2021-08-11 15:37

Introduction to Information Theory

--- Author: Alex Pak; a.pak at kbtu.kz ---

Тэги: #edu #infotheory #lec0

Текст

Information theory answes 2 fundamental questions in communication theory, namely:

1. What's the ultimate data compression? - The entropy H.
2. What's the ultimate transmission rate of communcation? - The channel capacity C

Therefore, the data compression rate is $minI(X, X^0)$, where X^0 is an original and X is compressed data. All data compression schemes require description rates at least equal to this minimum. And the data transmission is $maxI(X, Y)$ (known as the *channel capacity*), where X is input data and Y is output data.

For this reason sometimes Information Theory considered as a part of communication theory. Actually it's not true!

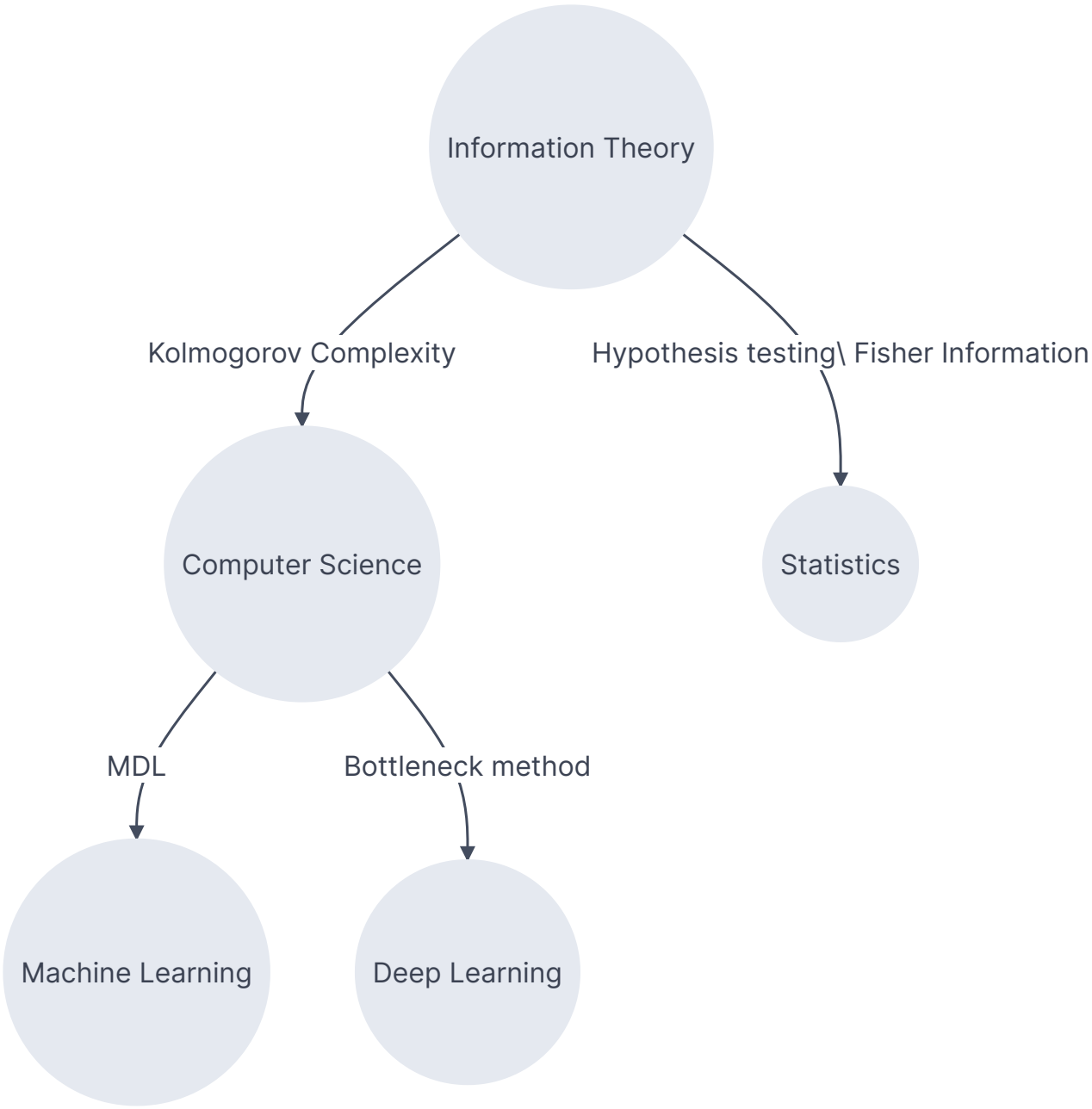

Figure 1. The connection between Information theory and others sci.

