Shannon Entropy and Kolmogorov Complexity

Ranitha Mataraarachchi

July 31, 2023

Abstract

Shannon entropy, named after mathematician Claude Shannon, was introduced in his 1948 paper "A Mathematical Theory of Communication"[1]. The concept of Shannon entropy has found diverse applications across various fields, such as computer science, data analysis, cryptography, and machine learning. Kolmogorov complexity, named after mathematician Andrey Kolmogorov, is a concept from algorithmic information theory. In this short article a brief introduction to both these concepts are given. Finally, the relationship between Shannon entropy and expected Kolmogorov complexity is explained.

1 Quantifying Information

First, we need to determine how to quantify information since Shannon entropy has its roots in Information theory. The information of an event is directly related the randomness of that event. If the event is very likely to happen, then the information content of that event is almost 0. However, if the event is less likely to happen, then the information of that event is very high. Therefore, the information content of an event indicates how random the event is. Nyquist[1924], Hartley[1928], and Shannon[1948] proposed the following equation to determine the information content present in an event E [2].

$$I(E) = log\left(\frac{1}{P(E)}\right)$$

where $P(E)$ is the probability of the event E, and $I(E)$ is the amount of information content in the event E. The base of the logarithm can be any value. However, Shannon has proposed base as 2 so that the unit of information is binary digits or simply, bits [2]. We will use this base throughout this article.

2 Shannon Entropy

In his paper, "The mathematical theory of communication"[1], Shannon proposed a measure of information in a distribution. For clarity, let us interpret this as the average information of a source emitting symbols. Assume that

a source is emitting a set of symbols $X = \{x_1, x_2, \ldots, x_N\}$ with probabilities $\{P_1, P_2, \ldots, P_N\}$, respectively. Then, the average information $H(X)$ of the source can be achieved by averaging the information content of symbols over their probability distribution. Mathematically, this can be represented by,

$$H(X) = \sum_{i=1}^{N} P_i \, I(x_i) = \sum_{i=1}^{N} P_i \, log\left(\frac{1}{P_i}\right).$$

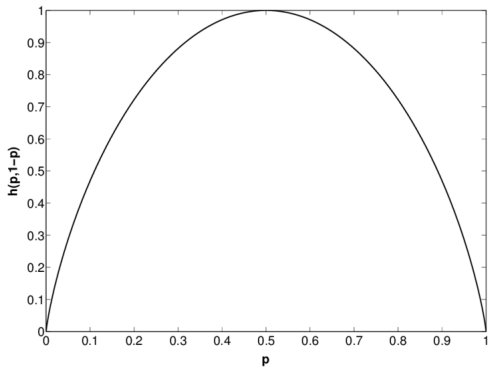
This measure of average information $H(X)$ is referred to as the Shannon entropy of the source. Entropy is essentially a functional map from a random variable distribution $\{P_1, P_2, \ldots, P_N\}$ to a real number $H(X)$.

Binary Entropy Function and maximum entropy

When only two symbols $\{x_1, x_2\}$ are present, their probabilities can be represented as $\{p, 1-p\}$. Thus, the entropy function $H(X)$ can be now represented as,

$$H(X) = p \, log(1/p) + (1-p) \, log(1/(1-p)).$$

If we plot this we get,



It clearly visible that the maximum entropy occurs when the probability is $0.5 = 1/2$. That is, when the events are equally likely.

In fact, for each N, $H(X)$ achieves its unique maximum for the uniform distribution $P_i = 1/N$. Therefore, an upper bound for entropy can be calculated as,

$$H(X) \leq \sum_{i=1}^{N} P_i \, log\left(\frac{1}{P_i}\right) = \sum_{i=1}^{N} \frac{1}{N} \, log\left(\frac{1}{(1/N)}\right) = log(N).$$

This shows that regardless of the distribution of P_i, entropy $H(X)$ never exceeds $log(N)$. It achieves this maximum only when the distribution P_i is a uniform distribution.

An extension of Shannon entropy is the source coding theorem (noiseless coding theorem) which was also proposed by Shannon. This theorem states that the average length of a code word $\bar{l}$ is lower bounded by the entropy $H(X)$ of the source, i.e.,

$$\bar{l} \geq H(x).$$

Additionally, it states that average length can be brought within one bit of the entropy, i.e.,

$$H(X) + 1 > \bar{l} \geq H(X).$$

3 Kolmogorov Complexity

The field now generally known as Kolomogorov complexity was discovered independently by R.N. Solomonoff [3], A.N. Kolomogorov [4] and G. Chaitin [5]. It is also called as, *algorithmic complexity, algorithmic information, algorithmic entropy, Solomonoff-Kolmogorov-Chaitin complexity, descriptional complexity, shortest program length, algorithmic randomness*[6].

In a broader sense, Kolmogorov complexity is the amount of information needed to uniquely describe a digital object. By definition, Kolmogorov complexity $K(x)$ of a digital object x is the length of the shortest possible binary description of x. We can think of this *digital object* as a finite binary string (a string of bits) for our convenience. Therefore, Kolmogorov complexity of a string of bits is the shortest computer program that prints the string of bits and stops running.

