Information Theory

Definitions

Information measure: \( I_j = \log_2 (\frac{1}{p_i}) \) bits

where \( p_i \) is the probability of the \( j \)th message \( A_j \)

Ex: let \( A_j \in \{0,1\} \) \( \Leftrightarrow \) binary symbol

\( I_0 = -\log_2(p_0) \) and \( I_1 = -\log_2(p_1) \)

so if \( p_0 = p_1 = 0.5 \) \( I_0 = I_1 = 1 \)

Ex: let \( A_j \in \{0,1,2,...,15\} \) \( \Leftrightarrow \) hex where \( p_i = \frac{1}{16} \)

The \( I_j = -\log_2 \frac{1}{16} = \log_2 16 = 4 \) bits
Average Information - Entropy

Average information in a message is called "Entropy".

Entropy is defined as:

$$H = \sum_{j=1}^{m} p_j I_j = \sum_{j=1}^{m} p_j \log_2 \left( \frac{1}{p_j} \right) \text{ bits}$$

Ex: $A_j \in \{0,1,3\}$

$$H = -p_0 \log_2 p_0 - p_1 \log_2 p_1,$$

where $p_0 + p_1 = 1$

and if $p_i = \frac{1}{2}$ then $H = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1$

Ex: $x = [000101001111]$

$$H(x) = -p \log_2 p_0 - p \log_2 p_1$$

$p_0 = \frac{\#0}{\text{total}} = \frac{6}{12} = 0.5, \ p_1 = 0.5 \Rightarrow H(x) = 1$

Max entropy

[Graph showing a curve with H on the y-axis and p on the x-axis, indicating max entropy at 1 when p = 0.5]
Example: Let \( A_j \in \{00, 01, 10, 11\} \)

\[
\begin{align*}
A &= \{00, 01, 01, 00, 11, 11\} \\
\rho_{00} &= \frac{2}{6}, \quad \rho_{01} = \frac{2}{6}, \quad \rho_{10} = \frac{1}{6}, \quad \rho_{11} = \frac{1}{6} \\
\text{Then } H &= \frac{1}{3}\log_2 3 + \frac{1}{3}\log_2 3 + 0\log_2 0 + \frac{1}{3}\log_2 3 = \log_2 3 \\
\log_2 3 &= \frac{1}{\ln(3)} \ln(3) = 1.58 \\
\end{align*}
\]

Ex: Given 4 unique symbols. A word is 6 symbols long.

The total number of permutations is \(4^6\) different words.

What is the probability of each word?

Since each symbol is equally likely and independent, the prob.

of a word is \(P_j = \frac{1}{4^6} = (\frac{1}{4})^6\)

What is the information measure of the word?

\(I_j = \log_2 \left(\frac{1}{(\frac{1}{4})^6}\right) = 6 \log_2 4 = 12 \text{ bits}\)
Discrete Channel Model

Discrete Memoryless channels are specified by a set of conditional probabilities that relate the input and output states.

The binary channel model is diagrammed as a butterfly structure:

\[ X_1 \xrightarrow{\alpha_1} Y_1 \xleftarrow{\alpha_2} X_2 \xrightarrow{\alpha_3} Y_2 \]

\[ \alpha_1 + \alpha_2 = 1 \quad \alpha_3 + \alpha_4 = 1 \]

The matrix form:

\[ P(Y|X) = \begin{bmatrix} P(Y_1|X_1) & P(Y_2|X_1) \\ P(Y_1|X_2) & P(Y_2|X_2) \end{bmatrix} = \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} \]

\[ P(Y) = [P(X_1)P(X_2)] \times [P(Y|X)] \]

\[ P(X, Y) = \begin{bmatrix} P(X_1) \times P(Y_1|X_1) \\ 0 \times P(Y_2|X_1) \end{bmatrix} \]

\[ P(X_2, Y) = \begin{bmatrix} 0 \times P(Y_1|X_2) \\ P(X_2) \times P(Y_2|X_2) \end{bmatrix} \]
Chained Binary Channels

\[ P(X) = [P(X_1), P(X_2)] \]
\[ P(Y) = [P(Y_1), P(Y_2)] \]
\[ P(Z) = [P(Z_1), P(Z_2)] \]
\[ P(Y) = P(X) P(Y|X) \]
\[ P(Z) = P(Y) P(Z|Y) = P(X) P(Y|X) P(Z|Y) \]
\[ P(Z|X) = P(Y|X) P(Z|Y) \]
Ex. Given a binary channel.

