Computational Statistics and Data Analysis 78 (2014) 176–185

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Variable and boundary selection for functional data via multiclass logistic regression modeling

Hidetoshi Matsui

Faculty of Mathematics, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan

ARTICLE INFO

Article history:
Received 8 April 2013
Received in revised form 17 April 2014
Accepted 22 April 2014
Available online 2 May 2014

Keywords:
Functional data analysis
Lasso
Logistic regression model
Model selection
Regularization

ABSTRACT

Penalties with an ℓ_1 norm provide solutions in which some coefficients are exactly zero and can be used for selecting variables in regression settings. When applied to the logistic regression model, they also can be used to select variables which affect classification. We focus on the form of ℓ_1 penalties in logistic regression models for functional data, in particular, their use in classifying functions into three or more groups while simultaneously selecting variables or classification boundaries. We provide penalties that appropriately select the variables in functional multiclass logistic regression models. Analysis of simulation and real data show that the form of the penalty should be selected in accordance with the purpose of the analysis.

1. Introduction

Variable selection is a crucial issue in regression analysis. Several methods have been proposed for the accurate and effective selection of appropriate variables (see, e.g., Burnham and Anderson, 2002). The lasso by Tibshirani (1996) and its extensions or refinements (Fan and Li, 2001; Zou and Hastie, 2005; Zou, 2006) provide a unified approach to problems of estimating and selecting variables, and for this reason they are broadly applied in several fields; an overview is provided in Hastie et al. (2009). In this paper, we consider the problem of classifying data while simultaneously selecting variables which affect the classification problem, by applying ℓ_1-type penalties to logistic regression models. The logistic regression model is one of the most useful tools for classifying data, and it does so by providing posterior probabilities which place the data in the appropriate group (McCullagh and Nelder, 1989).

Logistic regression models that use ℓ_1 regularization have been investigated as generalized linear models in Park and Hastie (2007). They considered binomial logistic regression models, and we consider classifying data into three or more groups using the multinomial or multiclass logistic regression model. Krishnapuram et al. (2005) and Friedman et al. (2010) applied ℓ_1-type penalties to the model as natural extensions of the binomial logistic regression models. On the other hand, there are also multiple parameters in each variable of the multinomial logistic regression model and the multivariate linear model. There have been several studies of the ℓ_1-type regularization for the multivariate linear model. Turlach et al. (2005) proposed a new penalty that can be used to estimate multivariate linear models. They imposed an ℓ_1 sum of the maximum absolute values (ℓ_∞ norm) of the coefficients with respect to multiple responses, and they also generalized it to the ℓ_1 sum of $\ell_q (q \geq 1)$ penalties. Following this, Yuan et al. (2007) and Obozinski et al. (2011) let the penalty be denoted by ℓ_1/ℓ_q and investigated its theoretical properties. It can be viewed as an extension of the group lasso (Yuan and Lin, 2006; Meier et al., 2008). Furthermore, Obozinski et al. (2010) proposed a new algorithm for estimating a multitask logistic regression model by using the ℓ_1/ℓ_q regularization for $q = 1, 2$.

E-mail addresses: hmatsui@math.kyushu-u.ac.jp, hidetoshi.matsui@gmail.com.

http://dx.doi.org/10.1016/j.csda.2014.04.015
0167-9473/

When the data to be classified have been measured repeatedly over time, they can be represented by a functional form. Ramsay and Silverman (2005) established this type of analysis and called it functional data analysis (FDA). FDA is one of the most useful methods for effectively analyzing discretely observed data, and it has received considerable attention in various fields (Ramsay and Silverman, 2002; Ferraty and Vieu, 2006). The basic idea behind FDA is to express repeated measurement data for each individual as a smooth function and then to draw information from the collection of these functions. FDA includes extensions of traditional methods, such as principal component analysis, discriminant analysis, and regression analysis (James et al., 2000; James, 2002). For regression models, there are various methods, such as a functional version of logistic regression models (Aguilera and Escabias, 2008; Aguilera-Morillo et al., 2013; Escabias et al., 2004, 2007), generalized linear models (Cardot and Sarda, 2005; Müller and Stadtmüller, 2005; Li et al., 2010; Goldsmith et al., 2011), and generalized additive models (Reiss and Ogden, 2010). Furthermore, the problem of variable selection for functional regression models using ℓ_1-type regularization is considered in Ferraty et al. (2010); Aneiros et al. (2011); Matsui and Konishi (2011); Zhao et al. (2012); Gertheiss et al. (2013), and Mingotti et al. (2013). However, these works do not include the multiclass logistic regression model. For this model, we may fail to select functional variables when we use existing types of penalties, since it has multiple coefficients for multiple classification boundaries.

