Journal of Computational and Applied Mathematics 261 (2014) 394–405

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

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

Least squares regression with l_1-regularizer in sum space

Yong-Li Xu [a,*], Min Han [b], Xue-mei Dong [c], Min Wang [d]

[a] Department of Mathematics, Beijing University of Chemical Technology, Beijing 100029, China
[b] College of Applied Science, Beijing University of Technology, Beijing 100024, China
[c] School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, 310018, China
[d] National Association of Financial Market Institutional Investors, Beingjing, China

ARTICLE INFO

Article history:
Received 30 January 2013
Received in revised form 3 October 2013

Keywords:
Learning theory
Least square regression
Regularization scheme
Sum space
Error analysis

ABSTRACT

In this paper, we propose a least squares regularized regression algorithm with l_1-regularizer in a sum space of some base hypothesis spaces. This sum space contains more functions than single base hypothesis space and therefore has stronger approximation capability. We establish an excess error bound for this algorithm under some assumptions on the kernels, the input space, the marginal distribution and the regression function. For error analysis, the excess error is decomposed into the sample error, hypothesis error and regularization error, which are estimated respectively. From the excess error bound, convergency and a learning rate can be derived by choosing a suitable value of the regularization parameter. The utility of this method is illustrated with two simulated data sets and one real life database.

1. Introduction

Let X be a compact metric space and $Y = \mathbb{R}$. Suppose that ρ is a fixed (but unknown) probability distribution on $Z := X \times Y$. The regression function is defined as

$$f_\rho(x) = \int_Y y d\rho(y|x), \quad x \in X,$$

where $\rho(y|x)$ is the conditional probability measure at x induced by ρ. The error for a function $f : X \to Y$ with squared loss is defined as

$$\mathcal{E}(f) = \int_Z (f(x) - y)^2 d\rho.$$

It is well known that the regression function minimizes the error. Indeed,

$$\|f - f_\rho\|_{\rho_X}^2 = \mathcal{E}(f) - \mathcal{E}(f_\rho),$$

where ρ_X is the marginal distribution of ρ on X and $\|f\|_{\rho_X}^2 = \int_X |f(x)|^2 d\rho_X$. The above difference is called the excess error of f.

The task is to find a good approximation $f_{\mathbf{z}}$ to f_ρ from a set of samples $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ which are drawn independently and identically distributed according to ρ. However, approximating a function from sparse samples is an ill-posed problem.

* Corresponding author. Tel.: +86 13552701365.
E-mail address: xuyongli2312@sina.com (Y.-L. Xu).

0377-0427/$ – see front matter
http://dx.doi.org/10.1016/j.cam.2013.11.029

To deal with it, a regularization technique is needed [1–3]. Given a set of functions $\mathcal{H}$ from X to Y called hypothesis space and a penalty functional $\Omega : \mathcal{H} \to R+$ called the regularizer, it searches for an approximation of f_ρ by the following scheme:

$$f_{\mathbf{z},\mathcal{H}} = \arg\min_{f\in\mathcal{H}} \{\mathcal{E}_{\mathbf{z}}(f) + \lambda\Omega(f)\},$$

where $\mathcal{E}_{\mathbf{z}}(f)$ is the empirical error

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2,$$

and $\lambda > 0$ is a regularization parameter.

In learning theory, the hypothesis space is commonly chosen as reproducing kernel Hilbert space (RKHS). Let $K : X \times X \to \mathbb{R}$ be continuous, symmetric and positive semidefinite. Such a function is called a Mercer kernel. The reproducing kernel Hilbert space $\mathcal{H}_K$ associated with the kernel K is defined to be the closure of the linear span of the set of functions $\{K_x := K(x,.) : x \in X\}$ with the inner product $\langle .,.\rangle_{\mathcal{H}_K} = \langle .,.\rangle_K$, satisfying $\langle K_x, K_{x'}\rangle_K = K(x,x')$. The reproducing property takes the form $\langle K_x, f\rangle_K = f(x)$, for $x \in X$, and $f \in \mathcal{H}_K$. The regularization in $\mathcal{H}_K$ with the norm square regularizer is given as

$$f_{\mathbf{z},\mathcal{H}_K} = \arg\min_{f\in\mathcal{H}_K}\left\{\mathcal{E}_{\mathbf{z}}(f) + \lambda\|f\|_K^2\right\}.$$

It has been well understood due to a lot of studies [4–6].

