11/03/2020

UNIT-4

CLASSIFICATION METHODS

→ classification can be done by different Algorithms

* Statistical Based Algorithm:
  - → Linear Regression
  - → Navie Bayes
* Distance based Algorithm:
  - → k-nearest neighbour (KNN)
* Decision tree based Algorithm:
  * ID3
  * C4.5
  * CART
* Neural Network based Algorithm
* Rule based Algorithm
* Combining Techniques

* classification: It is a form of data analysis that extract the models describing important data classes

Ex: A class student has A, B, C, D and E classes based on the grades.

→ By using above models, we can classify the students depends upon the Grade.

grade $\geq$ 90 $\Rightarrow$ class A

grade $\geq$ 80 & grade $<$ 90 $\Rightarrow$ class B

grade $\geq$ 70 & $<$ 80 $\Rightarrow$ class C

grade $\geq$ 60 & $<$ 70 $\Rightarrow$ class D

grade $<$ 60 $\Rightarrow$ class E

Applications:

→ Image and Pattern matching

→ Medical diagnosis

→ fraud detection

→ loan approvals

→ Classifying finanicial market, trends etc.

classification Problems:

Given a database 'D', $t_1, t_2, \ldots, t_n$ = T no of tuples (items, records) and a set of classes $C = C_1, C_2, \ldots, C_m$. The classification problem is to define the mapping $f: D \rightarrow C$ where each t_i is assigned to one class. A class C_j contains preciseley those tuples mapped to it

$$C_j = \{ t_i \mid f(t_i), \ 1 \leq i \leq n \ \{ t_i \in D \} \}$$

Scanned with CamScanner

General Approach to classification problem:

→ There are 2 ways to do the classification problem

* Learning (or) training stage
* classification stage.

Learning (or) training: where the classification model is constructed

classification: where the model is used to predict the class labels for the given data.

(1) Statistical based Algorithm:

→ Refer linear Regression topic.

(2) Base classification method / Baseian method / Navy Bayes classification method

→ It is statistical qualifier that can predict the class member probabilities such as the probability that given belongs to the particulae class

→ It is also called as Base theory

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

P(A/B) refers the probability of occurence of event A given that the event B is true

P(A) and P(B) refers to the probabilities of occurences of event A and event B

A is called preposition

B is called Evidence

P(A) is called prior probability of preposition.

P(B) is called prior probability of evidence

P(A|B) is called posterior/ posterior probability.

P(B|A) is called likelihood

$$\boxed{\text{posterior} = \frac{\text{likelihood} * \text{posterior probability of preposition}}{\text{prior probability of Evidence}}}$$

* consider the given weather dataset with attributes (outlook, temperature, windy, humidity and class label play with values {yes, No} find the person can play or Not with given weather conditions outlook = sunny, temperature, cool, windy = strong and humidity = high

SNO	outlook	temp	humidity	windy	play
1	sunny	Hot	Hot	weak	No
2	sunny	Hot	Hot	strong	No
3	overcast	Hot	Hot	weak	Yes
4	rainy	mild	Hot	weak	Yes
5	rainy	cool	nominal	weak	Yes
6	rainy	cool	nominal	strong	No
7	overcast	cool	nominal	strong	yes
8	sunny	Mild	Hot	weak	No
9	sunny	cool	nominal	weak	yes
10	rainy	mild	nominal	weak	Yes
11	sunny	mild	nominal	strong	yes
12	overcast	mild	Hot	strong	yes
13	overcast	hot	nominal	weak	yes
14	Rainy	mild	Hot	strong	No

SOL: $P(Y) = \frac{9}{14}$, $P(N) = \frac{5}{14}$

$\downarrow$ yes $\qquad$ $\downarrow$ No

outlook	Y	N	P(Y)	P(N)
sunny	2	3	2/9	3/5
overcast	4	0	4/9	0
rainy	3	2	3/9	2/5

Temperature	Y	N	P(Y)	P(N)
Hot	2	2	2/9	2/5
cold	4	2	4/9	2/5
mild	3	1	3/9	1/5

Humidity	Y	N	P(Y)	P(N)
Hot	3	4	3/9	4/5
Nominal	6	1	6/9	1/5

Windy	Y	N	P(Y)	P(N)
Weak	6	2	6/9	2/5
Strong	3	3	3/9	3/5

Probability that a player can play a game

$P(x/play = yes) \cdot P(Play = yes)$

$= \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}$.

$= 0.00529$

$= 0.0053$

Probability that player cannot play the game

$P(x/Play = No) \cdot P(Play = No)$

$= 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14$

$= \frac{18}{875}$

$= 0.0205$

$= 0.02$

Probability of class $P(x) = P(x/Play = yes) \cdot P(Play = yes)$

$+ P(x/Play = No) \cdot P(Play = No)$

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

$$P(Play = Yes/x) = \frac{P(x/Play = Yes) \cdot P(Play = Yes)}{P(x)}$$

$$= \frac{0.0053}{0.0253}$$

$$= 0.2094$$

$$P(Play = No/x) = \frac{P(x/Play = No) \cdot P(Play = No)}{P(x)}$$

$$= \frac{0.0205}{0.0253}$$

$$= 0.8102$$

since, P(NO) is having the maximum value when compared with yes, Hence the person cannot play

* find the person can go for play or not when outlook = rainy, temperature = mild, humidity = nominal, windy = weak

sol: $P(x/Play=yes)\cdot P(Play=yes) = \frac{3}{9} \times \frac{4}{9} \times \frac{6}{9} \times \frac{6}{9} \times \frac{9}{14}$

$= 0.0423$

$P(x/Play=No)\cdot P(Play=No) = \frac{2}{5} \times \frac{2}{9} \times \frac{1}{5} \times \frac{2}{5} \times \frac{5}{14}$

$= 2.539 \times 10^{-3}$

$P(x) = 0.0423 + 2.539 \times 10^{-3}$

$= 0.0448$

$P(Play = yes/x) = \frac{P(x/Play=yes)\cdot P(Play=yes)}{P(x)}$

$= \frac{0.0423}{0.0448}$

$= 0.9441$

$P(Play = No/x) = \frac{P(x/Play=No)\cdot P(Play=No)}{P(x)}$

$= \frac{0.00253}{0.0448}$

$= 0.0558$

Scanned with CamScanner

Since, P(yes) is having the maximum value when compared to NO. The person can play a game.

* find the patient having flue or No with the given conditions, cold = yes, runningnose = no, headache = mild, fever = yes.

