Entropy

In information theory, **entropy** is the average amount of information contained in each message received. Here, *message* stands for an event, sample or character drawn from a distribution or data stream. Entropy thus characterizes our uncertainty about our source of information. (Entropy is best understood as a measure of uncertainty rather than certainty as entropy is larger for more random sources). The source is also characterized by the probability distribution of the samples drawn from it. The idea here is that the less likely an event is, the more information it provides when it occurs. For some other reasons (explained below) it makes sense to define information as the negative of the logarithm of the probability distribution. The probability distribution of the events, coupled with the information amount of every event, forms a random variable whose average (a.k.a. expected) value is the average amount of information, a.k.a. entropy, generated by this distribution. The units of entropy are commonly referred to as bits, but entropy is also measured in shannons, nats, or hartleys, depending on the base of the logarithm used to define it.

The logarithm of the probability distribution is useful as a measure of information because it is additive

Entropy is a measure of *unpredictability* of *information content*

If a compression scheme is lossless—that is, you can always recover the entire original message by decompressing—then a compressed message has the same quantity of information as the original, but communicated in fewer characters. That is, it has more information, or a higher entropy, per character. This means a compressed message has less redundancy.

$$H(x) = H(p) = -\sum p(x) \log_2 p(x)$$

$$= \sum p(x) \log_2 \frac{1}{p(x)}$$

$$= E\left[\log_2 \frac{1}{p(x)}\right]$$

Scanned with CamScanner

To understand the meaning of $\sum p_i \log \frac{1}{p_i}$, at first, try to define an information function, I, in terms of an event i with probability p_i. How much information is acquired due to the observation of event i? Shannon's solution follows from the fundamental properties of information:[8]

1. $I(p) \geq 0$ – information is a non-negative quantity
2. $I(1) = 0$ – events that always occur do not communicate information
3. $I(p_1 p_2) = I(p_1) + I(p_2)$ – information due to independent events is additive

The last is a crucial property. It states that joint probability communicates as much information as two individual events separately. Particularly, if the first event can yield one of n equiprobable outcomes and another has one of m equiprobable outcomes then there are mn possible outcomes of the joint event. This means that if $\log_2(n)$ bits are needed to encode the first value and $\log_2(m)$ to encode the second, one needs $\log_2(mn) = \log_2(m) + \log_2(n)$ to encode both. Shannon discovered that the proper choice of function to quantify information, preserving this additivity, is logarithmic, i.e.,

$$I(p) = \log(1/p)$$

The base of the logarithm can be any fixed real number greater than 1. The different units of information (bits for $\log_2$, trits for $\log_3$, nats for the natural logarithm ln and so on) are just constant multiples of each other. (In contrast, the entropy would be negative if the base of the logarithm were less than 1.) For instance, in case of a fair coin toss, heads provides $\log_2(2) = 1$ bit of information, which is approximately 0.693 nats or 0.631 trits. Because of additivity, n tosses provide n bits of information, which is approximately $0.693n$ nats or $0.631n$ trits.

Now, suppose we have a distribution where event i can happen with probability p_i. Suppose we have sampled it N times and outcome i was, accordingly, seen $n_i = Np_i$ times. The total amount of information we have received is

$$\sum_i n_i I(p_i) = \sum Np_i \log(1/p_i)$$

The average amount of information that we receive with every event is therefore

$$\sum_i p_i \log \frac{1}{p_i}$$

Scanned with CamScanner

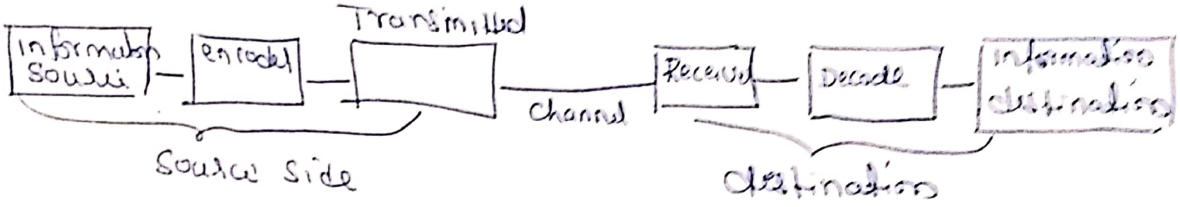
Information Theory and coding

Information: It is the intelligence / ideas / messages in information theory.

Messages:

1) Electrical Signals
2) Speech / Voice
3) Picture / image
4) Video
5) Text

In communication system information is transmitted from source to destination



Uncertainity

We have message $X = \{x_1, x_2, \ldots, x_n\}$

corresponding probabilities $P(x) = \{P_1, P_2, \ldots, P_n\}$

Total probability $P = \sum_{i=1}^{n} P_i$

Now Symbol x_0 < $P_0 = 0$ event never happen
$P_0 = 1$ event happen 100%.

$x_0 = P_0 = 0.5$ } uncertainity
$P_0 = 0.1$ } (No uncertainity) in both cases

As probability of message decreases then uncertainty increases.

Message & information

1) It is information content of messages
2) Consides an information source emitting independent messages.

