w_{ik} (x_i - v_k) = 0$$

$$u_k = \frac{\sum_{i=1}^{M} w_{ik} x_i}{\sum_{i=1}^{M} w_{ik}} \quad \text{— (3)}$$

DETAILED LECTURE NOTES

15, 15, 16, 19, 19, 20, 20, 21, 22, 28, 35, 40,
41, 42, 43, 44, 60, 61, 65

$\mu = 19$

initial cluster random.

$k = 2$

$C_1 = 16$

$C_2 = 22$

Distance 1 $(x_i - C_1)$

Distance 2 $(x_i - C_2)$

iteration 1 :-

x_i	C_1	C_2	Distance 1	Distance 2	nearest cluster	New Centroid
15	16	22	1	7	1	15.33
15	16	22	1	7	1	
16	16	22	0	6	1	
19	16	22	9	3	2	
19	16	22	9	3	2	
20	16	22	16	2	2	
21	16	22	16	1	2	
22	16	22	25	0	2	36.25
28	16	22	36	6	2	
35	16	22	12	13	2	
40	16	22	19	18	2	
41	16	22	24	19	2	
42	16	22	25	20	2	
43	16	22	26	21	2	
44	16	22	27	22	2	
60	16	22	28	38	2	
61	16	22	44	39	2	
65	16	22	45	43	2	
			49			

it is unsupervised learning technique which is used when you have unlabled data (i.e data without defined categories and groups) the goal of this algo. is to find groups in the data with the no. of groups represented by the value k.

PANIMALAR COLLEGE OF ENGINEERING

DETAILED LECTURE NOTES

PAGE NO. .....

~~Hueristic~~ Hyerical clustering

hierarchical clustering analysis is a method of cluster analysis which seeks to build a hierarchy of cluster example tree type structure based on the hierarchy.

two type clustering :-

1. Aggomerative clustering :- it is bottom up approach or hierarchical agglomerative clustering (HAC). A structure that is more informative than the unstructured set of cluster returned by flat clustering.

algorithm:-

given a dataset ($d_1, d_2, d_3 \cdots d_n$)

compute the distance matrix

for i=1 to N:

as the distance matrix is symmetric
the primary diagonal so we compute
only lower value.

part of the primary diagonal

for j=1 to i:

dis_mat[i][j] = distance[di, dj]

each data point is a singular cluster

repeat

merge the two cluster having minimum
distance

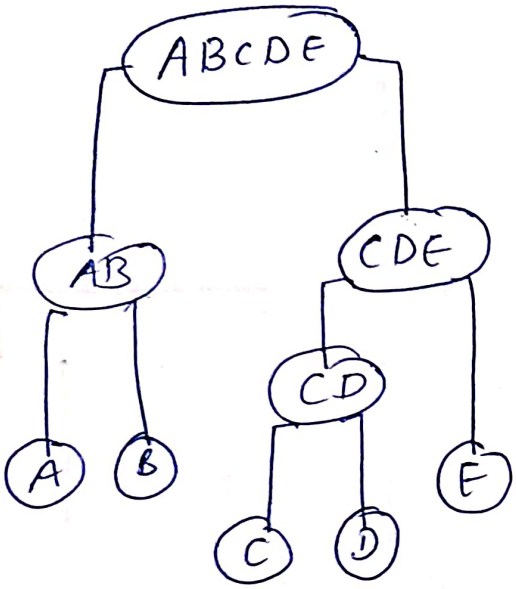
update the distance matrix

until only a single cluster remains

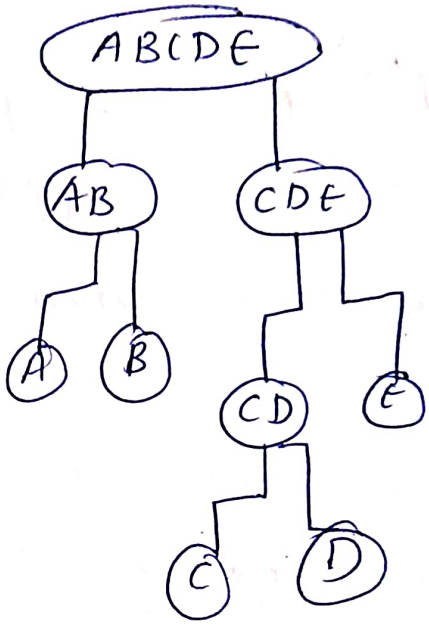
DETAILED LECTURE NOTES

Campus: ..... Course: ..... Class/Section: ..... Date: .....
Name of Faculty: ..... Name of Subject: ..... Code: .....

