Homework - 3

Part - 1

Q.1 (a)

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

S: $N^+ = 4$
$N^- = 5$
$N = N^+ + N^- = 9$

S': $N^+ = 3$
$N^- = 6$
$N = 9$

$$\text{Entropy}(S) = -\left[\frac{N^+}{N} \log_2 \frac{N^+}{N} + \frac{N^-}{N} \log_2 \frac{N^-}{N}\right]$$

$$= -\frac{1}{N}\left[N^+ \log_2 N^+ - N^+ \log_2 N + N^- \log_2 N^- - N^- \log_2 N\right]$$

$$= -\frac{1}{N}\left[N^+ \log_2 N^+ + N^- \log_2 N^- - (N^+ + N^-) \log_2 N\right]$$

$$= -\frac{1}{N}\left[N^+ \log_2 N^+ + N^- \log_2 N^- - N \log_2 N\right]$$

For S and S' n is same which is 9 so term to actually compare is $N^+ \log_2 N^+ + N^- \log_2 N^-$

Entropy (S)

$$-\frac{1}{9}(4\log_2 4 + 5\log_2 5 - 9\log_2 9)$$

$$-(4\log_2 4 + 5\log_2 5)$$

Entropy (S')

$$-\frac{1}{9}(3\log_2 3 + 6\log_2 6 - 9\log_2 9)$$

$$-(3\log_2 3 + 6\log_2 6)$$

$\log_2 (4^4 5^5)^{-1}$ $\qquad\qquad$ $\log_2 (3^3 6^6)^{-1}$

$\frac{1}{4^4 5^5}$ $\qquad\qquad$ $\frac{1}{3^3 6^6}$

$\frac{3^3 6^6}{4^4 5^5}$

$= \frac{3^3 2^6 3^6}{2^4 2^4 5^5}$

$= \frac{3^9}{2^2 5^5} > \frac{3^9}{2^2 3^5} = \frac{3^4}{2^2} = \frac{81}{4} > 1$

Entropy (S) > Entropy (S')

Here S is more balanced so it has higher entropy which is proven mathematically here

$\therefore$ The example which has 4 positives and 5 negatives has higher entropy.

(b) In x_1, we have $V = [F, T]$

$S_F = \{+, -, +, -\}$
$S_T = \{+, -, -\}$

$S = \{+, +, -, +, -, -, -\}$

$$x_1 = \begin{bmatrix} F \\ T \\ F \\ F \\ T \\ T \\ F \end{bmatrix}$$

$$\text{Entropy}(S_F) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}$$

$$= 1$$

$$\text{Entropy}(S_T) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$$

$$= 0.918$$

$$\text{Entropy}(S) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}$$

$$= 0.985$$

$$IG(S) = \text{Entropy}(S) - \frac{4}{7}\text{Entropy}(S_F) - \frac{3}{7}\text{Entropy}(S_T)$$

