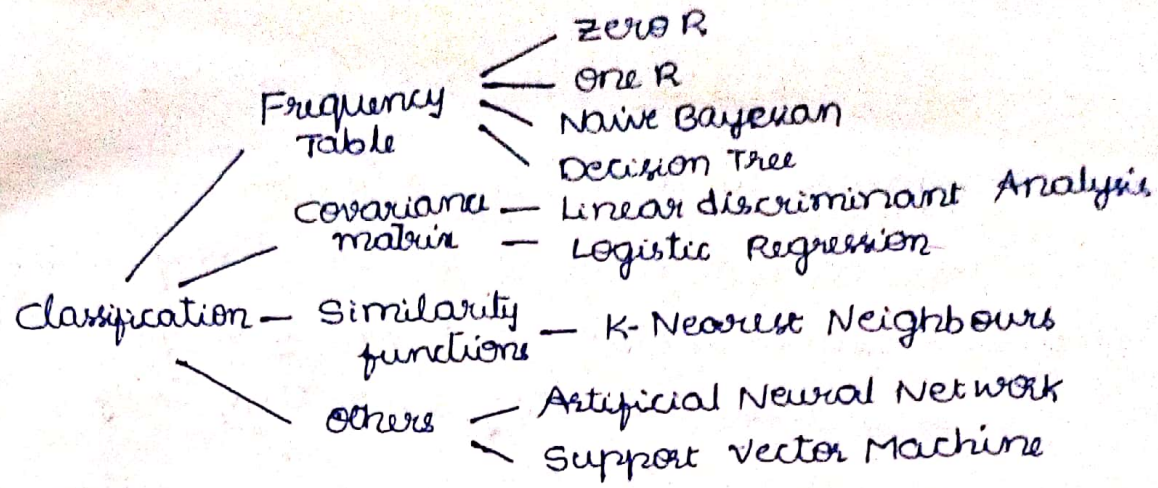
UNIT-4

Classification Methods :-



Classification :- It is the form of data analysis that Extracts models describing important data classes.

Ex:- A class has A, B, C, D, E classes based on their Grades. By using a model, we can classify the students depends upon grades.

$90 \geq$ grade	class A
$80 \geq$, < 90	class B
$70 \geq$, < 80	C
$60 \geq$, < 70	D
< 60	E

Applications of classification :-

- Image & pattern Matchings
- Medical diagnosis
- Fraud detections
- Loan Approvals
- Classifying financial market trends

Classification problem :-

Given a db $D = T_1, T_2 \cdots T_n$ of Tuples (items, records) and a set of classes $C = C_1, C_2 \cdots C_m$

The classification problem is to define mapping $f: D \rightarrow C$ where Each T_i is assigned to one class. A class C_j contains precisely those tuples mapped to it means

$$C_j = \{ t_i \mid f(t_i) = C_j \; ; \; i \leq i \leq n \; \& \; t_i \in D \}$$

General Approach to the classification :-

There are two ways to do the classification process - Learning stage (or) Training stage
- Classification Stage (or) Testing Stage.

Learning stage :- where a classification model is constructed

Classification stage :- where the model is used to predict the class labels for given data.

Ex :-

Training data Ex:

Name	age	income	loan-decision
A	youth	low	risk
B	youth	middle	risk
C	middle	high	Safe
D	middle	low	risk
E	senior	low	risk
F	senior	low	risk

↓

Classification Algo

⇓

Classification rules

↓

Stage 1 → Learning Stage

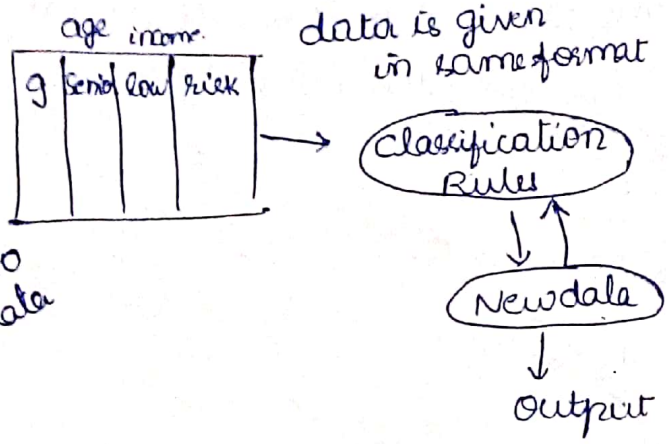
If age = youth then loan-decision = risk

If age = middle and income = high then loan decision = safe

If age = middle and income = low then loan = risk

If age = senior then loan-decision = risk

Stage 2
Testing stage



The data in testing should not match to training data

Issues in classification :-

- Missing data
- performance measurement

Missing data :-

. The missing value in the dataset create a problems during both training & classification problem itself.
. The missing values in training data must be handled otherwise produced inaccurate results.
. By using handling methods of missing data, we can overcome the problem.

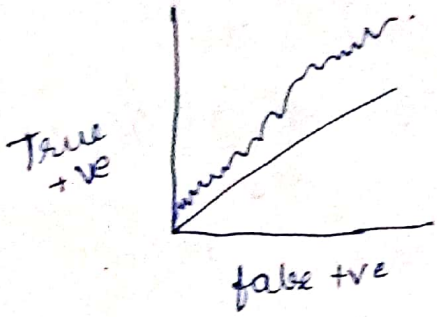
Measuring the performance :-

- The performance of a classification algo is usually examined by evaluating the accuracy of classification
- To Determine which is best algo depends on the interpretation of problem by users.
- The space & time complexity is used to Evaluate the algorithm but these approaches are usually secondary.

* By using the confusion matrix we can Estimate the accuracy of classification methods

	x/Ci	Yes	NO
Class A	Yes	True +ve	false -ve
	NO	false +ve	True -ve

ROC curve :- The ROC curve (Recieves Operata curve) shows the relationship b/w false +ve & true +ve



Statistical Based Algorithm :-

1. Regression :- 1st unit refer
2. Bayesian method / Baye's classification

Naive Baye's classification method :-

It is a statical classifier they can predict the class membership probabilities such as the probability that a given tuple belongs to the particular class.

The Bayesian classification is based on baye's Theorem.

→ Naive Bayesian classifier assume that the effect of an attribute value on a given class is independent of the value of other attributes. This assumption is called as class conditional Independence.

Baye's Theorem :-

$$P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

where $P(A/B)$ = The probability of occurance of Event A given the Event B is true.

$P(A) \& P(B)$ = The probability of occurence of Event A and Event B

A is called proposition
B is called Evidence

$P(A)$ is called prior probability of proposition
$P(B)$ is called prior probability of Evidence
$P(A/B)$ is called posterior or posterior probability.
$P(B/A)$ is called likelihood

$$\text{posterior} = \frac{(\text{likelihood})(\text{prior probability of proposition})}{\text{prior probability of Evidence}}$$

Ex:- Consider the given dataset apply the Naive baysian algo and predict that if a fruit has the following properities then which type of fruit it is.

fruit = {yellow, sweet, long}

Fruit	Yellow	Sweet	long	total
Mango	350	450	0	650
Banana	400	300	350	400
Other	50	100	50	150
Total	800	850	400	1200

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

P(x/mango)

P(Yellow/mango)

$$= \frac{P(m/y)\,P(y)}{P(m)}$$

$$= \frac{\frac{350}{800} \times \frac{800}{1200}}{\frac{650}{1200}}$$

$$= 0.52$$

P(Sweet/mango)

$$= \frac{P(\text{mango/sweet})\,P(\text{sweet})}{P(\text{mango})}$$

$$= \frac{\frac{450}{850} \times \frac{800}{1200}}{\frac{650}{1200}}$$

$$= 0.69$$

P(long/mango)

$$= \frac{P(\text{mango/long})\,P(\text{long})}{P(\text{mango})}$$

$$= \frac{\frac{0}{400} \times \frac{400}{1200} = 0}{\frac{650}{1200}}$$

Result : $\pi P(x_i/c_i)$

$= 0.52 \times 0.69 \times 0$

$= 0$

Ans: Banana

P(x/Banana)

$$P(Y/B) = \frac{P(B/Y)\,P(Y)}{P(B)}$$

$$= \frac{\frac{400}{800} \times \frac{800}{1200}}{\frac{400}{1200}} = 1$$

P(Sweet/Banana)

$$\frac{P(B/S)\,P(S)}{P(B)}$$

$$= \frac{\frac{300}{850} \times \frac{850}{1200}}{\frac{400}{1200}}$$

$$= 0.75$$

P(long/banana)

$$= \frac{P(B/l)\,P(l)}{P(b)}$$

$$= \frac{\frac{350}{400} \times \frac{400}{1200}}{\frac{400}{1200}}$$

$$= 0.87$$

Result : $0.75 \times 1 \times 0.87$

$= 0.65$

As it is highest Banana

P(x/other)

$$P(Y/O) = \frac{P(O/Y)\,P(Y)}{P(O)}$$

$$= \frac{\frac{50}{800} \times \frac{800}{1200}}{\frac{150}{1200}} = 0.33$$

$P(S/O) = 0.66$

$P(l/O) = 0.33$

Result : 0.072

Ex-2

RID	age	income	Student	Credit-rating	class: buys-computer
1	youth	high	NO	fair	NO
2	y	high	NO	Excellent	NO
3	middle	high	NO	f	yes
4	Senior	medium	y	f	y
5	S	low	y	f	y
6	S	low	y	ex	NO
7	m	low	N	ex	y
8	y	medium	y	f	NO
9	y	low	y	f	y
10	S	m	y	f	y
11	y	m	N	ex	y
12	m	m	y	ex	y
13	m	h	y	f	N
14	S	m	N	ex	