SNO	cold	running nose	Headache	fever	flu
1	yes	NO	Mild	yes	No
2	yes	yes	No	No	yes
3	yes	NO	strong	yes	yes
4	No	yes	mild	yes	yes
5	No	NO	no	NO	No
6	No	yes	strong	yes	yes
7	No	yes	strong	No	No
8	yes	yes	mild	yes	yes

Sol: $P(y) = \frac{5}{8}$, $P(n) = \frac{3}{8}$

$\downarrow$ yes $\downarrow$ No

cold	yes	no	P(y)	P(n)
yes	3	1	3/5	1/3
NO	2	2	2/5	2/3.

Running Nose	Y	N	P(Y)	P(N)
yes	4	1	4/5	1/3
No	1	2	1/5	2/3

Headache	Y	N	P(Y)	P(N)
Mild	2	1	2/5	1/3
strong	2	1	2/5	1/3
No	1	1	1/5	1/3

fever	Y	N	P(Y)	P(N)
yes	4	1	4/5	1/3
No	1	2	1/5	2/3

Probability having flu is as follows

$$P(x/flu = yes) \cdot P(flu = yes)$$

$$= 3/5 \times 1/5 \times 2/5 \times 4/5 \times 5/8$$

$$= 0.024$$

Scanned with CamScanner

Probability of a person not having flu is as follows.

$P(x|flu=No) \cdot P(flu=No)$

$= 1/3 \times 2/3 \times 1/3 \times 1/3 \times 3/8$

$= 9.25925 \times 10^{-3}$

Probability of class $P(x) = P(x|flu=yes) \cdot P(flu=yes)$
$+ P(x|flu=No) \cdot P(flu=No)$

$P(x) = 0.024 + 9.25925 \times 10^{-3}$

$P(x) = 0.033$

$$P(flu=yes|x) = \frac{P(x|flu=yes) \cdot P(flu=yes)}{P(x)}$$

$$= \frac{0.024}{0.033}$$

$= 0.727$

$$P(flu=No|x) = \frac{P(x|flu=No) \cdot P(flu=No)}{P(x)}$$

$$= \frac{9.25925 \times 10^{-3}}{0.033}$$

$= 0.2805$

→ since the probability of having flu is high
Hence the patient is having flu

→ The person is not having flu in database but is actually having flu. so, it is identified as False-positive (FP)

Advantages of Navie Bayes:

→ It is easy & fast to predict a class of test data set.

→ It performs better predictions compared to other models

→ It performs well in the place of categorical input variables compared to numerical data varcable

DisAdvantage of Navie Bayes:

→ It is also known as bad estimator.

* Distance based Algorithm:

k-Nearest Neighbour:

→ find the value of k that must be a positive integer

→ for the given datapoint, that needs to be classified using the algorithm computes the distance b/w the given datapoint and all other data of the given dataset

→ find the k nearest neighbours based on the distance (the least distance value is the first neighbour for given datapoint)

→ The given datapoint is placed in the majority of classes with respect to neighbours

Problem: find a person x belongs to which class based on the given values weight = 57 kgs. and height = 170 cm

weight	height	class
51	167	under weight
62	182	normal
69	176	normal
64	173	normal
65	172	normal
56	174	under weight
58	169	normal
57	173	normal
55	170	normal

Sol

Weight	Height	Class	distance
51	167	under weight	$\sqrt{(57-51)^2+(170-167)^2}$ $= 6.708$
62	182	Normal	$\sqrt{(62-51)^2+(170-182)^2}$ $=13$
69	176	Normal	$\sqrt{(69-51)^2+(170-176)^2}$ $=13.41$
64	173	Normal	$\sqrt{(64-51)^2+(170-173)^2}$ $=7.615$
65	172	normal	$\sqrt{(65-51)^2+(170-172)^2}$ $=8.24$
56	174	under weight	$\sqrt{(56-51)^2+(170-174)^2}$ $=4.12$
58	169	normal	$\sqrt{(58-51)^2+(170-169)^2}$ $=1.41$ N_1
57	173	Normal	$\sqrt{(57-51)^2+(170-173)^2}$ $=3.00$ N_3
55	170	normal	$\sqrt{(55-51)^2+(170-170)^2}$ $=2.00$ N_2

Assume $K=3$

Then the nearest neighbours are d_7, d_8, d_9.

Hence x belongs to Normal class when

$K=3$

Problem: find the person X fans following

With given values age = 20 Gender = Male.

Consider male = 1, female = 0, Assume K = 3

Name	Age	Gender	class fan.
A	32	M	X1
B	40	M	X1
C	16	F	X2
D	14	F	X1
E	55	M	X2
F	40	M	X1
G	20	F	X2
H	15	M	X2
I	55	F	X1
J	15	M	X2

Sol:

Name	Age	Gender	classfan	distance	
A	32	M	X1	12	
B	40	M	X1	20	
C	16	F	X2	4.12	N2
D	14	F	X2	6.08	
E	55	M	X2	35	
F	40	M	X1	20	
G	20	F	X2	1	N1
H	15	M	X2	5	N3
I	55	F	X1	35.01	
J	15	M	X2	5	N3

∴ The person fan following class is X_2

* Remarks of k:

→ k must be a positive integer

→ k must not be a multiples of the given class parameter.

* Fix and Hods proposed k-nearest neighbour classifier algorithm in year of 1951 for performing classifier class.

→ The simple version of this algorithm is to predict the target label by finding nearest neighbour class

→ The class will be identified by using distance Eucledian distance or Manhattan distance

Applications:

→ It is used in recommendation (Ex:- online shopping)

→ information storing and retreiving in the server

How to choose k value:

→ selecting the value k in knn is the most critical problem

→ small value of k means it produce noise that will have a higher noise in research

→ The value of k is large it makes computationally Expensive

→ The best way to choose k value in knn is: $\sqrt{n}$

* Decision tree Indexion:

→ It is a flow chart like tree structure where each internal node (non-leaf node) denotes test on an attribute.

