MULTIVARIATE ADAPTIVE REGRESSION SPLINE MODEL ESTIMATION FOR BINARY RESPONSE

Anna Islamiyati, Raupong

Abstract

MARS is a spline nonparametric regression model of multivariate data, which takes into knots and basis function. MARS model estimation can be approximated by least square or maximum likelihood methods. However, for binary data response, MARS model estimation should be completed with a numerical approach. This paper about MARS model estimation for binary response, it was solved by newton raphson methods, because the estimation outputs of maximum likelihood was implicit.

Key words: basis function, knots, maximum likelihood, MARS, and newton raphson

Introduction

Multivariate Adaptive Regression Spline (MARS) model was introduced by Friedman (1991) which is a nonparametric regression model between response and predictors variabel. Understanding of multivariate use the some predictor variable. MARS is development at Resurcive Partitioning Regression (RPR) with a combination of spline. The use spline in the MARS led to the nature of the data which is very flexible in finding the patterns that correspond to the data.

The use of MARS is already highly developed in various fields of aplication, such as by Xiong and Meullenet (2002), which compares between MARS with logistic regression on the effect of water activity, and producing method of MARS better than method of logistic regression. Other research by Leatwich, et al (2006), which the use of MARS in analysis of distribution relationship of 15 species of fresh water fish and their environment. Further by Otok (2010), research on the classification of the zone an area with MARS.

Required a study at the theory of MARS, apart from completion of data analysis software with MARS, expecially in the binary response, to be able to show a stage in getting the MARS model estimates through numerical methods. So expect the user is not only able to resolved the problem with the software of MARS, but they are cappable of estimating model with MARS trough an algorithm to another.

Material and Method

Multivariate Adaptive Regression Spline Model

Friedman (1991) introduced the method of the MARS as a method to build the accurate prediktif models to continue and binary response variables. The model MARS is focused to overcome the problems dimension is high and continue on the data. MARS is a combination of the complex of development RPR and spline. Both things that buil a MARS model is the knots and the function of basis.

Unknown:

$$y_i = f\left(x_i\right) + \varepsilon_i,\ i = 1, 2, \ldots, n, \tag{1}$$

which $f\left(x_i\right)$ is MARS of function:

$$f\left(x_i\right) = \beta_0 + \sum_{m=1}^{M} \beta_m \prod_{k=1}^{km} \left(s_{km} \left(x_{v(k,m)} - t_{km} \right) \right), \tag{2}$$

until MARS model can be written as follows:

$$y_i = f\left(x_i\right) = \beta_0 + \sum_{m=1}^{M} \beta_m \prod_{k=1}^{km} \left(s_{km} \left(x_{v(k,m)} - t_{km} \right) \right) + \varepsilon_i . \tag{3}$$

Where:

y_i : the variable response on $i = 1, 2, \ldots, n$

x_i : the predictor on $i = 1, 2, \ldots, n$

β_0 : the regression constan,

β_m : the regression coefficient $m = 1, 2, \ldots, \mathrm{M}$,

km : the degree of interaction,

t_{km} : the point of knots,

ε_i : the error on $i = 1, 2, \ldots, n$

The selection of the optimum model is that has a minimum GCV on basis function and point knots. GCV functions introduced by Wahba (1979), as follow:

$$GCV\left(M\right) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}_M\left(x_i\right)\right)^2}{\left(\frac{1 - \left(tr\left(B\left(B^T B\right)^{-1} B^T\right) + 1\right)}{n}\right)^2} . \tag{4}$$

Distribution of The Binary Response

The variable response of two categories to as binary response model of MARS, and which takes into account the binary response is called MARS binary response. The MARS binary

response of the bernoulli distribution, i.e of a random variable has only two categories, category 1 to a category of succes, and 0 to category of failure.

Hosmer & Lomeshow (2000), if a random gaussian variable y bernoulli distribution, $y \square b\left(1,\pi\left(x\right)\right)$ with $y \in \left(0,1\right)$ and $x \in \Re^p$, so probability density function:

$$f\left(y_i\right)=\left[\pi\left(x\right)\right]^{y_i}\left[1-\pi\left(x\right)\right]^{1-y_i}, \tag{5}$$

where $\pi\left(x\right)$ is success probability and $1-\pi\left(x\right)$ is failure probability.