$$= 0.985 - \frac{4}{7}(1) - \frac{3}{7}(0.918)$$

$$= 0.0202 \quad \underline{\text{Ans}}$$

(c) x_1, x_2, x_3

$IG(x_1) = 0.020$

$S_1 = + - -$

$S_2 = + - + -$

x_1	x_2	x_3	
F	F	F	+
F	F	T	-
F	T	F	+
F	F	F	-
T	F	T	+
T	T	F	-
T	T	T	-

x_2

x_3

For x_2, $S_1 = \{+\}$ $S_2 = \{+, -, -\}$

$Entropy(S_T) = 0$

$Entropy(S_F) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$
$= 0.918$

$Entropy(S) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}$

$Entropy(S) = 1$

$Ik(x_2) = Entropy(S) - \frac{3}{4}Entropy(S_F) - \frac{1}{4}Entropy(S_T)$

$= 1 - \frac{3}{4}(0.918)$
$= 0.3115$

$Ik(x_3)$ $S_1 = \{-\}$ $S_2 = \{+, +, -\}$

$Entropy(S_T) = 0$

$Entropy(S_F) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}$
$= 0.918$

$S = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$

$Ik(x_3) = 1 - \frac{3}{4}(0.918)$

$= 0.3115$

So, splitting on x_2

for other part,

For x_2, ~~$s_1=\{+,-\}$ $s_2=\{-\}$~~

$s_1=\{-,-\}$ $s_2=\{+\}$

Entropy$(S_1)=0$

Entropy$(S_2)=0$

Entropy$(S)=0.918 \longrightarrow IG(x_2)$

For x_3, $S_1=\{+,-\}$ $S_2=\{-\}$

Entropy$(S_2)=0$

Entropy$(S_1) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$

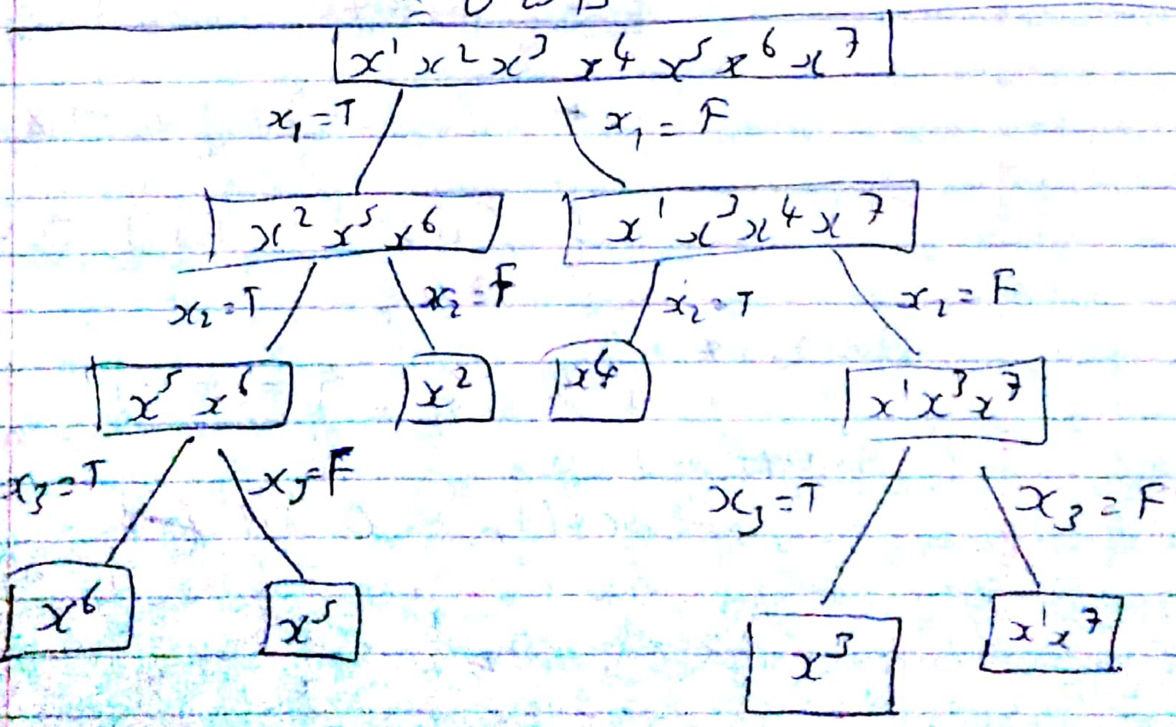
$= -\log_2\frac{1}{2}$

$= \log_2 2 = 1$

Entropy$(S) = 0.918$

$IG(x_3) = \cancel{0.198}\ 0.918 - \frac{2}{3}(1)$

$= 0.2513$



(d) $H(Y)$, $H(Y|X)$

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

$Y = [+, -]$

$H(Y) = -P(Y=+)\log P(Y=+) - P(Y=-)\log P(Y=-)$

$= -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.985$

$H(Y|X_1) = P(X_1=T)\left[-P(Y=+|X_1=T)\log_2 P(Y=+|X_1=T) + (-P(Y=-|X_1=T)\log_2 P(Y=-|X_1=T))\right]$

$+ P(X_1=F)\left[-P(Y=+|X_1=F)\log_2 P(Y=+|X_1=F) - P(Y=-|X_1=F)\log_2 P(Y=-|X_1=F)\right]$

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

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

$= \frac{3}{7}(0.918) + \frac{4}{7}.1 = 0.96486$

$= 0.965$

$H(Y) - H(Y|X_1) = 0.985 - 0.965$

$= 0.020$ Ans

(e) The entropy of a labeled dataset S with z possible labels

if $z = 2$ Total labels $= 4$
Each label appears equally so
$N_+ = 2$ $N_- = 2$ $N = 4$

$$\text{Entropy} = -\frac{N_1}{N} \log_2 \frac{N_1}{N}$$

$$= -\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4}$$

$$= -\log_2 \frac{2}{4} = 1$$ (which is equal to $\log_2(2)$ which is $\log_2(z)$)

if you take $N = 6$ then also answer remains same

$\therefore$ The entropy is $\log_2(z)$ Ans