Name: Rohit Gorla
NET ID: rg3365

Homework 3.

Part I.

1(a)

In Dataset I,

$$P(+) = \frac{4}{9} \qquad P(-) = \frac{5}{9}.$$

In Dataset II,

$$P(+) = \frac{3}{9} = \frac{1}{3} \qquad P(-) = \frac{6}{9} = \frac{2}{3}$$

As we can see, the data set 1 is rather close to a 50-50 split, so the uncertainity to choose a class is higher.

10) b)

A)

	x_1	x_2	x_3	y
x(1)	F	F	F	+
x(2)	T	F	T	+
x(3)	F	F	T	−
x(4)	F	T	F	+
x(5)	T	T	F	−
x(6)	T	T	T	−
x(7)	F	F	F	−

v - possible values for ~~label~~ attributes (T, F)

S_v - subset of eg S satisfying $x_i = v$

$$\text{Infogain} = \text{Entropy}(S) - \sum_{v \in V} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

For x_1, $S_T = 3$; $S_F = 4$

$$\text{Entropy}(S) = -\sum_{l \in L} \frac{N_l}{N} \log_2 \frac{N_l}{N}$$

l - class labels

N_l = no of eg in S that have label l

$N_+ = 3$, $N_- = 4$

$$\text{Entropy} = -\left[\frac{3}{7}\log_2\frac{3}{7} + \frac{4}{7}\log\frac{4}{7}\right]$$

$$= \frac{-1}{7}\left[3\log_2 3 - 3\log 7 + 4\log 4 - 4\log 7\right]$$

$$= -\frac{1}{7}\left[3\log_2 3 + 4\log_2 4 - 7\log 7\right]$$

$$= \log_2 7 - \frac{3}{7}\log_2 3 - \frac{4}{7}\log_2 4$$

$$= -[-0.1577 - 0.1388] - [(0.5238) - (0.4613)]$$

$$\approx 0.9851$$

$$\sum_{v \in V} \frac{|S_v|}{|S|} \text{Entropy}(S_v) = \frac{|S_F|}{|S|}\text{Entropy}\,S_F + \frac{|S_T|}{S}\text{Entropy}\,S_T$$

$$= \frac{4}{7}\text{Entropy}(S_F) + \frac{3}{7}\text{Entropy}(S_T)$$

$$= \frac{4}{7}\left[-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right] + \frac{3}{7}\left[-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}\right]$$

$$= \frac{4}{7}\left[-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right] + \frac{3}{7}\left[-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}\right]$$

$$= \frac{4}{7}\left[1\right] + \frac{3}{7}\left[0.9182\right] = \frac{4}{7} + 0.3935$$

$$= 0.9649$$

Info Gain $= 0.9851 - 0.9649$

$= 0.0202$

(c) We know from 6 that Information gain of X_1 is 0.0202

Similarly for infogain of X_2 is,

Entropy (S) = 0.9851

$$\sum_{v \in V} \frac{|S_v|}{|S|} \text{ Entropy } (S_v)$$

$$= \frac{4}{7} \text{ Entropy } (S_F) + \frac{3}{7} \text{ Entropy } (S_T)$$

$$= \frac{4}{7}\left[\frac{-2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right] + \frac{3}{7}\left[\frac{-2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right]$$

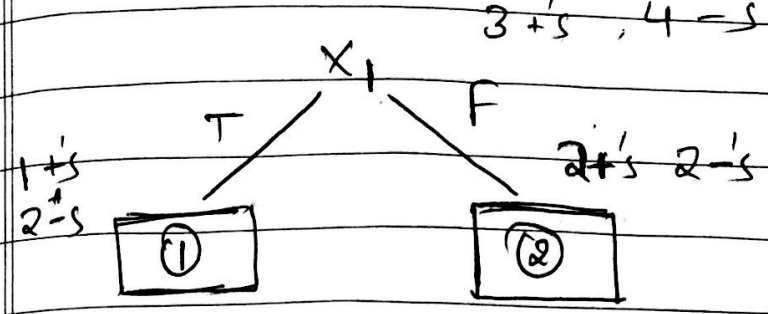
$$= 0.9649.$$

Info Gain of X_2 is = 0.0202.

