Regression Model

$$y = 0.9 + 1.2x_1 + 2x_2 + 4x_3 + 1x_4$$

* Simple linear Equation

$$y = \alpha_0 + \alpha_1 x_1$$

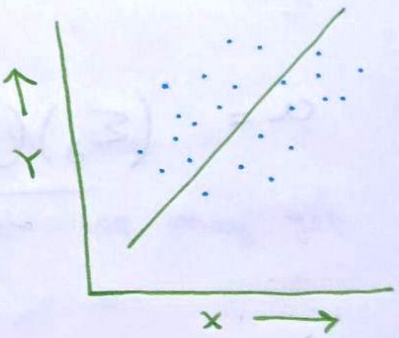
* Multiple Linear Equation

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \dots + \alpha_n x_n$$

α_i = Regression coefficient

x_i = Independent variable

y = Dependent variable.



* Regression Model provides a function that describes the relationship b/w one or more independent variables and a response variables or dependent variables.

eg. Relationship b/w height and weight may described by a linear regression model.

- Dependent variable is in continuous in nature eg. Data set;

Second call	Profit
23	1
28	2
39	3
48	3
64	4
75	4
88	6

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

Ques. Linear Regression Question

x	y	xy	x^2
1	3	3	1
2	4	8	4
3	5	15	9
4	7	28	16
10	19	54	30

$$a = \frac{(19)(30) - (10)(54)}{4 \times 30 - 100} = \frac{570 - 540}{120 - 100} = \frac{30}{20} = 1.5$$

$$b = \frac{(4)(54) - (10)(19)}{120 - 100} = \frac{26}{20} = 1.3$$

$$y = bx + a$$

$$y = 1.3x + 1.5$$

A linear regression is a model where the relationship between dependent & independent variable is a straight line

> one ex. around the no. of responses to the market campaign. If we send 1000 emails we may get 5 response.

If this relationship is modelled using this reg. model we could expect to get 10 responses for 2000 emails sent.

Simple linear Equation :~

$$y = \alpha_0 + \alpha_1 x_1$$

p	error
2.8	0.2
4.1	0.1
5.4	0.4
6.7	0.3

Logistic Regression :~

Time	clicked on AD
68.95	NO
80.25	NO
69.45	NO
74.15	NO
50	YES
55.5	YES
80.0	NO
70.5	NO

$$\boxed{y = \frac{1}{1 + e^{-x}}}$$

SIGMOID Eqⁿ

It estimates the probability of an event occuring such as voted or didn't vote based on given data set of independent variable.

Since outcome is probability, the dependent variable are bounded between 0 and 1.

It is simply trying to convert the inde var into exps of prob that ranges between 0 & 1 with dependent variable.

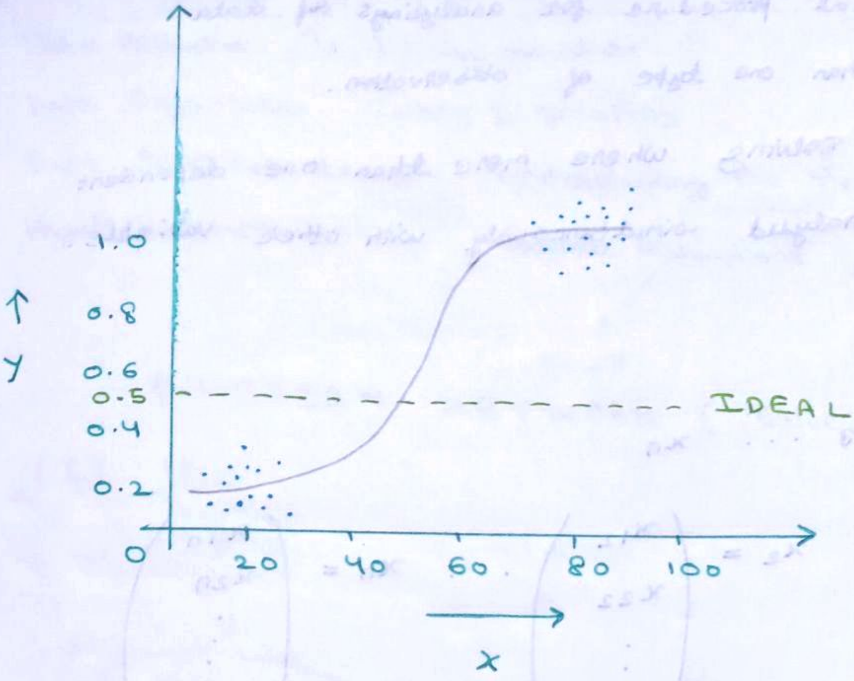
Applications :~

Fraud Detection

Disease diag

Emergency Detection

Spam or no spam.