→ Each branch represents an outcome of the test. Each leaf node represents class label

→ top most node in the tree is a root node

→ Internal node label with attribute A_i, Leaf node labeled with class C_j

Applications:

→ Manufacturing and production companies

→ financial Analysis

→ Astronomy

→ Molecular Biology

* Attribute selection :

→ An Attribute selection is a huriostic for selecting splitting criteria that seperate a given data (D) of class label training tuples into individual classes also called as splitting rules

→ It provides the ranking for Each attribute and select best score attribute as root node of a tree

→ Three popular Attribute selection methods

* Information gain (ID3)
* Gain Ratio (C4.5)
* Giny Index (CART)

Information Gain (ID3)

Gain of Attribute (A) = Information Gain(p, n) − Entropy (A)

IG

$$IG(P, n) = \frac{-P}{P+n} \log_2\left(\frac{P}{P+n}\right) - \frac{n}{P+n} \log_2\left(\frac{n}{P+n}\right)$$

+ve -ve

$$Entropy\ (A) = \sum_{i=1}^{v} \frac{P_i + n_i}{P+n} I(n_i, A)$$

	outlook	temperature	humidity	windy
1	sunny	Hot	High	weak
2	sunny	Hot	High	strong
3	overcast	Hot	High	weak
4	Rainy	mild	High	weak
5	Rainy	cool	nominal	weak
6	Rainy	cool	nominal	strong
7	overcast	cool	nominal	strong
8	sunny	mild	high	weak
9	sunny	cool	nominal	weak
10	Rainy	mild	nominal	weak
11	sunny	mild	nominal	strong
12	overcast	mild	high	strong
13	overcast	hot	nominal	weak
14	Rainy	mild	high	strong

Sol: Information gain(9,5) =

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

$= 0.940$

$$\text{Entropy(outlook)} = \sum_{i=1}^{3} \frac{p_i+n_i}{p+n} I(r, A)$$

Entropy (outlook):

	P_i	n_i	$IG(P_i, n_i)$
Sunny	2	3	0.97
overcast	4	0	0
Rainy	3	2	0.97

$$IG(2,3) = \frac{-2}{2+3} \log_2\left(\frac{2}{2+3}\right) - \frac{3}{2+3} \log_2\left(\frac{3}{2+3}\right)$$

$$= 0.97$$

$IG(4,0) = 0$ (if any value is 0 then $IG = 0$)

(If ~~be~~ two values are same the $IG = 1$)

$$IG(3,2) = \frac{-3}{2+3} \log_2\left(\frac{3}{3+2}\right) - \frac{2}{3+2} \log_2\left(\frac{2}{2+3}\right)$$

$$= 0.970$$

$$\text{Entropy (outlook)} = \frac{2+3}{14}(0.970) + 0 + \frac{3+2}{14}(0.970)$$

$$= 0.692$$

Gain of Autlook = 0.940 − 0.692

Gain of outlook = 0.248

Entropy (temperature)

	P_i	n_i	$IG(P_i, n_i)$
Hot	2	2	1
mild	4	12	0.918
COOL	3	1	0.811

$$IG(4,2) = \frac{-4}{4+2} \log_2\left(\frac{4}{4+2}\right) - \frac{2}{4+2} \log_2\left(\frac{2}{4+2}\right)$$

$$= 0.918$$

$$IG(3,1) = \frac{-3}{3+1} \log_2\left(\frac{3}{3+1}\right) - \frac{1}{3+1} \log_2\left(\frac{1}{3+1}\right)$$

$$= 0.811$$

Entropy (temperature) =

$$\frac{2+2}{14}(1) + \frac{4+2}{14}(0.918) + \frac{3+1}{14}(0.811)$$

$$= 0.910$$

Gain of temperature = 0.940 − 0.910

= 0.03

Gain of temperature = 0.03

Entropy (humidity)

	p_i	n_i	IG
High	3	4	0.985
nominal	6	1	0.591

$$IG(3,4) = \frac{-3}{3+4} \log_2\left(\frac{3}{3+4}\right) - \frac{4}{3+4} \log_2\left(\frac{4}{3+4}\right)$$

$$= 0.985$$

$$IG(6,1) = \frac{-6}{6+1} \log_2\left(\frac{6}{6+1}\right) - \frac{1}{6+1} \log_2\left(\frac{6}{6+1}\right)$$

$$=$$

$$\text{Entropy (humidity)} = \frac{3+4}{14}(0.985) + \frac{6+1}{14}(0.591)$$

$$= 0.788$$

$$\text{Gain of humidity} = 0.940 - 0.788 = 0.152$$

Gain of humidity $= 0.152$

Entropy (Windy)

	p_i	n_i	$IG(p_i, n_i)$
Weak	6	2	0.811
Strong	3	3	1

$$IG(6,2) = \frac{-6}{6+2}\log_2\left(\frac{6}{6+2}\right) - \frac{2}{6+2}\log_2\left(\frac{2}{6+2}\right)$$

$$= 0.811$$

$$IG(3,3) = 1$$

$$\text{Entropy (windy)} = \frac{6+2}{14}(0.811) + \frac{3+3}{14}(1)$$

$$= \cancel{1.106}\ 0.8952$$

$$\text{Gain of windy} = 0.940 - \cancel{1.106}\ 0.8952$$

$$= \cancel{-0.166}\ 0.048$$

→ Among all the 4 attributes the Gain value of outlook is high Hence we consider this as rootnode

	outlook	temperature	humidity	windy	Play
1	sunny	Hot	High	weak	No
2	sunny	Hot	High	strong	No
3	sunny	mild	High	weak	No
4	sunny	cool	nominal	weak	yes
5	sunny	mild	nominal	strong	yes
6	overcast	Hot	High	weak	yes
7	overcast	cool	nominal	strong	yes
8	overcast	mild	high	strong	yes
9	overcast	Hot	nominal	weak	yes
10	Rainy	mild	High	weak	yes
11	Rainy	cool	nominal	weak	yes
12	Rainy	cool	nominal	strong	No
13	Rainy	mild	nominal	weak	yes
14	Rainy	mild	High	strong	No

consider sunny dataset

yes = 2 No = 3

$$IG(2,3) = \frac{-2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$$

$$= 0.970$$

Temperature

	P_i	n_i	$IG(P_i, n_i)$
Hot	0	2	0
mild	1	1	1
cool	1	0	0

$$\text{Entropy (temperature)} = \frac{0+2}{5}(0) + \frac{1+1}{5}(1) + \frac{1+0}{5}(0)$$

$$= 0.4$$

$$\text{Gain (temperature)} = 0.970 - 0.4$$

$$\text{Gain (temperature)} = 0.57$$

Humidity

	P_i	n_i	$IG(P_i, n_i)$
High	0	3	0
nominal	0	2	0

$$\text{Entropy (humidity)} = 0$$

$$\text{Gain (humidity)} = 0.97 - 0$$

$$= 0.97$$

Windy

	P_i	n_i	$IG(P_i, n_i)$
weak	1	2	0.918
Strong	1	1	1

~~Entropy(windy)→~~

$$IG(1,2) = \frac{-1}{1+2}\log_2\left(\frac{1}{1+2}\right) - \frac{2}{1+2}\log_2\left(\frac{2}{1+2}\right)$$

