Journal of Statistical Planning and Inference 140 (2010) 65–74

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Bayesian bootstrap prediction

Tadayoshi Fushiki

The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan

ARTICLE INFO

Article history:
Received 3 September 2007
Received in revised form
17 June 2008
Accepted 17 June 2009
Available online 21 June 2009

Keywords:
Bagging
Bayesian bootstrap
Bootstrap
Kullback–Leibler divergence
Probabilistic prediction

ABSTRACT

In this paper, bootstrap prediction is adapted to resolve some problems in small sample datasets. The bootstrap predictive distribution is obtained by applying Breiman's bagging to the plug-in distribution with the maximum likelihood estimator. The effectiveness of bootstrap prediction has previously been shown, but some problems may arise when bootstrap prediction is constructed in small sample datasets. In this paper, Bayesian bootstrap is used to resolve the problems. The effectiveness of Bayesian bootstrap prediction is confirmed by some examples. These days, analysis of small sample data is quite important in various fields. In this paper, some datasets are analyzed in such a situation. For real datasets, it is shown that plug-in prediction and bootstrap prediction provide very poor prediction when the sample size is close to the dimension of parameter while Bayesian bootstrap prediction provides stable prediction.

1. Introduction

In this paper, bootstrap prediction is adapted to resolve some problems in small sample datasets. Bagging was proposed by Breiman (1996), and has been studied in machine learning (for example, Bühlmann and Yu, 2002). The bootstrap predictive distribution is obtained by applying the bagging algorithm to the plug-in distribution with the maximum likelihood estimator (MLE). The effectiveness of bootstrap prediction has been studied under the Kullback–Leibler loss (Fushiki et al., 2005; Fushiki, 2005; Harris, 1989). It is known that Bayesian prediction is admissible when a proper prior is used (Aitchison, 1975). However, sometimes it is computationally not very easy to obtain Bayesian prediction. We have shown that bootstrap prediction is considered to be an approximation of Bayesian prediction and provides asymptotically better prediction than plug-in prediction with the MLE (Fushiki et al., 2005).

Let N be the number of observations. The difference between the bootstrap predictive distribution and the plug-in distribution with the MLE is $O_p(N^{-1})$, thus the difference is negligible if N is quite large. However, the difference has significance in small sample datasets even if the difference between the true distribution and the plug-in distribution with the MLE/bootstrap predictive distribution is $O_p(N^{-1/2})$. These days, analysis of small sample datasets is quite important in various fields, but it is known that plug-in prediction sometimes provides poor prediction for small sample datasets, then other methods are required. Although bootstrap prediction is thought to be more effective in such a case, some problems may arise when bootstrap prediction is constructed in small sample datasets. For example, in regression problems, a matrix whose inverse is needed to obtain the MLE may be rank-deficient in a bootstrap sample because some observations may not appear in the bootstrap sample. To overcome such problems, we adapt bootstrap prediction in this paper. Instead of ordinary bootstrap, Rubin's (1981) Bayesian bootstrap is

E-mail address: fushiki@ism.ac.jp.

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2009.06.007

used to construct estimators. I learned that the same method is known as Bayesian bagging (Clyde and Lee, 2001; Lee and Clyde, 2004) in machine learning. The Bayesian bootstrap predictive distribution is obtained by applying Bayesian bagging to the plug-in distribution with the MLE. In this paper, we study the effectiveness of Bayesian bootstrap prediction.

The effectiveness of Bayesian bootstrap prediction is confirmed by some examples. For real datasets, plug-in prediction and bootstrap prediction provide very poor prediction when the sample size is close to the dimension of parameter while Bayesian bootstrap prediction provides stable prediction. The risk improvement is quite large in this case.

Although we are interested in nonasymptotic situations, the asymptotic properties of Bayesian bootstrap prediction are also studied. It is shown that Bayesian bootstrap prediction is asymptotically more effective than plug-in prediction with the MLE. This fact guarantees that Bayesian bootstrap prediction is effective even if the sample size is not small. Thus, Bayesian bootstrap prediction can be used in various situations. The connection between Bayesian procedures and bootstrapping is also an interesting topic. For example, Efron (2003, 2005) discussed the connection. Boos (2003) suggested some analogies between the parametric bootstrap in frequentist inference and Markov chain Monte Carlo methods in Bayesian analysis. Hastie et al. (2001) also pointed out the relation between bootstrap and noninformative Bayes in Section 8. In this paper, it is shown that Bayesian bootstrap prediction is second-order equivalent to Bayesian prediction with a prior satisfying a condition, and that the condition is an extension of the result obtained by Newton and Raftery (1994).

