Information Theory: Fundamental Equations

Jacob L. Fine

Jun 28th, 2024

Introduction

Molecular biologists often talk about 'the flow of biological information' and use words/phrases such as 'codes', 'decoding', 'signal processing', 'channels', and 'complexity' to describe biomolecules, organisms, and their properties. Unfortunately, they often do so without regard for the formal field of study that is information theory, which defines and characterizes the nature of information (and the above terms in particular) in a mathematically rigorous manner. Information theory is therefore well-suited for the study of biological systems. In this exercise, we will introduce key concepts of information theory and show their application in biological sequence analysis. Much of this is based on Claude Shannon's landmark paper 'A Mathematical Theory of Communication' (1948).

As will be seen, much of the foundational equations of information theory can be viewed as a natural extentions of mathematical statistics to describe probability distributions in terms of the information they contain within themselves, and with respect to other probability distributions.

Information content

Suppose we have a string s composed of a finite collection of symbols, $x_1, x_2, ..., x_n$. We can think of the event 'observing symbol x in some position in the string' as the outcome of a random variable X, such that the probability of each event x is given by the probability mass function $p(X)$. The string of interest might be a sequence of nucleotides, i.e., with the alphabet A, T, C, and G. Building on this statistics foundation, we can now introduce the concept of information, in relation to random variables.

Suppose we now wish to describe how much 'information' is contained in each outcome of the random variable X. Let us call this $I(X)$. We would like the values of $I(X)$ to be in some 'units of information' (such as bits), which we desire to be inversely proportional to the probablity of each event. This would enable us to use less bits of information to encode events that occur more frequently, which could be desirable for transmitting signals. It is also useful to think of $I(X)$ in relation to a hypothetical 'communication channel', where we

are transmitting the identity of each symbol to some reciever, and $I(X)$ is the information associated with transmitting or receiving each event.

To illustrate the notion of information content, let us consider a DNA sequence, consisting of four possible bases: A, T, C, and G, and let the random variable X denote the probability that each base is observed in some position of the sequence, such that each base is equally probable. Our pmf is therefore uniform as each event has the probability $p(x) = \frac{1}{4}$. It is convenient to represent the events in bits, where the four possible outcomes can be specified by two bits such that A = 00, T = 01, C = 10, and G = 11. Since our pmf is uniform, we require no additional or less information to encode each event other than the identity of the event itself, thus, the information content of each event is 2 bits. Here, $I(X)$ has the convenient interpretation as the number of bits needed to code each event in our pmf (though this interpretation is not always the case). Therefore,

$$I(X = A) = I(X = T) = I(X = C) = I(X = G) = 2 \text{ bits}$$

If our pmf was non-uniform, however, then each event may be associated with a greater or less amount of suprise. In such cases, we would desire that our 'communication channel' uses more bits for less frequent and therefore more suprising events, and vice versa. We therefore define an expression for $I(X)$ which generalizes to non-uniform pmfs, where

$$I(X = x) = -log_2[p(x)]$$

In our case with a uniform pmf of DNA bases, we obtain that $I(X = x) = -log_2(\frac{1}{4}) = 2$ which is consistent with the above. But in cases when the denominator is not a power of 2, the interpretation of $I(X)$ as the number of bits needed to encode each event is less intuitive. Thus, it is more useful to think of it as an index of suprisal associated with each event, that happens to be measured in bits. For instance, in the case of a fair die, the probability of each event is $p(X = x) = \frac{1}{6}$, so the information content of each event is $I(X = x) = -log_2(\frac{1}{6}) = 2.58$ bits.

