Entropy and Information

1. Uncertainty and Entropy

The primary characteristic of a random experiment is the inability to ascertain which outcome will occur before the experiment is conducted, meaning that a random experiment involves a degree of uncertainty. However, the extent of this uncertainty can vary greatly for different random experiments. For example, consider shooting: if there are two shooters, their shooting performances can be described using the following two random experiments：

$$\text{A:}\begin{pmatrix} A, \overline{A} \\ 0.5,0.5 \end{pmatrix} \qquad \text{B:}\begin{pmatrix} A, \overline{A} \\ 0.99,0.01 \end{pmatrix}$$

Here, A represents hitting the target, and $\overline{A}$ represents missing the target. The next line indicates the corresponding probabilities. Clearly, the levels of uncertainty in these two experiments are quite different, with experiment A having much greater uncertainty. If there is a third shooter, the random experiment describing their shooting proficiency would be:

$$\text{C:}\quad \begin{pmatrix} A, \overline{A} \\ 0.7,0.3 \end{pmatrix}$$

It is evident that the uncertainty of this experiment should be considered to lie between the levels of uncertainty described above. Therefore, it is necessary to numerically estimate the uncertainty of various random experiments. In other words, we hope to find a quantity that can reasonably serve as a measure of the degree of uncertainty. Such a quantity has already been found by the American mathematician Shannon.

Suppose we are studying a random experiment α with only a finite number of incompatible outcomes: $A_1, A_2, ..., A_n$, with corresponding probabilities $p(A_1), p(A_2), ..., p(A_n)$, satisfying $\sum_{i=1}^{n} p(A_i) = 1$, , abbreviated as follows:

$$\text{a}:\begin{pmatrix} A_1, A_2, ..., A_n \\ p(A_1), p(A_2), ..., p(A_n) \end{pmatrix}$$

We hope to find a quantity H(a) to measure the uncertainty of α. This quantity naturally depends on $p(A_1), p(A_2),..., p(A_n)$

and is denoted as $H(p(A_1), p(A_2),..., p(A_n))$

To precisely determine the expression for H(a), let's first consider what requirements H(a) should satisfy.

Firstly, we demand that:

(i) H being a continuous function of $p(A_i)$

On one hand, small variations in p(Ai) should not lead to significant changes in H, and on the other hand, only continuous functions are convenient for mathematical treatment.

Secondly, let's consider a special type of random experiment with n outcomes, each with a probability of 1/n. We'll refer to this as an experiment with n equiprobable outcomes. In this special experiment, H should only be a function of n, and as n increases, indicating more possible outcomes, the corresponding level of uncertainty should increase as well. Therefore, we demand from H:

(ii) For experiments with equiprobable outcomes, H is a monotonically increasing function of n.The third requirement for H(a) is more complex, involving the division of one experiment into two successive experiments. We will elucidate its meaning through a simple example.

$$a:\begin{pmatrix} A_1, A_2, A_3 \\ p_1, p_2, p_3 \end{pmatrix}$$

Consider an experiment with three outcomes. The uncertainty of this experiment is denoted as $H(\alpha) = H(p_1+ p_2+p_3)$. To determine which outcome occurs, we can also conduct two

successive experiments: In the first experiment, we first ascertain whether A_1 occurs or if A_2 or A_3 occurs, which can be represented by the following experiment:

$$a_1 : \begin{pmatrix} A_1, B \\ p_1, p_2 + p_3 \end{pmatrix}$$

Clearly, $H(a_1) = H(p_1, p_2 + p_3)$. If A_1 occurs (with probability p_1), the outcome of the experiment is already completely determined, and no further experiments are necessary. However, if B occurs (with probability $P_2 + P_3$), further experiments are needed to finally determine the outcome of the experiment:

$$a_2 : \begin{pmatrix} A_2, A_3 \\ \frac{p_2}{p_2 + p_3}, \frac{p_3}{p_2 + p_3} \end{pmatrix}$$

The uncertainty of this experiment is $H(a_2) = H\left(\frac{p_2}{p_2 + p_3}, \frac{p_3}{p_2 + p_3}\right)$

We could directly conduct experiment α to determine which of A_1, A_2, or A_3 occurs. However, if experiment B is conducted first and then if necessary (with probability $P_2 + P_3$), experiment α is conducted, the same result can be achieved. Therefore, it is natural to consider the uncertainty contained in these two sets of experiments to be the same, i.e.

$$H(p_1, p_2, p_3) = H(p_1, p_2 + p_3) + (p_2 + p_3) H\left(\frac{p_2}{p_2 + p_3}, \frac{p_3}{p_2 + p_3}\right)$$

These considerations lead us to propose the following requirement for H:

(iii) When an experiment is divided into successive experiments, the uncertainty H before division is the weighted sum of the uncertainties after division.

