Statistical Undecidability

Raphael Douady, CNRS & RiskData
Nassim N. Taleb, NYU-Poly

October 2010

Presentation of the result:

Using the metadistribution of possible distributions for a given measure, we define a condition under which it is possible to make a decision based on the observation of random variable, which we call "statistical decidability". We provide a sufficient condition on the metadistribution for the decision to be "statistically decidable" and conjecture that decisions based on a metadistribution with non compact support are always "statistically undecidable". There is the need for a strong *undefeasable a priori* without which decisions are not statistically justified — an effect that is very significant for decisions affected by small probabilities.

Decisions are not made on naive measure of True/False in simple cumulative probability space, but on a higher moments (say, expectation or some similar decision measure such as utility) —off some numerical decidability criterion. Unlike the Gödel result, which has not yet shown practical significance, the added dimension of consequence or utility of decision makes enormous consequences, making situations completely undecidable statistically.

Bayesian updating methods do not bring any remedy as they are much more prior-dependent than is thought naively by preselecting prior data and *a priori* (nonrevisable) distribution (i.e, without metadistribution). Maximum likelihood estimations are even worse as, by inverting the question of the distribution of the objective criterion and that of the sample conditionally to a choice of distribution, they provide absolutely no control on the objective criterion. In both cases, two observers can observe the same series, without ever converging.

Introduction

Let Ω be the space of possible eventualities (the *"random space"*) and μ be the (unknown) probability distribution on it. We need to take an "informed" decision, based on a criterion $\Phi(\mu)$ that depends on μ. Therefore Φ is a function defined on $\wp(\Omega)$ with values in a set V depending on the nature of the decision. For example:

- Yes/No decision: $\Phi : \wp(\Omega) \to V = \{0,1\}$
- Quantitative decision: $\Phi : \wp(\Omega) \to V = \mathbf{R}$ or $\mathbf{R}^d$

The decision will be taken with respect to the estimated distribution of $\Phi(\mu)$ knowing all or some of the available information.
Let us assume that Φ is continuous with respect to some norm $\|.\|_{\wp(\Omega)}$ on $\wp(\Omega)$. We shall assume that μ is drawn from an *a priori* distribution π on the σ-algebra spanned by this norm.

Let $\pi_\Phi = \Phi_* \pi$ be the image measure in V, that is, the distribution of $\varphi = \Phi(\mu)$ according to the distribution π. The decision will in fact not be taken with respect to $\Phi(\mu)$, which is unknown, but with respect to a criterion $\Psi(\pi_\Phi) \in V$, where the function $\Psi : \wp(V) \to V$ is assumed to be continuous with respect to a norm $\|.\|_{\wp(V)}$ and such that, for a Dirac mass δ_a on $a \in V$, one has $\Psi(\delta_a) = a$ (in other words, Ψ coincides with Φ when μ is perfectly known).
Let us now assume that the information is given by a sample of values of random variables $X_i(\omega)$, $i \in \{1,\ldots,n\}$, $\omega \in \Omega$, drawn at random from the probability distribution μ. Our decision question can be restated as:

- *What is the distribution of* $\Phi(\mu)$ *knowing* $(X_1,\ldots,X_n)$ *?* (Q_1)

Let us consider the compound random variables $(\xi_1,\ldots,\xi_n)$ defined by picking μ at random with respect to π, then ω at random with respect to μ and compute $X_i(\omega)$. Our question Q_1 can now be restated in questions Q_2 and Q_3 as follows:

- *What is the joint distribution of* $(\varphi, \xi_1,\ldots, \xi_n)$ *in* $V \times \mathbf{R}^{nd}$ *?* (Q_2)
- *What is the conditional distribution of* φ *in* V *knowing* $(\xi_1,\ldots,\xi_n)$ *?* (Q_3)

We can see Q_3 as a function $g_\pi : \mathbf{R}^{nd} \to \wp(V)$, then the decision criterion is the function $\psi = \Psi \circ g_\pi$. For this criterion to be usable, it must be well defined, continuous with respect to input values of $(\xi_1,\ldots,\xi_n)$ – hence g must be continuous when the image space $\wp(V)$ is equipped with the norm $\|.\|_{\wp(V)}$ – and converge to the criterion φ when n tends to $+\infty$.
Now comes the general question that π itself is generally unknown. At best, we assume that μ is picked within a certain class $C \subset \wp(\Omega)$.

Definition

A decision based on criteria Φ and Ψ and distribution π is *statistically decidable* if the following holds:

1. For any fixed n, the function $\psi : \mathbf{R}^{nd} \to V$ is well defined. If it is given as an integral with respect to π, then the integrand must be π-integrable.
2. For any fixed n, the function $\psi : \mathbf{R}^{nd} \to V$ is continuous with respect to the sample $(\xi_1,\ldots,\xi_n)$
3. Let us assume that $(X_1,\ldots,X_n)$ are drawn from a given measure μ and let us consider the sample error $\varepsilon(X_1,\ldots,X_n) = |\psi(X_1,\ldots,X_n) - \Phi(\mu)|$ and its expectation $\mathrm{Err}(\mu) = \mathrm{E}_\mu[\varepsilon(X_1,\ldots,X_n)]$. Then $\mathrm{Err}(\mu)$ must tend to 0 when n tends to $+\infty$ both π-almost surely and in $\mathrm{L}^1(\pi)$.

