Topic-14
Classical Information Theory

Dipan Kumar Ghosh
Indian Institute of Technology Bombay,
Powai, Mumbai 400076

April 15, 2017

1 Classical Information Theory

What is Information? We are all familiar with the word "information" and probably understand what it roughly means. However, in a technological context, we need to give it a precise meaning and also find a way to quantify the amount of information in a message. Suppose we make a statement like "there will not be an earthquake in Mumbai tomorrow". What is the information content of this statement? What additional knowledge it gives us that we did not have? Almost next to nothing because that is the way things are supposed to happen in Mumbai. On the other hand, supposing a statement is made that "there will be an earthquake in Mumbai tomorrow". This will count as a statement with significant amount of information because this event has a low probability of occurrence. Consider the statement "it is cold in Bombay today". What information does this statement convey? Is today cold compared to what it was yesterday or is it as compared to what it was a year before? Is it colder than it is in Alaska or as compared to Chennai? We do know that the statement has some information, though not very precise. Suppose we define a day to be cold if the temperature is less than 22°. Does this statement give more information? Yes, but only partly so. Let us make it a little more precise by saying that "cold" in Bombay means that the temperature T is between 16° and 22°, i.e. $17 < T \leq 22$. Assume that the temperature is measured only in integral values. Our information is better but now there is now an equal probability of the temperature being 17,18, 19, 20, 21 or 22, probability of each is $p = 1/6$. Thus, associated with a statement on information is an "uncertainty". If we try to guess the temperature, say by throwing a die, we can be right about the temperature with a probability of one sixth. Suppose, we are told that yesterday the temperature was 19° and that today is colder than it was yesterday. Our uncertainly has now decreased and we can be right about guessing the temperature by tossing a coin which will give us correct answer with a probability of $1/2$.

Thus the information content in a statement is a measure of uncertainty associated with an event. If we are going to construct a mathematical measure of information contained in a message, we need to specify the amount of uncertainty that is there in a message- the measure of uncertainty cannot be on intuitive ground. Further, we need a measure of this uncertainty so that when a message is transmitted through a channel, we should be able to measure the fidelity of transmission by computing the uncertainty at the receiving end. As the word "information" means different things in different context (for instance, a cognitive scientist may define information as a piece of knowledge), it is important to point out that our approach here is content neutral, we are only interested in problem of data communication and reception. In order to be be able to communicate some information, a system must be capable of being in at least two states These two states could be electrical levels with ground (zero volt) or +5 volts, an atomic system with spin up or down, an ammonia molecule with the nitrogen atom being above or below the plane defined by three hydrogen atoms etc. Mathematically we represent the two state systems by 0 and 1. We assume that we can encode the two states of the system by these two digits and define this as a "bit". (When the digits represent classical two state systems, we occasionally call it a cbit. If instead, we have a system which can exist in m different states, we will need more than a single bit to encode such a system. The minimum number n which satisfies $2^n < m$ is the minimum number of bits required to describe m different states. For instance a 4 state system requires 2 bits to describe it and we can represent the four states by 00, 01, 10 and 11. Hartley proposed the formula for the number of bits n that is required to send m different messages, the formula for the number of bits n that is required to send m different messages,

$$n = \log_2 m$$

Thus a byte (= 8 bits) can send up to 256 different messages. How does one measure information content in a signal? There are two basic approaches. The approach by Claude Shannon, the one that we will follow, defines the amount of information in a signal as the number of bits that needs to be transmitted in order to select it from a list of previously agreed choices. Suppose we want to invite a celebrity to an event out of a few previously agreed list of persons, which consists of Federer, Sania Mirza, Narendra Modi, Hillary Clinton, Bradd Pitt, Deepika Padukone, J. K. Rowling and Vikram Seth. How many questions do we need to ask, which can only be answered in Yes or No to arrive at our decision? Suppose, the number of choices is more, say selecting one candidate out of 8 in an election, the following decision tree can be used.

The second approach is due to Kolmogorov and Chaitin, where they consider the minimum number of bits required to store (compress) a given message. Suppose a source can only emit one of two possible messages. As per Shannon's approach, the message is

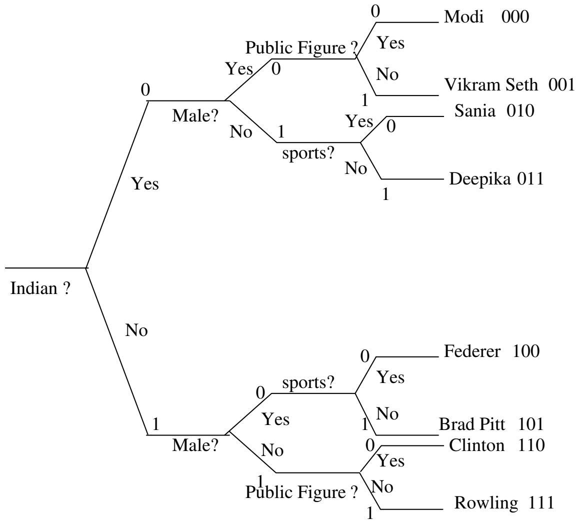

Figure 1: A Decision Tree

identified only by sending one bit of information. However, since the message is of arbitrary length, the corresponding Kolmogorov complexity can be arbitrarily large. Thus whereas Shannon ignores the message itself and only considers the characteristic of the source which emits the message, Kolmogorov's concern is only the message. Kolmogorov's approach is frequently termed as algoirithmic approach and we can define complexity of a message as the length of the shortest algorithm which can describe the message. Let us clarify this by an example. Consider the string which is 200 bit length 010101 . . . 0101. Though the string length is 200 bits, the complexity is quite low because one can transmit it by writing a short program which describes the string as "hundred 01s". On the other hand toss a coin 200 times a generate an arbitrary sequence 0110010110011101 . . .. One cannot write a short program to communicate this exact string by a short program. Thus the complexity is large.

