MODULE - IV

Classification & Prediction

Classification → Supervised
Clustering → Unsupervised

1) Decision tree Classification
↳ Attribute selectⁿ Measure → ID3 Method
↳ IBM Method

ID3

eg

A_1 A_2 A_3 A_4

Entropy(D)

Entropy (A_1) ⇒ Gain (A_1) → Ent(D) − Ent(A_1)
Entropy (A_2) ⇒ Gain (A_2)
Entropy (A_3) ⇒ Gain (A_3)
Entropy (A_4) ⇒ Gain (A_4)

- Highest gain attribute = ROOT
- Entropy (D) $= -\sum_{i=1}^{m} P_i \log_2 P_i$

Pg-26, 9 Yes, 5 - No, Total = 14
↳ 9/14 ↳ 5/14 ⇒ Probability (P_i)

Entropy (D) $= -\sum_{i=1}^{2} P_i \log_2 P_i$

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

$= 0.94$

Then, for Age → (3 No, 2 Yes) Y M S
↳ 3/5 ↳ 2/5 5/14 4/14 5/14

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

$= 0.347 + 0 + 0.347$

$= 0.694$

So, $\text{Gain}_{(D, Age)} = 0.94 - 0.694 = 0.246$ ✓

Then, for income,

(4/14) H   M (6/14)   L → (4/14)

N Y N Y N Y

2/4 2/4 2/6 4/6 1/4 3/4

$$Entropy(Income) = \frac{4}{14}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) + \frac{6}{14}\left(\frac{-2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6}\right) + \frac{4}{14}\left(\frac{-1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}\right)$$

$$= 0.2857 + 0.3935 + 0.2318$$

$$= 0.911$$

So, gain $= 0.94 - 0.911 = 0.029$

for student,

N   Y

7/14   7/14

Yes NO Yes NO

3/7 4/7 6/7 1/7

$$Entropy(Student) = \frac{7}{14}\left(\frac{-3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}\right) + \frac{7}{14}\left(-\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}\right)$$

$$= 0.4926 + 0.2958 = 0.7884$$

So, gain $= 0.94 - 0.7884 = 0.1516$

for Credit_Rating,

(8/14) F   E (6/14)

Y N Y N

6/8 2/8 3/6 3/6

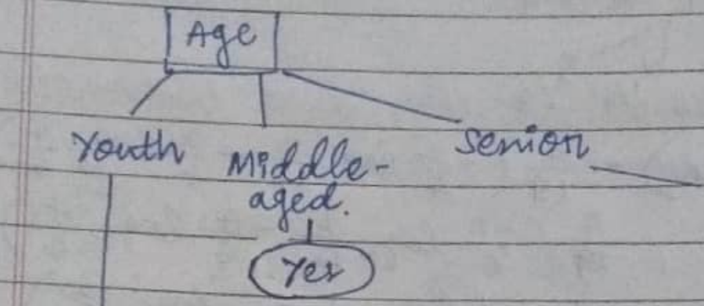
$$Entropy(Credit\text{-}Rating) = \frac{8}{14}\left(\frac{-6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}\right) + \frac{6}{14}\left(\frac{-3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}\right)$$

$$= 0.4636 + 0.4286 = 0.8922$$

So, gain $= 0.94 - 0.8922 = 0.0478$

∴ Age has highest gain ⇒ ROOT : Age

Decision Tree (ID3 Technique)



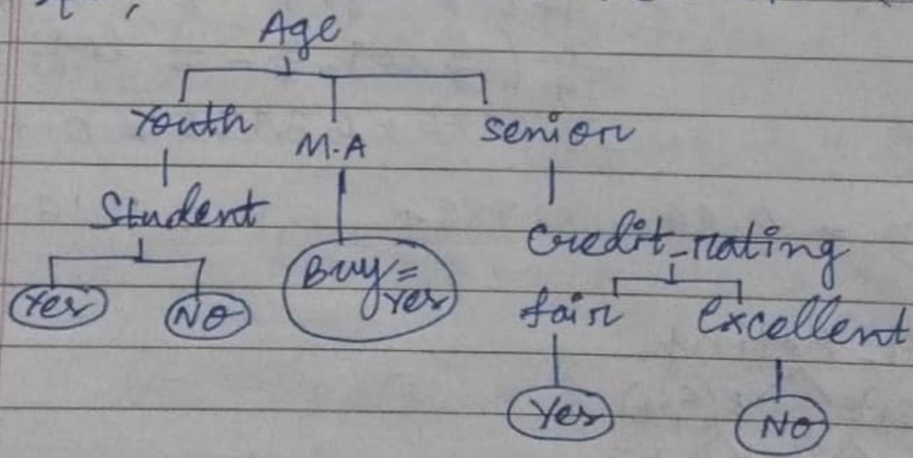
Income	Stud	CR	Class
High	No	Fair	No
High	No	Excel	No
Med^m	No	Fair	No
Low	Yes	Fair	Yes
Med^m	Yes	Excel	Yes

