Appl. Comput. Harmon. Anal. 37 (2014) 36–51

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Convergence rates of learning algorithms by random projection ☆

Di-Rong Chen [a,*], Han Li [a,b]

[a] Department of Mathematics, Beijing University of Aeronautics and Astronautics, Beijing 100191, China
[b] Department of Mechanical Engineering, Toyohashi University of Technology, Toyohashi, Aichi 441-8580, Japan

ARTICLE INFO

Article history:
Received 28 October 2011
Received in revised form 29 August 2013
Accepted 8 September 2013
Available online 16 September 2013
Communicated by Ding-Xuan Zhou

Keywords:
Random projection
Dimensionality reduction
Reproducing kernel Hilbert spaces
Learning rate

ABSTRACT

Random projection allows one to substantially reduce dimensionality of data while still retaining a significant degree of problem structure. In the past few years it has received considerable interest in compressed sensing and learning theory. By using the random projection of the data to low-dimensional space instead of the data themselves, a learning algorithm is implemented with low computational complexity. This paper investigates the accuracy of the algorithm of regularized empirical risk minimization in Hilbert spaces. By letting the dimensionality of the projected data increase suitably as the number of samples increases, we obtain an estimation of the error for least squares regression and support vector machines.

1. Introduction

In real world, data items, such as document analysis, hyperspectral image processing, are often curves and functions rather than standard vectors. This often leads to the curse of dimensionality. The curse of dimensionality can be a big obstacle in machine learning threatening the accuracy and computation complexity. To overcome these problems, some dimensionality reduction techniques are developed, among which is the method of random projection. Its key idea arises from the Johnson–Lindenstrauss lemma [3,17], which states that any m points in a Euclidean space can be embedded in $\mathbb{R}^n$ without distorting the distances between any pair of the m points by more than a factor $1 \pm \epsilon$ provided $n \gtrsim \epsilon^{-2} \log m$. The Johnson–Lindenstrauss lemma still holds if the m points are in a Hilbert spaces [2]. The method of random projection has been found substantial use in the area of algorithm design, by allowing one to substantially reduce dimensionality while still retaining a significant degree of problem structure.

In the past few years the method of random projection has received considerable interest in compressed sensing and in learning theory. A fundamental fact in compressed sensing is that, with high probability, one can recover a high-dimensional vector x by a relatively small number of random projection of x, if it is sparse [8,15], i.e., many components of x are zero. In other words, a sparse vector in high-dimensional space may be recovered, with high probability, from its random projection onto lower dimensional spaces.

Biau, Devroye and Lugosi [2] designed a clustering algorithm in Hilbert spaces based on Johnson–Lindenstrauss random projections and argued that the random projection works better than other simplistic dimension reduction schemes for clustering. The compressed learning algorithm introduced in [6] for dealing with sparse signals is also based on random

☆ Research supported in part by NSF of China under grants 11171014 and 91130009, National Basic Research Program of China under grant 973-2010CB731900, and National Technology Support Program under grant 2012BAH05B01.

* Corresponding author.
E-mail addresses: drchen@buaa.edu.cn (D.-R. Chen), lihan0809@gmail.com (H. Li).

1063-5203/$ – see front matter
http://dx.doi.org/10.1016/j.acha.2013.09.003

projection method. It is pointed out in [6] that the linear SVM classifier in the compressed domain performs almost as well as the best linear classifier in the data domain.

This paper considers learning problems in Hilbert space. As in [2,6], we first project randomly the dataset in Hilbert space into some lower dimensional space $\mathbb{R}^n$, and then perform the algorithms of regularized empirical risk minimization, with general loss functions, in $\mathbb{R}^n$. The problem of accuracy of the proposed algorithms arises naturally. We tackle it by estimating the generalization errors of the algorithms. An estimation bound, dependent of ϵ, m and regularizer parameter λ, is established for generalization errors provided $n \gtrsim \epsilon^{-2} \log m$. The asymptotic property of the error bounds is investigated as m tends to infinity. For least squared loss and hinge loss, by suitably choosing ϵ and λ according to m, we obtain convergence rates $(\frac{1}{m})^s$ of the generalization errors. It is worth to note that the reduced dimensionality n needs only to satisfy $n \gtrsim m^\gamma \log m$, $\gamma < 1$. We have not been aware of such a quantitative result for the generalization errors so far. And it is important to point out that, in our algorithms, the infinite-dimensional data are reduced to the finite-dimensional data. Thus, Johnson–Lindenstrauss type random projection is an effective tool for dimension reduction. Experiments performed on a real database demonstrate the effectiveness of our algorithms.

