

Information Theory

Started by Shannon in 1949.

We use it to measure how much (cipher) text is needed to break a cipher.

The amount of information in a message is measured by its *entropy*.

Definition: Let X_1, \dots, X_n be all possible ($n = a^k$, say) messages (of length k using an alphabet of a letters), with respective probabilities $p(X_1), \dots, p(X_n)$, so that $\sum_{i=1}^n p(X_i) = 1$. The entropy of a message $X \in \{X_1, \dots, X_n\}$ is

$$H(X) = - \sum_{i=1}^n p(X_i) \log_2 p(X_i)$$

$$= \sum_{i=1}^n p(X_i) \log_2 \left(\frac{1}{p(X_i)} \right).$$

Example: If $X \in \{H, T\}$, equally likely, then $H(X) = 2(-1/2) \log_2(1/2) = 1$.

Intuitively, $\log_2(1/p(X))$ is the number of bits needed to encode X in an optimal encoding (Huffman code). Thus, $H(X)$ is the expected number of bits needed to represent X . Short codes are used for popular messages.

In defining entropy, Shannon required that it satisfy three properties. First, it should be a continuous function of the variables p_1, \dots, p_n , subject to $p_1 + \dots + p_n = 1$. Second, when the messages are equally likely, that is, every $p_i = 1/n$, H should be an increasing function of n . He required this property because there is more choice, or uncertainty, when there are more equally likely messages. The third property said that if the choice of one among n messages is replaced by two successive choices, first of a subset of the messages and then a message in the chosen subset, then the entropy of the set of messages should be a weighted sum of the entropies of the two choices. For example, if there are four equally likely messages, we may choose one as of them follows: (1) Choose a subset of the messages, either the first one or the second one or the last two. (2) If the subset was the last two, choose one of them. Then the third property would say

$$H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) = H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right) + \frac{1}{2}H\left(\frac{1}{2}, \frac{1}{2}\right).$$

The coefficient of the last term is $\frac{1}{2}$ because the second choice is made half of the time.

From these three properties, Shannon proved that the entropy must be

$$H(p_1, \dots, p_n) = -K \sum_{i=1}^n p_i \log p_i,$$

where K is a positive constant, The constant K may be regarded as a choice of units for entropy. Choosing $K = 1/\log 2$ makes the binary digit the unit of entropy. His theorem motivates the definition of entropy.

More about Entropy

Example. n equally likely messages have entropy $H(X) = -\sum_1^n (1/n) \log_2(1/n) = \log_2 n$, so we need $\log_2 n$ bits to encode each message.

What if n is not a power of 2? Then we have to approximate. For example, if $n = 3$, code each group of 3 messages in 5 bits. Note that $\log_2 3 \approx 1.5849 \approx 5/3$.

Given fixed n and variable probabilities $p(X_i)$, the entropy $H(X)$ is maximum in the equally likely case. $H(X) = 0$ in case $p(X_i) = 1$ for some i .

Entropy measures uncertainty in that it gives the number of bits which are learned when X is received. Public-key systems are particularly vulnerable to a ciphertext-only attack if the plaintext entropy is small: Just try all cases.

Example. Salary data is an integer $\leq \$100000$. Improve security by adding a random string.