Shannon's Entropy Suppose we want to observe a discrete random variable X which can take finite number of values $x_1, x_2, \ldots, x_M$ with probabilities $p_1, p_2, \ldots, p_M$ respectively with the proviso $\sum p_i = 1$. Let $H(p_1, p_2, ..., p_M)$ be the average uncertainty associated with the event $X = x_i$, i.e., it is the average uncertainty removed when the result of the experiment is known. We would like to obtain a mathematical expression for H_M. Let us first discuss the properties that such a function should satisfy.

1. Suppose $f(M)$ is average uncertainty associated with M equally likely events, i.e.,

$$f(M) = H\left(\frac{1}{M}, \frac{1}{M}, \cdots \frac{1}{M}\right)$$

   This means $f(2)$is the average uncertainty associated with a coin toss, $f(6)$ the uncertainty associated with throw of a die, etc. Clearly $f(M) > f(M')$ if $M > M'$. In otherwards f(M) is a monotonic function of its argument M.

2. Consider two random variables X and Y , the former taking values $x_1, x_2, \ldots x_M$ with equal probability and the latter taking values $y_1, y_2, \ldots y_N$, also with equal probability. The joint experiment XY has MN events with equal probability. The average uncertainty associated with the joint experiment XY is $f(MN)$. Suppose now, the result of X is revealed. This will not remove the uncertainty in Y and we will be left with an uncertainty $f(N)$. Thus, we have $f(MN) - f(M) = f(N)$, so that
$$f(MN) = f(M) + f(N)$$

3. Let us now relax the condition of equal probability. Consider an experiment for measuring a random variable X which has M possible outcomes. The events are divided into two groups A and B, the former having r outcomes $x_1, x_2, \ldots x_r$ and the latter with $M - r$ outcomes $x_{r+1}, x_{r+2}, \ldots x_M$. If the group A is chosen, the outcome of the event $X = x_i$ is $\frac{p_i}{\sum_{i=1}^r p_i}$ and if group B is chosen, the outcome of the event $X = x_i$ is $\frac{p_i}{\sum_{i=r+1}^M p_i}$. Suppose Y is the result of the combined experiment in which we first choose the group (A or B) and then determine the event. Let $Y = x_1$ be the result. We can determine the probability of the combined experiment, by using Baye's theorem,
$$\begin{aligned} P(Y = x_1) &= P(\text{A is chosen \& x}_1 \text{ is selected}) \\ &= P(\text{A is chosen}) \ \times P(x_1 \mid A) \\ &= \sum_{i=1}^r p_i \times \frac{p_1}{\sum_{i=1}^r p_i} \\ &= p_1 \end{aligned}$$
This shows that X and Y have the same distribution. This implies that $P(Y = x_i) = p_i \ \forall i = 1, M$.

Before the compound experiment is performed, the uncertainty associated with the outcome is $H(p_1, p_2, ..., p_M)$. If the group that has been chosen (A or B) is revealed, the uncertainty removed is $H(p_A, p_B) = H\left(\sum_{i=1}^r p_i, \sum_{i=r+1}^M p_i\right)$ so that the uncertainty remaining after the group chosen is revealed is given by
$$H(p_1, p_2, \ldots, P_M) - H\left(\sum_{i=1}^r p_i, \sum_{i=r+1}^M p_i\right)$$

This must be equal to
$$\sum p_i \times H\left(\frac{p_1}{\sum_{i=1}^r p_i}, \frac{p_2}{\sum_{i=1}^r p_i} \ldots, \frac{p_r}{\sum_{i=1}^r p_i}\right) + \sum p_i \times H\left(\frac{p_{r+1}}{\sum_{i=r+1}^M p_i}, \frac{p_{r+2}}{\sum_{i=r+1}^M p_i} \ldots, \frac{p_M}{\sum_{i=r+1}^M p_i}\right)$$

Example:

Suppose the group A has two possible outcomes with $p_1 = 1/2, p_2 = 1/4$ and the group B also has two outcomes with $p_3 = p_4 = 1/8$, we then have,

$$H\left(\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{8}\right) - H\left(\frac{3}{4},\frac{1}{4}\right) = \frac{3}{4}H\left(\frac{2}{3},\frac{1}{3}\right) + \frac{1}{4}\left(\frac{1}{2},\frac{1}{2}\right)$$

The above is known as **Grouping Theorem**. We also require $H(p, 1-p)$ to be a continuous function of p. This is intuitive as a small change in p should lead to a small change in uncertainty. Thus there are four requirement for a function in order to be a measure of uncertainty:

1. $f(M) = H(\frac{1}{M}, \frac{1}{M} \ldots, \frac{1}{M})$ is a non-negative, monotonic and continuously increasing function of M.

2. $f(1) = 0$. This is because, if an event is certain then there is no uncertainty.

3. $f(MN) = f(M) + f(N)$

4. The grouping theorem stated and proved above.

In the next lecture we will find the explicit form of a function which satisfies the above.