Mathematical and Computer Modelling 57 (2013) 1395–1407

Contents lists available at SciVerse ScienceDirect

Mathematical and Computer Modelling

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

Learning performance of elastic-net regularization

Yu-long Zhao *, Yun-long Feng

Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China

ARTICLE INFO

Article history:
Received 13 June 2011
Received in revised form 22 November 2012
Accepted 25 November 2012

Keywords:
Learning theory
Elastic-net regularization
ℓ^2-empirical covering number
Learning rate

ABSTRACT

In this paper, within the framework of statistical learning theory we address the elastic-net regularization problem. Based on the capacity assumption of hypothesis space composed by infinite features, significant contributions are made in several aspects. First, concentration estimates for sample error are presented by introducing ℓ^2-empirical covering number and utilizing an iteration process. Second, a constructive approximation approach for estimating approximation error is presented. Third, the elastic-net learning with infinite features is studied and the role that the tuning parameter ζ plays is also discussed. Finally, our learning rate is shown to be faster compared with existing results.

1. Introduction and main results

During the last few decades, several regularized methods for linear regression have been adopted to overcome deficiencies of ordinary least square regression on prediction and interpretation. Shrinking coefficients toward zero, ridge regression [1] achieves better prediction performance through a bias-variance trade-off. However, ridge regression is not able to provide a sparse model which can be interpreted better since the coefficients are shrunken toward zero but never become zero exactly. Aiming at continuous shrinkage and automatic variable selection simultaneously, a penalized least squares method called LASSO is proposed [2] by imposing an ℓ^1-regularizer on regression coefficients. Different from ridge regression, coefficients in LASSO can be shrunken toward zero exactly, which leads to much better interpretability. However, in some special cases the LASSO also shows its deficiency, for example when the variables have group effects or the number of predictors is much larger than the number of observations [3]. Mindful of these flaws, a regularized regression scheme generated by a combination of the LASSO and ridge penalty is proposed. It was first introduced in [3] and then analyzed in [4]. It is demonstrated that the elastic net often outperforms LASSO and simultaneously preserves the sparse property [4,3]. The advantages of this regularization scheme have been also confirmed by various applications [4,3,5,6].

In this paper, we focus on the statistical properties of this scheme and in particular its consistency property, which is studied within the framework of statistical learning theory. To address this problem, we first present a mathematical setup, which follows the setting in [4].

The regression problem aims at learning a regression function on a separable metric space X (called the input space) with values in $Y = \mathbb{R}$. The elastic net algorithm is given in terms of finite set $\{\varphi_k\}_{k=1}^N$ of continuous functions on X with sufficiently large N, which is a subset of a dictionary $\{\varphi_k\}_{k\in\Gamma}$ with cardinality $|\Gamma|$ countable, where $|\Gamma| \geq N$. Its regularizer is an elastic net penalty on $\mathbb{R}^N$. In fact the learning algorithm can be extended to the infinite case, as we will explain later. We first present the definition of elastic net penalty.

* Corresponding author.
E-mail addresses: zhaoyulong@gmail.com, zyulong2@student.cityu.edu.hk (Y.-l. Zhao), yunlfeng@cityu.edu.hk (Y.-l. Feng).

0895-7177/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.mcm.2012.11.028

Definition 1. Let $\zeta > 0$, the *elastic net penalty* $p_\zeta : \mathbb{R}^N \to [0, \infty)$ is defined as

$$p_\zeta(\beta) = \sum_{k=1}^{N} \{|\beta_k| + \zeta \beta_k^2\}$$

where N is a positive integer.

The hypothesis space $\mathcal{H}_N$ for the regularization scheme consists of linear combinations of N features: $f_\beta = \sum_{k=1}^N \beta_k \varphi_k$. We adopt the setting introduced in [4], which also can be found in [7]. Explicitly, $\mathcal{H}_N = \left\{ f : f = \sum_{k=1}^N \beta_k \varphi_k \right\}$ is a subset of $\mathcal{H}_\Gamma$, which is defined as

$$\mathcal{H}_\Gamma = \left\{ f : f = \sum_{k \in \Gamma} \beta_k \varphi_k, \beta_k \in \mathbb{R} \right\}. \tag{1.1}$$

Then elastic net algorithm is now defined for a given sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in (X \times Y)^m$ by

$$f_{\mathbf{z}} = \arg\min_{f_\beta \in \mathcal{H}_N} \left(\frac{1}{m} \sum_{i=1}^{m} (f_\beta(x_i) - y_i)^2 + \lambda \mathcal{P}_\zeta(f_\beta) \right), \tag{1.2}$$

where $\lambda = \lambda(m) \geq 0$ is a regularization parameter and $\mathcal{P}_\zeta(f_\beta) := p_\zeta(\beta)$.

