International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878
Volume 2 Issue 2, February 2014

Support Vector Regression for Outliers Removal

G. Malik[1], M. Tarique[2]

[1]R.L.A (Eve.) College, University of Delhi

[2]Dyal Singh College, University of Delhi

Abstract: ***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. The model produced by support vector classification only depends on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin. The regression analysis gives absurd results, if there are outlier's presents in the data sets.***

Keywords: Classification, Data Mining, Regression, SVM

1. Introduction

The Support Vector Machine (SVM) is a universal approach for solving the problems of multidimensional function estimation. Those approaches are all based on the Vapnik–Chervonenkis (VC) theory. Initially, it was designed to solve pattern recognition problems, where in order to find a decision rule with good generalization capability, a small subset of the training data, called the support vectors are selected. Experiments showed that it is easy to recognize high-dimensional identities using a small basis constructed from the selected support vectors. Recently, SVM has also been applied to various fields successfully such as classification, time prediction and regression. When SVM is employed to tackle the problems of function approximation and regression estimation, the approaches are often referred to as the Support Vector Regression (SVR). The SVR type of function approximation is very effective, especially for the case of having a high-dimensional input space.

In general, for any real-world applications, observations are always subject to noise or outliers. The intuitive definition of outliers is that "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism". Outliers may occur due to various reasons, such as erroneous measurements or noisy phenomenon appearing in the tail portion of some noise distribution functions. However, the traditional SVR is not effective in dealing with outliers in training data commonly encounter in practical applications. Thus few outliers result in a poor regression. The basic idea of the proposed method consists in gradually partitioning data into outliers and inliers, and thus refining the estimation with the inliers.

2. Linear Regression

Regression analysis includes any techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps us understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables — that is, the average value of the dependent variable when the independent variables are held fixed.

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships.

Regression models involve the following variables:

- The unknown parameters denoted as β; this may be a scalar or a vector of length k.
- The independent variables X.
- The dependent variable, Y.

A regression model relates *Y* to a function of X and β.

$$Y \approx f(\mathbf{X}, \boldsymbol{\beta})$$

In linear regression, the model specification is that the dependent variable, y_i is a linear combination of the parameters (but need not be linear in the independent variables).

Suppose we are given a data set

$$\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$$

of *n* statistical units, a linear regression model assumes that the relationship between the dependent variable y_i and the *p*-vector of regressors x_i is approximately linear. This approximate relationship is modeled through a so-called "disturbance term" ε_i — an unobserved random variable that adds noise to the linear relationship between the dependent variable and regressors.

Thus the model takes form

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i'\beta + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $'$ denotes the transpose, so that $x_i'\beta$ is the inner product between vectors x_i and β.

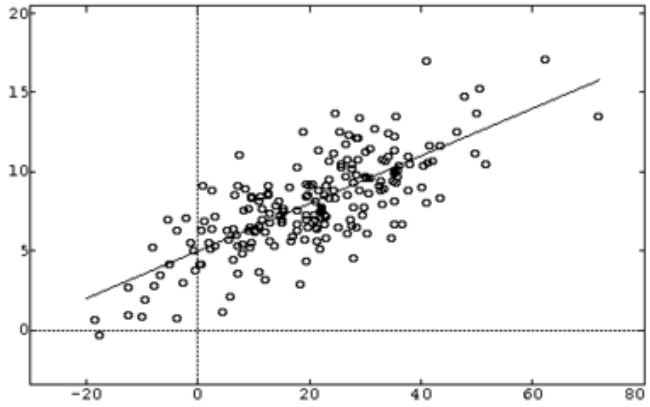
Often these n equations are stacked together and written in vector form as

$$y = X\beta + \varepsilon,$$

Where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Example of linear regression with one independent variable



3. Support Vector Regression

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. A version of a SVM for regression has been proposed in 1997 by Vapnik, Steven Golowich, and Alex Smola . This method is called support vector regression (SVR). The model produced by support vector classification only depends on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin. Analogously, the model produced by SVR only depends on a subset of the training data, because the cost function for building the model ignores any training data that is close (within a threshold ε) to the model prediction.