Shannon entropy

In the case of a non-uniform pmf, $I(X)$ clearly differs for each event, so it would be convenient to define 'expected information', which we will define as the expected value of the $I(X)$ function, $E[I(X)]$. To find its closed-form expression, we consider the law of the unconscious statistician (LOTUS) for a discrete random variable, which is a convenient way to derive the expected value of a transformation of a random variable

$$E[g(X)] = \sum_x g(x)p(x)$$

Here, $I(X)$ was just the transformation of our random variable's pmf $p(X = x)$ such that $g(X) = I(X) = -log_2[p(x)]$. By the LOTUS, we have

$$E[I(X)] = \sum_x -log_2[p(x)]p(x)$$

The quantity $E[I(X)]$ is referred to as Shannon entropy. It is often written as $H(X)$, or $H(p)$ to specify our pmf p, and we use p_i instead of $p(x)$ for convenience. Therefore,

$$H(X) = H(p) = -\sum_{p_i} p_i log_2(p_i)$$

The Shannon entropy of a discrete random variable is the expected information of each event represented by the random variable. This enables us to conveniently talk about the information content of a non-uniform pmf. In the case of a uniform pmf with $p_i = \frac{1}{n}$, it can easily be shown that the Shannon entropy of the random variable is equivalent to the information content of every outcome of the random variable. In such cases, we expect every event to have the same information content. Furthermore, while there are various definitions of complexity, in this context, Shannon entropy is often used to define complexity.

Let us consider an example with a non-uniform pmf representing a string of DNA bases, where $p(X = x) = \{0.2, 0.1, 0.4, 0.3\}$ for bases A, T, C, G, respectively. The Shannon entropy of the pmf is therefore

$$H(X) = -0.2 * log_2(0.2) - 0.1 * log_2(0.1)$$

$$-0.4 * log_2(0.4) - 0.3 * log_2(0.3) = 1.84 \text{ bits}$$

Notice how this is less than the uniform case. To make sense of this, we can reason that since two bases, C and G, had sufficiently greater probabilities, there is less suprise associated with this DNA sequence. In the real biological world, this may be due to a CpG island, which is less 'complex' and thus has lower expected information.

Cross-entropy

Now that we have introduced Shannon entropy, it is useful to introduce the concept of cross-entropy, which deals with cases where we may wish to 'predict' the pmf of some random variable, and compare it with some 'ground truth' distribution. Let us call these two distributions q and p, respectively. The equation for cross-entropy is almost identical as Shannon entropy, with the expection of using estimated rather than true probabilities in the argument inside the logarithm. Cross entropy is therefore defined as

$$H[p(x), q(x)] = -\sum_{x} p(x) log_2[q(x)]$$

We can usually write this without explictly mentioning x, as

$$H(p, q) = -\sum p_i log_2(q_i)$$

We can think of cross-entropy as the expected information of a distribution of symbols with pmf p but using a different probability distribution, q, as a 'reference'.

Kullback-Leilber divergence

Suppose now we wish to know how much 'informational difference' there is if we use q as our distribution rather than p. This simply the difference between cross entropy and Shannon entropy, which is

$$H(p,q) - H(p) = -\sum p_i log_2(q_i) - (-\sum p_i log_2(p_i)) = \sum p_i log_2\Big[\frac{p_i}{q_i}\Big]$$

This quantity is referred to as the Kullback-Leilber divergence (KL-divergence), defined as

$$D_{KL}(p||q) = \sum p_i log_2\Big[\frac{p_i}{q_i}\Big]$$

We can think of the KL-divergence as the difference in expected information (measured in bits) if we assume our distribution is q, when it is really p. The KL-divergence between two distributions is effectively an information-centric distance measure. For instance, we might want to know how much our DNA sequence deviates from a 'uniform null distribution' where $p(X) = 0.25$ for all bases. The KL-divergence between the previous distribution $p(X) = \{0.2, 0.1, 0.4, 0.3\}$ and $q(X) = \{0.25, 0.25, 0.25, 0.25\}$ is therefore

$$D_{KL}(p||q) = -0.2 * log_2(0.2/0.25) - 0.1 * log_2(0.1/0.25)$$

$$-0.4 * log_2(0.4/0.25) - 0.3 * log_2(0.3/0.25) = \text{-0.15 bits}$$

We therefore require 0.15 bits less of information to encode the distribution p relative to the background q.

