Journal of Korean
Data & Information Science Society
2007, Vol. 18, No. 3, pp. 767 ~ 772

Kernel Machine for Poisson Regression[1]

Changha Hwang[2]

Abstract

A kernel machine is proposed as an estimating procedure for the linear and nonlinear Poisson regression, which is based on the penalized negative log-likelihood. The proposed kernel machine provides the estimate of the mean function of the response variable, where the canonical parameter is related to the input vector in a nonlinear form. The generalized cross validation (GCV) function of MSE-type is introduced to determine hyperparameters which affect the performance of the machine. Experimental results are then presented which indicate the performance of the proposed machine.

Keywords : Canonical Parameter, Generalized Cross Validation Function, Kernel Machine, Penalized Negative Log-Likelihood, Poisson Regression.

1. Introduction

Poisson regression is widely used to analyze the event count data (Vermunt, 1996). It can be used to model the number of occurrences of an event of interest or the rate of occurrence of an event of interest, as a function of some input variables. A nonparametric estimation for the canonical parameter (logarithm of the mean function) of Poisson process based on penalized likelihood smoothing spline models was proposed by O'Sullivan et al.(1986). When canonical parameter is of a linear form, it is known that the generalized linear model with a Poisson likelihood is useful to Poisson process (McCullagh and Nelder, 1983). But the linear assumption on the canonical parameter is strict in some cases. Various parametric approaches are proposed by Wei(1998) to attain more flexibility than a simple linear model. Yuan(2005) studied smoothing spline models for nonparametric Poisson regression along with the adaptive choice of smoothing parameters.

[1] The present research was conducted by the research fund of Dankook university in 2006.

[2] Professor, Division of Information and Computer Science, Dankook University, Seoul 140-714, Korea. Correspondence : chwang@dankook.ac.kr

In this paper we propose a kernel machine for Poisson regression to obtain the estimate of the mean function of response variable. For the easy selection of appropriate hyperparameters to achieve high generalization performance, we propose the generalized cross validation (GCV) function of MSE-type, which does not use the log-likelihood of the Poisson distribution. The rest of this paper is organized as follows. In Section 2 we propose a kernel machine for Poisson regression, which is based on the penalized negative log-likelihood. In Section 3 we propose GCV function of MSE-type for the model selection. In Section 4 we perform the numerical studies through an example. In Section 5 we give the conclusion.

2. Poisson Regression via Kernel Machine

In Poisson regression it is assumed that the response variable $y_i \in \{1, 2, \cdots\}$, number of occurrences of an event, has a Poisson distribution given the input vector $\boldsymbol{x}_i \in \boldsymbol{R}^d$ including a constant 1,

$$p(y_i) = \frac{e^{-\mu(\boldsymbol{x}_i)}\mu(\boldsymbol{x}_i)^{y_i}}{y_i!}, \quad i = 1,2,\cdots,n.$$

The negative log-likelihood of the given data set can be expressed as (a constant term is omitted)

$$\ell(\mu) = \frac{1}{n}\sum_{i=1}^{n}(\mu(\boldsymbol{x}_i) - y_i \log \mu(\boldsymbol{x}_i)).$$

We write the canonical parameter(logarithm of $\mu(\boldsymbol{x}_i)$) as $\eta(\boldsymbol{x}_i)$, then the negative log-likelihood can reexpressed as

$$\ell(\eta) = \frac{1}{n}\sum_{i=1}^{n}(e^{\eta(\boldsymbol{x}_i)} - y_i \eta(\boldsymbol{x}_i)).$$

The canonical parameter given $\boldsymbol{x}_i$ is estimated by a linear model, $\eta(\boldsymbol{x}_i) = \omega'\phi(\boldsymbol{x}_i)$, conducted in a high dimensional feature space. 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. It suffices to know and use $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)'\phi(\boldsymbol{x}_j)$ instead of defining $\phi(\cdot)$ explicitly. Note that the identity map ϕ leads nonlinear model to linear model. The kernel function used here is the Gaussian kernel,

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{-\frac{1}{\sigma^2}\| \boldsymbol{x}_i - \boldsymbol{x}_j \|^2}, \quad i, j = 1, \cdots, n,$$

where σ^2 is the kernel parameter.

