Information Theory

Elif Yılmaz

What is information theory?

Information theory examines the quantification, storage, and communication of information. It was firstly asserted by Claude Shannon in 1948 to look for fundamental limits on signal processing and communication operations such as data compression in the article "A Mathematical Theory of Communication".

Let's begin with considering a discrete random variable x and asking how much information is received when we observe a specific value for this variable.
The measure of information content depends on the probability distribution p(x), and we therefore look for a quantity h(x) that is a monotonic function of the probability p(x) and that expresses the information content.
For the form of h(·);

- if we have two events x and y that are unrelated, then the information gain from observing both of them should be the sum of the information gained from each of them separately, so that h(x, y) = h(x) + h(y) and also two unrelated events will be statistically independent so p(x, y) = p(x)p(y).
- By using these two equations, we get $h(x) = -log_2 p(x)$ where the negative sign provides that information is positive or zero.

* Note that low probability events x correspond to high information content.

Entropy

To find the average amount of information, take the expectation of $h(x) = -log_2 p(x)$ with respect to the distribution p(x). Therefore, we get

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

and H(x) is called the entropy of the random variable x.

Examples

Consider a random variable x having 8 possible states, each of which is equally likely. In order to communicate the value of x to a receiver, we would need to transmit a message of length 3 bits. Notice that the entropy of this variable is given by

$$H(x) = -8(\frac{1}{8}\log_2\frac{1}{8}) = 3 \ bits$$

Now consider an example (Cover and Thomas, 1991) of a variable having 8 possible states $\{a, b, c, d, e, f, g, h\}$ for which the respective probabilities are given by $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$. The entropy in this case is given by

$$H(x) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{4}{64}\log_2\frac{1}{64} = 2 \ bits$$

* We see that the nonuniform distribution has a smaller entropy than the uniform one.

Alternative view of entropy

Consider a set of N identical objects that are to be divided among a set of bins, such that there are n_i objects in the i^{th} bin. Keep in mind the number of different ways of allocating the objects to the bins.
There are N ways to choose the first object, (N - 1) ways to choose the second object, and so on, therefore there are N! ways to allocate all N objects to the bins. However, we don't wish to distinguish between rearrangements of objects within each bin. In the i^{th} bin there are $n_i!$ ways of reordering the objects, and so the total number of ways of allocating the N objects to the bins is given by

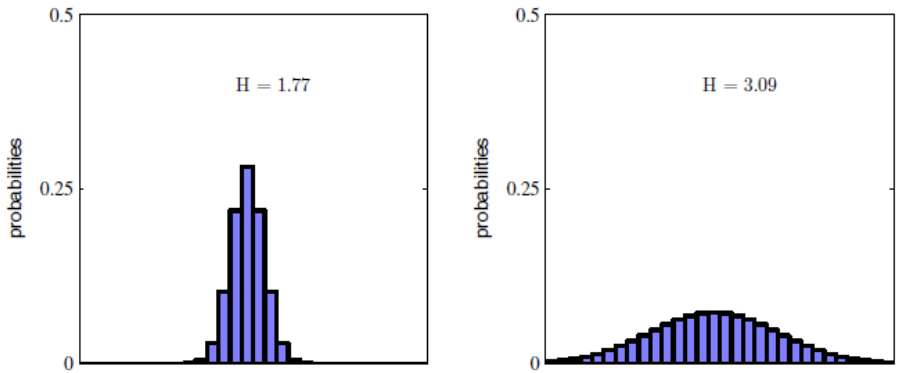
$$W = \frac{N!}{\prod_i n_i}$$

Then, the entropy is $H = \frac{1}{N} \ln W = \frac{1}{N} \ln N! - \frac{1}{N} \sum_i \ln n_i!$.

$$H = -\lim_{N\to\infty} \sum_i (\frac{n_i}{N}) \ln(\frac{n_i}{N}) = -\sum_i p_i \ln p_i$$

by taking limit as N goes to infinity and using Stirling's approximation ($\ln N! \simeq N \ln N - N$) and $p_i = \lim_{N\to\infty} \frac{n_i}{N}$, which is probability of an object being assigned to i^{th} bin.
We can interpret the bins as the states xi of a discrete random variable X, where $p(X = x_i) = p_i$. The entropy of the random variable X is then

$$H[p] = -\sum_i p(x_i) \ln p(x_i)$$



Because $0 \leq p_i \leq 1$, the entropy is nonnegative, and it will equal its minimum value of 0 when one of the $p_i = 1$ and all other $p_{j \neq i} = 0$. The maximum entropy configuration can be found by maximizing H using a Lagrange multiplier to enforce the normalization constraint on the probabilities. Thus we maximize

$$\tilde{H} = -\sum_i p(x_i) \ln p(x_i) + \lambda(\sum_i p(x_i) - 1)$$

from which we find that all of the $p(x_i)$ are equal and are given by $p(x_i) = 1/M$ where M is the total number of states x_i. The corresponding value of the entropy is then $H = \ln M$.

