Vsevolod Shegolev *Updated: 2024-10-18*

Information Theory

Messages

Communication. Information theory was originally developed to provide a theoretical framework for addressing an engineering problem of communication. Formally, consider a scenario where a sender wants to transmit a message, denoted as ω, to a receiver. In information theory, the semantic content of messages is irrelevant; the focus is solely on the engineering aspects of transmission. Here, the message is treated as a sequence of *symbols*.

Symbols. Symbols are distinct, identifiable entities that form the alphabet used to represent a message. For instance, the digits from 0 to 9 can be used to represent any integer number (the message). Similarly, nodding or shaking one's head can be considered symbols representing the messages "yes" and "no." For hearing-impaired individuals, hand gestures are used as symbols to communicate, with each gesture representing a specific symbol.

Encoding. The simplest symbolic system capable of transmitting information must have at least two distinct symbols, such as "0" and "1." Using this simple formal language, we can encode basic messages.

Example. A message about whether it will rain tomorrow can be encoded with just two symbols:

$$0 = \text{[It will not rain tomorrow]}, \quad 1 = \text{[It will rain tomorrow]}.$$

Example. The outcome of a coin flip can be encoded similarly:

$$0 = \text{[The coin landed on tails]}, \quad 1 = \text{[The coin landed on heads]}.$$

Messages can be represented by the codes used for their encoding. More complex messages can be encoded as sequences of symbols.

Example. For four possible messages:

$$\begin{aligned}\Omega = \{ & \\ & \omega_1 = \text{[It will not rain tomorrow, and the coin landed on heads]}, \\ & \omega_2 = \text{[It will not rain tomorrow, and the coin landed on tails]}, \\ & \omega_3 = \text{[It will rain tomorrow, and the coin landed on tails]}, \\ & \omega_4 = \text{[It will rain tomorrow, and the coin landed on heads]} \\ \},&\end{aligned}$$

The following encoding can be used:

$$00 = \omega_1, \quad 01 = \omega_2, \quad 10 = \omega_3, \quad 11 = \omega_4.$$

These binary sequences carry *information*. Importantly, there is no way to transmit this information in a more compact form; we need at least a sequence of length 2 binary symbols to encode the weather and coin states. Thus, this code is *optimal* and cannot be further compressed.

Uncertainty. A particular message ω^* can be considered successfully transmitted if and only if the receiver can identify this specific message ω^* from the set of all possible messages $\omega in \Omega$. The larger the set Ω is, the harder it is to select ω^* from the set of all possible messages. Therefore, the larger Ω, the greater the *uncertainty* associated with determining the transmitted message. Uncertainty is directly related to the number of possible messages.

Bits. Various encoding schemes can be used to transmit the same message. To abstract away from the details of specific encodings and focus on the information content itself, we introduce the concept of *bits*. A bit is the minimum amount of information required to eliminate uncertainty between two possibilities. The sender must transmit at least 1 bit so that the receiver can distinguish the sent message ω^* from a set of two possible messages $\Omega = \{\omega_0, \omega_1\}$.

If the set of all possible messages contains $N > 2$ messages, each bit of information can be used to eliminate half of the remaining possibilities.

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$$

$$0 = [\omega^* \text{ is in the first half of } \Omega], \quad 1 = [\omega^* \text{ is in the second half of } \Omega].$$

- The first bit divides the set into two halves:

$$\Omega = \{ \quad \overbrace{\omega_1, \omega_2}^{0}, \quad \overbrace{\omega_3, \omega_4}^{1} \quad \}$$

- The second bit further divides the set:

$$\Omega = \{ \quad \overbrace{\underbrace{\omega_1}_{00}, \underbrace{\omega_2}_{01}}^{0}, \quad \overbrace{\underbrace{\omega_3}_{10}, \underbrace{\omega_4}_{11}}^{1} \quad \}$$

Information. The amount of bits required to identify a specific message ω is called the *information content* $I(\omega)$. The sender sends $I(\omega)$ bits of information by transmitting the message ω, and the receiver receives $I(\omega)$ bits of information when decoding the message to identify ω from the set Ω.