Scanned with CamScanner

$X = \{x_1, x_2, \ldots\ldots x_n\}$ with probability of occurs

$P = \{P_1, P_2, \ldots\ldots P_n\}$

$\therefore P_1 + P_2 + \ldots + P_n = 1$

The amount of information is given by

$$I_k = \log_2\left(\frac{1}{P_k}\right) = \frac{\log_e\left(\frac{1}{P_k}\right)}{\log_e 2} \text{ bits/message}$$

Property of information

Thm More uncertainty of message then information is more

Let $P_1 = \frac{1}{4}$ having information U_1

$P_2 = \frac{1}{2}$ - - - - U_2

$\therefore$ Prove that $U_1 > U_2$

The amount of information of P_1 and P_2 are

$$I_1 = \log_2\left(\frac{1}{P_1}\right) = \log_2\left(\frac{1}{4}\right) = \log_2 2^2 = 2\log_2 2 = 2 \text{ bits}$$

$$I_2 = \log_2\left(\frac{1}{P_2}\right) = \log_2\left(\frac{1}{1/2}\right) = \log_2 2 = 1 \text{ bit.}$$

$\therefore I_1 > I_2 \Rightarrow U_1 > U_2$

If Receiver knowns message is transmitted, then there is no uncertinity & information is zero.

$\therefore P = 1 \quad \therefore I = \log_2\left(\frac{1}{P}\right) = \log_2(1) = 0$ bit.

Let I_1 is the information carried by message M_1

I_2 - - - - M_2

The combined information $I_{12} = I_1 + I_2$ (T.P.T)

Let $I_1 = \log_2(1/P_1) \qquad I_2 = \log_2(1/P_2)$

ie Individual amount of message M_1 and M_2

Scanned with CamScanner

Ex Since messages m_1 and m_2 are independent

$\therefore$ combined probability $p = p_1 p_2$

$\therefore I = \log_2\left(\frac{1}{p}\right) = \log_2\left(\frac{1}{p_1 p_2}\right) = \log_2\left(\frac{1}{p_1} \cdot \frac{1}{p_2}\right)$

$= \log_2\left(\frac{1}{p_1}\right) + \log_2\left(\frac{1}{p_2}\right)$

$= I_1 + I_2$

If there are $M = 2^N$ equally likely messages, then amount of information carried by each message will be equal to N bits.

The probability of each message $= \frac{1}{M}$

$\therefore I = \log_2\left(\frac{1}{p}\right) = \log_2\left(\frac{1}{1/M}\right) = \log_2 M = \log_2 2^N = N$ bits

Ex Calculate the amount of information if binary digits occur with equal likelihood in binary P.C.M.

Soln In a binary p.c.m. there are only two binary levels i.e 1 or 0

Since, they occur with equal likelihood, their probabilities of occurence will be

P_1 ('0' level) = P_2 ('1' level) $= \frac{1}{2}$

hence amount of information carried

$I_1 = \log_2\left(\frac{1}{p_1}\right)$ & $\log_2\left(\frac{1}{p_2}\right) = I_2$

$I_1 = \log_2\left(\frac{1}{1/2}\right)$ $I_2 = \log_2\left(\frac{1}{1/2}\right)$

$= \log_2 2$ $I_2 = \log_2 2$

$I_1 = 1$ bit $I_2 = 1$ bit

$\therefore I_1 = I_2 = 1$ bits information

Scanned with CamScanner

Ex: A fair coin is tossed until the first head occurs. Let X denotes the number of tosses required. Find the entropy H(X) in bits.

Use $\sum_{r=0}^{\infty} r^n = \frac{1}{1-r}$ and $\sum_{n=0}^{\infty} n r^n = \frac{r}{(1-r)^2}$

Soln: Let X be a head occurs

$\therefore$ X is a geometric random variable

$\therefore f(x) = f(n) = (1-p)^{n-1} p = q^{n-1} p.$

$$\therefore H(X) = -\sum p(x) \log p(x) = -\sum_{n=1}^{\infty} (q^{n-1} p) \log_2 (q^{n-1} p)$$

$$= -\sum q^{n-1} p \left[(n-1) \log_2 q + \log_2 p \right]$$

$$= -\sum_{n=1}^{\infty} q^{n-1} p (n-1) \log_2 q + \sum_{n=1}^{\infty} q^{n-1} p \log_2 p$$

Put $n-1 = t$ if $n=1$, $t=0$

$$\therefore H(X) = -\sum_{t=0}^{\infty} p q^t t \log_2 q - \sum_{t=0}^{\infty} q^t p \log_2 p$$

$$= -p \log_2 q \sum_{t=0}^{\infty} q^t \cdot t - p \sum_{t=0}^{\infty} q^t$$

$$= -p \log_2 q \sum_{t=0}^{\infty} q^t t - p \log_2 p \sum_{t=0}^{\infty} q^t$$

$$= -p \log_2 q \cdot \frac{q}{(1-q)^2} - p \log_2 p \cdot \frac{1}{1-q}$$

$$= -\frac{pq \log_2 q}{p^2} - \frac{p \log_2 p}{p} \quad \because 1-q = p$$

$$= -\frac{1}{p} \left[p \log_2 p + q \log_2 q \right]$$

$$= -\frac{1}{p} H(p) \text{ bits}$$

Rem. If $p = \frac{1}{2}$ then $H(X) = 2$

$p = 1/4$ then $H(X) = 3.2451$

$p = \frac{1}{8}$ $H(X) = 4.38$

Scanned with CamScanner

Ex: In a binary P.C.M if '0' occu probability $\frac{1}{4}$ and '1' occu with probability $3/4$, then calculate amount of information conveyed by each digits (binit)

Soln:

digit '0' has prob :- $P_1 = 1/4$

'1' has prob. $P_2 = 3/4$

$\therefore I_K = \log_2\left(\frac{1}{P_K}\right)$

$\therefore I_1 = \log_2\left(\frac{1}{1/4}\right) = \log_2 4 = \log_2 2^2 = 2$ bits

$I_2 = \log_2\left(\frac{4}{3}\right) = 0.415$ bits

This shows that if probability of occurence is less, information carried is more and vice versa

Ex:

Let $X = \begin{cases} a & \text{with prob. } \frac{1}{2} \\ b & \text{" } \frac{1}{4} \\ c & \text{" } \frac{1}{8} \\ d & \text{" } \frac{1}{8} \end{cases}$

Find the entropy $H(X)$.