X = (age = youth, income = medium, Student = yes, Credit-rating = fair)

p(buys-computer = yes) = 9/14

p(buys-computer = NO) = 5/14

p(X / buys-computer = yes)

p(age = youth / buys-computer = yes) = 2/9

p(age = youth / buys-computer = NO) = 3/5

p(income = medium / buys-computer = y) = 4/9

p(income = medium / buys-computer = N) = 2/5

p(student = yes / b-c = yes) = 6/9

p(St = y / buys comp = NO) = 2/5 total no. of NO's

p(R = fair / bc = y) = 6/9 total no. of yes's

p(R = fair / bc = N) = 2/5

yes → 2/9 × 4/9 × 6/9 × 6/9 × 9/14 = 0.044

NO → 3/5 × 2/5 × 2/5 × 3/5 × 5/14 = 0.038.

∴ X will buy the computer

Decision tree Induction :-

A decision tree is a flowchart like tree structure where each internal node (non leaf node) denotes a test or an attribute. Each branch represents an outcome of the tests and the leaf node represents a class label. The top most node in tree is the root node.

In this decision tree a tuple X for which class label is unknown, attribute values of the tuple are tested against the decision tree. A path is traced from root to the leaf node which holds the class prediction for the tuple.

In this decision tree doesnot require any domain knowledge or parameter setting to classify the data (so it is very popular)

Applications :-

- Manufacturing & production
- Financial Analysis
- Astronomy
- Molecular Biology.

KNN → K-Nearest Neighbours

The KNN classifies to predict the target label by finding nearest neighbour class.

The closest class will be identified using the distance measures like Euclinear distance or Manhatten distance or Jacad distance

Application:-

- used in Recommendation systems
- To store info in servers

How to choose the K-value:-

. (Select the) selecting the K-value in KNN is the most Critical problem.

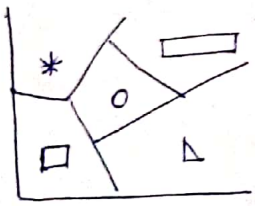
. The simple approach to select the K-value is $\sqrt{n}$ where n is no of observations / data points

. By using the cross validation also, we can select the value of K.

NOTE:-

The KNN Algorithm also called as lazy learner algo because we are not used to learn any function or to store values.

Case 1:-



called as voronaiEffect as all classes are nearer to the obj we are unable to select the class.

Perform KNN Algo on following dataset & predict the class for $X(P_1 = 3 \text{ \& } P_2 = 7)$ & $K = 3$

	P_1	P_2	classes
①	7	7	False
②	7	4	False
③	3	4	True
④	4	4	True

Euclinear distance

$$\sqrt{(x_H - H)^2 + (y_H - H)^2}$$

Observed or given value — Actual value

Given $P_1 = 3$, $P_2 = 7$

As K = 3
consider 3 values from low to hig

① $\sqrt{(3-7)^2 + (7-7)^2} = (4)$ — True ✓

② $\sqrt{(3-7)^2 + (7-4)^2} = 5$

③ $\sqrt{(3-3)^2 + (7-4)^2} = (3)$ — False

④ $\sqrt{(3-1)^2 + (7-4)^2} = \sqrt{13} = (3.6)$ — True ✓ highest category

Case 2 :-

Line encoding

Gender

female — 0
male — 1
male — 1
female — 0

} consider one as zero's & other to one's

Range Encoding

consider Avg

Height (cm)	Avg	Lower	Upper
100-110	105	100	110
110-120	115	110	120
120-130	125	120	130
130-140	135	130	140

Case 3 :-

divide them

Age	Age youth	Age middle	Age senior
youth	1	0	0
middle age	0	1	0
Senior	0	0	1
youth	1	0	0
middle age	0	1	0
Senior	0	0	1

Attribute selection :- (Decision Tree)

An attribute selection is a heuristic for selecting the decision tree spliting criteria that best separate the given data partition (D) of class label training tuples into Individual classes.

It is also called as splitting rules because they determine how the tuples of given node are to be split.

- The attribute selection provides a ranking for Each attribute and select the best score attribute as a root node of trees.
- There are three popular attribute selection methods.
  - Information gain
  - Gain ratio
  - Gini Index

$$\text{Information gain}(D,A) = \text{Entropy}(D) - \sum_{j=1}^{y} \frac{|D_j|}{|D|} \text{entropy}(D_j)$$

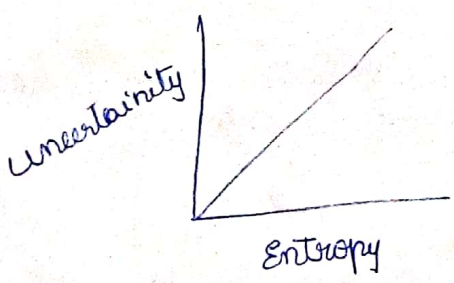
$$\text{Entropy}(D) = \sum_{i=1}^{c} -p_i \log_2(p_i)$$

(OR)

$$= \frac{-p_i}{p+n} \log_2\left(\frac{p_i}{p+n}\right) - \frac{n_i}{p+n} \log_2\left(\frac{n_i}{n+p}\right)$$

* Information gain is used as Attribute selection measure
* Entropy is used for to findout uncertainity associated with a random variable.

The values of an entropy is in b/w 0 to 1.



From previous example

RID	age	income	student	credit rating	buys computer

$p_i \rightarrow$ yes $\Rightarrow 9$

$n_i \rightarrow$ no $\Rightarrow 5$

entropy of data $(D) = \frac{-p_i}{p+n}\log_2\left(\frac{p_i}{p+n}\right) - \frac{n_i}{p+n}\log_2\left(\frac{n_i}{p+n}\right)$

$= -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$

$= \frac{9}{14}\log_2(0.6428) - \frac{5}{14}\log_2(0.3571)$

$= -0.6428 \times \frac{\log(0.6428)}{\log 2} - \frac{5}{14} \times \frac{\log(0.3571)}{\log 2}$

$= -0.6428 \times \frac{-0.1919}{0.3010} - 0.3571 \times \frac{-0.4472}{0.3010}$

$= -0.6428 \times -0.6375 - 0.3571 \times -1.4857$

$= 0.40977 + 0.5305$

$= 0.9402$

Info age (D) youth middle senior

youth	middle	senior
yes = 2	yes = 4	yes = 3
no = 3	no = 0	no = 2

$= \frac{5}{14}\left(\frac{-2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) + \frac{4}{14}\left(\frac{-4}{4}\log\frac{4}{4+0} - 0\right)$

$+ \frac{5}{14}\left(\frac{-3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}\right)$

$= 0.694$ bits

Information gain $(D, age) = info(D) - info_{age}(D)$

$= 0.940 - 0.694$

$= 0.246$ bits

Info Income (D) High medium low

Yes = 2 Yes = 4 Yes - 3

NO = 2 NO = 2 NO = 1

$$= \frac{4}{14}\left(\frac{-2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) + \frac{6}{14}\left(\frac{-4}{6}\log\frac{4}{6} - \frac{2}{6}\log\frac{2}{6}\right)$$

$$+ \frac{4}{14}\left(\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4}\right)$$

$$= 0.909$$

Information gain (D, Income) = info(D) − infoIncome(D)

$$= 0.940 - 0.909$$

$$= 0.029 \text{ bits}$$

Info Student (D) Yes NO

Yes = 6 Yes = 3

NO = 1 NO = 4

$$= 0.786$$

info gain = 0.151 bits.

Credit rating fair Excellent

Y = 6 yes = 3

N = 2 NO = 3

$$= 0.892$$

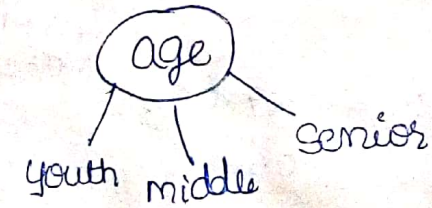
Info gain = 0.940 − 0.892 = 0.048.

