Computational Statistics and Data Analysis 104 (2016) 183–196

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Using the Bayesian Shtarkov solution for predictions

Tri Le *, Bertrand Clarke

Department of Statistics, University of Nebraska-Lincoln, 340 Hardin Hall North, Lincoln, NE, USA

ARTICLE INFO

Article history:
Received 19 October 2015
Received in revised form 11 May 2016
Accepted 30 June 2016
Available online 15 July 2016

Keywords:
Bayes
Prequential
Model average
Stacking
Shtarkov predictor
Bagging

ABSTRACT

The Bayes Shtarkov predictor can be defined and used for a variety of data sets that are exceedingly hard if not impossible to model in any detailed fashion. Indeed, this is the setting in which the derivation of the Shtarkov solution is most compelling. The computations show that anytime the numerical approximation to the Shtarkov solution is 'reasonable', it is better in terms of predictive error than a variety of other general predictive procedures. These include two forms of additive model as well as bagging or stacking with support vector machines, Nadaraya–Watson estimators, or draws from a Gaussian Process Prior.

1. Introduction

What kind of inference can we do when we do not believe the data were generated by a model? The most obvious answer to this question is prediction[1]: As long as there is something to measure we can make a guess as to its next value. The act of making the guess does not by itself even require there be anything stable enough about the data generator (DG) to make good prediction feasible. Moreover, predicting is more general than modeling because every model corresponds to a predictor but not every predictor corresponds to a model. Hence, if there is no model for a given DG we are essentially forced to predict using a larger class of predictors than models represent, i.e., it is not reasonable to limit ourselves to models for prediction. One effect of this in a Bayes context is to change the meaning of the prior.

This situation is far from unusual. Indeed, one can argue that many of the most important data that are gathered were not generated by a model, or, more precisely, they were not generated by any mechanism for which modeling per se is likely to be helpful. We use the term $\mathcal{M}$-open to label this class of problems, see Bernardo and Smith (2000) and Clyde and Iversen (2013). Specifically, we say a problem is $\mathcal{M}$-open when there is no model that accurately describes the mechanism by which the DG generated the data. Operationally, when we say this we mean that on intuitive and pragmatic grounds it is more reasonable to abandon rather than continue the search for a true model.

Let us review three techniques that have been proposed for $\mathcal{M}$-open data.

One of the earliest techniques intended for $\mathcal{M}$-open data is due to Shtarkov (1987). He recognized that if there is no model it may make sense to imagine a collection of 'experts' regarded as density functions who issue predictions. Then, at each time step the best expert can be identified in the sense of regret under log-loss. This approach has been extended in Vovk (2001) and Cesa-Bianchi and Lugosi (2006). Techniques for computing the Shtarkov solution were first presented in Kontkanen and Myllymaki (2007). Although Shtarkov's formulation was not Bayesian, the frequentist Shtarkov predictor

* Corresponding author.
E-mail addresses: tle20@unl.edu (T. Le), bclarke3@unl.edu (B. Clarke).

[1] R code for generating Bayes Shtarkov predictions presented here is in given in an annex to the electronic version of this paper.

http://dx.doi.org/10.1016/j.csda.2016.06.018

is asymptotically Bayes (see Clarke, 2007) and it is easy to write down the Bayes version. (Here and below, we abbreviate 'Bayesian' to 'Bayes' wherever possible, for brevity.)

A second approach for prediction is stacking, due to Wolpert (1992). Given a list of candidate models, weights for their predictions can be derived by minimizing a criterion similar to cross-validation. There are several versions of this minimization problem depending on the constraints imposed on the weights. Stacking has been studied by Breiman (1996b), Ting and Witten (1999) amongst others and explicitly extended to $\mathcal{M}$-open problems by Clyde and Iversen (2013). Le and Clarke (2015) showed that stacking can be asymptotically regarded as the Bayes action under several loss functions.

A third technique is bagging, see Breiman (1996a). Despite its origins in classification, bagging has also been used, usually without comment, for regression problems that are $\mathcal{M}$-open. For instance, Strobl et al. (2009) provide several examples as well as a good discussion of the key features of bagging in practice. It can be shown that bagging is asymptotically a specific form of Bayes model averaging (BMA), see Le and Clarke (unpublished).