Hartley function

Bits. When there are multiple possible outcomes, we can distinguish between them if we have the necessary information. The minimal amount of information is 1 bit. **By definition**, each bit of information distinguishes between 2 possibilities. For example, 1 bit of information is required to unambiguously identify the sex of a child. The event:

$$B = \{\text{Masha gave birth to a boy}\}$$

corresponds to exactly 1 bit of information, and the inverse event similarly corresponds to 1 bit of information:

$$\bar{B} = \{\text{Masha gave birth to a girl}\}$$

Additivity. For two **independent and equally probable** events:

$$B = \{\text{Masha gave birth to a boy}\}, \quad G = \{\text{Lena gave birth to a girl}\}$$

we expect the total amount of information received when both events have occurred to be additive:

$$I(AB) = I(A) + I(B)$$

Since the logarithm satisfies this property, the possible choice is:

$$\log f(AB) = \log f(A) + \log f(B)$$

Probability. The probability p characterizes the frequency of an event. An event with a high probability provides a small amount of information. For instance, the probability of the sunrise is nearly one, so the information that there will be a sunrise tomorrow carries little value. In contrast, the information that there will be no sunrise tomorrow conveys a significant amount of information. Thus, **the lower the probability, the more information is conveyed**:

$$I(A) = I(\mathbb{P}[A]) \qquad \text{and} \qquad p \uparrow \Leftrightarrow I \downarrow .$$

The appropriate formula that satisfies these conditions is:

$$I(f(A)) = \log \frac{1}{\mathbb{P}[A]}, \quad f(A) = \mathbb{P}(A).$$

$$I(AB) = \log \frac{1}{\mathbb{P}[A]} + \log \frac{1}{\mathbb{P}[B]}$$

Inverse probability. The inverse probability $\frac{1}{p}$ represents the **expected number of trials** needed to achieve **one occurrence** of an event with probability p. For example, if $p = 0.01$, the event occurs, on average, once every 100 trials.

Information content. Information is the **capacity to distinguish** between possibilities. Each bit of information distinguishes 2 possibilities, and it can assume 2 different values, 0 and 1. n bits of information distinguish 2^n possibilities. Hence, the amount of information required to distinguish between 2^n possibilities is n bits.

In k-valued logic, each k-valued digit (0, 1, ..., $k-1$) represents information:

- In binary logic, each digit is a bit (0 or 1).
- In ternary logic, each digit is a trit (0, 1, or 2).

If there are N outcomes, each time you assign a bit value to an outcome, you divide all outcomes into 2 sets corresponding to the bit values:

$$\Omega = \underbrace{\{\omega \in \Omega \mid \omega\text{'s bit value} = 0\}}_{\Omega_0} \cup \underbrace{\{\omega \in \Omega \mid \omega\text{'s bit value} = 1\}}_{\Omega_1}$$

Now, for any ω, knowing the corresponding bit value allows you to determine whether $\omega \in \Omega_0$ or $\omega \in \Omega_1$, thereby halving the uncertainty.

Repeating this I times, you partition Ω into 2^I disjoint sets, or more precisely, $\min(2^I, N)$, as Ω contains only N elements:

$$\Omega = \{\omega_1\} \cup \{\omega_2\} ... \cup \{\omega_N\}$$

Once the sets contain only one element, further bits do not provide additional meaningful information. Therefore, the amount of information is proportional to the size of Ω. Each bit splits Ω into 2 parts, and each subsequent bit continues dividing the sets into 2 parts. However, it is only meaningful to repeat these binary divisions up to $\log_2 N$ times.

Thus, the exact number of bits needed to distinguish all outcomes is:

$$I = \log_2 N = \log_2 \frac{1}{p}.$$

Shannon self-information

Self-information, introduced by Claude Shannon, quantifies the amount of information or "surprise" associated with the occurrence of an event. The key properties of Shannon's self-information are:

- An event with a probability of 100% is unsurprising and thus carries no information.
- Events that are less probable yield more information when they occur.
- For two independent events, the total information is the sum of their individual self-informations.

The self-information for an event A is defined as:

$$I(A) \coloneqq \lg_2 \frac{1}{\mathbb{P}[A]}.$$

For a random variable X taking a specific value x with probability $\mathbb{P}[X = x]$, the self-information is:

$$I_X(x) \coloneqq \lg_2 \frac{1}{\mathbb{P}[X = x]}.$$

Odds ratio

The **odds** of an event A is defined as the difference in self-information (also known as surprisal) between the event A and its complement $\bar{A}$:

$$\begin{aligned}\text{Odds } A &:= I(A) - I\left(\bar{A}\right) \\ &= \log \frac{\mathbb{P}[A]}{1 - \mathbb{P}[A]}.\end{aligned}$$

Shannon entropy

Information associated with a probability distribution. Information corresponds to the amount of uncertainty: the more uncertain (less probable) an outcome is, the more information it carries.