Soln: By definition $H(X) = \sum P(x_i)\log_2 \frac{1}{P(x_i)}$

$= \frac{1}{2}\log_2\left(\frac{1}{1/2}\right) + \frac{1}{4}\log_2\left(\frac{1}{1/4}\right) + \frac{1}{8}\log_2\left(\frac{1}{1/8}\right) + \frac{1}{8}\log_2\left(\frac{1}{1/8}\right)$

$= \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 2^2 + \frac{1}{8}\log_2 2^3 + \frac{1}{8}\log_2 2^3$

$= \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{3}{8} = \frac{4+4+3+3}{8} = \frac{14}{8} = 7/4$ bits

Ex: Find the entropy of the random variable

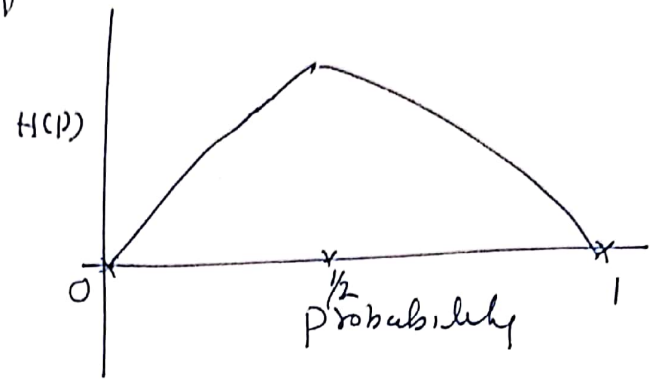
$X = \begin{cases} 1 & \text{with prob. } \frac{1}{2} \\ 0 & \text{with prob } \frac{1}{2} \end{cases}$

Scanned with CamScanner

Sol: By definition

$$H(X) = \sum p(x) \log_2 \frac{1}{p(x)} = \frac{1}{2}\log_2 2 + \frac{1}{2}\log_2 2 = 1 \text{ bit.}$$

The graph of the function H(P) is as shown in figure



Basic properties of entropy

It is a concave function of the distribution and equals 0 when $p=0$ or $p=1$, i.e the variable is not random and there is no uncertainity.
Uncertainity is maximum when $p=\frac{1}{2}$ which is corresponds to the maximum value of the entropy.

Joint entropy: Let X and Y are two discrete random variables with the joint probability $p(x,y)$ then the joint entropy of X, Y is denoted by $H(X,Y)$ and is defined as

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

which can also be expressed as

$$H(X,Y) = -E[\log_2 p(x,y)] = E\left[\log_2 \frac{1}{p(x,y)}\right]$$

Conditional entropy Let X and Y are two discrete random variable with probability $p(y/x)$ or $p(x/y)$.
Then the condition entropy $H(Y/X)$ is defined as

$$H(Y/X) = -\sum_x \sum_y p(x,y) \log_2 p(y/x)$$

$$= -E[\log_2 (p(y/x))]$$

Scanned with CamScanner

Thm: The entropy of a pair of a random variables is the entropy of one plus the conditional entropy of the other.

i.e. $H(X,Y) = H(X) + H(Y/X)$ or $H(Y) + H(X/Y)$

Proof: $\left[\text{W.K.T } P(A \cap B) = P(A)\,P(B/A) = P(B)\,P(A/B) \right]$

We have from the definition

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log p(x,y)$$

$$= -\sum_x \sum_y p(x,y) \log \left[p(x)\, p(y/x) \right]$$

$$= -\sum_x \sum_y p(x,y) \left[\log p(x) + \log p(y/x) \right]$$

$$= -\sum_x \sum_y p(x,y) \log p(x) - \sum_x \sum_y p(x,y) \log p(y/x)$$

$$= -\sum_x p(x) \log p(x) - \sum_x \sum_y p(x,y) \log p(y/x)$$

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

Similarly

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

Ex: The input source to a noisy communication channel is a random variable X over the four symbols a, b, c, d. The output from this channel is a random variable Y over these same four symbols. The joint distribution of these two random variables is as follows

y \ x	a	b	c	d
a	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
b	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	0
c	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$	0
d	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$	0

Scanned with CamScanner

a) Find the marginal distribution for x and y and compute marginal entropy H(x) and H(y) in bits

The joint distribution of x and y are

x	a=1	b=2	c=3	d=4
P(x)	1/4	1/4	1/4	1/4

y	a	b	c	d
P(y)	1/2	1/4	1/8	1/8

Marginal entropy of x is $H(x) = \frac{4}{4}\left[\frac{1}{4}\log_2 \frac{1}{4}\right]$

$= 4\left[\frac{1}{2}\right] = 2$ bits

y : $H(y) = \frac{1}{2}\log 2 + \frac{1}{4}\log 2^2 + \frac{2}{8}\log 2^3$

$= \frac{1}{2} + \frac{1}{2} + \frac{3}{4} = 7/4$ bits

b) Find the joint entropy H(x,y)