(Here, class has both No & Yes so again division)

Gain of student is ↑ among Income, Stud & C (compute again with new table)

So new root = Student.

So, final decision tree ⇒



- Yes, Yes, No } Go with Yes

  Yes, Yes, No, No } Can't take a decision.

Majority Voting

Naive Bayesian Classification

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)\,P(B|A)}{P(B)}$$ } (dependent) Posterior (conditional) Probability

- Total = 14
  Yes = 9 ⇒ Prob = 9/14
  No = 5 ⇒ Prob. = 5/14

Q X = Age = Youth
Income = Medium
Stud = Yes
C-R = Fair . Class = 2
Ans Yes

Class : Yes	Class : No
P Age = Youth ⇒ 2/9	P Age = Youth ⇒ 3/5
P Income = Med^m ⇒ 4/9	P Income = Med^m ⇒ 2/5
P Stud = Yes ⇒ 6/9	Stud = Yes ⇒ 1/5
P C-R = fair ⇒ 6/9	C-R = fair ⇒ 2/5
288/6561 = 0.044	12/625 = 0.0191
$\frac{9}{14} \times 0.044$ = 0.028 ✓ more	$\frac{5}{14} \times 0.0191$ = 0.0067

★ When one sample has Prob. of 0 then Naive Bayes Algorithm will not work.
Sol^n :- Laplace correction
(Adding 1 to each class).

Q- C = Y, S = Y, S = N → Pass

Total = 5

Pass = 3

Fail = 2

Class: Pass	Class: Fail
Confident = Yes	
⇒ 2/3	⇒ 1/2
Studied = Yes	
⇒ 2/3	⇒ 1/2
Sick = No	
⇒ 1/3	⇒ 1/2
$\therefore \frac{2\times2\times1}{3\times3\times3} = 0.148$	$\therefore \frac{1\times1\times1}{2\times2\times2} = 1/8 = 0.125$
$\frac{3}{5}\times 0.148 = 0.0888$	$\frac{2}{5}\times 0.125 = 0.05$

(lazy learner)

↳ **KNN** (K nearest neighbour) → only for numeric values

Distance Calculation →

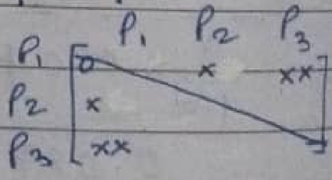
① **Ecludian Distance** :-

$$dist(P_1, P_2) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

	A_1	A_2
P_1	2	5
P_2	3	7

$$= \sqrt{(2-3)^2 + (5-7)^2} = \sqrt{5} = 2.24$$

P_3 : 5 & 8



② **Manhattan Distance** :-

$$dist(P_1, P_2) = \sum_{i=1}^{n} |x_i - y_i|$$

$$= |2-3| + |5-7| = |-1| + |-2| = 3$$

	attr 1	attr 2
P_1	2	5
P_2	3	7
P_3	5	8
P_4	6	3

$$\begin{array}{c|cccc} & P_1 & P_2 & P_3 & P_4 \\ \hline P_1 & 0 & 2.24 & & \\ P_2 & 2.24 & 0 & & \\ P_3 & 4.24 & 2.24 & 0 & \\ P_4 & 4.47 & 5 & 5.1 & 0 \end{array}$$