Otherwise it is said *statistically undecidable.* The latter condition is probably the most important of all: it means that no uncertainty on the distribution is left aside when the sample is large enough, so that the decision criterion corresponds to that originally fixed by the problem.
When π is unknown within a class $\Gamma \subset \wp(\wp(\mathbf{R}^d))$, then for the decision to be *statistically decidable,* functions $\psi = \Psi \circ g_\pi$ must be equi-continuous and the convergence of errors to 0 must be uniform in the class Γ.

Bayesian Statistics

Bayesian statistics are based on a prior distribution μ_0 then, given a sample X, the probability is modified to a posterior distribution μ_1 that depends on the prior probability of the sample:

$$\mu_1(A) = \frac{\mu_0(X \mid A)}{\mu_0(X)} \mu_0(A)$$

Explain why the knowledge of $\Phi(\mu_1)$ doesn't give any info the distribution of $\Phi(\mu)$ knowing X.

Maximum Likelihood

Given a sample $X = (X_1,\ldots,X_n)$, one defines the likelihood of a distribution $L(\mu) = \prod_{i=1}^{n} f_\mu(X_i)$ where f_μ is the pdf of μ. Then assuming $\mu = \mu_\alpha$ depends on a parameter $\alpha \in \mathbf{R}^d$ with $d < n$, one selects the parameter $a_{\max}$ that maximizes the likelihood $L(\mu_{\alpha\max})$.
Explain why the knowledge of $\Phi(\mu_{\alpha\max})$ doesn't give any info the distribution of $\Phi(\mu)$ knowing X.

Fourier Transform

Let us consider question (Q_3). By definition of conditional distributions, for any test functions $h(\mu)$ and $u_i(\xi_i)$, $i = 1\ldots n$, one has:

$$\int h(\mu)u_1(\xi_1)\ldots u_n(\xi_n)d\pi_\xi(\mu) = \int h(\mu)u_1(X_1)\ldots u_n(X_n)d\mu(X_1)\ldots d\mu(X_n)d\pi(\mu)$$

Assume that $\Psi(\pi) = \int U(\varphi(\mu))d\pi(\mu)$ and set $\psi(\xi) = \Psi \circ g_\pi(\xi)$. One has:

$$\begin{aligned}\int \psi(x)u_1(x_1)\ldots u_n(x_n)dx_1\ldots dx_n &= \int U(\Phi(\mu))u_1(X_1)\ldots u_n(X_n)d\mu(X_1)\ldots d\mu(X_n)d\pi(\mu)\\ &= \int U(\mathrm{E}_\mu(f))\mathrm{E}_\mu(u_1)\ldots\mathrm{E}_\mu(u_n)d\pi(\mu)\end{aligned}$$

Where $\Phi(\mu) = \int f d\mu$.
Using functions $u(x) = \exp(itx)$, we get the Fourier transform of ψ:

$$\begin{aligned}\hat{\psi}(t_1,\ldots,t_n) &= \int U(\mathrm{E}_\mu(f))\hat{\mu}(t_1)\ldots\hat{\mu}(t_n)d\pi(\mu)\\ &= \int U\left(\frac{1}{2\pi}\int \hat{f}(s)\hat{\mu}(s)ds\right)\hat{\mu}(t_1)\ldots\hat{\mu}(t_n)d\pi(\mu)\end{aligned}$$

We can therefore deduce the following:

Theorem
The function ψ is continuous – hence the statistical problem is decidable *– if:*

$$\int\left|\int U\left(\frac{1}{2\pi}\int \hat{f}(s)\hat{\mu}(s)ds\right)\hat{\mu}(t_1)\ldots\hat{\mu}(t_n)d\pi(\mu)\right| dt_1\ldots dt_n < +\infty$$

Conversely, if ψ is continuous – i.e. if the problem is decidable *– then:*

$$\lim_{\sum t_i^2 \to +\infty} \int U\left(\frac{1}{2\pi}\int \hat{f}(s)\hat{\mu}(s)ds\right)\hat{\mu}(t_1)\ldots\hat{\mu}(t_n)d\pi(\mu) = 0$$

Would this condition not be satisfied, then the problem would be undecidable.

Conjectures

Here is a list of conjectures that express "generic statistical undecidability":

1. If, for any criterion Ψ of the form $\Psi(\pi) = \int U(\varphi(\mu))d\pi(\mu)$, the problem is statistically decidable, then the metadistribution π has compact support in $\wp(\Omega)$. This result would show that for a problem to be statistically decidable, one needs either to make assumptions on the growth of the criterion at infinity, or strong a priori assumptions, such as a finitely parameterized class, on the acceptable measures.
2. Whatever the norm on $\wp(\wp(\Omega))$, the map $\pi \to \Psi \circ g$ is generically discontinuous. This means that very minor changes in the a priori distribution π lead to completely different decision criteria.
3. If the class C of possible π is not compact (a set with non empty interior in $\wp(\wp(\Omega))$ is not compact, whatever the norm), then the set of corresponding criteria is generically not uniformly continuous. This means that even when assuming that π is close to a given a priori probability measure π_0, one cannot control the sensitivity of the decision to inputs.
4. The more Φ depends on areas where μ has low probability, the less $\Psi \circ g$ is continuous, i.e. very close input samples can lead to very different decisions. This assertion, which needs a precise definition of "depending on where μ has low probability", exactly express the fact that small probabilities are harder to estimate than large ones. More precisely, let us assume that the norm $\|.\|_{\wp(\Omega)}$ is the dual of the standard max norm on $L^\infty(\Omega)$. Then the modulus of continuity of $\Psi \circ g$ is generically no better than that of Φ.