Raping the Likelihood Principle

Carlos C. Rodríguez

August 30, 2006

Abstract

Information Geometry brings a new level of objectivity to bayesian inference and resolves the paradoxes related to the so called Likelihood Principle.

Introducing the Bayesian Acolyte

The acolyte has come from far away to learn from the famous wise men of modern inference. At the time of this tale, the subject of statistics is in the middle of a tumultuous transformation with statisticians metamorphosing away from bookkeepers of statistical tables into bookkeepers of all human knowledge. There is a revision of the foundations of everything and the Likelihood Principle is at the center of it. We find the acolyte trying to explain the problem and his solution to a small group of friends...

Two Descriptions of a coin flip

Consider the classical canonical experiment of flipping a coin (a U.S. quarter) with unknown probability of tails $t \in [0,1]$. We provide two descriptions, A, and B, of the outcomes of the same coin flip. Description A is the usual one. The outcomes are either tails, $x = 0$ or heads, $x = 1$. The alternative description B, also labels tails with $x = 0$, but when heads come up, description B labels the outcome with either $x = 1$ or $x = 2$ depending on the direction of Washington's head w.r.t. fix cartesian coordinate directions on the floor, with origin right on the center of the coin. Thus, if an imaginary arrow, going from Washington's neck to the top of his head, points westwards (i.e. has a positive east-west component) we label the outcome of heads with $x = 1$ and if it points eastwards (i.e., negative east-west component) we label it with $x = 2$.

Let us suppose that we are somehow able to flip the coin in such a way that the two likelihoods, defined for $x \in R$ and $t \in [0,1]$ are,

$$p(x|t, A) = t\delta(x) + (1-t)\delta(x-1)$$

and,

$$p(x|t,B) = t\delta(x) + t(1-t)\delta(x-1) + (1-t)^2\delta(x-2)$$

The LP's Sex Appeal

Our illustrious bayesian forefathers (no known moms out there...) have warned us about the perils of using priors that depend on anyway on the likelihood. These, we are told, will open Pandora's box of paradox.

To begin with, it is claimed, it is obviously schizophrenic to assign different prior probabilities to the exact same parameter $t = p(x = 0|A) = p(x = 0|B)$. For it is the same coin, the same observed tails under both descriptions. Nevertheless, suppose that bad influences push you to use different priors, $p(t|A)$ and $p(t|B)$ for the parameter t in the two descriptions above and that you flip the coin and you get tails, i.e., $x = 0$. Now, we are told, you could easily expose your mistake by just computing the posterior probabilities for A and for B. So, you go ahead and start computing: Shaking the bayes wand (i.e., applying bayes theorem) obtain,

$$p(A|x) = \frac{p(A)}{p(x)} \int_0^1 p(x|t,A)p(t|A)dt$$

and,

$$p(B|x) = \frac{p(B)}{p(x)} \int_0^1 p(x|t,B)p(t|B)dt.$$

At this point of your calculation you hit a small wall. To get numbers, at least for the ratio $p(A|x)/p(B|x)$, you would need the prior probabilities for the descriptions A and B that no one gave you. You ask the forefathers for help but here the forefathers run out of magic things to shake and start shaking themselves instead, claiming "ignorance" or an appeal to the principle of insufficient reason or plainly telling you to assume $p(A) = p(B)$ to make them (puff!) disappear. You do just that and replace $x = 0$ in the above formulas to obtain,

$$\frac{p(A|x=0)}{p(B|x=0)} = \frac{\int_0^1 tp(t|A)dt}{\int_0^1 tp(t|B)dt}$$

Hmm, you say. I see, if the two different priors happen to give different expected values for t then that would mean that the observation of tails (i.e., $x = 0$) gives some evidence in favor of one of the two descriptions... And the forefathers quickly add: Yeah, and that's crazy! For, there is nothing in the coin, nor in the way you flipped it, that could distinguish between descriptions A and B. PERIOD.

The Acolyte Learns Geometry

Before coming to this country I studied a fair amount of math and physics. I knew about Hausdorff and Banach and Hilbert spaces and I was supposed to also know about riemannian manifolds but later I realized that I didn't. Why do I tell you this?. Because I think you may find my experience useful.

To continue... the concept of manifold, and with it the importance of non euclidean geometries, hit me unexpectedly one morning on a train along the Hudson river. I ow it to Dubrovin, Fommenko and Novikov volume 1. I realized then, that it was possible to study shapes, volumes, curvatures, intrinsically, without reference to any outside, and of objects made out of abstract things that we still call points but that could be very different from the familiar points of physical space. Shun-ichi Amari's lecture notes on Differential Geometry revealed to me that these objects could be collections of probability distributions, in fact, most of the hypothesis spaces in statistical inference were riemannian manifolds. More over, it turns out, the intrinsic geometries of regular hypothesis spaces are fixed by simple consistency requirements and the only admissible metric is proportional to Fisher information. This is an objective purely mathematical fact that we must take into account an try to exploit to our advantage.

