Shannon entropy and mutual information

Or, how to play "guess who"

Emanuele Crosato and Richard G. Morris

November 16, 2023

Abstract

These notes are about one of the most fashionable, yet misinterpreted words in science: *information*. The purpose is to introduce Information Theory, the mathematical framework that allows us to unambiguously define and deal with information. We describe the two of the main quantities of the field, Shannon entropy and mutual information, which characterise information about a random variable in terms of uncertainty reduction. Intuitive everyday examples, such as the roll of a die, are used for explaining the main concepts. Biological application are also discussed.

1 A framework for information

In everyday language, the word *information* is often used loosely, letting the context clarify its meaning. Colloquial definitions of information may include:

- "Things that are or can be known about a given topic", from wikitionary.com;
- "Knowledge communicated or received concerning a particular fact or circumstance" from dictionary.com;
- "advice, clue, data, instruction, intelligence, knowledge, ..." from thesaurus.com.

In most scenarios, these definitions work perfectly fine, and there typically isn't much confusion about what is meant by information. For scientific purposes, however, a rigorous definition is needed. In mathematics, information is unambiguously defined as a measure of resolution of uncertainty or, using the language of Information Theory, a measure of reduction of entropy. Let's dig in the details.

As a first step, we need to construct a framework, and set some ground rules.

Observers and observable variables

In mathematics, information only makes sense if we can identify 1) an observable variable and 2) an observer who witnesses the realisation of the variable.

In other words, we can only talk about information if we can answer the questions *about what?* and *to whom?* This might sound pedantic, but that's because we want to highlight a very important point: information is not a property of an object, but rather a property of an observation. We are familiar with physical quantities, such as the energy of a system, its pressure, volume, *etc.* There is no such a thing as the information of a system! The correct framework for talking about information is: someone (or something) observes the realisation of a random variable and, in doing so, acquires a certain amount of information.

Let's consider a simple example of a person who is watching a die being rolled. The observable variable is the die, whose outcomes (or realisations) are 1, 2, 3, 4, 5 or 6. The person, of course, is the observer, who is unaware of the the outcome of the roll until he/she observes it, acquiring information.

A second essential ingredient are probability distributions:

Probability distributions

Information can only be quantified if we know the probabilities associated with the observations.

From the perspective of the observer, the observable is a so-called *random* variable: its realisations cover a set of possible values (the *range* of the variable) each of which is associated with a probability of occurrence. A random variable is commonly denoted using uppercase, say X. It is defined over a *set* of possible values, for which we use calligraphic typeface, in this case $\mathcal{X}$. The individual realisations (also called outcomes or events) that make up this set are then denoted using lowercase, so that we may write $x \in \mathcal{X}$. We assign a probability $p(x)$ to each such realisation, which is a number from 0 (no chance of happening) to 1 (certainty of happening). The distribution of probabilities over $\mathcal{X}$ must always be normalised, so that they 'sum to one', *i.e.*

$$\sum_{x \in \mathcal{X}} p(x) = 1. \tag{1}$$

If X is our die, then the range $\mathcal{X} = \{1, .., 6\}$ corresponds to the six faces. Assuming the die is a fair one, we have $p(x) = 1/6$ for all $x \in \mathcal{X}$.

2 Information content and Shannon entropy

We now have all the ingredients to define the information content associated to an observation.

Information content

When the probability distribution of a random variable X is known, the *information content* associated to each individual realisation x is defined as:

$$h(x) = \log_2 \frac{1}{p(x)} = -\log_2 p(x). \tag{2}$$

The information content is proportional to the 'surprise' $1/p(x)$ of observing the outcome — another name for the information content is the *surprisal*. Considering again the example of a die, we can see that, since all outcomes have the same probability, they all carry the same information content:

$$h(1) = \cdots = h(6) = -\log_2 \frac{1}{6} \approx 2.585 \text{ bits}. \tag{3}$$

However, when the die is biased (suppose $p(1) = \ldots = p(5) = 0.1$ while $p(6) = 0.5$) then different observations would carry different information content:

$$h(1) = \cdots = h(5) = -\log_2 0.1 \approx 3.322 \text{ bits}, \tag{4}$$

$$h(6) = -\log_2 0.5 = 1 \text{ bit}. \tag{5}$$

The biased die example highlights something very important:

Rare and common events

Rare events carry more information content than common events, i.e., the information content is inversely proportional the probability.