Result and Discussion

Estimatio of binary response model of MARS in the stage of the maximum likelihood. This is reveated by Otok (2009) that to get value estimate on a binary model of MARS can be done by using the method of maximum likelihood.

MARS model is given in (3), where y_i the follow the bernoulli distribution with parameter $\pi\left(x\right)$, so probability density functions y_i is given in (5).

The cumulative density function for random variable $y_1, y_2, \ldots, y_n$,

$$\begin{aligned} f\left(y_1, y_2, \ldots, y_n \left|\boldsymbol{\beta}\right.\right) &= \left[\pi\left(x\right)\right]^{\sum_{i=1}^{n} y_i}\left[1-\pi\left(x\right)\right]^{\left(n-\sum_{i=1}^{n} y_i\right)} \\ &= \prod_{i=1}^{n}\left[\left(\pi\left(x\right)\right)^{y_i}\left(1-\pi\left(x\right)\right)^{1-y_i}\right] \end{aligned}.$$

Futhermore, likelihood function can be made as follow:

$$\begin{aligned} L\left(\boldsymbol{\beta}\left|y_i\right.\right) &= \prod_{i=1}^{n}\left[\left(\pi\left(x\right)\right)^{y_i}\left(1-\pi\left(x\right)\right)^{1-y_i}\right] \\ &= \prod_{i=1}^{n}\left[\left(\frac{\pi\left(x\right)}{1-\pi\left(x\right)}\right)^{y_i}\left(1-\pi\left(x\right)\right)\right], \end{aligned}$$

so, the logarithm natural of likelihood function is:

$$\ln L\left(\boldsymbol{\beta}\left|y_i\right.\right)=\sum_{i=1}^{n}\left[y_i\left(\beta_0+\sum_{m=1}^{M}\beta_m\prod_{k=1}^{Km}\left(S_{km}\left(x_{v(k,m)}-t_{km}\right)\right)\right)+\ln\left(1+\exp\left(\beta_0+\sum_{m=1}^{M}\beta_m\prod_{k=1}^{Km}\left(S_{km}\left(x_{v(k,m)}-t_{km}\right)\right)\right)\right)\right] \tag{6}$$

Futhermore:

$$\frac{\partial \ln L\left(\boldsymbol{\beta}\left|y_i\right.\right)}{\beta_0}=\sum_{i=1}^{n}\left[y_i-\frac{\exp\left(\beta_0+\sum_{m=1}^{M}\beta_m\prod_{k=1}^{Km}\left(S_{km}\left(x_{v(k,m)}-t_{km}\right)\right)\right)}{1+\exp\left(\beta_0+\sum_{m=1}^{M}\beta_m\prod_{k=1}^{Km}\left(S_{km}\left(x_{v(k,m)}-t_{km}\right)\right)\right)}\right]$$

$$\frac{\partial \ln L(\boldsymbol{\beta}|y_i)}{\beta_1} = \sum_{i=1}^{n}\left[y_i \prod_{k=1}^{Km}\left(S_{k1}\left(x_{v(k,1)} - t_{k1}\right)\right) - \prod_{k=1}^{Km}\left(S_{k1}\left(x_{v(k,1)} - t_{k1}\right)\right)\frac{\exp\left(\beta_0 + \sum_{m=1}^{M}\beta_m \prod_{k=1}^{Km}\left(S_{km}\left(x_{v(k,m)} - t_{km}\right)\right)\right)}{1+\exp\left(\beta_0 + \sum_{m=1}^{M}\beta_m \prod_{k=1}^{Km}\left(S_{km}\left(x_{v(k,m)} - t_{km}\right)\right)\right)}\right]$$