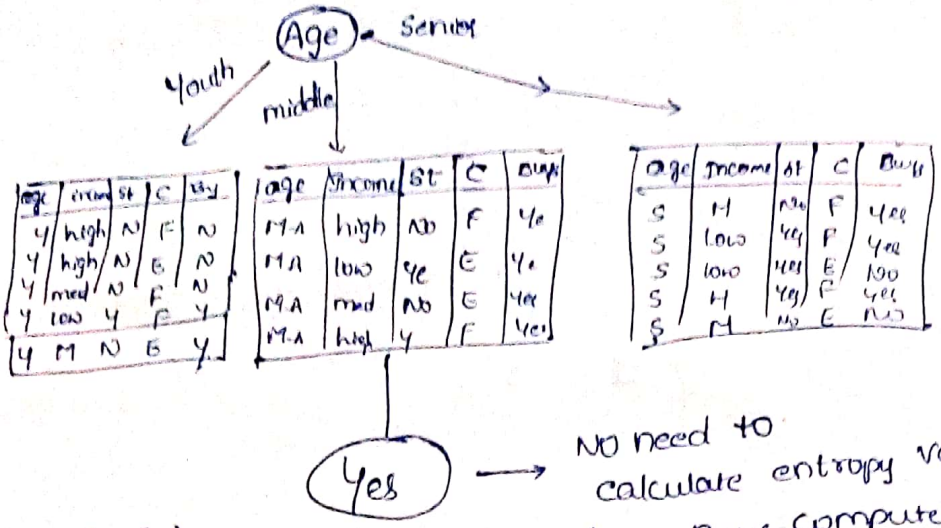
Gain y, age = 0.246

Gain Income = 0.029

Student = 0.154

Gain rating = 0.048.





→ No need to calculate entropy values bcz Buys-computer = yes on all tuples.

Entropy Youth (D) $= -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}$

NO = 3
yes = 2 $= 0.2922$

Next attribute Income

High	Medium	LOW
yes = 0	yes = 2	yes = 1
NO = 2	NO = 1	NO = 0

$$= \frac{2}{5}\left(\frac{-0}{2}\log\frac{0}{2} - \frac{2}{2}\log\frac{2}{2}\right) + \frac{2}{5}\left(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right) + \frac{1}{5}\left(-\frac{1}{1}\log\frac{1}{1} - 0\right)$$

$$= \frac{2}{5}\left(-\log\frac{1}{2}\right) = 0.1204$$

Information gain $= 0.29 - 0.1204 = 0.17$

Entropy Student (D) ⇒

yes	NO
yes = 2	yes = 0
NO = 0	NO = 3

$$\frac{2}{5}\left(\frac{-2}{2}\log\left(\frac{2}{2}\right) - \frac{0}{2}\log\left(\frac{0}{2}\right)\right) + \frac{3}{5}\left(\frac{-0}{3}\log\frac{0}{3} - \frac{3}{3}\log\frac{3}{3}\right) = 0$$

Info gain $= 0.29 - 0 = 0.29$

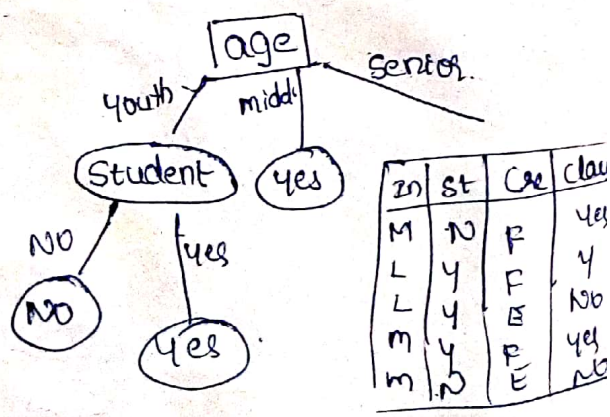
Entropy credit rating (D) :-

Fair	Excellent
yes = 1	yes = 1
NO = 2	NO = 1

$= 0.95$

Info gain $= 0.29 - 0.95$
$= -0.66$

Among all student have high Information gain so next node is Student



Entropy Senior (D) NO=2 Yes=3

$$-\frac{3}{5}\log(3/5) - 2/5\log(2/5) = 0.29$$

Entropy Income (D) Medium Low

Yes=2 Yes=1

NO=1 NO=1

$$\frac{3}{5}\left(-2/3\log(2/3) - 1/3\log(1/3)\right) + 2/5\left(-1/2\log(1/2) - 1/2\log 1/2\right) = 0.286$$

Info gain (Income) = 0.29 - 0.268 ⇒ 0.006.

Entropy Credit rating (D) Fair Excellent.

Yes=3 Yes=0

NO=0 NO=2

$$\frac{3}{5}\left(-3/3\log 3/3 - 0/3\log 0/3\right) + 2/5\left(0/2\log 0/2 - 2/2\log 2/2\right) = 0$$

Info gain (Credit rating) = 0.29 - 0 = 0.29

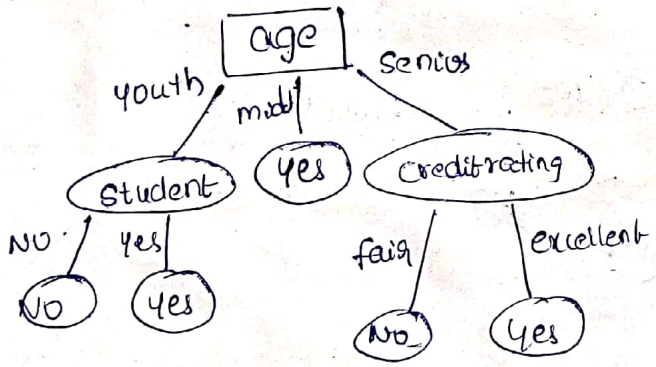
Entropy Student (D) Yes NO

Yes=2 Yes=1

NO=1 NO=1

$$\frac{3}{5}\left(-2/3\log 2/3 - 1/3\log 1/3\right) + 2/5\left(-1/2\log 1/2 - 1/2\log 1/2\right) = 0.286$$

Info gain = 0.292 - 0.286 = 0.006



✓ Gain Ratio (C4.5), Gini Index

Refer TB.

Artificial
Neural Networks :-

Neural networks are parallel computing devices which is basically an attempt to make a computer model of the brain.

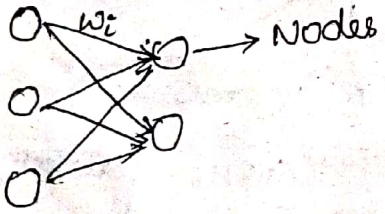
The main objective of neural networks to perform the various computation tasks faster than tradition system

The tasks are - pattern Recognition
- classification
- Approximation
- clustering & optimization

* ANN also called as Artificial neural system, parallel Distributed processing system. connectionist systems.

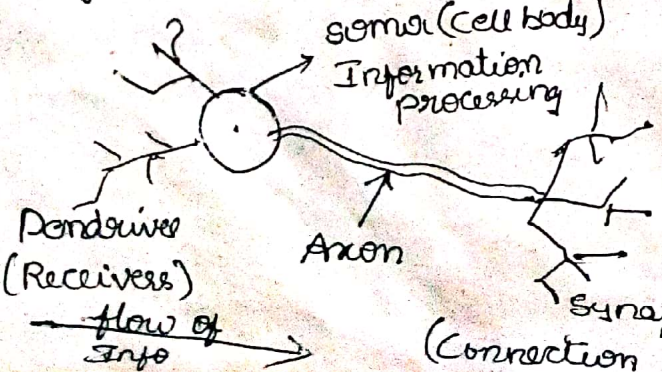
Working of ANN:-

* ANN requires the large collection of units that are interconnected in some pattern to allow communication b/w the units. Each unit is called as neuron / node. which are simple processors operates in parallel.
* Each connection link is associated with weight that has information about the input signal



History of ANN:-

- In 1943, Started the origin of networks using electric circuits in order to describe how neurons in brain might work.

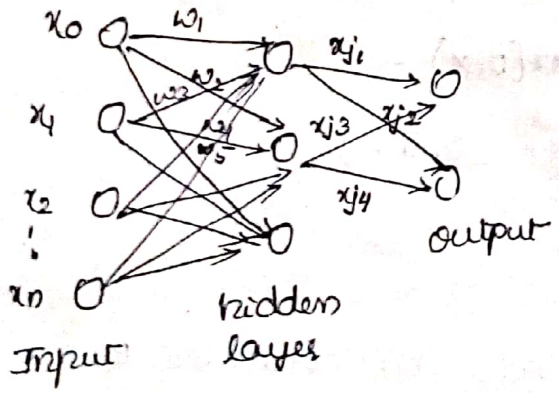


BNN	ANN
Soma	Node
Dendrites	Input
Synapse	weights / Interconnection
Axon	output

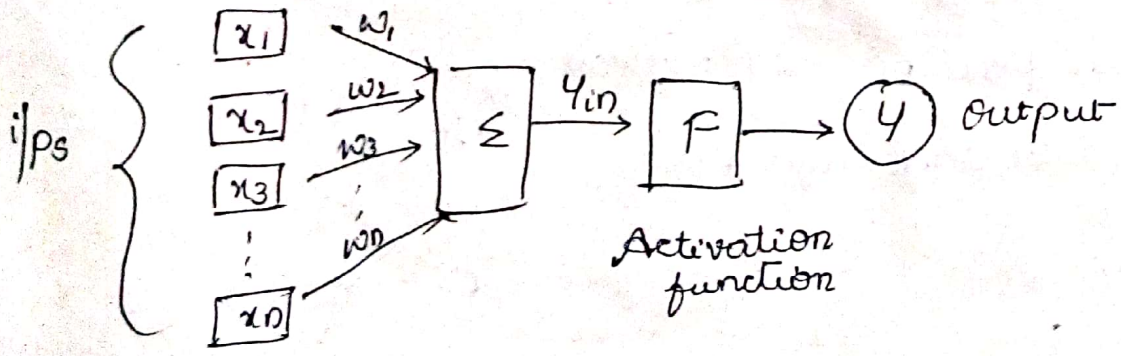
The difference b/w BNN & ANN :-

Criteria	BNN	ANN
processing	Massively parallel, slow but superior than ANN	Massively parallel fast but inferior than BNN
Size	10^{11} neurons & 10^{15} connection	10^2 to 10^4 nodes depends on applications & N/w design
Fault tolerance	performance degrades with even partial damage	performance is very fast

