Information Sciences 453 (2018) 239–262

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

The randomness of the inferred parameters. A machine learning framework for computing confidence regions.

Bruno Apolloni[a,*], Simone Bassis[b]

[a] Social Things S.r.l., Viale Umbria 632013, 5 Milano, Italy
[b] PerceptoLab S.r.l., Via Giuseppe Mercalli 12, Giussano (MB), Italy

ARTICLE INFO

Article history:
Received 24 March 2017
Revised 10 April 2018
Accepted 14 April 2018
Available online 16 April 2018

Keywords:
Algorithmic inference
machine learning
parameter distribution
confidence intervals
learning Boolean functions

ABSTRACT

We start from the very operational perspective – having data, organize them in a suitable way to be used in the future – to enter the long standing fray on the nature of inferred parameters within a machine learning thread. Still in an operational perspective, we introduce a parametric inference approach that unprecedentedly gets rid of most drawbacks incurred by current methods to compute confidence intervals. The key idea is to consider the parameters of the distribution underlying a sample to be random, where randomness is expressed in terms of a probability measure of the compatibility of the parameter values with the actually observed data. The probability is understood, in a frequentist acceptation, in terms of the asymptotic frequency of those parameter values *matching* the observed sample in a story of infinite observations. The aim of this paper is to recap and complete theoretical results obtained through our approach as presented in preceding papers. In particular, here we focus on statistical tools both for computing confidence regions, at the basis of appraising the learnability of a function, and for checking their efficacy. We basically support our theory with a series of well-known benchmarks where, as for both volume and coverage of the confidence regions, our method proves superior – with very few ties – to those of competitors. Then we mention some results in computational learning theory that have been achieved recently exactly by adopting our approach, with a special focus on a new data_ accuracy - sample_complexity trade off.

1. Introduction

I, at any rate, am convinced that He does not throw dice[1]. The idea of randomness is deeply rooted in human thought. However, since the first haruspices trying to divine upcoming events from the flight of birds, it is indissolubly connected with the willing of profiting from any information hidden in a random sign. The ways to achieve this goal may depend on the available technology. Confining ourselves to a scientific context, we gather these ways in the statistical domain. In this framework, we see the computer as one of the most relevant technological leaps of the last century that is reflected in the evolution of methodologies from inference to learning. Hiding epistemological aspects in favor of operational ones, we may frame these methodologies in the following two essential scenarios, with some abuse of sharpness. The common features

* Corresponding author. Tel.: +393337469878; fax: +39283538781.
E-mail addresses: apolloni@di.unimi.it (B. Apolloni), bassis@di.unimi.it (S. Bassis).
URL: http://www.socialthingum.com/ (B. Apolloni), http://www.perceptolab.com (S. Bassis)

[1] Letter from A. Einstein to M. Born dated December 12, 1926, as reproduced in [8] on page 113.

https://doi.org/10.1016/j.ins.2018.04.047

are: a distribution law with a cumulative distribution function (CDF) F_{X_θ} that is known as for the shape yet unknown for the parameter θ and a set of observed data $\boldsymbol{x}$. Then[2]:

- Statistical inference: we assume the data to be the realization of a random sample $\boldsymbol{X}$ drawn from the above distribution law given a fixed value θ of its parameter that we try to appreciate in spite of the sample randomness (*fixed parameters – random data*).
- Statistical learning: we assume the data to be conveniently described in terms of a sample $\boldsymbol{x}$ drawn from F_X whose parameter Θ may assume different values with proper probabilities, given the values of the data we have observed (*fixed data – random parameters*).

