Conditionality principle under unconstrained randomness

Vladimir Vovk

практические выводы
теории вероятностей
могут быть обоснованы
в качестве следствий
гипотез о *предельной*
при данных ограничениях
сложности изучаемых явлений

On-line Compression Modelling Project (New Series)

Working Paper #40

First posted February 7, 2024. Last revised February 10, 2024.

Project web site:
http://alrw.net

Abstract

A very simple example demonstrates that Fisher's application of the conditionality principle to regression ("fixed x regression"), endorsed by Sprott and many other followers, makes prediction impossible in the context of statistical learning theory. On the other hand, relaxing the requirement of conditionality makes it possible via, e.g., conformal prediction.

Contents

1 Introduction

The main goal of this note is to draw the reader's attention to the fact that the conditionality principle is not compatible with statistical learning theory, in which the task is to predict the label y of an object x. Two characteristic features of statistical learning theory are that the labelled objects (x, y) are only assumed to be independent and identically distributed (the unrestricted *assumption of randomness*) and that the objects x are complex (such as videos), so that we are unlikely to ever see identical objects. These two features make prediction impossible if we want to condition on the observed xs as recommended by Fisher. This is not a new observation, but it might not be as widely known as it deserves.

2 Assumption of randomness and conformal prediction

In statistical learning theory (see, e.g., [17, 16]) we consider *observations* (x, y) each consisting of two components: an *object* $x \in \mathbf{X}$ and its *label* $y \in \mathbf{Y}$. In general, the *object space* $\mathbf{X}$ and *label space* $\mathbf{Y}$ are arbitrary measurable spaces, but to discuss the relevance to Fisher's ideas it will often be convenient to concentrate on the case of *regression* $\mathbf{Y} = \mathbb{R}$.

The simplest setting is where we are given a training sequence

$$(x_1, y_1), \ldots, (x_n, y_n)$$

and the problem is to predict the label y_{n+1} of a test object x_{n+1}. The (unrestricted) *assumption of randomness* is that the observations $(x_1, y_1), \ldots, (x_{n+1}, y_{n+1})$ are generated independently from an unknown probability measure on $\mathbf{X} \times \mathbf{Y}$. This assumption is standard in machine learning and popular in nonparametric statistics.

One way to make predictions with validity guarantees under unconstrained randomness is *conformal prediction* [18]: given a target probability of error $\epsilon > 0$ conformal prediction produces a prediction set $\Gamma \subseteq \mathbf{Y}$ such that $y_{n+1} \in \Gamma$ with probability at least $1-\epsilon$. The basic idea of conformal prediction is familiar (see, e.g., [6, Sect. 7.5]): we fix a statistical test of the null hypothesis of randomness, go over all possible labels y for the test object x_{n+1}, and include in Γ all labels y for which the augmented training set $(x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, y)$ does not lead to the rejection of the null hypothesis. In many interesting cases this idea has computationally efficient implementations (see, e.g., [18, Sects 2.3 and 2.4] and [12]).

3 A simple conditionality principle and example

In this section we will only need a special case of the conditionality principle, which I will call the "fixed x principle" following Aldrich [1]. The fixed x principle says that in regression problems (or any other prediction problems of the kind described in the previous section) we should consider the observed sequence of objects $x_1, \ldots, x_{n+1}$ as fixed, even if they were in fact generated from some probability distribution (known or unknown).

Fisher was a life-long promoter of both the general conditionality principle (which was introduced formally only in 1962 by Birnbaum [5]) and the fixed x principle as its special case. When H. Fairfield Smith asked Fisher about the origin of the fixed x principle (treating "the independent variable as fixed even although it might have been observed as a random sample of some variate population") in his letter of 6 August 1954 [4, pp. 213–214], Fisher responded with a reference to his 1922 paper [7, p. 599].

See Aldrich [1] about the development of Fisher's fixed x regression. Aldrich quotes Fisher [8, Sect. 2, p. 71] (in the context of regression with y having a Gaussian distribution given x): "The qualitative data may also tell us how x is distributed, with or without specific parameters; this information is irrelevant."

David Sprott, a prominent follower of Fisher's, also promoted the conditionality principle in his work. In his 1989 interview with Mary Thompson [15], he remembers a case when, as a student, he was tempted to take into account the variation in the xs in a practical regression problem. His statistics professor said, "No, you wouldn't do that, you'd condition on the xs". Sprott couldn't find out why conditioning on the xs was the right thing to do until he went to London a few years later to work with Fisher, but then he was fully convinced by Fisher.