Info Gain of X_3 is = 0.0202.

Comparing all we can see that they are of the same information so we can choose either for variable.

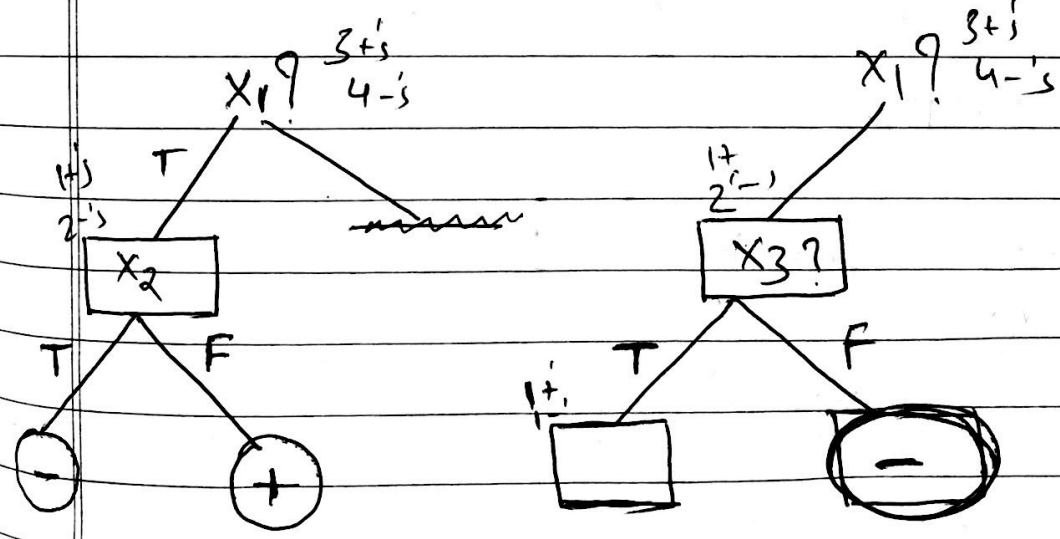
Choosing X_1 as root we get.



Let's recurse the left sub-tree.

If $X_1 = T$;

X_1	X_2	X_3	y
T	F	T	+
T	T	F	−
T	T	T	=



Now we calculate ~~for~~ info gain for x_2 and x_3 at position ①

Infogain for x_2 at (1)

$= \text{Entropy}(S) - \sum_{v \in V} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$

$= \left[\frac{-2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right] -$

$= \left[\frac{2}{3}\left[\frac{-2}{2}\log_2\left(\frac{2}{2}\right) - 0\log_2(0)\right] + \frac{1}{3}\left[\frac{-1}{1}\log_2 1 - 0\log_2 0\right]\right]$

$= 0.9182 - [0] + [0]$

$= 0.9182$

Infogain for x_3 at (1)

$= \text{Entropy}(S) - \sum_{v \in V} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$

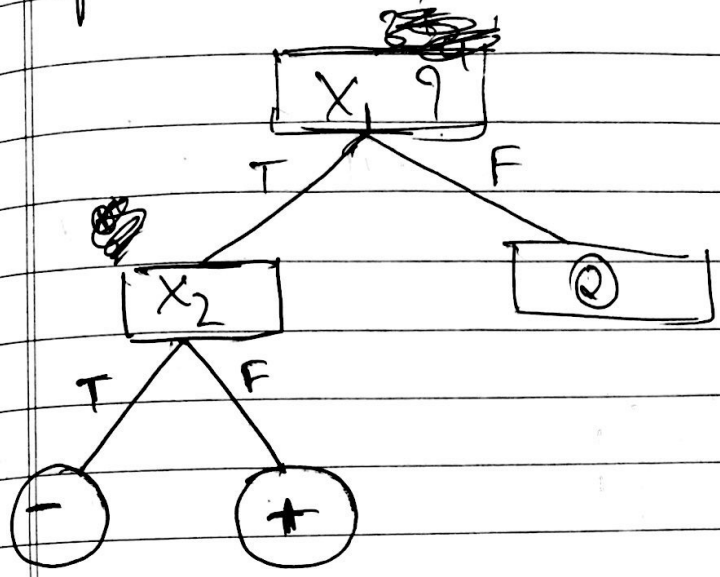
$= \left[\frac{-2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right] -$