General Model of ANN :-



$$y_{in} = \sum_{i=1}^{m} x_i w_i$$



Activation Functions :-

- To check node is working or not
- Convert the data into non-linear form

Types:-

- Step Function
- Sigmoid Function
- Linear function
- Tanh
- ReLU

Step Function : A = 1 if y > Threshold otherwise A = 0

Sigmoid function :-

$\frac{1}{1+e^{-x}}$ 2 types - binary 0 to 1
bipolar -1 to 1

Linear function : $y = mx$ $(-\infty \text{ to } +\infty)$

Tanh : Tangent hyperbolic function

$$\text{Tanh} = \frac{2}{1+e^{-2x}} - 1 = 2 * \text{Sigmoid}(2x) - 1$$

ReLU (Rectified Linear unit)

It is implemented in Hidden layers

$A(x) = \max(0, x)$

value range = $[0, \inf)$

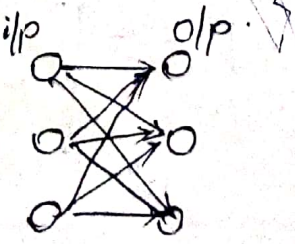
Network Topology :-

/

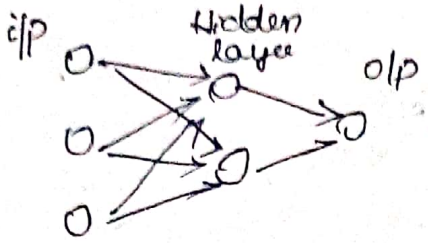
How the nodes are connected to Each other.

- Feed Forward Network

(1) Single layer F.F.N

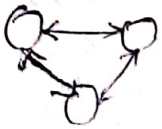


(iv) Multilayer F.F.N

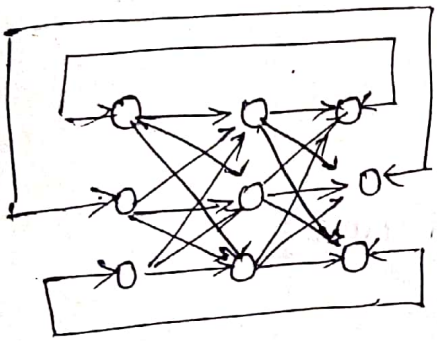


- Feed back N/w

(i) Fully recurrent N/w



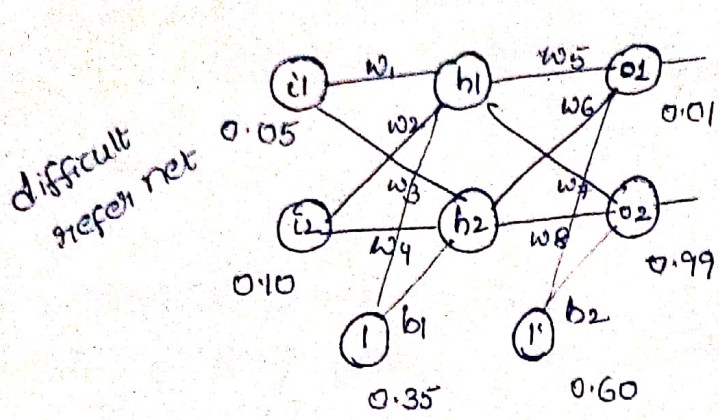
(ii) Jordan N/w



Issues in ANN :-

(i) Attributes (No. of source nodes)
(ii) No. of hidden layers
(iii) No. of hidden nodes.
(iv) Training data.
(v) No. of sinks
(vi) Interconnections
(vii) Weights
(viii) Activation function
(ix) Learning Techniques
(x) Stop

Ex:- Back propagation



difficult refer net

$w_1 = 0.15$
$w_2 = 0.20$
$w_3 = 0.25$
$w_4 = 0.30$
$w_5 = 0.40$
$w_6 = 0.45$
$w_7 = 0.50$
$w_8 = 0.55$

Back propagation :-

Algorithm –

- For each input layer unit j

$$O_j = I_j$$

- For each hidden or output layer unit j

$$I_j = \Sigma_i w_{ij} + \theta_j$$

Back propagation Errors :-

For each unit j in the O/p layer

$$Err_j = O_j(1-O_j)(T_j - O_j)$$

For each unit j in the hidden layer

$$Err_j = O_j(1-O_j)\Sigma_k Err_k w_{jk}$$

For each weight w_{ij} in the N/w

$$\Delta w_{ij} = (l) Err_j O_j$$

$$w_{ij} = w_{ij} + \Delta w_{ij}$$

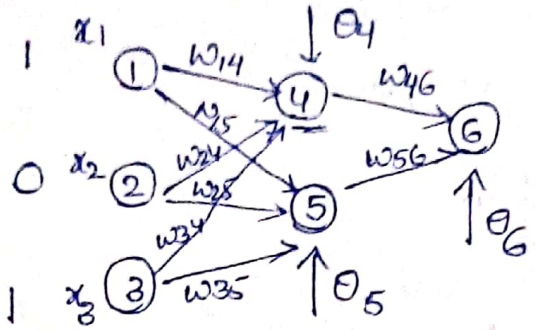
For each bias θ_j in N/w

$$\Delta \theta_j = (l) Err_j$$

$$\theta_j = \theta_j + \Delta \theta_j$$

[For each hidden or output layer unit j]

Ex:-



we are needed to get o/p as 1 if not modify the weights & modi

Given

x_1	x_2	x_3
1	0	1

$w_{14} = 0.2$ $\quad w_{34} = -0.5$ $\quad \theta_4 = -0.4$

$w_{15} = 0.3$ $\quad w_{35} = 0.2$ $\quad \theta_5 = 0.2$

$w_{24} = 0.4$ $\quad w_{46} = -0.3$ $\quad \theta_6 = 0.1$

$w_{25} = 0.1$ $\quad w_{56} = -0.2$

$O_j = I_j$

$O_1 = x_1$

$O_2 = x_2$

$O_3 = x_3$

unit j	input layer	output
4	$x_1 \times w_{14} + x_2 \times w_{24} + x_3 \times w_{34} + \theta_4$ $= 1 \times 0.2 + 0 \times 0.3 + (0.5)1 + (-0.4)$ $= -0.7$	$\frac{1}{1+e^{-x}} = \frac{1}{1+e^{-(-0.7)}}$ $= 0.332$
5	$x_1 \times w_{15} + x_2 \times w_{25} + x_3 \times w_{35} + \theta_5$ $= 1 \times 0.3 + 0 \times 0.1 + 1 \times 0.2 + 0.2$ $= 0.7$	$\frac{1}{1+e^{-x}} = \frac{1}{1+e^{-0.7}}$ $= 0.668$
6	$\text{Output}(4) \times w_{46} + O(5) \times w_{56} + \theta_6$ $= (-0.7)(-0.3) + (0.7)(-0.2) + 0.1$ $= 0.332(-0.3) + 0.668(-0.2) + 0.1$ $= -0.13 \neq 1$	$\frac{1}{1+e^{-(-0.13)}}$ $= 0.467$

so, we need to calculate Errors from back

$Err_j = O_j(1-O_j)(T_j - O_j)$

$Err_6 = O_6(1-O_6)(T_6-O_6) = 0.467(1-0.467)(1-0.467)$

$= 0.1326$

Errors for hidden layers → one layer means eliminated Σ_K as it has only one node blw

$Err_j = O_j(1-O_j) \sum_K err_k w_{jK}$

~~$Err_4 = 0.332(1-0.332)$~~

$err_5 = O_5(1-O_5)\, err_6 w_{56}$

$= 0.668(1-0.668)(0.132)(-0.2)$

$= -0.0058$

$err_4 = O_4(1-O_4)\, err_6\, w_{46}$

$= 0.332(1-0.332)(0.132)(-0.3)$

$= -0.0087$

Calculation for weights & bias (Modification)

~~Δw_{ij}~~ $= w_{ij} = w_{ij} + (l)\, err_j O_i$ → represents learning rate given in question

