Neurocomputing 108 (2013) 13–22

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Fast learning rates for sparse quantile regression problem

Shao-Gao Lv [a],*, Tie-Feng Ma [a], Liu Liu [b], Yun-Long Feng [c]

[a] Statistics School, Southwestern University of Finance and Economics, ChengDu 611130, China
[b] V.C. and V.R. Key Lab and College of Mathematics and Software Science, SiChuan Normal University, ChengDu, China
[c] Department of Mathematics, City University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 27 March 2012
Received in revised form
10 October 2012
Accepted 19 October 2012
Communicated by X. Gao
Available online 20 November 2012

Keywords:
Quantile regression
Sparse property
Regularization theory
Empirical covering number
Reproducing kernel Hilbert spaces (RKHS)

ABSTRACT

Learning with coefficient-based regularization has attracted a considerable amount of attention recently in both machine learning and statistics. This paper presents a kernelized version of a quantile estimator integrated with coefficient-based regularization, which can be solved efficiently by a simple linear programming. Fast convergence rates are obtained under mild condition on the underlying distribution. Besides, this algorithm can be adapted easily to large scale problems and sparse solution is often achieved as that of Lasso. In our work we make the following main contributions: girst, improved learning rates are obtained by employing so called variance bounds, which is optimal in the literatures of learning theory; second, we establish stronger convergence rates by employing self-calibration inequalities; third, our learning rates can also be derived by a simple data-dependent parameter selection method; finally, the performance of the classical and our new algorithms are compared respectively in a simulation study and an actual problem.

1. Introduction

Quantile regression has emerged as a comprehensive approach for analyzing the impact of regressors on the conditional distribution of a response variable [12,13] and others. In this paper, we consider quantile regression from a learning theory viewpoints. We begin with a supervised learning problem over a collection of observational data $\mathbf{z}=\{(x_i,y_i)\}_{i=1}^m$, drawn from an unknown distribution ρ on $X\times Y$, where $x_i\in X\subset\mathbb{R}^n$ is an input space and $y_i\in\mathbb{R}$ is the corresponding output for the regression problem. Given a loss function $L: Y\times\mathbb{R}\to[0,\infty)$, we hope that the following risk functional

$$\mathcal{R}(f) := \int_{X\times Y} L(y,f(x))\, d\rho(x,y)$$

is small over all measurable functions on X, and the least square loss $(L(y,f(x))=(y-f(x))^2)$ is the most commonly used, which corresponds to the conditional mean function. The conditional mean regression describes the centrality of the conditional response distribution. However, sometimes one wants to obtain a good estimate satisfying a certain proportion of $y|x$ below the estimate. For example, for financial risk management, an investor may need to estimate a lower bound on the changes in the value of its portfolio with high probability, so as to take positive measure beforehand to avoid big asset volatility. In this case the mean function is no longer valid since it cannot provide more complete descriptions of the conditional response distribution. The above problem can be solved by means of quantile regression, which has an equivalent relationship with the so-called τ-pinball loss, defined by

$$L_\tau(y,t)=\begin{cases}(1-\tau)(t-y) & \text{if } y<t,\\ \tau(y-t) & \text{if } y\ge t.\end{cases} \tag{1.1}$$

Let $\rho(\cdot|x)$ be the conditional distribution of ρ, and with fixed any constant $\tau\in(0,1)$, the set-valued function is defined as

$$F^*_{\tau,\rho}(x) := \{t\in\mathbb{R} : \rho((-\infty,t]|x)\ge\tau \quad\text{and}\quad \rho([t,\infty)|x)\ge 1-\tau\},\quad x\in X.$$

If $\rho(\cdot|x)$ has finite support, it is well understood that $F^*_{\tau,\rho}(x)$ is a bounded and closed interval at any $x\in X$ in [20]. We write

$$t^*_{min}(x) := \min\{F^*_{\tau,\rho}(x)\} \quad\text{and}\quad t^*_{max}(x) := \max\{F^*_{\tau,\rho}(x)\},$$

which imply that $F^*_{\tau,\rho}(x)=[t^*_{min}(x),t^*_{max}(x)]$. Moreover, it is easy to check that the interior of $F^*_{\tau,\rho}(x)$ is a $\rho(\cdot|x)$-zero set, namely, $\rho((t^*_{min}(x),t^*_{max}(x))|x)=0$. In this paper we assume that $F^*_{\tau,\rho}(x)$ consists of singletons, namely, there exists a function $f^*_{\tau,\rho}: X\to\mathbb{R}$ called the conditional τ-quantile function, such that $F^*_{\tau,\rho}(x)=\{f^*_{\tau,\rho}(x)\}$ for almost every $x\in X$.

