Robust regression algorithms with kernel functions in Support Vector Regression Models

Muthukrishnan. R , Kalaivani. S

•

(1) Professor, Department of Statistics, Bharathiar University, Tamil Nadu
(2) Research Scholar, Department of Statistics, Bharathiar University, Tamil Nadu
E-mail: muthukrishnan1970@gmail.com, kalaivanistatistics1994@gmail.com

Abstract

In machine learning, support vector machines (SVM) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. SVM is one of the most robust prediction method based on statistical learning frameworks. Regression is a statistical method that attempts to determine the strength and character of the relationship between dependent and independent variables. This paper explores the idea of support vector Regression. The most commonly used classical procedure is Least Squares, which is less efficient and very sensitive when the data contains outliers. To overcome this limitations, alternative robust regression procedures exist such as LMS regression, S-estimator, MM-estimator and Support Vector Regression (SVR). In this study, the comparisons have made for the classical regression procedure and the robust regression procedures. In that, various measures of errors are much efficient when we work with robust regression procedures. In this paper, an attempt has been made to review the existing theory and methods of SVR.

Keywords: Linear regression, Robust regression, kernels, Support Vector Regression

1. Introduction

Support vector regression is a feature of support vector machines. It's worth mentioning that the support vector machine (SVM) is a concept that may be utilized to analyze both regression and classification data [4] and [12]. Support vector classification is the name given to the support vector machine when it is used for classification, while support vector regression is the name given to it when it is used for regression [5]. SVM is a new machine learning technique based on the statistical learning theory proposed by Vapnik and Wolfe dual programming theory. SVM has a robust mathematical theory base, well-generalized ability, and global optimum, as compared to other learning algorithms, and is widely used in pattern recognition and functional regression as a result are seen in [9].

Support vector machine (SVM) has been first introduced by Vapnik. There are two main categories for support vector machines: support vector classification (SVC) and support vector regression (SVR). SVM is a learning system using a high dimensional feature space. It yields prediction functions that are expanded on a subset of support vectors. SVM can generalize complicated gray level structures with only a very few support vectors and thus provides a new mechanism for image compression. A version of a SVM for regression has been proposed in 1997 by Vapnik, Steven Golowich, and Alex Smola [11]. This method is called support vector regression (SVR).

The support vector machine is a more advanced version of the support vector classifier, resulting from the use of kernels to enlarge the feature space in a specified fashion [1] and [6].

The feature space is expanded in this technique to allow a non-linear boundary between the classes. The kernel technique is a computationally efficient way to put such a notion into action [2]. Support vector regression allows us to specify how much error in our model is acceptable, and it will identify an appropriate line (or hyperplane in higher dimensions) to fit the data. See, [3] and [10].

The manuscript of this paper is laid out as follows. The concept of regression procedures is defined in Section 2. This session also covers the LS, LMS, S, MM, and SVR methodologies utilized in this article for examining these ideas. Support vector regression types and various kernel functions are also covered. Section 3 summarizes results of the numerical study of comparative analysis under various kernels along with regression procedures. Section 4 ends with a summary and conclusion.

2. REGRESSION PROCEDURES

The conventional regression procedure, namely, Least Squares Method (LS), the robust procedures, Least Median Squares Method (LMS), S-Estimator (S), and MM-Estimator (MM), and Support Vector Regression (SVR) are briefly discussed in this section.

2.1. Least Squares Method (LS)

A fundamental statistical method for determining a regression line or the best-fit line for a given pattern is the least-squares approach. An equation with specified parameters are described in this procedure. This method is considered a typical strategy in regression analysis for approximating sets of equations with more equations than unknowns. The least squares method determines the best results by minimizing the sum of squares of deviations or errors in each equation's result. Least-square method is the curve that best fits a set of observations with a minimum sum of squared residuals or errors. The exercise of minimizing these residuals would be the trial and error fitting of a line "through" the Cartesian coordinates representing these values. One way to proceed with the Least Squares Method is to solve using matrix multiplication.
he least squares method can more formally be described. Given a dataset of points (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) and derive the matrices:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix}, y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix}, A = \begin{bmatrix} b \\ m \end{bmatrix}, E \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ e_n \end{bmatrix} \tag{1}$$

Then set up the matrix equation,

$$Y = XA + E \tag{2}$$

where matrix Y contains the y_n values, matrix X contains a row of 1™s and along with the x_n values, matrix A consists of the Y-intercept and slope, and matrix E is the errors. Then solve for A,

$$A = (X^T X)^{-1} X^T Y \tag{3}$$

This is the matrix equation ultimately used for the least squares method of solving a linear system.