Euclidean

$P_1 \& P_2 \rightarrow \sqrt{5} = 2.24$

$P_1 \& P_3 = \sqrt{3^2 + 3^2} = \sqrt{18} = 4.24$

$P_1 \& P_4 = \sqrt{4^2 + 2^2} = 4.47$

$P_2 \& P_3 = \sqrt{2^2 + 1^2} = 2.24$

$P_2 \& P_4 = \sqrt{3^2 + 4^2} = 5$

$P_3 \& P_4 = \sqrt{1^2 + 5^2} = 5.1$

Manhattan

$P_1 \& P_2 = 1 + 2 = 3$

$P_1 \& P_3 = 3 + 3 = 6$

$P_1 \& P_4 = 4 + 2 = 6$

$P_2 \& P_3 = 2 + 1 = 3$

$$\begin{array}{c|cccc} & P_1 & P_2 & P_3 & P_4 \\ \hline P_1 & 0 & & & \\ P_2 & 3 & 0 & & \\ P_3 & 6 & 3 & 0 & \\ P_4 & 6 & 7 & 6 & 0 \end{array}$$

③ Minkowski Distance

(Generalizat^n of Euclidean & Manhattan)

$$\text{dist}(P_1, P_2) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

If $p = 1 \Rightarrow$ Manhattan

$p = 2 \Rightarrow$ Euclidean

Q-

	Temp	Wind	Class
R_1	20	30	$\boxed{\text{Rain}}$
R_2	40	20	No
R_3	45	35	No
R_4	30	40	$\boxed{\text{Rain}}$
R_5	35	45	$\boxed{\text{Rain}}$
Test	28	51	?

Ans- (K)NN ⟶ Lazy Learner
Odd number (K ≠ 1 as wrt 1 sample comput
Not accurate)

distance (Only on Euclidean dist)

r_1	22.47	(dist bet^n r_1 & test) (Rank 3)
r_2	33.24	(Rank 5)
r_3	23.34	(Rank 4)
r_4	11.18	(Rank 2)
r_5	9.21	(Rank 1)

Let K = 3, so consider 3 sample i.e.
R_1, R_2 & R_3 i.e. r_5, r_4, r_1
All have class = Rain.
So, Test sample class = Rain.

If Rain, Rain, No } Rain (Majority Voting)

Q- 2024 Spring Endsem. Find Euclidean dist. bet^n (24, 12, 40, 16) & (22, 9, 37, 8) 2

$$\sqrt{(24-22)^2 + (12-9)^2 + (40-37)^2 + (16-8)^2}$$

$$\sqrt{2^2 + 3^2 + 3^2 + 8^2} = 9.27$$

Q-

	Customer	Age	Loan	Default
1	Suman	25	4000	$\boxed{\text{No}}$
2	Arya	30	4000	$\boxed{\text{No}}$
3	Sarthak	35	8000	No
4	Rohit	23	2000	$\boxed{\text{No}}$
5	Hardik	26	2500	$\boxed{\text{Yes}}$
6	Binit	31	1800	$\boxed{\text{Yes}}$
7	Suryansh	22	9000	Yes
8	Shagun	40	4500	No
9	Suneha	42	5600	No
10	Yash	45	7000	Yes

Test
Sumit
38
1400
2

K = 5

Ans

	distance			
1	2600.03	(R_5)	No	
2	2600.01	(R_4)	No	
3	6600.00		No	No (Ans.).
4	600.18	(R_2)	Yes	
5	1100.06	(R_3)	Yes	
6	400.06	(R_1)		
7	7600.01			
8	3100.00			
9	4200.00			
10	5600.00			

Rule Based Classification

If conditⁿ then conclusⁿ

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

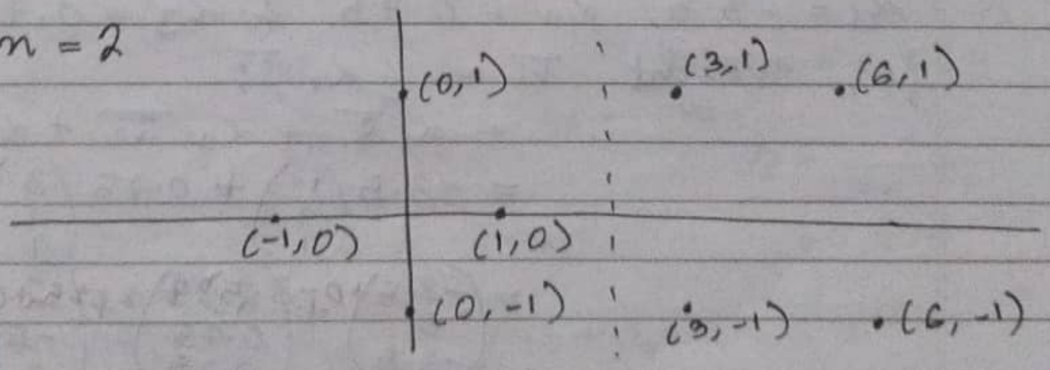
$$\text{Accuracy} = \frac{n_{correct}}{n_{covers}} = \frac{2}{2} = 100\%$$

R_2: If age = 'senior' AND income = 'medium' then buy comp' = 3/14

Accuracy (R_2) = $2/3$ = 66.67 %.