$H(x,y) = 1 \times \frac{1}{4}\log_2 2^2 4 + 2\left[\frac{1}{8}\log_2 2^3 8\right] + 6 \times \left[\frac{1}{16}\log_2 2^4 16\right] + 4\left[\frac{1}{32}\log_2 2^5 32\right]$

$= \frac{1}{4} \cdot 2 + 2 \times \frac{1}{8} \times 3 + 6 \times \frac{1}{16} \times 4 + 4 \times \frac{1}{32} \times 5$

$= \frac{1}{2} + \frac{3}{4} + \frac{3}{2} + \frac{5}{8} = \frac{4+6+12+5}{8} = \frac{27}{8}$ bits

c) Find H(y/x) and H(x/y)

We have

$H(x,y) = H(x) + H(y/x)$ &

$\therefore H(y/x) = H(x,y) - H(x)$

$= \frac{27}{8} - 2$

$= \frac{11}{8}$ bits

$H(x,y) = H(y) + H(x/y)$

$H(x/y) = H(x,y) - H(y)$

$= \frac{27}{8} - \frac{7}{4}$

$= \frac{13}{8}$ bits

Scanned with CamScanner

Ex: Given the joint probability $p(x,y)$

x \ y	0	1	f(x)
0	1/3	1/3	2/3
1	0	1/3	1/3
g(y)	1/3	2/3	1

Find $H(X)$, $H(Y)$, $H(X,Y)$, $H(Y/X)$, $H(X/Y)$

Soln: The marginal probability distribution of random variable x and y are

x	0	1
p(x)	2/3	1/3

y	0	1
p(y)	1/3	2/3

∴ The marginal entropy of x and y are

$$H(X) = -\sum p(x) \log_2 \frac{1}{p(x)} = -\left[\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right] = 0.9183$$

$$H(Y) = -\sum p(y) \log_2 \frac{1}{p(y)} = -\left[\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right] = 0.9183$$

Joint entropy

$$H(X,Y) = -3 \times \left[\frac{1}{3}\log_2\frac{1}{3}\right] = 1.5849$$

The conditional entropy

$$H(Y/X) = H(X,Y) - H(X) = 1.5849 - 0.9183 = 0.6666$$

$$H(X/Y) = H(X,Y) - H(Y) = 0.6666$$

Scanned with CamScanner

Ex: Consider the joint entropy probability $P(x,y)$

x \ y	a	b	c	f(x)
1	1/6	1/12	1/12	1/3
2	1/12	1/6	1/12	1/3
3	1/12	1/6	1/12	1/3
g(y)	1/3	5/12	1/4	1

Find $H(X)$, $H(Y)$, $H(X,Y)$, $H(X/Y)$ & $H(Y/X)$

Soln: The Marginal distribution of x and y are

x	1	2	3
p(x)	1/3	1/3	1/3

y	a	b	c
p(y)	1/3	5/12	1/4

The Marginal entropy

$$H(X) = -3\left[\frac{1}{3}\log_2\frac{1}{3}\right] = 1.5849$$

$$H(Y) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{5}{12}\log_2\frac{5}{12} - \frac{1}{4}\log_2\frac{1}{4}$$

$$= \frac{1}{3}\log_2 3 + \frac{5}{12}\log_2\frac{12}{5} + \frac{1}{4}\log_2 4 = 1.5546$$

Joint Entropy

$$H(X,Y) = 3\left[\frac{1}{6}\log_2 6\right] + 6\times\left[\frac{1}{12}\log_2 12\right] = 3.0849$$

$H(Y/X) = H(X,Y) - H(X) =$

$H(X/Y) = H(X,Y) - H(Y) =$

Note:

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

$H(X)$, $H(Y)$ → uncertainty without knowing anything about Recieved

$H(Y/X)$, $H(X/Y)$ → uncertainty after knowing about Recieved

$H(X,Y)$ → uncertainty towards complete Recieved

Scanned with CamScanner

Relative Entropy and Mutual information

The relative entropy is a measure of the distance between distributions. It arises as an expected logarithm of the likelihood ratio.

The relative entropy denoted by $D(p||q)$

Defn: The relative entropy or Kullback-Leibler distance between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$D(p||q) = \sum_x p(x) \log_2 \left[\frac{p(x)}{q(x)}\right]$$

$$= E_p \log_2 \frac{p(x)}{q(x)} = -E_q \left[\log \frac{q(x)}{p(x)}\right]$$

Defn: Two random variables x and y with a joint probability mass function $p(x,y)$ and marginal probability mass functions $p(x)$ and $p(y)$.

The Mutual information $I(X:Y)$ is the relative entropy between the joint distribution and the product distribution $p(x)p(y)$

i.e. $$I(X:Y) = \sum_x \sum_y p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

$$= D\left[p(x,y) || p(x)p(y)\right]$$

$$= E_{p(x,y)} \log \frac{p(x,y)}{p(x)p(y)}$$

Scanned with CamScanner

Properties of mutual information

1) Mutual is symmetric
$$I(X;Y) = I(Y;X)$$

2) Mutual information is always non-negative
$$I(X;Y) \geq 0$$

3) Mutual information may be written as entropies
$$I(X,Y) = H(X) - H(Y/Y) = H(Y) - H(Y/X)$$

4) Mutual information is related to joint entropy $H(X,Y)$
$$I(X:Y) = H(X) + H(Y) - H(X,Y)$$

Relations:

The mutual information is

$$I(X:Y) = \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

$$= \sum_{x,y} p(x,y) \log_2 \frac{p(x/y)}{p(x)}$$

$\because p(x/y) = \frac{p(x,y)}{p(y)} = \frac{p(x,y)}{p(x)}$

$$= \sum_{x,y} p(x,y) \left[\log_2 p(x/y) - \log_2 p(x)\right]$$

$$= -\sum p(x,y) \log_2 p(x) + \sum p(x,y) \log_2 p(x/y)$$

$$= -\sum p(x) \log_2 p(x) - \left[-\sum p(x,y) \log_2 p(x/y)\right]$$

$$I(X:Y) = H(X) - H(X/Y)$$

Thus, the mutual information $I(X:Y)$ is the reduction in the uncertainity of X due to the knowledge of Y.