MULTIVARIATE ANALYSIS :~

$X = \overset{\text{Salary}}{x}$; $x_1, x_2 \dots x_n$ ⟶ n employ (univariant)

$X \leftarrow$ Economic situation, academics etc.

$X \rightarrow X_1, X_2 \dots X_p$ p = number of Random variables

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_p \end{pmatrix}$$

x_{ij} ; i ↙ variable, j ↘ Individual

	ind 1	ind 2	.	.	ind n
var 1	x_{11}	x_{12}	.	.	x_{1n}
var 2	x_{21}	x_{22}	.	.	x_{2n}
var 3	x_{31}	x_{32}	.	.	x_{3n}
⋮	⋮	⋮			⋮
var p	x_{p1}	x_{p2}	.	.	x_{pn}
	↓ obser 1	↓ obs 2			↓ obs n

* It is statistical procedure for analysis of data involving more than one type of observation.

* It is also solving where more than one dependent variables is analysed simultaneously with other variable.

Techniques :~

$X = x_1 , x_2 , x_3 \dots x_n$

$$x_1 = \begin{pmatrix} x_{11} \\ x_{21} \\ x_{31} \\ \vdots \\ x_{p1} \end{pmatrix} \quad x_2 = \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{p2} \end{pmatrix} \dots \quad x_n = \begin{pmatrix} x_{1n} \\ x_{2n} \\ \vdots \\ x_{pn} \end{pmatrix}$$

① Sample Mean for Multivariate Analysis

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

② Sample Variance

Transpose

$$S_n = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$$

$$\bar{x} = \frac{1}{n} \begin{bmatrix} x_{11} + x_{12} + \dots + x_{1n} \\ x_{21} + x_{22} + \dots + x_{2n} \\ \vdots \\ x_{p1} + x_{p2} + \dots + x_{pn} \end{bmatrix} = \begin{pmatrix} \bar{x_1} \\ \bar{x_2} \\ \vdots \\ \bar{x_n} \end{pmatrix}$$

Obj

- Data Red
- Data Or
- Data I
- Hypothes

(I)

(II)

*

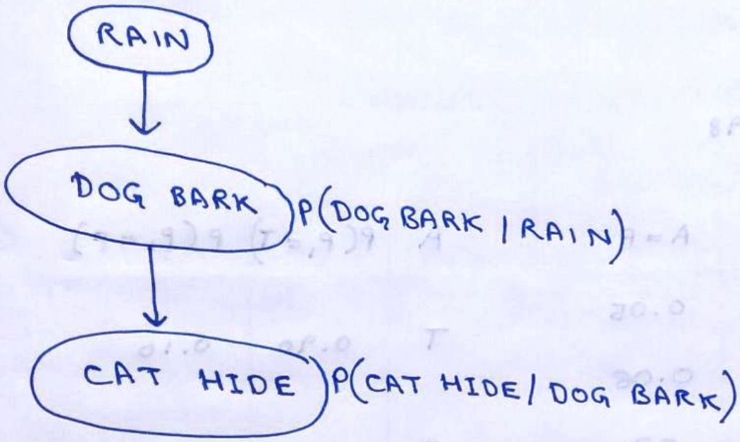
*

Objective of Multivariate Analysis

- Data Reduction - To simplify the data
- Data Organisation - Sorting & grouping
- Data Interdependency - understanding the relationship b/w variables.
- Hypothesis construction - Helps validate Assumptions

BAYESIAN NETWORK (Belief Network)

