Statistics and Probability Letters 139 (2018) 40–46

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Classification with incomplete functional covariates

Majid Mojirsheibani *, Crystal Shaw

Department of Mathematics, California State University Northridge, CA, 91330, USA

ARTICLE INFO

Article history:
Received 23 March 2017
Received in revised form 6 March 2018
Accepted 13 March 2018
Available online 4 April 2018

Keywords:
Classification
Pattern recognition
Functional covariates
Supervised learning

ABSTRACT

We consider the problem of functional classification when the covariate may be unavailable (unobservable) on some subsets of its domain. Given the observed fragments of the functional covariates, we propose a strongly consistent nonparametric classifier based on local averaging.

1. Introduction

This article focuses on the following standard two-group functional supervised classification problem. Let $(\mathbf{X}, Y)$ be a random pair, where $\mathbf{X}$ is a functional covariate taking values in a metric space $(\mathcal{F}, \rho)$ and $Y \in \{0, 1\}$, called the class variable, has to be predicted based on $\mathbf{X}$. In classification one seeks to find a function (a classifier) $g : \mathcal{F} \to \{0, 1\}$ whose misclassification error, i.e., the probability of incorrect prediction, $L(g) = P\{g(\mathbf{X}) \neq Y\}$, is as small as possible. Similar to the case where $\mathcal{F} = \mathcal{R}^d$, here the best classifier is given by

$$g_B(\mathbf{x}) = \begin{cases} 1 & \text{if } P\{Y = 1|\mathbf{X} = \mathbf{x}\} > \dfrac{1}{2} \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

i.e., g_B has the smallest error probability given by $L(g_B) = \inf_{g:\mathcal{F}\to\{0,1\}} P\{g(\mathbf{X}) \neq Y\}$; see, for example, Cérou and Guyader (2006), Abraham et al. (2006), and Devroye et al. (1996, Ch. 2). In practice the distribution of $(\mathbf{X}, Y)$ is almost always unknown and therefore finding the function g_B is impossible. However, one typically has available n independent and identically distributed (iid) observations, i.e., the data, $\mathbb{D}_n := \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$, where $(\mathbf{X}_i, Y_i) \overset{\text{iid}}{=} (\mathbf{X}, Y),\ i = 1, \ldots, n,$. The aim of classification is to construct a function $\widehat{g}$, based on the data $\mathbb{D}_n$, whose probability of error, defined by $L_n(\widehat{g}) = P\{\widehat{g}(\mathbf{X}) \neq Y|\mathbb{D}_n\}$, is in some sense small. We say a classification rule $\widehat{g}$ is consistent if $E[L_n(\widehat{g})] \to L(g_B)$, as $n \to \infty$. If $L_n(\widehat{g}) \to L(g_B)$, with probability one, then $\widehat{g}$ is said to be strongly consistent.

Many different methods have been proposed in the literature for the classification of functional data. These methods may be divided roughly into two approaches. (I) Those methods that are based on the entire curve $\mathbf{X}$; these include the kernel classification method of Ferraty and Vieu (2003), the naive kernel classifier of Abraham et al. (2006), the nearest neighbor approach of Cérou and Guyader (2006), the depth-based classifier of López-Pintado and Romo (2006), the robust functional classification of Cuevas et al. (2007), the wavelet approach of Chang et al. (2014), and the work of Meister (2016) on the optimality properties of kernel regression and classification for functional data. (II) Those approaches that are based

* Corresponding author.
E-mail addresses: majid.mojirsheibani@csun.edu (M. Mojirsheibani), crystal.shaw.81@my.csun.edu (C. Shaw).

https://doi.org/10.1016/j.spl.2018.03.010
0167-7152/

on the classification of *filtered curves* or, perhaps, a finite set of points on the curves $\mathbf{X}$. More specifically, when $\mathbf{X}$ belongs to an infinite-dimensional separable Hilbert space with inner product $\langle\cdot,\cdot\rangle$ then each predictor $\mathbf{X}_i$ can be expressed by the expansion $\mathbf{X}_i = \sum_{j=1}^{\infty}\xi_{ij}\phi_j$, where $\xi_{ij} = \langle \mathbf{X}_i, \phi_j\rangle$, and where $\{\phi_1, \phi_2, \dots\}$ is a complete orthonormal basis for the underlying space. Here, the random variables $\xi_{i1}, \xi_{i2}, \dots$ are viewed as the surrogates for the datum $\mathbf{X}_i$ in the sense that estimation based on $\mathbf{X}_i$ is equivalent to that based on $\xi_{i1}, \xi_{i2}, \dots$. Truncating the infinite sum after d terms, one is in effect replacing $\mathbf{X}_i$ by a d-dimensional vector $(\xi_{i1}, \dots, \xi_{id})$, where d typically increase with n. Relevant work along these lines includes the signal discrimination method of Hall et al. (2001), the functional classification method of Biau et al. (2005), the classification of gene expression data of Leng and Müller (2006) as well as that of Song et al. (2008), the wavelet approach of Berlinet et al. (2008), and the componentwise classification approach of Delaigle et al. (2012).

Our approach in this paper is more in the spirit of (I) above, where classification is carried out based on the entire curve (instead of filtered curves), except that we allow for some segments of the covariate curve $\mathbf{X}$ to be missing. In the rest of this paper we develop a framework for classification with possibly incomplete covariate curves; this is discussed in Section 2. Here we present the form of the optimal classifier for the current setup and proceed to construct strongly consistent kernel classifiers. A supplementary file available online contains the proofs and additional details. Numerical examples are given in Section 3.