$w_{46} = w_{46} + 0.9\, err_6 O_4$

$= -0.3 + (0.9)(0.132)(0.332)$

$= -0.261$

$w_{56} = w_{56} + 0.9\, err_6 O_5 = -0.12$

$w_{14} = w_{14} + 0.9\, err_4 O_1 = 0.192$

$w_{15} = w_{15} + 0.9\, err_5 O_1 = 0.294$

$w_{24} = w_{24} + 0.9\, err_4 O_2 = 0.4$

$w_{25} = w_{25} + 0.9\, err_5 O_2 = 0.1$

$w_{34} = w_{34} + 0.9\, err_4 O_3 = -0.507$

$w_{35} = w_{35} + 0.9\, err_5 O_3 = 0.194$

Changing bias values

$\theta_j = \theta_j + (l)\, err_j$

$\theta_4 = \theta_4 + 0.9\, err_4$

$= -0.407$

$\theta_5 = \theta_5 + 0.9\, err_5 = -0.205$ ~~0.194~~

$\theta_6 = \theta_6 + 0.9\, err_6$

$= 0.219$

o/p

$\frac{1}{1+e^{-(-0.122)}} = 0.53$

0.32

dp 4 $\Rightarrow$ $w_{14} x_1 + w_{24} x_2 + w_{34} x_3 + \theta_4$

$= 0.192(1) + 0 + (0.507) + -0.407$

$= -0.722$

0.56

5 $\Rightarrow$ $w_{15} x_1 + w_{25} x_2 + w_{35} x_3 + \theta_5$

O(4) w_{46} O(5) w_{56}

6 $\Rightarrow$ $w_{46} \times x_4 + w_{56} \times x_5 + \theta_6$

$\frac{1}{1+e^{(0.539)}}$

~~$0.32 + 0.219 = 0.539$~~

$(0.32)(-0.261) + (0.56)(-0.12) + 0.219$ o/p: 0.5

$= 0.06$

Rule Based Classification :-

By using If-then, we make the rules to classify the data.

Ex:- If age = middle then buys-computer = yes

It will be work like Divide & conquer method

There are two methods in rule-based classification

- zero R
- one R

The General syntax for the Rule is IF condition -THEN Conclusion

Antecedent (or) precondition — condition

Rule consequent — Conclusion

(or)

(condition) $\Rightarrow$ y

class label

- To find out the rule is satisfied or not, we use coverage & accuracy to validate the rule.

$$\text{Coverage}(R) = \frac{n_{covers}}{|D|}$$ — no. of Tuples in data

$$\text{Accuracy}(R) = \frac{n_{correct}}{n_{covers}}$$

- Rules are constructed two ways
  - By using decision Tree
  - By using Neural Networks

5. Clustering

Collection of similar objects contains strong relationship and more dissimilar compare to other clusters
- Intra relation is stronger than inter relation in cluster

K-Means clustering Algo:-

↘ No. of clusters, we are going to dividing given large data sets

Ex:- Divide the given sample data into two clusters using K-Means Algorithm.

Height (H)	weight (w)
185	72
170	50
168	60
179	68
182	72
188	77
180	71
180	70

Consider any one as centroid

	Height	weight
C_1	185	72
C_2	170	50

C_2 ↓ centroid

Euclidean distance = $\sqrt{(x_H - H_1)^2 + (x_w - w_1)^2}$

(Observed value, Centroid value)

$C_1 = \{1\}$

$C_2 = \{2, 3\}$ → as C_2 gets 2 values, centroid changes

$C_2 \left(\frac{170+168}{2}, \frac{60+50}{2} \right)$

$C_2 (169, 55)$

row(R_3) = $\sqrt{(168-185)^2 + (60-72)^2}$

= 20.8

(C_1, C_2)

$C_2 \rightarrow \sqrt{(168-190)^2 + (60-50)^2} = 10.48$ ✓ least so write 3 under C_2

row(R_4): $C_1 \rightarrow \sqrt{(179-185)^2 + (68-72)^2} = \sqrt{52} = 7.2$ ✓ least go to C_1

$C_2 \rightarrow \sqrt{(179-169)^2 + (68-55)^2} = \sqrt{269} = 16.4$

C_1 {1, 4

C_2 {2, 3

Centroid

$C_1 \left(\frac{185+179}{2}, \frac{72+68}{2}\right) = (182, 70)$

Row (R_5) $C_1 \rightarrow \sqrt{(182-182)^2 + (72-70)^2} = 2$ ✓least

$C_2 \rightarrow \sqrt{(182-169)^2 + (72-55)^2}$

C_1 {1, 4, 5

C_2 {2, 3

Centroid

$\left(\frac{182+182}{2}, \frac{70+72}{2}\right) = (82, 71)$

Row (R_6) $C_1 \rightarrow \sqrt{(188-182)^2 + (77-71)^2} = \sqrt{36+36}$ ✓least

$C_2 \rightarrow \sqrt{(188-169)^2 + (77-55)^2}$

Row (R_7) C_1 {1, 4, 5, 6

C_2 {2, 3

Centroid

$\left(\frac{182+188}{2}, \frac{71+77}{2}\right) = (185, 74)$

$C_1 \rightarrow \sqrt{(180-185)^2 + (71-74)^2} = 5.83$ ✓least

$C_2 \rightarrow \sqrt{(180-169)^2 + (71-55)^2} = 19.41$

C_1 {1, 4, 5, 6, 7

C_2 {2, 3

Centroid

$\left(\frac{185+180}{2}, \frac{74+71}{2}\right) = (182.5, 72.5)$

Row (R8)

$C_1 \rightarrow \sqrt{(180-82.5)^2 + (70-72.5)^2} = 35$ ✓least

$C_2 \rightarrow \sqrt{(80-169)^2 + (70-55)^2} = 18.6$

C_1 {1, 4, 5, 6, 7, 8}

C_2 {2, 3}

Clustering : 1) partitional Based 2) hierarchial
3) density based 4) clustering lB Agglomerative Divisive

- In hierarchial clustering method works by grouping data objects into hierarchy or tree of clusters
- There are different sub methods in hierarchial clustering.
  1. Agglomerative Nesting VS Divisive analysis hierarchial clustering.
  2. Distance measures in algorithmic methods
  3. BIRCH (balanced iterative reducing & clustering using hierarchies) multiphase hierarchial clustering
  4. chameleon multiphase hierarchial clustering.

Agglomerative Nesting VS Divisive :-

A hierarchial clustering method can be Either agglomerative or divisive, depending on whether the hierarchial decomposition is formed in a bottom up (merging) or Top down (spliting) approach.

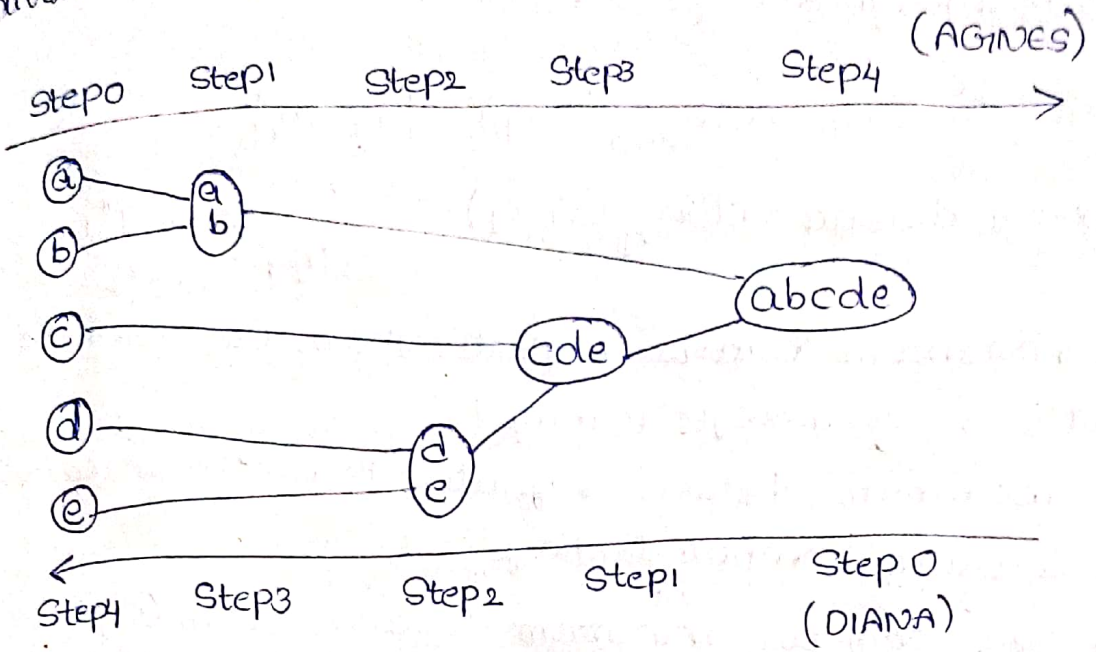
Agglomerative hierarchial clustering method :-

This method uses bottom up strategy Initially, Each object having own cluster and Each object combined with other object to form a new cluster until all the objects are in a single cluster. Or certain Termination conditions are satisfied. The single cluster becomes the hierarchy root

Divisive hierarchial clustering method :-

This method uses top down strategy it starts by placing all objects in one cluster which is a hierarchies root it then divides the root cluster into several smaller sub cluster and recursively partitions those clusters into small ones. The partitioning process continues until Each cluster at lowest level is coherent Enough Either containing only one objects.