By Symmetry, it is also prove that

$$I(X:Y) = H(Y) - H(Y/X) = I(Y:X)$$

Thus, X says as much about Y as Y says about X.

Scanned with CamScanner

Since $H(X;Y) = H(X) + H(Y/X)$, then

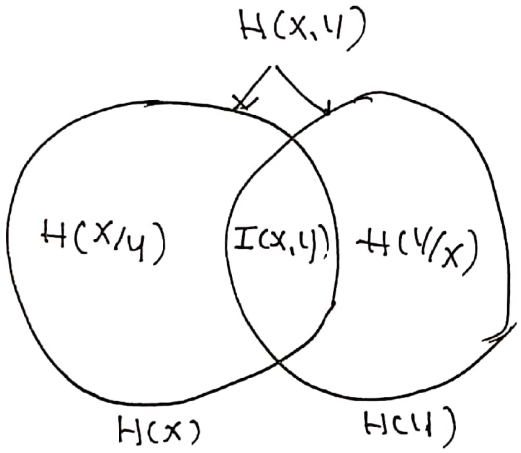
$$I(X:Y) = H(Y) - H(Y/X)$$

$$= H(Y) - \left[H(X,Y) - H(X)\right]$$

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

and $$I(X,X) = H(X) \qquad \because H(X/X) = 0$$

Thus, the mutual information of a random variable with itself is the entropy of the random variable. Therefore, such entropy is called self-information.



Chain rule for entropy:

The entropy of a collection of random variables is the sum of the conditional entropies

Thm: The messages $x_1, x_2 \cdots x_n$ be drawn according to $p(x_1, x_2 \cdots x_n)$. Then

$$H(X_1, X_2 \cdots X_n) = \sum_{i=1}^{n} H\left(\frac{X_i}{X_{i-1}, X_{i-2} \cdots X_1}\right)$$

proof: By repeated application of the two-variable expansion rule of entropies, we have

$$H(X_1, X_2) = H(X_1) + H(X_2/X_1)$$

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2, X_3/X_1) = H(X_1) + H(X_2/X_1) + H\left(\frac{X_3}{X_2, X_1}\right)$$

So on

$$H(X_1, X_2, \cdots X_n) = H(X_1) + H(X_2/X_1) + H\left(\frac{X_3}{X_2, X_1}\right) + \cdots + H\left(\frac{X_n}{X_{n-1}, X_{n-2} \cdots X_1}\right)$$

$$= \sum_{i=1}^{n} H\left(\frac{X_i}{X_{i-1}, X_{i-2} \cdots X_1}\right)$$

Scanned with CamScanner

Ex: Let $X=\{0,1\}$ and two distributions p and q on X. Let $p(0)=1-r$, $p(1)=r$, and $q(0)=1-s$, $q(1)=s$. Then

$$D(p\|q) = (1-r)\log\frac{(1-r)}{(1-s)} + r\log\frac{r}{s} \quad ---(1)$$

and

$$D(q\|p) = (1-s)\log\frac{(1-s)}{(1-r)} + s\log\frac{s}{r} \quad --(2)$$

Verify that $D(p\|q) = D(q\|p)$.

Soln: If $r=s$, then $p=q$

$\therefore D(p\|q) = D(q\|p) = 0$,

If $r\neq s$, let $r=\frac{1}{2}$, $s=1/4$, then

$$D(p\|q) = \frac{1}{2}\log_2\left(\frac{1/2}{3/4}\right) + \frac{1}{2}\log_2\left(\frac{1/2}{1/4}\right)$$

$$= \frac{1}{2}\log_2\left(\frac{4}{6}\right) + \frac{1}{2}\log_2(4/2)$$

$$= \frac{1}{2}\log_2\left(\frac{2}{3}\right) + \frac{1}{2}\log_2 2$$

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

$$= 1 - \frac{1}{2}\log_2 3 = 0.2073 \text{ bits}$$

&

$$D(q\|p) = \frac{3}{4}\log\left(\frac{3/4}{1/2}\right) + \frac{1}{4}\log\left(\frac{1/4}{1/2}\right)$$

$$= \frac{3}{4}\log_2\left(\frac{3}{2}\right) + \frac{1}{4}\log_2\left(\frac{1}{2}\right)$$

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

($\log_2 2 = 1$, $\log_2 1 = 0$)

$$= -\frac{3}{4} + \frac{3}{4}\log_2 3 - \frac{1}{4} = \frac{3}{4}\log_2 3 - 1$$

$$= 0.1887 \text{ bits}$$

$$\therefore D(p\|q) = 0.2073 \neq 0.1887 = D(q\|p)$$

Scanned with CamScanner

Def? The conditional mutual information of random variables X and Y given Z is defined as

$$I(X:Y|Z) = H(X|Z) - H(X|Y:Z)$$
$$= E_{p(x,y,z)}\left[\log \frac{p(x,y|z)}{p(x|z)\,p(y|z)}\right]$$

which is also satisfies a chain rule:

Thm? prove that

$$I(x_1, x_2, \dots x_n : y) = \sum_{i=1}^{n} I\left(x_i : y \,/\, x_{i-1}, x_{i-2}, \dots x_1\right)$$

proof: By definition, we have

$$I(x_1, x_2 \dots x_n ; y) = H(x_1, x_2 \dots x_n) - H(x_1, x_2 \dots x_n | y)$$
$$= \sum_{i=1}^{n} H(x_i / x_{i-1}, x_{i-2} \dots x_1) - \sum_{i=1}^{n} H(x_i / x_{i-1}, \dots x_1 : y)$$
$$= \sum_{i=1}^{n} I(x_i ; y / x_1, x_2 \dots x_{i-1})$$

Def? Let $p(x,y)$ and $q(x,y)$ are joint probability mass functions, the conditional relative entropy $D[p(y|x) \| q(y|x)]$ is the average of the relative entropies between the conditional probability mass functions $p(y|x)$ and $q(y|x)$ averaged over the probability mass function $p(x)$

$$D[p(y|x) \| q(y|x)] = \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)}$$
$$= E_{p(x,y)} \log \frac{p(y|x)}{q(y|x)}$$

Thm: prove that $D[p(x,y) \| q(x,y)] = D[p(x) \| q(x)] + D[p(y|x) \| q(y|x)]$

proof: By definition, we have

$$D[p(x,y) \| q(x,y)] = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x,y)}$$
$$= \sum_x \sum_y p(x,y) \log \frac{p(x)\,p(y|x)}{q(x)\,q(y|x)} = \sum_x \sum_y p(x,y) \left[\log p(x) p(y|x) - \log q(x) q(y|x)\right]$$
$$= \sum_x \sum_y p(x,y) \left[\log p(x) - \log q(x) + \log p(y|x) - \log q(y|x)\right]$$
$$= \sum \sum p(x,y) \left[\log \frac{p(x)}{q(x)} + \log \frac{p(y|x)}{q(y|x)}\right]$$

Scanned with CamScanner

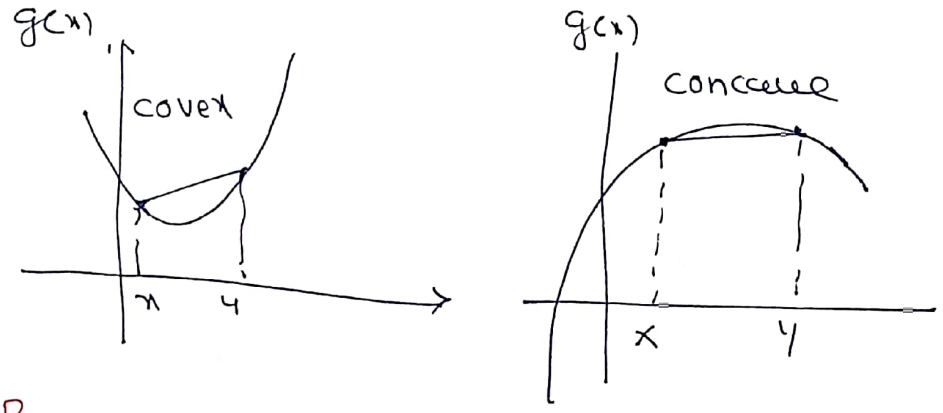
$$= \sum_x \sum_y p(x,y) \log \frac{p(x)}{q(x)} + \sum_x \sum_y p(x,y) \log \frac{p(y|x)}{q(y|x)}$$

$$= D\left[p(x) \| q(x)\right] + D\left[p(y|x) \| q(y|x)\right]$$

Scanned with CamScanner

Jensen's inequality and its consequences

A function is convex, if when you pick any two points on the graph of the function and drawn a line segment between two points, the entire segment lies above the graph. If ~~does~~ the line segment always lies below the graph, the function is said to be concave. If $g(x)$ is convex if and only if $-g(x)$ is ~~convex~~ concave.



OR A function $f(x)$ is said to be convex over an interval (a,b) if for every $x_1, x_2 \in (a,b)$ and $0 \le \lambda \le 1$

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2)$$

A function f is said to be strictly convex if equality holds if $\lambda = 0$ or $\lambda = 1$.

Thm: If the function f has a second derivative that is non-negative (+ve) over an interval, the function is convex over that interval.

proof: The Taylor series expansion of the function $f(x)$ around x_0.

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x^*)}{2!}(x-x_0)^2 + \cdots$$

Scanned with CamScanner

where x^* lies between x_0 and x, by hypothesis $f''(x^*) \geq 0$ and thus the last term is non-negative for all x.