Example of Dendrogram



Agglomerative



Divisive

(single link Technique)

(complete link)

	P_1	P_2	P_3	P_4	P_5
P_1	0				
P_2	9	0			
P_3	3	7	0		
P_4	6	5	9	0	
P_5	11	10	2	8	0

min value

	P_1	P_2	$[P_3\ P_5]$	P_4
P_1	0			
P_2	9	0		
$\Rightarrow [P_3, P_5]$	3	7	0	
P_4	6	5	8	0

$\Rightarrow d[P_1, [P_3, P_5]]$

$\Rightarrow \min(d(P_1P_3), d(P_1P_5))$

$\Rightarrow \min(3, 11)$ $\min \Rightarrow \underline{3}$

$\Rightarrow d[P_2, [P_3, P_5]]$

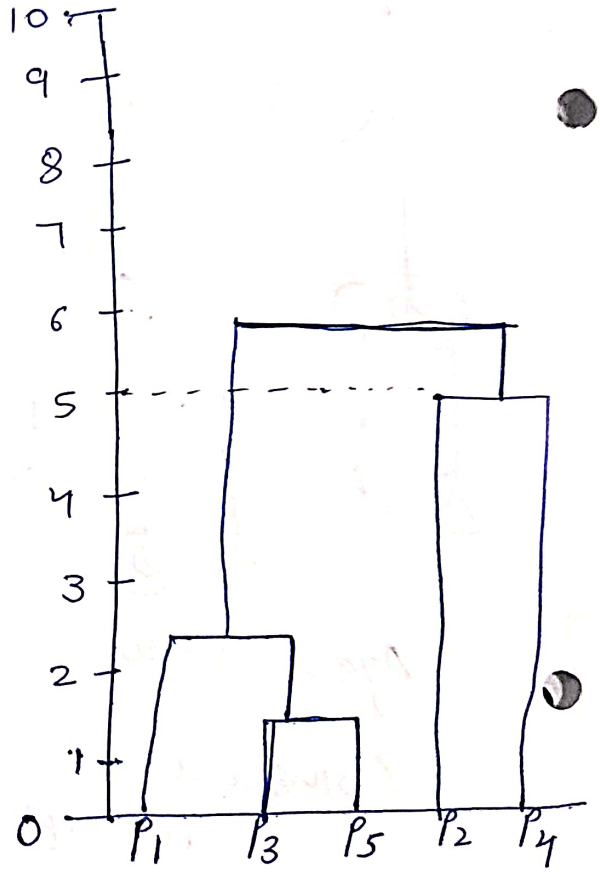
$\min d(P_2, P_3), (P_2, P_5)$

$\min [7, 10] \Rightarrow \underline{7}$

$\Rightarrow d[P_4, [P_3, P_5]$

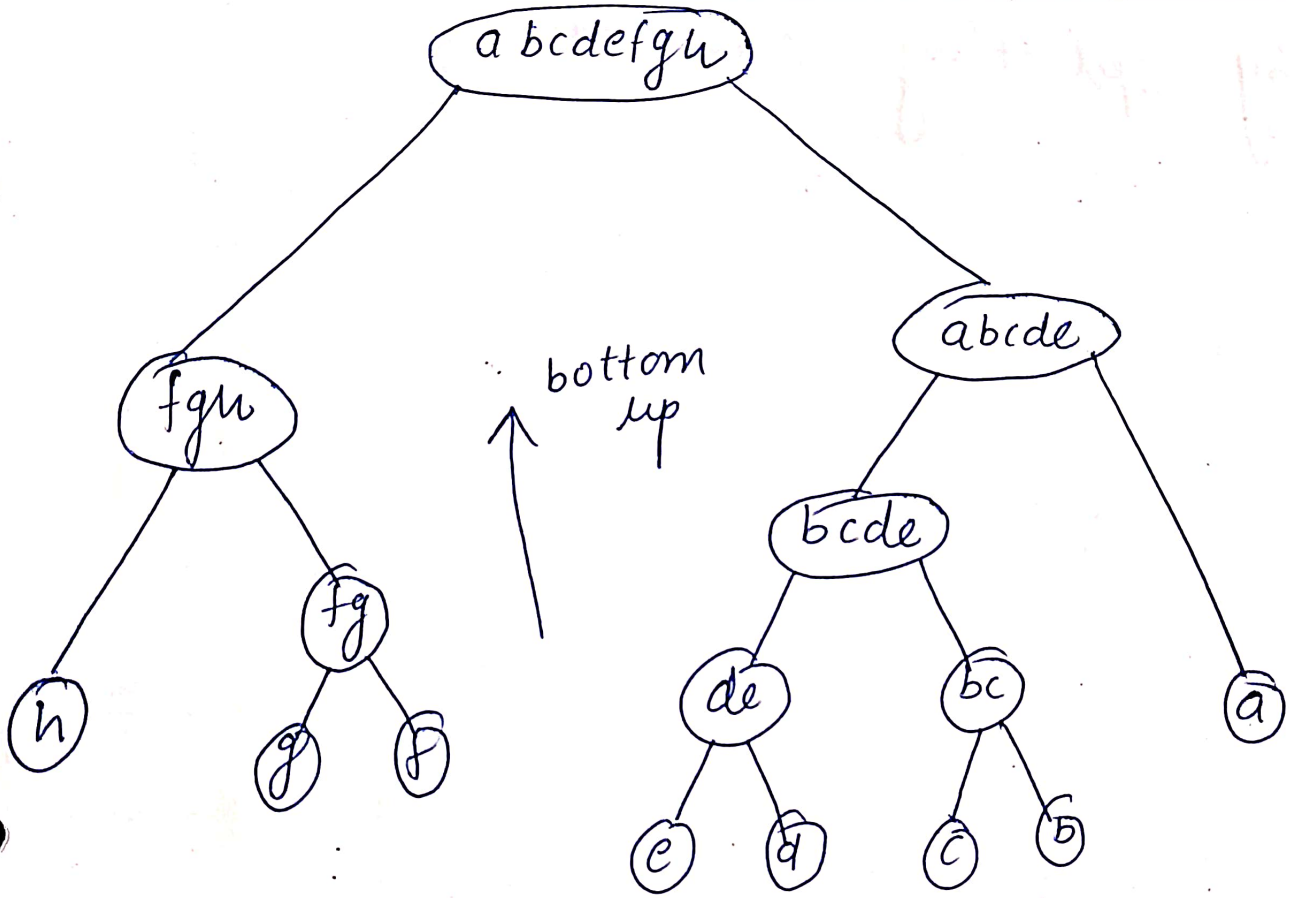
$\min d[P_4, P_3], d[P_4, P_5]$

$\min (9, 8) \Rightarrow \underline{8}$



DETAILED LECTURE NOTES

Campus: ..... Course: .....
Name of Faculty: .....
Class/Section: .....
Name of Subject: .....
Date: .....
Code: .....



Divisive / विभाजित

(2) Divise clustering :- it is top-down approach this algorithm also does not require to no. of cluster. top-down clustering requires a method for spitting a cluster the contain the whole data and proceeds by splitting clusters.

DETAILED LECTURE NOTES

Campus: ..... Course: ..... Class/Section: ..... Date: .....
Name of Faculty: ..... Name of Subject: ..... Code: .....

