(12)

DataSet	T_{10}	T_{100}
A	0.86	0.97
B	0.84	0.77

1) From the table we can assume T_{100} has low train Error & high test Error implying Overfit.

From this i Conclude that T_{10} Predicts better on unSeen data

2) I would choose T_{10} over T_{100} as the total accuracy of T_{100} is only Sightly better than T_{10}. So choosing T_{10} is Computationally good decision

(8)

(a) Split A

	T	F
+	25	25
-	0	50

$IG = \epsilon_{Total} - \epsilon_{Split}$

$= 0.5 - \epsilon_{Split}$

$\epsilon_{Split} = \epsilon_{A=T} + \epsilon_{A=F}$

$\epsilon_{A=T} = 1 - \max\left(\frac{25}{25}, \frac{0}{25}\right)$

$= 1 - 1 = 0$

$\epsilon_{A=F} = 1 - \max\left(\frac{25}{75}, \frac{50}{75}\right)$

$= 1 - \frac{50}{75} = \frac{25}{75}$

$IG = 0.5 - \left(\frac{100}{100}\right)\epsilon_A^2 - \epsilon_{orig} - \left(\frac{25}{100}\right)\epsilon_{A=T} - \frac{75}{100}\epsilon_{A=F}$

$= 0.5 - 0 - \frac{75}{100} \times \frac{25}{75}$

$= 0.25$

Split at B

	T	F
+	30	20
-	20	30

$E_{B=T} = 1 - \max\left(\frac{30}{50}, \frac{20}{50}\right) \Rightarrow 1 - \frac{3}{5} = 2/5 = 0.4$ //

$E_{B=F} = 1 - \max\left(\frac{20}{50}, \frac{30}{50}\right) \Rightarrow 1 - \frac{3}{5} = 0.4$ //

$IG = 0.5 - \frac{50}{100}(0.4) - \frac{50}{100}(0.4)$

$= 0.5 - (0.2) - 0.2 = 0.1$

$IG_B = 0.1$

$\boxed{IG = 0.1}$

Split at C

	T	F
+	25	25
-	25	25

$E_{C=T} = 1 - \max\left(\frac{1}{2}, \frac{1}{2}\right) = 0.5$

$E_{C=F} = 1 - \max\left(\frac{1}{2}, \frac{1}{2}\right) = 0.5$

$IG = 0.5 - \frac{50}{100}\left(\frac{1}{2}\right) - \frac{50}{100}\left(\frac{1}{2}\right)$

$= 0.5 - 0.5 =$

$\boxed{IG_C = 0}$

Feature "A" is chose as it has highest Gain.

(b)

When A=T

B	C	+	-
T	T	5	0
F	T	20	0
T	F	0	0
F	F	0	0

Seeing above table as it contains only "+" class label no Split is required

When A=F

B	C	+	-
T	T	0	20
F	F	0	5
T	F	25	0
F	F	0	25

$E_{original_{A=F}} = \frac{25}{75} = \frac{1}{3}$

$= 1 - \max\left(\frac{25}{75}, \frac{50}{75}\right)$

$= 1 - \frac{50}{75}$

Split at B

	T	F
+	25	0
-	20	30

$E_{B=T} = 1 - \max\left(\frac{25}{45}, \frac{20}{45}\right) = 1 - \frac{5}{9} = \frac{4}{9}$

$E_{B=F} = 1 - \max\left(\frac{0}{30}, \frac{30}{30}\right) = 0$

$$E_{original} = \frac{45}{75}\left(\frac{4}{15}\right) - \frac{30}{75}(0)$$

$$\frac{1}{3}\,\frac{4}{15} = \frac{15-12}{3(15)} = \frac{1}{15}$$

$$IG_B = \frac{1}{15}$$

Split at C

	T	F
+	0	25
−	25	25

$$E_{C=T} = 1 - \max\left(\frac{0}{25}, \frac{25}{25}\right) = 0$$

$$E_{C=F} = 1 - \max\left(\frac{25}{50}, \frac{25}{50}\right) = 0.5$$

$$IG = E_{original} - \frac{25}{75}(0) - \frac{50}{75}(0.5)$$

$$= \frac{1}{3} - \frac{2}{3}(0.5)$$

$$IG_C = 0$$

Feature 'B' will be chosen as it has highest IG.

(c)

Error rate = 0.2

$50 \times 0.2 \times 100 = 20$

20 instances are misclassified.