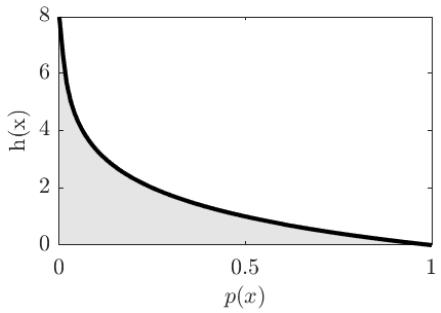

Figure 1: Information content as a function of probability.

This is an obvious property of Eq. (2), but let's spend a moment on it. A very common event, such as *not* winning the lottery, can be predicted with great accuracy without requiring the acquisition of any information, simply because of its high a priori probability to happen. Events that always happen, such as the sun rising in the morning, carry no information content at all to us, the observers. On the other hand, rare events (e.g., actually winning the lottery) are very hard to predict and thus their observation carry a lot of information content.

The example of the biased die also highlights another important idea in information theory: the difference between information and data.

Data is not information

The amount of symbols (e.g., bits) contained in a message quantifies the data. The amount of data doesn't always (almost never, actually) correspond to the amount of information.

Suppose that the outcome of rolling a die is communicated to the observer via a message containing a number 1 to 6. The content of the message, which we call the data, would be of one symbol, regardless of the outcome. However, as per Eqs. (4) and (5), the information content of the message would be different (i.e., lower for the outcome 6). In general, the relation between data and information depends on many aspects, including the mechanism used for encoding the data (see the Advanced box below) as well as the a priori knowledge of the receiver of the data. For the moment, it is just important to keep in mind that information and data are two different concepts.

Advanced: Optimal code

It can be shown that the information content of a message can be equal (or close to equal) to the amount of data if a so called 'entropy encoding' method, such as the Huffman code, is used. This is a variable length encoding in which symbols that are very frequent are encoded by short codes, while rare symbols are encoded by longer codes (as shorter ones become unavailable). Consider a message which contains only four symbols, 'a', 'b', 'c' and 'd', which appear with the following frequencies:

symbol	probability	code
a	0.42	0
b	0.37	10
c	0.16	110
d	0.05	111

Table 1: Huffman encoding for a message made of four symbols, given the frequencies of such symbols.

The Huffman code in Table 1 is an example of entropy encoding method, and is close to optimal. Note that the end of each code is easily identified by either a 0 or a 111 (no need for an extra end-of-symbol code).

The mechanism to construct such encoding is the following:

1. Set each symbol as a leaf of a tree. Build a binary tree by iteratively connecting the two nodes with the lowest probability. The probability of a new node is the sum of the probabilities of the two children.

2. Label each node starting from the root: Label the left edge with 0 and the right edge with 1 and repeat for all children.

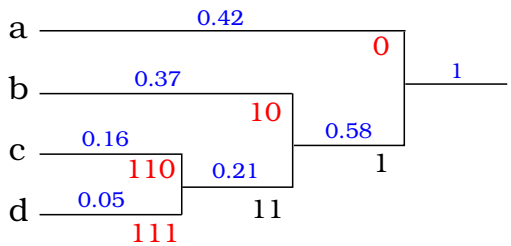

Figure 2: Huffman code construction and reading for the source in Table 1.

At this point, the reader might have a question: If information content and data aren't the same, why are they both measured in bits? This is a very good question indeed, although the answer may be disappointing. The actual unit of measure of information is not the bit but the Shannon: one Shannon is the information content of an event *that has probability 0.5*. However, the name bit is typically (mis)used. The bit is used for data and no assumption is made on the sequences of symbols in the message being equally likely.

