The Korean Communications
in Statistics Vol. 6, No. 2, 1999
pp. 337-343

Support Vector Machine for Linear Regression

Changha Hwang[1], Kyungha Seok[2]

Abstract

Support vector machine(SVM) is a new and very promising regression and classification technique developed by Vapnik and his group at AT&T Bell Laboratories. This article provides a brief overview of SVM, focusing on linear regression. We explain, from statistical point of view, why SVM might be attractive and how this could be compared with other linear regression techniques. Furthermore, we explain model selection based on VC-theory.

1. The Basic Idea

There are two types of SVMs, i.e., SVM for classification and SVM for regression. However, SVM classification can be viewed as a special case of SVM regression. SVM can be used for both linear and nonlinear regression. In this paper we explain SVM for linear regression. For details, see Gunn(1998) and Smola & Scholkopf(1998).

Suppose we are given training data $\{(\boldsymbol{x}_i, y_i),\ i=1,\cdots,n\} \subset \mathfrak{X} \times R$, where $\mathfrak{X}$ denotes the space of the input vectors, R^d. Our goal is to find a function $f(\boldsymbol{x})$ that has at most ε deviation from the actually obtained targets y_i's for all the training data, and at the same time, is as flat as possible. We now take the form

$$f(\boldsymbol{x}) \ = \ \boldsymbol{w}^t\boldsymbol{x} + b \text{ with } \boldsymbol{w} \in \mathfrak{X},\ b \in R$$

where superscript t represents the transpose of a vector. Flatness here means that one seeks small $\boldsymbol{w}$. One way to ensure this is to minimize the Euclidean norm $\|\boldsymbol{w}\|^2$. Formally we can write this problem as a convex optimization problem by requiring:

$$\text{minimize } \ \frac{1}{2}\|\boldsymbol{w}\|^2,$$

$$\text{subject to } y_i - \boldsymbol{w}^t\boldsymbol{x}_i - b \le \varepsilon \text{ and } \boldsymbol{w}^t\boldsymbol{x}_i + b - y_i \le \varepsilon$$

The underlying assumption here is 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. To make it feasible, we introduce slack variables ξ, ξ_i^*. Hence we arrive at the

[1] Dept. of Statistical Information, Catholic University of Taegu-Hyosung, Kyungbuk, Korea.

[2] Dept. of Data Science, Inje University, Kyungnam, Korea.

formulation stated in Vapnik(1995).

$$\text{minimize} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*),$$

$$\text{subject to} \quad \begin{cases} y_i - \boldsymbol{w}^t\boldsymbol{x}_i - b \le \varepsilon + \xi_i \\ \boldsymbol{w}^t\boldsymbol{x}_i + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i,\ \xi_i^* \qquad \ge 0 \end{cases}$$

The constant $C > 0$ determines the trade off between the flatness of f and the amount up to which deviations larger than ε are tolerated. Here, ξ, ξ_i^* are slack variables representing upper and lower constraints on the outputs. The formulation above corresponds to dealing with a so called ε-insensitive loss function $|\xi|_\varepsilon$ described by

$$|\xi|_\varepsilon = \begin{cases} 0 & \text{if } |\xi| \le \varepsilon \\ |\xi| - \varepsilon & \text{otherwise} \end{cases} .$$

The key idea is to construct a Lagrange function. Hence we proceed as follows:

$$L = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}\alpha_i(\varepsilon + \xi_i - y_i + \boldsymbol{w}^t\boldsymbol{x}_i + b)$$
$$- \sum_{i=1}^{n}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \boldsymbol{w}^t\boldsymbol{x}_i - b) - \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*)$$

We notice that the positivity constraints α_i, α_i^*, η_i, $\eta_i^* \ge 0$ should be satisfied. Hence we arrive at the optimization problem below.

$$\text{maximize} \quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\boldsymbol{x}_i^t\boldsymbol{x}_j - \varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n}y_i(\alpha_i - \alpha_i^*),$$

$$\text{subject to} \quad \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \quad \text{and} \quad \alpha_i,\ \alpha_i^* \in [0\ ,\ C]$$