SVM (Support vector machine)

Q- margin = 2



$S_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

$S_2 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$

$S_3 = (3$

Bias value added = 50% of margin $= \frac{50}{100}(2)$.
x coordinate same hence,

$$\bar{s}_1 = \begin{pmatrix}1\\0\\1\end{pmatrix}, \quad \bar{s}_2 = \begin{pmatrix}3\\1\\1\end{pmatrix}, \quad \bar{s}_3 = \begin{pmatrix}3\\-1\\1\end{pmatrix}$$

$$\alpha_1 \bar{s}_1 \bar{s}_1 + \alpha_2 \bar{s}_2 \bar{s}_1 + \alpha_3 \bar{s}_3 \bar{s}_1 = -1 \quad (\text{-ve class})$$

$$\Rightarrow \alpha_1 \begin{pmatrix}1\\0\\1\end{pmatrix}\begin{pmatrix}1\\0\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}3\\1\\1\end{pmatrix}\begin{pmatrix}1\\0\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}3\\-1\\1\end{pmatrix}\begin{pmatrix}1\\0\\1\end{pmatrix} = -1$$

$$\Rightarrow \alpha_1 (1.1+0.0+1.1) + \alpha_2 (3.1+1.0+1.1) + \alpha_3 (3.1+(-1).0+1.1) =$$

$$\Rightarrow 2\alpha_1 + 4\alpha_2 + 4\alpha_3 = -1 \quad \text{--- (1)}$$

$$\alpha_1 \bar{s}_1 \bar{s}_2 + \alpha_2 \bar{s}_2 \bar{s}_2 + \alpha_3 \bar{s}_3 \bar{s}_2 = +1 \quad (\text{+ve class})$$

$$\Rightarrow \alpha_1 \begin{pmatrix}1\\0\\1\end{pmatrix}\begin{pmatrix}3\\1\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}3\\1\\1\end{pmatrix}\begin{pmatrix}3\\1\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}3\\-1\\1\end{pmatrix}\begin{pmatrix}3\\1\\1\end{pmatrix} = 1$$

$$\Rightarrow \alpha_1 (3+1) + \alpha_2 (9+1+1) + \alpha_3 (9-1+1) = 1$$

$$\Rightarrow 4\alpha_1 + 11\alpha_2 + 9\alpha_3 = 1 \quad \text{--- (2)}$$

$$\alpha_1 \bar{s}_1 \bar{s}_3 + \alpha_2 \bar{s}_2 \bar{s}_3 + \alpha_3 \bar{s}_3 \bar{s}_3 = +1 \quad (\text{+ve class})$$

$$\Rightarrow \alpha_1 \begin{pmatrix}1\\0\\1\end{pmatrix}\begin{pmatrix}3\\-1\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}3\\1\\1\end{pmatrix}\begin{pmatrix}3\\-1\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}3\\-1\\1\end{pmatrix}\begin{pmatrix}3\\-1\\1\end{pmatrix} = 1$$

$$\Rightarrow \alpha_1 (3+1) + \alpha_2 (9-1+1) + \alpha_3 (9+1+1) = 1$$

$$\Rightarrow 4\alpha_1 + 9\alpha_2 + 11\alpha_3 = 1 \quad \text{--- (3)}$$

Solving eq^n (1), (2) & (3),

$\alpha_1 = -3.5$, $\alpha_2 = 0.75$ & $\alpha_3 = 0.75$

So, weight $\bar{w} = \sum \alpha_i \bar{s}_i$

$$= \alpha_1 \bar{s}_1 + \alpha_2 \bar{s}_2 + \alpha_3 \bar{s}_3$$

$$= -3.5 \begin{pmatrix}1\\0\\1\end{pmatrix} + 0.75 \begin{pmatrix}3\\1\\1\end{pmatrix} + 0.75 \begin{pmatrix}3\\-1\\1\end{pmatrix}$$

$$= \begin{pmatrix}-3.5\\0\\-3.5\end{pmatrix} + \begin{pmatrix}2.25\\0.75\\0.75\end{pmatrix} + \begin{pmatrix}2.25\\-0.75\\0.75\end{pmatrix}$$

$$= \begin{pmatrix}-3.5+2.25+2.25\\0+0.75-0.75\\-3.5+0.75+0.75\end{pmatrix} = \begin{pmatrix}1\\0\\-2\end{pmatrix}$$