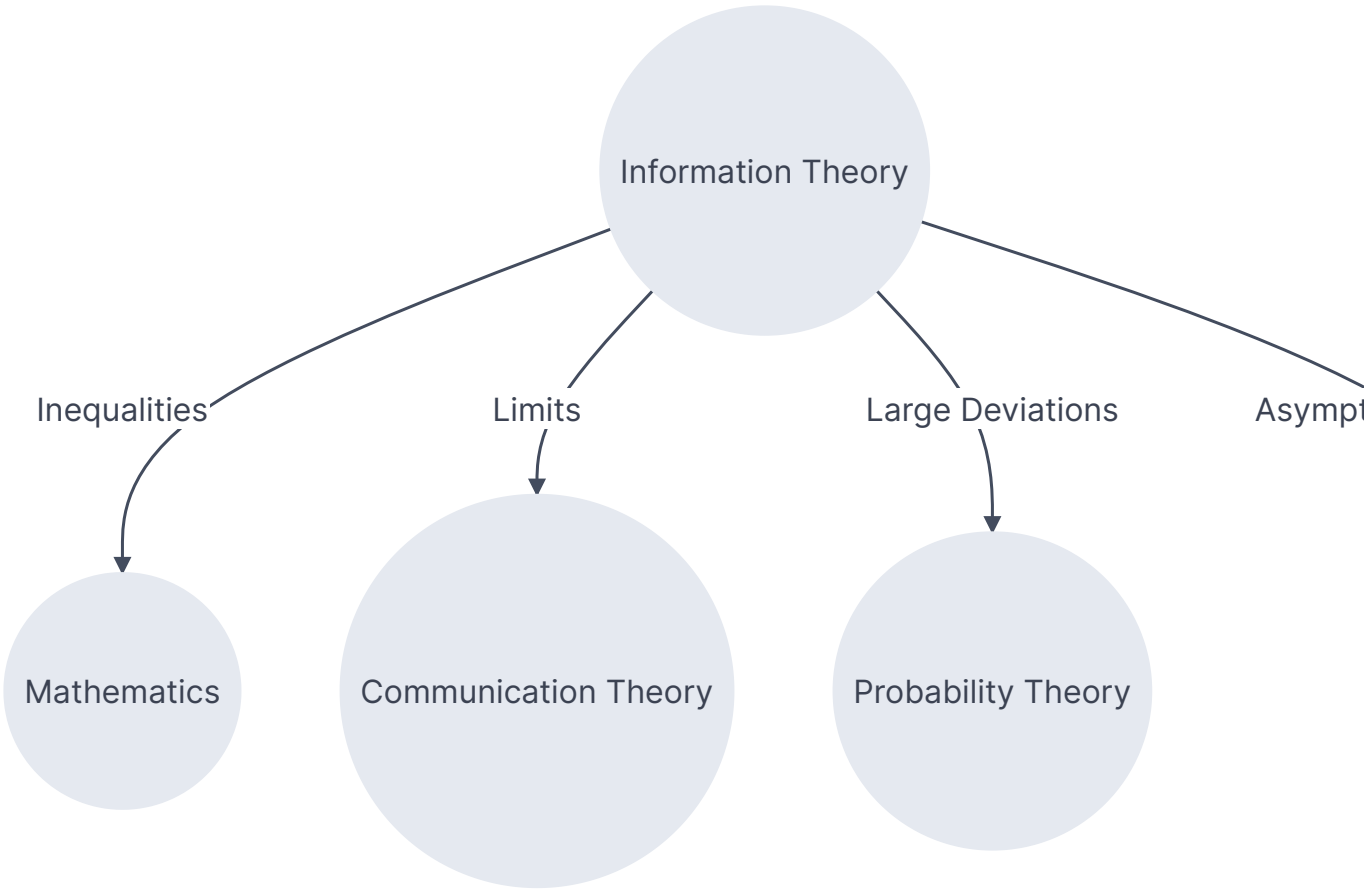

Figure 1b. The connection between Information theory and others sci.