As mentioned above, in this paper we are interested in the learning ability of the algorithm (1.2). To this end, we take a common model in learning theory and assume that ρ is a Borel probability measure on $Z := X \times Y$ and the *regression function* is defined by

$$f_\rho(x) = \int_Y y d\rho(y|x), \quad x \in X, \tag{1.3}$$

where $\rho(\cdot|x)$ is the conditional probability measure induced by ρ at $x \in X$.

In the supervised learning framework, ρ is unknown and one cannot obtain the regression function f_ρ directly. Indeed, we learn the regression function from the sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in Z^m$, which is assumed to be drawn independently according to the measure ρ. Throughout this paper, we assume that $\rho(\cdot|x)$ is supported on $[-M, M]$, for some $M > 0$. The learning ability of the algorithm (1.2) is measured by the error $\|f_{\mathbf{z}} - f_\rho\|_{L^2_{\rho_X}}$ of the difference function $f_{\mathbf{z}} - f_\rho$ in the space $L^2_{\rho_X}$ where ρ_X is the marginal distribution of ρ on X.

Considering that the analysis in this paper is based on the complexity assumption of the hypothesis space, we need the following capacity condition for $\mathcal{H}_\Gamma$ in terms of ℓ^2-empirical covering numbers.

Definition 2. Let $(\mathcal{M}, d)$ be a pseudo-metric space and $S \subset \mathcal{M}$ a subset. For every $\epsilon > 0$, the covering number $\mathcal{N}(S, \epsilon, d)$ is defined as the minimal number of balls of radius ϵ whose union covers S, that is,

$$\mathcal{N}(S, \epsilon, d) = \min \left\{ \ell \in \mathbb{N} : S \subset \bigcup_{j=1}^{\ell} B(s_j, \epsilon) \text{ for some } \{s_j\}_{j=1}^{\ell} \subset \mathcal{M} \right\},$$

where $B(s_j, \epsilon) = \{s \in \mathcal{M} : d(s, s_j) \leq \epsilon\}$.

Let d_2 be the normalized metric on the Euclidian space $\mathbb{R}^n$ given by

$$d_2(\mathbf{a}, \mathbf{b}) = \left(\frac{1}{n} \sum_{i=1}^{n} |a_i - b_i|^2 \right)^{1/2} \quad \text{for } \mathbf{a} = \{a_i\}_{i=1}^n, \ \mathbf{b} = \{b_i\}_{i=1}^n \in \mathbb{R}^n.$$

Definition 3. Let $\mathcal{F}$ be a set of functions on X, $\mathbf{x} = (x_i)_{i=1}^n \subset X^n$ and $\mathcal{F}|_{\mathbf{x}} = \{(f(x_i))_{i=1}^n : f \in \mathcal{F}\} \subset \mathbb{R}^n$. Set $\mathcal{N}_{2,\mathbf{x}}(\mathcal{F}, \epsilon) = \mathcal{N}(\mathcal{F}|_{\mathbf{x}}, \epsilon, d_2)$. The ℓ^2-empirical covering number of $\mathcal{F}$ is defined by

$$\mathcal{N}_2(\mathcal{F}, \epsilon) = \sup_{n \in \mathbb{N}} \sup_{\mathbf{x} \in X^n} \mathcal{N}_{2,\mathbf{x}}(\mathcal{F}, \epsilon), \quad \epsilon > 0.$$

Assumption 1. The space $\mathcal{H}_\Gamma$ has empirical polynomial complexity with exponent p, where $0 < p < 2$. That is, there exists a constant $c_{p,\mathcal{H}_\Gamma} > 0$ such that

$$\log \mathcal{N}_2(B_1, \epsilon) \leq c_{p,\mathcal{H}_\Gamma} \left(\frac{1}{\epsilon} \right)^p, \quad \forall \epsilon > 0, \tag{1.4}$$

where B_1 is the subset of $\mathcal{H}_\Gamma$ defined by $B_R = \{f_\beta \in \mathcal{H}_\Gamma : \|\beta\|_{\ell^1} \leq R\} \cap \{f_\beta \in \mathcal{H}_\Gamma : \|\beta\|_{\ell^2} \leq \sqrt{\frac{R}{\zeta}}\}$ with $R = 1$.