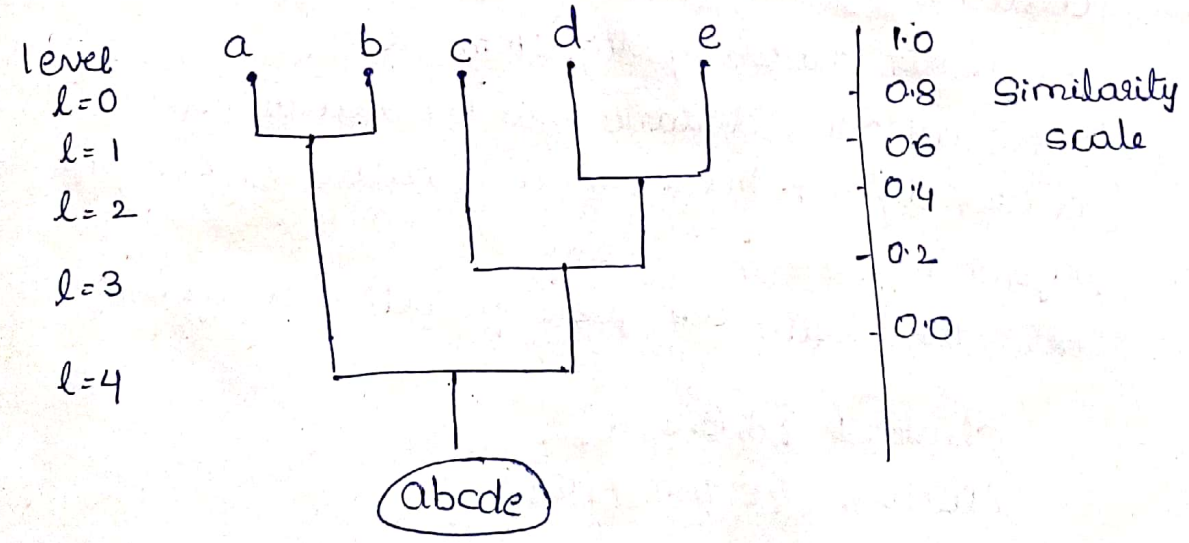
The Agglomerative hierarchial clustering also called as AGNES and divisive hierarchial clustering also called as divisive analysis Diana.



Dendogram :-

A Tree structure called a dendogram is commonly used to represent the process of hierarchial clustering.

It shows how objects are grouped Together on partitioned step by step.



Dendogram Representation

2) Distance measures in Alg methods :-

→ Nearest Neighbour

minimum distance $dist_{min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} \{|p - p'|\}$

Maximum Distance : $dist_{max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} \{|p - p'|\}$

/

Farthest Neighbour

Mean Distance : $dist_{mean}(C_i, C_j) = |m_i - m_j|$

Average distance : $dist_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i, p' \in C_j} |p - p'|$

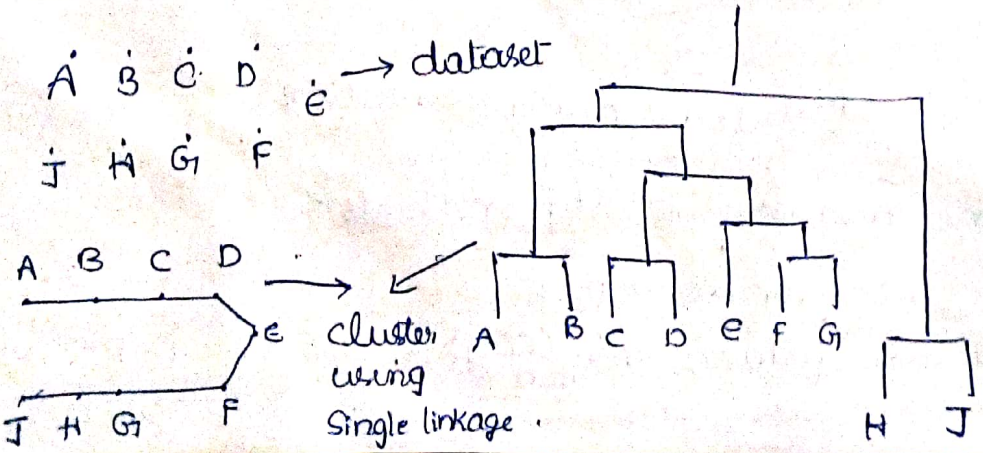
If minimum distance is greater than the threshold value is called single leakage

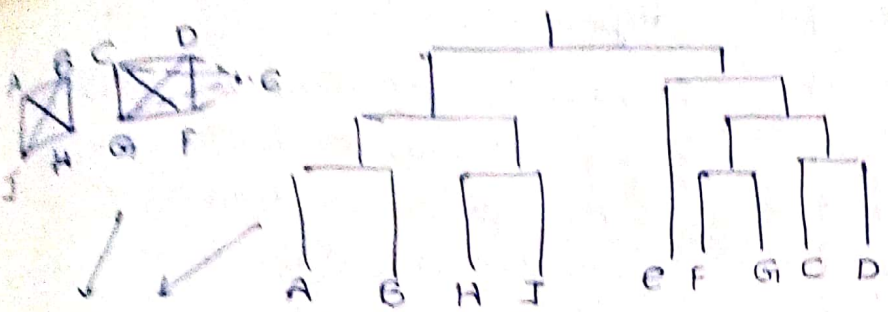
If maximum distance is greater than threshold value it is called complete linkage.

- An Algorithm uses minimum distance, $d_{min}(C_i, C_j)$ to measure the distance b/w clusters, it is sometimes called as nearest neighbour clustering Algorithm.
- When the min distance b/w nearest clusters exceeds a user defined threshold is called single linkage Algo.
- When an algorithm uses $d_{max}(C_i, C_j)$ to measure the distance b/w clusters it is sometimes called as farthest neighbour clustering Algorithm.
- When the max distance b/w the nearest clusters exceeds a user defined threshold it is called as complete linkage Algorithm.
- Here the clusters are given for below diagrams.

Cluster 1 {A, B, I, H}

Cluster 2 {C, D, G, F, E}





cluster using complete linkage.

partitional based clustering :-

K-means clustering :-

K → No. of clusters, means → mean

Ex:- {2, 4, 10, 12, 3, 20, 30, 11, 25} apply K-means algo and grp the data into two clusters

$K = 2 \quad m_1 = 2 \quad m_2 = 4$

$K_1 = \{2, 3\}$

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

$m_1 = 2.5 \quad m_2 = 10$

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

m_1	m_2	K_1	K_2
3	18	{2, 3, 4, 10}	{12, 20, 30, 11, 25}
4.75	19.6	{2, 3, 4, 10, 11, 12}	{20 30 25}
7	25	{2, 3, 4, 10, 11, 12}	{20 30 25}

K-medoids Algorithm (or) PAM Algorithm :-

PAM (partitioning Around mediods)

medioid is nothing but a data point which contains the most similarity with other datapoints in cluster

while making the data into clusters a medioid is a compulsory datapoint in a cluster

Ex:-

Document	x	y
d_1	2	6
d_2	3	4
d_3	3	8
d_4	4	7
d_5	6	2
d_6	6	4
d_7	7	3
d_8	7	7

Algo:-

1. Select k no of mediods randomly
2. we calculate the distance from mediods to other data points
3. calculate the total cost in making clusters
4. If previous total cost is less than the present one while making the clusters stop the procedure.

Sol:- K = 2 (3,4) (7,4)

Manhalton distance method $|x_1-y_1| + |x_2-y_2|$

	(3,4)	(7,4)	
d_1	$\|2-3\|+\|6-4\|=3$	$\|2-7\|+\|6-4\|=7$	
d_3	4	8	Cluster 1 = {d_1, d_3, d_4}
d_4	4	6	cluster 2 = {d_5, d_6, d_7}
d_5	5	3	Total cost = 16
d_6	3	1	we will choose some other
d_7	5	1	mediod (3,4) (7,3)

Iteration 2:

	(3,4)	(7,3)	
d_1	3	8	Cluster 1 = {d_1, d_3, d_4}
d_3	4	9	Cluster 2 = {d_5, d_6, d_8}
d_4	4	7	Total cost = 16
d_5	5	2	
d_6	3	2	
d_8	4	1	

Hierarchial:-

By using following Technique we develop hierarchial for in distance measure.

Clustering using single link.

	x	y
P_1	0.40	0.53
P_2	0.22	0.38
P_3	0.35	0.32
P_4	0.26	0.19
P_5	0.08	0.41
P_6	0.45	030

1. find out the weighted distance matrix for the given data by using Euclidean distance
2. Find out the min value among all the distances b/w two data points & make them cluster.
3. continue the procedure until all the datapoints merge into single data cluster

Sol:

	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0					
P_2	0.234	0				
P_3	0.22	0.15	0			
P_4	0.37	0.20	0.15	0		
P_5	0.34	0.14	0.28	0.29	0	
P_6	0.23	0.25	0.11	0.22	0.39	0.

cluster small.