(I) DAG



(II) CONDITIONAL PROBABILITY :~

	R	~R
B	9/48	18/48
~B	3/48	18/48

(B = T & R = T) = 0.19

(B = T & R = F) = 0.375

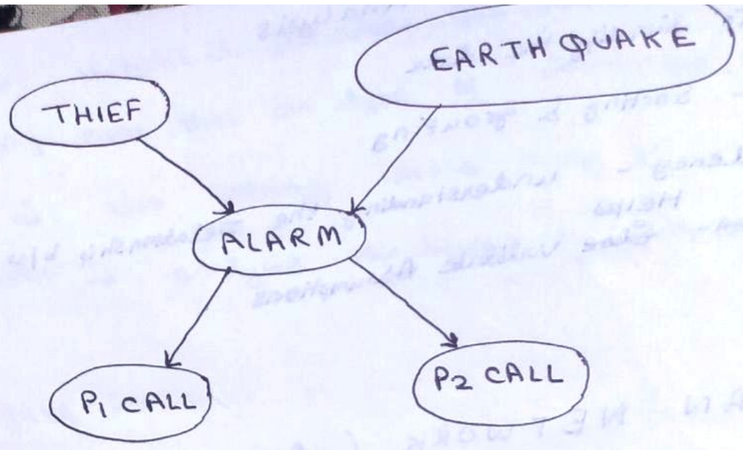
(B = F & R = T) = 0.06

(B = F & R = F) = 0.375

* It is probabilistics graphical model representation of set of variables and condition via Dag.

* It can be used for building for Models from data & experts opinions and it contains of two parts

(i) Dag (ii) Table of conditional probability



$P(TH=T) = 0.001$

$P(E=T) = 0.002$

$P(TH=F) = 0.999$

$P(E=F) = 0.998$

TH	E	A = T	A = F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

A	$P(P_1=T)$	$P(P_1=F)$
T	0.90	0.10
F	0.050	0.95

A	$P(P_2=T)$	$P(P_2=F)$
T	0.70	0.30
F	0.01	0.99

$$P(P_1, P_2, A, \sim TH, \sim E) = P(P_1/A) . P(P_2/A) . P(A/\sim TH \sim E) . P(\sim TH)\, P(\sim E)$$

$$= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998$$

$$= 0.00062$$

In the mul

dependent

as follow

No. of

Sold

4

5

TIME SERIES :~

understand , interact and access chronological changes in the values of a variables in the past , so the reliable prediction can be made about future value.

Component of Time Series :~

(i) T(Secular Trends)

* movement over all long terms

(ii) S(Seasional variations)

* Variation with one year that Repeated more or less regular.

(iii) c (Cyclical Variation)

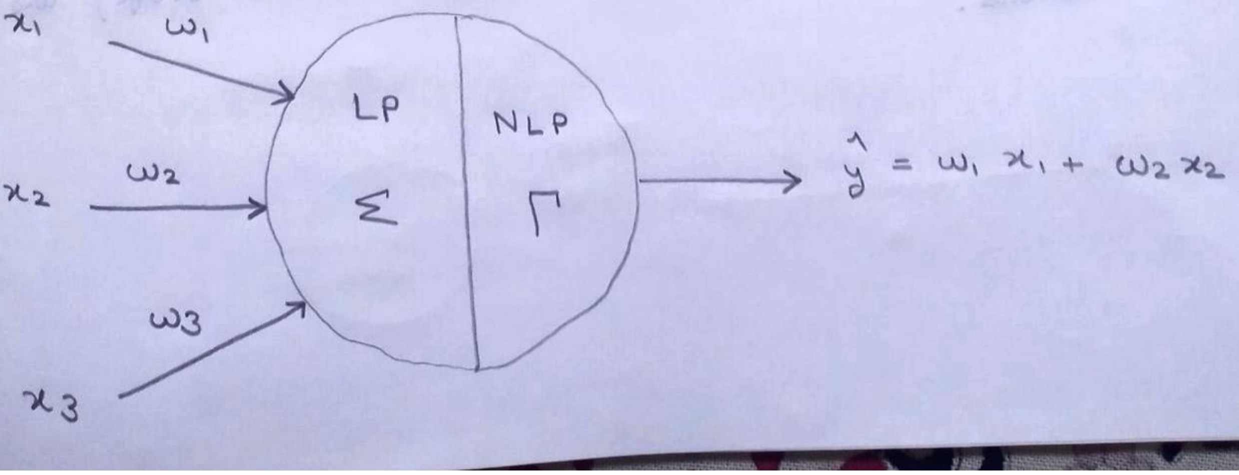
Due to ups & down after a period from time to time.