The intuition behind the fixed x principle is that only the observed objects are relevant for predicting the label of x_{n+1}. In Sect. 5 we will discuss this in a wider context. And in this section, we discuss the paralysing effect of the fixed x principle under unrestricted randomness using the following example.

Example 3.1. Consider the problem of regression with $\mathbf{X} = \mathbf{Y} = \mathbb{R}$, and suppose:

- the training sequence is such that $y_i = x_i$ for all $i = 1, \ldots, n$, for a large n;
- $x_1, \ldots, x_{n+1}$ are all different.

What can we say about y_{n+1} knowing x_{n+1}?

Example 3.1 may describe a situation where the observations are independent and coming from the same continuous distribution. Under the assumption of randomness, we can confidently claim that $y_{n+1} = x_{n+1}$; otherwise, the last observation looks strange and leads to a p-value of $1/(n+1)$ for a fixed statistical test. As the test statistic T for such a test we can take, e.g.,

$$T := \begin{cases} 1 & \text{if } |y_n - x_n| > |y_i - x_i| \text{ for all } i = 1, \ldots, n-1 \\ 0 & \text{otherwise.} \end{cases}$$

For any $\epsilon > 1/(n+1)$, we have $[x_{n+1}, x_{n+1}]$ as the prediction interval for y_{n+1} at confidence level $1 - \epsilon$. Intuitively, this follows from the expectation that the future will be similar to the past (Laplace's rule of succession). Conformal prediction extends this idea greatly.

However, we can say nothing whatsoever about y_{n+1} if we condition on the observed $x_1, \dots, x_{n+1}$. The problem with conditioning on x is that it destroys the assumption of randomness. The assumption of randomness becomes the following *assumption of conditional randomness*: the data-generating distribution P is determined by an arbitrary sequence $(x_1, \dots, x_{n+1}) \in \mathbf{X}^{n+1}$ and an arbitrary family of probability measures $\{Q_x \mid x \in \mathbf{X}\}$ on $\mathbf{Y}$ (measurable in the sense of being a Markov kernel with $\mathbf{X}$ as source and $\mathbf{Y}$ as target); we then have $P = (\delta_{x_1} \times Q_{x_1}) \times \cdots \times (\delta_{x_{n+1}} \times Q_{x_{n+1}})$, where δ_x is the distribution on $\mathbf{X}$ that is concentrated at x. When $x_1, \dots, x_{n+1}$ are all different, the true conditional distribution of y_{n+1} can be any probability measure on $\mathbf{Y}$.

4 More advanced results

In this section I will briefly describe several results that shed light on the phenomenon illustrated by Example 3.1. Lei and Wasserman [13, Lemma 1] show that efficient set prediction under unrestricted randomness is impossible even if we only condition on the test object x_{n+1}, provided it is not an atom of the data-generating distribution.

Barber et al. [2] show that non-trivial set prediction becomes possible when we condition on $x_{n+1} \in \mathcal{X}$ for sets $\mathcal{X}$ of probability at least δ for some threshold $\delta > 0$, but it can always be accomplished as a trivial corollary of unconditional prediction procedures such as conformal prediction.

Despite these negative results, designing conformal predictors that are conditional in a weaker sense is an active area of research, starting from a basic idea of Mondrian conformal prediction [18, Sect. 4.6]. Asymptotically, conditional conformal prediction is possible in a very strong sense [13, Theorem 1]. Finite-sample results are mathematically less satisfying but may hold great promise in practice; see, e.g., [14, 3, 9, 10].

5 General conditionality principle and its difficulties

Cox and Hinkley [6, Sect. 2.3(iii)] point out two versions of the conditionality principle, basic and extended. In the basic version, we are given an *ancillary statistic* C, i.e., a random variable with a known distribution, and the conditionality principle says that our analysis should be conditional on the observed value of C. This prescription is very compelling in some cases, such as Cox's famous example of choosing one of two measuring instruments at random and then observing its reading (knowing which instrument has been chosen); see [6, Example 2.33].