P_3P_6 cluster

	P_1	P_2	$\{P_3\ P_6\}$	P_4	P_5
P_1	0				
P_2	0.234	0			
$\{P_3,P_6\}$	0.22	0.15	0		
P_4	0.37	0.20	0.15	0	
P_5	0.34	0.14	0.28	0.29	0

For $\{P_3,P_6\} \rightarrow \min\{\text{dist}\{P_3,P_1\}\ \{P_6,P_1\}\}$

$\{0.22\ \ 0.23\} = 0.22$

$\min\{(P_3,P_1),(P_6,P_2)\} = 0.15$

	P_1	P_2P_5	P_3P_6	P_4
P_1				
P_2P_5				
P_3P_6				
P_4				

process until we get single cluster continues

Nearest Neighbour clustering Method:-

It is an example for the partitioned based clustering Method. By using single link techniques, we calculate the clusters for given data sets.

In this Techniques, Items are iteratively merged into the existing clusters which are closest.

In this Algorithm, a Threshold value t is used to determine, If items will be added to existing clusters/not. If the distance b/w the two items is less than the threshold value, we club into single cluster.

Ex:- Threshold $t = 2$

data = {1, 2, ③, 4, 5, 7}

Taken value

$K_1 = \{1, 2, 3\}$ $2-1 = 1 < t$

$3-1 = 2 = t$

$K_2 = \{4, 5\}$ ← $4-1 = 3 > t$

$K_3 = \{7\}$

Density Based clustering:-

clustering based on density (local cluster criteria) such as density connected points are based on Explicitly constructed density function.

Advantages -

- Discover clusters of different shapes
- Handle noise
- It uses single scan
- It requires Density parameters

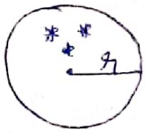
Ex:- DB Scan

DBSCAN :-

Density Based Spatial clustering of applications with Noise

Density : No of points within a specified radius(r) ~~area~~ :-

– ESP or r or ϵ



Core point – A point is a core point, if it has more than specified no of points (Minpts) within r

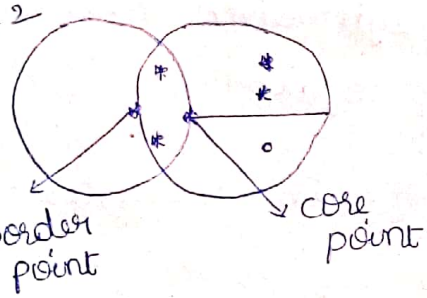
These are the points that are interior of a cluster

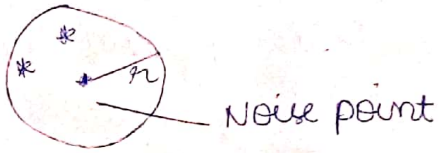
Border point – A border point has less than minpoints within r but is in the neighbourhood of a core point

Noise point :- It is not a core point nor a border point-

Minpts = 4

eps(or) r = 2





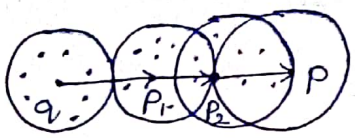
Algorithm :-

- for each a ∈ D do
  - if a is not yet classified then
  - if a is a core point then
    - collect all objects density reachable from a
    - and assign them to a new cluster
  - Else
    - assign to Noise

Density - Reachability :-

→ A point q is directly density reachable from a point p if p is a core point and q is p's ϵ-Neighbourhood

Ex:-

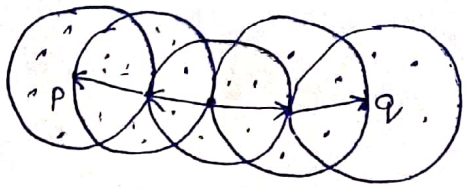


Minpts = 7

$q \rightarrow P_1 \rightarrow P_2 \rightarrow P$

Density connectivity :-

A pair of points p & q are Density-connected, If they are commonly density-Reachable from a point x.



minpts = 7

Density-connectivity - Symmetric

Density-Reachability - Asymmetric

BIRCH - Balanced Iterative Reducing & clustering using hierarchies

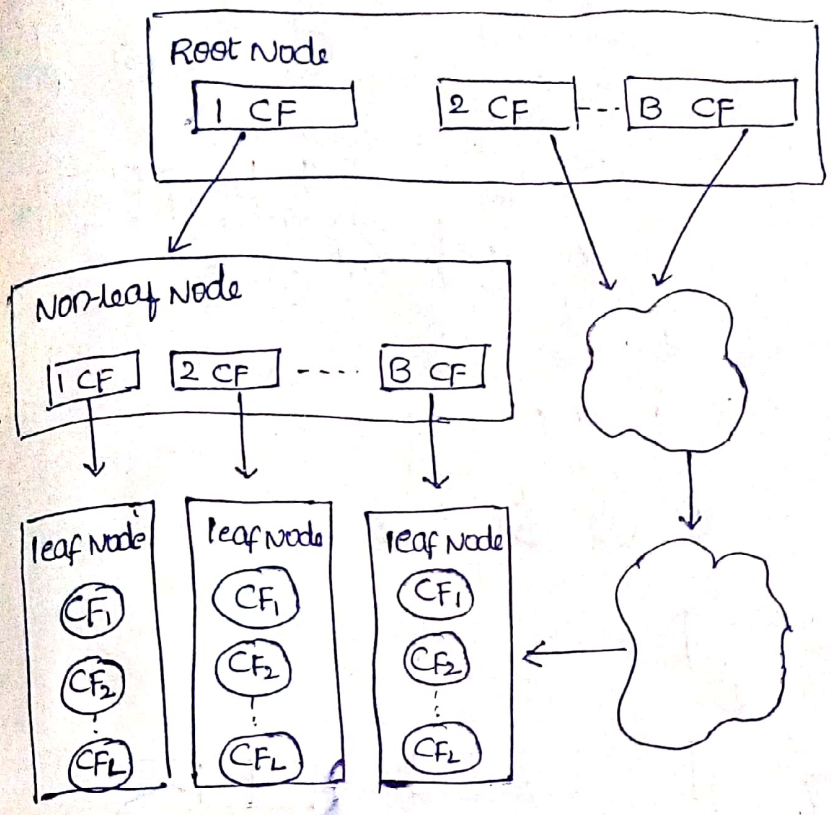
- It is the scalable clustering method
- Designed for very large datasets
- It requires only one scan of data if necessary
- It is based on the notation of CF (clustering Feature) of a CF tree
- A CF tree is a height balanced Tree that stores the clustering features for a hierarchial clustering
- Cluster of data points is represented by Triple no's (N, LS, SS)

N = No. of items in sub-cluster

LS = Linear sum of the points

SS = Sum of the squares of the points

CF-Tree



B - Balanced factor

L - Threshold value

∴ Height Balanced Tree with two parameters

1. Balancing factor (B) which indicates atmost the pre-Entries in non-leaf Node
2. ~~(Each Leaf Node has atmost)~~ Threshold (L) which indicates Each leaf Node has atmost L Entries (L CF Entries)

The Main Advantage of BIRCH Algo is to handle memory restrictions
- It finds a gud clustering with a single scan & improve quality.

Disadvantage - handles only numerical data

Applications :-
- pixel classification in Images
- Image compression
- Speech Enhancement

$CF = (N, LS, SS)$

$LS: \sum_{i=1}^{N} x$

$SS: \sum_{i=1}^{N} x^2$

Ex: (3,4)(2,6)(4,5)(4,7)(3,8)

$N = 5$

$LS = 3+2+4+4+3 = 16$

$4+6+5+7+8 = 30$

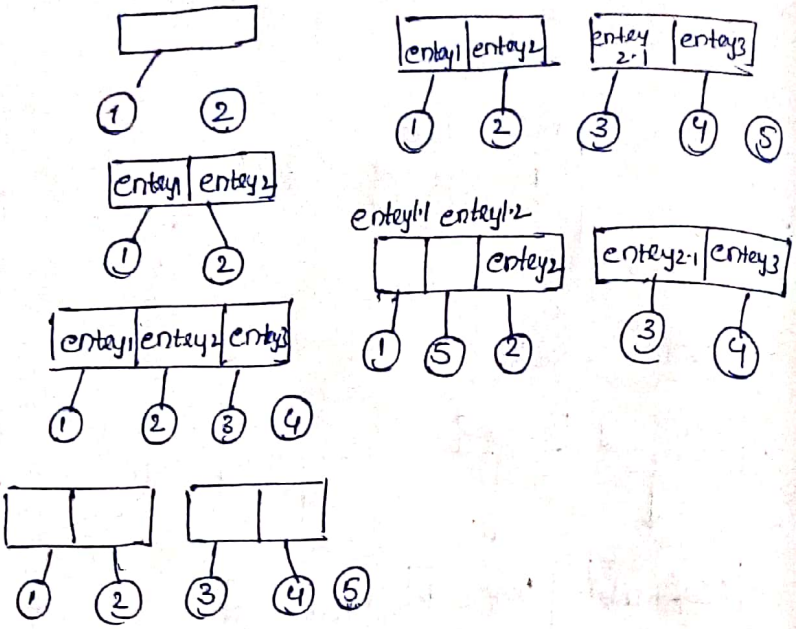
$SS = 3^2+2^2+4^2+4^2+3^2 = 54$

$4^2+6^2+5^2+7^2+8^2 = 190$

(5, (16,30)(54,190))

