

Basics of information theory

Today we live in the information age. The internet has become an integral part of our lives, making this, the third planet from the sun, a global village. People talking over the cellular phones is a common sight, sometimes even in cinema theatres. Movies can be rented in the form of a DVD disk. Email addresses and web addresses are common on business cards. Many people prefer to send emails and e-cards to their friends rather than the regular snail mail. Stock quotes can be checked over the mobile phone.

Information has become the key to success (it has always been a key to success, but in today's world it is *the* key). And behind all this information and its exchange lie the tiny 1's and 0's (the omnipresent bits) that hold information by merely the way they sit next to one another. Yet the information age that we live in today owes its existence, primarily, to a seminal paper published in 1948 that laid the foundation of the wonderful field of **Information Theory**—a theory initiated by one man, the American Electrical Engineer Claude E. Shannon, whose ideas

appeared in the article “The Mathematical Theory of Communication” in the *Bell System Technical Journal* (1948). In its broadest sense, information includes the content of any of the standard communication media, such as telegraphy, telephony, radio, or television, and the signals of electronic computers, servo-mechanism systems, and other data-processing devices. The theory is even applicable to the signals of the nerve networks of humans and other animals.

The chief concern of information theory is to discover mathematical laws governing systems designed to communicate or manipulate information. It sets up quantitative measures of information and of the capacity of various systems to transmit, store, and otherwise process information. Some of the problems treated are related to finding the best methods of using various available communication systems and the best methods for separating wanted information or signal, from extraneous information or noise. Another problem is the setting of upper bounds on the capacity of a given information-carrying medium (often called an information channel). While the results are chiefly of interest to communication engineers, some of the concepts have been adopted and found useful in such fields as psychology and linguistics.

Uncertainty and Information

Any information source, analog or digital, produces an output that is random in nature. If it were not random, i.e., the output were known exactly, there would be no need to transmit it! We live in an analog world and most sources are analog sources, for example, speech, temperature fluctuations etc. The discrete sources are man-made sources, for example, a source (say, a man) that generates a sequence of letters from a finite alphabet (typing his email).

Before we go on to develop a mathematical measure of information, let us develop an intuitive feel for it. Read the following sentences:

- (A) Tomorrow, the sun will rise from the East.
- (B) The phone will ring in the next one hour.
- (C) It will snow in Delhi this winter.

The three sentences carry different amounts of information. In fact, the first sentence hardly carries any information. Everybody knows that the sun rises in the East and the probability of this happening again is almost unity. Sentence (B) appears to carry more information than sentence (A). The phone may ring, or it may not. There is a finite probability that the phone will ring in the next one hour (unless the maintenance people are at work again!). The last sentence probably made you read it over twice. This is because it has never snowed in Delhi, and the probability of a snowfall is very low. It is interesting to note that the amount of information carried by the sentences listed above have something to do with the probability of occurrence of the events stated in the sentences. And we observe an inverse relationship. Sentence (A), which talks about an event which has a probability of occurrence very close to 1 carries almost no information. Sentence (C), which has a very low probability of occurrence, appears to carry a

information. Sentence (C), which has a very low probability of occurrence, appears to carry a lot of information (made us read it twice to be sure we got the information right!). The other interesting thing to note is that the length of the sentence has nothing to do with the amount of information it conveys. In fact, sentence (A) is the longest but carries the minimum information.

We will now develop a mathematical measure of information.

Definition 1.1 Consider a discrete random variable X with possible outcomes x_i , $i = 1, 2, \dots, n$.

The **Self-Information** of the event $X = x_i$ is defined as

$$I(x_i) = \log \left(\frac{1}{P(x_i)} \right) = -\log P(x_i) \quad (1.1)$$

We note that a high probability event conveys less information than a low probability event. For an event with $P(x) = 1$, $I(x) = 0$. Since a lower probability implies a higher degree of uncertainty (and *vice versa*), a random variable with a higher degree of uncertainty contains more information. We will use this correlation between uncertainty and level of information for

We would like to develop a usable measure of the information we get from observing the occurrence of an event having probability p . Our first reduction will be to ignore any particular features of the event, and only observe whether or not it happened. Thus we will think of an event as the observance of a symbol whose probability of occurring is p . we will thus be defining the information in terms of the probability p . A specific special case of interest is probabilities (i.e., real numbers between 0 and 1), We will

want our information measure $I(p)$ to have several properties:

1. Information is a non-negative quantity: $I(p) \geq 0$.
2. If an event has probability 1, we get no information from the occurrence of the event: $I(1) = 0$.
- 3- If two independent events occur (whose joint probability is the product of their individual probabilities), then the information we get from observing the events is the sum of the two information:

$$I(p_1 * p_2) = I(p_1) + I(p_2) : (\text{This is the critical property . . .})$$

- Summarizing: from the four properties,

$$1. I(p) \geq 0$$

$$2. I(p_1 * p_2) = I(p_1) + I(p_2)$$

$$3. I(p) \text{ is monotonic and continuous in } p$$

$$4. I(1) = 0$$

we can derive that

$$I(p) = \log_b(1/p) = -\log_b(p),$$

$$I(p^{n/m}) = \frac{n}{m} * I(p)$$

4. And thus, by continuity, we get, for $0 < p \leq 1$, and $a > 0$ a real number:

$$I(p^a) = a * I(p)$$

- From this, we can derive the nice property:

$$I(p) = -\log_b(p) = \log_b(1/p)$$

for some base b .

Some probability ideas

At various times in what follows, I may float between two notions of the probability of an event happening. The two general notions are:

1. A frequentist version of probability: In this version, we assume we have a set of possible events, each of which we assume occurs some number of times. Thus, if there are N distinct possible events ($x_1; x_2; \dots; x_N$); no two of which can occur simultaneously, and the events occur with frequencies ($n_1; n_2; \dots; n_N$), we say that the probability of event x_i is given by

$$P(x_i) = \frac{n_i}{\sum_{j=1}^N n_j}$$

This definition has the nice property that

$$\sum_{i=1}^N P(x_i) = 1$$

Example 1.1 Consider a binary source which tosses a fair coin and outputs a 1 if a head (H) appears and a 0 if a tail (T) appears. For this source, $P(1) = P(0) = 0.5$. The information content of each output from the source is

$$\begin{aligned} I(x_i) &= -\log_2 P(x_i) \\ &= -\log_2(0.5) = 1 \text{ bit} \end{aligned} \quad (1.2)$$

Indeed, we have to use only one bit to represent the output from this binary source (say, we use a 1 to represent H and a 0 to represent T).

Example 1.2 Consider a discrete, memoryless source (DMS) (source C) that outputs *two* bits at a time. This source comprises two binary sources (sources A and B) as mentioned in Example 1.1, each source contributing one bit. The two binary sources within the source C are independent. Intuitively, the information content of the aggregate source (source C) should be the *sum* of the information contained in the outputs of the two independent sources that constitute this source C . Let us look at the information content of the outputs of source C . There are four possible outcomes $\{00, 01, 10, 11\}$, each with a probability $P(C) = P(A)P(B) = (0.5)(0.5) = 0.25$, because the source A and B are independent. The information content of each output from the source C is

$$\begin{aligned} I(C) &= -\log_2 P(x_i) \\ &= -\log_2(0.25) = 2 \text{ bits} \end{aligned} \quad (1.4)$$

We have to use two bits to represent the output from this combined binary source.