$$\vdots$$

$$\frac{\partial \ln L(\boldsymbol{\beta}|y_i)}{\beta_M} = \sum_{i=1}^{n}\left[y_i \prod_{k=1}^{Km}\left(S_{kM}\left(x_{v(k,\mathrm{M})} - t_{kM}\right)\right) - \prod_{k=1}^{Km}\left(S_{kM}\left(x_{v(k,\mathrm{M})} - t_{kM}\right)\right)\frac{\exp\left(\beta_0 + \sum_{m=1}^{M}\beta_m \prod_{k=1}^{Km}\left(S_{km}\left(x_{v(k,m)} - t_{km}\right)\right)\right)}{1+\exp\left(\beta_0 + \sum_{m=1}^{M}\beta_m \prod_{k=1}^{Km}\left(S_{km}\left(x_{v(k,m)} - t_{km}\right)\right)\right)}\right]$$

The first differential of the ln likelihood function gives no explicit completion, so the assessment parameter $\boldsymbol{\beta}$ is done by Newton Raphson iteration method.

$$\begin{bmatrix} \hat{\beta}_{0(\mathrm{h}+1)} \\ \hat{\beta}_{1(\mathrm{h}+1)} \\ \hat{\beta}_{2(\mathrm{h}+1)} \\ \vdots \\ \hat{\beta}_{\mathrm{M}(\mathrm{h}+1)} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_{0(\mathrm{h})} \\ \hat{\beta}_{1(\mathrm{h})} \\ \hat{\beta}_{2(\mathrm{h})} \\ \vdots \\ \hat{\beta}_{\mathrm{M}(\mathrm{h})} \end{bmatrix} - \mathbf{H}^{-1}\mathbf{d} \ ,$$

where $\begin{bmatrix} \hat{\beta}_{0(\mathrm{h})} & \hat{\beta}_{1(\mathrm{h})} & \hat{\beta}_{2(\mathrm{h})} & \dots & \hat{\beta}_{\mathrm{M}(\mathrm{h})} \end{bmatrix}^T$ is the regression parameters in the iteraton to h, $\mathbf{d}$ is the first derivative of the vector of parameters $\boldsymbol{\beta}$, and second derivative matrix $\mathbf{H}$ is againts the parameter $\boldsymbol{\beta}$.

Iterations stop when the value $\hat{\boldsymbol{\beta}}$ convergent, that is $\hat{\boldsymbol{\beta}}_{h+1} - \hat{\boldsymbol{\beta}}_h = 0$, and binary response model MARS estimates obtained:

$$\hat{y}_i = \hat{\beta}_0 + \sum_{m=1}^{M}\hat{\beta}_m \prod_{k=1}^{km}\left(s_{km}\left(x_{v(k,m)} - t_{km}\right)\right). \tag{7}$$

Without the use of software on MARS, determination of the basis functions $\prod_{k=1}^{km}\left(s_{km}\left(x_{v(k,m)} - t_{km}\right)\right)$ done through a system of trial an error by starting the lowest basis, by considering the point of knots *t*.

Conclusion

This paper only examines the estimation function mars binary response, so it is advisable to study higher on multi categories response, taking into account the nominal and ordinal data.

Bibliography

Friedman, JH. 1991. Multivariate Adaptive Regression Spline. *The Annals of Statistics*. 19(1):1-68.

Leatwhick, J.R, et.al. 2006. Comparative Performance of Generalized Additive Models and Multivariate Adaptive Regression Splines for Statistical Modeling of Species Distribution. *Ecological Modelling* 199:188-196, Hamilton. New Zealand.

Otok, B. W. Et al. 2008. Asimtotik Model Multivariate Adaptive Regression Spline. *Jurnal Natur Indonesia* Vol 10 No. 2:112-119, Yogyakarta.

Otok, B. W. 2010. Approach Multivariate Adaptive Regression Spline on the Classification of The Zone an Area of The Season. *Jurnal Statistika* Vol 10 No. 2:107-120, Surabaya.

Xiong R & Meulenet JF, 2002. *Comparasion of Logistik Regression and MARS In Modeling The Effects of Water Activity, pH, and Potssium Sorbate on Growth-No Growth of Sacchararomyces cerevisiae*. Food Science Department University of Arkansan.