(d)

C = T

A	B	+	−
T	T	5	0
F	T	0	20
T	F	20	0
F	F	0	5

$\epsilon_{init} = \epsilon_{original} = 1 - \max\left(\frac{25}{50}, \frac{25}{50}\right) = \frac{25}{50} = \frac{1}{2}$

Split A:

	T	F
+	25	0
−	0	25

$\epsilon_{A=T} = 1 - \max\left(\frac{25}{25}, \frac{0}{25}\right)$

$= 1 - 1 = 0$

$\epsilon_{A=F} = 1 - \max\left(\frac{0}{25}, \frac{25}{25}\right)$

$= 1 - 1 = 0$

$G_{original} = \frac{25}{50}(0)\ \frac{25}{50}(0)$

$\frac{1}{2} - 0 - 0 = \frac{1}{2}$

$I.G_A = \frac{1}{2} = 0.5$

Split B

	T	F
+	5	20
−	20	5

$G_{original} = \frac{1}{2}$

$G_{B=T} = 1 - Max\left(\frac{5}{25}, \frac{20}{25}\right)$

$= 1 - \frac{20}{25} = \frac{5}{25}$

$G_{B=F} = 1 - Max\left(\frac{20}{25}, \frac{5}{25}\right)$

$= 1 - \frac{20}{25} = \frac{5}{25}$

$IG_B = \frac{1}{2} - \frac{25}{50}\left(\frac{5}{25}\right) - \frac{25}{50}\left(\frac{5}{25}\right)$

$= \frac{1}{2} - \frac{2}{10} = \frac{5-2}{(10)} = 3/10$

$IG_B = 0.3$

$IG_A > IG_B$ so A is chosen for next Split

for C=E

A	B	+	−
T	T	0	0
F	T	25	0
T	F	0	0
T	F	0	25

When

$$\epsilon_{original} = 1 - \max\left(\frac{25}{50}, \frac{25}{50}\right)$$

$$= 1 - \frac{25}{50} = \frac{1}{2}$$

When $A=T$

	T	F
+	0	25
−	0	25

$$\epsilon_{A=T} = 1 - \max\left(\frac{0}{50}, \frac{0}{50}\right)$$

$$= 1$$

$$\epsilon_{A=F} = 1 - \max\left(\frac{25}{50}, \frac{25}{50}\right)$$

$$= 1 - \frac{1}{2} = \frac{1}{2}$$

$$IG = \frac{1}{2} - \frac{0}{50}(1) - \frac{50}{50}\left(\frac{1}{2}\right)$$

$$\boxed{IG_A = 0}$$

Split At B

When

B=T

	T	F
+	25	0
−	0	25

$E_{original} = 0.5$

$E_{B=T} = 1 - \max\left(\frac{25}{25}, \frac{0}{25}\right) = 1 - 1 = 0$

$E_{B=F} = 1 - \max\left(\frac{0}{25}, \frac{25}{25}\right) = 1 - 1 = 0$

$IG_B = 0.5 - 0 - 0 = 0.5$

$IG_B > IG_A$ So Attribute B is chosen.

(c) The Greedy Algorithm does not always give the best result

Question 5

(10)

(a) for Split A

When A=T

	B=T	B=F
+	3	1
−	1	5

When A=F

	B=T	B=F
+	4	0
−	3	3

$$E_{original} = -[0.4 \log_2 0.4 + 0.6 \log_2 0.6]$$

$$= 0.9709$$

IG at Split A

$$E_{original} - (E_{A=T}) + (E_{A=F})$$

$$E_{A=T} = -\frac{4}{7}\log\frac{4}{7} - \frac{3}{7}\log\frac{3}{7} = 0.985$$

$$E_{A=F} = -0 - \frac{3}{3}\log\frac{3}{3} = 0$$

$$IG_A = E_{Orig} - E_{A=T} - E_{A=F}$$

$$= 0.9709 - \frac{7}{10}(0.985)$$

$$= 0.2814$$

$$\boxed{IG = 0.2814}$$

$$E_{B=T} = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.811$$

$$E_{B=F} = -\frac{1}{6}\log\frac{1}{6} - \frac{5}{6}\log\frac{5}{6} = 0.65$$

$$IG_B = E_{original} - \frac{4}{10}(0.811) - \frac{6}{10}(0.65)$$

$$= 0.9709 - 0.7174$$

$$IG_B = 0.2535$$

$IG_A > IG_B$ So feature A is chosen

(12) Gini G_{am} $G_{am}_{original} = 1 - (\text{Positive Proba..})^2 - (\text{Negative prob.})^2$

$$= 1 - (0.4)^2 - (0.6)^2$$

$$= 0.48$$

Gini when Split At "A"

$$G_{A=T} = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = 0.4897$$

$$G_{A=F} = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0$$

$$IG_A = 0.48 - \frac{7}{10}(0.4897) - \frac{3}{10}(0) = 0.13$$

Gain When Split at 3

$$G_{B=1} = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.375$$

$$G_{B=2} = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2 = 0.277$$

$$\Delta = 0.48 - \left(\frac{4}{10}\right)(0.375) - \frac{6}{10}(0.277)$$

$$= 0.48 - 0.3162$$

$$= 0.1638$$

$\Delta_{GB} > \Delta_{GA}$ so feature B is chosen

(c) Yes, As we can see from the figure as the respective measures and monotonically increasing have different values even though it has same range which is shown in the figure.