2.2. Least Median Squares Method (LMS)

Rousseeuw was the first to introduce the method of Least Median Squares in 1984 [7]. The least median of squares approach estimates the parameters by solving the nonlinear minimization

problem. To put it another way, the estimator must produce the minimum number for the median of squared residuals computed throughout the entire data set. It turns out that this strategy is extremely resistant to false matches and outliers caused by poor localization, implying that it is unaffected by outliers or other violations of the typical normal model's assumptions. The LMS regression was developed to optimise the median of the squares of residuals. The LMS estimate can be obtained as the solution of the following optimization problem.
Let ${x_i}^T = (x_{i1}, x_{i2},, x_{ip})$, i=1,2,¦.,n and

$$y = (y_1, y_2,, y_n)^T \tag{4}$$

be given real vectors.
It is assumed that $\frac{n}{p} \geq p$ and the (nxp) matrix, X= $[x_{ij}]$ is offull rank to avoid degenerate cases

$$\theta = (\theta_1, \theta_2,, \theta_p)^T \tag{5}$$

be a vector of regression parameters. The optimization problem that arises out of the LMS method is to estimate θ^* providing

$$^{min}{}_{\theta} med(y_i - {x_i}^T\theta)^2 \tag{6}$$

LMS is designed to have a high breakdown point, which is commonly defined as the minimum percentage of "contaminated" data required to change an estimate by a given amount. The LMS breakdown point is 50%, while the comparable LS breakdown point is zero.

2.3. S-Estimator (S)

Rousseeuw and Yohai (1984) proposed the S-estimator, which minimises a scale estimator [8]. S-estimators combine the flexibility and asymptotic features of M-estimators to provide a simple high-breakdown regression estimator. Because they are based on scale estimators, the name S-estimators was chosen. The S-estimator minimizes an estimator of scale, which is given by

$$\hat{\beta_n} = argmin\hat{\alpha_n}(\beta) \tag{7}$$

The estimator of scale may be defined by a function ρ, For any sample $r_1, r_2,, r_n$ of real numbers, we define the scale estimate $s(r_1, r_2,, r_n)$ as the solution of

$$\frac{1}{n}\sum_{i=1}^{n}\rho\frac{r_i}{s} = k \tag{8}$$

where k is the expectation value of ρ for a standard normal distribution. Let $(x_1, y_1),, (x_n, y_n)$ be a sample of regression data with p-dimensional x_i. For each vector θ, obtain the residuals (r$_1(\theta)$,r$_2(\theta)$,¦,r$_n(\theta)$) by solving the above equation and hence the estimator of scale may be defined by a function . Further the function ρ must satisfy the conditions, such as symmetric, continuously differentiable and ρ(0)=0 and also there exists c $>$ 0 such that ρ is strictly increasing on [c, ∞]..
Thus, the S-estimator $\hat{\theta}$ is defined by

$$\hat{\theta} = {}^{min}{}_{\theta} s(r_1(\theta), r_2(\theta), ..r_n(\theta) \tag{9}$$

and the final scale estimator $\hat{\sigma}$ is then

$$\hat{\sigma} = s(r_1(\hat{\theta}), ..., r_n(\hat{\theta})) \tag{10}$$

In least Squares, least absolute deviation estimation, and even generalized M-estimators, outlying observations sometimes strongly influence the estimation result, making an important and interesting relationship existing in the majority of observations. The S-estimators are a class of estimators that overcome this difficulty by smoothly down-weighting outliers in fitting regression functions to data.

2.4. MM-Estimator (MM)

Such estimators are interesting as they combine high efficiency and high breakdown point in a simple and intuitive way. Typically one starts first with a highly-robust regression estimator, typically an S-estimator. Then one can use the scale based upon this preliminary fit along with a better-tuned ρ function to obtain a more efficient M-estimator of the regression parameter. An MM-estimator of α then defined as any solution of an M-type equation where

$$\sum_{i=1}^{n} \rho_1{}' [\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta_j}{S_{MM}}] x_{ij} = 0 \tag{11}$$