As for a single outcome $\omega \in \Omega$, we can define the information associated with the probability distribution $\{\mathbb{P}[\omega] \mid \omega \in \Omega\}$. Let it be the expected value of the self-information:

$$\begin{aligned}H[\mathbb{P}, \Omega] &:= \sum_{\omega \in \Omega} \mathbb{P}[\omega] \cdot I(\omega) \\ &= \sum_{\omega \in \Omega} \mathbb{P}[\omega] \cdot \log_2 \frac{1}{\mathbb{P}[\omega]}.\end{aligned}$$

This quantity is known as the Shannon entropy, it can be interpreted as the average amount of information produced by the probability distribution.

As a random variable X induces its own probability distribution $\{\mathbb{P}[X = x] \mid x \in \operatorname{supp}(X)\}$, the entropy can be defined specifically for the random variable:

$$\begin{aligned}H(X) &:= \sum_{x \in \operatorname{supp}(X)} \mathbb{P}[X = x] \cdot I(X = x) \\ &= \sum_{x \in \operatorname{supp}(X)} \mathbb{P}[X = x] \cdot \log_2 \frac{1}{\mathbb{P}[X = x]},\end{aligned}$$

which is equivalent to the distribution of the events $\{[X = x] \mid x \in \operatorname{supp}(X)\}$.

Practical interpretation.

Conditional entropy

Entropy H can be generalized to the conditional case. Suppose that we have random variables X and Y, with their (marginal) distributions $\mathbb{P}[X = x]$ and $\mathbb{P}[Y = y]$ and their joint distribution $\mathbb{P}[X = x, Y = y]$.

Specific conditional entropy can be trivially defined by replacing the probability $\mathbb{P}[X = x]$ with the conditional probability $\mathbb{P}[X = x | Y = y]$:

$$H(X|Y = y) := \sum_{x \in \operatorname{supp} X} \mathbb{P}[X = x|Y = y] \cdot \log \frac{1}{\mathbb{P}[X = x|Y = y]}.$$

Conditional entropy (non specific) is can be defined as the expected value of the specific conditional entropy over all possible values of $y \in \operatorname{supp} Y$:

$$\begin{aligned}H(X|Y) &:= \mathbb{E}_Y[H(X|Y = y)] \\ &= \sum_{y \in \operatorname{supp} Y} \mathbb{P}[Y = y] \cdot H(X|Y = y)\end{aligned}$$

Boltzmann distribution maximizes thermodynamic probability W and provides the **probability for each state**, so the Boltzmann formula defines a **probability distribution**:

$$\frac{n_i}{N} = \frac{e^{-\beta E_i}}{\sum_i e^{-\beta E_i}}.$$

Shannon's information entropy can be applied to this probability distribution:

$$p_i := \frac{e^{-\beta E_i}}{Z}, \quad Z := \sum_i e^{-\beta E_i},$$

$$\begin{aligned}H &= \sum_i p_i \cdot \ln p_i \\ &= \sum_i \frac{e^{-\beta E_i}}{Z} \cdot (-\beta E_i - \ln Z) \\ &= -\beta \sum_i \underbrace{\frac{e^{-\beta E_i}}{Z}}_{p_i} \cdot E_i - \ln Z \cdot \underbrace{\sum_i \frac{e^{-\beta E_i}}{Z}}_{=1} \\ &= -\beta \langle E \rangle - \ln Z\end{aligned}$$

As a result, we get statistical entropy expressed via the partition function:

$$S_{\text{stat}} = k_B \cdot H$$

The amount of information needed to encode the probability distribution by energy levels can be calculated via Shannon's formula. This amount of information directly corresponds to the statistical thermodynamic entropy with k_B units, which translates bits into energy per temperature units.

Fair dice induces the uniform distribution over the set of possible outcomes $p_1 = ... = p_6 = \frac{1}{6}$.

We need exactly $I = \log_2 6 \approx 2.58$ bits to encode each outcome. All outcomes are equally probable, so we need the same I bits of information to encode any outcome on average.

On the other hand, when we roll a fair dice, we receive $\log_2 6$ bits of information from any outcome, and on average we receive the *entropy* H amount of information:

$$H = \sum_{i=1}^{6} \frac{1}{6} \cdot \log_2 \frac{1}{6} = \log_2 \frac{1}{6} \approx 2.58 \text{ bits}.$$