(3)

(a) Probability of + class = 4/9

Probability of - class = 5/9

$$H(x) = -\sum (P_i * \log_2 P_i)$$

$$= -((4/9) * \log_2(4/9)) - ((5/9) * \log_2(5/9))$$

$$= 0.991$$

(b) for feature a_1

a_1 = True 3 positive 1 Negative

a_1 = False 1 positive 4 Negative

$$(4/9) * (-(3/4)\log_2(3/4) - (1/4)\log_2(1/4)) - (5/9) * (-(1/5)\log_2(1/5) - (4/5)\log_2(4/5))$$

$$= 0.7616$$

$$IG = Entropy_{total} - Entropy_{a_1}$$

$$= 0.991 - 0.7616$$

$$= 0.2294$$

← for ← feature a_2

If a_2 = True 2 POSITIVE 3 NEGATIVE

If a_2 = True 2 POSITIVE 3 NEGATIVE

$$\left(\frac{5}{9}\right) * \left(\left(\frac{-2}{5}\right) * \log_2\left(\frac{2}{5}\right) - \left(\frac{3}{5}\right) * \log_2\left(\frac{3}{5}\right)\right) + \frac{4}{9}\left(\left(\frac{-2}{4}\right)\log_2\left(\frac{2}{4}\right) - \left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right)\right)$$

$$= 0.9838$$

$$IG = Entropy_{Total} - Entropy_{a_1}$$

$$= 0.991 - 0.9838$$

$$= 0.0072$$

(c)

1	2	4	5	6	7	8
0.5	2	3.5	4.5	5.5	6.5	7.5
0 4 / 0 5	1 3 / 0 5	1 3 / 1 4	2 2 / 1 4	2 2 / 3 2	3 1 / 3 2	4 0 / 1 4

Split 1 $= -\left[\left(\frac{4}{9}\right)\log_2\left(\frac{4}{9}\right) + \left(\frac{5}{9}\right) * \log_2\left(\frac{5}{9}\right)\right]$ ←

$IG = 0$ //

Split 2 $\overset{W.A}{=} -\frac{1}{9}\left[1\log(1) + 0\log(0)\right] + \left(\frac{8}{9}\right)(-)\left[\left(\frac{3}{8}\right)\log_2\frac{3}{8} + \left(\frac{5}{8}\right)\log_2\frac{5}{8}\right]$

$IG = 0.1428$

Split 3 $= -\left(\frac{2}{9}\right)\left[\frac{1}{2}\log\left(\frac{1}{2}\right) + \frac{1}{2}\log\left(\frac{1}{2}\right)\right] - \left(\frac{7}{9}\right)\left(\frac{3}{7}\log_2\left(\frac{3}{7}\right) + \left(\frac{4}{7}\right)\log_2\left(\frac{4}{7}\right)\right)$ ⑧

$\therefore$ I.G = 0.00248

Split 4 $= -\left(\frac{3}{9}\right)\left(\left(\frac{2}{3}\right)\log\left(\frac{2}{3}\right) + \frac{1}{3}\log\left(\frac{1}{3}\right)\right) + \left(\frac{6}{9}\right)\log\left(\frac{2}{6}\right) + \left(\frac{4}{6}\right)\log_2\left(\frac{4}{6}\right)$

I.G = 0.0727

Step-5 $= -\left(\frac{5}{9}\right)\left(\frac{2}{5}\log\left(\frac{2}{5}\right) + \frac{3}{5}\log\left(\frac{3}{5}\right)\right) - 4/9\left(\frac{2}{4}\log\left(\frac{2}{6}\right) + \frac{2}{4}\log\frac{2}{4}\right)$

I.G = 0.0071

Step-6 $= -\left(\frac{6}{9}\right)\left(\left(\frac{3}{6}\right)\log\left(\frac{3}{6}\right) + \frac{3}{6}\log\left(\frac{3}{6}\right)\right) - \frac{3}{9}\left(\frac{1}{3}\log\frac{1}{3} + \frac{2}{3}\log\frac{2}{3}\right)$

I.G = 0.01

Step-7 $= -\frac{8}{9}\left(\left(\frac{4}{8}\right)\log\frac{4}{8} + \frac{4}{8}\log\frac{4}{8}\right) - \frac{1}{9}\left(\left(\frac{0}{1}\right)\log\left(\frac{0}{1}\right) + \left(\frac{1}{1}\right)\log\left(\frac{1}{1}\right)\right)$

I.G = 0.1022

From all the Splits "Split 2" has Highest I.G.

(d) The best split is the one which has highest Information gain (16)

← from Given "a_1", "a_2", "a_3" "a_1" has Information Gain.