Solving the above equation with these constraints determines the Lagrange multipliers, α_i, α_i^*, and the optimal regression function is given by

$$\overline{\boldsymbol{w}} = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\boldsymbol{x}_i, \quad \overline{b} = -\frac{1}{2}\overline{\boldsymbol{w}}^t[\boldsymbol{x}_r + \boldsymbol{x}_s],$$

where $\boldsymbol{x}_r$ and $\boldsymbol{x}_s$ are support vectors. The Karush-Kuhn-Tucker(KKT) conditions that are satisfied by the solution are,

$$\overline{\alpha}_i\overline{\alpha}_i^* = 0, \ i = 1, \cdots, n.$$

Hence the support vectors are points where exactly one of the Lagrange multipliers is greater than zero.

2. VC-theory Based Model Selection

For SVM regression, there are two parameters ε and C need to be defined by the user. Parameter ε controls the precision of the data fitting by adjusting the loss function.

Parameter C can be considered as the regularization parameter, which controls the trade-off between model complexity and the fitting of the data. Both parameters will affect the final model complexity, and therefore should be used in model selection. Model selection in SVM is still an open issue. Statistics is usually based on asymptotic properties of data. The Statistical Learning Theory (also known as VC-theory) provides comprehensively mathematical and conceptual framework for predictive learning with finite samples. VC-theory provides a very general and powerful framework for complexity control called Structual Risk Minimization(SRM). In VC-theory, the model complexity is defined as the VC-dimension, which coincides with the classical definition (the number of parameters) for linear parametrization.

Under SRM, a set of possible models (approximating functions) are ordered according to their complexity (or flexibility to fit the data). Specifically under SRM the set S of approximating functions $f(\boldsymbol{x}, \boldsymbol{w})$, $\boldsymbol{w} \in \Omega$ has a structure, that is, it consists of the nested subsets (or elements) $S_k = \{f(\boldsymbol{x}, \boldsymbol{w}), \boldsymbol{w} \in \Omega_k\}$ such that

$$S_1 \subset S_2 \subset \cdots \subset S_k \subset \cdots$$

where each element of the structure S_k has finite VC-dimension h_k. By design, a structure provides ordering of its elements according to their complexity (i.e., VC-dimension):

$$h_1 \le h_2 \le \cdots \le h_k \le \cdots$$

For regression problems with squared loss the following bound on prediction risk holds with probability $1-\eta$:

$$\textit{prediction risk} \le \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \left(1 - r\sqrt{\frac{h\left(\ln\left(\frac{an}{h}\right)+1\right) - \ln\eta}{n}}\right)_+^{-1}$$

where h is the VC-dimension of the set of approximating functions and r is a constant which reflects the tails of the loss function distribution, i.e., the probability of observing large values of the loss, and a is a theoretical constant. Here, $(u)_+ = \max(u, 0)$. Vapnik(1998) shows that we can use this bound with constant a close to 1, so we choose $a = 1$. We can also set $r = 1$. From a practical viewpoint, the confidence level of the above bound should depend on the sample size n, i.e., for larger sample sizes we should expect higher confidence level. So we set $\eta = 1/\sqrt{n}$. Making all these substitutions into the above bound gives the following penalization factor which we call Vapnik's measure:

$$r(p, n) = \left(1 - \sqrt{p - p\ln p + \frac{\ln n}{2n}}\right)_+^{-1}$$

where $p = h/n$. The common constructive implementation of SRM is to choose the optimal structure S_{opt} minimizes this prediction risk bound.

As we have seen, applying VC-based model selection requires the knowledge of the model complexity measure, i.e., the VC-dimension. Estimating the true VC-dimension of SVM for regression is not easy, since the empirical VC-dimension estimation method will result in tremendous amount of computation. We would like to introduce the following approximate VC-dimension measure for SVM. For details, see Shao(1999). Approximate VC-dimension for SVM is given by

$$h = \text{number}\{\alpha_i^{(*)} : 0 < \alpha_i^{(*)} < C\} - 1$$

$$\text{or} \quad h = \text{number}\{\text{support vectors}\} - \text{number}\{\alpha_i : \alpha_i^{(*)} = C\} - 1.$$

Here, $\alpha_i^{(*)}$ represents α_i or α_i^*. Since we have the estimate of VC-dimension, we can apply VC-bound for model selection.