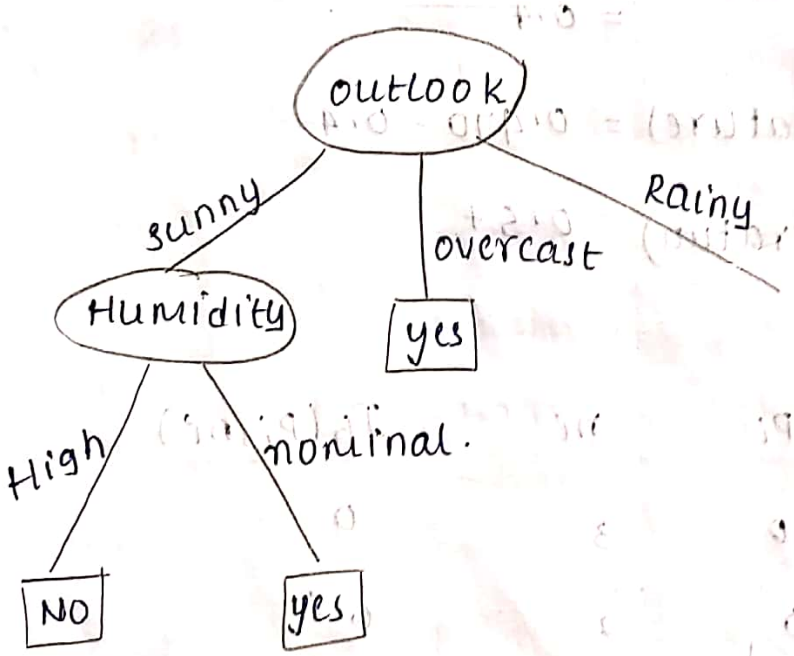
$$= 0.918$$

$$\text{Entropy (windy)} = \frac{3}{5}(0.918) + \frac{2}{5}(1)$$

$$= 0.950.$$

Gain (windy) = 0.97 − 0.95

Gain (windy) = 0.03



consider Rainy dataset

Temperature

	Pi	ni	IG(Pi,ni)
cool	1	1	~~0.918~~ 1
mild	2	1	0.918

Entropy (Temp) = $\frac{2}{5}(1) + \frac{3}{5}(0.918)$

Entropy (Temp) = 0.950.

Gain (Temp) = 0.970 − 0.950

G(Temp) = 0.03

Humidity

	Pi	ni	IG(Pi, ni)
High	1	1	1
nominal	2	1	0.918.

Entropy (humidity) = 0.950

G(Humidty) = 0.03

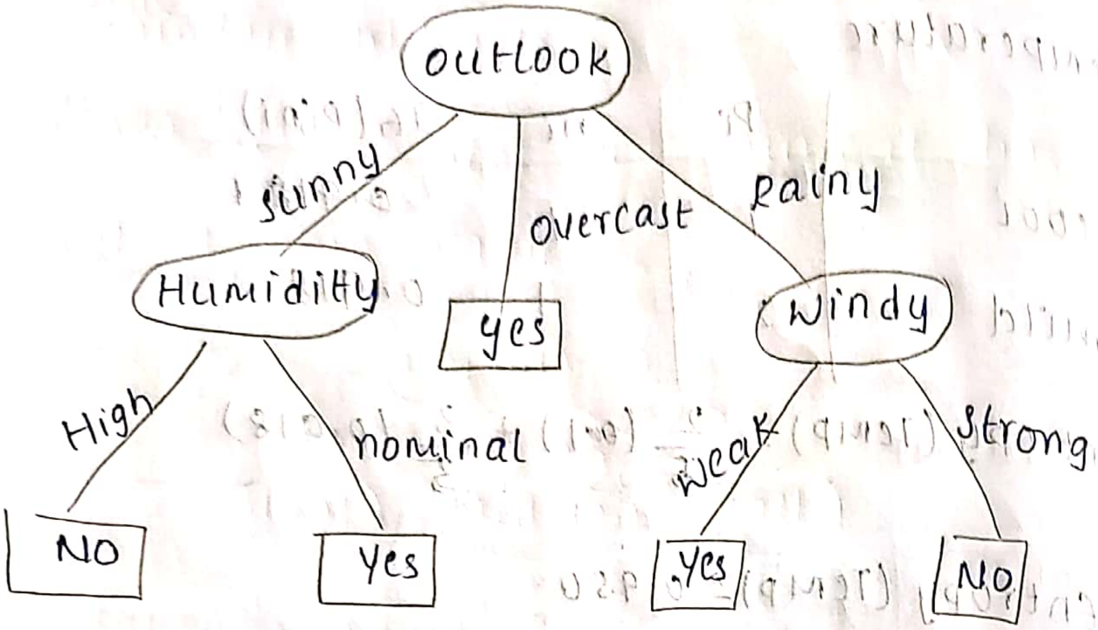
Windy

	Pi	ni	IG(Pi, ni)
weak	3	0	0
strong	0	2	0

Entropy (windy) = 0

G(windy) = 0.97

final decision tree is as follows



→

Gender	Height	class
F	1.6m	short
M	2M	Tall
F	1.9m	Medium
F	1.88M	medium
F	1.7M	short
M	1.85m	medium
F	1.6M	short
M	1.7M	short
M	2.2M	Tall
M	2.1m	Tall
F	1.8m	Medium
M	1.8M	Medium
F	1.95m	medium
F	1.9m	medium
F	1.8m	medium

Sol: Short = 4, Tall = 3, Medium = 4

$$IG(S,T,M) = \frac{-4}{8+4+3} \log_2\left(\frac{4}{15}\right) - \frac{8}{15} \log_2\left(\frac{8}{15}\right) - \frac{4}{15} \log_2\left(\frac{4}{15}\right)$$

$= 1.456$

Gender

	Si	Ti	Mi	IG (Si, Ti, Mi)
M	1	3	2	1.459
F	3	0	6	0

$$IG(1,3,2) = \frac{-1}{6} \log_2\left(\frac{1}{6}\right) - \frac{3}{6} \log_2\left(\frac{3}{6}\right) - \frac{2}{6} \log_2\left(\frac{2}{6}\right)$$

