Mathematical and Computer Modelling 54 (2011) 1347–1364

Contents lists available at ScienceDirect

Mathematical and Computer Modelling

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

Classification with non-i.i.d. sampling

Zheng-Chu Guo [a,b], Lei Shi [b,*]

[a] School of Mathematics and Computational Science, Sun Yat-sen University, Guangzhou 510275, PR China

[b] Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China

ARTICLE INFO

Article history:
Received 22 December 2010
Received in revised form 31 March 2011
Accepted 31 March 2011

Keywords:
Learning theory
Regularized classification
β-mixing sequence
Reproducing kernel Hilbert spaces
ℓ^2-empirical covering number
Capacity dependent error bounds

ABSTRACT

We study learning algorithms for classification generated by regularization schemes in reproducing kernel Hilbert spaces associated with a general convex loss function in a non-i.i.d. process. Error analysis is studied and our main purpose is to provide an elaborate capacity dependent error bounds by applying concentration techniques involving the ℓ^2-empirical covering numbers.

1. Introduction

In this paper, we consider learning algorithms for classification with non-i.i.d. sampling processes.

In a binary classification problem, the input space is a compact subset $X \subset \mathbb{R}^d$ and the outputs space $Y = \{-1, 1\}$ represents two classes. Classification algorithms produce binary classifiers $\mathcal{C} : X \to Y$. Let ρ be a probability measure defined on $Z := X \times Y$. The prediction ability of a classifier $\mathcal{C}$ is measured by the *misclassification error* which is defined as

$$\mathcal{R}(\mathcal{C}) = \mathrm{Prob}_{(x,y)\in(Z,\rho)}\{\mathcal{C}(x) \neq y\} = \int_X \rho_x(y \neq \mathcal{C}(x))\mathrm{d}\rho_X. \tag{1.1}$$

Here ρ_X is the marginal distribution of ρ on X and ρ_x is the conditional distribution at $x \in X$. The best classifier that minimizes the misclassification error is the *Bayes rule* given by

$$f_c(x) = \begin{cases} 1, & \text{if } \rho_x(y = 1) \geq \rho_x(y = -1), \\ -1, & \text{if } \rho_x(y = 1) < \rho_x(y = -1). \end{cases} \tag{1.2}$$

Since ρ_x is unknown, f_c cannot be computed directly. The goal of classification algorithms is to find classifiers which approximate f_c from a finite sample $\mathbf{z} = \{z_i = (x_i, y_i)\}_{i=1}^m \in Z^m$. The classifiers considered here are induced by real-valued functions $f : X \to \mathbb{R}$ as $\mathcal{C} = \mathrm{sgn}(f)$ which is defined by $\mathrm{sgn}(f)(x) = 1$ if $f(x) \geq 0$ and $\mathrm{sgn}(f)(x) = -1$ otherwise. We define a loss function $\phi : \mathbb{R} \to \mathbb{R}_+$ and use the error $\phi(yf(x))$ to measure the difference between the output y and the prediction $\mathrm{sgn}(f)(x)$.

Definition 1. A function $\phi : \mathbb{R} \to \mathbb{R}_+$ is called a classifying loss function if it is convex, differentiable at 0 with $\phi'(0) < 0$, and the smallest zero of ϕ is 1.

* Corresponding author.
E-mail addresses: gzhengchu@gmail.com (Z.-C. Guo), leishi@cityu.edu.hk, sl1983@mail.ustc.edu.cn (L. Shi).

0895-7177/$ – see front matter
doi:10.1016/j.mcm.2011.03.042