SNLP Assignment - 4

ANKIT AGRAWAL (2581532)
AKSHAY JOSHI (2581346).

Information Theory :-

(1) $P(M) = 0.5 \qquad P(F) = 0.5$

$P(T/M) \rightarrow T = 20\%$ of M

$P(T/F) \rightarrow T = 6\%$ of f

$P(M/T) = ?$

using Bayes Theorem,

$$P(M|T) = \frac{P(T|M)\,P(M)}{P(T)} \rightarrow ?$$

we know,

$$P(T) = P(T|M)\,P(M) + P(T|F)\,P(F).$$

$$= 0.2 \times 0.5 + 0.06 \times 0.5$$

$$= 0.13$$

$$P(M|T) = \frac{0.2 \times 0.5}{0.13}$$

$$\boxed{P(M|T) = 0.77}$$

$\Rightarrow$ info gain in bits $\Rightarrow$ on $\log_2 (P(E))$

so, info gain that a male is tall is $= -\log_2 P(T|M)$

$\rightarrow -\log_2(0.2) = \boxed{2.32 \text{ bits}}$

now, info gain from learning that a female is tall

$= -\log_2 P(T \mid F)$

$= -\log_2 (0.06)$

$= \boxed{4.058 \text{ bits}}$

now, Info gain from learning that a tall person is female,

$= -\log_2 P(F \mid T)$

Again using Bayes theorem :-

$$P(F \mid T) = \frac{P(T \mid F)\, P(F)}{P(T)}$$

$$= \frac{0.06 \times 0.5}{\cancel{0} \; 0.13}$$

$P(F \mid T) = \underline{0.2307}$

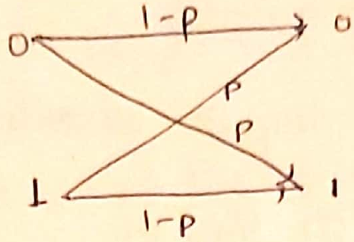
now, $-\log_2 P(F \mid T)$

$= -\log_2 (0.2307)$

$= \boxed{2.114 \text{ bits}}$.

Scanned with CamScanner

Ex 2 Entropy



→ (a) $H(x)$:- ?

Channel matrix:

	$x=0$	$x=1$
$y=0$	$1-p$	p
$y=1$	p	$1-p$

$$H(x) = -\sum_{i=0}^{1} p_i \log_2 p_i$$

$$= -\left[\tfrac{1}{2} \log_2(1/2) + 1/2 \log_2(1/2) \right]$$

$$\Rightarrow -\left[\tfrac{1}{2}(-1) + \tfrac{1}{2}(-1) \right)$$

$$\boxed{H(x) \Rightarrow 1 \text{ bits}}$$

→ (b) Probability distribution of output $P(y)$
entropy of this output distribution $H(Y)$

$Y = 0, 1$

Output probabilities

$$P(y=0) = (0.5)(1-p) + (0.5)p$$

$$\Rightarrow 0.5 - 0.5p + 0.5p$$

$$\underline{P(y=0) = 0.5}$$

Also $P(y=1) = (0.5)(1-p) + (0.5)p$

$$\boxed{P(y=1) = 0.5}$$

now, entropy of output distribution $H(Y) =$

$$H(Y) = -\sum_{i=0}^{1} p_i \log_2 p_i$$

$$= -\left[\tfrac{1}{2} \log_2 (1/2) + \tfrac{1}{2} \log_2 (1/2) \right]$$

$$\boxed{H(Y) = 1 \text{ bit}}$$

→ (c) joint probability distribution for the source and the output $P(X, Y)$.

Also joint entropy $H(X,Y)$.

using channel matrix:

$$P(X,Y) = \begin{bmatrix} 0.5(1-p) & 0.5p \\ 0.5p & 0.5(1-p) \end{bmatrix}$$

now,

joint entropy $H(X,Y) = -\sum_{x,y} p(x,y) \log_2 p(x,y)$

$$= -\left[0.5(1-p)\log_2 0.5(1-p) + 0.5p \log_2 0.5p \right.$$
$$\left. + 0.5p \log_2 0.5(p) + 0.5(1-p) \log_2 0.5(1-p) \right]$$

$$\Rightarrow -\left[(1-p)\log_2 0.5(1-p) + p\log_2 0.5p \right]$$

$$\Rightarrow -\left[(1-p)\left(\log_2 0.5 + \log_2 (1-p) \right) + (p)\left(\log_2 0.5 + \log_2 p \right) \right]$$

$$\Rightarrow -\left[(1-p)\left(\log_2 2^{-1} + \log_2(1-p)\right] + p\left[\log_2 2^{-1} + \log_2 p\right]\right)$$

$$\Rightarrow \left[(1-p)\left(-1 + \log_2(1-p)\right)\right] + p\left(-1 + \log_2 b\right)\Big]$$

$$\Rightarrow -\left[-(1-p) + (1-p)\log_2(1-p) - p + p\log_2 p\right]$$

$$\Rightarrow -\left[-1 + \cancel{p} + (1-p)\log_2(1-p) - \cancel{p} + p\log_2 p\right]$$

$$\boxed{H(x,y) = 1 - (1-p)\log_2(1-p) - p\log_2 p}$$

→(d) Mutual information of this channel $I(x;y)$.

using formula from wikipedia:-

$$I(x;y) = H(x) + H(y) - H(x,y)$$

$$\Rightarrow 1 \text{ bit} + \cancel{1} \text{ bit} - \left(\cancel{1} - (1-p)\log_2(1-p) - p\log_2 p\right)$$

$$\boxed{I(x;y) = 1 + (1-p)\log_2(1-p) - p\log_2 p}$$

Ex 3 <u>K-L Divergence:</u>

To proof K-L divergence does not follow triangular inequality.