$= 1.459$

$$\text{Entropy (gender)} = \frac{1+3+2}{15}(1.459)$$

$= 0.583$

Gain (gender) = 0.873

Height : consider it in ranges of <1.7, $>1.788 < 2.0$ and >2.0

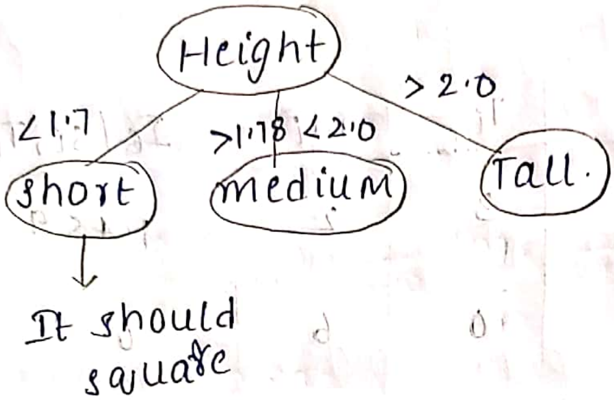
	Si	Ti	Mi	IG (Si, Ti, Mi)
≤ 1.7	4	0	0	0
$<1.788 < 2.0$	0	0	8	0
>2.0	0	3	0	0

Scanned with CamScanner

Entropy (Height) = 0

Gain (Height) = 1.456

Height is the root node since it has highest gain value



→ construct decision tree for below dataset :

Rid	age	income	Student	Credit Rating	class
1	young	High	NO	fair	No
2	young	High	NO	Excleent	NO
3	middleage	High	NO	fair.	yes
4	senior	medium	yes	fair	yes
5	senior	Low	yes	fair	yes
6	senior	Low.	yes	Excellent	No
7	middle age	Low	NO	Excellent	yes
8	young	medium	yes	fair	No
9	young	Low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	young	medium	No	Excellent	yes
12	Middle age	medium	yes	Excellent	yes
13	middle age	high	yes	fair	yes
14	senior	medium	No	Excellent	No.

Scanned with CamScanner

Sol: consider overall data set

yes = 9, No = 5

$$IG(9,5) = \frac{-9}{9+5}\log_2\left(\frac{9}{9+5}\right) - \frac{5}{9+5}\log_2\left(\frac{5}{9+5}\right)$$

$$= 0.940$$

Entropy (Age):

Age	Pi	ni	IG
young	2	3	0.970
middleage	4	0	0
senior	3	2	0.970

$$IG(2,3) = \frac{-2}{2+3}\log_2\left(\frac{2}{2+3}\right) - \frac{3}{2+3}\log_2\left(\frac{3}{3+2}\right) = 0.970$$

$$\text{Entropy (age)} = \frac{2+3}{4}(0.970) + 0 + \frac{3+2}{14}(0.970) = 0.692$$

$$\text{Gain (Age)} = 0.940 - 0.692 = 0.248$$

Entropy (Income):

Income	Pi	ni	IG
High	2	2	1
medium	4	2	0.918
Low	3	1	0.811

$$IG(4,2) = \frac{-4}{4+2}\log_2\left(\frac{4}{4+2}\right) - \frac{2}{2+4}\log_2\left(\frac{2}{2+4}\right) = 0.918.$$

$$IG(3,1) = -\frac{3}{3+1}\log_2\left(\frac{3}{3+1}\right) - \frac{1}{3+1}\log_2\left(\frac{3}{3+1}\right) = 0.811$$

Entropy (income) = 0.910

Gain (income) = 0.03

Entropy (student):

Student	Pi	ni	IG
No	3	3	1
yes	6	2	0.811

$$IG(6,2) = \frac{-6}{6+2}\log_2\left(\frac{6}{6+2}\right) - \frac{2}{6+2}\log_2\left(\frac{2}{6+2}\right) = 0.811$$

Entropy (student) = 0.892

Gain (student) = 0.940 − 0.892 = 0.048.

Entropy (credit Rating):

credit Rating	Pi	ni	IG
fair	6	2	0.811
Excellent	3	3	1

$$IG(6,2) = \frac{-6}{6+2}\log_2\left(\frac{6}{6+2}\right) - \frac{2}{6+2}\log_2\left(\frac{2}{6+2}\right)$$

$$= 0.811$$

Entropy (credit rating) = 0.892

Gain (Credit rating) = 0.048.

→ Among all gains, Age has highest gain value. So, Age will be root node.

Scanned with CamScanner

Age

dataset-1:

Age	Income	Student	Creditrating	class
young	High	No	fair	No
young	High	No	excellent	No
young	medium	yes	fair	No
young	low	yes	fair	yes
young	medium	No	excellent	yes.

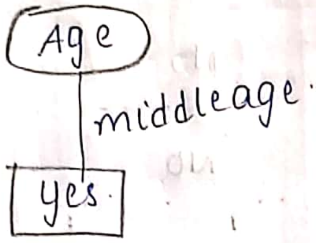
dataset-2:

Age	Income	Student	Credit-rating	class.
middleage	High	No	fair	yes
middleage	low	No	Excellent	yes
middleage	medium	yes	Excellent	yes
middleage	High	yes	fair	yes.

data set-3:

Age	Income	Student	Creditrating	class
senior	medium	yes	fair	yes
senior	low	yes	fair	~~No~~ yes
senior	low	yes	Excellent	~~yes~~ No
senior	medium	yes	fair	yes
senior	medium	No.	Excellent	No.

Since middle age have all 'yes' classes then the tree is as follows



Consider data-set -①:

yes = 2, NO = 3

$$IG(young) = \frac{-2}{2+3}\log_2\left(\frac{2}{2+3}\right) - \frac{3}{3+2}\log_2\left(\frac{3}{3+2}\right) = 0.970$$

Income:

Income	Pi	ni	IG
High	0	2	0
medium	1	1	1
LOW	1	0	0

Entropy (income) = 0.40

Gain(income) = 0.57

Student:

Student	Pi	ni	IG
No	1	2	0.918
yes	1	1	1

$IG(1,2) = 0.918$.

Entropy (student) = 0.946

Gain (student) = 0.03.

Credit rating:

Creditrating	Pi	ni	IG
fair	1	2	0.918
Excellent	1	1	1

IG(1,2) = 0.918

Entropy (credit Rating) = 0.946.

