What is Information? How to measure it?

Venkaesh Vinayakarao
Chennai Mathematical Institute
venkateshv@cmi.ac.in

1 INTRODUCTION

Shannon argued that information is *uncertainty*. If everything is deterministic, there is no information in it. Consequently, information source is modeled as a random process. Claude Shannon described a point-to-point communication system that has a source, destination and a noisy channel with specific capacity of transmission. Infomation flows from source to destination digitally. His *source coding theorem* introduced entropy as the measure of information. It became the basis for lossless data compression. In another theorem called the *channel coding theorem*, he showed that information can be reliably communicated as along as the rate is less than the capacity. Information theory builds on these fundamental theorems.

2 ENTROPY

We assume all random variables to be discrete. Let X be a random variable whose probability distribution is denoted as $\{Pr(X = x), x \in \mathcal{X}\}$.The support of X denoted as S_X is the set of all $x \in \mathcal{X}$ such that $p(x) > 0$.

The entropy H(X) is defined by

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

Entropy is undefined if p(x) = 0. The base of the log is usually taken as the size of the alphabet $\mathcal{X}$. Here, we restrict our discussion to a binary alphabet $\mathcal{X} = \{0, 1\}$. We assume that all information is transmitted digitally in 0s and 1s. Say, $p(x) = p$, and $\mathcal{X} = \{0, 1\}$, the probability distribution of $\mathcal{X}$ would be $\{p, 1 - p\}$. So,

$$H_2(X) = -plog_2 p - (1 - p) lop_2(1 - p)$$

Notice that if $p = 0$, we have $h_2(0) = h_2(1) = 0$.

3 MAXIMUM ENTROPY

Say, we have a six sided dice.The probability distribution of seeing the faces when the dice is rolled is $p = \{p_1, p_2, ..., p_6\}$. Mathematically, finding the maximum entropy [1] distribution over die faces such that the expected die roll is d corresponds to the following optimization problem:

Maximize H(p) such that:

$$\sum_{i=1}^{6} p_i = 1$$

and

$$\sum_{i=1}^{6} i * p_i = d$$

This is a constrained optimization problem which can be solved using the method of Lagrange multipliers.

4 INFORMATIONAL DIVERGENCE

Let p and q be two probability distributions on a common alphabet $\mathcal{X}$. We are interested in the measure of how much p is different from q. The informational divergence also known as Kullback-Leibler distance

$$D(p||q) = -\sum_{x} plog\frac{p}{q}$$

Note that this is an asymmetic measure [3]. $D(p||q) \neq D(q||p)$.

In many cases, we prefer natural unit of information known as Naperian Digit or nit instead of **b**inary dig**its** called bits.

As an example, consider the following information where $\mathcal{X} = \{0, 1, 2\}$ is given to you.

x	0	1	2
$p(\mathcal{X} = x)$	9/25	12/25	4/25
$q(\mathcal{X} = x)$	1/3	1/3	1/3

The relative entropy of Q to P denoted as $D(p||q)$ is also called the information gain achieved if P would be used instead of Q. $D(p||q) = 0.085$ nats whereas $D(q||p) = 0.097$. Refer to wikipedia[1] for the calculations.

5 APPLICATIONS OF ENTROPY IN IR

The concept of information measures, specifically entropy finds several applications in the area of information retrieval. Aslam et al [1] apply it to compare evaluation metrics. Greiff and Ponte [2] apply it for ranking. The KL divergence is a commonly used measure for comparing query and document language models in the language modeling framework to ad hoc retrieval. We can encode a set of possible events produced with the distribution p using entropy encoding. This is a common idea deployed in compression. We can compress the data by replacing each input symbol with a variable length bit code where the length is determined based on p. For example, the most frequently occurring symbol takes the one bit, say 0. An example for such a scheme is Huffman coding[2]. Thus the messages we encode will have shortest length on average.

REFERENCES

[1] Javed A. Aslam, Emine Yilmaz, and Virgiliu Pavlu. 2005. The Maximum Entropy Method for Analyzing Retrieval Measures. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Salvador, Brazil) *(SIGIR '05)*. Association for Computing Machinery, New York, NY, USA, 27–34. https://doi.org/10.1145/1076034.1076042

[2] Warren R. Greiff and Jay M. Ponte. 2000. The Maximum Entropy Approach and Probabilistic IR Models. *ACM Trans. Inf. Syst.* 18, 3 (July 2000), 246–287. https://doi.org/10.1145/352595.352597

[3] Raymond W. Yeung. 2006. *A First Course in Information Theory (Information Technology: Transmission, Processing and Storage).* Springer-Verlag, Berlin, Heidelberg.

[1]https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence
[2]https://en.wikipedia.org/wiki/Huffman_coding