(e) misClassification error$_{a_1}$: $1 - \max\left(\frac{7}{9}, \frac{2}{9}\right) = 1 - \frac{7}{9} = \frac{2}{9}$

misClassification error$_{a_2}$ $1 - \max\left(\frac{5}{9}, \frac{4}{9}\right) = 1 - \frac{5}{9} = \frac{4}{9}$

(f) $G_{ini(a_1)} = 1 - \left(\frac{7}{9}\right)^2 - \left(\frac{2}{9}\right)^2 \approx 0.34$

$G_{ini(a_2)} = 1 - \left(\frac{4}{9}\right)^2 - \left(\frac{5}{9}\right)^2 = 0.493$

(7)

misclassification Error$_x$ = $1 - \max(40/80, 40/80) = 1 - 1/2 = 0.5$

misclassification Error$_{(y)}$ = $1 - \max\left(\frac{60}{100}, \frac{40}{100}\right) = 1 - \frac{60}{100} = 0.4$

misclassification Error$_{(z)}$ = $1 - \max\left(\frac{70}{100}, \frac{30}{100}\right) = 1 - 70/100 = 03$

We pick as z it has lowest error rate i.e. 03

X=0 Z=0

misclassification Error = $1 - \max\left(\frac{15}{60}, \frac{45}{60}\right) = 1 - 45/60 = 025$

misclassification Error = $1 - \max\left(\frac{15}{40}, \frac{25}{40}\right) = 3/8 = 0375$

x=1 z=0

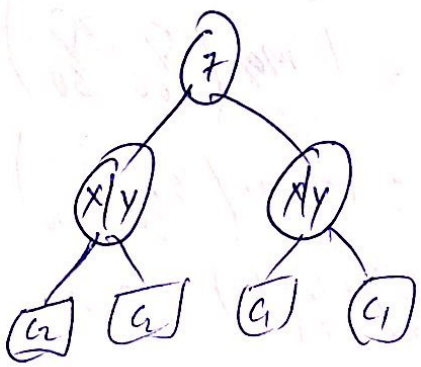
misclassification Error = $1 - \left(\frac{15}{60}, \frac{45}{60}\right) = 0.25$

misclassification Error = $1 - \left(\frac{15}{60}, \frac{25}{40}\right) = 0.375$

As the both Classification Error are Same the next split

Can be either x (or) y

Weighted avg = $\left[\left(\frac{60}{100}\right)\left(\frac{1}{4}\right)\right] + \left(\left(\frac{40}{100}\right) \times \frac{3}{8}\right) = 0.3$



(b)

$Y=1 \quad Z=1$

Misclassification Error $= 1 - \max\left[\frac{45}{60}, \frac{15}{60}\right] = \left[1 - \left(\frac{3}{4}\right)\right] = \frac{1}{4} = 0.25$

$Y=0 \quad Z=1$

Misclassification Error $= 1 - \max\left[\frac{25}{40}, \frac{15}{40}\right] = 1 - \left(\frac{5}{8}\right) = 0.375$

Weighted Avg. $= \left[\left(\frac{60}{100}\right) \times \left(\frac{1}{4}\right)\right] + \left[\left(\frac{40}{100}\right) \times \left(\frac{3}{8}\right)\right] = 0.3$

Total Weighted Error $= 0.3 + 0.3 = 0.611$

(5)

$X=0 \quad Y=0$ Misclassify Error $= 1 - \max\left(\frac{55}{60}, \frac{15}{60}\right)$

$X=0 \quad Y=1$ Misclassify Error $= 1 - \max\left(\frac{55}{60}, \frac{5}{60}\right) = 0.083$

Weighted Average: $\frac{1}{2}(0.083) + \frac{1}{2}(0.083) = 0.0833$

$X=0 \quad Z=0$ Misclassify Error $= 1 - \max\left(\frac{45}{60}, \frac{15}{60}\right) = \frac{1}{4}$

$X=0 \quad Z=1$ Misclassify Error $= 1 - \max\left(\frac{15}{60}, \frac{45}{60}\right) = \frac{1}{4}$

Weighted Average $= \left(\frac{1}{2}\right)\left(\frac{1}{4}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{4}\right) = \frac{1}{4} = 0.25$

$X=0 \quad Y=0$ Misclassify Error $= 1 - \max\left(\frac{5}{60}, \frac{35}{60}\right) = \frac{1}{8}$

$X=1 \quad Y=1$ Misclassify Error $= 1 - \max\left(\frac{35}{60}, \frac{15}{60}\right) = \frac{1}{8}$

Weighted Average $= \frac{1}{2}\left(\frac{1}{8}\right) + \frac{1}{2}\left(\frac{1}{8}\right) = 0.125$

(19)

The lowest error is for x=1 for "y"

(c) No, greedy heuristic does not create optimum result.