(iv) I (Irregular variation)

$$x_t = T \times S \times C \times I \quad \text{(Multiplicative Model)}$$

$$x_t = T + S + C + I \quad \text{(Additive Model)}.$$

Neural Network :~

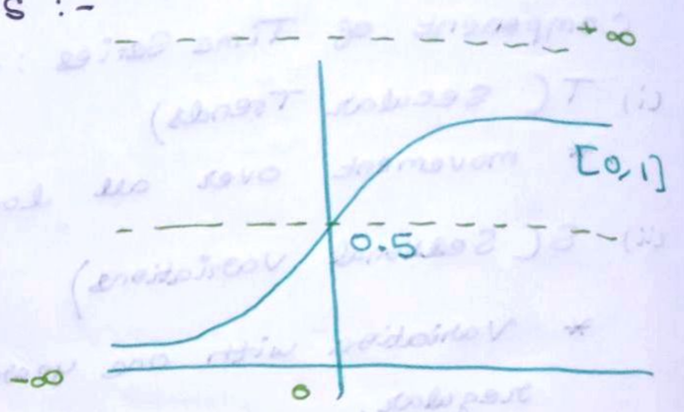


Mc CuLLocH PITTS Newron (M CP) 1943

Residual Network

Google Inspection

STRUCTURE OF NEURONS :-



* LF → Linear function

* NLF → Non-Linear function

Activation function

$$\hat{y} = \frac{1}{1+e^{-z}}$$

$$z = x_1\omega_1 + x_2\omega_2 + x_3\omega_3 + \cdots x_n\omega_n$$

- It is computational learning system that uses a network of function to understand & translate a data form into one form into desired form.
- MCP in 1943 is discovered. It may be divided into two parts. The first part takes an input and perform an aggregation & based on aggregated value
- And second part to make a decision.
- Non-linear function helps any neuron from collapsing.

for eg. Google Inspection & Residual Network (Microsoft) are two neurons.

Learning & Generalisation :~

A network will produce for input pattern that it was not originally setup for classifying.

LEARNING :~

The network must learn decision surfaces from a set of training patterns so that these training patterns are classified correctly.

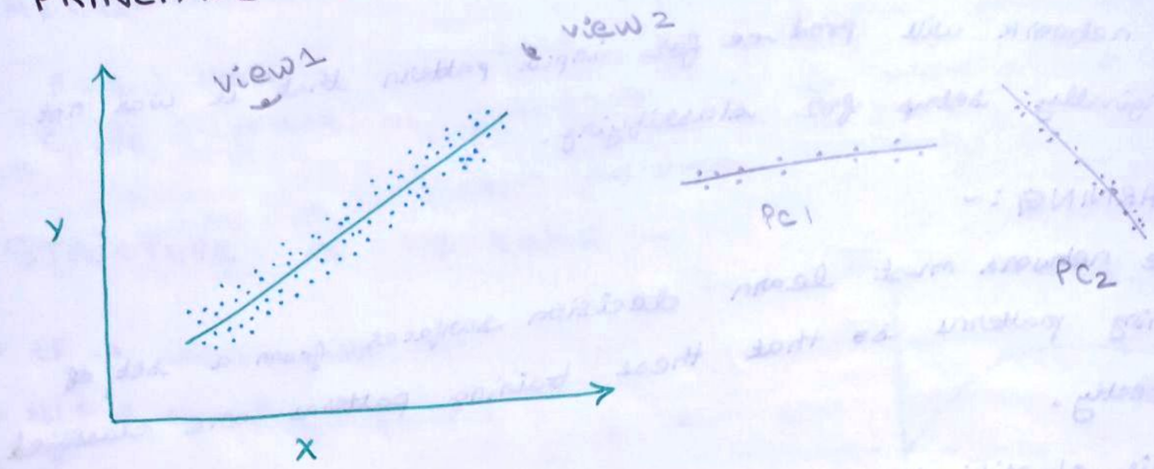
This classification is used for binary classification.

GENERALISATION :~

After training the network must able to generalize ie. correctly classify test pattern it has never seen before.

We wish the network to learn perfectly and then generalisation well with small generalisationerror.

