About some properties of the Kullback-Leibler divergence

Angel Garrido, Facultad de Ciencias de la UNED

Abstract

Our paper analyzes some aspects of of a very important Uncertainty Measure, one that belongs to the so-called Entropy; more concretely, the Kullback-Leibler divergence measure. We need to obtain new ways to model adequate conditions or restrictions, constructed from vague pieces of information. For this, it will be very necessary to analyze different type of measures; in particular, to consider these fuzzy measures.

Keywords: Mathematical Analysis, Measure Theory, Fuzzy Measures, Information Theory, Entropy.

Mathematics Subject Classification: 68R10, 68R05, 05C78, 78M35.

1. Introduction to Entropy and Information

Entropy and related information measures provide descriptions of the long term behavior of random processes, and that this behavior is a key factor in developing the Coding Theorems of Information Theory (IT, in acronym).

The contributions of *Andrei N. Kolmogorov* (1903-1987) to the mathematical IT produces great advances to the Shannon formulations, proposing a new complexity theory, now translated to Computer Sciences. According such theory, the complexity of a message is given by the size of the program necessary to be possible the reception of such message.

From these ideas, Kolmogorov also analyzes the entropy of literary texts. More concretely, on Pushkin poetry. Such entropy appears as a function of the semantic capacity of the texts, depending of factors as their extension and also the flexibility of the corresponding language.

Also may be mentioned *Norbert Wiener* (1894-1964), considered the founder of Cybernetics, who in 1948 also propose a similar vision of such problem.

But the approach used by Shannon differs from that of Wiener in the nature of the transmitted signal and in the type of decision made at the receiver.

In the Shannon model messages are first encoded and then transmitted, whereas in the Wiener model the signal is communicated directly through the channel without being encoded.

The initial studies on IT were undertaken by *Harry Nyquist* (1889-1976) in 1924. And later by *Ralph Hartley* (1888-1970), who in 1928 recognized the

[1] AMO - Advanced Modeling and Optimization. ISSN: 1841-4311

logarithmic nature of the measure of information. Later, it appears the key, with the essential *Shannon* and *Wiener* papers.

About some apparent "evidences" prescribing that the Shannon information measure is the only possible one, it must be clear that it will be only valid within the more restricted scope of coding problems which the own C. E. Shannon had see in his time. As pointed out by *Alfred Rényi* (1961), in his essential paper on generalized information measures, in other sort of problems other quantities may serve just as well, or even better, as measures of information. This should be supported either by their operational significance or by a set of natural postulates characterizing them, or, preferably, by both. Thus, the idea of generalized entropies arises in the scientific literature.

The name of Entropy proceeds indeed from the resemblance between Shannon´s formula and some similar formulae which are usual in Thermodynamics. So, in Statistical Thermodynamics we will take the Gibbs Entropy. The standard Boltzmann-Gibbs entropy may be generalized to the so-called *Tsallis Entropy* (1988). It is also possible to translate the Gibss Entropy to Quantum Physics, giving us the *Von Neumann Entropy (1927).*

2. Kullback-Leibler divergence

We will define the *Relative or Differential Entropy.* It will be also called with many other different names, as *Kullback-Leibler* (1951) *"distance"* (pseudo-distance, indeed), or *divergence K-L,* either *relative entropy,* or *information gain.* It is denoted by D_{KL}.

Given two probability distributions, p and q, it will be defined by

$$D_{KL}\left(p \shortparallel q\right) = \sum_{x \in X} p\left(x\right) \log_2 \left(\frac{p(x)}{q(x)}\right) = E_{p(x)}\left[\log_2 \frac{p(x)}{q(x)}\right]$$

A very essential property of D will be that the K-L divergence is always non-negative, i.e.

$$D_{KL}\left(p \shortparallel q\right) \geq 0$$

The equality is reached when both distribution coincides, i.e. $p\left(x\right) = q\left(x\right), \forall x$.

