A Brief Introduction to Shannon Entropy

March 15,2019

1 Definition

What is information and how to calculate it?

In information theory, entropy is the measure of the amount of the information that is missing before reception and is sometimes referred to as Shannon entropy. Shannon entropy is a broad and general concept which finds application in information theory as well as thermodynamics. It was originally devised by Claude Shannon in 1948 to study the amount of information in a transmitted message.

2 Why we need the Shannon entropy?

Why is there a large amount of information and a small amount of information?

Some things are not quite certain, such as whether the stock is going up or down tomorrow. And some things are certain, for example, the sun rises from the east, and you tell me a hundred times that the sun rises from the east, and you have no information at all, because it can't be more sure.

Therefore, the amount of information is related to the uncertainty. So what does the uncertainty have to do with it? First, it is related to the number of possible results; second, it is related to probability.

For example, we discuss where the sun rises. There is only one result, and we already know that no matter who sends any information, there is no information. When the number of possible results is large, the new information we get has the potential to have a large amount of information. Second, it also depends on the initial probability distribution. For example, I knew that Xiaoming watched a movie in the A-Room, which had 15 seats in the cinema. Xiao Ming

can sit 225 positions, may result in a large number of. But if we knew from the that Xiao Ming was probably 99% on the far left of the first row, and there was little chance of sitting in any other position, then in most cases, it would not be very useful for you to tell me anything about Xiao Ming. Because we are almost sure Xiaoming is on the far left in the first now.

3 How to calculate the Shannon entropy?

For a discrete random variable x with probability distribution p(x), the average amount of the information transmitted by x is:

$$H = -\sum_{X} P(X) log P(X)$$

H is called the entropy of probability distribution p.

Following are three important properties of this format:

A.Monotonicity, that is, the higher the probability of occurrence, the lower the information entropy. The extreme case is the Sun rising from the East, because it does not carry any information to determine the event. From the perspective of information theory, this sentence does not eliminate any uncertainty.

B.Non-negativity, that is, information entropy can not be negative. This is easy to understand, because negative information, that is, when you know a certain information, it is illogical to increase uncertainty.

C. the measurement of the total uncertainty of simultaneous occurrence of multiple random events, can be expressed as the sum of the uncertainties of each event.

4 Applications of Shannon entropy

One of the important application fields of information entropy is natural language processing. For example, a 500000-word Chinese book has an average amount of information. We know that the commonly used Chinese characters are about 7000 characters. If each Chinese character is equal to probability, about 13 bits is required to represent a Chinese character. The application of information entropy is that a Chinese character has 7000 possibilities, each of which

is equal to probability, so the information entropy of a Chinese character is:

$$H = -\left[\left(\frac{1}{7000}\right) * log(\frac{1}{7000}) + \cdots + \left(\frac{1}{7000}\right) * log\left(\frac{1}{7000}\right)\right] = 12.77(bit)$$

In fact, since the first 10% of Chinese characters account for more than 95% of the common text, considering the context of words and other words, the information entropy of each Chinese character is about 5 bits. So a 500000-word Chinese book, the amount of information is about 2.5 million bits. Note that the 2.5 million bits here are an average. Information entropy is high, does not mean that you say, written text contains more information than others. More precisely, information entropy is used to compress language data in natural language processing and has nothing to do with language literacy. Of course, conditional entropy, relative entropy and so on should be more useful concepts in natural language processing. I'll talk about it again

references:

Jiamingmao.github.io/data-analysis

Zhihu,http://zhihu.com/question/22178202

CSDN.http://blog.csdn.net/saltriver/article/detailes/53056816