A separate issue from how predictions from components are combined is the selection of the components themselves. While it is desirable to choose components that will yield good predictors, this cannot be known in advance. So, for $\mathcal{M}$-open DG's we want components that are flexible.

For bagging and stacking we have used three classes of components. First is the Nadaraya–Watson (NW) estimator. Given a data set, we can draw, say, ten bootstrap samples and therefore generate ten NW estimators. In the $\mathcal{M}$-open case it does not make sense to call them estimators since there is nothing to estimate. However, we do so for convenience. Now we can 'bag' the ten NW estimators by taking the average of the predictions they make at a new value of the explanatory variables. Alternatively, we can stack the ten NW estimators or more precisely the predictors they generate. NW estimators can be regarded as Bayes by using a prior on the smoothing parameter. This is computationally infeasible for the scale of our work here.

Second, given a kernel function, we can obtain the posterior from a Gaussian process prior (GPP) and use it to generate predicted values similar to the way NW estimators were used.

A third class of components that we use is support vector machines (SVM's). These are also based on kernel functions. Although it does not seem to have been formally proved, SVM regression and Gaussian process regression are not equivalent, see Rasmussen and Williams (2006) Sec. 6.4.1. Moreover, while more familiar from classification than regression, SVM's do give a regression function under an ϵ-insensitive loss. Again, taking ten bootstrap samples leads to ten SVM regression functions that can be bagged or stacked. The form of solution is based on the Representer Theorem, see Kimeldorf and Wahba (1973). More recently, Chakraborty et al. (2012) developed Bayes estimation in this context.

A fourth class of components that we use only with the Shtarkov predictor is the multinomial. Even though the experts combined in a Shtarkov predictor may be discrete or continuous, here we must discretize any explanatory variables so that predictions can be computed. For independent data, the 'experts' then naturally assume a multinomial distribution. The dependent variable must also be discretized to minimize computational difficulties and avoid having to choose specific parametric forms for the experts. While the computational procedure was proposed by Kontkanen and Myllymaki (2007) in the frequentist case, we are the first to evaluate how well it performs in contrast to other techniques.

Given independent data of the form $(y_i, x_i)|_{i=1}^n$ where x_i is a vector of explanatory variables, we form various predictors $\hat{y}_{i+1}(\cdot)$ for $Y_{i+1}(x_{i+1})$ using the first i data points. We evaluate these predictors by their cumulative squared prediction error, namely

$$CPE = \sum_{i=1}^{n} (\hat{y}_i(x_i) - y_i)^2. \tag{1}$$

Thus our evaluation is prequential, see Dawid (1984).

For a sequence of $\mathcal{M}$-open data sets we compare the CPE's of 10 different predictors using up to two explanatory variables (chosen by highest correlation with the y_i's). Six predictors come from the combination of bagging or stacking with NW, GPP, and SVM predictors. The other four are forms of the Shtarkov predictor: no side information, one of two explanatory variables as side information, and two explanatory variables as side information. For comparison purposes, we also generate predictions using additive models via the Bayes LASSO and the horseshoe prior, even though these are intended for use outside the class of $\mathcal{M}$-open problems. To get a better assessment of CPE, often (1) is averaged over several permutations of the data. Also, if several permutations of the data are used the standard deviation of the errors can be found at each time step, $i = 1, \ldots, n$. We have not done this here because it was too computationally demanding and probably not necessary given the sample sizes we have used in our examples.

Our main finding here is that when Bayes Shtarkov solutions are feasible to compute reliably at least one of them outperforms the other six methods. We attribute this to the fact that the optimality property satisfied by Shtarkov predictors is the most desirable one for $\mathcal{M}$-open problems. This is not to say that the naive use of Bayes Shtarkov solutions will always be the best. Indeed, we found many cases where side information made for a worse predictor than the absence of side information. Moreover, computing Shtarkov solutions reliably is very difficult with existing data storage. Amongst the other predictors, we also noted some regularities but they were not as strong. First, for $\mathcal{M}$-open problems, stacking with SVM's tended to do well in the sense that stacked SVM's were always one of the top three methods in terms of CPE. Second, in some cases (not shown here) where the problems were $\mathcal{M}$-open but not as hard the best results were typically obtained by stacking NW estimators.