With the help of the τ-pinball loss, one can easily verify that $f^*_{\tau,\rho}$ is a minimizer of $\mathcal{R}(f)$ associated with L_τ, and for almost every $x\in X$

$$f^*_{\tau,\rho}(x)=\arg\min_{t\in\mathbb{R}}\int_Y L_\tau(y,t)\, d\rho(y|x).$$

Quantile regression is an important statistical methods for analyzing the impact of the regressors on the conditional distribution of a response variable. Compared with the conditional mean, one

* Corresponding author.
E-mail address: kenan716@mail.ustc.edu.cn (S.-G. Lv).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.10.015

of the advantages using quantile regression is that quantile regression is more robust in response to large outliers. In recent years, quantile regression has been widely used in many practical applications, such as reference charts in medicine [2,9] and economics [11]. For comprehensive reviews of quantile regression, refer to the articles by Koenker and Hallock [11,30] and well-written book by Koenker [12].

Throughout this paper, we assume that $\rho(\cdot|x)$ is supported on $[-M,M]$ for some constant $M>0$, and it follows that $|f^*_{\tau,\rho}(x)|\le M$ for $x\in X$ almost surely. A special case with $\tau=\frac{1}{2}$ of (1.1) gives the absolute value function, that is, $L_{1/2}(y,t)=\frac{1}{2}|y-t|$, and the minimizer $f^*_{1/2,\rho}$ becomes the median function $\mathcal{M}_\rho$ defined at each $x\in X$ by

$$\rho(y\ge\mathcal{M}_\rho(x)|x)\ge\frac{1}{2}\quad\text{and}\quad\rho(y\le\mathcal{M}_\rho(x)|x)\ge\frac{1}{2}.$$

Usually $f^*_{\tau,\rho}$ cannot be computed directly, due to the unknown distribution ρ. Instead, one seeks to minimize the *empirical error* associated with the sample $\mathbf{z}$ and τ-pinball loss, which is given by

$$\mathcal{R}_{\mathbf{z}}(f)=\frac{1}{m}\sum_{i=1}^{m}L_\tau(y_i,f(x_i)).$$

Unfortunately, minimizing $\mathcal{R}_{\mathbf{z}}(f)$ may lead to overfitting, that is, complex functions fit well on training data but not be able to generalize to unseen data. A promising approach to avoid this is to minimize the following regularized risk:

$$\min_{f\in\mathcal{H}}\{\mathcal{R}_{\mathbf{z}}(f)+\lambda\Omega(f)\},\tag{1.2}$$

where $\mathcal{H}$ is a pre-specified hypothesis function space and Ω is a functional on $\mathcal{H}$.

Kernel methods have been widely used in many areas of machine learning and achieved great success. Among others kernel regression has drawn much attention including quantile regression and least square regression. The study has focused on the application of Mercer kernels and regularization in the associated reproducing kernel Hilbert space (RKHS). Let $K:X\times X\to\mathbb{R}$ be a bounded, symmetric, and positive semi-definite function. The RKHS denoted by $\mathcal{H}_K$ associated with the kernel K is the completion of the linear span of functions $K_x:=K(x,\cdot),x\in X$ with the inner product given by $\langle K_x,K_y\rangle_{\mathcal{H}_K}=K(x,y)$. With these preparation, the learning algorithm for the quantile regression is given by the regularization scheme (see [23,16,20])

$$f_{\mathbf{z}}=\arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{m}\sum_{i=1}^{m}L_\tau(y_i,f(x_i))+\lambda\|f\|_K^2\right\},\tag{1.3}$$

where $0<\lambda\le1$, a regularization parameter controlling the trade-off between the empirical error and the penalty. In practice, we need to choose an adaptive parameter λ which usually depends on the sample, and probably the cross validation approach is the most commonly used technique aiming at this, but burdens upper computational complexity when the sample size is large. To this end, in Section 5 we will employ a standard training-validation approach to choose a suitable λ, also we give the convergence rate of our proposed algorithm (1.4) below based on such adaptive parameter.