In 1940s at **Communication Theory** there was an opinion that:
It's impossible to send info at a positive rate with zero-like error.

And then Shannon stated:

- the probability of error can be made nearly zero for all communication rates below the channel capacity. The capacity can be computed simply from the noise characterstics of the channel.
- The random process such as music or speech have an irreducible complexity below which the signal cannot be compressed, so called Entropy

---

Computer science (Kolmogorov Complexity)
Kolmogorov, Chaitin and Solomonoff stated that the complexity of a random string can be defined by the length of the simplest binary computer program that generates a particular string. Such idea is pretty universal for various area of computer science. Furthermore, the Kolmogorov complexity K is bound to the Shannon entropy H if the sequence is drawn at random from a distribution that has entropy H.

In addition there is connection between algo (Kolmogorov) and computational complexity (time complexity, so called O-notation). In other words, there are program length and program running time. It's still open research question how to minimize the both of them.

Physics (Thermodynamics). Statistical mechanics is the birthplace of entropy and the second law of thermodynamics. Entropy always increases. Among other things, the second law allows one to dismiss any claims to perpetual motion machines.

Mathematics (Probability Theory and Statistics). The fundamental quantities of information theory—entropy, relative entropy, and mutual information—are defined as functionals of probability distributions. In turn, they characterize the behavior of long sequences of random variables and allow us to estimate the probabilities of rare events (large deviation theory) and to find the best error exponent in hypothesis tests.

Philosophy of Science (Occam's Razor). William of Occam said "Causes shall not be multiplied beyond necessity," or to paraphrase it, "The simplest explanation is best." Solomonoff and Chaitin argued persuasively that one gets a universally good prediction procedure if one takes a weighted combination of all programs that explain the data and observes what they print next.

Economics (Investment). Repeated investment in a stationary stock market results in an exponential growth of wealth. The growth rate of the wealth is a dual of the entropy rate of the stock market. The parallels between the theory of optimal investment in the stock market and information theory are striking. We develop the theory of investment to explore this duality.

Computation vs. Communication. As we build larger computers out of smaller components, we encounter both a computation limit and a communication limit. Computation is communication limited and communication is computation limited. These become intertwined, and thus all of the developments in communication theory via information theory should have a direct impact on the theory of computation.

Entropy

$$H(X) = -\sum_{x} p(x) \log_2 p(x)$$

Entropy is the uncertainty of a single random variable.

I should say a few words about 2 machines

Example 1. Consider a random variable that has a uniform distribution over 32 outcomes. To identify an outcome, we need a label that takes
on 32 different values. H-?

Example 2. Suppose that we have a horse race with eight horses taking part. Assume that the probabilities of winning for the eight horses are (1/2, 1/4, 1/8, 1/16, 1/64, 1/64, 1/64, 1/64) H-?
Binary encoding - ?

Mutual Information

We can define conditional entropy $H(X|Y)$, which is the entropy of a random variable conditional on the knowledge of another random variable. The reduction in uncertainty due to another random variable is called the mutual information. For two random variables X and Y this reduction is the mutual information.

$$I(X;Y) = H(X) - H(X \mid Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

The mutual information $I(X;Y)$ is a measure of the dependence between the two random variables. It is symmetric in X and Y and always non-negative and is equal to zero if and only if X and Y are independent.

Relative Entropy

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

The Relative Entropy is a measure of the “distance” between two probability mass functions p and q.

Communication channel

A communication channel is a system in which the output depends probabilistically on its input. It is characterized by a probability transition matrix $p(Y|X)$ that determines the conditional distribution of the output given the input. For a communication channel with input X and output Y, we can define the capacity C by

$$C = max_{p(x)} I(X;Y)$$

The capacity is the maximum rate at which we can send information over the channel and recover the information at the output with a vanishingly low probability of error. We illustrate this with a few examples.

Figure 2. Noiseless binary channel. C = 1 bit.

Example 3. (Noiseless binary channel) For this channel, the binary input is reproduced exactly at the output. This channel is illustrated in Figure 2. Here, any transmitted bit is received without error. Hence, in each transmission, we can send 1 bit reliably to the receiver, and the capacity is 1 bit. We can also calculate the information capacity $C = max I(X;Y) = 1$ bit.