	$[P_1\ P_3\ P_5]$	P_2	P_4
$[P_1\ P_3\ P_5]$	0		
P_2	(7)	0	
P_4	(6)	5	0

$d\ (P_2, [P_1, P_3, P_5]$

$\min(d(P_2, P_1), d(P_2 P_3), d(P_2 P_5)$

$\min(9, 7, 10) \Rightarrow 7$

$d(P_4 (P_1 P_3 P_5]$

$\min d(P_4, P_1), (P_4 P_3)$

$d(P_4, P_5]$

$\min [6, 9, 8]$

$\min = 6$

	$[P_1\ P_3\ P_5]$	$[P_2, P_4)$
$[P_1, P_3, P_5]$	0	
$[P_2, P_4]$	(6)	0

$d\ [P_1, P_3, P_5]\ [P_2\ P_4]$

$\min d\ [P_2, P_1], d[P_2 P_3], d[P_2, P_5]\ d\ [P_4, P_1]\ d[P_4, P_3], d\ [P_4, P_5]$

$\min [9, 7, 10, 6, 9, 8]\quad \min [6]$

Probb · Probabilistic clustering :-

In probabilistic clustering is estimated for each point x (1 & 0) the P is a probability of set of k element in data sampling space.

$$P = \frac{\sqrt{(x_2 - x_1)^2}}{\text{total no. of attributes.}}$$

k mean $O(n)$ hierarchical $O(n^2)$

Application of clustering :- market segmentation
Social network analysis
image segmentation

Apriori algorithm:- Apriori algo. was the first algo. that was proposed for frequent itemset mining. This algo. uses two steps 'join' and 'prune' to reduce the search space.

F-P growth algorithm:- FP growth algo is an efficient algo. for calculating frequently co-occuring items in a transaction database. F-P grow algo. used database without candidate generation.

Advantage of fp growth:-

1. Faster than apriori algo.
2. No. candidate generation.
3. only two passes over dataset.

Disadvantage:-

1. FP tree may not fit in memory.
2. FP tree is expensive to build.

Q. Generate F-P tree for the following transaction

Data set [minimum support = 30%]

Transaction id	items
1	E, A, D, B
2	D, A, C, E, B
3	C, A, B, E
4	B, A, D
5	D
6	D, B
7	A, D, E
8	B, C

minimum no. of transaction $\frac{30}{8} = 2.4$ - (3) remove

items	Frequency	Priority
A	5	3
B	6	1
C	3	5
D	6	2
E	4	4

Lower priority no. means higher priority.

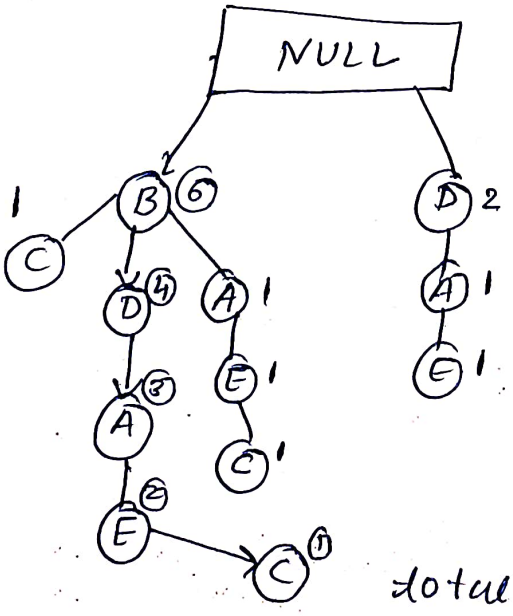
⇒ order the items according to the priority.

DETAILED LECTURE NOTES

Step-2

Trans. id	items	ordered items
1	E, A, D, B	B, D, A, E
2	D, A, C, E, B	B, D, A, E, C
3	C, A, B, E	B, A, E, C
4	B, A, D	B, A, D
5	D	D
6	D, B	B, D
7	A, D, E	D, A, E
8	B, C	B, C

Step:-3 draw F-P tree



B, D, A, E = 1 — compute

B : 1, 1, 2, 3, 4, 5, 6
D : 1, 1, 2, 3, 4
A = 1 1 2, 3
E 1, 2
C = 1

A = 1
E = 1 — (3) transaction.
C = 1

D : 1, 2 — 4
A = 1
E = 1
C : 1

total occurence.-