The reader may also be wondering why the information content is the *logarithm* of the surprise, rather than simply the surprise. Intuitively, the reason for using the logarithm is that it allows to capture relative variations of the surprise at any scale. For instance, the information content carried by an observation that has a certain probability is always exactly 1 bit greater that the information content carried by an observation that has a half the probability, regardless of how large or small such probabilities are:

$$h(x) + 1 = h\left(\frac{x}{2}\right). \tag{6}$$

It is also important to note that

The base is not important

We don't need to use base 2. In fact, sometimes it is more convenient to use other bases. This means that information does not need to be measured in bits (or Shannon).

Any base is equally correct! In physics, for example, the natural logarithm (which base is the Euler's number, $e = 2.71828182\ldots$) is typically used. In this case, the unit of measure of the information content is the *nat*, which stands for 'natural unit of information'. More rarely (mostly in finance), base 10 is used and the unit of measure takes the name *Hart* (short for Hartley), also called *dit* (decimal digit).

Now that we know everything about the information content, we can introduce the main quantity in information theory: the *Shannon entropy*.

Shannon entropy

An average information content can be assigned to a whole random variable, given its probability distribution. This quantity is called the Shannon entropy:

$$H(X) = \sum_{x \in \mathcal{X}} p(x)h(x) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x). \tag{7}$$

The Shannon entropy (from Claude Shannon, who started the field of information theory in the late 1940s) represents the uncertainty associated with the entire random variable, rather than a single realisation of it. Think of it this way: The more oblivious the observer is on average in predicting the realisations of a random variable, the greater is the Shannon entropy associated to that variable.

Considering again the die example, we can see that the Shannon entropy is the highest when the die if fair, i.e., when it is hardest to predict the outcome of the outcome:

$$H(X) = -\log_2 \frac{1}{6} \approx 2.59 \text{ bits}. \tag{8}$$

Contrarily, the biased die has a lower Shannon entropy:

$$H(X) = -0.5 \log_2 0.5 - 5 \times 0.1 \log_2 0.1 \approx 2.16 \text{ bits}. \tag{9}$$

It is easy to show that the more biased the coin is, the lower the Shannon entropy. Fig. 3 shows the Shannon entropy associated to the die at different bias levels, from never 6 ($p(6) = 0$, $p(1) = \cdots = p(5) = 0.2$) to fair ($p(1) = \cdots = p(6) = 1/6$), to always 6 ($p(6) = 1$, $p(1) = \cdots = p(5) = 0$).

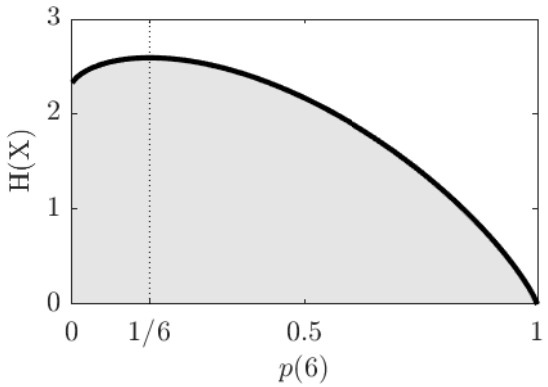

Figure 3: Entropy associated with a biased die over different amount of bias.

The Shannon entropy should not to be confused with the thermodynamic entropy, although the two quantities are intimately related. Thermodynamic entropy is a specific instance of Shannon entropy, applied to physical systems. More about thermodynamic entropy can be found in the lectures to come.

3 Information channels and mutual information

In the previous section we have introduced the information content and the Shannon entropy using a framework that involves an observable random variable and an observer. In most circumstances, however, the observer is not interested in the observable variable itself, but rather in another variable that statistically depends on it. Therefore, a more common framework is: Someone (or something) observes a realisation of a random variable and in doing so acquires a certain amount of information about a quantity of interest that is related.

When this is the case, a different information theoretic quantity, called the *pointwise mutual information*, is used, which involves not one, but two variables.