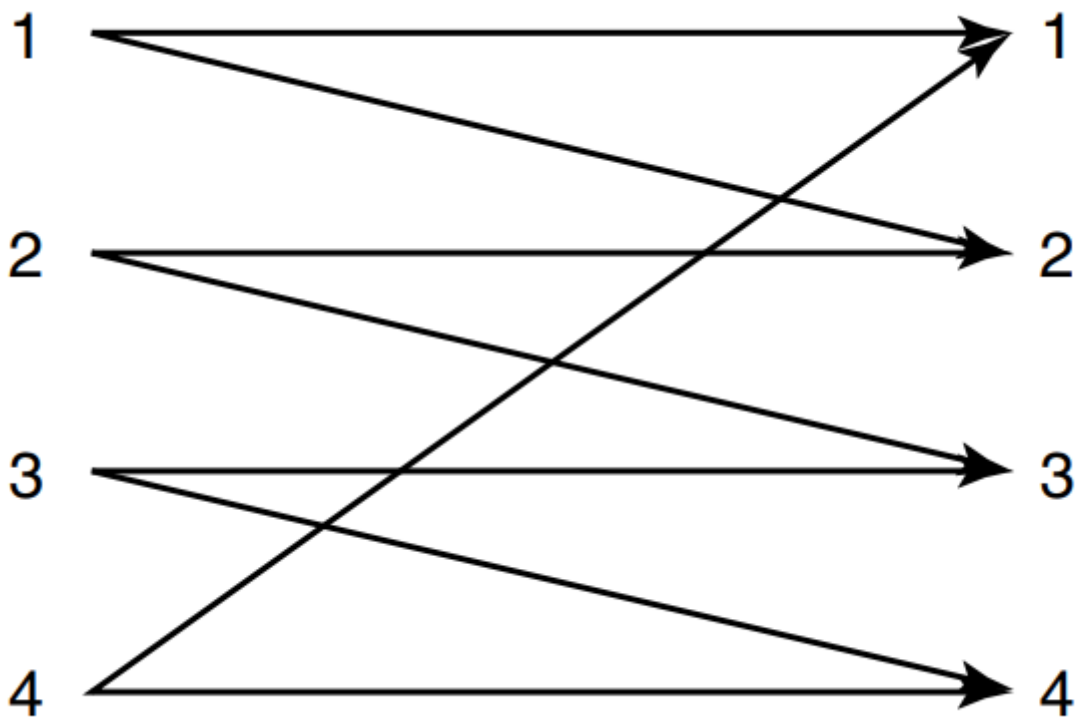

Figure 3. Noisy channel. C = 1 bit.

Example 4 (Noisy four-symbol channel) Consider the channel shown in Figure 3. In this channel, each input letter is received either as the same letter with probability 1/2 or as the next letter with probability 1/2 . If we use all four input symbols, inspection of the output would not reveal with certainty which input symbol was sent. If, on the other hand, we use only two of the inputs (1 and 3, say), we can tell immediately from the output which input symbol was sent. This channel then acts like the noiseless channel of Example 1.1.3, and we can send 1 bit per transmission over this channel with no errors. We can calculate the channel capacity $C = \max I(X; Y)$ in this case, and it is equal to 1 bit per transmission, in agreement with the analysis above.

