Assignment - II

PAGE No
DATE : / 202

Q1

↳

Obsv.	x	y	xy	x^2	Y_p	$(Y_a - Y_p)^2$
1	10	25	2500	100	26.25	1.576
2	15	28	420	225	30.18	4.75
3	20	35	700	400	34.11	0.79
4	25	42	1050	625	38.04	15.65
5	30	43	1290	900	41.97	1.06
6	35	45	1575	1225	45.90	0.81
7	40	50	2000	1600	49.83	0.03
8	45	52	2340	2025	53.76	3.1
n=8	220	320	9625	7100		

i) $Y = mx + c$

$$m = \frac{n\Sigma xy - \Sigma x \Sigma y}{\Sigma x^2 - (\Sigma x)^2}, \quad c = \frac{\Sigma y - m\Sigma x}{n}$$

$$m = \frac{8 \cdot 9625 - 220 \cdot 320}{8 \cdot 7100 - 220}, \quad c = \frac{320 - 0.7857 \cdot 220}{8}$$

$$m = 0.7857, \quad c = 18.39.$$

$\therefore Y = 0.7857x + 18.39.$

(2) Predicted sales for 32000 spend

$x = 32$

$Y_p = 0.7857 \cdot 32 + 18.39$

$= 25.1424 + 18.39$

$= 43.53$ Lakhs

(c) MSE

MSE = ~~$\sum (Y_a - Y_p)^2$~~

MSE = $\frac{\sum (Y_a - Y_p)^2}{n}$

$= \frac{27.78}{8} = 3.47$

for $R^2 = 1 - \frac{\text{Sum of squared error}}{(\text{Total variance in } Y)^2}$

$= 1 - \frac{\sum (Y_a - Y_p)^2}{\sum (Y - \bar{Y})^2}$

$\bar{Y} = \frac{320}{8} = 40$

$= 1 - \frac{27.78}{676}$

$= 1 - 0.041$

$= 0.959$

Q2

↳

Model	Training Acc.	Test Acc.
A	98%	70%
B	85%	83%
C	60%	58%

i) Model A show high variance i.e. overfitting because it has very high training accuracy but lower test accuracy

PAGE No
DATE : / 202

Model B shows well balanced because good accuracies and close accuracies in both training and testing.

Model C shows high bias i.e. underfitting i.e. both training and testing accuracies are low and close.

(ii)

As complexity increases means from linear to polynomial, biases decrease because the model can fit training data more closely while its variance increase because the model become more sensitive to small changes.

(iii) validation or regularisation can be used

→ It helps to control overfitting and keep small model generalised.

PAGE No
DATE : / 202

Q3

Outlook	Temp.	Humidity	Wind	Play
Sunny	Hot	High	Weak	N
Sunny	Hot	High	Strong	N
Overcast	Hot	High	Weak	Y
Rain	Mild	High	Weak	Y
Rain	Cool	Normal	Weak	Y

(i) Entropy of Target (Play)

Entropy is $S(\text{play}) = S(3,2) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$

$= -0.6(-0.73) - 0.4(-1.32)$

$= 0.4421 + 0.5288$

$= 0.9709$

(ii) IG (Outlook)

Entropy of Sunny $S(2,2) = -0 - \frac{2}{2}\log_2\frac{2}{2}$

$= 0$

Entropy of Overcast $(1,0) = -1.0 - 0.$

$= 0$

Entropy of Rain $(2,0) = \frac{2}{2}.0 - 0 = 0$

IG of Outlook $= S_{bs} - 0 - 0 - 0$

$= 0.9709$

PAGE No
DATE : / 202

(iii)

IG of temp

Entropy of HOT (2,2) $= -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}$
$= 1$

Entropy of Cool (3,1) $= -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$
$= 0.811$

Entropy of Mild (4,2) $= -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}$
$= 0.918$