Pointwise mutual information

Given two random X and Y, the pointwise mutual information associated to a pair of outcomes (x, y) is:

$$\begin{aligned} i(x;y) &= h(x) + h(y) - h(x,y) \\ &= \log_2 \frac{p(x,y)}{p(x)p(y)}. \end{aligned} \tag{10}$$

The pointwise mutual information tells us how much information about the outcome of a variable is carried by observing the outcome of another variable.

Let's apply it to a simple scenario. Elwood lives in Massachusetts, where it rains 20% of the days. Elwood loves riding his unicycle, in fact he rides it 90% of the days. The join probability of the two variables X (it rains in Massachusetts) and Y (Elwood rides his unicycle) is also known:

$$p(x = 0, y = 0) = 0.01 \tag{11}$$

$$p(x = 1, y = 0) = 0.09 \tag{12}$$

$$p(x = 0, y = 1) = 0.79 \tag{13}$$

$$p(x = 1, y = 1) = 0.11 \tag{14}$$

Let's suppose we observe that it is not raining ($x = 0$) and Elwood is riding his unicycle ($y = 1$). The pointwise mutual information would be:

$$i(x = 0; y = 1) = \log_2 \frac{p(x = 0, y = 1)}{p(x = 0)p(y = 1)} = \log_2 \frac{0.79}{0.8 \times 0.9} \approx 0.13 \text{ bits}. \tag{15}$$

This means that the observer gets approximately 0.13 bits of information about not raining in Massachusetts from observing that Elwood is riding his unicycle. Vice versa is also true: the observer gets approximately 0.13 bits of information about Elwood riding his unicycle from observing that it isn't raining in Massachusetts. In fact, one important property of the mutual information is symmetry.

Just like the Shannon entropy is the average information content over all possible outcomes, the *mutual information* is the average pointwise mutual information.

Mutual information

The mutual information between two random varialbles X and Y is

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}. \end{aligned} \tag{16}$$

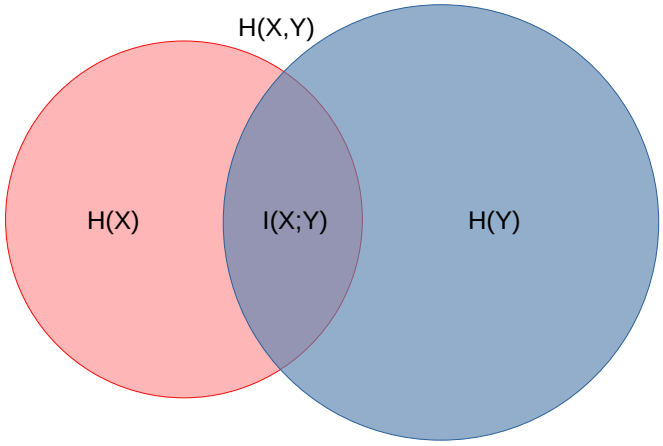

Figure 4: Venn diagram representation of the mutual information.

The mutual information between two variables is the amount of information about one variable that is on average obtained by observing the other[1]. One intuitive representation of the mutual information uses Venn diagrams (see Fig. 4): The two circles represent the uncertainty associated with the two variables X and Y, while the union of the two circles represents the joint Shannon entropy (the uncertainty associated with the two variables taken together). The intersection is the mutual information. Observing Y (i.e., removing the blue circle representing the uncertainty about Y) decreases the uncertainty about X (the red circle) by an amount that is the mutual information (the intersection). This diagram is useful for understanding a few properties of the mutual information. First, just like its pointwise version, the mutual information is symmetric: observing X instead of Y would decrease the uncertainty about Y by the same amount. Second, when the two variables X and Y are completely independent (i.e., there is no overlapping between the two circles) the mutual information is zero. This mean that no information about one variable is obtained by observing the other. Last, the maximum mutual information between X and Y is equal to the smallest of the Shannon entropies. This would happen, for example, if $H(Y)$ contained $H(X)$ entirely: the knowledge of Y would completely determine X.