3. Numerical Illustrations

To see how SVM performs in the linear regression problem on real data, let us look at three examples. For the comparisons on three data sets, we use least squares(LS) regression, least absolute deviations(LAD) regression, M-regression, nonparametric regression and ride regression besides SVM. Three data sets are acid content data, turnip green data and stack loss data. For details, see Birkes and Dodge(1993). Some results are taken from their book.

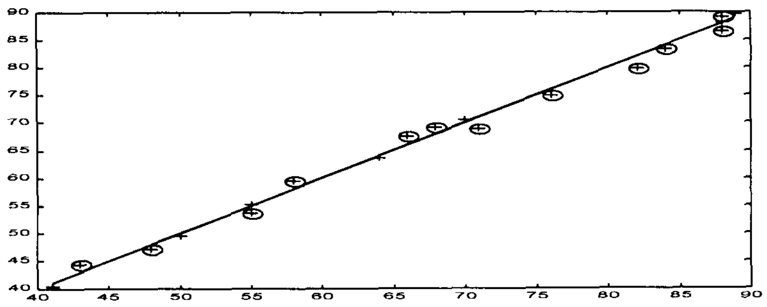

Fig. 1. SVM Linear Regression for Acid Content Data

Example 1. Consider the acid content data with no outliers. We see from Birkes and Dodge(1993) that all the data points fall closely around a straight line. For such a well-behaved data set, all the regression methods give very similar results. LS estimates of β_0, β_1 are 35.46 and 0.3216, respectively. Compared with the LS estimates of β_0 and β_1, the LAD estimates were within 2%, the M-estimates were exactly the same, the nonparametric estimates were within 1%, and the ridge estimates differed only in the fourth significant digit. SVM estimates of β_0, β_1 are 35.35 and 0.3219, respectively. All the estimates of σ were not so close; they were 1.230, 1.233, 1.433, 1.595, and 1.364 for LS, SVM, LAD, M-, and nonparametric regression, respectively. Hereafter, the estimate of σ for LS and SVM is the square root of mean squares error. The size of the intensive zone ε, and the constraint parameter C minimizing VC-bound were 0.7143 and 9890, respectively. To conclude, we can say SVM works as well as LS does for the acid content data with no outliers. In Fig. 1, support vectors are circled. In many practical problems, only a small amount of input vectors turn into support vectors. By the way, in this example 12 out of 20 points were support vectors. ■

Example 2. Let us apply all six regression methods to turnip green data. Table 1 lists the estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$ of the regression coefficients, the estimate $\hat{\sigma}$ of the standard deviation of the error population, and number N_o of standardized residuals with absolute value larger than 2.5.

Table 1. Results on Turnip Green Data

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\sigma}$	N_o
LS	119.6	-0.03367	5.425	-0.5026	-0.1209	6.104	0
LAD	133.8	-0.03367	6.635	-0.6974	-0.1460	4.140	4
M-regression	122.7	-0.03967	5.763	-0.5443	-0.1282	4.177	4
Nonparametric	123.7	-0.04478	6.043	-0.5583	-0.1339	4.509	3
Ridge	115.9	-0.02805	4.807	-0.4363	-0.1089		
SVM	116.9	-0.04060	4.653	-0.4001	-0.1071	6.386	

LAD, M-, and nonparametric regression are especially suitable when there are outliers in the data. In all the three methods the data points numbered 10, 19, and 20 had standardized residuals more than 2.5 in absolute value and hence may be regarded as outliers. The point numbered 15 was also detected as an outlier by LAD and M-regression and was almost detected by the nonparametric procedure. All of these points were detected as support vectors by SVM. SVM usually picks the outlier as one of the support vectors. However, the fitting result has not really been pulled up or down a lot. This good virtue is due to the robust loss function and complexity control. Here, ε and C minimizing VC-bound were 0.4490 and 1, respectively. This small amount of C is due to collinearity.