But note that in general,

$$D\left(p \shortparallel q\right) \neq D\left(q \shortparallel p\right)$$

Therefore, it does not symmetrical. Neither verifies the triangular inequality. So, it is not really a metric, but a premetric. Hence, it specifies a topology. Furthermore, such topology strictly dominates the topology of total variation, due to the well-known inequality of Pinsker.

It report us the measure of inefficiency when supposing q as the "true", or correct, distribution, being so indeed p.

3. Mutual Information

The *mutual info of X on Y* is the measure of the info which X has on Y. Usually, it is denoted by

$$I\ (X;\ Y)$$

If we write instead *I (Y; X),* we have the info which Y posess on X. But they give us the same value; so, it is *a symmetrical measure.*

The *relationship between mutual info and entropy* is

$$I\left(X;Y\right)=H\left(X\right)-H\left(X/Y\right)=H\left(Y\right)-H\left(Y/X\right)=I\left(Y;X\right)$$

And also

$$I\left(X;Y\right)=H\left(X\right)+H\left(Y\right)-H\left(X,Y\right)$$

Therefore,

$$I\left(X;X\right)=H\left(X\right)$$
$$I\left(Y;Y\right)=H\left(Y\right)$$

In general, conditioning a random variable on another, we reduce the uncertainty of the last variable

$$H\left(Y/X\right)\leq H\left(X\right)$$
$$H\left(X/Y\right)\leq H\left(Y\right)$$

It is possible to generalise the *Chain Rule for n variables*

$$H\left(X_1,H_2,...,X_n\right)=\sum_{i=1}^{n}H\left(X_i/X_{i-1},X_{i-2},...,X_1\right)$$

And therefore, in the conditional case

$$H\left(X_1,H_2,...,X_n/Y\right)=\sum_{i=1}^{n}H\left(X_i/X_{i-1},X_{i-2},...,X_1,Y\right)$$

4. Generalizing the Kullback-Leibler divergence

It is possible to generalize the K-L divergence.
In the *discrete case,* we have

$$D_{_{KL}}\left(P\ ||\ Q\right)=\sum_{i}p\left(i\right)\log_2\left(\tfrac{P(i)}{Q(i)}\right)$$

Whereas in the *continuum case*

$$D_{KL}\left(P \mid\mid Q\right) = \int_{-\infty}^{+\infty} p\left(x\right) \log_2 \left(\frac{p(x)}{q(x)}\right) dx$$

Being p and q the density functions corresponding to both, P and Q distributions.

Let P and Q be two probability measures over a set X, and Q is absolutely continuous w.r.t. P, then the *K-L div from P to Q* is given by

$$D_{KL}\left(P \mid\mid Q\right) = -\int_X \log \left(\frac{dQ}{dP}\right) dP$$

Analogously, if P is absolutely continuous w.r.t. Q, then it holds

$$D_{KL}\left(P \mid\mid Q\right) = \int_X \log \left(\frac{dQ}{dP}\right) dP = \int_X \frac{dP}{dQ} \log \left(\frac{dP}{dQ}\right) dQ$$

Let

$$dP = pd\mu$$
$$dQ = qd\mu$$

be two probability measures, on the set X, such that they are absolutely continuous with respect to the measure.

Then, we define the *divergence of Kullback-Leibler, or K-L* (if such integral exist) as

$$D_{KL}\left(P \mid\mid Q\right) = \int_X p \log \left(\frac{p}{q}\right) d\mu$$

where

$$\frac{p}{q} = \frac{dP}{dQ}$$

is the *Radon-Nikodym derivative* of P with respect to Q. Then, the final expression should be independent of measure μ.

Given two joint probability mass unctions, *p(x/y)* and *q(x/y),* the Conditional Relative Entropy between them may be denoted by

$$D_{KL}\left(p\left(y/x\right) \parallel q\left(y/x\right)\right)$$

It will be the average on the relative entropiesbetween the conditional probability mass functions, *p(x/y)* and *q(x/y),* averaged over the probability mass function, *p(x).*

I.e.

$$D_{KL}\left(p\left(y/x\right) \parallel q\left(y/x\right)\right) = \sum_{x} p\left(x\right) \sum_{y} p\left(y/x\right) \log \frac{p(y/x)}{q(y/x)} = E_{p(x,y)}\left[\log_2 \frac{p(Y/X)}{q(Y/X)}\right]$$

So, the *Chain Rule for Relative Entropy* can be expressed as

$$D_{KL}\left(p\left(y/x\right) \parallel q\left(y/x\right)\right) = D_{KL}\left(p\left(y/x\right) \parallel q\left(y/x\right)\right) + D_{KL}\left(p\left(x\right) \parallel q\left(x\right)\right)$$

Some other interesting measures of divergence.