Let's go back to Elwood and his unicycle. According to Eq. (16), the mutual information is

$$I(X,Y) = 0.01 \log_2 \frac{0.01}{0.8\times 0.1} + 0.09 \log_2 \frac{0.09}{0.2\times 0.1} + 0.79 \log_2 \frac{0.79}{0.8\times 0.9} + 0.11 \log_2 \frac{0.11}{0.2\times 0.9} \approx 0.1929 \text{ bits}. \tag{17}$$

This means that the weather in Massachusetts and Elwood's unicycling schedule are not statistically independent variables. Therefore, observing one of the two events provides us with some information about the other. On average, this information is around 0.1929 bits.

[1]The more advanced reader will notice that the mutual information is the Kullback–Leibler divergence between the joint distribution and the product of the marginals. Indeed, the mutual information can be interpreted as a measure of how different these two probability distributions are.

At this point, it should be understood that

Mutual information measures statistical dependence

The mutual information is a useful tool to estimate the statistical dependence between variables. The two most desired features of the mutual information compared to other measures of statistical dependence are that it is *non-parametric* and *model-free*.

The reader may wonder why we need the mutual information, since there are already many kinds of correlation (Pearson, Spearman, etc.). The reason is that these correlations do not test *any* statistical dependence between the variables, instead, they only test *specific* types of correlation (e.g., linear ones for Person correlation). In contrast, the mutual information does not make any assumption and fully captures statistical dependence. So, next time you need to check if there's a relation between two variables, do keep the mutual information in mind!

Here is an example of how to calculate the mutual information between two variables from experiments. For simplicity, we again consider dice: Two dice X and Y that somehow (the actual mechanism is not important here) stochastically depend on each other are observed thousands of times and the outcomes are summarised in Table 2.

$_{\mathbf{Y}}\backslash^{\mathbf{X}}$	**1**	**2**	**3**	**4**	**5**	**6**
1	651	7	2	0	18	101
2	15	344	25	17	864	22
3	1	12	213	789	6	1
4	0	1	32	28	0	0
5	0	0	2	1	0	0
6	0	0	0	0	0	0

Table 2: Occurrences of two statistically dependent dice, experimentally observed.

Notice that there is a clear dependence between the two dice as, for example, small outcomes of Y (i.e., 1 and 2) are mostly observed for either small or large outcomes of X (i.e., 1, 2, 5 and 6) but not for middle outcomes of X (i.e., 3 and 4). This dependence is clearly non-linear and therefore methods that assume a linear dependence, such as the Pearson correlation, would not perform well. At the contrary, the mutual information can capture the statistical dependence between the two dice. In order to compute the mutual information, we first need to estimate from the experiment the probabilities $p(X)$, $p(Y)$ and $p(X, Y)$. This can be easily from Table 2 and the result is summarised in Table 3.

$_{\mathbf{p(Y)}}\backslash^{\mathbf{p(X)}}$	**1**	**2**	**3**	**4**	**5**	**6**
1	0.2065	0.0022	0.0006	0	0.0057	0.0320
2	0.0048	0.1091	0.0079	0.0054	0.2741	0.0070
3	0.0003	0.0038	0.0676	0.2503	0.0019	0.0003
4	0	0	0.0006	0.0003	0	0
5	0	0	2	1	0	0
6	0	0	0	0	0	0

	p(X)	**p(Y)**
1	0.2116	0.2471
2	0.1155	0.4083
3	0.0869	0.3242
4	0.2649	0.0194
5	0.2817	0.0010
6	0.0393	0

Table 3: Probabilities estimated from the experiment.

Applying Eq. 16 using the probabilities in Table 3 we obtain a mutual information $I(X;Y) \approx 1.318$ bits.

More in general, the mutual information is often used in the context of *communication channels*.

Mutual information and communication channels

Let's consider two interacting systems, whose statistical behaviour is modelled using two random variables. The mutual information between these two variables quantifies the minimum capacity of a communication channel between the two systems that can explain the observed interaction.

