A Brief Review of Entropy

Li Zichao
WISE 2016 IUEC

March 15, 2019

Abstract

This is a short theoretical report about the definiton as well as some characteristics of information entropy and maxent and minxent .

1 Information and Entropy

Imagine the example below. We have the empirical mean value of tossing a six-sides die for a million time and we also know the sum of the probabilities of all results is one. And we want to predict the result of the next throw. Obviously we can not make certain prediction with only two knowns. To get greater certainty, more information, such as the air condition, the motion of toss and the weight of the die is required. In the theory of information, **information is defined as the solution of uncertainty**. In the process of reduction in uncertainty, we need a measure of it thus we can quantify the decrease of it.

In 1948, Claude Elwood. Shannon[1] introduced the concept of Entropy[2] into the theory of Information in his landmark paper, *A Mathematical Theory of Communication*. Shannon concluded three properties that a reasonable measurement of uncertainty, says H should satisfy:

1. H should be continuous.
2. H is monotonic increasing of the number of events.

[1] An American mathematician, electrical engineer, and cryptographer known as "the father of information theory".

[2] A definition in thermodynamics to measure the disorder of a system.

3. H should be additive.

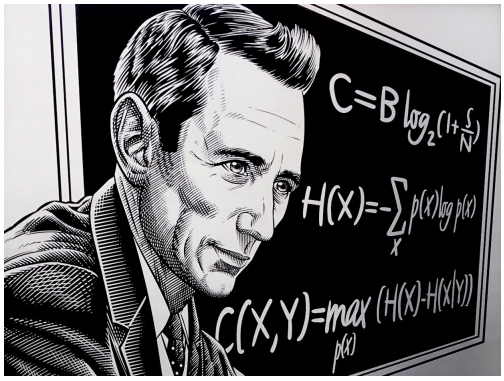

Figure 1: Claude E. Shannon

The H satisfying the three above assumptions is of the form

$$H(p) = -E_p[\log(p(x))] = -\int p(x)\log(p(x))\,\mathrm{d}x \quad OR \quad -\sum_i p_i \log(p_i) \tag{1}$$

where p_i is the probability of x in cell i of its phase space. Larger entropy implies greater uncertainty, and it increases as the probability distribution becomes more even, since it is harder to tell which case is most likely to happen next time. Take the entropy of a Bernoulli distribution with probability p as an example, the entropy

$$H = -(p\log p + q\log q) \tag{2}$$

is plotted in Figure 2. Intuitively, the entropy is maximized when $p = (1-p) = \frac{1}{2}$. Provement is in the next section

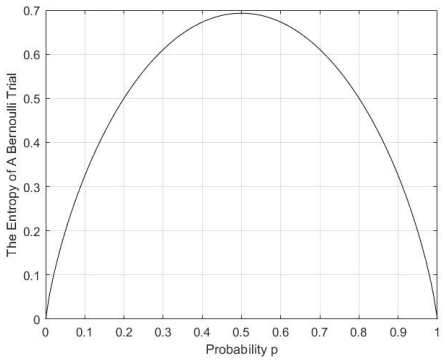

Figure 2: The entropy of a Bernoulli distribution

2 Maxent

Maximum entropy, as known as maxent, is **the notion of measuring the consistence between a distribution and the current state of our knowledge**. To maximize

$$H = -\sum_i p_i \log(p_i) \tag{3}$$

we can apply the lagrange-multiplier method as

$$\begin{aligned} &\max_{p_i} H = -\sum_i p_i \log(p_i) \\ &Subject \quad to \quad \sum_i p\left(x_i\right) - 1 = 0 \end{aligned} \tag{4}$$

$$L = -\sum_i p_i \log(p_i) + \lambda \left(\sum_i p\left(x_i\right) - 1 \right) \tag{5}$$

$$\frac{\partial H}{\partial p_j} = 0 \implies -\left(1 + \log(p_j)\right) + \lambda = 0 \tag{6}$$

Thus, p_j are all equal for $\forall j \in n$ and must be $\frac{1}{n}$ to maximize the entropy because uncertainty reaches its peak when all possible outcomes are equiprobable.

3 Cross Entropy

In information theory, the cross entropy between two probability between two probability distributions measures the performance of identification of an event drawn from the underlying set when the coding scheme[3] is optimal for estimated distribution rather than the true one.

[3] In Shannon's theory, when information is transmitting through channel to receiver, first it will be encoded then decoded after receive.

Using the Kullback-Leibler divergence $D_{KL}(p||q)$[4], the cross entropy is defined as

$$\begin{aligned} H(p,q) &= H(p) + D_{KL}(p||q) \\ &= E_p[-\log(q)] \end{aligned} \tag{7}$$

For discrete probability p and q,

$$H(p,q) = -\sum p(x) \log(q(x)) \tag{8}$$

For continous distribution the situation is analogous. With the definition it is easy to discover that the cross entropy is equivalent to the KL divergence when comparing two distributions. Therefore the cross entropy is used to measure the closeness between a distribution and a prior reference distribution. The cross entropy reaches its minimal value when $p = q$. This principle is called as **Principle of Minimun Cross-Entropy**, or **Minxent**

References

[1] Cross entropy. https://en.wikipedia.org/wiki/Cross_entropy.

[2] Rahul Dave. Entropy. http://am207.info/wiki/Entropy.html, 2019.

[3] Amos Golan, editor. *Information and entropy econometrics*. Elsevier B. V., Amsterdam, 2002. J. Econometrics **1**07 (2002), no. 1-2.

[4] Jiaming Mao. Foundations of statistical learning. https://jiamingmao.github.io/data\discretionary{-}{}{}analysis/assets/Lectures/Foundations_of_Statistical_Learning.pdf, 2019.

[5] C. E. Shannon. A mathematical theory of communication. 1948.

[4] p is a fixed reference distribution and q is the distribution we use to estimate p.