$\left[\frac{2}{3}\left[\frac{-1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right)\right] + \frac{1}{3}\left[\frac{-1}{1}\log(1) - 0\log\right]\right]$

$= 0.9182 - \left[\frac{2}{3}[1] - \frac{1}{3}[0]\right] = 0.9182 - \frac{2}{3}$

Infogain for X_3 at ① = 0.2515

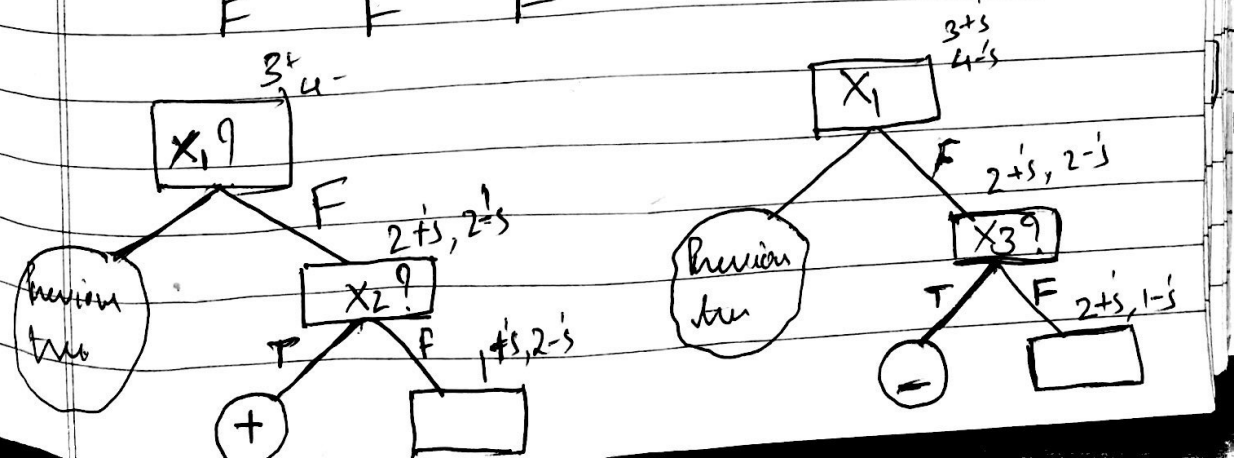
So, we can see that infogain is more for X_2 at ①.



Now, let recurse the right subtree

If $X_1 = F$

X_1	X_2	X_3	o/p
F	F	F	+
F	F	T	−
F	T	F	+
F	F	F	−



Infogain with x_2 at (2)

$$= \text{Entropy}(S) - \sum_{v \in v} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$= \left[-\frac{2}{4} \log\left(\frac{2}{4}\right) - \frac{2}{4} \log \frac{2}{4} \right]$$

$$- \left[\frac{1}{4} (\text{Entropy}(S_T) + \frac{3}{4} (\text{Entropy}(S_F)) \right]$$

$$= \left[1 \right] - \left[\frac{1}{4} \left[\frac{1}{1} \log_2 \frac{1}{1} - \frac{0}{0} \log_2 0 \right] + \frac{3}{4} \left[\frac{-2}{3} \log \frac{2}{3} - \frac{1}{3} \log \frac{1}{3} \right] \right]$$

$$= 1 - \left[0 + \frac{3}{4} \left[0.9182 \right] \right]$$

$$1 - \frac{3}{4}(0.9182) = 0.31135$$

Info gain with $x_2 = 0.31135$

Infogain with X_3 at (2)

$$= \text{Entropy}(S) - \sum_{v \in V} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