Let $x_0 = \lambda x_1 + (1-\lambda) x_2$

and take $x = x_1$, then

$$f(x_1) \geq f(x_0) + f'(x_0)(1-\lambda)(x_1 - x_2) \quad -- (1)$$

Similarly $x = x_2$

$$f(x_2) \geq f(x_0) + f'(x_0)\lambda(x_2 - x_1) \quad -- (2)$$

Multiplying (1) by λ and (2) by $(1-\lambda)$ and adding, we get

$$\lambda f(x_1) + (1-\lambda) f(x_2) \geq \lambda f(x_0) + \lambda f'(x_0)(1-\lambda)(x_1 - x_2) + (1-\lambda) f(x_0) + (1-\lambda) f'(x_0) \lambda (x_2 - x_1)$$

$$= \lambda f(x_0) + f(x_0) - \lambda f(x_0) + \lambda(1-\lambda) f'(x_0)(x_1 - x_2) - \lambda(1-\lambda) f'(x_0)(x_1 - x_2)$$

$$\geq f(x_0)$$

$$\therefore \lambda f(x_1) + (1-\lambda) f(x_2) \geq f\left(\lambda x_1 + (1-\lambda) x_2\right)$$

Thm: If f is a convex function and X is a random variable

$$E[f(x)] \geq f[E(x)]$$

If f is strictly convex $[f''(x) > 0]$, then $X = E(X)$ with probability 1 (ie X is a constant)

proof: It can be prove by mathematical induction on the number of mass point for discrete distribution

Scanned with CamScanner

For a two-mass-point distribution, the inequality is

$$P_1[f(x_1)] + P_2[f(x_2)] \geq f(p_1x_1 + p_2x_2) \quad -(1)$$

by definition of convex function by J.E

Suppose that the theorem is true for distribution with $(k-1)$ mass points. Then

$$\text{i.} \sum_{i=1}^{k-1} p_i f(x_i) \geq f\left[\sum_{i=1}^{k-1} p_i x_i\right] \quad -(2)$$

Then To prove that it is true for k mass points distribution

Let $p_i' = \frac{p_i}{(1-p_k)}$ for $i = 1, 2 \dots k-1$

$$\therefore \sum_{i=1}^{k} p_i f(x_i) \geq p_k f(x_k) + (1-p_k) \sum_{i=1}^{k-1} p_i' f(x_i)$$

$$\geq p_k f(x_k) + f\left[p_k x_k + (1-p_k) \sum_{i=1}^{k-1} p_i' x_i\right]$$

$$\geq f\left[\sum_{i=1}^{k} p_i x_i\right]$$

i Expected Value of the function is greater than or equal to function of expected value.

Thm Let $p(x), q(x), x \in X$ be two probability mass functions, then

$$D(p\|q) \geq 0$$

with equality if and only if $p(x) = q(x), \forall x$.

Proof Let $A = \{x : p(x) > 0\}$ be the support set of $p(x)$, then

$$-D(p\|q) = -\sum p(x) \log \frac{p(x)}{q(x)}$$

$$= \sum p(x) \log \frac{q(x)}{p(x)}$$

Scanned with CamScanner

$$-D(p\|q) \leq \log \sum_{x \in A} p(x) \cdot \frac{q(x)}{p(x)}$$

$$\leq \log \sum q(x)$$

$$\leq \log 1$$

$$= 0$$

$$\therefore D(p\|q) \geq 0$$

OR From Jensen's inequality, we have

$$D(p\|q) \leq \log \sum p(x) \frac{q(x)}{p(x)} \quad -(1)$$

Since $\log t$ is a strictly concave function of t, Equality holds good iff $\frac{q(x)}{p(x)}$ is constant

i.e. $q(x) = c\,p(x)$ every where

Thus $\sum_{x \in A} q(x) = c \sum_{x \in A} p(x) = c$

But $\log \sum q(x) = \sum q(x) = 1 \Rightarrow c = 1$

Note $\sum_{x \in A} q(x) = c =$

Here, we have

$D(p\|q) = 0$ iff $p(x) = q(x)$, $\forall x$.

Cor: [Non-negative of Mutual information]

For any two random variables X, Y

$I(X,Y) \geq 0$

If and only if X and Y are independent

Proof: By Definition

$$I(X,Y) = D(p(x,y) \,\|\, p(x)\,p(y)) \geq 0$$

Since X and Y are independent

$$p(x,y) = p(x)\,p(y)$$

Scanned with CamScanner

Cor 2: $D[p(y/x) \| q(y/x)] \geq 0$ iff $p(y/x) = q(y/x)$

Cor 3: $I(x; y/z) \geq 0$ iff x and y are conditionally independent given z.

Ex Let x, y have the following joint distribution

x \ y	1	2	f(x)
1	0	1/8	1/8
2	3/4	1/8	7/8
f(y)	3/4	1/4	1

$$H(x) = H\left(\tfrac{1}{8}, \tfrac{7}{8}\right) = +\frac{1}{8}\log_2 8 + \frac{7}{8}\log_2 \frac{8}{7}$$

$$= \frac{1}{8}\log_2 2^3 + \frac{7}{8}\left[\log_2 8 - \log_2 7\right]$$

$$= \frac{3}{8} + \frac{7}{8}\left[3 - \log_2 7\right]$$

$$H(y) = \frac{3}{4}\log_2 \frac{4}{3} + \frac{1}{4}\log_2 4 = \frac{3}{4}\left[\log_2 4 - \log_2 3\right] + \frac{1}{2} \times 2$$

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

$$H(x,y) = 0 + \frac{1}{8}\log_2 8 + \frac{3}{4}\log_2 \frac{4}{3} + \frac{1}{8}\log_2 8$$

$$= \frac{3}{8} + \frac{3}{8} + \frac{3}{4}\left[\log_2 4 - \log_2 3\right]$$

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

$$H(x/y) = H(x,y) - H(y) = \frac{3}{4}\left[3 - \log_2 3\right] - \frac{3}{4}\left[2 - \log_2 3\right] - \frac{1}{2}$$

$$= \frac{9}{4} - \frac{6}{4} - \frac{1}{2} = \frac{9-6-2}{4} = +\frac{1}{4} = 0.25$$

$$H(y/x) = H(x,y) - H(x)$$

$$= \frac{3}{4}\left[3 - \log_2 3\right] - \frac{3}{8} - \frac{7}{8}\left[3 - \log_2 7\right]$$

Scanned with CamScanner

Theorem: [Independent bound on entropy]

Let $x_1, x_2, \ldots x_n$ be drawn according to $p(x_1, x_2 \ldots x_n)$, then

$$H(x_1, x_2 \ldots x_n) \leq \sum_{i=1}^{n} H(x_i)$$

If and only if, x_i are independent.