lets use an example of 3 biased bag, each having 1 black and 1 red balls.

let P(black) for each bag be 0.1, 0.2, & 0.3 respectively
B1 B2 B3

now, KL divergence:

$$D(B_1 \| B_2) = \sum_i B_i \log_2\left(\frac{B_{1i}}{B_{2i}}\right)$$

$$\Rightarrow 0.1 \log_2\left(\frac{0.1}{0.2}\right) + 0.9 \log_2\left(\frac{0.9}{0.8}\right)$$

$$\Rightarrow 0.016 \quad \ldots (1)$$

Similarly,

$$D(B_1 \| B_3) = 0.05 \quad \ldots (2)$$

$$D(B_2 \| B_3) = 0.011 \quad \ldots (3)$$

for triangular equality, sum of any two probabilities should be greater than the third one,

but here we can see

$$D(B_1 \| B_2) + D(B_2 \| B_3) < D(B_1 \| B_3)$$

$$0.016 + 0.011 < 0.05$$

$\Rightarrow$

$\therefore$ KL divergence does not follow triangular inequality.

and thus cannot be considered a true distance metric.

Symbol	$p(x)$	$q(x)$
a	1/2	1/3
b	1/4	1/3
c	1/4	1/3

Now,

$$H(p) = 1/2 \log_2(2) + 1/4 \log_2(4) + 1/4 \log_2(4)$$

$$\Rightarrow 1/2(1) + 1/4(2) + 1/4(2)$$

$$\boxed{H(p) \Rightarrow 1.5}$$

$$H(q) = 1/3 \log_2(3) + 1/3 \log_2(3) + 1/3 \log_2(3)$$

$$\boxed{H(q) \Rightarrow 1.584}$$

Now,

$$D(p\|q) = \sum_{i=1}^{3} p_i \log\left(\frac{p_i}{q_i}\right)$$

$$\Rightarrow 1/2 \log_2\left(\frac{0.5}{1/3}\right) + 1/4 \log_2(0.25 \times 3) + 1/4 \log_2(0.25 \times 3)$$

$$\boxed{D(p\|q) \Rightarrow 0.084}$$

$$D(q\|p) = \sum_{i=1}^{3} q_i \log\left(\frac{q_i}{p_i}\right)$$

$$= 1/3 \log_2(2/3) + 1/3 \log_2(4/3) + 1/3 \log_2(4/3)$$

$$\boxed{D(q\|p) = 0.081}$$

$\therefore$ we get $D(p\|q) \neq D(q\|p)$ #

or, $0.084 \neq 0.081$

4] Codes & Entropy :-

→ Given,

A	B	C	D
1/4	1/8	1/2	1/8

(a) We know the formula to calculate the entropy per alphabet as :

$$\boxed{H(X) = -\sum_{x \in X} P(x) \cdot \log_2 P(x)}$$

For the above data

$$H(P) = -\sum_{x \in \{A,B,C,D\}} P(x) \cdot \log_2$$

$$= -\left[2 \times (1/8) \cdot \log_2 (1/8) + 1 \times (1/4) \cdot \log_2 (1/4) + 1 \times (1/2) \log_2 (1/2)\right]$$

$$= -\left[2 \times (0.125) \cdot \log_2 (0.125) + 1 \times (0.25) \cdot \log_2 (0.25) + 1 \times (0.5) \log_2 (0.5)\right]$$

$$= -\left[2 \times (0.125)(-3) + 1 \times (0.25)(-2) + 1 \times (0.5)(-1)\right]$$

$$= -\left[(-0.75) + (-0.5) + (-0.5)\right]$$

$$= -\left[-1.75\right]$$

$$= 1.75 \text{ bits / letter}$$

(b) Fixed length coding are inefficient for alphabets whose letters are not equiprobable because of the following reasons :

- The time & cost utilized to code improbable letters is equivalent to the cost spent to code for letters with much higher probability. This is the major step towards inefficiency.
- The more efficient approach would be to use variable length encoding where a smaller code is assigned to a highly probable letter which would produce a higher probability of words & a larger code to less-frequent letters.
- The same analogy / technique is already used in morse code where frequent letters in english such as e, i or s are coded with shorter codes & letters such as J, Y, U with longer codes.

(e) Uniquely decodable prefix codes for letters are:

- A : 10 or 01
- B : 110 or 111
- C : 0 0
- D : 111 or 000

Features of this unique decodable prefix codes:

- Each code represents only a single letter & is unique
- Since the codes are unique, there is no possibility of ambiguity of prefix codes for different letters.

5] Martian Codes and Kraft's inequality

→ Given,

- $S = \begin{vmatrix} S_1, \ldots\ldots S_m \\ P_1, \ldots\ldots P_m \end{vmatrix}$

- $m=6$
  - Length of code words : $(l_1, \ldots l_6) = (1,1,2,3,2,3)$

Solution :- [Ref :- How many fingers a martian - Univ of Spain]

WKT, Kraft's inequality is satisfied by a uniquely decodable code if :

$$f(D) = \underbrace{D^{-1} + D^{-1} + D^{-2} + D^{-3} + D^{-2} + D^{-3}}_{m=6}$$

& this $f(D) \leq 1$ (Prefix codes over an alphabet of size D)

* The length of codewords $l_1, \ldots l_6 \leq 1$

- At depth, D=1
  $f(1) = 3/2 > 1$ [Doesn't satisfy]

- At depth, D=2
  $f(2) = 7/4 > 1$ [Doesn't satisfy]

- At depth, D=3
  $f(2) = 26/27 \leq 1$ [Satisfies!]
  (0.962)

- From the inequality it is easy to understand that, we need base 3 to understand martian code & this is perhaps the best lower bound for D with the given length of code words!