Vsevolod Shegolev

Updated: 2024-10-02

Information Theory

Hartley Information

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's 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:

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

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

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

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 $\bar{A}$:

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

Shannon entropy

Shannon cross entropy

Kullback-Leibler divergence

Mutual information

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:

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

$$H = \sum p_i \cdot \ln p_i$$