Ignorance

One very useful fact that follows from all of this is an extension of Laplace's principle of insufficient reason to general regular statistical models. In every regular hypothesis space, there is a fixed objective notion of uniformity, i.e. of what it is meant to choose a point (a probability distribution in this case) uniformly over the whole space. This objective underlying measure of uniformity, provides the reference vacuum of knowledge (i.e., ignorance) for the space. The instrument, the equipment that will be used to run the experiment, fixes the likelihood and that in turn fixes the meaning of no-thing! ($\alpha \to 0$ for the cognoscenti). It fixes the vacuum.

The Acolyte Shows Off His New Tricks

Let's do all the computations for our coin problem. We have here two hypothesis spaces. The hypothesis space associated to the standard description A is the set of all probability distributions on the data space $\{0, 1\}$. This is a compact one dimensional manifold with boundary. We could think of it as a curve inside the space of all possible distributions on the real line. This line exists independent of any parametrization. In fact, we have just defined it without introducing any parameters. We can compute its length by integrating its element of length,

$$l(A) = \int_A dl = \int_0^1 \frac{dt}{\sqrt{t(1-t)}} = \pi$$

Thus, even though we can label the points along this line with $t \in [0,1]$, the line doesn't have length one but about 3.14 because it curves. Geometrically this is not unlike half a circle on the xy-plane, i.e., the points in euclidean space with coordinates,

$$p_A(t) = \cos(\pi t)\ e_1 + \sin(\pi t)\ e_2 \text{ for } 0 \leq t \leq 1$$

where $\{e_1, e_2, e_3\}$ is a fix orthonormal frame. Straight forward computations provide the corresponding length associated to B as,

$$l(B) = \int_B dl = \int_0^1 \sqrt{\frac{2-t}{t(1-t)}}\ dt \approx 3.82$$

Geometrically, the hypothesis space B is like the curve of points,

$$p_B(t) = \cos(ct)\ e_1 + \sin(ct)\cos(ct)\ e_2 + \sin^2(ct)\ e_3, \text{ for } 0 \leq t \leq 1$$

where $c = \pi$ makes the length to be about 3.82. Hence, under the assumption of total ignorance, the geometry demands to assign,

$$p(A) = \frac{l(A)}{l(A) + l(B)} \approx 0.45$$

and,

$$p(B) = \frac{l(B)}{l(A) + l(B)} \approx 0.55$$

for the apriori probabilities for the two descriptions! That's pure objective magic.

After observing the tails, $x = 0$ the posterior probabilities (using the above prior probabilities $p(A)$ and $p(B)$) are,

$$p(A|x=0) \propto \frac{0.45}{3.14} \int_0^1 \frac{t}{\sqrt{t(1-t)}}\ dt \approx 0.23$$

and,

$$p(B|x=0) \propto \frac{0.55}{3.82} \int_0^1 t\sqrt{\frac{2-t}{t(1-t)}}\ dt \approx 0.25$$

producing,

$$p(A|x=0) \approx 0.48 \text{ and } p(B|x=0) \approx 0.52$$

But what do these numbers mean? You ask.... and I explain.

These numbers have a very simple and totally objective interpretation free of paradox. These are the approximate frequencies that would be observed if you collect together all the probability distributions in A and in B and choose uniformly among them. Think of all these points as the set,

$$S = \{p_A(t) : t \in [0,1]\} \cup \{p_B(t) : t \in [0,1]\}$$

of vectors in euclidean space. Choose uniformly among them, to obtain a point p say. Then, sample the probability distribution that corresponds to the selected point. Repeat over and over again and count the frequency of occurrence of A when $x = 0$. You'll get 48%. No paradox.

Moral: *Probabilities are NOT physical*

Even though we used the same letter t to label both $p(x = 0|A)$ and $p(x = 0|B)$ that was only a parameter, a coordinate that lacks any meaning beyond a label for a probability distribution. Probabilities do not attach to physical reality. No matter what experiments we do now or in the future we'll never be able to see a single probability. QM included. Because probabilities do not attach to the world but to our descriptions of the world, to logical propositions in a given domain of discourse. Thank you Ed Jaynes.

Not Physical but Objective

How come?... Just like the rest of mathematics actually (See *What's Mathematics, Really?*)