Then the estimate of canonical parameter η is obtained by minimizing the penalized negative log-likelihood,

$$\ell(\omega) = \frac{1}{n}\sum_{i=1}^{n}(e^{\omega'\phi(x_i)} - y_i\,\omega'\phi(x_i)) + \frac{\lambda}{2}\parallel \omega \parallel^2,$$

where λ is a regularization parameter which controls the trade-off between the goodness-of-fit on the data and the complexity of η. The representation theorem (Kimeldorf and Wahba, 1971) guarantees the minimizer of the penalized negative log-likelihood to be $\eta(\boldsymbol{x}_i) = \boldsymbol{k}_i'\boldsymbol{\alpha}$ for some $n \times 1$ vector $\boldsymbol{\alpha}$, where $\boldsymbol{k}_i$ is the ith column of the $n \times n$ kernel matrix $\boldsymbol{K}$ with elements $K(\boldsymbol{x}_k, \boldsymbol{x}_l)$, $k,l = 1,\cdots,n$.

Now the penalized negative log-likelihood (4) becomes

$$\ell\ (\boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^{n}(e^{\boldsymbol{k}_i'\boldsymbol{\alpha}} - y_i\boldsymbol{k}_i'\boldsymbol{\alpha}) + \frac{\lambda}{2}\boldsymbol{\alpha}'\boldsymbol{K}\boldsymbol{\alpha}.$$

By minimizing the penalized negative log-likelihood (5) we obtain the estimator of parameter vector $\boldsymbol{\alpha}$ using Newton-Raphson method, which is not given in a explicit form. At the $(t+1)$th iteration, the parameter vector is estimated as follows:

$$\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)} + \boldsymbol{H}^{-1}\,\boldsymbol{G},$$

where $\boldsymbol{H} = \boldsymbol{W}\boldsymbol{K} + n\lambda \boldsymbol{I}_n$, $\boldsymbol{G} = \boldsymbol{y} - \boldsymbol{\mu}^{(t)} - n\lambda\boldsymbol{\alpha}^{(t)}$, $\boldsymbol{W} = \mathrm{diag}(\boldsymbol{\mu}^{(t)})$. Here $\boldsymbol{\mu}^{(t)}$ is the $n \times 1$ vector with elements $\mu(\boldsymbol{x}_i) = e^{\boldsymbol{k}_i'\boldsymbol{\alpha}^{(t)}}$ and $\boldsymbol{y} = (y_1,\cdots,y_n)'$. With the optimal values of $\boldsymbol{\alpha}$, the predicted mean function given the input vector $\boldsymbol{x}_0$ is obtained as follows:

$$\hat{\mu}(\boldsymbol{x}_0) = e^{\boldsymbol{k}_0'\boldsymbol{\alpha}},$$

where $\boldsymbol{k}_0 = (K(\boldsymbol{x}_1, \boldsymbol{x}_0), \cdots, K(\boldsymbol{x}_n, \boldsymbol{x}_0))'$.

3. GCV function of MSE-type

The functional structure of the kernel machine for Poisson regression is characterized by hyperparameters which are the regularization parameter λ and the kernel parameter σ^2.

For the model selection of the kernel machine, we define the leave-one-out cross validation(CV) function of MSE-type for a set of hyperparameters, θ, as follows:

$$CV(\theta) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\eta}_\theta(\boldsymbol{x}_i) - \hat{\eta}_\theta^{(-i)}(\boldsymbol{x}_i))^2,$$

where $\hat{\eta}_\theta(\boldsymbol{x}_i)$ is the estimate of $\eta(\boldsymbol{x}_i)$ from full data and $\hat{\eta}_\theta^{(-i)}(\boldsymbol{x}_i)$ is the estimate of $\eta(\boldsymbol{x}_i)$ obtained from data without the ith observation. Since for each candidate of hyperparameter sets, $\hat{\eta}_\theta^{(-i)}(\boldsymbol{x}_i)$'s should be evaluated, selecting parameters using CV function is computationally formidable. By leaving-out-one lemma (Kimeldorf and Wahba, 1971),

$$(y_i - \hat{\eta}_\theta{}^{(-i)}(\boldsymbol{x}_i)) - (y_i - \hat{\eta}_\theta(\boldsymbol{x}_i)) = \hat{\eta}_\theta(\boldsymbol{x}_i) - \hat{\eta}_\theta{}^{(-i)}(\boldsymbol{x}_i) \approx \frac{\partial \hat{\eta}_\theta(\boldsymbol{x}_i)}{\partial y_i}(y_i - e^{\hat{\eta}_\theta{}^{(-i)}(\boldsymbol{x}_i)})$$