The first framework proves ideal for testing hypotheses (for $\theta \in \mathit{\Theta}$ compute extremal quantile s_α of a given statistic). The same framework entails inversion problems to compute confidence intervals (for which set of parameters a function $s = \rho(\boldsymbol{x})$ of the sample – a statistic therefore – may result in an extremal quantile s_α?) with a special semantic (the unknown θ will fall a fraction α of times in the confidence intervals along an infinite series of i.i.d. random samples drawn exactly with θ), that often prove intractable without proper approximations. The second framework appears more suitable to compute confidence intervals. Once the Θ distribution law is determined, it is just a matter of computing its quantile θ_α per usual. The main benefit is that the semantic is much cleaner. In a long series of samples of whatever distribution law and related parameters – possibly changing from one instance to another – if we compute quantiles to obtain a confidence interval of level α, we expect these parameters to fall inside the interval with probability exactly equal to α. This is the typical goal of a computational learning algorithm, whose success rate is related to the high confidence with which we confine the generalization error in a tight interval close to 0, where the confidence is generally expected to be distribution-free over the course of human life [25]. We do not escape the necessity of solving inversion problems. Instead, thanks to the availability of powerful computational tools, we may successfully locate this problem at the beginning of the learning process – by computing the parameter distribution law, rather than at its end – by inverting the quantile formulas, provided it is possible.

Within this framework we formalize a class of procedures for computing confidence intervals that unprecedentedly get rid of the crucial drawbacks and paradoxes variously affecting all current methods on the same task (as we will show in the next Section). On the one hand, the key to the success of our approach lies in the linking of the notion of suitable statistics to draw inference on parameters to the features of the algorithms that compute them from data. On the other hand, precisely this link is at the core of the merger between computational and statistical disciplines underpinning computational learning theory. Accordingly, in this paper we provide a set of theorems and numerical benchmarking examples to firmly establish the bedrock of our approach and exhibit its benefits with respect to competitors. Then we focus the inference on a Bernoulli distribution in order to review some results recently achieved by the authors on the computational learning of boolean functions. They establish a further dependence of the sample complexity on the numerical accuracy with which data underpinning a learning problem is represented. Therefore, the paper is structured as follows. In Section 3 we recall a theory for computing parameter distribution laws introduced in previous papers [2,3] as well as in a couple of books of ours [4,5]; afterward, we extend it to vectorial parameters and strengthen the whole through formal statements. We then formalize two general methods for computing parameter distribution laws in both scalar and multidimensional domains. In Section 4 we adapt the confidence interval definition to the new framework and review some methods for computing it operatively. Section 5, which is devoted to the numerical results, first highlights the benefit of our procedures in three benchmarks; then in a special subsection it extends these results to the computation of the accuracy distribution law in the task of learning Boolean functions; finally it mentions some advanced results we got on computational learning theory through our approach. The paper ends in Section 7 with conclusions and references to future work. Some supplementary material is reported in the Appendix.

2. A narrow survey on confidence interval methods

Let us recall the basic definition of confidence interval for a scalar parameter θ.

Definition 1. Given a r.v. with a scalar parameter Θ and a real number $0 \leq \delta \leq 1$, $(\theta_{\mathrm{dw}}, \theta_{\mathrm{up}})$ is called a $1-\delta$ *confidence interval* for Θ if:

$$\mathrm{P}(\theta_{\mathrm{dw}} \leq \Theta \leq \theta_{\mathrm{up}}) = 1 - \delta \tag{1}$$

The quantity δ is called the confidence *level* of the interval.

The two scenarios mentioned in the introduction, (*fixed data – random parameters, fixed parameters – random data*), are reflected on the epistemological question about which is random, the parameter or the bounds. This entailed an extensive debate on how to compute confidence intervals. Cramer approach [11] is manifestly rooted on the former. Even fiducial intervals introduced by Fisher [17] are rooted on the former scenario, in spite of the fact that they deal with parameter

[2] By default, capital letters (such as U, X) will denote random variables and small letters (u, x) their corresponding realizations; the sets which the realizations belong to will be denoted by capital gothic letters ($\mathfrak{U}$, $\mathfrak{X}$); bold-faced characters will denote vectorial quantities. We denote by P the probability of an event.