One of the main approaches is an extension of Johnson–Lindenstrauss lemma to kernels. This is also of independent interest. Usually a kernel $K(x, x')$ is introduced to measure the similarity, in some sense, between x and x'. We are restricted with kernels of forms $h(\langle \cdot, \cdot \rangle)$ and $h(\|x - x'\|)$, where h is a suitable function.

The paper is organized as follows. In Section 2, after a brief review of algorithms of empirical risk minimization, we propose the learning algorithms with random projection of data. In Section 3, we introduce some notions and present our main results. Furthermore, by letting some parameters change with the number m of samples, we obtain learning rates of least square regression and SVM. Section 4 contains the proofs of the main results and examples. Finally, experiments are given in Section 5 to illustrate our results.

2. Learning with random projection of data

We first review the algorithms of regularized empirical risk minimization and then propose the learning algorithms with random projection.

Let $(\mathbb{H}, \|\cdot\|)$ be a separable Hilbert space (possibly infinite dimensional), and let $X \subseteq \{x \in \mathbb{H}\colon \|x\| \leqslant B\}$ and $Y \subseteq \mathbb{R}$, where B is a positive number. Denote by $Z = \{(x, y)\colon x \in X, y \in Y\}$ the instance space. Let $\rho(x, y)$ be the unknown probability distribution on $Z = X \times Y$ describing the relation between $x \in X$ and $y \in Y$ and $\rho_X(x)$ the marginal probability distribution of Z on X. Note that $\rho(x, y)$ and $\rho_X(x)$ are related via $\rho(x, y) = \rho(y|x)\rho_X(x)$ where $\rho(y|x)$ is the conditional probability distribution on Y.

Let $\ell : \mathbb{R}^2 \to [0, +\infty)$ be a loss function. The target function f^* that we want to learn or approximate is a minimizer (may not be unique) of the error risk functional, referred as to generalization error [22],

$$\mathcal{E}(f) = E\ell\big(y, f(x)\big) = \int_Z \ell\big(y, f(x)\big)\, d\rho(x, y) \tag{1}$$

over the set of all measurable functions. That is

$$f^* = \arg\min \mathcal{E}(f). \tag{2}$$

Some specific loss functions are considered in learning algorithms. For regression we usually use

- the square loss $\ell(y, f(x)) = (y - f(x))^2$;
- the absolute value loss $\ell(y, f(x)) = |y - f(x)|$ and
- the ϵ-insensitive loss $\ell(y, f(x)) = |y - f(x)|_\epsilon := \max\{|y - f(x)| - \epsilon, 0\}$.

For classification, $Y = \{1, -1\}$, we usually use

- the square loss $\ell(y, f(x)) = (1 - yf(x))^2$ and
- the hing loss $\ell(y, f(x)) = (1 - yf(x))_+ := \max\{1 - yf(x), 0\}$.

The target function f^* cannot be obtained exactly since the probability distribution $\rho(x, y)$ is unknown. Instead, we give a set of samples $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ drawn independently according to ρ and f^* is learned from these samples. Here we will consider the regularized empirical risk minimization (ERM) algorithm associated with reproducing kernel Hilbert space (RKHS) [1,13], which is formulated as follows.

Let $K : X \times X \to \mathbb{R}$ be continues, symmetric, and positive semi-definite, i.e., given any finite set $\{x_1, \ldots, x_m\} \subset X$ of points, the matrix $(K(x_i, x_j))_{i,j=1}^m$ is positive semi-definite. Such a function is called a Mercer Kernel. The RKHS $\mathcal{H}_K$ associated with the kernel K is defined to be the closure of the linear span of the set of functions $\{K_x := K(x, \cdot)\colon x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_K} = \langle \cdot, \cdot \rangle_K$, satisfying $\langle K_x, K_y \rangle_K = K(x, y)$. The reproducing property takes the form

$$f(x) = \langle f, K_x \rangle_K, \quad \forall x \in X, f \in \mathcal{H}_K.$$