Mutual information

We often would like to know whether two events are independent or not. If two events are independent, then their joint probability is simply the product of their marginal probabilities, so

$$p_{(X,Y)}(x,y) = p_X(x)p_Y(y) \iff X \perp Y$$

Suppose we now have two strings of symbols, whose emissions are specified by the random variables X and Y with pmfs $p_X(x)$ and $p_Y(y)$, with a joint pmf $P_{(X,Y)}(x,y)$.

We introduce 'mutual information' $I(X,Y)$ as the information 'contained' in one random variable about another. We desire this quantity to be zero when the variables are independent, in which case, the variables posses no information about each other. It can therefore be seen that a convenient way to express this would be the informational distance based on the true joint pmf evaluated at each pair of outcomes X, Y compared to the 'expected' joint probabilities assuming the events are independent, i.e., $p(X = x)p(Y = y)$. Mutual information $I(X,Y)$ is therefore defined as the KL-divergence between the true joint probability distribution and the expected joint distribution under indepdence:

$$I(X,Y) = D_{KL}\Big(p_{(X,Y)}(x,y)||p_X(x) \otimes p_Y(y)\Big)$$

In the above, $p_X(x) \otimes p_Y(y)$ is the outer product of the vectors $p_X(x)$ and $p_Y(y)$, such that a value $p(X = x)p(Y = y)$ is assigned for each pair of symbols x, y in the domain. Mutual information therefore becomes

$$I(X,Y) = \sum_y \sum_x p_{(X,Y)}(x,y) * log_2 \left[\frac{p_{(X,Y)}(x,y)}{p_X(x)p_Y(y)} \right]$$

As desired, $X \perp Y$ then we can see that $I(X,Y) = 0$ since both the numerator and denominator will be equivalent in the argument of the logarithm.

Joint entropy

When dealing with a pair of random variables with a joint pmf $p_{(X,Y)}(x,y)$ the entropy of the joint pmf, or the joint entropy with respect to X and Y is simply based on Shannon entropy but using the joint pmf:

$$H(X,Y) = -\sum_{x,y} p(x,y) log[(p(x,y)]$$

This gives an information-centric basis of joint probability.

Conditional entropy

Continuing to extend our mathematical statistics framework into information theory, we can also build up the notion of conditional entropy from conditional probability. This quantity would represent the information gained by observing X once we know Y.

We first consider the entropy of X conditioned on some value of $Y = y$, as

$$H(X|Y = y) = -\sum_x p(x|y) log_2[p(x|y)]$$

We then sum over all possible values of y to obtain the full conditional entropy $H(X|Y)$

$$\implies H(X|Y) = -\sum_y p(y) \sum_x p(x|y) log_2[p(x|y)]$$

By Bayes' rule we have that $p(y)p(x|y) = p(x,y)$, which allows us to write that

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

or equivalently

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

We can also rearrange the above to obtain

$$H(X|Y) = -\sum_{x,y} p(x,y) log_2\Big[p(x,y)\Big] + \sum_{x,y} p(x,y) log_2\Big[p(y)\Big]$$

Since $\sum_x p(x,y) = p(y)$, we have

$$\iff H(X|Y) = -\sum_{x,y} p(x,y) log_2\Big[p(x,y)\Big] + \sum_{y} p(y) log_2\Big[p(y)\Big]$$

$$\iff H(X|Y) = H(X,Y) - H(Y)$$

which may be convenient.

Conclusion

We have gone through some of the basic equations of information theory, showing how fundamental concepts of mathematical statistics can be expressed in terms of information. It should be clearler how information theory serves to unify mathematical statistics with theories pertaining strings and their constituent symbols. Our example with DNA bases underscores the role information theory ought to play in molecular biology's central dogma, which is all about biological sequences and their properties. Though we can in principle extend these equations to any domain of science with well-defined symbols with known probability distributions, to model their informational properties.