These might be, for instance, exchange rates for some currency measured at subsequent days together with corresponding econometric indicators. The main goal of regression problems is to find a function f (x) that can correctly predict the observation values, y, of new input data points, x, by learning from the given training data set, S.

Here, learning from a given training data set means finding a linear or nonlinear surface that tolerates a small error in fitting this training data set. In ε -SV regression, our goal is to find a function f(x) that has at most ε deviation from the actually obtained targets yi for all the training data, and at the same time is as flat as possible. In other words, we do not care about errors as long as they are less than ε , but will not accept any Example of linear regression with one independent variable deviation larger than this. This may be important if you want to be sure not to lose more than ε money when dealing with exchange rates, for instance.

Also, applying the idea of support vector machines (SVMs) the function f(x) is made as flat as possible in fitting the training data. This problem is called ε-support vector regression (ε-SVR) and a data point $x_i \in R^n$ is called a **support vector** if $|f(x_i) - y_i| \geq \varepsilon$.

Conventionally, ε -SVR is formulated as a constrained minimization problem, namely, a convex quadratic programming problem or a linear programming problem. Such formulations introduce 2m more nonnegative variables and 2m inequality constraints that enlarge the problem size and could increase computational complexity for solving the problem.

For pedagogical reasons, we begin by describing the case of linear functions f, taking the form f(x)=<w,x>+b

where <. , .> denotes the dot product in X. *Flatness* in the case of (1) means that one seeks a small w. One way to ensure this is to minimize the norm i.e. $\| w \|^2 = \langle w,w \rangle$. We can write this problem as a convex optimization problem.

minimize $\frac{1}{2} \| w \|^2$

subject to $y_i - \langle w,x_i \rangle - b \leq \varepsilon$;
$\langle w,x_i \rangle + b - y_i \leq \varepsilon$;

The tacit assumption was that such a function f actually exists that approximates all pairs (xi; yi) with ε precision, or in other words, that the convex optimization problem is feasible. Sometimes, however, this may not be the case, or we also may want to allow for some errors. Analogously to the εsoft margin, loss function which was adapted to SV machines, one can introduce slack variables ξ_i and ξ_i^* to cope with otherwise infeasible constraints of the optimization problem. Hence we arrive at the formulation stated in

$$\min_{w,b,\xi_i,\xi_i^*} R(w,b,\xi_i,\xi_i^*) = C\sum_{i=1}^{n} (\xi_i + \xi_i^*) + \frac{1}{2} \| w \|^2$$

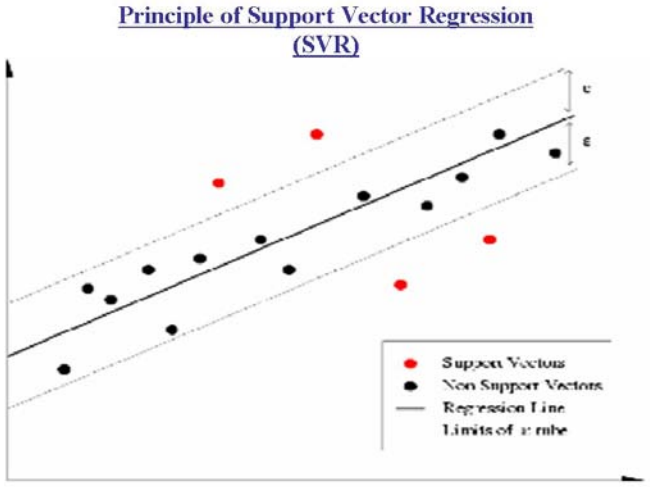
Subject to $y_i - \langle w,x_i \rangle - b \leq \varepsilon + \xi_i$; $\langle w,x_i \rangle + b - y_i \leq \varepsilon + \xi_i^*$; $\xi_i \geq 0$; $\xi_i^* \geq 0$;

The constant C > 0 determines the trade-off between the flatness of f and the amount up to which deviations larger than are tolerated.

