Numerical Information Theory

Cesare Magri

May 10, 2014

Introduction

Back in 2008 I released a small software package for the estimation of the mutual-information from data. This toolbox (essentially a collection of software-tools that I was using for my research work) attracted quite some interest within the neuroscience community to which it was aimed.

Receiving feedback from users all around the world, I started to realize how hard it was for people without a strong background in statistics and information-theory to apply numerical-information-theory techniques to actual data. The problem, however, did not lie with the users, but rather with the required information being scattered over a multitude of extremely technical papers comprehensible only to few specialized scholars.

However, it is my belief that the fundamental concepts behind numerical-information-theory techniques and their applications can be presented in an intuitive, easy-to-understand way accessible also to non-experts. Everyone can learn to apply these methods correctly and analyze critically their results without having to attain a degree in statistics.

This guide aims at providing a straightforward introduction to the fundamental concepts of numerical-information-theory with tips-and-tricks for the estimation of the information quantities from actual data. The guide also serves as an step-by-step manual for the new toolbox (the `InfoToolbox`), a software suite that aims at providing an exhaustive and unified set of user-friendly numerical information-theory tools for a variety of computational environment.

Chapter 1

Entropy

1.1 Measuring uncertainty

Suppose you were asked to forecast tomorrow's weather. In particular, you might be asked to bet on whether tomorrow the weather will be warm. At the time when I am writing, it is summer here in Frankfurt. Last week it has been the hottest week of the year, almost unbearably warm, with temperatures reaching 35°C. Not bad for Germany. I really expect tomorrow to be warm too. Maybe not as warm as today but definitely not cold. If I were asked to bet today, I would quite confidently say "*tomorrow the weather is going to be nice and warm*!" and I would bet my money on it[1].

What have I based my forecast on? I have used the current available data about today's weather to try estimating how likely it is that tomorrow it will be warm. Since the probability of this outcome appears high, I have decided to bet on it. Just a couple of weeks ago, when rain and sun were switching daily, I would not have been so confident in my prediction.

The process that I have just described sounds quite intuitive and we can summarize it as follows: *The higher the probability of an event, the more confident we are in predicting it, and vice-versa.*

Before continuing let me modify the above statement only slightly. The fact is that statisticians do not like to speak of *more confidence*. They rather prefer to say *less uncertainty*, which has exactly the same meaning but hints to a greater modesty. We will thus conform to this rule and rephrase the above concept in terms of uncertainty: *The higher the probability of an event, the less uncertain we are in predicting it, and vice-versa.*

Now let's take another small step, and let's express our concept with a very simple formula. If x is the event that we want to predict (in our example, the event "*tomorrow it will be warm*") and $P(x)$ be the probability that x will occur, we can define our uncertainty, $U(x)$, in predicting x simply the inverse of $P(x)$

$$U(x) = \frac{1}{P(x)}. \tag{1.1}$$

While this simple formula captures our intuition regarding the relationship between probability and uncertainty, in that the uncertainty decreases with increasing probability, it does not, however, really qualify as a good measure of uncertainty.

1.1.1 A unit of measurement

First of all, an event that we are absolute sure about, that is an event which has a 100% probability of occurring, should have an uncertainty of zero because we have absolutely no doubt that this event will occur. However, for $P(x) = 1$ equation (1.1) returns $U(x) = 1$, which is not really what we were expecting.

[1] In case you are curious to know, the day after I wrote this paragraph temperatures dropped by 10 degrees Celsius and it rained. However, with a cozy 22°C, I could still say that I won my bet.

An additional question arises when trying to interpret the results computed through (1.1): what do these value actually mean? For example, the probability of the coveted event *winning the lottery*" is something around one in seventeen million. If we substitute $P(\mathsf{winning\ the\ lottery}) = 1/17\,000\,000$ in (1.1) we obtain, again, seventeen million, which is a very large number but does not provide us with much additional information.

This is because measuring something requires comparing it to a standard unit that we are familiar with. For example, if I tell you that the temperature tomorrow is going to be high, this will not give you a very precise idea of how warm it is going to be. However, if I tell you that tomorrow it is going to be 35°C, then you immediately get a sense of how warm tomorrow will be since I am quantifying tomorrow's temperature in terms of a standard unit, 1°C, that you are familiar with.

Similarly, if we want to measure the uncertainty of an event, we need to compare this event with some fixed simple outcome that everyone is familiar with and the uncertainty of which we can intuitively grasp. In statistics, this unit it the king of all random events, one of the simplest examples one can think of, namely, the toss of a fair coin.

Flipping a head when tossing a fair coin has a probability $P(\mathsf{head}) = 1/2$ of occurring. How does an event x with probability $P(x) = 1/4$ compare to it? Since flipping two heads when tossing *two* coins also has probability 1/4, we could say that the uncertainty of an event with probability 1/4 is twice that of a single coin toss. Similarly, an event with $P(x) = 1/8$ will have an uncertainty which is three-times that of a single coin toss, since 1/8 is the probability of flipping three heads when tossing *three* coins at the same time. We start to notice a relationship.

If we want to include these observations in our measure of uncertainty we need to modify (1.1) to include the logarithm in base 2, as follows:

$$H(x) = \log_2\left(\frac{1}{P(x)}\right) \tag{1.2}$$

This way we obtain $H(x) = 2$ for an event with $P(x) = 1/4$ and $H(x) = 3$ when $P(x) = 1/8$, which is indeed what we computed above. We also obtain $H(x) = 0$ for $P(x) = 1$, which is the result we were looking for the case of no uncertainty.

We will say that $H(x)$ is the uncertainty[2] associated to an event x and we will measure it in *bits*, the name of the unit originating from the use of base 2 in informatics.

We can now use (1.2) to get an actual sense of the huge uncertainty associated to the event "*winning the lottery*". If we substitute $P(\mathsf{winning\ the\ lottery}) = 1/17\,000\,000$ in (1.2) we obtain $H(x) \approx 24$ bits: Winning the lottery is as unlikely as tossing 24 coins at the same time and having them all 24 flip a head!

[2] $H(x)$ is also called the *self-information* of the event x. We will, however, refrain from using this nomenclature to avoid confusing this quantity with the mutual information introduced in later chapters.

1.1.2 Toolbox Example

You can compute the above example using Matlab and the InfoToolbox by running the following script:

```
p = 1/17000000;     % Assingn the probability
h = aentrdiscr(p); % Compute the uncertainty
disp(['Uncertainty of winning the lottery: ' num2str(h) ' bits']);
```