Statistics and Probability Letters 95 (2014) 33–38

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

An asymptotically minimax kernel machine

Debashis Ghosh *

Department of Biostatistics and Informatics, Colorado School of Public Health, United States

ARTICLE INFO

Article history:
Received 29 April 2014
Received in revised form 2 August 2014
Accepted 2 August 2014
Available online 11 August 2014

Keywords:
Data mining
Decision theory
Hard-thresholding
Nonparametric regression
Support vector machines

ABSTRACT

Recently, a class of machine learning-inspired procedures, termed kernel machine methods, has been extensively developed in the statistical literature. In this note, we construct a so-called 'adaptively minimax' kernel machine. Such a construction highlights the limits on the interpretability of such kernel machines.

1. Introduction

With the availability of massive datasets from scientific and medical disciplines, increasing attention is being paid to the use of data mining techniques. This has in turn sparked interest as to the statistical properties of the methodologies. One example is support vector machines (Cristianini and Shawe-Taylor, 2000). This is a supervised learning procedure that attempts to find a margin-maximizing hyperplane that separates two groups. Liu et al. (2007) developed a statistical framework and equivalence in which the support vector machine regression with a continuous outcome is identical to a certain mixed effects model. This equivalence is reviewed in Section 2. The kernel machine has been utilized heavily with applications to genomics (Liu et al., 2007, 2008; Kwee et al., 2008; Cai et al., 2011; Wu et al., 2010, 2011; Pan, 2011; Kim et al., 2012) and imaging (Ge et al., 2012). These articles typically show power gains for the kernel machine-based tests relative to their fixed-effects counterparts due to shrinkage brought on by the use of random effects.

Given the recent popularity of the kernel machine methodology, it is important to understand its theoretical foundations. Many of the previous authors have used estimation and attendant inference results from the mixed model framework. In this article, we seek to offer a viewpoint on the kernel machine methodology. The concept of minimaxity is well-studied in statistics and has also been addressed in the context of nonparametric density estimation and regression problems by many authors (e.g., Fan, 1992). For recent problems in high-dimensional data analysis, there is a lot of interest in understanding rates of convergence for minimax estimators (e.g., Cai and Zhou, 2012; Birnbaum et al., 2013; Dicker, in press; Raskutti et al., 2011).

In this paper, the adaptively minimax estimator will be very different in structure relatively to previously studied adaptive kernel methods. The paper proceeds as follows. In Section 2 we review the kernel machine methodology as previously developed in the literature. Section 3 features the construction of the adaptive kernel machines using simple thresholding ideas, for which we prove asymptotic minimaxity results. Section 4 concludes with some discussion.

* Correspondence to: Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Center, 30000 17th Avenue East, Aurora, CO 80045, United States. Tel.: +1 303 734 4365.
E-mail address: debashis.ghosh@ucdenver.edu.

http://dx.doi.org/10.1016/j.spl.2014.08.005
0167-7152/

2. Kernel machine methodology: model and estimation

We first review the kernel machine framework of Liu et al. (2007). For the sake of exposition, we will work in the case that there is no parametric component. For i ($i = 1, \ldots, n$), we observe $(Y_i, \boldsymbol{X}_i)$, where Y_i is a normally distributed continuous outcome, and $\boldsymbol{X}_i$ is a $p \times 1$ vector of covariates. We assume the following model:

$$Y_i = \beta_0 + h(\boldsymbol{X}_i) + e_i, \tag{1}$$

where β_0 is an intercept term, $h(\boldsymbol{z}_i)$ is an unknown centered smooth function, and the errors e_i are assumed to be independently and identically distributed from a $N(0, \sigma^2)$ distribution. Assume that y_i ($i = 1, \ldots, n$) are centered so that β_0 drops out of the model (1).

One issue that arises in (1) is how to specify basis functions for h, especially in the case of high-dimensional $\mathbf{X}$. The advantage of kernel methods as defined in machine learning contexts is that one specifies a kernel function $K(\mathbf{x}, \mathbf{x}')$ instead of the basis functions. Specifically, a kernel function $K(\mathbf{x}, \mathbf{x}')$ is a bounded, symmetric, positive function satisfying

$$\int K(\mathbf{x}, \mathbf{x}')g(\mathbf{x})g(\mathbf{x}')d\mathbf{x}d\mathbf{x}' \geq 0, \tag{2}$$

for any arbitrary square integrable function $g(\mathbf{x})$ and all $\mathbf{x}, \mathbf{x}' \in \mathbf{R}^p$. The kernel function can be viewed as a measure of similarity between the covariate vectors $\mathbf{x}$ and $\mathbf{x}'$. From Mercer's Theorem (Cristianini and Shawe-Taylor, 2000, p. 33), any kernel function satisfying some regularity conditions implicitly specifies a unique function space spanned by a particular set of basis functions (features), and vice versa. We note now that the conditions for a proper kernel function imply that the observed data matrix $\mathbf{K}$, with (i, j)th entry $K(\mathbf{X}_i, \mathbf{X}_j)$, will be positive definite.