After solving this optimization problem one can get the function f(x)

as $f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) < x_i, x > + b$

and this is the equation of hyper plane.



4. Support Vector Regression for Outliers Removal

An outlier is an observation that is numerically distant from the rest of the data. In larger samplings of data, some data points will be further away from the sample mean than what is deemed reasonable. This can be due to incidental systematic error or flaws in the theory that generated an assumed family of probability distribution, or it may be that some observations are far from the center of the data. Outlier points can therefore indicate faulty data, erroneous procedures, or areas where a certain theory might not be valid. However, in large samples, a small number of outliers is to be expected (and not due to any anomalous condition). Grubbs defined an outlier as: "An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs." Outliers can occur by chance in any distribution, but they are often indicative either of measurement error or that the population has a heavy-tailed distribution.

5. Results and Comparisons

Data Set Generation with and without Outliers:

Here we are working on 4 dim artificial data. To generate the data we need x1, x2, x3, x4 and a linear relationship between them. We generate x1 from a uniform distribution on the unit interval i.e. on the interval [0,1], x2 from a uniform distribution on the interval [1,4], x3 from a uniform distribution on the interval [-1,2], and x4 from a uniform distribution on the unit interval i.e. on the interval [0,1].

Y=2-3*x1+4*x2+x3+0*x4 is the linear relationship between x1,x2,x3 and x4 and our regression function y is given by normal distribution with the mean of Y and variance of 1.

6. Conclusion and Future Discussion

In this approach, we are finding the value of maximum epsilon by using the usual Support Vector Regression Method. And for value of epsilon less than maximum epsilon see the performance. If outliers are not present in the data set, then usual Support Vector do the better regression in compare to Multivariate Regression. But if outliers are present in the dataset then first we will have to remove that outlier by the above method and then do the regression on the inliers data points. So ultimately we are using Support Vector Regression Method twice so its complexity will increase by the factor of 2. SVR performance depends on a good setting of meta-parameters parameters C, ε and the kernel parameters. Parameter C determines the tradeoff between the model complexity (flatness) and the degree to which deviations larger than are tolerated in optimization formulation for example, if C is too large (infinity), then the objective is to minimize the empirical risk only, without regard to model complexity part in the optimization formulation. Here we are dealing with the value of C to be 100 and linear kernel.

References

[1] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression,"Royal Holloway College, London, U.K., Neuro COLT Tech. Rep.TR-1998-030, 1998.
[2] A. J. Smola, "Regression estimation with support vector learning machines,"Master's thesis, Technical Univ. Munchen, Munich, Germany,1998
[3] Barnett, V. and Lewis, T.: 1994, Outliers in Statistical Data. John Wiley & Sons, 3rd edition.
[4] C.J.C. Burges, (1998), "A tutorial on support vector machines for pattern recognition", Knowledge Discovery and Data Mining,
[5] H. Drucker et al., "Support vector regression machines," in Neural information Processing Systems. Cambridge, MA: MIT Press, 1997, vol. 9.
[6] S. Mukherjee, E. Osuna, and F. Girosi, "Nonlinear prediction of chaotic time series using a support vector machine," in Proc. NNSP, 1997, pp.24–26.

Author Profile

G. Malik received M.Sc in Mathematics from A.M.U and M.Tech in C.S from I.S.I. Kolkata India currently he is working as an Asst Professor in deptt of Mathematics R.L.A College (E), University of Delhi India.

M Tarique received M.Sc in Mathematics from B.H.U and M.Tech in C.S from I.S.I. Kolkata, India, currently he is working as an Asst Professor in department of Mathematics Dyal Singh College, University of Delhi India.