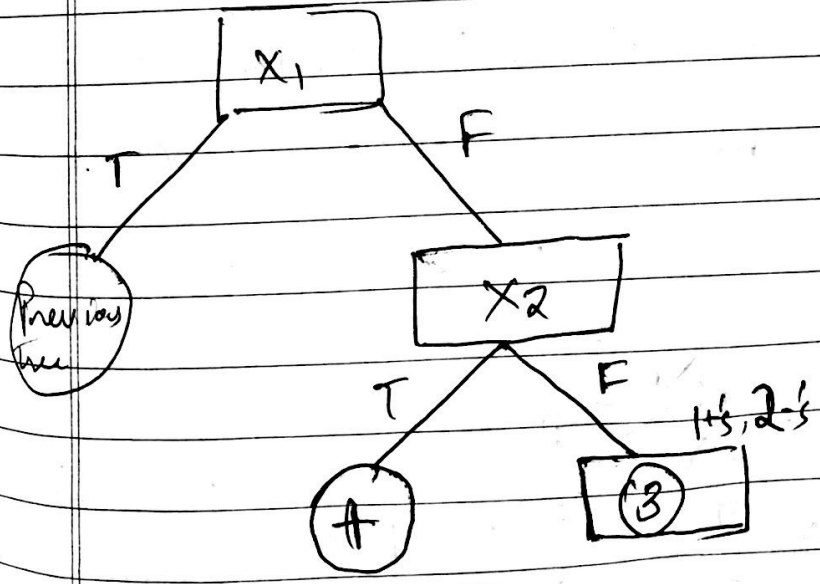
$$= 1 - \left[\frac{1}{4} \left(-\frac{1}{1} \log_2 \frac{1}{1} - \frac{0}{0} \log_2 0 \right) + \frac{3}{4} \left(\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log \frac{1}{3} \right) \right]$$

$$= 1 - \frac{3}{4}(0.9182) = 0.31135$$

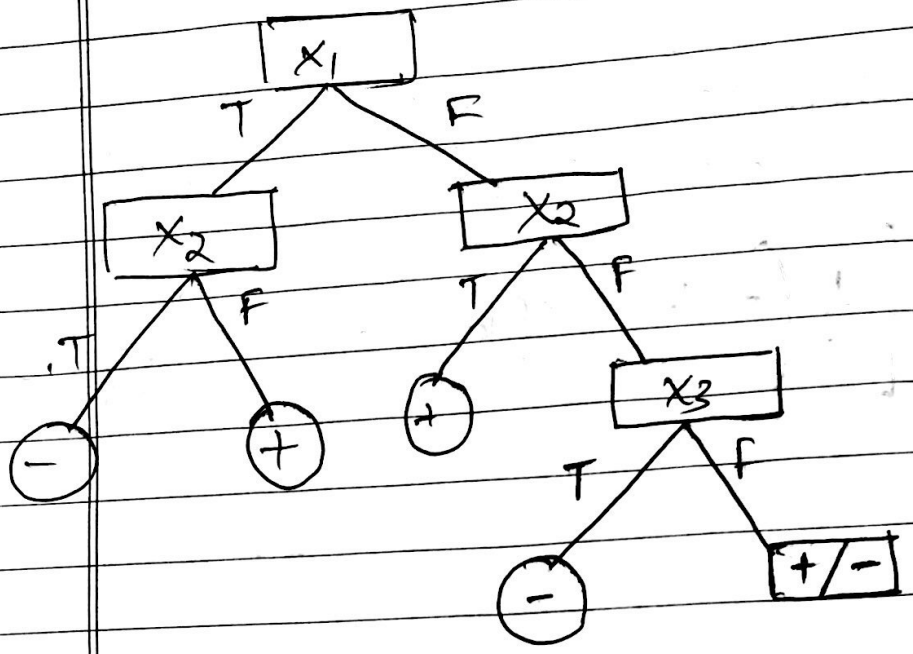
Now, we can see that both infogain are equal