Assume that the nonparametric function $h(\cdot) \in \mathcal{H}_K$, where $\mathcal{H}_K$ is a reproducing kernel Hilbert space (e.g., Wahba, 1990). Then there is a 1–1 correspondence between K and the corresponding RKHS. Estimation of $\boldsymbol{\beta}$ and $h(\cdot)$ proceeds by maximizing the scaled penalized likelihood function

$$J(h, \beta) = -\frac{1}{2}\sum_{i=1}^{n}[y_i - h(\mathbf{x}_i)]^2 - \frac{1}{2}\lambda\|h\|^2_{\mathcal{H}_K}, \tag{3}$$

where $\lambda > 0$ is a tuning parameter and controls the tradeoff between goodness of fit and complexity of the model. Exploiting a primal/dual equivalence from Karush–Kuhn–Tucker theory, one can show that the estimator of the nonparametric function $h(\cdot)$ evaluated at the design points $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ is estimated as

$$\widehat{\boldsymbol{h}} = \lambda^{-1}\boldsymbol{K}(\boldsymbol{I} + \lambda^{-1}\boldsymbol{K})^{-1}\boldsymbol{y}, \tag{4}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)$. In Liu et al. (2007), it was shown that the estimates of h in (4) can be derived as arising from a random effects model of the following form:

$$\boldsymbol{y} = \boldsymbol{h} + \boldsymbol{e}, \tag{5}$$

where $\boldsymbol{h}$ is an $n \times 1$ vector of random effects following $\boldsymbol{h} \sim N(\mathbf{0}, \tau\boldsymbol{K})$, τ is a scale parameter, and $\boldsymbol{e} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$. Because of this equivalence, all regression parameters in the model can be estimated by maximum likelihood, while the variance component parameters can be estimated by restricted maximum likelihood.

Remark 1. Liu et al. (2007, 2008) used the standard mixed effects model framework for estimation and attendant inference result. While they did not prove asymptotic normality results in that work, one could use Theorems 1 and 2 from Mardia and Marshall (1984) or the work of Cressie and Lahiri (1993) to derive consistency and asymptotic normality results for the kernel machine estimators of the fixed and random effects. Here, we will investigate different properties of the kernel machine relative to those studied by the previously mentioned authors.

3. Shrinkage, decision-theoretic framework and minimaxity

We will begin by assuming that $\lambda > 0$ is fixed. In addition, we assume the model (1) without the intercept holds and further that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are treated as fixed. We will use lower-case notation, i.e. $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, to denote this. We can write $\widehat{\boldsymbol{h}}$ from (4) as $\mathbf{H}(\mathbf{I} + \mathbf{H})^{-1}\mathbf{Y}$, where $\mathbf{H} = \lambda^{-1}\boldsymbol{K}$ and $\mathbf{I}$ is the $n \times n$ identity matrix. Thus, we have defined the kernel machine in terms of an operator $\mathbf{P}$ mapping from $\mathbf{R}^n$ to $\mathbf{R}^n$, where $\mathbf{P} = \mathbf{H}(\mathbf{I} + \mathbf{H})^{-1}$. By the positive definiteness of $\mathbf{K}$, $\mathbf{H}$ is also positive definite so that the following equivalences hold:

$$\mathbf{H} = \mathbf{U}\mathbf{D}_n\mathbf{U}' \tag{6}$$

$$\mathbf{I} + \mathbf{H} = \mathbf{U}(\mathbf{I} + \mathbf{D}_n)\mathbf{U}' \tag{7}$$

where $\mathbf{U}$ is an $n \times n$ orthonormal matrix such that $\mathbf{U}\mathbf{U}' = \mathbf{U}'\mathbf{U} = \mathbf{I}$, and $\mathbf{D}_n$ is a diagonal matrix with entries equaling the eigenvalues of $\mathbf{H}$. Because $\mathbf{H}$ is symmetric and positive definite, all the eigenvalues will be positive. Note that these results