Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 4, pp. 1045~1052

Censored Kernel Ridge Regression

Jooyong Shim[1)]

Abstract

This paper deals with the estimations of kernel ridge regression when the responses are subject to randomly right censoring. The weighted data are formed by redistributing the weights of the censored data to the uncensored data. Then kernel ridge regression can be taken up with the weighted data. The hyperparameters of model which affect the performance of the proposed procedure are selected by a generalized approximate cross validation(GACV) function. Experimental results are then presented which indicate the performance of the proposed procedure.

Keywords: Generalized approximate cross validation function, Kernel ridge regression, Randomly right censoring

1. Introduction

Regression model has been studied extensively with the data subject to randomly right censoring. Koul et al.(1981) proposed a simple estimation in censored regression model by applying the least square method on the weighted observations. Zhou(1992) proposed the M-estimators of regression parameter with the weights suggested by Koul et al.(1981). Yang(1999) proposed a censored median regression model as an alternative to the mean regression model for examining the input vector effect with the data subject to randomly right censoring and showed that the estimators are consistent and asymptotically distributed. Heuchenne and Keilegom(2005) proposed an estimation procedure which extends the least squares procedures for nonlinear regression with censored data.

Ridge regression(Hoerl and Kennard, 1970) is the classical statistical technique which implements a regularized form of the least squares regression. Kernel ridge regression(Saunders et al., 1998), which is a nonlinear form of ridge regression, is

1) Department of Applied Statistics, Catholic University of Daegu, Kyungbuk, 712-702, Korea.
E-mail : ds1631@hanmail.net

developed by introducing kernel functions satisfying Mercer conditions(Mercer, 1909). The least squares support vector machine(LS-SVM), a formulation of kernel ridge regression including a bias term has been proposed for classification and regression by Suykens and Vanderwalle(1999). In kernel ridge regression the solution is given by a linear system instead of a quadratic program problem. The fact that kernel ridge regression has an explicit formulations has a number of advantages.

It is well known that the prediction performance of kernel ridge regression is affected the hyperparameters. We apply the cross-validation method(Green and Silverman, 1994) to kernel ridge regression for censored data.

In this paper we present the estimation procedure of the nonlinear regressions by utilizing kernel ridge regression with the data subject to randomly right censoring. The rest of this paper is organized as follows. In Section 2 we give a brief review of kernel ridge regression. In Section 3 we present the estimation procedure of kernel ridge regression for the censored data and GACV function for selecting hyperparameters. In Section 4 we perform the numerical studies through examples. In Section 5 we give the concluding remarks.

2. Kernel Ridge Regression

Let the training data set be denoted by $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$, with each input vector $\boldsymbol{x}_i \in R^d$ including a constant 1 and the response $y_i \in R$ which is the output corresponding to $\boldsymbol{x}_i$. For kernel ridge regression, we can assume the functional form of unknown mean function f for given input vector $\boldsymbol{x}$ by $f(\boldsymbol{x}) = \boldsymbol{w}'\phi(\boldsymbol{x})$ where $\boldsymbol{w}$ is an appropriate weight vector. Here the feature mapping function $\phi(\cdot): R^d \rightarrow R^{d_f}$ maps the input space to the higher dimensional feature space where the dimension d_f is defined in an implicit way. The optimization problem is defined with a regularization parameter C as

$$\text{minimize } \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + \frac{C}{2}\sum_{i=1}^{N} e_i^2 \tag{1}$$

over $\{\mathbf{w}, \mathbf{e}\}$ subject to equality constraints,

$$y_i = \boldsymbol{w}'\phi(\boldsymbol{x}_i) + e_i, \quad i = 1, \cdots, N.$$

The Lagrangian function can be constructed as

$$L(\boldsymbol{w}, \boldsymbol{e}; \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + \frac{C}{2}\sum_{i=1}^{N} e_i^2 - \sum_{i=1}^{N}\alpha_i(\boldsymbol{w}'\phi(\boldsymbol{x}_i) + e_i - y_i), \tag{2}$$

where α_i's are the Lagrange multipliers. The Karush-Kuhn-Tucker(Smola and Scholkopf, 1998) conditions for optimality are given by

$$\begin{aligned}
&\frac{\partial L}{\partial \boldsymbol{w}} = 0 \rightarrow \boldsymbol{w} = \sum_{i=1}^{N}\alpha_i\phi(\boldsymbol{x}_i)\\
&\frac{\partial L}{\partial e_i} = 0 \rightarrow \alpha_i = Ce_i, \quad i = 1, \cdots, n\\
&\frac{\partial L}{\partial \alpha_i} = 0 \rightarrow \boldsymbol{w}'\phi(\boldsymbol{x}_i) + e_i - y_i = 0, \quad i = 1, \cdots, n,
\end{aligned}$$

leading to the solution

$$\boldsymbol{\alpha} = (K + C^{-1}\boldsymbol{I})^{-1}\boldsymbol{y} \tag{3}$$

with $\boldsymbol{y} = (y_1, \cdots, y_n)'$, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)'$, and $K = \{K_{kl}\}$ where $K_{kl} = \phi(\boldsymbol{x}_k)'\phi(\boldsymbol{x}_l)$, $k, l = 1, \cdots, n$, which are obtained from the application of Mercer's conditions(1909). Several choices of the kernel $K(\cdot, \cdot)$ fucntions are possible.