The basic version does not imply the fixed x principle, since the distribution of the xs does depend on the unknown data-generating distribution. In the extended conditionality principle, the unknown parameter is split into two parts, and only one of those parts is of direct interest to us. In the context of the assumption of randomness, the parameter is the data-generating distribution $P := R^{n+1}$, where R is the probability measure on $\mathbf{X} \times \mathbf{Y}$ generating one observation. Split R into the marginal distribution $R_{\mathbf{X}}$ on $\mathbf{X}$ and the conditional distribution Q_x of y given x, for each $x \in \mathbf{X}$; Q is a Markov kernel. This splits P into the marginal distribution $R_{\mathbf{X}}^{n+1}$ on $\mathbf{X}^{n+1}$ and the family

$$Q := \{Q_{x_1} \times \cdots \times Q_{x_{n+1}} \mid (x_1, \ldots, x_{n+1}) \in \mathbf{X}^{n+1}\};$$

only Q is of interest to us in our prediction problem. Then $C := (x_1, \ldots, x_{n+1})$ is ancillary for Q in the following extended sense:

- The distribution of C does not depend on Q (and only depends on $R_{\mathbf{X}}^{n+1}$).
- The conditional distribution of the remaining part of the data $y_1, \ldots, y_{n+1}$ given the value of C depends only on Q (and does not depend on $R_{\mathbf{X}}^{n+1}$).

The extended conditionality principle, requiring analysis to be conditional on C, becomes the fixed x principle when adapted to the assumption of randomness.

Lehmann and Scholz [11, Sect. 1] point out that conditional inference can be less efficient for small samples, although the difference tends to disappear as the sample size increases. However, Example 3.1 is more serious, since it demonstrates a complete failure of the conditionality principle. Another instance of a comparable complete failure of this principle is where the experimental design involves deliberate randomization, as in a random assignment of subjects to treatments in randomized clinical trials [11, end of Sect. 3]. The conditionality principle then forces us to disregard randomization disabling the most standard and powerful statistical tools in medicine.

6 Conclusion

This note observes that the extended conditionality principle prevents successful prediction in statistical learning theory, which is the basic setting of machine learning. (Other varieties of machine learning usually make prediction even more difficult; e.g., they may allow different distributions for the training and test observations.) A natural way out is to relax the conditionality principle, making *approximate* conditionality one of the desiderata for prediction algorithms (see, e.g., [18, Sect. 1.4.4]).

Acknowledgments

Thanks to Changbao Wu, Chair of the Department of Statistics and Actuarial Science (University of Waterloo, Canada), for inviting me to deliver the David Sprott Distinguished Lecture in June 2023. I am grateful to my listeners, especially Ruodu Wang, for their perceptive comments. My research has been partially supported by Mitie.

References

[1] John Aldrich. Fisher and regression. *Statistical Science*, 20:401–417, 2005. Original title: "The origins of fixed X regression".

[2] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10:455–482, 2021.

[3] Anthony Bellotti. Approximation to object conditional validity with inductive conformal predictors. *Proceedings of Machine Learning Research*, 152:4–23, 2021. COPA 2021.

[4] J. H. Bennett, editor. *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher*. Clarendon Press, Oxford, 1990.

[5] Allan Birnbaum. On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association*, 57:269–326, 1962.

[6] David R. Cox and David V. Hinkley. *Theoretical Statistics*. Chapman and Hall, London, 1974.

[7] Ronald A. Fisher. The goodness of fit of regression formulae, and the distribution of regression coefficients. *Journal of the Royal Statistical Society*, 85:597–612, 1922.

[8] Ronald A. Fisher. Statistical methods and scientific induction. *Journal of the Royal Statistical Society B*, 17:69–78, 1955.

[9] Leying Guan. Localized conformal prediction: a generalized inference framework for conformal prediction. *Biometrika*, 110:33–50, 2023.

[10] Rohan Hore and Rina Foygel Barber. Conformal prediction with local weights: randomization enables robust guarantees. Technical Report arXiv:2310.07850 [stat.ME], arXiv.org e-Print archive, October 2023.

[11] Erich L. Lehmann and F. W. Scholz. Ancillarity. In Malay Ghosh and Pramod K. Pathak, editors, *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, volume 17 of *Lecture Notes—Monograph Series*, pages 32–51. Institute of Mathematical Statistics, Hayward, CA, 1992.

[12] Jing Lei. Fast exact conformalization of lasso using piecewise linear homotopy. *Biometrika*, 106:749–764, 2019.

[13] Jing Lei and Larry Wasserman. Distribution-free prediction bands for nonparametric regression. *Journal of the Royal Statistical Society B*, 76:71–96, 2014.

[14] Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.

[15] Mary Thompson. A conversation with David A. Sprott. *Liaison*, 3(2), February 1989. Available (in January 2024) on the web.

[16] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[17] Vladimir N. Vapnik and Alexey Y. Chervonenkis. Теория распознавания образов *(Theory of Pattern Recognition)*. Nauka, Moscow, 1974. German translation: *Theorie der Zeichenerkennung*, Akademie, Berlin, 1979.

[18] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, Cham, second edition, 2022.