Conditions (i), (ii), and (iii) completely determine the form of H. For convenience, let's denote $P(A_i)$ as P_i:

Theorem 4.3.1 (Shannon):

The only function H satisfying conditions (i), (ii), and (iii) has the following form:

$$H = -C\sum_{i=1}^{n} p_i \log p_i \qquad (4.3.1)$$

where C is a positive constant.

To prove this theorem, we need the following analytical lemma:

Lemma 4.3.1:

If f(n) is a monotonically increasing function of n and for all positive integers m, n,

the following holds:

$$f(mn) = f(m) + f(n) \qquad (4.3.2)$$

then, $f(n) = C \log n$

Where C is a positive constant.

[Proof]

From (4.3.2), we obtain f(1) = 0. Thus, for other positive integers m, we have f(m) > 0. On the

other hand,

$$f(n^2) = f(n) + f(n) = 2f(n)$$
$$f(n^3) = f(n^2) + f(n) = 3f(n)$$

generally,

$$f(n^k) = kf(n) \qquad (4.3.3)$$

if n and m are any two positive integers with m ≠ 1, we select any sufficiently large positive

integer k and then choose a positive integer l such that

$$m^l \le n^k < m^{l+1} \quad (4.3.4)$$

Due to the monotonicity of the function, we have:

$$f\left(m^l\right) \le f\left(n^k\right) < f\left(m^{l+1}\right)$$

Therefore:

$$\frac{l}{k} \le \frac{f(n)}{f(m)} < \frac{l+1}{k}$$

Taking logarithms(4.3.4), we get:

$$l \log m \le k \log n < (l+1) \log m$$

Thus, we also have:

$$\frac{l}{k} \le \frac{\log n}{\log m} < \frac{l+1}{k}$$

In this way:

$$\left| \frac{f(n)}{f(m)} - \frac{\log n}{\log m} \right| < \frac{1}{k}$$

The above inequalities hold for any sufficiently large k. Therefore:

$$\frac{f(n)}{f(m)} = \frac{\log n}{\log m}$$

Due to the arbitrariness of m and n, we know that:

$$f(n) = C \log n$$

Where C is a constant, and given that f(n) is an increasing function of n, it follows that C is positive.

Now we can proceed with the proof of the theorem.

[Proof]

Firstly, let's denote $H\left(\frac{1}{n},\frac{1}{n},...,\frac{1}{n}\right)=f(n)$

.According to condition (ii), we know that f (n) is a monotonically increasing function of n. For an experiment with mn equiprobable outcomes, it can be decomposed into m experiments each with n equiprobable outcomes.

Therefore, according to condition (iii), we have

$$f(mn)=f(m)+m\bullet\frac{1}{m}f(n)=f(m)+f(n)$$

Utilizing Lemma 4.3.1, we immediately obtain:

$$H\left(\frac{1}{n},\frac{1}{n},...,\frac{1}{n}\right)=C\log n$$

Secondly, when $p_1,p_2,...,p_n$ are rational numbers,

we may denote $p_i=\frac{n_i}{\sum_{i=1}^{n}n_i}$

Consider an experiment with $\sum_{i=1}^{n}n_i$ equiprobable outcomes, which can be viewed as two successive experiments. In the first experiment, outcome A_i occurs with probability P_i, while in the second experiment, given the occurrence of outcome A_i, it is further examined which of the n_i equiprobable outcomes occurs. Thus, according to condition (iii), we have:

$$C\log\sum_{i=1}^{n}n_i=H(p_1,p_2,...,p_n)+C\sum_{i=1}^{n}p_i\log n_i$$

Therefore:

$$
\begin{aligned}
H(p_1, p_2, ..., p_n) &= C\left[\log \sum_{i=1}^{n} n_i - \sum_{i=1}^{n} p_i \log n_1\right] \\
&= C\left[\sum_{i=1}^{n} p_i \left(\log \sum_{j=1}^{n} n_j - \log n_i\right)\right] \\
&= -C \sum_{i=1}^{n} p_i \log p_i
\end{aligned}
$$

Lastly, for any real numbers $p_1, p_2, ..., p_n$ we can approximate them with rational numbers. However, according to condition (i), H is a continuous function of its variables. Therefore, the above expression still holds true.