and $\dfrac{\partial \hat{\eta}_\theta(\boldsymbol{x}_i)}{\partial y_i}(y_i - e^{\hat{\eta}_\theta{}^{(-i)}(\boldsymbol{x}_i)}) \approx \dfrac{\partial \hat{\eta}_\theta(\boldsymbol{x}_i)}{\partial y_i} \dfrac{y_i - e^{\hat{\eta}_\theta(\boldsymbol{x}_i)}}{1 - e^{\hat{\eta}_\theta(\boldsymbol{x}_i)} \dfrac{\partial \hat{\eta}_\theta(\boldsymbol{x}_i)}{\partial y_i}}$, we have

$$\hat{\eta}_\theta(\boldsymbol{x}_i) - \hat{\eta}_\theta{}^{(-i)}(\boldsymbol{x}_i) \approx \frac{\partial \hat{\eta}_\theta(\boldsymbol{x}_i)}{\partial y_i} \frac{y_i - e^{\hat{\eta}_\theta(\boldsymbol{x}_i)}}{1 - e^{\hat{\eta}_\theta(\boldsymbol{x}_i)} \dfrac{\partial \hat{\eta}_\theta(\boldsymbol{x}_i)}{\partial y_i}}$$

Then the ordinary cross validation(OCV) function can be obtained as

$$OCV(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \hat{\eta}_\theta(\boldsymbol{x}_i)}{\partial y_i} \frac{y_i - e^{\hat{\eta}_\theta(\boldsymbol{x}_i)}}{1 - e^{\hat{\eta}_\theta(\boldsymbol{x}_i)} \dfrac{\partial \hat{\eta}_\theta(\boldsymbol{x}_i)}{\partial y_i}} \right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \hat{\eta}_\theta(\boldsymbol{x}_i)}{\partial y_i} \frac{(y_i - e^{\hat{\eta}_\theta(\boldsymbol{x}_i)})}{1 - e^{\hat{\eta}_\theta(\boldsymbol{x}_i)} s_{ii}} \right)^2 \quad (11)$$

where $s_{ij} = \partial \hat{\eta}_\theta(\boldsymbol{x}_i)/\partial y_j$ is the (i,j)th element of $\boldsymbol{S}$ which is the matrix such that $\hat{\boldsymbol{\eta}}_\theta = \boldsymbol{K}\boldsymbol{H}^{-1}(\boldsymbol{y} + const) = \boldsymbol{S}(\boldsymbol{y} + const)$. Replacing s_{ii} by $tr(\boldsymbol{S})/n$ and $e^{\hat{\eta}_\theta(\boldsymbol{x}_i)} s_{ii}$ by $tr(\boldsymbol{VS})/n$, the generalized cross validation(GCV) function can be obtained as

$$GCV(\theta) = \frac{tr^2(\boldsymbol{S})\sum_{i=1}^{n}\left(y_i - e^{\hat{\eta}_\theta(\boldsymbol{x}_i)}\right)^2}{n(n - tr(\boldsymbol{VS}))^2},$$

where $\boldsymbol{V}$ is a diagonal matrix with the (i,i)th entry $e^{\hat{\eta}_\theta(\boldsymbol{x}_i)}$.

4. Numerical Studies

We illustrate the performance of the kernel machine for Poisson regression through the simulated data on the nonlinear case. 100 data sets are generated to present the estimation performance of the proposed machine. Each data set consists of 40 x's and 40 y's. Here x's are randomly generated from $U(0,1)$ and y's are generated from a Poisson distribution with the canonical parameter $\eta(x_i) = 2 + 2\sin(2\pi x_i)$. The canonical parameter given x_i can be modelled as $\eta(x_i) = \boldsymbol{k}_i'\boldsymbol{\alpha}$ with $\boldsymbol{k}_i = (K(\boldsymbol{x}_1, \boldsymbol{x}_i), \cdots, K(\boldsymbol{x}_{40}, \boldsymbol{x}_i))'$ and $\boldsymbol{x}_i = (1, x_i)'$ for $i = 1, \cdots, 40$. Note that we use Gaussian kernel function in this example.

Figure 1(Left) shows the values of CV function(solid line), OCV function(dashed line), and GCV function(dotted line) for one of 100 data sets. We found that the proposed machine is not much sensitive to the choice of the regularization parameter but sensitive to the kernel parameter and that the values of CV function and OCV function are very close with $\lambda = 10$ for each data set. In Figure 1(Left) we can see that CV function and GCV function show similar behaviors.