2. Incomplete data

2.1. The setup

The great majority of existing functional classification methods are based on the implicit assumption that the covariate functions are fully observable on their domain. Here we consider the situation where parts of $\mathbf{X}$ may be unavailable, i.e., the situation where one may only be able to observe certain fragments of the full curve $\mathbf{X}$. Results along these lines are available for the simpler case of $\mathbf{X} \in \mathbb{R}^p$, $p > 1$; see, for example, Pawlak (1993), Mojirsheibani and Montazeri (2007), Reese and Mojirsheibani (2017), and Demirdjian and Mojirsheibani (2017).

Let $(\Omega, \mathcal{A}, P)$ be the underlying probability space. We take $\mathcal{F}$ to be the space of absolutely integrable functions defined on an interval of the real line, i.e., $\mathbf{X}$ is a random function on $(\Omega, \mathcal{A}, P)$ with values (sample paths) in $L^1(\mathcal{I})$, where $\mathcal{I}$ is an interval on the real line. However, instead of observing the full curve $\mathbf{X} : \Omega \to L^1(\mathcal{I})$, one might only be able to observe the fragment(s) of the curve denoted by $\mathbf{X}|_s$, i.e., the restriction of $\mathbf{X}(t)$ to $t \in s$, for some $s \subseteq \mathcal{I}$. To present the structure of the missing patterns (in the functional covariates), we follow the missing data setup of Bugni (2012). More specifically, we assume that each covariate function $\mathbf{X}$ is such that for a fine enough partition of $\mathcal{I}$ into $J < \infty$ subintervals $\mathcal{I}_1, \dots, \mathcal{I}_J$, each sample function of $\mathbf{X}$ is either completely observed or completely unobserved within each of these J subintervals. Here $\mathcal{I}_i \cap \mathcal{I}_j = \emptyset$ whenever $i \neq j$, and $\cup_{j=1}^{J}\mathcal{I}_j = \mathcal{I}$. Examples of such functional variables can be found in, for example, Bugni (2012, page 965). In what follows, we also denote by s_k the subset of $\mathcal{I}$ where a covariate corresponding to the kth missing pattern is observed. Without loss of generality we put $s_1 = \mathcal{I}$, which corresponds to the case where the functional predictor $\mathbf{X}$ is observable over the entire interval $\mathcal{I}$. Next, let Δ be the $\{1, \dots, M\}$-valued random variable defined by $\Delta = k$, if pattern k is observed, $k = 1, \dots, M$, where $M < 2^J$; in practice M is substantially smaller that 2^J. Thus if $\mathbb{I}\{\Delta = k\} = 1$ then one observes $\mathbf{X}|_{s_k}$, i.e., the restriction of $\mathbf{X}(t)$ to $t \in s_k$. If we denote the observed functional covariate by $\mathbf{X}^*$ (which is one, and only one, of $\mathbf{X}|_{s_1}, \dots, \mathbf{X}|_{s_M}$) then one can equivalently write $\mathbf{X}^* = \sum_{k=1}^{M}\mathbb{I}\{\Delta = k\} \cdot \mathbf{X}|_{s_k}$. We can also represent the data according to $\mathbb{D}_n = \{(\Delta_1, \mathbf{X}_1^*, Y_1), \dots, (\Delta_n, \mathbf{X}_n^*, Y_n)\}$. In this paper, we allow the presence of incomplete covariates (incomplete curves) in both the data $\mathbb{D}_n$ and the unclassified observation $\mathbf{X}$. In passing we also note that since s_k, $k = 1, \dots, M$, is the subset of $\mathcal{I}$ associated with the kth missing pattern, a non-data-based classifier corresponding to pattern k is any function of the form $g_k : L^1(s_k) \to \{0, 1\}$. Thus, with M missing patterns, any classifier is necessarily of the form

$$\psi(\mathbf{X}^*, \Delta) = \sum_{k=1}^{M}\mathbb{I}\{\Delta = k\} \cdot g_k(\mathbf{X}|_{s_k}), \tag{2}$$

for some $g_k : L^1(s_k) \to \{0, 1\}, k = 1, \dots, M$. Now suppose that g_k is the best classifier corresponding the kth missing pattern, i.e.,

$$g_k(\mathbf{x}|_{s_k}) = \begin{cases} 1 & \text{if } E\big[Y \big| \mathbf{X}(t) = \mathbf{x}(t),\ \forall t \in s_k\big] > \dfrac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Then one may be tempted to consider ψ in (2), with g_k's as in (3), as an intuitively reasonable candidate for the theoretically optimal classifier in the current setup with missing covariate segments. Unfortunately, this is not true in general. To find the optimal classifier, let

$$T_k(\mathbf{x}|_{s_k}) = E\left[(2Y - 1)\mathbb{I}\{\Delta = k\}\big|\mathbf{X}(t) = \mathbf{x}(t),\ \forall t \in s_k\right], k = 1, \dots, M, \tag{4}$$

and define the classifier

$$\psi_B(\mathbf{X}^*, \Delta) = \sum_{k=1}^{M}\mathbb{I}\{\Delta = k\} \cdot \mathbb{I}\{T_k(\mathbf{X}|_{s_k}) > 0\}, \tag{5}$$