Note that the resulting function $f_{\mathbf{z}}$ generated by (1.3) is given as a dense expansion in terms of the training patterns, which adds to computational cost greatly. To solve large-data problems, a promising fact is that sparse solutions are often achieved in statistics and compress sensing settings by imposing an L^1-regularizer on the expansion coefficients [3,18,24]. The L_1-norm penalty not only shrinks the fitted coefficient toward zero, but also causes some of the fitted coefficient to be exactly zero when λ is chosen to be large enough. Thus a lot of irrelevant noise variable or useless data can be removed mostly. In order to introduce L^1-regularizer in a machine learning setup, we need to study the coefficient-based regularization scheme, which can be written as

$$f_{\mathbf{z},\lambda}=\arg\min_{f_\alpha\in\mathcal{H}_{K,\mathbf{z}}}\{\mathcal{R}_{\mathbf{z}}(f_\alpha)+\lambda\Omega_{\mathbf{z}}(f_\alpha)\},\tag{1.4}$$

where

$$\mathcal{H}_{K,\mathbf{z}}=\left\{f_\alpha(x)=\sum_{i=1}^{m}\alpha_iK(x,x_i):\alpha=(\alpha_1,\ldots,\alpha_m)\in\mathbb{R}^m\right\}$$

and

$$\Omega_{\mathbf{z}}(f_\alpha)=\sum_{i=1}^{m}|\alpha_i|\quad\text{for } f_\alpha(x)=\sum_{i=1}^{m}\alpha_iK(x,x_i).$$

Algorithm (1.4) is a linear programming which can be solved efficiently by existing codes for large scale problems. For completeness, we also give the concrete optimization procedure in Section 6. This approach is similar to Linear Programming Regularization by proposed Smola et al. [23]. However, the main difference lies in that some convex function class substitutes the set of functions $K(\cdot,x_i)$ here, rather than acting automatically on data point x_i. It is worth noting that our choice of basis functions $K(x,x_i)$ is more natural since the solution of (1.3) can be found in a finite-dimensional space spanned by the set of functions $\{K(x,x_i)\}_{i=1}^m$ [16,23]. Also we find later that this ensures good generalization ability theoretically.

It is worth noting that some previous works have considered the format: L_1 loss + L_1 penalty structure. For example, Koenker et al. [14] used $\lambda\int_0^1|f''(x)|\,dx$ as the penalty, namely,

$$\min_{f\in\mathcal{F}}\sum_{i=1}^{m}L_\tau(y_i,f(x_i))+\lambda\int_0^1|f''(x)|\,dx,$$

where $\mathcal{F}$ is a certain function space. As is shown that the solution is a linear spline with knots x_i $(i=1,\ldots,m)$ if $\mathcal{F}$ is chosen appropriately. This motivates partially us to study the coefficient-based regularization scheme, considering that smoothing spline function can be viewed as a special kernel.

In the field of machine learning, the coefficient regularization was first introduced in [6] to design linear programming support vector machine. In [7], the sparse property of estimation coefficients with least square regression was discussed by a spectral decomposition technique. In terms of theoretical analysis, [27] derived some learning rates by using a local polynomial reproduction formula in approximation theory. In their analysis, the quantile regression was paid especially under rather general conditions. Ref. [17] provided a unified analytical framework for coefficient-based regularization with strict convex regularizer and indefinite kernel. As we notice, the previous works mentioned above did not consider the effect of the conditional distribution to the convergence rates. In this paper, we can improve the corresponding learning rates sharply by making fully use of the information on the conditional distribution and the so-called comparison theorem first proposed in [32]. In addition, we apply the empirical covering numbers to measure the functional complexity instead of uniform covering numbers [27]. This ensures the entropy integral finiteness with respect to the space $\mathcal{H}_{K,\mathbf{z}}$, which is shown in Section 4.

From the algorithmic point of view, one expects that (1.4) would be more stable and computational efficiency, and most importantly $f_{\mathbf{z},\lambda}$ can approximate the target function $f^*_{\tau,\rho}$ well in the whole space X as the sample size m tends to infinity. Error analysis for regularization (1.4) aims at estimates of the excess generalization error $\mathcal{R}(f_{\mathbf{z},\lambda})-\mathcal{R}(f^*_{\tau,\rho})$ in terms of the kernel K, the τ-pinball loss and the underlying measure ρ through the proper choice of the regularization parameter λ. However, this only implies that $f_{\mathbf{z},\lambda}$ is close to $f^*_{\tau,\rho}$ in a weak form. Recently, some stronger convergence rates are derived from establishing so called self-calibration inequalities (see [20,25]).