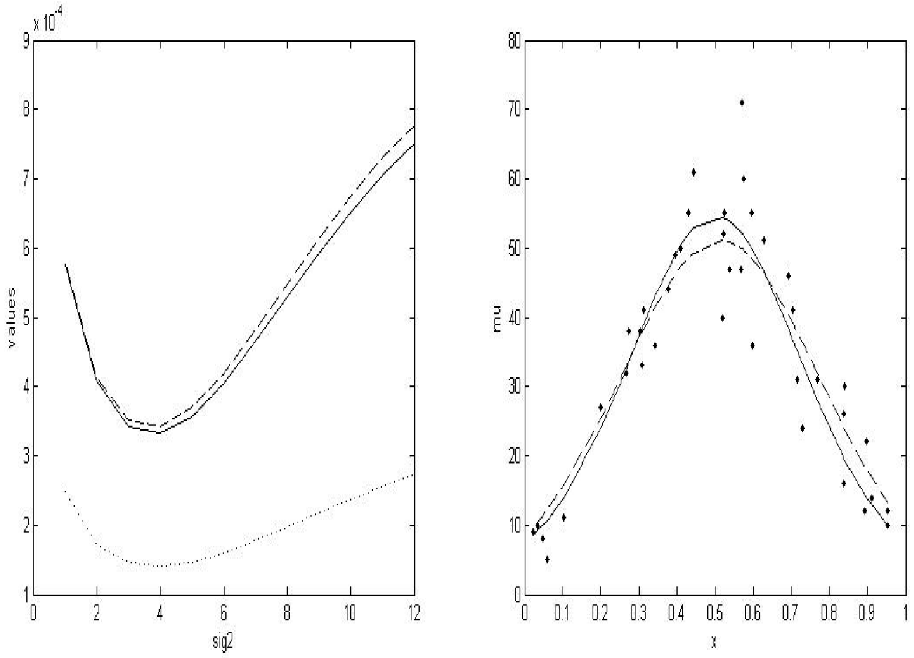

Figure 1. The values of CV function(solid line), OCV function(dashed line) and GCV function(dotted line) on various values of kernel parameter σ^2 with $\lambda = 10$ (Left). Mean function imposed on the scatter plot of one of 100 data sets (Right).

To illustrate the estimation performance of the kernel machine for Poisson regression, the fraction of variance unexplained (FVU) is used as the estimation performance measure, which is introduced by Roosen and Hastie(1994) as follows:

$$FVU = \frac{\sum_{i=1}^{n} (\hat{\mu}(x_i) - \mu(x_i))^2}{\sum_{i=1}^{n} (\mu(x_i) - \bar{\mu})^2},$$

where $\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mu(x_i)$.

Figure 1(Right) shows the true mean function(solid line) and the estimated mean function(dashed line) imposed on the scatter plot of one of 100 data sets. In the figure we can see that the estimated mean function behaves similarly as the true mean function does. From 100 data sets we obtain the average of FVUs as 0.0305, which indicates that the proposed machine provide satisfying results.

5. Conclusions

In this paper, we have dealt with estimating the mean function of Poisson regression by the kernel machine and have provided GCV function for choosing regularization and kernel parameters which affect the performance of the proposed machine. Through the example we have showed that the proposed machine derives the satisfying results and has an advantage of an easy model selection method based on GCV technique.

References

1. Kimeldorf, G. S. and Wahba, G. (1971). Some results on Tchebycheffian spline functions, *Journal of Mathematical Analysis and its Applications*, 33, 82-95.
2. McCullagh, P. and Nelder, J. A. (1983). *Generalized linear models, Monographs on Statistics and Applied Probability*, London: Chapman and Hall.
3. O'Sullivan, F., Yandell, B. S. and Raynor, W. J., Jr. (1986). Automatic smoothing of regression functions in generalized linear models, *Journal of American Statistical Associations*, 81, 96-103.
4. Roosen, C. B. and Hastie, T. J. (1994), Automatic smoothing spline projection pursuit, *Journal of Computational and Graphical Statistics*, 3, 235-248.
5. Vermunt, J. K. (1996). Log-linear event history analysis, *Series on Work and Organization*, Tilburg: Tilburg University Press.
6. Wei, B. C. (1998). Exponential family nonlinear models, *Lecture Notes in Statistics*, 130, Singapore: Springer-Verlag Singapore.
7. Yuan, M. (2005). Automatic smoothing for Poisson regression, *Communications in Statistics - Theory and Methods*, 34, 603 - 617.

[received date : June 2007, accepted date : June 2007]