To understand this concept further, let us revisit a thought experiment brought forward by Solomonoff. Suppose a scientist is tracking the outcome of an experiment he is conducting. He records his observations with 1s and 0s and keeps track of all his observations as a binary string. The scientist tries to model his observations using an algorithm. By doing so, he can predict the future observations. Out of the variety of suitable algorithms that he develops, his motivation is to come up with the smallest algorithms possible. That is, the algorithm with the least amount of bits.

For clarity, let us look at some examples of his observations that he gets and try to write an algorithm to produce each. We will be using *python* but it should be noted that any computer language can be used here instead.

Observation 1: $'111111111111'$
Algorithm 1:
```
'1'*12
```

Observation 2: $'101010101010'$
Algorithm 2:
```
'10'*6
```

Observation 3: $'101100111000'$
Algorithm 3:

```
a=''
for i in range(4):
   a=a+'1'*i+'0'*i
```

Observation 4: $'111001010110'$
Algorithm 4:

```
'111001010110'
```

It should be noted that, for any two programming languages, the resulting algorithm lengths differ at most by a constant not depending on the observation.

For the first three observations there was a convenient way to describe the observation and to predict the future observations. We could have printed the whole string itself, but there was a more efficient way to achieve this result by doing a computation. On the contrary, the last observation has no easy algorithm due to its randomness (unpredictability). Hence, the algorithm must print all the digits in the string rather than doing a computation. For random data the most compact way for the scientist to communicate his or her observations is to publish them in their entirety [6]. Therefore, random strings require the longest algorithms. This in a way is a measure of randomness. More specifically, the *Kolmogorov randomness*. If a string has a length n, and if there exists an algorithm which is shorter than n that can compute that string, then can call this string to be *nonrandom*. If for another string the algorithm cannot be made shorter than n, then we say that string is *random*. A random string is more *complex* than a nonrandom string. This is the basic idea behind *Kolmogorov complexity*. If a string can be achieved easily using an algorithm then the *Kolmogorov complexity* of that string is lower and vice versa.

Mathematically, the *Kolmogorov complexity* of a string can be stated as follows. Suppose an algorithm p is run on a machine. The outcome of the function A_T produces the string s. The, the $KolmogorovComplexity$ of the string s is defined as,

$$K(s) = min\{|p| : s = A_T(s)\}$$

,

where $|p|$ is the length of p.

We can also define a measure of *randomness* as we discussed earlier. A string r is defined as *random* if the shortest length of the algorithm required to produce it is almost equal to the length of itself, i.e.,

$$K(r) \approx |r|.$$

4 Expected Kolmogorov complexity and Shannon entropy

Kolmogorov complexity is a measure of *absolute information* of a single object. In contrast, Shannon entropy is a measure of *average information* in a source emitting random symbols.

We will revisit our example of a source X emitting N symbols $\{x_1, x_2, \ldots, x_N\}$ with probabilities $P = \{P_1, P_2, \ldots, P_N\}$, respectively. While $K(x)$ depend on individual symbols, it is independent of the probability distribution P. But, it is natural to examine the relationship between the expected Kolmogorov complexity of the source and the average information of the source (the entropy $H(X)$).

Recall that the entropy of the source is,

$$H(X) = \sum_{i=1}^{N} P_i \, log\left(\frac{1}{P_i}\right)$$

and the expected Kolmogorov complexity can be computed by weighing the Kolmogorov complexity of each symbol with their respective probability, i.e.,

$$\sum_{i=1}^{N} P_i K(x_i).$$

The relationship as it turns out to be is the following,

$$0 \leq \sum_{i=1}^{N} P_i K(x_i) - H(X) \leq K(P) + \mathcal{O}(1).$$

Here, $K(P)$ is the Kolmogorov complexity of the probability distribution P. The Kolmogorov complexity of a probability distribution P is the shortest program that outputs $P(x)$ to precision q on input $< x, q >$[7]. The difference between the *expected Kolmogorov complexity* and *Shannon entropy* is upper bounded by the Kolmogorov complexity of the probability distribution of the source (and an additive constant $\mathcal{O}(1)$).

This implies that for simple distributions which are less complex, the expected Kolmogorov complexity is close to Shannon entropy. On the other hand, if the probability distribution is complex then they are far apart.

References

[1] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[2] A. Gunawardena, "Lecture notes for ee518 (digital communication)," *Lecture Note 01*, 2021.

[3] R. Solomonoff, "A formal theory of inductive inference. part i," *Information and Control*, vol. 7, no. 1, pp. 1–22, 1964.

[4] A. Kolmogorov, "Three approaches to the definition of the concept "quantity of information"," *Probl. Peredachi Inf.*, vol. 1, pp. 3–11, 1965.

[5] G. J. Chaitin, "On the length of programs for computing finite binary sequences: Statistical considerations," *J. ACM*, vol. 16, p. 145–159, jan 1969.

[6] V. Korotkich, *Kolmogorov complexity*, pp. 1191–1196. Boston, MA: Springer US, 2001.

[7] P. M. V. Peter D. Grunwald, "Algorithmic information theory," *arXiv:0809.2754*, 2008.