# Some bits of Information Theory

François Fleuret

January 19, 2024

Information Theory is awesome so here is a TL;DR about Shannon's entropy.

The field is originally about quantifying the amount of "information" contained in a signal and how much can be transmitted under certain conditions.

What makes it awesome is that it is very intuitive, and like thermodynamics in Physics, it gives exact bounds about what is possible or not.

## 1 Shannon's Entropy

Shannon's entropy is the key concept from which everything is defined.

Imagine that you have a distribution of probabilities $p$ on a finite set of symbols, and that you generate a stream of symbols by sampling them one after another independently with that distribution.

To transmit that stream, for instance with bits

over a communication line, you can design a coding that takes into account that the symbols are not all as probable, and decode on the other side.

For instance if $\mathrm{P}('A') = 1/2$, $\mathrm{P}('B') = 1/4$, and $\mathrm{P}('C') = 1/4$ you would transmit "0" for a "A" and "10" for a "B" and "11" for a "C", 1.5 bits on average.

If the symbol is always the same, you transmit nothing, if they are equiprobable you need $\log_2$(nb symbols) etc.

Shannon's Entropy (in base 2) is the minimum number of bits you have to emit on average per symbol to transmit that stream.

It has a simple analytical form:

$$\mathbb{H}(p) = -\sum_k p(k) \log_2 p(k)$$

where by convention $0 \log_2 0 = 0$.

It is often seen as a measure of randomness since the more deterministic the distribution is, the less you have to emit.

The examples above correspond to "Huffman coding", which reaches the Entropy bound only for some distributions. A more sophisticated scheme called "Arithmetic coding" does it always.

From this perspective, many quantities have an

intuitive value. Consider for instance sending pairs of symbols $(X, Y)$.

If these two symbols are independent, you cannot do better than sending one and the other separately, hence

$$\mathbb{H}(X, Y) = \mathbb{H}(X) + \mathbb{H}(Y).$$

However, imagine that the second symbol is a function of the first Y=f(X). You just have to send $X$ since $Y$ can be computed from it on the other side.

Hence in that case

$$\mathbb{H}(X, Y) = \mathbb{H}(X).$$

An associated quantity is the mutual information between two random variables, defined with

$$\mathbb{I}(X; Y) = \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X, Y),$$

that quantifies the amount of information shared by the two variables.

## 2 Conditional Entropy

Conditional entropy is the average of the entropy of the conditional distribution:

$$\mathbb{H}(X \mid Y) = \sum_y \mathrm{P}(Y = y)\mathbb{H}(X \mid Y = y)$$

with

$$\mathbb{H}(X \mid Y = y) = \sum_x \mathrm{P}(X = x \mid Y = y) \log \mathrm{P}(X = x \mid Y = y)$$

Intuitively it is the [minimum average] number of bits required to describe $X$ given that $Y$ is known.

So in particular, if $X$ and $Y$ are independent, getting the value of $Y$ does not help at all, so you still have to send all the bits for $X$, hence

$$\mathbb{H}(X \mid Y) = \mathbb{H}(X),$$

and if $X$ is a deterministic function of $Y$ then

$$\mathbb{H}(X \mid Y) = 0.$$

And if you send the bits for $Y$ and then the bits to describe $X$ given that $Y$, you have sent $(X,Y)$, hence the chain rule:

$$\mathbb{H}(X,Y) = \mathbb{H}(Y) + \mathbb{H}(X \mid Y).$$

And then we get

$$\begin{aligned} I(X;Y) &= \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y) \\ &= \mathbb{H}(X) + \mathbb{H}(Y) - (\mathbb{H}(Y) + \mathbb{H}(X \mid Y)) \\ &= \mathbb{H}(X) - \mathbb{H}(X \mid Y). \end{aligned}$$

# 3 Kullback-Leibler divergence

Imagine that you encode your stream thinking it comes from distribution $q$ while it comes from $p$. You would emit more bits than the optimal $\mathbb{H}(p)$, and that excess of bits is $\mathbb{D}_{\mathsf{KL}}(p||q)$ the Kullback-Leibler divergence between $p$ and $q$.

In particular if $p = q$

$$\mathbb{D}_{\mathsf{KL}}(p\|q) = 0,$$

and if there is a symbol $x$ with $q(x) = 0$ and $p(x) > 0$, you cannot encode it and

$$\mathbb{D}_{\mathsf{KL}}(p\|q) = +\infty.$$

Its formal expression is

$$\mathbb{D}_{\mathsf{KL}}(p\|q) = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

that can be understood as a value called the cross-entropy between $p$ and $q$

$$\mathbb{H}(p,q) = -\sum_x p(x) \log q(x)$$

minus the entropy of p

$$\mathbb{H}(p) = -\sum_x p(x) \log p(x).$$

Notation horror: if $X$ and $Y$ are random variables $\mathbb{H}(X,Y)$ is the entropy of their joint law, and if $p$ and $q$ are distributions, $\mathbb{H}(p,q)$ is the cross-entropy between them.