Information Theory

listenzcc

March 11, 2020

Abstract

This manuscript gives a brief introduction of *Information Entropy*. It can be used as a manual of how to calculate *Shannon Entropy*, *Mutual Information*, *Transfer Entropy* and *Relative Entropy*. Brief proofs of the key concepts are also provides. In practice, one can also use the *entropy.py* besides this manuscript to calculate the Information Entropy above.

Contents

1 Information Entropy **1**
1.1 Shannon Entropy 2

2 Mutual Information **3**
2.1 Independent variables 3
2.2 Dependent variables 3

3 Transfer Entropy **4**
3.1 Definition 4
3.2 Calculating 5

4 Relative Entropy **5**
4.1 Definition 5
4.2 Non-negative 5

1 Information Entropy

Information Theory is one of the few scientific fields fortunate enough to have an identifiable beginning - Claude Shannon's 1948 paper.

> What made possible, what induced the development of coding as a theory, and the development of very complicated codes, was Shannon's Theorem: he told you that it could be done, so people tried to do it. [Interview with Fano, R. 2001]

For a given distribution X $P(X)$, a single symbol x can explain the amount of uncertainty as the information quantity

$$I(x) = \log \frac{1}{p(x)} = -\log p(x) \tag{1}$$

where the base of log is arbitrary. Usually we use 2 as the base, and unit of the information quantity is called *bytes*.[1]

[1] The log in this manuscript uses the base of 2 if not specified.

1.1 Shannon Entropy

The expectation of the mean of the information quantity is information entropy (or *Shannon entropy*)

$$H(X) = \int_X p(x)I(x)dx \tag{2}$$

Theorem 1.1. *The information entropy is maximized when all the symbols occurs in equal probabilities. In a discrete situation, X has n possible values. When $p(x) = \frac{1}{n}$, the information entropy is maximized.*

Proof. Re-write information entropy as

$$H(x) = -\mathcal{C}\sum_{i=1}^{n} p(x_i)\ln p(x_i)$$

where $\mathcal{C}$ is a constant which guarantee $\mathcal{C}\ln p = \log p$ when $0 < p < 1$. To maximizing the information entropy, there is another constraint that $\sum_{i=1}^{n} p(x) = 1$.

Use Lagrangian method to solve the constrained maximizing problem. Formulate Lagrangian function

$$\mathcal{L}(x) = H(x) + \lambda(\sum_{i=1}^{n} p(x_i) - 1)$$

where λ is unsolved constant.

Calculate the partial differential of $\mathcal{L}(x)$ to $p(x_i)$

$$\frac{\partial}{\partial p(x_i)} H = \lambda - \mathcal{C}\ln p(x_i) + \mathcal{C}$$

the maximizing of information entropy is equivalent to the partial differentials equal to zero for each $i \in [1, 2, \ldots, n]$.

Since λ is constant, we have

$$p(x_i) = p(x_j) = p(x)$$

for each $i \neq j$ and $i, j \in [1, 2, \ldots, n]$. Hence proved. □

In the sense of above analysis, we can see that the information entropy can be considered as the minimized code length of a communication system. To make sure the system reaches the minimized code length, an efficient way is to design it making sure all the symbols are happening with equal possibility. Here is another question to be answered: how many symbols do we have to use in the system?

Theorem 1.2. *The best number of symbols in a equal possibility system is e. The value can maximize the information quantity of a single symbol.*

Proof. Re-write the information entropy in a equal possibility discrete system.

$$H(x) = -\mathcal{C}\sum_{i=1}^{n} p\ln p$$

where $p = p(x_i) = \frac{1}{n}$ for $i \in [1, 2, \ldots, n]$.

Since all the symbols have the same probability, the information quantity of a single symbol can be wrote as

$$I = \mathcal{C}\frac{1}{n}\ln n$$

Calculate the partial derivative by n, we have

$$\frac{\partial}{\partial n} I = \mathcal{C}(\frac{1}{n^2} - \frac{1}{n^2} \ln n)$$

one can see that $n = e$ can make $\frac{\partial}{\partial n} I = 0$, and the $2^{nd} - order$ partial is negative when $n = e$. It turns out that the value maximizes the information being carried by single symbol. Hence proved. □

2 Mutual Information

In practice, one may concerns the interaction between several variables. We can start with two variables. The simplest situation is that two variables are independent with each other.

2.1 Independent variables

If X and Y are independent with each other, the joint probability can be expressed as

$$P(X, Y) = P(X)P(Y) \tag{3}$$

which is a necessary condition of independence, although not sufficient.

Theorem 2.1. *The information entropy of independent variables equals to the summation of each information entropy.*

Proof. The information entropy of X and Y can be expressed as

$$H(X, Y) = -\int_X \int_Y P(x, y) \log(P(x, y)) dxdy \tag{4}$$

use (3), we have

$$H(X, Y) = -\int_X P(x) \log(P(x)) dx - \int_Y P(y) \log(P(y)) dy \tag{5}$$

the equation also uses the fact that $\int_Y P(x, y)dy = P(x)$. Use the definition in (2) we have

$$H(X, Y) = H(X) + H(Y)$$

Hence proved. □

2.2 Dependent variables

If X and Y are not independent, the mutual information can be expressed as

$$I(X; Y) = \int_X \int_Y p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dxdy \tag{6}$$

The meaning of mutual information is the uncertainty of one variable solved by the fact of knowing another variable.