We know that

\[ P(x_1) = \rho, \quad P(x_2) = 1-P(x_1) = 1-\rho \]

\[ P(x_1|x_1) = 1 \]

\[ P(x_2|x_2) = 1 \]

\[ P(x_1|x_2) = \phi \]

\[ P(x_2|x_1) = \gamma \]

\[ P(x_2|x_1)P(x_1) = \gamma P(x_1) = \gamma \rho \]

\[ P(x_2|x_2)P(x_2) = 1-P(x_1)P(x_2) = (1-\rho) (1-\phi) = \frac{x_2}{x_2^2} + \frac{d_2}{x_2^2} \]

\[ P(y_2) = P(y_2|x_1)P(x_1) + P(y_2|x_2)P(x_2) \]

\[ = \frac{x_2 \rho + x_2 \phi}{x_2^2} \]

\[ P(x_1|y_2) = \frac{P(x_1,y_2)}{P(y_2)} \]

\[ = \frac{x_2 \rho}{x_2 \rho + x_2 \phi} \]

\[ P(x_2|y_2) = \frac{P(x_2,y_2)}{P(y_2)} \]

\[ = \frac{x_2 \phi}{x_2 \rho + x_2 \phi} \]
Joint and Conditional Entropy

Given \( p(x_i), p(y_j), p(y_j | x_i) \) and \( p(x_i, y_j) \), we can define joint and conditional entropy.

We know:

\[
H(x) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)
\]

average uncertainty

\[
H(y) = -\sum_{j=1}^{m} p(y_j) \log_2 p(y_j)
\]

are uncertainty

of the source

then the conditional entropy is

\[
H(y|x) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log_2 p(y_j | x_i)
\]

and joint

\[
H(x,y) = -\sum_{j=1}^{m} p(x_i, y_j) \log_2 p(x_i, y_j)
\]

and

\[
H(x|y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log_2 p(x_i | y_j)
\]
After we receive symbol $Y$, the joint entropy of the system uncertainty $H(X,Y)$ is given by

$$H(X,Y) = H(X|Y) + H(Y)$$

where $H(X|Y)$ is the conditional entropy and $H(Y)$ is the entropy of the received symbol $Y$. The mutual information $I(X;Y)$ measures the amount of information that one random variable contains about the other.

Other relationships:

$H(X,Y) = H(X) + H(Y) - I(X;Y)$

and

$H(Y) = H(Y|X) + I(X;Y)$
Channel Capacity

Before receiving data through a channel, an observer has an uncertainty of $H(x)$. As data is received, the uncertainty decreases so

$$H(x|y) \leq H(x)$$

The decrease in entropy is mutual information

$$I(x;y) = H(x) - H(x|y) = H(y) - H(y|x)$$

Mutual info. is a function of source prob. and channel transition prob.

We can show $H(x) \geq H(x|y)$ by showing

$$H(x|y) - H(x) = -I(x;y) \leq 0$$
This can be written as

\[ -I(X;Y) = - \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log \left( \frac{p(x_i) p(y_j)}{p(x_i, y_j)} \right) \]

since \( \log_2 x = \frac{\ln x}{\ln 2} \) and \( \frac{p(x_i)}{p(x_i, y_j)} = \frac{p(x_i) p(y_j)}{p(x_i, y_j)} \)

\[ -I(X;Y) = \frac{1}{\ln 2} \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \ln \left( \frac{p(x_i) p(y_j)}{p(x_i, y_j)} \right) \]

Using the identity \( \ln x \leq x - 1 \) and \( f(x) = \ln(x) - (x-1) \)

where \( f(1) = 0 \) the \( f(x) \) is max value of \( f(x) \)

so \[ -I(X;Y) \leq \frac{1}{\ln 2} \left[ \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i) p(y_j) - \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \right] \]

\[ I(X;Y) \geq 0 \text{ so } H(X) \geq H(X|Y) \]

The channel capacity is the maximum value of the mutual information \( I(X;Y) \).
\[ C = \max \{ I(X;Y) \} \quad \text{with respect to source prob.} \]

Ex: given a noiseless channel

\[ X_1 \rightarrow Y_1 \quad I(X_1;Y_1) = H(X_1) - H(X_1|Y_1) \]

\[ X_2 \rightarrow Y_2 \quad H(X_1Y_1) = -\sum_{i=1}^{n} p(x_i, y_i) \log_2 p(x_i, y_i) \]

\[ X^n \rightarrow Y^n \quad \text{no noise} \Rightarrow p(x_i, y_i) = p(x_i|y_i) = 0 \quad \text{for } i \neq i' \]

so \[ I(X;Y) = H(X) \]

The source entropy is maximum if all source prob. are equal

\[ C = \sum_{i=1}^{n} \frac{1}{n} \log_2 n = \log_2 n \]
Ex: Binary symmetric channel

\[ p(x_1) = x \quad \text{and} \quad p(x_2) = 1 - x \]

\[ H(Y|X) = -p \log_2 p - q \log_2 q \]

\[ I(X,Y) = H(Y) + p \log_2 p + q \log_2 q - 3 \quad \text{maximum when} \quad H(Y) \text{ is max} \]

So \[ C = 1 + p \log_2 p + q \log_2 q = 1 - H(p) \quad \text{or} \quad p(x_1) = p(x_2) = \frac{1}{2} \]

Error Probability

For a binary symmetric symmetric channel, the error prob. is

\[ P_e = \prod_{i=1}^{2^n} p(e|x_i) \cdot p(x_i) \]
where $p(e|x_i)$ is error prob. given $x_i$ input

$$p_e = q \ p(x_1) + q \ p(x_2) = q \ [p(x_1) + p(x_2)]$$

so $p_e = q$