Gain (credit Rating) = 0.03

consider dataset-3:

yes = 3 , NO = 2

IG(3,2) = 0.970

Income:

Income	Pi	ni	IG
medium	2	1	0.918
low	1	1	1

IG(2,1) = 0.918

Entropy (income) = 0.950

Gain (income) = 0.970 − 0.950

= 0.03

Student:

Student	P_i	n_i	I_G
yes	3	1	0.811
No	0	1	0

$I_G(3,1) = 0.811$

Entropy (Student) = 0.648

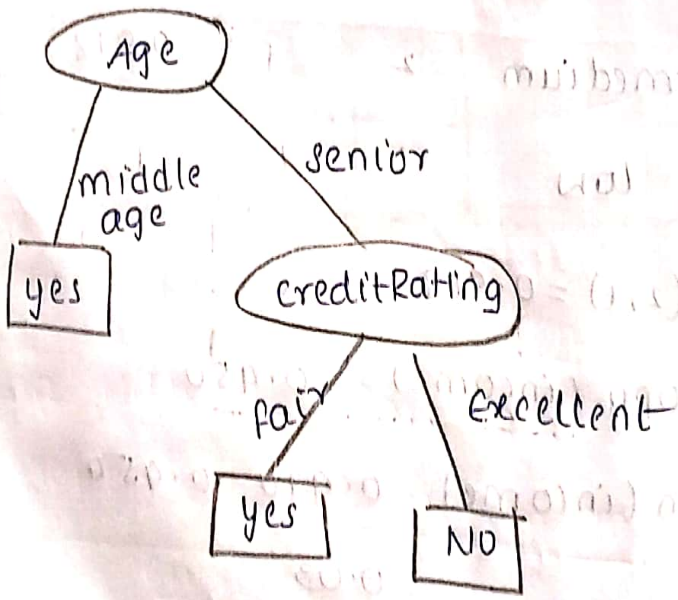
Gain (student) = 0.322

Credit Rating:

Credit Rating	P_i	n_i	I_G
fair	3	0	0
Excellent	0	2	0

Entropy (credit Rating) = 0

Gain (credit Rating) = 0.970



from data-set-①: In the Income, we have one yes class and one No class in the middle. so, we have to consider subdataset from data-set ①.

Income	Student	Credit Rating	class
medium	yes	fair	No
medium	No	Excellent	yes

consider Age = young
Income = middle.

Medium:

yes = 1, No = 1

$IG(1,1) = 1$

Student:

Student	p_i	n_i	IG
yes	0	1	0
No	0	0	0.

Entropy (Student) = 0

Gain (student) = 1.

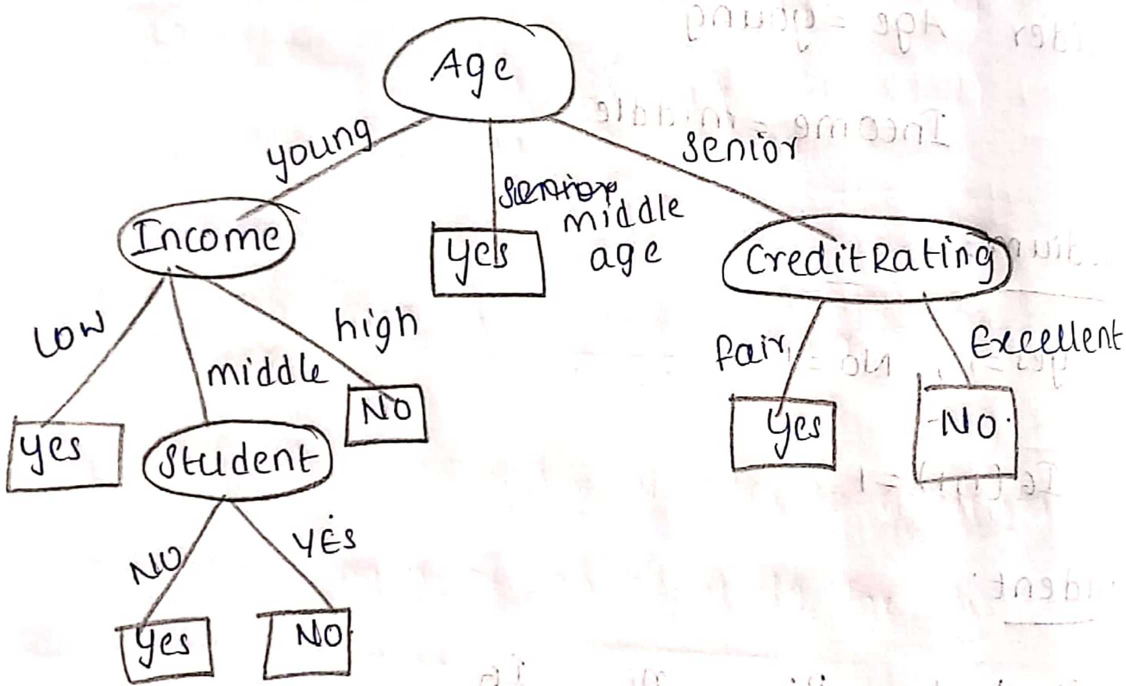
Credit Rating :

Credit Rating	Pi	ni	IG
fair	0	1	0
Excellent	0	0	0

Entropy(Credit Rating) = 0

Gain = 0.

decision tree :



→ In ID_3 we can construct a decision tree by considering with its class label with consideration of information gain and entrophy

→ drawback of information gain is it prefers to select attributes with having large number of values.

→ C4.5 uses an extension of information gain known as gain ratio which attempts to over come the drawback of Id3.

→ C4.5 improves id3 with the following ways:

* missing data.
* continous data
* Tree pruning (post and pre pruning)

present overfiting to noise in the data	take a fully grown decision tree and discard the unreliable parts.	stop growing a branch when information gain becomes unreliable.

→ Rules

→ splitting

* C4.5 applies a kind of normalization to information gain using split information gain. value defined with information gain (D) where. D indicates data set.

$$\boxed{\text{split info}(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \log_2 \left(\frac{|D_j|}{|D|} \right)}$$

$$\boxed{\text{gain ratio}(A) = \frac{\text{gain}(A)}{\text{splitinfo}(D)}}$$

Where A indicates particular attribute.

Ex: caluculate split information for the attribute income and caluculate gain ratio.

$$\text{Split info}_{\text{income}}(D) = \frac{-4}{14} \log_2\left(\frac{4}{14}\right) - \frac{6}{14} \log_2\left(\frac{6}{14}\right)$$

$$= \frac{-4}{14} \log_2\left(\frac{4}{14}\right)$$

$$= 1.557$$

$$\text{gainratio (income)} = \frac{0.029}{1.557} = 0.019$$

CART: Here the gini index inorder to caluculate information gain of an attribute for the decision tree.