In this paper, we consider the problem of using ℓ_1-type regularization to select the variables for classifying functional data by using the multiclass logistic regression model. Data from repeated measurements are represented by basis expansions, and the functional logistic regression model is estimated by the penalized maximum likelihood method with the help of ℓ_1-type penalties. By extending the ℓ_1/ℓ_q penalties, we propose a new class of penalties, denoted by $\ell_1\ell_2/\ell_q$, for appropriately estimating and selecting variables or boundaries for the functional multiclass logistic regression model. Since the basis expansion produces multiple parameters for each variable and each classification boundary, we use the group lasso to treat them as grouped parameters. We here consider the cases for $q = 1$ and $q = 2$. When $q = 1$, instead of selecting the variables themselves, we select classification boundaries for each variable; however, when $q = 2$, we can select the variables that are given as functions by grouping all the coefficients for each variable. The estimated model is evaluated by a selection criterion, since its evaluation is a crucial issue. In order to investigate the effectiveness of the proposed penalty, we conducted Monte Carlo simulations and analyzed actual data.

This paper is organized as follows. Section 2 provides a multiclass logistic regression model for functional data. Section 3 shows a method for estimating and evaluating the model. We apply the proposed method to the analysis of simulated and real data in Sections 4 and 5, respectively. Concluding remarks are given in Section 6.

2. Multiclass logistic regression model for functional data

Suppose that we have n sets of functional data and a class label $\{(x_\alpha(t), g_\alpha); \alpha = 1, \ldots, n\}$, where $x_\alpha(t) = (x_{\alpha 1}(t), \ldots, x_{\alpha p}(t))^T$ are predictors given as functions and $g_\alpha \in \{1, \ldots, L\}$ are the classes to which x_α belongs. In the classification setting, we apply the Bayes rule, which assigns x_α to class $g_\alpha = l$ with the maximum posterior probability given x_α, denoted by $\Pr(g_\alpha = l|x_\alpha)$. Then the logistic regression model is given by the log-odds of the posterior probabilities:

$$\log\left\{\frac{\Pr(g_\alpha = l|x_\alpha)}{\Pr(g_\alpha = L|x_\alpha)}\right\} = \beta_{l0} + \sum_{j=1}^{p}\int x_{\alpha j}(t)\beta_{lj}(t)dt, \tag{1}$$

where β_{l0} is the intercept and $\beta_{lj}(t)$ are the coefficient functions. We assume that $x_{\alpha j}(t)$ can be expressed by basis expansions as

$$x_{\alpha j}(t) = \sum_{m=1}^{M_j} w_{\alpha jm}\phi_{jm}(t) = w_{\alpha j}^T\phi_j(t), \tag{2}$$

where $\phi_j(t) = (\phi_{j1}(t), \ldots, \phi_{jM_j}(t))^T$ are vectors of basis functions, such as B-splines or radial basis functions, and $w_{\alpha j} = (w_{\alpha j1}, \ldots, w_{\alpha jM_j})^T$ are coefficient vectors. Since the data are originally observed at discrete time points, we smooth them with a basis expansion prior to obtaining the functional data $x_{\alpha j}(t)$. In other words, $w_{\alpha j}$ are obtained before constructing the functional logistic regression model (1). Details of the smoothing method are described in Araki et al. (2009b). Furthermore, $\beta_{lj}(t)$ are also expressed by basis expansions

$$\beta_{lj}(t) = \sum_{m=1}^{M_j} b_{ljm}\phi_{jm}(t) = b_{lj}^T\phi_j(t), \tag{3}$$

where $b_{lj} = (b_{lj1}, \ldots, b_{ljM_j})^T$ are vectors of the coefficient parameters.

Using the notation $\pi_l(x_\alpha; b) = \Pr(g_\alpha = l|x_\alpha)$, where $b = (b_1^T, \ldots, b_{(L-1)}^T)^T$ and $b_l = (\beta_{l0}, b_{l1}^T, \ldots, b_{lp}^T)^T$ since it is controlled by b, we can express the functional logistic regression model (1) as

$$\log\left\{\frac{\pi_l(x_\alpha; b)}{\pi_L(x_\alpha; b)}\right\} = \beta_{l0} + \sum_{j=1}^{p} w_{\alpha j}^T\Phi_j b_{lj} = z_\alpha^T b_l, \tag{4}$$