Proof: By Chain rule of entropy

$$H(x_1, x_2, \ldots x_n) = \sum_{i=1}^{n} H\left(\frac{x_i}{x_{i-1}, x_{i-2} \ldots x_1}\right)$$

$$\leq \sum_{i=1}^{n} H(x_i)$$

Since x_i are independent, then

$$H(X|X_2) = H(X_1)$$

Scanned with CamScanner

Log Sum inequality and its applications

Thm: For non-negative numbers $a_1, a_2, \ldots a_n$ and $b_1, b_2 \ldots b_n$

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{j=1}^{n} b_j}$$

Iff $\frac{a_i}{b_i} = \text{constant}$

proof: Without loss of generality $a_i > 0$ and $b_i > 0$.

The function $f(t) = t \log t$ is strictly convex,

Since $f''(t) = \frac{1}{t} \log e > 0$, $\forall\ t > 0$.

Hence by Jensen's inequality, we have

$$\sum \alpha_i f(t_i) \geq f\left(\sum \alpha_i t_i\right),$$

for $\alpha_i \geq 0$, $\sum \alpha_i = 1$, let $\alpha_i = \frac{b_i}{\sum_{j=1}^{n} b_j}$ and $t_i = \frac{a_i}{b_i}$

We obtain

$$\sum \frac{a_i}{\sum b_j} \log \frac{a_i}{b_i} \geq \sum \frac{a_i}{\sum b_j} \cdot \log \sum \frac{a_i}{\sum b_j}$$

which is the log sum inequality.

Corollary: Using log sum inequality, prove that

$$D(p \| q) \geq 0$$

proof: By defn, we have

$$D(p \| q) = \sum p(x) \log \frac{p(x)}{q(x)}$$

$$\geq \sum p(x) \log \frac{\sum p(x)}{\sum q(x)} = 1 \log 1 = 0 \quad \because \frac{p}{q} = c$$

Since both p and q are probability mass functions, $c = 1$, hence, we have

$$D(p \| q) = 0 \quad \text{iff} \quad p(x) = q(x),\ \forall\ x.$$

Scanned with CamScanner

Data processing Inequality

The data processing inequality can be used to show that no clever manipulation of the data can improve the inference that can be made from the data

Def^n Random variables X, Y, Z are said to form a Markov chain in that order $X \to Y \to Z$, if the conditional distribution of Z depends on Y and is conditionally independent of X.

i.e. X, Y, Z form a Markov chain $X \to Y \to Z$ if the joint probability mass function is

$$p(x,y,z) = p(x)\,p(y/x)\,p(z/y)$$

Consequences

i) $X \to Y \to Z$ iff X and Z are conditionally independent given Y. Markovity implies conditional independence because

$$p(x,z/y) = \frac{p(x,y,z)}{p(y)} = \frac{p(x,y)\,p(z/y)}{p(y)} = p(x/y)\,p(z/y)$$

This is the characterization of Markov chains that can be extended to define Markov fields, which are n-dimensional random processes in which the interior and exterior are independent given the values on the boundary.

2) $X \to Y \to Z$ implies that $Z \to Y \to X$. Thus, the condition is sometimes written $X \leftrightarrow Y \leftrightarrow Z$

3) If $Z = f(Y)$, then $X \to Y \to Z$.

Scanned with CamScanner

Sufficient Statistics

A family of a probability mass functions $\{f_\theta(x)\}$ indexed by θ, and let X be a sample from a distribution family. Let $T(X)$ be any statistic [μ or σ^2]. Then $\theta \to X \to T(X)$ and by data-processing inequality, we have

$$I(\theta; T(X)) \leq I(\theta; X)$$

for any distribution on θ, however, if equality holds, no information is lost.

A statistic $T(X)$ is called sufficient for θ, if it contains all the information in X about θ.

Defn: A function $T(X)$ is said to be a sufficient statistic relative to the family $\{f_\theta(x)\}$, if X is independent of θ given $T(X)$ for any distribution on θ [i.e. $\theta \to T(X) \to X$ form a Markov chain]

Ex Let $x_1, x_2 - - x_n$, $x_i \in \{0,1\}$ be an independent and identically distributed (i.i.d) sequence of coin tosses of a coin with unknown parameter $\theta = P_r(x_i = 1)$. Given n, the number of 1's is a sufficient statistic for θ.

i.e. $$T(x_1, x_2 - - x_n) = \sum_{i=1}^{n} x_i \qquad (TS.T)$$

ii. All sequences having that many 1's are equally likely and independent of the parameter θ, then

$$P_r(x_1, x_2 - - x_n) = (x_1, x_2 - - x_n) \Big/ \sum_{i=1}^{n} x_i = k$$

$$= \begin{cases} \dfrac{1}{\binom{n}{k}} & \text{if } \sum x_i = k \\ 0 & \text{otherwise} \end{cases}$$

Thus $\theta \to \sum x_i \to (x_1, x_2 - - x_n)$ form a Markov chain and T is a sufficient statistic for θ.

Scanned with CamScanner