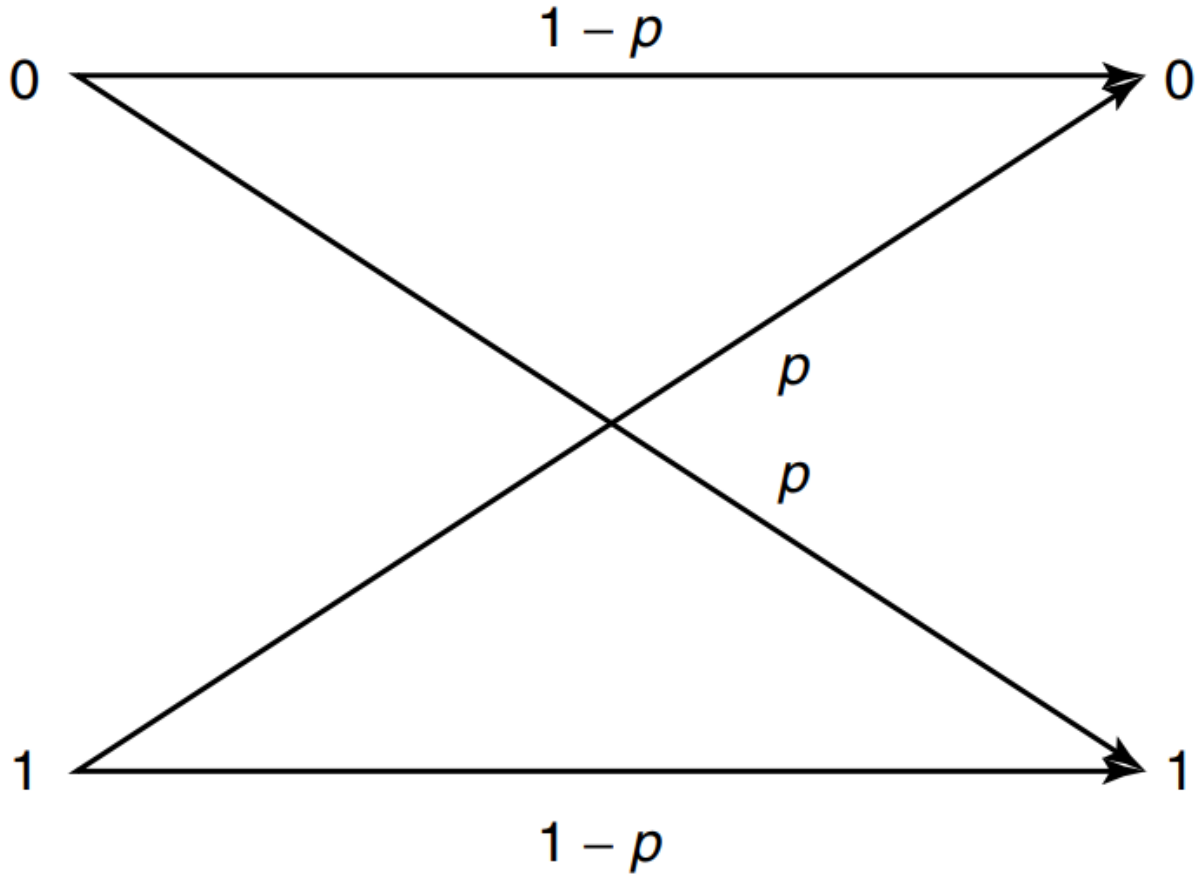

Figure 4. Binary symmetric channel

Example 5. (Binary symmetric channel) This is the basic example of a noisy communication system. The channel has a binary input, and its output is equal to the input with probability $1-p$. With probability p, on the other hand, a 0 is received as a 1, and vice versa. In this case, the capacity of the channel can be calculated to be $C = 1 + plogp + (1-p)log(1-p)$ bits per transmission.

However, it is no longer obvious how one can achieve this capacity. If we use the channel many times, however, the channel begins to look like the noisy four-symbol channel of Example 4, and we can send information at a rate C bits per transmission with an arbitrarily low probability of error.

Conclusion

- Data compression. The entropy H of a random variable is a lower bound on the average length of the shortest description of the random variable.

- Data transmission. The problem of transmitting information so that the receiver can decode the message with a small probability of error.

- Network information theory. Each of the topics mentioned previously involves a single source or a single channel. What if one wishes to compress each of many sources and then put the compressed descriptions together into a joint reconstruction of the sources?

- Ergodic theory. The asymptotic equipartition theorem states that most sample n-sequences of an ergodic process have probability about 2^{-nH} and that there are about 2^{nH} such typical sequences.

- Hypothesis testing. The relative entropy D arises as the exponent in the probability of error in a hypothesis test between two distributions. It is a natural measure of distance between distributions.

- Inference. We can use the notion of Kolmogorov complexity K to find the shortest description of the data and use that as a model to predict what comes next. A model that maximizes the uncertainty or entropy yields the maximum entropy approach to inference.

- Probability theory. The asymptotic equipartition property (AEP) shows that most sequences are typical in that they have a sample entropy close to H. So attention can be restricted to these approximately 2^{nH} typical sequences. In large deviation theory, the probability of a set is approximately 2−nD, where D is the relative entropy distance between the closest element in the set and the true distribution.

- Complexity theory. The Kolmogorov complexity K is a measure of the descriptive complexity of an object. It is related to, but different from, computational complexity, which measures the time or space required for a computation.