Gini index: It measures the impurity of the data partion (or) set of training tuples.

$$\boxed{\text{Gini}(D) = 1 - \sum_{i=1}^{m} P_i^2}$$

Scanned with CamScanner

where p can be yes (or) No gini index for a particular attribute

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

where D_1 and D_2 are the partion data values (or) partion of training data.

$$\boxed{\Delta Gini(A) = Gini(D) - Gini_A(D)}$$

Problem:

calucculate Gini index for the attribute income for the given data set which consits of the class labels yes = 9 and No = 5

$$Gini(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459$$

Gini index for the partion D_1 which consits of low and medium.

$$Gini_{income\ \{low,\ medium\}}(D) = \frac{10}{14} \underset{\{low,\ medium\}}{Gini(D_1)} + \frac{4}{14} \underset{High}{Gini(D_2)}$$

$$= \frac{10}{14}\left[1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2\right] + \frac{4}{14}\left[1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right]$$

$$= 0.443$$

$$Gini(income) = 0.459 - 0.443$$

$$= 0.026$$

Scanned with CamScanner

(1) Rule based classification : (IF - then)

IF part - rule antecedent

Then part - rule consequent

Using if then rules for classification : A rule based classifier uses a set of if - then rules for classification. An if - then rule is an expensive of the form

$\boxed{\text{if condition then conclusion}}$

Ex:

Rule R_1 : IF (age = YOUTH AND student = yes)

then buys-computer = yes

Rule R_2 : IF (age = middleage AND student = NO)

then buys-computer = No.

→ IF part (or) Left side of a rule is known as rule antecedent (or) precondition.

→ then part (or), right side of a rule is known as rule consequent

→ In the rule antecedent, the condition consists of one or more attribute tests (Ex: age = young or student = yes) that are logically ANDed

→ It can also be written as

If (age = youth) ∧ (student = yes) ⟹ buys_computer = yes

→ A Rule R can be assessed by its coverage and accuracy. Given a tuple X from a class labelled data set D, $n_{correct}$ be the No of tuples correctly classified. We can define the coverage and accuracy of R(rule)

$$\boxed{\text{coverage}(R) = \frac{n_{covers}}{|D|}}$$

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

Ex: lets consider student dataset are class labelled tuples from the all electronics customer database our task is to be predict whether a customer will buy a computer

→ consider rule R_1 which covers 2 of the 14 tuples. It can correctly classify both the tuples.

Therefore,

$$\text{coverage}(R_1) = \frac{2}{14} \times 100$$
$$= 0.1428 \times 100$$
$$= 14.28\%$$

$$\text{Accuracy}(R_1) = \frac{2}{2} = 100\%$$

Scanned with CamScanner

Neural Network:

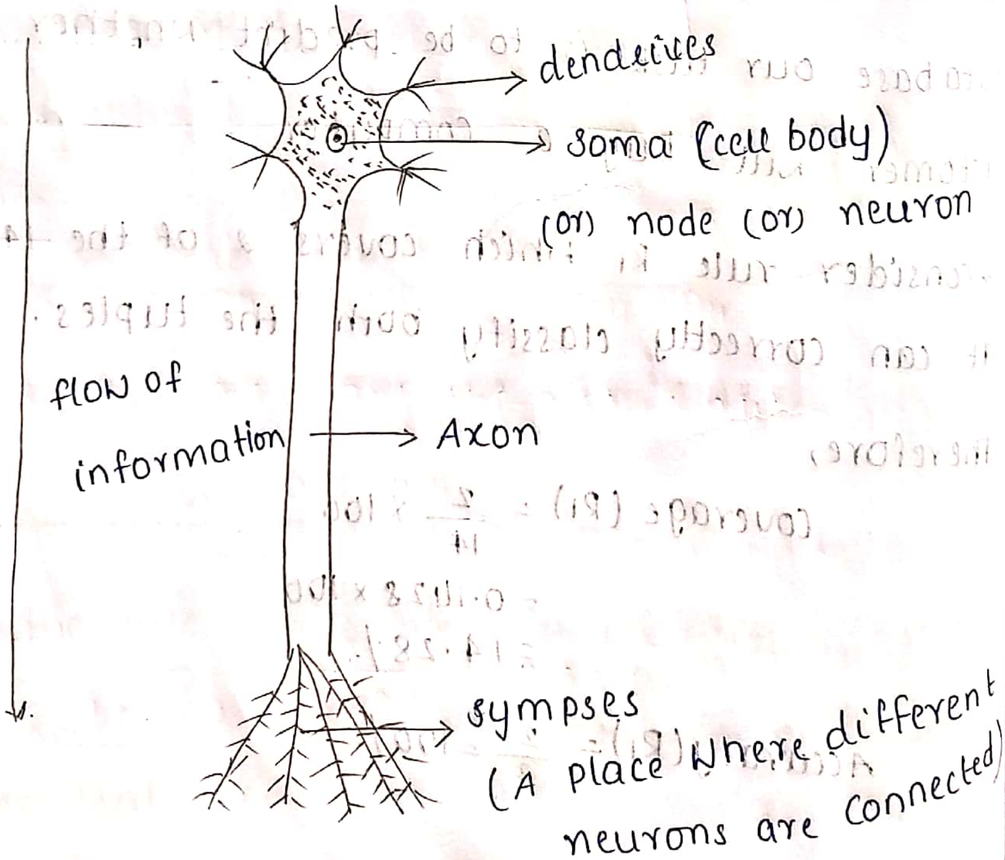
→ ANN (Artificial Neural Network): The main objective of ANN is to develop a computer (or) to design a computer that work like a brain.