This paper is organized as follows. In Section 2, we formulate the statistical prediction problem discussed in this paper. In Section 3, the Bayesian bootstrap predictive distribution is defined. In Section 4, the effectiveness of Bayesian bootstrap prediction is shown by some examples. In Section 5, the asymptotic properties of Bayesian bootstrap prediction are studied. Section 6 is the Discussion section.

2. Problem formulation

In this paper, the following statistical prediction problem is considered. Observations $x^N=\{x_1,\dots,x_N\}$ are drawn independently from an unknown distribution p_0. We discuss a probabilistic prediction of x_{N+1} based on x^N. A statistical model $\mathscr{P}=\{p(x;\theta)|\theta\in\Theta\}$ is used to predict x_{N+1}. In this paper, it is assumed that there exists a unique θ_0 such that $p_0(x)=p(x;\theta_0)$. Let $\hat{p}(x_{N+1};x^N)$ be a predictive distribution. The Kullback–Leibler divergence is used to measure the predictive performance of $\hat{p}$. The loss function is given by

$$D(p_0(x_{N+1}),\hat{p}(x_{N+1};x^N))=\int p_0(x_{N+1})\log\frac{p_0(x_{N+1})}{\hat{p}(x_{N+1};x^N)}\,\mathrm{d}x_{N+1}.$$

The risk function is

$$\mathrm{E}_{x^N}[D(p_0(x_{N+1}),\hat{p}(x_{N+1};x^N))]=\int p_0(x^N)\left\{\int p_0(x_{N+1})\log\frac{p_0(x_{N+1})}{\hat{p}(x_{N+1};x^N)}\,\mathrm{d}x_{N+1}\right\}\mathrm{d}x^N.$$

The Bayesian predictive distribution

$$p_\pi(x_{N+1}|x^N)=\int p(x_{N+1};\theta)\pi(\theta|x^N)\,\mathrm{d}\theta,\quad \pi(\theta|x^N)=\frac{p(x^N;\theta)\pi(\theta)}{\int p(x^N;\theta)\pi(\theta)\,\mathrm{d}\theta}\tag{1}$$

is an optimal predictive distribution in the sense that it minimizes the following Bayes risk:

$$\mathrm{E}_\theta[\mathrm{E}_{x^N}[D(p(x_{N+1};\theta),\hat{p}(x_{N+1};x^N))]]=\int\pi(\theta)\left[\int p(x^N;\theta)\left\{\int p(x_{N+1};\theta)\log\frac{p(x_{N+1};\theta)}{\hat{p}(x_{N+1};x^N)}\,\mathrm{d}x_{N+1}\right\}\mathrm{d}x^N\right]\mathrm{d}\theta.$$

Thus, the Bayesian predictive distribution is admissible when the prior is proper, but sometimes it is computationally not very easy to calculate (1).

3. Bayesian bootstrap prediction

The bootstrap predictive distribution (Fushiki et al., 2005) is defined by

$$\hat{p}_{\mathrm{B}}(x_{N+1};x^N)=\int p(x_{N+1};\hat{\theta}_{\mathrm{MLE}}(x^{*N}))\hat{p}(x^{*N})\,\mathrm{d}x^{*N},$$

where $\hat{p}$ is the empirical distribution $\hat{p}(x)=(1/N)\sum_{i=1}^N\delta(x-x_i)$ and $\hat{\theta}_{\mathrm{MLE}}(x^{*N})$ is the MLE based on a bootstrap sample x^{*N}. This predictive distribution is obtained by applying Breiman's bagging to the plug-in distribution with the MLE. We have shown that the bootstrap predictive distribution is considered to be an approximation of the Bayesian predictive distribution (Fushiki et al., 2005).