Line equation :- $ax + b$

$$\Rightarrow (1)x + (-2) = 0$$
$$\Rightarrow x - 2 = 0$$
$$\Rightarrow x = 2$$

Backpropagation Algorithm

Step 1 : Initialization (assign the weights, inputs & biasing values)

Step 2: Feed forward (output gets computed)

Step 3: Backward computation (to improvise)

Step 4: Iterate

$$I_j = \sum_i w_i x_i + \theta_i$$

Bias value

weight

input from i^{th} unit (prev.)

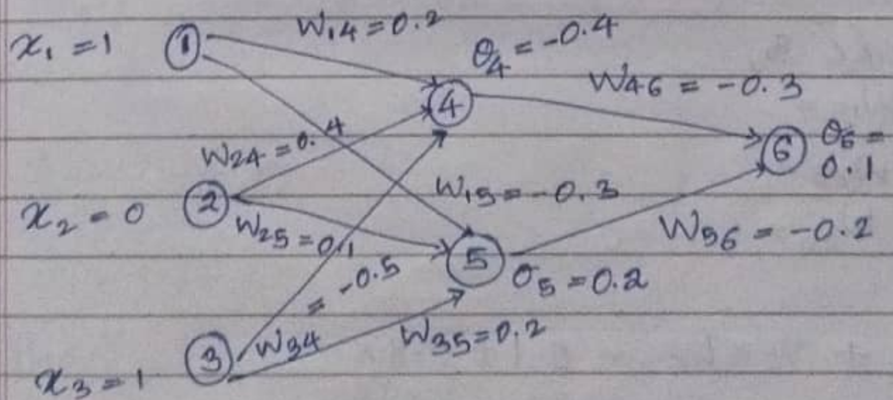
Activation

$$Y_j = f(I_j) = \frac{1}{1-}$$

Q- $x_1 = 1$ Target = 1

$x_2 = 0$ Learning rate = 0.9

$x_3 = 1$



$$Y_4 = \frac{1}{1+e^{-I_4}} = 0.332$$

$$Y_5 = \frac{1}{1+e^{-I_5}} = 0.525$$

$$Y_6 = \frac{1}{1+e^{-I_6}} = 0.474$$

$$I_4 = x_1 W_{14} + x_2 W_{24} + x_3 W_{34} + \theta_4 = -0.7$$

$$I_5 = x_1 W_{15} + x_2 W_{25} + x_3 W_{35} + \theta_5 = 0.1$$

$$I_6 = (W_{46} \times Y_4 + W_{56} \times Y_5) + \theta_6 = -0.105$$

Scanned with OKEN Scanner

$Error_6 = Y_6(1-Y_6)(T-Y_6)$ $= 0.474$

$= 0.1311$

$E_5 = Y_5(1-Y_5) \times W_{56} \times E_6$

$=$

$= -0.0065$

$E_4 = Y_4(1-Y_4) \times W_{46} \times E_6$

$=$

$= -0.0087$

Update, ~~for Node 6,~~

$W_{46} = W_{46} + \eta * E_6 \times Y_4$

$= -0.3 + 0.9 \times 0.1311$

$= -0.261$

$W_{56} = W_{56} + \eta \times E_6 \times Y_5$

$= -0.2 + 0.9 \times 0.1311 \times 0.525$

$=$

Then, for Node 4,

$W_{14} =$

$W_{24} =$

$W_{34} =$

for Node 5,

$W_{15} =$

$W_{25} =$

$W_{35} =$

Then,

$O_6 = O_6 + \eta \times E_6 = 0.1 + 0.9 \times$

$O_5 =$

$O_4 =$

Then, iterate using updated weights &
till values of error are in acceptable range

Genetic Algorithm (GA)

- Initial population (choosing sample)
- fitness function (conditions for calculation)
- Crossover (replacing one substring with another)
- Mutation (replacing randomly)
- Selection (to find result)
- Termination

Slow but parallelization is possible.