So we can choose either

If we choose X_2



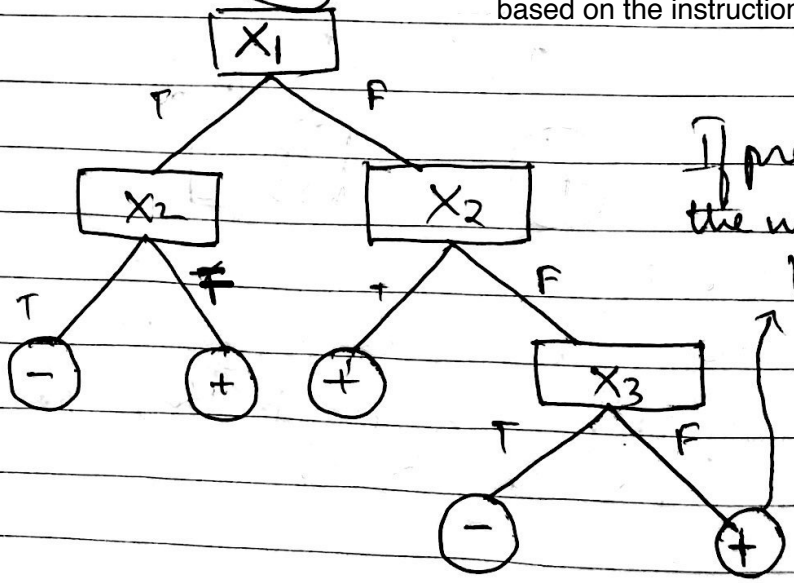
Now for position (3) we only have attribute X_3 left so, the resulting tree would be,



This will be the resulting tree, the final value is [+/-] since there are two labels assigned to examples with similar attributes, so, the tree can never be 100% accurate if we choose only one value.

There is a tie with 1+ and 1-. We choose + based on the instructions

Answer



d)

From the given information we can say that,

$$H(X) = -\sum_{i=1}^{n} P[X=i] \log_2 P[X=i]$$

and conditional entropy is

$$H(Y/X) = \sum_{x} P[X=x] * \left(\sum_{y} -P[Y=y|X=x] * \log_2 P[Y=y|X=x] \right)$$

Now, we calculate $H(X)$

$$H(X) = -\sum_{i=1}^{n} P[X=i] \log_2 P[X=i]$$

$$i \in (T, F)$$

$$= -\left[P[X=T] \log_2 P[X=T] + P[X=F] \log_2 P[X=F] \right]$$

$$= -\left[\frac{3}{7} \log_2 \frac{3}{7} + \frac{4}{7} \log_2 \frac{4}{7} \right]$$

$$= 0.9851$$

$$H(Y|X) = \sum_{x} P[X=x] * \left(\sum_{y} -P[Y=y|X=x] * \log_2 P[Y=y|X=x] \right)$$

$$x \in \{T, F\}, \quad y \in \{+, -\}$$

$$= P[X=T] * \left(-\left[P[Y=+/X=T] * \log_2 P[Y=+/X=T] \right. \right.$$

$$\left. + P[Y=-/X=T] * \log_2 P[Y=-/X=T] \right]$$

$$+ P[X=F] * \left(-\left[P[Y=+/X=F] * \log_2 P[Y=+/X=F] + \right. \right.$$

$$\left. \left. P[Y=-/X=F] * \log_2 P[Y=-/X=F] \right] \right)$$

$= \frac{3}{7} * \left[-\left[\frac{1}{3}\log\left(\frac{1}{3}\right) + \frac{2}{3}\log\left(\frac{2}{3}\right) \right] \right] +$

$\frac{4}{7} * \left[-\left[\frac{2}{3}\log\left(\frac{2}{3}\right) + \frac{2}{3}\log\left(\frac{1}{3}\right) \right] \right]$