From (3) the predicted mean function for the given $\boldsymbol{x}_t$ is obtained as

$$\hat{f}(\boldsymbol{x}_t) = K_t\boldsymbol{\alpha} = K_t(K + C^{-1}\boldsymbol{I})^{-1}\boldsymbol{y} \tag{4}$$

where $K_t = \{K(\boldsymbol{x}_t, x_i)\}_{i=1}^{n}$.

3. Kernel Ridge Regression for Censored Data

Consider the linear regression model for the response variables t_i, $i = 1, 2, \cdots, n$,

$$t_i = \boldsymbol{x}_i\boldsymbol{\alpha} + \varepsilon_i, \tag{5}$$

where $\boldsymbol{x}_i$ is the input vector including a constant 1, $\boldsymbol{\alpha}$ is the regression parameter of the model, and ε_i's are unobservable errors assumed to be independent with zero means and bounded variances. c_i's are the censoring variables assumed to be independent and identically distributed and follows a distribution with distribution function $G(t) = P(c_i \le t)$. t_i is not observed but

$$\delta_i = I_{(t_i < c_i)} \quad \text{and} \quad y_i = \min(t_i, c_i), \tag{6}$$

where $I_{(\cdot)}$ denotes the indicator function. In most practical cases G is not known and needs to be estimated by the Kaplan-Meier(1958) estimator or its variation. The problem considered here is that of the estimation of α based on $(\delta_1, y_1, \boldsymbol{x}_1), \cdots, (\delta_n, y_n, \boldsymbol{x}_n)$. Koul et al.(1981) defined new observable responses y_i^* as $y_i^* = wt_i y_i$ with $wt_i = \frac{\delta_i}{1 - G(y_i)}$, and showed y_i^* has the same mean as t_i and thus follows the same linear model as t_i does.

Zhou(1992) proposed M-estimator of the regression parameter α with a loss function $\rho(\cdot)$ using the weights,

$$\text{minimize} \sum_{i=1}^{n} wt_i \rho(y_i - \boldsymbol{x}_i \alpha).$$

We consider the similar weighting scheme as Zhou(1992) with quadratic loss function for censored nonlinear case, replacing the optimal problem (1) by

$$\text{minimize} \quad \frac{1}{2} \boldsymbol{w}' \boldsymbol{w} + \frac{C}{2} \sum_{i=1}^{N} wt_i e_i^2 , \tag{7}$$

which leads to the solution,

$$\alpha = (KWK + C^{-1}K)^{-1} KW\boldsymbol{y} \quad \text{with} \quad W = diag\{wt\}. \tag{8}$$

The predicted mean function for the given $\boldsymbol{x}_t$ is obtained as

$$\hat{f}(\boldsymbol{x}_t) = K_t \alpha = K_t (WKW + C^{-1}K)^{-1} KW\boldsymbol{y}. \tag{9}$$

The functional structures of kernel ridge regression is characterized by hyperparameters - the regularization parameter C and the kernel parameters. We consider the cross validation(CV) function as follows:

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} wt_i (y_i - \hat{f}^{(-i)}(\boldsymbol{x}_i | \lambda))^2,$$

where λ is a set of hyperparameters and $\hat{f}^{(-i)}(\boldsymbol{x}_i | \lambda)$ is the mean function estimated without i-th observation. Since for each candidates of parameters,

$\hat{f}^{(-i)}(\boldsymbol{x}_i|\boldsymbol{\lambda})$ for $i=1,\cdots,n$, should be evaluated, selecting parameters using CV function is computationally formidable. To select the smoothing parameter $\boldsymbol{\lambda}$ for the smoothing spline estimates Nychka et al. (1995) proposed the following approximate cross validation(ACV),

$$ACV(\boldsymbol{\lambda})=\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{y_i-\hat{f}(\boldsymbol{x}_i|\boldsymbol{\lambda})}{1-h_{ii}}\right),$$

where $\rho(\cdot)$ is a differentiable loss function and $h_{ij}=\partial\hat{f}(\boldsymbol{x}_i|\boldsymbol{\lambda})/\partial y_j$. Here h_{ii} can be replaced by their average $tr(H)/n$.