Performance Measure

Confusion Matrix →

Prediction \ Actual	Class 1 'Yes'	Class 2 'NO'
Class 1 'Yes'	TP	FP
Class 2 'NO'	FN	TN

FN → false -ve
Actual ✓ Pred ×
TP → True +ve
Predictⁿ ✓
Actual ✓
TN → True -ve
Actual × Pred ×
FP → false +ve
Actual × Pred ✓

TP from table → 3 (Both Yes)
TN → 0
FP → 5
FN → 6 (Actual-Yes, Pred-No)

Precisⁿ

3	5
6	0

Recall

Here, $\sum TP + TN + FP + FN$ = Total no. of records

1) Accuracy

$$\frac{TP+TN}{TP+TN+FP+FN}$$, error = 1 − Accuracy

2) Precision $= \dfrac{TP}{TP+FP}$

3) Recall $= \dfrac{TP}{TP+FN}$

4) F1 - Score = $\dfrac{2}{\dfrac{1}{Precis^n} + \dfrac{1}{Recall}}$

$= \dfrac{2\ Precis^n \times Recall}{Precis^n + Recall}$

5) Sensitivity = $\dfrac{TN}{FP + TN}$

eg

Actual

3 TP	6 FP
5 FN	0 TN

A = $\dfrac{3+0}{14}$ ⇒ Error = 100 − 21.4 = 78.6 ≈ 79%

Prec^n = $\dfrac{3}{9}$ = 33%

Recall = $\dfrac{3}{8}$ = 38%

F1-Score = $\dfrac{2 \times 3/9 \times 3/8}{3/9 + 3/8}$

Specificity = 0 (as TN = 0)

eg Yes = 40 → Pred^n Yes = 30 (TP), → No = 10 (FN)

No = 60 → Yes = 5 (FP), → No = 55 (TN)

30	5
10	55

A = $\dfrac{85}{100}$ = 85%, Error = $1 - \dfrac{85}{100}$ = 15%

Prec^n = $\dfrac{30}{35}$

Recall = $\dfrac{30}{40}$

F1-Score = $\dfrac{2 \times 30/35 \times 30/40}{30/35 + 30/40}$

Scanned with OKEN Scanner

- Underfitting → training is not proper works on given data but not real data.

- Overfitting → Very complex, gives proper answer but in long period of time.

★ Only linear regress^n in syllabus.

- Correlation varies from -1 to $+1$.

$$r = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{(n\Sigma X^2 - (\Sigma X)^2) \cdot (n\Sigma Y^2 - (\Sigma Y)^2)}}$$

↳ Pearson's correlat^n

Q- $n = 10$, $\Sigma X = 80$, $\Sigma Y = 255$, $\Sigma Y^2 = 7097$

$\Sigma XY = 2289$, $\Sigma X^2 = 756$

$$r = \frac{10(2289) - 80(255)}{\sqrt{(10(756) - 80^2)(10(7097) - 255^2)}}$$

$= 0.95$

So, +vely correlated.

Performance Evaluat^n of Regression

$$\text{Mean Square Error (MSE)} = \frac{\Sigma(\overset{\text{Predict}^n}{X'(t)} - \overset{\text{Actual}}{X(t)})^2}{N}$$

Root MSE $= \sqrt{MSE}$

Mean Absolute Percentage Error (MAPE)

$$= \frac{100}{N} \Sigma \frac{|X'(t) - X(t)|}{X(t)}$$

Q- 1 2

42 45

44 46

-2 -1

4 1

$$MSE = \frac{\Sigma 4+1+1+25+ \cdots +4}{12} = \frac{56}{12} = 4.6$$

$$RMSE = \sqrt{4.67} = 2.15$$

$$MAPE = \frac{100}{12}\left(\frac{-2}{42} + \frac{-1}{45} + \cdots \right)$$

$$= 3.64\%.$$

Linear Regression

$y = bx + a$ } equation

b → slope, a → intercept

Here, $b = \dfrac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma x^2 - (\Sigma x)^2}$

$$a = \frac{\Sigma Y}{n} - b\frac{\Sigma X}{n}$$

Q- $n = 10$, $\Sigma XY = 2289$, $\Sigma X = 80$, $\Sigma Y = 255$, $\Sigma x^2 = 756$

$$b = \frac{10(2289) - 80 \times 255}{10 \times 756 - 6400} = 2.146$$

$$a = \frac{255}{10} - 2.146\frac{80}{10} = 8.33$$

$Y = 2.146X + 8.33$ (Ans.).

Scanned with OKEN Scanner