→ It is also called as parallel distributing processing system

→ ANN works like a large collection of units that are inter connected in some patterns to allow the communication between the units

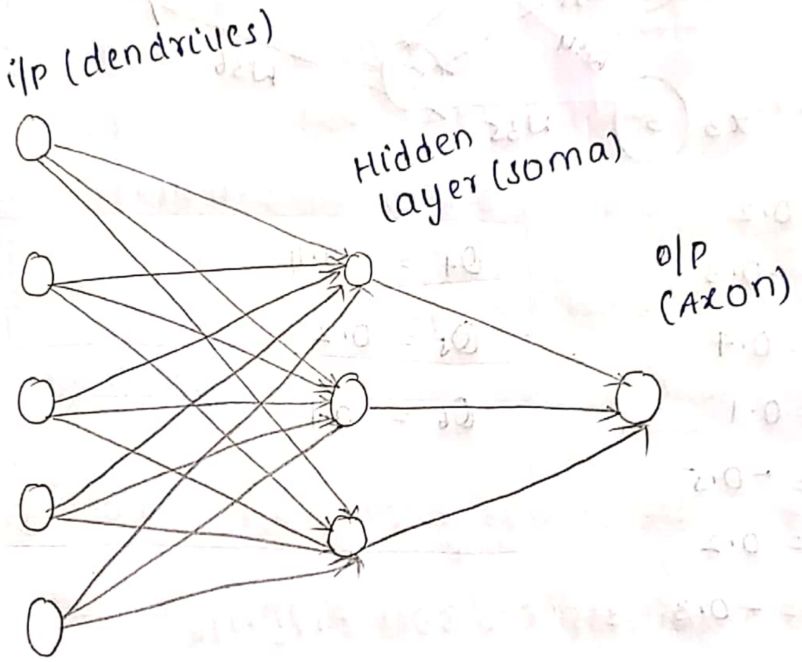
→ The units are called as Nodes (or) Neurons

Biological Neural Network (BNN)



BNN		ANN
dendrives	—	input's
Sympses	—	Weights (or) inter connection
soma	—	node
Axon	—	o/p

* ANN:

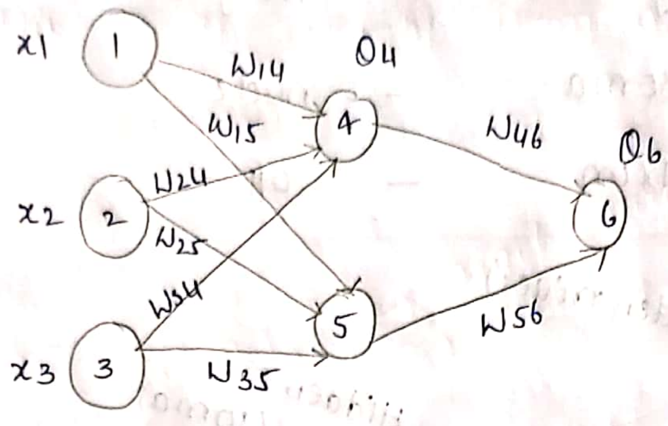


* Differences between BNN and ANN:

function	BNN	ANN
processing	It is a parallel distributing system	It is also parallel but slower than BNN.
size	In BNN, there are 10^{11} neurons and 10^{15} inter connections	10^2 to 10^4 neurons. (depending on application we can change the structure of ANN)
fault tolerance	very High & degrades the performance of object	It's not affected the performance as compared to BNN.

Problem:

$x_1, x_2, x_3 = 1, 0, 1$ target $= 1$, learning rate $= 0.9$



$W_{14} = 0.2$

$W_{15} = -0.3$

$W_{24} = 0.4$

$W_{25} = 0.1$

$W_{34} = -0.5$

$W_{35} = 0.2$

$W_{46} = -0.3$

$W_{56} = -0.2$

$O_4 = -0.4$

$O_5 = 0.2$

$O_6 = 0.1$

$$\boxed{\text{Activation function of } x = \frac{1}{1+e^{-x}}}$$

At Node-4:

node $4 = x_1 W_{14} + x_2 W_{24} + x_3 W_{34} + O_4$

$= 1 \times 0.2 + 0 \times 1 + 1 \times (-0.5) + (-0.4)$

$= 0.2 - 0.5 - 0.4$

$= -0.7$

$O_4 = \frac{1}{1+e^{-x}} = \frac{1}{1+e^{0.7}} = 0.33\%$

At node-5:

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

$$= 1 \times (-0.3) + 0 + 1 \times (0.2) + 0.2$$

$$= -0.3 + 0.2 + 0.2$$

$$= 0.1$$

$$O_5 = \frac{1}{1+e^{-x}} = \frac{1}{1+e^{-0.1}} = 0.524$$

At node-6:

$$node\text{-}6 = O_4 w_{46} + O_5 w_{56} + \theta_6.$$

$$(0.331)(-0.3) + (0.524)(-0.2)$$

$$= -0.104.$$

$$O_6 = \frac{1}{1+e^{+0.104}} = 0.474.$$

→ The output of node 6 and target value are not matching. Hence, adjust the weights and bias values to adjust the weights, the formula is as follows:

$$\boxed{w_{ij}\,new = w_{ij}\,old + l\; Err_j\, O_i}$$

to calculate the error at the o/p node, the formula is as follows:

$$\boxed{Err_j = O_j(1-O_j) * (T-O_j)}$$

If error at Hidden nodes

$$\boxed{Err_j = O_j(1-O_j) \sum_{K} Err_K W_{jK}}$$

Where W_{jk} is the weight from node j to node k

Error at node 6:

target value

$$Err_6 = O_6(1-O_6) * (T-O_6)$$

$$= 0.474(1-0.474)(1-0.474)$$

$$= 0.131$$

Error at node 4:

$$Err_4 = O_4(1-O_4) \sum_{R} Err_R W_{jR}$$

$$= (0.331)(1-0.331) Err_6 W_{46}$$

$$= (0.331)(1-0.331)(0.131)(-0.3)$$

$$= -0.008$$

Error at node 5:

$$Err_5 = O_5(1-O_5) Err_6 W_{56}$$

$$= (0.524)(1-0.524)(0.131)(-0.2)$$

$$= -0.006$$

To adjust the bias, the formula is as follows:

$$\boxed{W_{ij}new = W_{ij}old + l \; Err_j O_i}$$

$W_{46} = W_{46} + (0.9) \times Err_6 O_4$

$= W_{46} + (0.9) \times (0.131)(0.331)$

$= -0.260$

$W_{56} = W_{56} + (0.9) \times Err_6 O_5$

$= (-0.2) + (0.9) \times (-0.008) \times 1$

$= 0.1938$

$W_{24} = W_{14} + (0.9) \times Err_4 \times O_1$

$= 0.2 + (0.9) \times (-0.008) \times 1$

$= 0.1928$

$W_{24} = W_{24} + (0.9) \times Err_4 \times O_2$

$= 0.4$

$W_{34} = W_{34} + (0.9) \times Err_4 \times O_3$

$= -0.507$

$W_{15} = -0.305$

$W_{25} = 0.1$

$W_{35} = 0.191$

$$\boxed{O_j new = O_j odd + l \; Err_j}$$

$O_6 = 0.218$

$O_5 = 0.194$

$O_4 = -4.08$

Similarly, continue the process until O_6 is approximately to target value.

Scanned with CamScanner