An intuitive approach to Information Theory

When teaching information theory, certain entities (functions) such as entropy and mutual information are first defined in terms of the probabilistic parameters of the source and channel and then their practical relevance is demonstrated.

In this note, I start by stating quantities as symbols and then discuss their intuitive meaning, derive certain relationship between them based on the common sense and then, finally, express them in terms of probabilistic parameters. The latter being mainly necessary for computational purposes.

In most texts entropy is defined first and then the mutual information. I find mutual information more intuitively explainable. So, I start with the concept of mutual

information.

Mutual information quantifies the amount of information that an event (or a process) provides about another event (or process).

Take a process X with events $x \in X$. For example X can represent the weather and $x \in X$ can be "cold", "hot", "very cold", "cool", etc.

Another process Y can model the trend of clothing purchase by people. So $y \in Y$ can be "coat", "jacket", "pants", "shorts", etc.

The joint information between an event $x \in X$ and an event $y \in Y$, is denoted by $I(x;y)$ and is the amount of information the knowledge of y gives about x.

For example, knowing that people buy more coats than other type of apparel points to the possibility that the weather is cold and vice versa, i.e., the weather

being predicted to be cold, the vendors stock coats instead of other clothing items.

The amount of information x gives about y (or y gives about x), i.e., $I(x;y)$ logically should depend on how much y is dependant on x. The extreme case is when x and y are independent. In such a case, it is natural to expect $I(x;y)=0$.

Just a brief mention of probability here: When two events are independent their joint probability mass (or density) function can be written as $p(x,y)=p(x)p(y)$. But when x and y are not independent $p(x,y)=p(x)p(y|x)$ or $p(y)p(x|y)$. So, in a sense mutual information is a quantification of how $p(x,y)$ is different from $p(x)p(y)$. We will come back to this later.

What usually is considered as mutual inform is actually the average of $I(x;y)$ over all possible values of $x \in X$ and $y \in Y$ and is shown as:

$$I(X;Y) = \underset{x,y}{E}\{I(x;y)\}$$

and is called [average] mutual information and is a measure of the average amount of information that observing X provides about the process Y and vice versa.

Now, let's look at $I(X;X)$, i.e., the amount of information the observation of X gives about X!?

$I(X;X)$ is all you need to know (or like to know) about X. Having seen X, there is no uncertainty about X. So, $I(X;X)$ is the uncertainty about X (or the amount of information contained in X). It is given a specific symbol $H(X)$ and is called the entropy of X.

In order to be useful, we expect that, the mutual information between two processes to be positive (non-negative):

$$I(X;Y) \geq 0$$

This is to say that knowing something at worst can be useless (cannot be harmful)

Conditional Mutual information:

$I(X;Y|Z)$ is the mutual information between X and Y conditioned on Z, i.e., the average amount of information X provides about Y given that we have already observed Z.

Conditional entropy:

$H(X|Y)$ is the entropy of X conditioned on Y. That is, $H(X|Y)$ is the uncertainty about X given that we have observed Y.

$I(X;Y)$ is the information that Y gives about X. So, it is clear that it can be written as the difference between the uncertainty we have about X before and after observing Y, i.e.,

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

and since $I(X;Y) = I(Y;X)$ we can write

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

Unlike mutual information, in general, entropy is decreased (or stays unchanged with conditioning).

$$I(X;Y) = H(X) - H(X|Y) \geq 0$$

So

$$H(X) \geq H(X|Y)$$

or

$$H(X|Y) \leq H(X).$$

Now, let's consider a channel with input X and output Y

$H(X)$ is the uncertainty that we have about the input a priori and $H(X|Y)$ is the uncertainty about the input after observing the received signal. So $H(X)-H(X|Y)$ is the rate of information through the channel That is

$R = H(X) - H(X|Y) = I(X;Y)$ is the rate of information transfer for a given input X (a given probability distribution on the input). So, if we find the maximum of R, we have the capacity of the channel, i.e., the highest rate at which communication is possible over the channel:

$$C = \max_{p(x)} R = \max_{p(x)} I(X;Y)$$

Entropy of multiple events:

$H(X, Y)$ is the entropy of the pair (X, Y) that is, the uncertainty about (X, Y) prior to observing them.

Similarly $H(X_1, X_2, \cdots X_n)$ is the entropy of the vector of processes $(X_1, X_2, \cdots X_n)$.

It is intuitive to expect the following to hold

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

It says that the uncertainty about X and Y is the uncertainty about X plus the uncertainty about Y when we know X.

In general:

$$H(X_1, X_2, \cdots X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \cdots + H(X_n|X_1, \cdots X_{n-1})$$

or

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

This is called the <u>chain rule</u> of entropy.

Chain rule of mutual information:

$I(X; Y, Z)$ is the mutual information between the process X and the Compound process (Y, Z), i.e., it is the amount of information that observation of (Y, Z) gives about X or observation X reveals about (Y, Z).

It is intuitively clear that:

$$I(X; Y, Z) = I(X; Y) + I(X; Z|Y)$$

That is the information that (Y, Z) give about X is the sum of the information provided by Y plus what else remains to be revealed by Z give that we have observed Y.

We can similarly write:

$$I(X; Y, Z) = I(X; Z) + I(X; Y|Z).$$

Note the distinction between comma and semi-colon (;) in $I(X; Y, Z)$. It is important not to mix the two. $I(X; Y, Z)$ is quite different from $I(X, Y; Z)$.

The chain rule of mutual information, in general, can be expressed as:

$$I(X; Y_1, Y_2, \cdots Y_n) = I(X; Y_1) + I(X; Y_2 | Y_1) + \cdots$$

$$\cdots + I(X; Y_n | Y_1, \cdots Y_{n-1})$$

$$= \sum_{i=1}^{n} I(X; Y_i | Y_1, \cdots Y_{i-1})$$

Similarly:

$$I(X_1, \cdots X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_1, \cdots, X_{i-1})$$

In other notes, we will give the expressions for the above discussed entities in terms of the probabilistic model of source and channel and verify the relationships we derived (mainly accepted) intuitively.