→ IG of temp $= 0.971 - \frac{2}{5}(1) - \frac{4}{5}(0.811) - \frac{6}{5}(0.918)$
$= 0.971 - 0.4 - 0.6488 - 0.7344$
$= 0.971 - 1.7832$
$= 0.1878$

IG of Humidity

Entropy of High (3,4) $= -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}$
$= 0.9852$

Entropy of Normal (6,1) $= -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}$
$= 0.5916$

→ IG of Humidity $= 0.971 - \frac{7}{10}(0.9852) - \frac{3}{10}(0.5916)$
$= 0.971 - 0.6896 - 0.1774$
$= 0.971 - 0.867$
$= 0.1040$

IG of wind

Entropy of Weak (6,2) $= -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}$
$= 0.8112$

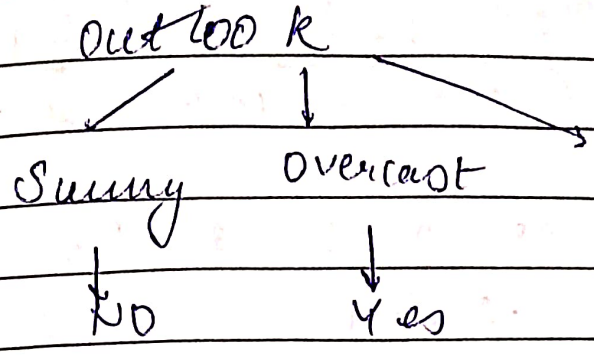
Entropy of Strong (3,3) $= 1$

IG of wind $= 0.971 - \frac{8}{10}(0.8112) - \frac{2}{10}(1)$
$= 0.971 - 0.6489 - 0.2$
$= 0.971 - 0.8489$
$= 0.1221$

Since IG of Outlook is max.
So root node is outlook

PAGE No
DATE : / 202

(4)



(iv) Outlook = Rain, Temp = Cool, Humidity = N, wind = Strong.

Since entropy of rain is 0, so we can directly say from decision tree, play. it = Yes.

Q4

Age	Income	Bought
22	20	0
25	30	0
28	35	1
35	50	1
40	60	1

→ For a new customer (Age = 30, Income = 40)

⇒ Step 1: Distance calculation

$d_1 = (22,20) \& (30,40)$

$= \sqrt{8^2 + 20^2} = \sqrt{464} = 21.5$

d_2 = (28, 30) & (30, 40)
= $\sqrt{5^2 + 10^2}$ = $\sqrt{125}$ = 11.18

d_3 = (28, 35) & (30, 40)
= $\sqrt{2^2 + 5^2}$ = 5.39

d_4 = (35, 50) & (30, 40)
= $\sqrt{5^2 + 10^2}$ = 11.18

d_5 = (50, 60) & (30, 40)
= $\sqrt{10^2 + 20^2}$ = $\sqrt{500}$ = 22.36

∴ K = 3

1st → (28 35) → Bought = 1
2nd → (25, 30) → Bought = 0
3rd → (35, 50) → Bought = 1

Final Classification : Bought = 1
Age = 30 & Income = 4

Q5

Model	TP	FP	FN	TN
A	40	10	20	30
B	45	25	15	15

N = 100

$$\text{Accuracy} = \frac{TP + TN}{FP + FN + TP + TN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

for Model A

$$A_A = \frac{40 + 30}{100} = 0.7 = 70\%$$

$$P_A = 40 / 40 + 10 = 0.8 = 80\%$$

$$R_A = 40 / 60 = 0.67 = 67\%$$

for Model B

$$A_B = 45 + 15 / 100 = 0.6 = 60\%$$

$$P_B = 45 / 45 + 25 = 0.64 = 64\%$$

$$R_B = 45 / 45 + 15 = 0.75 = 75\%$$

(11) FN (false negative) i.e. actually positive but predicted negative. Model B will be choose where FN are more costly because it has high recall than model A.

PAGE No
DATE : / 202

(ii) FP → False positive i.e. Actually negative but predicted positive. Model A will be choosen when FP are more costly because of higher precision