$= \frac{3}{7}[\quad] \quad 0.3935 + 0.5714$

$= \quad 0.96492$

$H(x) = 0.9851$

$H(y|x) = \quad 0.96492$

$H(Y) - H(Y|X) =$

$H(Y)$ means $\rightarrow$ y takes values $(+,-)$

$H(Y) = -\left[P[x=+]\log_2 P[x=+] + P[x=-]\log_2 P[x=-] \right.$

$= -\left[\frac{3}{7}\log_2\frac{3}{7} + \frac{4}{7}\log_2\frac{4}{7} \right]$

$= 0.9851$

$H(Y) - H(Y|X) = 0.9851 - 0.96492 = 0.0202$

$= 0.0202$

(e) Let the no. of examples be n.

$$\text{Entropy} = -\sum_{i=1}^{2} P(x=i) \log_2 P(x=i)$$

$$= -\left[\frac{\frac{n}{2}}{n} \log \frac{\frac{n}{2}}{n}\right]$$

$$= -\sum\left[\frac{1}{2} \log \frac{1}{2}\right]$$

$$= -\left[2 \times \left(\frac{1}{2} \log \frac{1}{2}\right)\right]$$

$$= -\left[\log \frac{1}{2}\right]$$

$$= -\left[\log 1 - \log 2\right)$$

$$= \log 2.$$

So entropy is $\log 2$.

[illegible]

[illegible]

Part II : Programming exercise.

2 (a) (i)

$$A = \begin{bmatrix} 0 & 14 \\ 6 & 9 \end{bmatrix}$$

the characteristic polynomial of matrix A,

$$\det [A - KI]$$

$$= \det \begin{bmatrix} 0-K & 14 \\ 6 & 9-K \end{bmatrix}$$

= [illegible]

= [illegible]

= [illegible] —— (1)

$$-K(9-K) - 84$$

$$= K^2 - 9K - 84.$$

(b) The eigen values are the solutions to the equation $\lambda^2 - 9\lambda - 84 = 0$

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{-9 \pm \sqrt{81 - 4(-84)}}{2}$$

$\lambda_1 = 14.71$, $\lambda_2 = -5.71$

(c) Eigen vectors.

$\lambda_1 = 14.71$

$6x + (9 - 14.71)y = 0$

$6x = 5.71y$

$x = 0.9516y$.

$\sqrt{x^2 + y^2} = 1$

$\Rightarrow \sqrt{(0.9516y)^2 + y^2} = 1$

$y = 0.724$

$x = 0.9516 \times 0.724$

$= 0.6893$

$\Rightarrow [0.6893, 0.724]$

$\lambda_2 = -5.71$

$6x + (9 + 5.71)y = 0$

$x = -2.451y$

$\sqrt{(-2.451y)^2 + y^2} = 1$

$2.6497y = 1$

$y = 0.3776$

$x = -2451 \times 0.3776$

$= -0.9256$

$[-0.9256 \quad 0.3376]$

(d) In the notebook.

(3)

a) Matrix X

$$= \begin{bmatrix} 5 & 2 & 4 \\ 9 & 6 & 4 \\ 7 & 1 & 0 \\ 2 & 5 & 6 \end{bmatrix}$$

Calculating Sample mean.

$$S_1 = \frac{5+9+7+2}{4} = 5.75$$

$$S_2 = \frac{2+6+1+5}{4} = 3.5$$

$$S_3 = \frac{4+4+0+6}{4} = 3.5$$

$$B = \begin{bmatrix} -0.75 & -1.5 & 0.5 \\ 3.25 & 2.5 & 0.5 \\ 1.25 & -2.5 & -3.5 \\ -3.75 & 1.5 & 2.5 \end{bmatrix}$$

(b) Sample covariance of $S_{1,3}$.

$$= \frac{\Sigma_+ (x_i^+ - S_i) * (x_j^+ - S_j)}{N-1}$$

$$= \frac{(0.75 \times 0.5) + (3.25 \times 0.5) + (1.25 \times 3.5) + (-3.75 \times 2.5)}{4-1}$$

$= -4.1667$

= -4.1667 .

(c) The value is equal to the 1,3 entry which is -4.1667

Largest value is 12.97881

(d)

First two columns are:

0.26018674	-1.41900435
-0.8735472	4.03721245
-4.04749635	-1.8486773
4.66084433	-0.7695308 .