In censored kernel ridge regression GACV function can be described as

$$GACV(\boldsymbol{\lambda})=\frac{n\sum_{i=1}^{n}wt_i(y_i-\hat{f}(\boldsymbol{x}_i|\boldsymbol{\lambda}))^2}{(n-tr(H))^2}, \tag{10}$$

where H is obtained from (8) as

$$H=K(KWK+C^{-1}K)^{-1}KW. \tag{11}$$

4. Numerical Studies

We illustrate the performance of the censored regression estimation using kernel ridge regression through the simulated example on the nonlinear regression cases. For the training data set of the nonlinear censored regression case, 100 of x's are generated from a uniform distribution, $U(0,1)$, and $(t,c)'$s are generated as follows.

$$t_i=\sin(0.75\pi x)+0.5+\varepsilon_{t_i},\quad c_i=\sin(0.75\pi x)+0.5+\varepsilon_{c_i},\quad i=1,\cdots,100,$$

where ε_{t_i}'s and ε_{c_i}'s are generated from normal distributions, $N(0,0.1)$ and $N(cc,0.1)$, respectively. cc is chosen for 20% censoring proportion. For the testing data set, 100 of $(x,t,c)'$s are generated by the same way as for the training ata set. Then the mean function of t_i given x can be modelled as

$$f(x)=\sin(0.75\pi x)+0.5\ .$$

Solving the optimal problem of equation (7) with the weighted training data the optimal Lagrange multipliers, α_i's can be obtained. Then by the equation (8) and (9), we have the predicted mean function given x. The Gaussian kernel is utilized in this example, which is

$$K(x_1, x_2) = \exp(-\frac{1}{\sigma^2}(x_1 - x_2)^2),$$

The regularization parameter C and the kernel parameter σ^2 are obtained as 1.0 and 0.9, respectively, from GACV function (10).

The figure 1 shows true mean function(dotted line) and predicted mean function(solid line) for the training data and testing data, respectively. Uncensored data points are denoted by "·" and those by "o" are censored. In the figure we can see that the predicted mean function behaves similarly as the true mean function does.

Predicted mean squared error(PMSE) is used for the performance metric,

$$PMSE = \frac{1}{n_t}\sum_{i=1}^{n_t}(\hat{f}(\boldsymbol{x}_{t_i}) - f(\boldsymbol{x}_{t_i}))^2,$$

where $\boldsymbol{x}_{t_i}$ is the testing input vector, $i = 1, \cdots, n_t$. From 100 pairs of training and testing data sets we obtained the average of PMSEs as 0.0109, which indicates that the proposed procedure provide satisfying results.

Figure 1. The true and the predicted mean functions for training data (Left) and testing data(Right)

Low-cycle fatigue data(Heuchenne and Keilegom, 2005) for a strain-controlled test on 26 cylindrical specimens of nikel-base superalloy, which include 4 censored data, are used for the next example. The polynomial kernel with degree 2 is utilized in this example. The regularization parameter C is obtained as 800 from GACV function (10). Figure 2 shows that the logarithms of thousands of cycles before fatigue against pseudostress. The predicted mean functions for pseudostress between 80 and 85 by the proposed procedure show lower values than those by (Heuchenne and Keilegom, 2005).

Figure 2. The predicted mean functions for fatigue data by censored kernel ridge regression

5. Concluding Remarks

In this paper, we dealt with estimating the mean of the censored regression model using kernel ridge regression and obtained GACV function for the proposed procedure. By using GACV function the model selection becomes easier and faster than that by a leave-one-out cross validation. Through the example we showed that the proposed procedure derives the satisfying solutions and is attractive approaches to modelling of the censored data. We found that the model is not much sensitive to the choice of the regularization parameter but it is sensitive to the choice of the kernel parameter. Thus a consideration such as standardization of input vectors is required for the choice of the kernel parameter.

References

1. Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Moldels*. Chapman and Hall, London.
2. Hoerl, A. E. and Kennard, R. W.(1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 12(1), 55 - 67.
3. Heuchenne, C and Van Keilegom, I.(2005). *Nonlinear Regression with Censored Data. Technical Report 520,* Universite catholique de Louvain.
4. Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations., *Journal of American Statistical Association,* 53, 457-481.
5. Koul, H., Susarla, V., Van Ryzin J. (1981). Regression analysis with randomly right censored data. *The Annal of Statistics*, 9, 1276-1288.
6. Mercer, J. (1909). Functions of Positive and Negative Type and Their Connection with Theory of Integral Equations. *Philosophical Transactions of Royal Society,* A, 415-446.
7. Nychka, D., Gray, G., Haaland, P., Martin, D., O'Connell, M. (1995). A Nonparametric Approach Syringe Grading for Quality Improvement. *Journal of American Statistical Association*, 432, 1171-1178.
8. Saunders, C., Gammerman, A and Vovk, V. (1998), Ridge Regression Learning Algorithm in Dual Variables. *In Proceedings of 15th International Conferenceon Machine Learning*. Madison, WI, 515 - 521.
9. Smola, A. and Scholkopf, B. (1998). On a Kernel-Based Method for Pattern Recognition, Regression, Approximation and Operator Inversion. *Algorithmica*, 22, 211-231.
10. Suykens, J.A.K. and Vanderwalle, J. (1999). Least Square Support Vector Machine Classifier, *Neural Processing Letters*, 9, 293-300.
11. Yang, S. (1999). Censored Median Regression Using Weighted Empirical Survival and Hazard Functions, *Journal of the American Statistical Association,* 94, 137-145.
12. Zhou, M. (1992). M-estimation in censored linear models., *Biometrika*, 79, 837-841.

[received date : Sep. 2005, accepted date : Oct. 2005]