Information Theory

listenzcc

March 9, 2020

Abstract

Not done yet.

Contents

1 Information Theory

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]

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

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

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

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.

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. □

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 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

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

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