For instance, we have the *symmetrized distance*

$$D\left(P \shortparallel Q\right) + D\left(Q \shortparallel P\right)$$

It will be very useful, for instance, in Feature Selection, into Classification Problems.

An alternative distance is the $\lambda -$ div *(lambda divergence),*

$$D_{\lambda}\left(P \shortparallel Q\right) = \lambda D\left[P \shortparallel \lambda P + \left(1-\lambda\right) Q\right] + \left(1-\lambda\right) D\left[Q \shortparallel \lambda P + \left(1-\lambda\right) Q\right]$$

This signifies the gaining expectation of info about that X is obtained from P or Q, with respective probabilities p and q.

In particular, when $\lambda = 1/2$, we found the *Jensen-Shannon divergence*

$$D_{JS}\left(P \shortparallel Q\right) = \tfrac{1}{2} D\left(P \shortparallel M\right) + \tfrac{1}{2} D\left(Q \shortparallel M\right)$$

Where M is the promediate value of probability distributions P and Q.

This *divergence of Jensen-Shannon* can be interpreted as the capability of a noisy channel of info with two entries and giving as output the probability distributions P and Q.

5. Concavity and Convexity. On Jensen Inequality

A very important ineqation shows interesting consequences. It is the so-called *Jensen inequality.*

But previously, we may recall the definitions of convex/concave function.

A function, f(x), is *convex,* over an interval (a,b),if for every $u, v \in (a,b)$, and $\lambda \in [0,1]$, we have

$$f\left[\lambda u + \left(1-\lambda\right) v\right] \leq \lambda\ f\left(u\right) + \left(1-\lambda\right)\ f\left(v\right)$$

And it is said to be *strictly convex,* if the equality holds only if $\lambda = 0$ or $\lambda = 1$.

A function, f(x), is *concave,* over an interval (a, b),if for every $u, v \in (a, b)$, and $\lambda \in [0, 1]$, we have

$$f\ [\lambda u + (1 - \lambda)\, v] \geq \lambda\ f\,(u) + (1 - \lambda)\ \ f\,(v)$$

I.e. f is concave, if $-f$ is convex.

Therefore, f is convex, if it always lies below any chord. And f is concave, if it always lies above any chord.

Passing to the Jensen Inequality, we can say that if f is a convex function, and X a random variable, then

$$E\,[f\,(X)] \geq f\,(E\,[X])$$

And in the particular case of a strictly convex function f, from the equality in the Jensen Inequality we may deduce that

$$X = E\,[X]$$

with probability equal to one. Therefore, X will be a constant.

$D_{KL}\ (p \parallel q)$ is convex in the pair $(p \parallel q)$,i.e. if (p_1, q_1) and (p_2, q_2) both are pairs of probability mass functions, then

$D_{KL}\,(\lambda p_1 + (1 - \lambda)\, p_2 \parallel \lambda q_1 + (1 - \lambda)\, q_2) \leq \lambda D_{KL}\,(p_1 \parallel q_1) + (1 - \lambda)\, D_{KL}\,(p_2 \parallel q_2)$

for each $\lambda \in [0, 1]$.

But in the case of entropy the situation is reversed, because

$H\,(p)$ is a *concave* function of the probability distribution, p

We have a result connected with them, but now about the *Mutual Information* (denoted by *MI,* in acronym).

Let *(X, Y)* be a joint probability distribution with

$$p\,(x, y) = p\,(x)\, p\,(y/x)$$

Then, the *Mutual Information on X on Y,*

$I\,(X; Y)$ is a concave function of p(x), for fixed p(y/x)

and

$I\,(X; Y)$ is a convex function of p(y/x), for fixed p(x).