Such estimators are interesting as they combine high efficiency and high breakdown point in a simple and intuitive way. Typically one starts first with a highly-robust regression estimator, typically an S-estimator. Then one can use the scale based upon this preliminary fit along with a better-tuned ρ function to obtain a more efficient M-estimator of the regression parameter. An MM-estimator of α then defined as any solution of an M-type equation where

$$\psi_M M(y, x : \alpha) = u_{MM}(X^T \sum_{s}^{-1}(y - x\alpha)) \tag{12}$$

2.5. Support Vector Regression (SVR)

Support Vector Regression is a method for estimating a function that maps from an input item to a real integer. SVR has the same qualities as the classifying SVM, as well as the margin maximization and kernel technique for non-linear mapping.
A dataset for regression is represented as follows,

$$D = (x_1, y_1), (x_2, y_2), ..., (x_m, y_m) \tag{13}$$

where x_i is a n-dimensional vector, y is the real number for each x_i. The SVR function $F(x_i)$ makes a mapping from an input vector x_i to the target y_i and takes the form.

$$F(x) = w.x - b \tag{14}$$

where w is the weight vector and b is the bias. The goal is to estimate the parameters (w and b) of the function that give the best fit of the data. An SVR function F(x) approximates all pairs (x_i, y_i) while maintaining the differences between estimated values and real values under precision.
Unlike LS, SVR's goal is to minimise the coefficients, specifically the L_2 norm of the coefficient vector rather than the squared error. In SVR, we can adjust epsilon to achieve the model's desired accuracy. SVR is an advanced regression technique that makes use of the concept of hyperplane to perform well with large datasets. Simple regression strives to lower error rates, whereas SVR aims to fit the error with a specific threshold.

2.5.1 Kernel function in Support Vector Regression

SVM can be used as a classification machine, as a regression machine, or for novelty detection. Kernel functions, a group of mathematical functions play a significate role for getting better accuracy in SVM. The function of a kernel is to require data as input and transform it into the desired form. Different kernel functions are used by different SVM algorithms. There are several types of these functions, including linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid. The most preferred kind of kernel function is RBF. Because it's localized and has a finite response along the complete x-axis. The kernel functions return the scalar product between two points in an exceedingly suitable feature space. The most widely used kernel functions are briefly furnished as follows.
Linear kernel, which is the most basic sort of kernel and is usually one dimensional. When there are a lot of features, it proves to be the best function. For text-classification tasks, the linear kernel

Table 1: ***Computed error measures under conventional, robust and support vector regression***

Errors	**LS**	**LMS**	**S**	**MM**	**SVR**							
					Linear		**Polynomial**		**Radial**		**Sigmoid**	
					γ	ϵ	γ	ϵ	γ	ϵ	γ	ϵ
MDAE	0.47 (0.23)	0.26 (0.22)	0.31 (0.24)	0.47 (0.23)	0.57 (0.24)	0.44 (0.23)	0.40 (0.31)	0.38 (0.29)	0.30 (0.23)	0.23 (0.22)	0.48 (0.49)	0.48 (0.43)
MSE	0.30 (0.09)	2.20 (0.10)	1.60 (0.09)	0.30 (0.09)	0.30 (0.09)	0.31 (0.09)	0.20 (0.12)	0.25 (0.12)	0.10 (0.09)	0.12 (0.09)	1.96 (0.44)	2.10 (0.68)
RMSE	0.55 (0.30)	1.48 (0.32)	1.26 (0.31)	0.55 (0.30)	0.55 (0.30)	0.56 (0.30)	0.50 (0.35)	0.50 (0.35)	0.40 (0.30)	0.35 (0.30)	1.40 (0.66)	1.45 (0.83)
(.)	without outliers											

is usually favoured because most of these problems can be linearly split. Linear kernel functions, denoted by $u^{'}$ v, are faster than other functions.

The polynomial kernel, which is a more generalised form of the linear kernel, is the second kernel. It is not as preferred as other kernel functions as it is less efficient and accurate. Polynomial Kernel function is denoted by $(\gamma u^{'} v + coef0)^{degree}$.

The radial basis function kernel, which is one of the most popular and widely utilized in SVM. It's typically used with non-linear data. When there is no prior knowledge of data, it aids in proper separation. The gamma value ranges from 0 to 1. Radial basis Kernel function is denoted by $e^{((-\gamma|u-v^2))}$.

The sigmoid kernel is mostly preferred for neural networks. This kernel function is similar to a two-layer perceptron model of the neural network, which works as an activation function for neurons. Sigmoid Kernel function is denoted by $tanh(\gamma u^{'} v + coef0)$.