PRINCIPAL COMPONENT ANALYSIS (PCA) :-



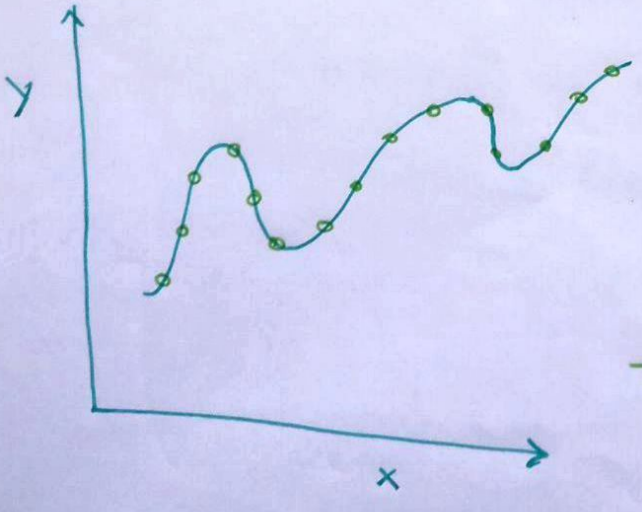
* It is unsupervised learning Algorithm that is used for dimension reduction in machine learning.

* It is statistical process that converts the observations of correlated features into a set of linear uncorrelated features with the help of orthogonal transformation.

* All principal component should be less than or equal to no. of attributes of data set.

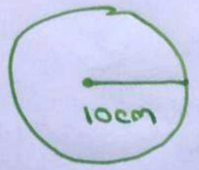
* It is based on variance, eigen values & eigen vector.

overfitting

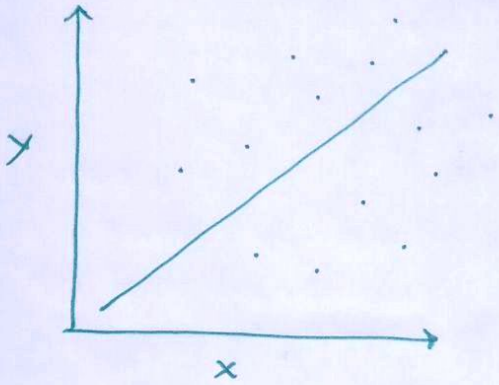


Sphere
Play
eat
Radius = 5cm

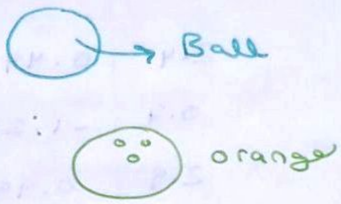
Test
Set



underfitting



Sphere



Test set :

Q. Given data sets. Compute PCA

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

$\bar{x} = 1.18$ $\bar{y} = 1.91$

$$C = \begin{bmatrix} \text{cov}(x, x) & \text{cov}(x, y) \\ \text{cov}(y, x) & \text{cov}(y, y) \end{bmatrix}$$

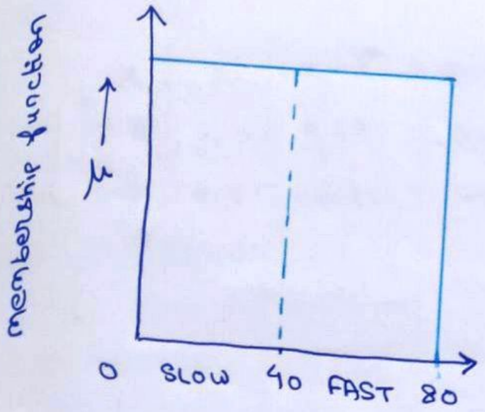
$$\text{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

x	$x - \bar{x}$	$(x - \bar{x})(x - \bar{x})$
2.5	0.69	0.48
0.5	-1.31	1.716
2.2	1.02	1.04
1.9	0.72	0.52
3.1	1.92	3.69
2.3	1.12	1.25
2	0.82	0.67
1	-0.18	0.03
1.5	0.32	0.102
1.1	-0.08	0.0064

Sum = 5.54 90

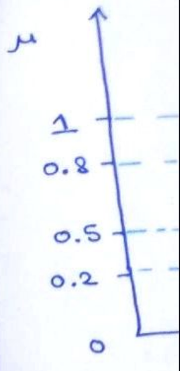
FUZZY LOGIC :~

- * It was discovered by Lotfi Zadeh.
- * It represents uncertinity [0, 1]
- * It represents with degree.
- * It represents the belongingness of member of crisp set to fuzzy set.
- * It is mathematical language.
- * Relational logic + Boolean logic + predicate logic