6. Generalizing the Entropy

And generalizing, from the Shannon Entropy measure, we can found the *Rényi Entropy,* or Entropy due to Alfred Rényi.

Let be a random sample, $\{x_i\}_{i=1}^n$, with probabilities $\{p_i\}_{i=1}^n$.

We will define the *Rényi´s Entropy* as

$$H_\alpha(X) = \frac{1}{1-\alpha} \log\left(\sum_{i=1}^{n} p_i^\alpha\right)$$

If they are equal all the above probabilities, then

$$H_\alpha(X) = \log n, \ \forall\alpha$$

The entropies, as functions of α, are weakly decreasing.

So, for instance,

$$H_0(X) \geq H_1(X) \geq H_2(X) \geq ... \geq H_\infty(X)$$

A particular case should be the *Hartley´s entropy,*

$$If \ \alpha = 0, then \ H_0(X) = \log \ n\,(\log\,[card\,(X)])$$

There exists these relation between entropies

$$H_\infty < H_2 < 2H_\infty$$

Furthermore, the *Generalyzed Divergence of Rényi,* of order α, of a distribution Q, relative to P, the *"authentic",* will be

$$D_\alpha(P \,||\, Q) = \frac{1}{\alpha-1} \log\left(\sum_{i=1}^{n} \frac{p_i^\alpha}{q_i^{i-1}}\right) = \frac{1}{\alpha-1} \log\left(\sum_{i=1}^{n} p_i^\alpha q_i^{1-i}\right)$$

So, we have

$$D_\alpha(P \,||\, Q) \geq 0, \forall P, Q$$

Conclusions

Statistical entropy is a probabilistic measure of uncertainty, or ignorance about; whereas, *Information* is a measure of a reduction in that uncertainty. For this, we must to ignore particular features of such event, only observing whether or not it happened. So, we can consider the event as the observance of a symbol whose probability of occurring is p. Whereas the Entropy of a probability distribution is just the expected value of the information of such distribution.

References

De Luca, and Termini: "A definition of non-probabilistic entropy in the setting of Fuzzy Set theory". *Inform. and Control*, Vol. 20, pp. 301-312, 1972.

Dubois and Prade: *Fundamentals of Fuzzy Sets.* Series: The Handbooks of Fuzzy Sets. Vol.7. Springer-Kluwer, 2000.

Garmendia: "The Evolution of the Concept of Fuzzy Measure". *Studies in Computational Intelligence,* Vol. 5, pp. 186-200, 2005.

Garrido: "Additivity and Monotonicity in Fuzzy Measures". Plenary Talk in ICMI45, at Bacau University, then published at SCSSM journal, Vol. 16, pp. 445-458, 2006.

Jaynes: *Probability Theory: The Logic of Science.* Cambridge University Press, 2003.

Jeffreys: *Theory of probability.* Oxford University Press, 1948.

Kullback: *Information Theory and Statistics.* Wiley, 1959.

Lambert: "Entropy is Simple, Qualitatively", *Journal of Chemical Ed.,* Vol. 79, pp. 1241-1246, 2002 [it is also disposable online].

Rényi: "On measures of information and entropy". *Proc. of the 4th Berkeley Symposium on Mathematics, Statistics and Probability,* pp. 547-561, 1961.

Shannon: "A Mathematical Theory of Communication", Bell System Technical Journal, pp. 379-423 and 623-656, 1948. Later, it appears as book, at Illinois Press, 1963. Also E-print.

Tribus, and Irvine: "Energy and Information". *Scientific American* 225, No. 3: 179-188, 1971.

Volkenstein: *Entropy and Information.* Series: Progress in Mathematical Physics, Vol. 57, Birkhäuser Verlag, 2009.

Wang, and Klir: *Fuzzy Measure Theory.* Plenum Press, New York, 1992.

Wang, and Klir: *Generalized Measure Theory.* Springer Verlag, Berlin-New York, 2008.