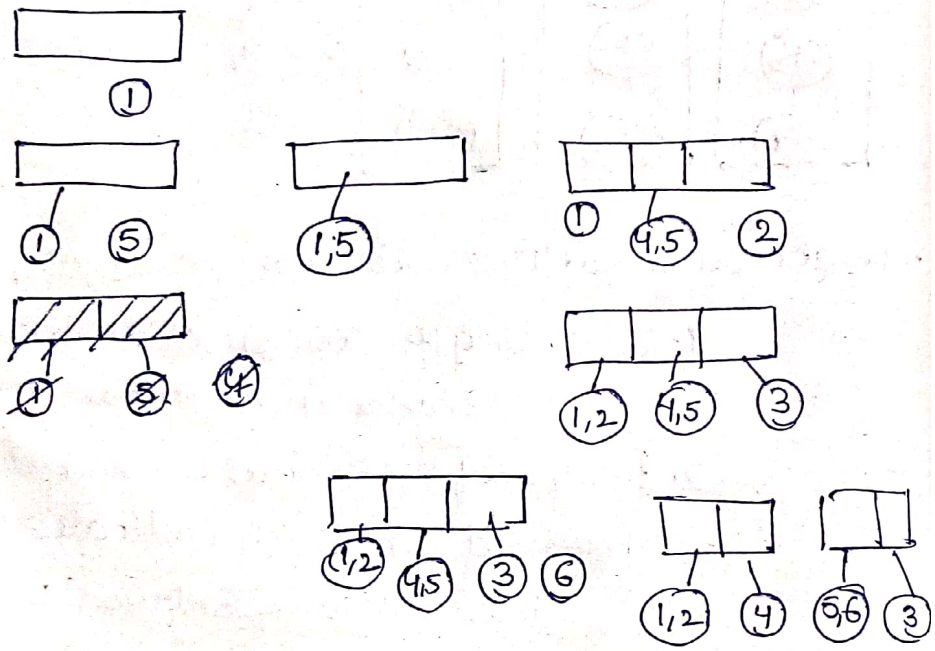
Ex:- B=3
T=1

data ①
②
③
④
⑤



Ex:- B=3
T=2

① ⑤ ④ ② ③ ⑥



Similarity & Distance Measures :-

Similarity Indicates a tuple within one cluster is more similar within that cluster than it is similar to outside of clusters. by using distance measures, we can divide the data into the different clusters

$$dist(t_i, t_j) \rightarrow t_i \in K_i,\ t_j \in K_j,\ t_i \notin K_j,\ t_j \notin K_i$$

Techniques to find Distances :-

1. Centroid $C_m = \frac{\sum_{i=1}^{N} (t_m)}{N}$

2. radium $R_m = \sqrt{\frac{\sum_{i=1}^{N} (t_{mi} - c_m)^2}{N}}$

3. Diameter $D_m = \sqrt{\frac{\sum_{i=1}^{N} \sum_{j=1}^{N} (t_{mi} - t_{mj})^2}{(N)(N-1)}}$

- The Radius is the square root of the Average mean square distance from any point to clusters to the centroid.
- Diameter is the square root of the Average mean square distance b/w all pairs of points in the clusters.

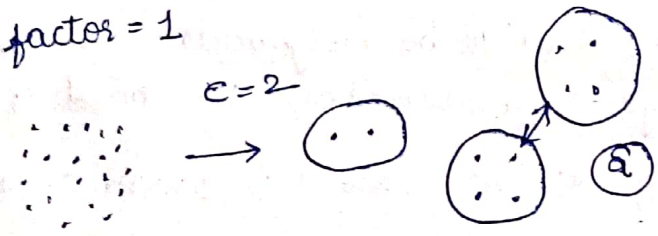
the Drawback of these methods is Time consuming process and costly operation. So replace, these methods by following techniques
- Single link
- complete link
- Average link
- centroid
- Medoid

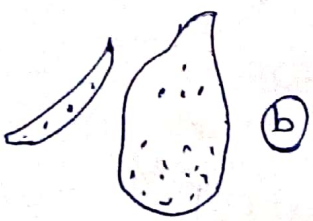
CURE (Clustering with Representatives) :-

It is an Example for both partitional & hierarchial cluster.

- A constant no of points c is choosen from each clusters and shrink towards the each clusters, If shrinking factor = 1

Ex:-





Clustering with categorical data :-

{ water, book, sun, sand, swim, read }

doc1 = { book } ≅ (1,0,0,0,0,0)

doc2 = { water, sun, sand, swim } = (0, 1, 1, 1, 1, 0)

doc3 = { water, sun, swim, read } = (0, 1, 1, 0, 1, 1)

doc4 = { read, sand } = (0,0,0,1,0,1)

By using Euclidean distance

	1	2	3	4
1	0			
2	2.24	0		
3	~~2.24~~	1.41	0	
4	1.73	2	2	0

→

	1	[2,3]	4
1	0		
(2,3)	2.24	0	
4	1.73	2	0

But there is no similarity b/w doc1 & 4, so, we are going to Rock algorithm.

ROCK Algo :- (Robest clustering with links)

- This Algo works for both categorical & boolean data.
- It is identifying the no of links b/w the nodes (or) docs.
- If the no of links greater than the threshold value, we can combine the data into single cluster.
- The pair of items are said to be neighbours, if they exceeds the links b/w objects greater than threshold value.
- Instead of Euclidean distance, we use Jacad Quotient to find out the similarity b/w two documents.

$$Sim(t_i, t_j) = \frac{|t_i \cap t_j|}{|t_i \cup t_j|}$$

	1	2	3	4
1	1	0	0	0
2	0	1	0.6	0.2
3	0	0.6	1	0.2
4	0	0.2	0.2	1

2,3 → $\frac{3}{5}$ = 0.6

2,4 → 1/5 = 0.2

(2,4) (2,3) (3,4) for threshold t = 0.2

Bond Energy clustering Algo:-

- It is used in the data base design to determine how to group the data and how to place physically data on to the disk.
- with BEA, Find out the affinity (bond) b/w the db attributes based on common usage.
- The basic objective of BEA Algo is Attributes that are used together form a cluster and should be stored together and resulting cluster called as vertical fragments.

$$Aff = \sum_{i=1}^{n} (bond(A_i, A_{i-1}) + bond(A_i, A_{i+1}))$$

Clustering with Genetic Algo:-

- To Apply this convert the given data into binary format

Ex:- {A, B, C, D}

{A, D} {B, C}

{1, 0, 0, 1} {0, 1, 1, 0}

Algorithm:-

Step 1: Randomly Create an initial solution.

Repeat:

use crossover to create a new step

until terminator criteria is met

Ex:- {A, B, C, D, E, F, G, H}

⇓, 3 clusters

{A, C, E} {B, F} {D, G, H}

Cross-over :-

{A B C D E F G H}

3-clusters

{A, C, E} {B, F} {D, G, H}

⇓ ⇓ ⇓

1010|1000 01000100 0001|0011

bit-map representation

{A, B, C, D}

{A, D} {B, C}

{1001} {0110}

① & ③ are clusters apply crossover

00011000 01000100 10100011

{D, E} {A, C, G, H}

Clustering with neural Networks :-

The neural networks that used unsupervised learning attempt to find the feature in data that characterize the design o/p.

The neural networks search for the clusters with requir links to ~~get~~ the given data.

There are two types of neural networks in unsupervise learning - Non-competitive

- competitive

- Non competitive or Hebbian learning :-

The weight b/w two nodes ~~is~~ of change is proportional to the o/p values.

$$\Delta w_{ji} = \eta y_j y_i$$

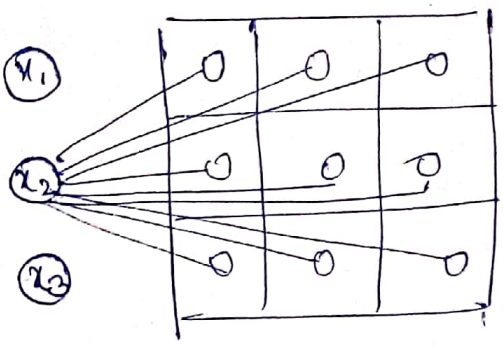
These types of n/w also called as Self-Organising neural networks

To construct Self-organising neural n/w, we use Self-organising feature maps.

This algorithm works like the behavior of a node should impact only those nodes and edges which are nearer to that node.

- The best ex of Self-organising neural n/w is

Kohonen Self-organising map



$$Sin(x, i) = \sum_{j=1}^{h} x_j w_{ji}$$

$$\Delta w_{kj} = \begin{cases} c(x_{ik} - w_k) & \text{if } j \in N_i \\ 0 & \text{otherwise} \end{cases}$$

In Kohonen Self organising map there are two layers.

1. Input layer
2. competitive layer

- In competitive layer the nodes are arranged in 2D format every node is associated with activation function.
- every layer in the competitive one produces an output and which layer output is equal to the expected output make all the nodes in the layer into single cluster