The nu- SVR and eps-SVR has been taken for comparing the performance of various kernel based SVR. In nu-SVR, the parameter γ is used to determine the proportion of the number of support vectors desire to keep in solution with respect to the total number of samples in the dataset. Also the parameter ϵ is introduced into the optimization problem formulation and it is estimated automatically. But in eps-SVR, there is no control on how many data vectors from the dataset become support vectors, it could be a few, it could be many. Nonetheless, the total control of how much error will allow the model to have, and anything beyond the specified ϵ will be penalized in proportion to C, which is the regularization parameter.

3. NUMERICAL STUDY

A real data is used to demonstrate the performance of various approaches by computing various measures of errors values. The dataset used in the numerical analysis is StarsCYG, which is available in a package namely, robustbase in R. The data describes the Hertzsprung-Russell diagram of the star cluster CYG OB1, which contains 47 stars in the direction of Cygnus, the predictor variable, the logarithm of the effective temperature at the star's surface (log.Te), and the response variable, the logarithm of the star's light intensity (log.light).

The experimental study has been carried out to study the performance of various procedures, such as Least Squares (LS), Least Median Squares (LMS), S-Estimator (S), MM-estimator (MM), and Support Vector Regression (SVR) with various kernels by computing error measures for the dataset under with/without outliers and thus obtained results are summarized in the table 1.

(.) without outliers

The result reveals that, robust procedures provide better results when compared with the conventional least square approach. Further, it is observed that $\gamma-$ and $\epsilon-$ type radial kernel based SVR outperforms over other kernels.

4. CONCLUSION

Regression analysis is one of the supervised learning techniques in the context of statistical learning. This paper explores the conventional, robust and support vector based regression procedures. In the context of support vector regression, the study has been carried out under the most widely used kernels. The efficiency of these algorithms have been studied under a real study, with and without outliers. The results indicate that the robust regression procedures more efficient than the traditional regression procedure under with and without outliers. Further, the study reveals that SVR delivers much superior outcomes when compared with the others. In the context of kernels, the γ and ϵ based radial kernel has the maximum efficiency when compared to other kernels, regardless of whether or not the data contains outliers. The study can be extended by incorporating the robust kernel in support vector regression for better accuracy.

References

[1] Collobert.R, Bengio.S (2001), SVM Torch: support vector machines for large-scale regression problems *Journal of Machine Learning Research*,Vol.1, pp.143-160.

[2] Chih-Chung Chang, Chih-Jen Lin, (2001), Training -support vector classifiers: Theory and algorithms *Neural Computation*,Vol.13(9), pp.2119-2147.

[3] Flake.G.W, Lawrence.S (2002), Efficient SVM regression training with SMO *Machine Learning*,Vol.46, pp.271-290.

[4] Hiroyuki Takeda, Sira Farsiu, Peyman milanfar (2007), kernel Regression for Image processing and reconstruction *IEEE transactions on Image processing*,Vol.16 (2), pp.349-366.

[5] Harris Drucker, Chris.J.C.Burges, Linda kaufman, Alex smola, Vladimir Vapnik (1996), Support vector regression machines *Proceedings of the 9th International conference on Neural Information processing systems*,pp.155-161.

[6] Keerthi.S.S, Shevade.S, Bhattacharyya.C, and Murthy.K (2000), Improvements to SMO algorithm for SVM regression *IEEE Transactions on Neural Networks*,Vol.11 (5), pp.1188-1193. .

[7] Rousseeuw.P.J (1984), Least Median of Squares Regression *Journal of the American Statistical Association*,Vol.79 (388), pp. 871-880.

[8] Rousseeuw.P.J, Yohai.V (1984), Robust regression by means of S-estimators *Robust and non-linear time series analysis*,Vol. 26, pp.256-272 .

[9] Shital Jore, Badadapure.P.R (2014), Remote sensing Image segmentation using linear regression *International Journal of Engineering research Technology*,Vol.3 (4) , pp. 2279-2282.

[10] Smola.A.J, Scholkopf.B (2004), A tutorial on support vector regression *statistics and computing*,Vol.14, pp.199-122.

[11] Vapnik.V.N, Golowich.S and Smola.A (1997), Support vector method for function approximation, regression estimation, and signal processing *Proceedings of Advances in Neural Information Proceedings of advances in Neural Information Processing systems*,Vol.9, pp.281-287.

[12] Yohai.V (1987), High breakdown point and high efficiency robust estimates for regression *The annals of Statistics*,Vol.15 (2), pp. 642-656.