Prediction is not everything, but everything is prediction

Leonardo Egidi

Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche 'Bruno de Finetti', Università degli Studi di Trieste, Trieste, Italy.

E-mail: legidi@units.it

Jonah Sol Gabry

Department of Statistics, Columbia University, New York, USA.

E-mail: jgabry@gmail.com

Abstract.

1. Introduction

2. Prediction for science or science for prediction?

2.1. It is prediction part of the science design?

Bertrand Russell's scheme, Cartesio

3. The role of prediction in statistics

Statistics has always been thought as the *science of inference*, or *science of estimates*, and inference is always seen as separate from prediction. Inference is based on an underlying mathematical model for the data-generating process (Bzdok et al., 2018), its main task is to describe an unknown mechanism working through generalization: the inferential laws should in fact be as broad as possible, ideally valid for the population of interest, and not symptomatic of the observed data (it is out of the scope of this aper to review the distinct inferential approaches). Prediction moves from the observed to the unobserved, being the action designed to forecast future events without requiring a full understanding of the data- generation process. Each person is more or less confident with the weather's predictions or with presidential election predictions, but rarely that person is aware of the underlying statistical model required to produce that forecast, unless he is a statistician/data scientist. In such a view, inference seems hard and obscure, and prediction easy and close to the people. This is often a paradoxical argument, since the inference is often associated to the *explanation* of the problem, and should be relevant and available to the majority of the population

As statisticians, we are often faced with a double task. First, we must create a sound mathematical model to accomodate the data and retrieve useful inferences for our parameters—there is not distinction here between classical and Bayesian statistics, they are both designed to draw inferential conclusions, either in form of point estimates/confidence intervals or in terms of posterior quantiles/credibility intervals. Then, we should use this model to make predictions, but this is rarely accounted by the statisticians in a transparent way.

For illustration purposes only, we consider the classical linear regression model, where y_n denotes the response variable, X is the $n \times p$ predictor matrix, and $\alpha, \beta_1, \beta_2, \dots, \beta_p$ are the $p+1$ parameters—the intercept α and the p regression parameters, we have

$$y_n = \alpha + \sum_{k=1}^{p} \beta_j x_{nk} + \varepsilon_n, \; n = 1, 2, \dots, N, \tag{1}$$

with $\varepsilon_n \sim \mathcal{N}(0, \sigma^2_\varepsilon)$. Once the model has been estimated and we have retrieved some parameters' estimates $\hat{\alpha}, \hat{\beta}_1, \beta_2, \dots, \hat{\beta}_p$, we could use them to provide a point forecast $\tilde{y}_{n+1}$ for the unobserved y_{n+1} associated to the predictor vector $\tilde{x}_{n+1k}$. To account for uncertainty, rather than using a point forecast, we could also compute the prediction interval for $\tilde{y}_{n+1}$.

Once the value y_{n+1} is known, the statistician can validate his prediction and check its plausibility. This *ex post* validation is not available when fitting the model, and as such should not represent the plausibility of the model fit itself, since the model construction did not require y_{n+1}. In this misalignment between the fitting of the model without the $n+1$-th unit and the forecast validation of y_{n+1} there is space for an infinite debate about the scientific role of prediction.

3.1. Overfitting and data accomodation

Even a simple linear regression case poses many challenges: should the statistician use all the N data to build a reasonable/useful model, or could he take only a portion of the sample to accomodate the model (the training set) and use the remaining values to validate the model (the test set)? This apparently naive question pushed many scholars to debate about the presumed supremacy of prediction over accomodation (Maher, 1988; Hitchcock and Sober, 2004; Worrall, 2014). According to this competition, the statistician should ask himself whether he wants models that are true—or approximately true—or predictively accurate.

It is well-known that more complex models tend to yield poor predictive results

3.2. Prediction as a confirmation theory approach

Popper, Kuhn, Mayo

3.3. The Bayesian framework

Gelman's scheme of Bayesian inference. Draw connection with Bertrand Russell

As stated by Gelman et al. (2013), the process of Bayesian data analysis can be idealized by dividing it into the following three steps:

(a) Setting up a full probability model–a joint probability distribution—for all observable and unobservable quantities in a problem. The model should be consistent with knowledge about the underlying scientific problem and the data collection process.

(b) Conditioning on observed data: calculating and interpreting the appropriate posterior distribution, i.e. the conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data.

(c) Evaluating the fit of the model and the implications of the resulting posterior distribution: how well does the model fit the data, are the substantive conclusions reasonable, and how sensitive are the results to the modeling assumptions in step 1? In response, one can alter or expand the model and repeat the three steps.

In the above paradigm, predictions are never mentioned. But this does not mean that predictions are not relevant in the Bayesian paradigm. Denoted by $\tilde{y}$ the unobserved vector of future values, we may derive the posterior predictive distribution as

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta, \tag{2}$$

where $p(\theta|y)$ is the posterior distribution for θ, whereas $p(\tilde{y}|\theta)$ is the likelihood function for future observable values. In the linear regression case (1), the posterior predictive distribution for the future observation $\tilde{y}_{n+1}$ is given by:

$$p(\tilde{y}_{n+1}|y) = \int \mathcal{N}(\alpha + \sum_{k=1}^{p} \beta_p \tilde{x}_{n+1k}, \sigma^2_\varepsilon) p(\alpha, \beta_1, \beta_2, \ldots, \beta_p|y) d\alpha d\beta_1 d\beta_2 \ldots d\beta_p.$$

3.4. *Communication duties*

4. The role of prediction in machine learning

Breiman, Bzdok, Popper (we could argue that ML procedures are not falsifiable!)

5. Going beyond inference and prediction: a tentative unifying approach

6. Applied examples

7. Conclusions

References

Bzdok, D., N. Altman, and M. Krzywinski (2018). Points of significance: statistics versus machine learning.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Hitchcock, C. and E. Sober (2004). Prediction versus accommodation and the risk of overfitting. *The British journal for the philosophy of science 55*(1), 1–34.

Maher, P. (1988). Prediction, accommodation, and the logic of discovery. In *PSA: Proceedings of the Biennial meeting of the philosophy of science association*, Volume 1988, pp. 273–285. Philosophy of Science Association.

Worrall, J. (2014). Prediction and accommodation revisited. *Studies in History and Philosophy of Science Part A 45*, 54–61.