Theorem 2.2. *The mutual information is symmetrical*

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{7}$$

Proof. Re-write (6) we have

$$I(X;Y) = \int_X \int_Y p(x,y) \log p(x,y) dxdy - \int_X p(x) \log p(x) dx - \int_Y p(y) \log p(y) dy$$

Start with $H(Y|X)$, it is the information entropy of conditional probability.

$$H(Y|X) = -\int_X \int_Y p(y,x) \log p(y|x) dxdy$$

calculate further

$$H(Y|X) = -\int_X \int_Y p(y,x) \log p(y,x) dxdy + \int_X p(x) \log p(x) dx$$

Easy to obtain

$$H(Y|X) + I(X;Y) = H(Y)$$

Reverse is the same. if we start with $H(X|Y)$, it will yield $H(X|Y) + I(X;Y) = H(X)$. Hence proved. □

Theorem 2.3. *The entropy of conditional distribution is*

$$H(X|Y) = H(X,Y) - H(Y) \tag{8}$$

Proof. It is more like a definition, we can only provide simple proof here. The entropy of conditional distribution can be *defined* as

$$H(X|Y) = -\int_X \int_Y p(x,y) \log \frac{p(x,y)}{p(y)} dxdy$$

thus we have

$$H(X|Y) = -\int_X \int_Y p(x,y) \log p(x,y) dxdy + \int_X p(x) \log p(x) dx$$

Hence proved. □

Lemma 2.1. *The mutual information can also being expressed as following*

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{9}$$

$$I(X;Y) = H(X,Y) - H(X|Y) - H(Y|X) \tag{10}$$

3 Transfer Entropy

Every system has its own trivial dynamic. If we want to measure the impact from *input*, the self-dynamic should be *zeroed-out*.

3.1 Definition

The *Transfer Entropy* is an useful measurement for the *pure* impact.

$$T_{X \to Y} = H(Y|\overline{Y}) - H(Y|\overline{Y}, X) \tag{11}$$

where Y refers the variable we are interested in, $\overline{Y}$ refers the history state of the variable Y, X refers the impact factor.

To be more clear, $H(Z|X,Y)$ means the conditional entropy of Z given (X,Y), not any other wise.

One meaning of transfer entropy is the amount of uncertainty solved by input variable X regardless the history state of the system.

3.2 Calculating

Applying (8) we have the method of computing transfer entropy

$$T_{X\to Y} = H(Y,\overline{Y}) - H(\overline{Y}) - H(Y,\overline{Y},X) + H(\overline{Y},X) \tag{12}$$

$$T_{X\to Y} = \int_X \int_{\overline{Y}} \int_Y p(x,\overline{y},y)\log\frac{p(x,\overline{y},y)p(y)}{p(x,y)p(\overline{y},y)}dxd\overline{y}dy \tag{13}$$

4 Relative Entropy

4.1 Definition

The *Relative Entropy* is a measurement of the different between two distributions. In formal words, relative entropy is the additional information we need to fully solve the uncertainty of the distribution $P(X)$ using the optimized symbol system derived from distribution $Q(X)$.

The entropy of an already known distribution $Q(X)$ is

$$H(X) = -\int_X q(x)\log q(x)dx$$

according to the meaning of shannon entropy, the entropy represents the optimizing minimized coding length of the system sending the symbols following $Q(X)$. However, the minimization can be reached only when the underlying unknown distribution $P(X)$ matches with the known one $Q(X)$. When the condition is not met, we have

$$H^{'}(X) = -\int_X p(x)\log q(x)dx \tag{14}$$

where the new entropy is definitely not smaller than the original one of $Q(X)$. The subtraction is *Relative Entropy*

$$H_{q\mapsto p} = \int_X p(x)\log\frac{p(x)}{q(x)}dx \tag{15}$$

In practice, the Relative Entropy is widely used for measuring the *distance* between two distributions. The smaller value of Relative Entropy, the closer. The minimal value of Relative Entropy is 0.

4.2 Non-negative

Theorem 4.1. *The relative entropy is non-negative*

$$H_{q\mapsto p} \geqslant 0 \tag{16}$$

Proof. Re-write relative entropy as

$$H_{q\mapsto p} = -\int_X p(x)\log\frac{q(x)}{p(x)}dx$$

since log is convex function, we have

$$\log\sum x,y,\ldots \geqslant \sum \log(x),\log(y),\ldots$$

thus we can exchange the log outside the integral, and transform the equational into following

$$H_{q\mapsto p} \geqslant -\log\int_X p(x)\frac{q(x)}{p(x)}dx$$

use the trivial condition that $\int_X q(x)dx \leqslant 1$ and $\log 1 = 0$, we have $H_{q\mapsto p} \geqslant 0$. Hence proved. □