For the ridge regression, only the regression coefficients are given because this method is intended to be used only for estimation. The first five estimation methods in the table produce estimated coefficients that are noticeably different. SVM, LAD, M-, and nonparametric regression are especially suitable when there are outliers in the data. SVM and Ridge regression are especially suitable when there is collinearity among the explanatory variables. Because of the high correlation of 0.997 between X_2 and $X_4 (= X_2^2)$, we expect ridge regression and SVM to give more accurate estimates of β_2 and β_4 than LS. Note that ridge estimates $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\beta}_4$ are all closer to 0 than the corresponding LS estimates. We can see the same phenomena for SVM. This agrees with the description of ridge regression as a procedure that shrinks the LS estimates. Actually, SVM is a kind of ridge regression. Therefore, we can conclude SVM works well when there are outliers in the data and is collinearity among the explanatory variables. In this example 25 out of 27 points were support vectors. ■

Examples 3. Next we consider a data set that appeared as example in many books and articles. The data consist of measurements from a factory for the oxidation of ammonia to nitric acid. On 21 different days, measurements were taken of the air flow(X_1), the temperature of cooling water(X_2), the concentration of acid(X_3), and the amount of ammonia that escaped before being oxidized, called stack loss(Y). All six regression methods were applied using the model $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e$. Table 2 shows the estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\sigma}$, number N_o of standardized residuals with absolute value larger than 2.5.

Table 2. Results on Stack Data

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_4$	$\hat{\sigma}$	No
LS	-39.92	0.7156	1.295	-0.1521	3.243	0
LAD	-39.69	0.8319	0.574	-0.0609	2.171	3
M-regression	-41.17	0.8133	1.000	-0.1324	2.661	1
Nonparametric	-40.16	0.8155	0.888	-0.1202	2.920	1
Ridge	-40.62	0.6861	1.312	-0.1273		
SVM	-38.00	0.8431	0.730	-0.1280	3.476	

There are substantial differences in the estimates of $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ for the six methods. This is at least partly due to outliers. As in Example 2, the M- and nonparametric estimates are similar to one another. Here, ε and C minimizing VC-bound were 0.8572 and 16632, respectively. In this example 15 out of 21 points were support vectors. It is hard to say whether SVM is good for this example or not. However, we can conclude that SVM works well for this example because there are many support vectors contain the information required to summarize the data. ■

In the above three examples we have applied six different mothods of regression to the same set of data for the purpose of comparing the methods. If our only purpose is to analyze the data, it would still be good practice to apply more than one regression methods. If you use several methods to analyze a data set and they all lead to similar results, you can feel confident about your conclusion. If there are serious disagreements between the results of the different methods, you should examine the data more closely to see the reason.

We would recommend using least squares and one another method. The personal preference of the authors is SVM because it focuses on both data fitting and generalization.

References

[1] Birkes, D. and Dodge, Y.(1993). Alternative Methods of Regression, John Wiley and Sons, Inc., New York.

[2] Gunn, S. (1998). Support Vector Machines for Classification and Regression, ISIS Technical Report, U. of Southampton.

[3] Shao, X. (1999). Model Selection Using Statistical Learning Theory, Ph. D. Thesis, U. of Minnesota.

[4] Smola, A.J. and Scholkopf, B. (1998). A Tutorial on Support Vector Regression, NeuroCOLT2 Technical Report, NeuroCOLT.

[5] Vapnik, V., Levin, E. and LeCun, Y. (1994). Measuring the VC-Dimension of a Learning Machine, *Neural Computation*, **6**, 851-876.

[6] Vapnik, V. (1995). The Nature of Statistical Learning Theory, Springer.

[7] Vapnik, V. (1998). Statistical Learning Theory, Springer.