To confirm that the stationary point is a maximum, evaluate the second derivative of the entropy;

$$\frac{\partial \tilde{H}}{\partial p(x_i)\partial p(x_j)} = -I_{ij}\frac{1}{p_i}$$

where I_{ij} are elements of identity matrix.
For continuous random variables x, we can write entropy as by dividing x into bins of width Δ and assuming p(x) is continuous;

$$H_\Delta = -\sum_i p(x_i)\Delta \ln(p(x_i)\Delta) = -\sum_i p(x_i)\Delta \ln p(x_i) - \ln \Delta$$

where there must exist a value x_i for each such bin such that $\int_{i\Delta}^{(i+1)\Delta} p(x)dx = p(x_i)\Delta$ and $\sum_i p(x_i)\Delta = 1$.
* H_Δ has a discrete distribution.

Differential Entropy

Let's discuss $H_\Delta = -\sum_i p(x_i)\Delta \ln p(x_i) - \ln \Delta$. By omitting the second term $-\ln \Delta$ with the limit $\Delta \to 0$,

$$\lim_{\Delta \to 0} \{\sum_i p(x_i)\Delta \ln p(x_i)\} = -\int p(x) \ln p(x) dx$$

where the quantity on the right-hand side is called the differential entropy. So,

$$H[x] = -\int p(x) \ln p(x) dx$$

Let us now consider the maximum entropy configuration for a continuous variable. In order for this maximum to be well defined, it will be necessary to constrain the first and second moments of p(x) as well as preserving the normalization constraint. We therefore maximize the differential entropy with three constraints

$$\int_{-\infty}^{\infty} p(x)dx = 1 \tag{1}$$

$$\int_{-\infty}^{\infty} xp(x)dx = \mu \tag{2}$$

$$\int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx = \sigma^2 \tag{3}$$

We can perform the constrained maximization by using Lagrange multipliers so we maximize following with respect to p(x)

$$-\int_{-\infty}^{\infty} p(x) \ln p(x) dx + \lambda_1 \left(\int_{-\infty}^{\infty} p(x) dx - 1 \right)$$

$$+\lambda_2 \left(\int_{-\infty}^{\infty} x p(x) dx - \mu \right) + \lambda_3 \left(\int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx - \sigma^2 \right)$$

Using the calculus of variations, we set the derivative of this functional to zero giving

$$p(x) = exp\{-1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2\}$$

The Lagrange multipliers can be found by back substitution of this result into the three constraint equations, leading finally to the result

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$$

and so the distribution that maximizes the differential entropy is the Gaussian.
If we evaluate the differential entropy of the Gaussian, we obtain

$$H[x] = \frac{1}{2}\{1 + \ln(2\pi\sigma^2)\}$$

* This result also shows that the differential entropy, unlike the discrete entropy, can be negative because $H(x) < 0$ for $\sigma^2 < 1/(2\pi e)$.

Suppose we have a joint distribution p(x, y) from which we draw pairs of values of x and y. Then, the conditional entropy of y given x can be written as

$$H[y|x] = -\int\int p(y,x) \ln p(y|x) dydx$$

It is easily seen, using the product rule, that the conditional entropy satisfies the relation

$$H[x,y] = H[y|x] + H[y]$$

where H[x, y] is the differential entropy of p(x, y) and H[x] is the differential entropy of the marginal distribution p(x).

Relative entropy

Consider some unknown distribution p(x), and suppose that we have modelled this using an approximating distribution q(x). If we use q(x) to construct a coding scheme for the purpose of transmitting values of x to a receiver, then the average additional amount of information required to specify the value of x as a result of using q(x) instead of the true distribution p(x) is given by

$$KL(p||q) = -\int p(x) \ln q(x) dx - (-\int p(x) \ln p(x) dx)$$

$$= -\int p(x) \ln(\frac{q(x)}{p(x)}) dx$$

This is known as the relative entropy or Kullback-Leibler divergence (Kullback and Leibler, 1951) between the distributions p(x) and q(x).
* Note that it is not a symmetrical quantity, that is to say $KL(p||q) \not\equiv KL(q||p)$.

Mutual Information

Now consider the joint distribution between two sets of variables x and y given by p(x, y). If the sets of variables are independent, then their joint distribution will factorize into the product of their marginals p(x, y) = p(x)p(y).
If the variables are not independent, we can gain some idea of whether they are 'close' to being independent by considering the Kullback-Leibler divergence between the joint distribution and the product of the marginals, given by

$$I[x, y] \equiv KL(p(x, y)||p(x)p(y))$$

$$= -\int\int p(x, y) \ln(\frac{p(x)p(y)}{p(x, y)})dxdy$$

which is called the mutual information between the variables x and y.