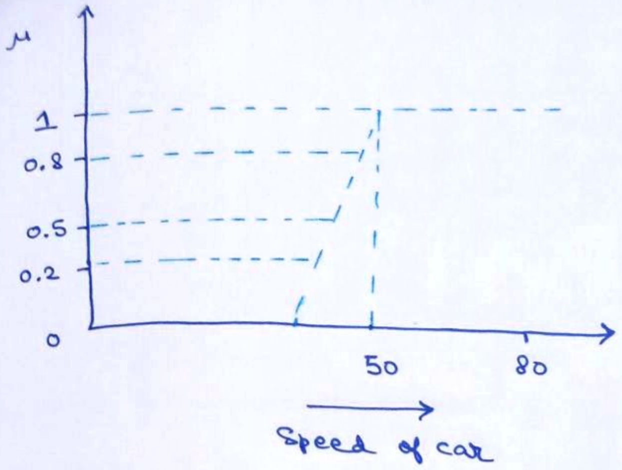
Check the Degree of fastness :~

$$\begin{cases} 0 & \text{if speed}(x) \leq 40 \\ \dfrac{\text{Speed}(x) - 40}{10} ; & \text{if } 40 < \text{speed}(x) < 50 \\ 1 & \text{if speed}(x) \geq 50 \end{cases}$$



C

Item



$x = 30 \quad (30,0)$

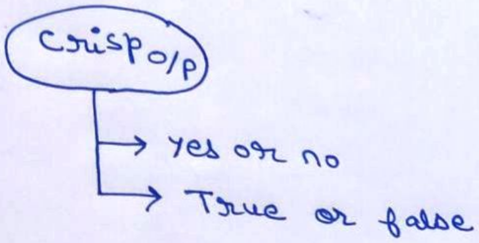
$x = 60 \quad (60,1)$

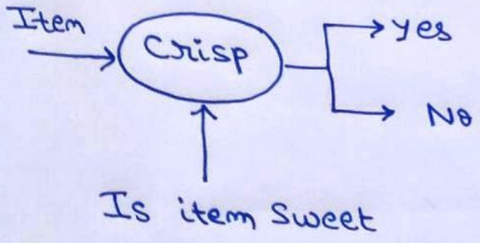
$x = 42 \quad \frac{42-40}{10} = 0.2$

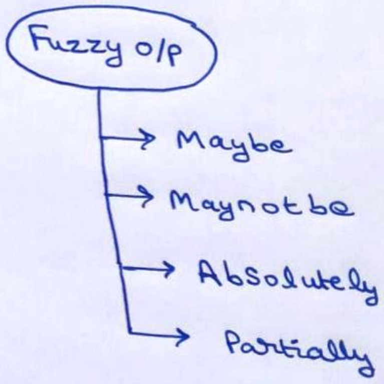
$x = 45 \quad \frac{45-40}{10} = 0.5$

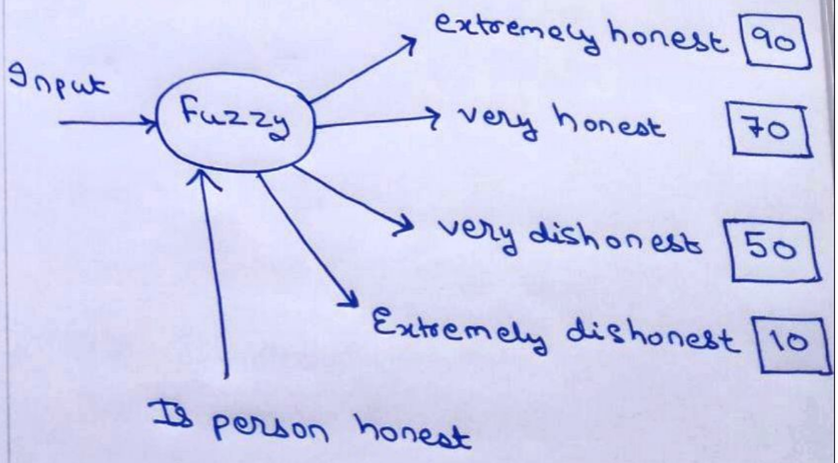
$x = 48 \quad \frac{48-40}{10} = 0.8$

CRISP LOGIC V/S FUZZY LOGIC


Crisp o/p
yes or no
True or false


Item
Crisp
yes
No
Is item Sweet


Fuzzy o/p
Maybe
Maynotbe
Absolutely
Partially


Input
Fuzzy
extremely honest 90
very honest 70
very dishonest 50
Extremely dishonest 10
Is person honest