Recently, the coefficient-based regularizer has attracted more and more attention. Xiao and Zhou proposed a regularization scheme with l_1-regularizer [7]. They defined a data dependent hypothesis space and chose the l_1 norm of the coefficients as the regularizer. The l_1-regularizer is attractive because of its sparse property in experiments, i.e., most coefficients in the solution vanish. Shi et al. studied the sparsity of this algorithm in theory based on their error analysis [8]. Sun and Wu studied regularized learning schemes with l_2-regularizer and obtained faster learning rates than that with l_1-regularizer while their assumptions were less restrictive [9]. Tong et al. considered the coefficient-based regularized least-squares regression problem with the l_p-regularizer for $1 \le p \le 2$ [10].

All the above-mentioned coefficient-based regularization algorithms are implemented in a hypothesis space generated by a single kernel function. However, in some complicated cases, it is found that kernel methods with a single kernel function cannot meet some practical requirements such as heterogeneous information, unnormalized data, large scale problems and non-flat distribution of samples [11–14]. In these cases, multiple kernel methods are needed, which search for a linear combination of base kernel functions [15,16].

In this paper, we consider a regularized least-squares regression problem with l_1-regularizer in a hypothesis space trained from samples by multi-scale kernels. Let $\{K^j\}_{j=1}^l$ be a sequence of Mercer kernel functions. The hypothesis space on sample set $\mathbf{z}$ is defined as

$$\mathcal{F}_{\oplus,\mathbf{z}} = \left\{\sum_{j=1}^{l}\sum_{k=1}^{m}\alpha_k^j K_{x_k}^j : \alpha_k^j \in R\right\}.$$

For $f = \sum_{j=1}^{l}\sum_{k=1}^{m}\alpha_k^j K_{x_k}^j$, define the regularizer as:

$$\Omega_{\mathbf{z}}(f) = \sum_{j=1}^{l}\sum_{k=1}^{m}|\alpha_k^j|.$$

The learning algorithm is given by

$$f_{\oplus,\mathbf{z},\lambda} = \arg\min_{f\in\mathcal{F}_{\oplus,\mathbf{z}}}\{\mathcal{E}_{\mathbf{z}}(f) + \lambda\Omega_{\mathbf{z}}(f)\},$$

where $\lambda > 0$ is a regularization parameter.

In [7], a hypothesis space was proposed, which depended on samples. In our setting, the hypothesis space is a sum space of multiple base hypothesis spaces in [7]. This sum space contains more functions than every base hypothesis space and therefore have a stronger approximation capability. Every hypothesis function is determined by its $l \times m$ coefficients and the penalty is imposed on all these coefficients. In fact, we have

$$f_{\oplus,\mathbf{z},\lambda} = \arg\min_{\alpha_k^j\in R}\left\{\frac{1}{m}\sum_{i=1}^{m}\left(\sum_{j=1}^{l}\sum_{k=1}^{m}\alpha_k^j K_{x_k}^j(x_i) - y_i\right)^2 + \lambda\sum_{j=1}^{l}\sum_{k=1}^{m}|\alpha_k^j|\right\}.$$

For a non-flat function approximation problem, multi-scale kernels learning is more efficient, in which the kernels with small and large scales can deal with the steep and smooth variations, respectively [17,18,14,19]. The multi-scale kernels could be chosen as Gaussian kernels with different widths, wavelet-based kernels, frame-based kernels, and so on. In numerical experiments, existing multi-scale kernel methods perform better than single kernel methods in reducing prediction errors. However, there are few papers that have considered convergency and learning rates for multi-scale kernels methods.