Consider two systems that interact with each other. These can be anything, e.g., two computers in a network,

two animals crossing the same spot, two cells in your body, etc. We can imagine a channel between the two systems through which information needed for their interaction is communicated. Two important things need to be emphasized. First, this channel needs not to be an actual connection (e.g., a wire or bluetooth), instead, it can be an abstract medium relating the signals from a systems to the response of the other. In fact, the communication channel abstraction is typically used when two systems interact via an unknown mechanism, whose functioning we don't fully understand, although we do have empirical understanding of the relationship between the systems' signals and responses. Second, a very important, we are not focussing on the data (of whatever sort) that is crossing the channel, we are interested in information as reduction of uncertainty about the behaviour of a system given the behaviour of the other. In this setup, the mutual information takes the role of the channel capacity: Whatever interaction mechanism is between the two system, it needs to be complex enough to resolve a minimum amount of uncertainty given by the mutual information, otherwise the dependence between the two system cannot be explained.

A recent approach to understanding interactions between cells or cells' components uses the communication channel approach. You can check the details of all individual studies, but here is a schematic example that capture their essence:

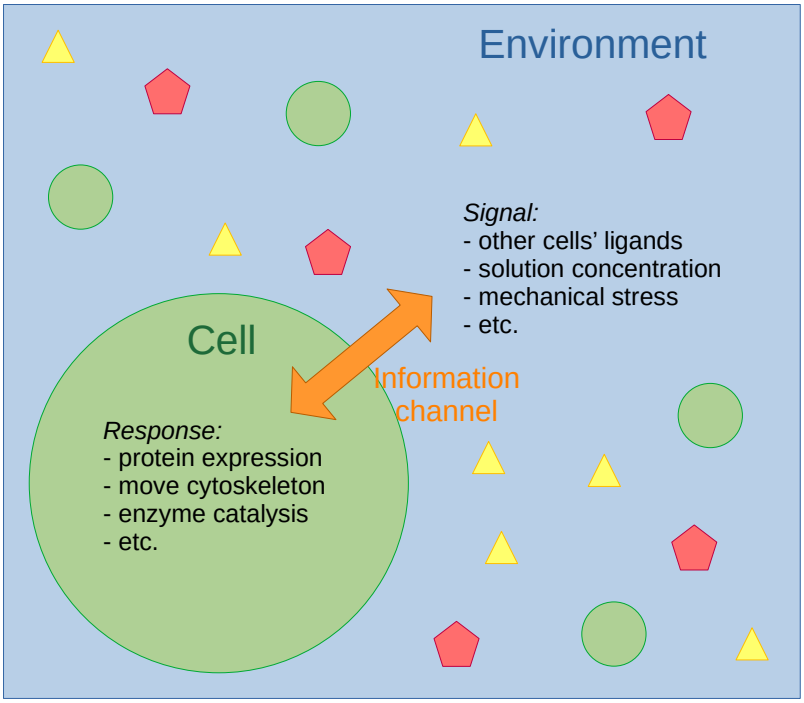

Figure 5: Cartoon of an information channel between a cell and its environment.

It is well known that the internal behaviour of a cell depends on its surrounding. For example, a cell can express some proteins instead of others depending on the concentration of the solutions it is submerged within, or it can remodel its cytoskeleton after sensing other cells or in response of mechanical stresses (see Fig. 5). The mechanisms thought which information is passed from the environment to the inside of a cell are typically complicated, involving complex sequences of chemical reactions. Many of these mechanisms are not fully understood yet and are currently being investigated. However, the question of *how much* information is conveyed between a cell and its surroundings is one that can in principle be answered using information theory. In fact, this is the equivalent of asking what is the capacity of the information channel between the cell and the environment.

If we know the probability distribution of a signal (e.g., the rate of binding with some ligands), the probability distribution of a response (e.g., the rate of a protein expression) and the joint distribution of signal and response (e.g., the probability of the cell binding with the ligand *end* expressing the protein), then we can ask the question "What it the maximum mutual information between the response and the signal?" or equivalently "What is the capacity of the communication channel between the response and the signal?".