Measuring Information

Thomas Krichel

2002–04–15

1 The problem

Our problem is to define a measure of information. To simplify, we start with a simple case.
Let there be a random variable X that can take a finite set of value $x_1, x_2, \ldots, x_M$. Each of the outcomes has probabilities $p_1, p_2, \ldots, x_M$. We assume that $p_i > 0 \; \forall i$ Of course we have

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

Example: X is the result of throwing a dice $x_1 = 1$, $x_2 = 2$, $\ldots x_6 = 6$ and $p_i = 1/6$.
Problem: How much information have we gained when we know the value of X.

2 Strategy to resolve the problem

To solve our problem, we will assume that we have a function that expresses the information that we gain from knowing X. We then look for reasonable properties that this function would have. Then we go out and ask a mathematician to point out useful functions that have this properties.
Let $h(p_1)$ be the information gained from knowing that the value of X is x_i. Let $H(X)$ be the average value of uncertainty removed by knowing that the value of X, i.e.

$$H(X) = \sum_{i=1}^{M} p_i \, h(p_i)$$

Let us first consider the very simple case where all events are equally probable. In that case

$$H(X) = \sum_{i=1}^{M} \frac{1}{M} \, h\left(\frac{1}{M}\right) = f(M)$$

In that simple case $H(X)$ is only a function $f(M)$ of M. If M is 2, we talk about what outcome we have when throwing a coin. If M is 10^8 we talk about picking a person at random in NYC. One idea is that the more outcomes that there are to choose form, the more information is gained by knowing the outcome. This is the first requirement.

Requirement 1 $H(X) = f(M)$ should be an increasing function of M.
Next consider the case where we run two independent experiments X and Y. X that can take a finite set of value $x_1, x_2, \ldots, x_M$, each with probability $1/M$. Y that can take a finite set of value $y_1, y_2, \ldots, y_L$, each with probability $1/L$. The joint experiment has $M\,L$ possible outcomes. Since we assume that both experiments are independent, knowing what value X has taken will not tell us anything about the value of Y. That is, when we remove from the average uncertainty of the joint experiment, $H(L\,M)$ the value of the information gained by knowing x, which is $H(X) = f(M)$, we should still have the information of Y, that is $H(Y) = f(L)$. Therefore

$$f(M\,L) - f(M) = f(L)$$

or in other words

Requirement 2 For two independent experiments that give M and L equiprobable outcomes,

$$f(M\,L) = f(M) + f(L)$$

Next, let us consider that the result of the experiment is revealed gradually and relax the assumption that all outcomes have the same probability. Assume that we group the possible outcomes into two groups. Group A has the events $x_1, x_2, \ldots, x_r$

and group B has the events $x_{r+1}, x_{r+2}, \ldots, x_M$. The probability of choosing an element in group A is

$$p(A) = \sum_{i=1}^{r} p_i$$

and the probability of picking an element in group B is

$$p(B) = \sum_{i=r+1}^{M} p_i$$

Imagine to pick event is chosen, but the information about the chosen element in revealed in two stages. Before the experiment is chosen, the information to be gained by knowing the result is $H(p_1, p_2, \ldots, p_M)$. If we discover the group that is being selected, we gain the amount of information $H(p(a), p(b))$. If group A was chosen, the remaining uncertainty is

$$H\left(\frac{p_1}{p(A)}, \frac{p_2}{p(A)}, \ldots, \frac{p_r}{p(A)}\right)$$

If group B was chosen, the remaining uncertainty is

$$H\left(\frac{p_{r+1}}{p(A)}, \frac{p_{r+2}}{p(A)}, \ldots, \frac{p_M}{p(A)}\right)$$

Thus, on average, the remaining information that will be revealed when the full outcome is announced, after the announcement of the group, will be

$$p_A\, H\left(\frac{p_1}{p(A)}, \frac{p_2}{p(A)}, \ldots, \frac{p_r}{p(A)}\right) + p_B\, H\left(\frac{p_{r+1}}{p(A)}, \frac{p_{r+2}}{p(A)}, \ldots, \frac{p_M}{p(A)}\right)$$

Requirement 3 We require that the full information gained from the revelation of the result, is the information about which group has been chosen, plus the average information to be gained after the group was chosen.

$$H(p_1, \ldots, p_M) = H(p(a), p(b)) + p_A\, H\left(\frac{p_1}{p(A)}, \frac{p_2}{p(A)}, \ldots, \frac{p_r}{p(A)}\right) + p_B\, H\left(\frac{p_{r+1}}{p(A)}, \frac{p_{r+2}}{p(A)}, \ldots, \frac{p_M}{p(A)}\right)$$

Finally, we will make a technical "sanity" requirement.
Requirement 4: $H(p, 1-p)$ should be a continuous function.
There are the requirements. We can now ask mathematicians which functions out there satisfy these requirements.

3 Theorem and interpretation

Theorem The only functions satisfying the requirements 1–4 are or the form

$$H(p_1, \ldots, p_M) = -C \sum_{i=1}^{M} p_i \log(p_i)$$

where C is a positive number and the log is taken to a basis that is larger then one. Remember the definition of the logarithm: if $a^b = c$, then $b = log_a(c)$.
We omit the proof of that theorem.
If we chose the base of the logarithm as 2 and set $C = 1$, we get a special measure of information called the entropy. It is measured in "binary digits" or short, in bits. The resulting quantity is called the entropy.
Let us construct a little table with values of the logarithm to the bases 2

p	p	$log_2(p)$
$3/4$	$2^{-.415}$	$-.415$
$1/2$	2^{-1}	-1
$3/8$	$2^{-1.415}$	-1.415
$1/4$	2^{-2}	-2
$1/8$	2^{-3}	-3
$1/16$	2^{-4}	-4
$1/64$	2^{-6}	-6
$1/256$	2^{-8}	-8

Assume I toss a coin once. What is the information gained?

$$H = -\frac{1}{2}\, log_2\left(\frac{1}{2}\right) - \frac{1}{2}\, log_2\left(\frac{1}{2}\right) = 1$$

Now suppose that the coin is a fake, it shows head with the probability $3/4$ and tail with the probability $1/4$, in that case

$$H = -\frac{1}{4}\, log_2\left(\frac{1}{4}\right) - \frac{3}{4}\, log_2\left(\frac{3}{4}\right) = .811$$

As the probability for head increases and the probability of tail falls, we are more and more certain that head will fall. At the limit, the probability of head becomes 1 and the probability of tail becomes 0, and nothing is learned out of knowing the result.

$$H = -log_2(1) - 0\, log_2(0) = 0$$

Another interpretation goes as follows. Suppose that I have a random variable X that will take the values $1, 2, 3, 4, 5$ with probabilities $0.3, 0.2, 0.2, 0.15, 0.15$, respectively. Imagine a quiz, where the quiz master knows the answer and only answers "yes" or "no".
1st question: "Is it 1 or 2?"
if we get "yes" to first question, second question will be "Is it 1?" and from the answer we will know the result.
if we get "no" from the first question, the we ask
2nd question: "Is it 3?"
if we get "yes" to second question, we have the result. if we get "no" we have to ask a third question.
The average number of questions is

$$(0.3 + 0.2 + 0.2)\, 2 + (0.15 + 0.15)\, 3 = 2.3$$

The entropy of X is 2.27. This is a general result, the number of questions remains above the entropy but it can be quite close.