Housing Price Prediction: a comparison among ε-Support Vector Regression, standard and regularized regression models

Yuanjie Jin
Department of Biology
Duke University
Durham, NC 27708
yj20@duke.edu

Abstract

ε-Support Vector Regression (ε-SVR) is a regression method that uses a unique ε-insensitivity loss function[1]. This study compares the predictions of Boston housing prices by ε-SVR using the Radial Basis Function (RBF) kernel and the linear kernel with those by the standard multivariate linear regression (MLR), ridge and lasso regression models. Among the five models, ε-SVR with the RBF kernel predicts best, linear ε-SVR, ridge regression and MLR intermediate and lasso regression the worst, demonstrating the power of nonlinear regression implemented by ε-SVR with specific kernel functions in modeling and predicting housing price.

1 Introduction

1.1 Background

To successfully sell a house, the house owner must offer a reasonable price. In order to determine the right price, it is necessary to have a rough prediction of it based on all relevant characteristics of the house. Given a dataset comprising known local house prices and the values of covariates that presumably affect the house pricing, a common approach to predicting the price of a future house is through multivariate linear regression (MLR). However, there are usually a large number of factors that drive the price, which can make standard linear regression difficult to fulfill the task. For example, if the mapping from predictors to the response is not very smooth, the linear paradigm needs an exponentially increasing number of samples with an increasing number of independent variables. This is known as the curse of dimensionality[2]. In comparison, Support Vector Machine (SVM) is a machine learning method that has been developed to work with data sets that are typically high dimensional and sparse[1] (data set contains a small number of the training data pairs). In addition to its use in classification problems, SVM can be extended to cases where the response variable is continuous, an application named Support Vector Regression[1] (ε-SVR). In this project, I propose to compare the performance between ε-SVR and other canonical regression models including ridge and lasso in housing price prediction. I am also interested in studying differences between linear and nonlinear kernels in ε-SVR. The dataset used in this study is obtained from the UCI machine learning repository, which contains housing prices in suburbs of Boston[3]. There are 506 data points with 12 numeric attributes (crime rate, age, etc.), and one binary categorical attribute (tract bounds Charles River or not).

1.2 Ridge and Lasso Regression

For the linear model $Y = X\beta + \varepsilon$, where β is a $(p+1) \times 1$ vector containing the coefficients of covariates, X is a $n \times (p+1)$ design matrix, and $\varepsilon \sim N(0, \sigma^2 I)$. The standard MLR tries to minimize the quadratic error:

$$\min_{\beta} \frac{1}{n} \|Y - X\beta\|^2 \tag{1}$$

Ridge regression, on the other hand, seeks to minimize the quadratic error and a L_2 penalty:

$$\min_{\beta} \frac{1}{n} \|Y - X\beta\|^2 + \lambda \|\beta\|^2 \tag{2}$$

whereas Lasso regression seeks to minimize the quadratic error plus a L_1 penalty, that is,

$$\min_{\beta} \frac{1}{n} \|Y - X\beta\|^2 + \lambda \|\beta\| \tag{3}$$

By penalizing L_2 and L_1 norm of coefficients β, respectively, ridge and lasso regression shrink the estimated coefficients toward zero, and thus reduce the variance in the prediction, especially when some covariates are highly correlated.

1.3 Support Vector Regression

ε-SVR is a regression technique different from abovementioned regression methods in that it uses a novel loss function called Vapnik's ε-insensitivity loss function.

Consider the problem of approximating the set of data $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, with a linear function

$$f(x, w) = \langle w, x \rangle + b \tag{4}$$

then the ε-insensitivity error function is defined as[4]

$$|y - f(x, w)|_\varepsilon = \begin{cases} 0, & \text{if } |y - f(x, w)| \leq \varepsilon \\ |y - f(x, w)| - \varepsilon, & \text{otherwise} \end{cases} \tag{5}$$

where ε specifies the width of the margin.

The optimal regression function is given by[4]

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \tag{6}$$

$$\text{subject to : } \begin{cases} y_i - \langle w, x_i \rangle - b \geq \varepsilon + \xi_i & i = 1, \ldots, n \\ \langle w, x_i \rangle + b - y_i \geq \varepsilon + \xi_i^* & i = 1, \ldots, n \\ \xi_i, \xi_i^* \geq 0 & i = 1, \ldots, n \end{cases} \tag{7}$$

where $\frac{1}{2}\|w\|^2$ is the regularization term that reflects the model complexity, ξ_i and ξ_i^* are slack variables representing upper and lower constraints on the outputs of the system, respectively, and the cost $C > 0$ is a pre-specified parameter which controls the trade-off between the model complexity and the amount up to which deviations larger than ε are tolerated.

Such optimization problem could be solved by Lagrange Multiplier, which gives the following solution[4]:

$$\alpha, \alpha^* = \arg\min_{\alpha, \alpha^*} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle - \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) y_i + \sum_{i=1}^{n} (\alpha_i + \alpha_i^*) \varepsilon \tag{8}$$

$$\text{subject to: } 0 \leq \alpha_i, \alpha_i^* \leq C, \quad i = 1, \ldots, n \tag{9}$$

$$\sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0 \tag{10}$$

Solving Equation (8) with constraints Equations (9) and (10) determines the Lagrange multipliers, α and α^*, and then we have[4]

$$w = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) x_i, \tag{11}$$

$$\text{and thus } f(x) = \langle w, x \rangle + b = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \langle x_i, x \rangle + b \tag{12}$$

where b is given by[4]

$$\max_{\alpha_i < C \text{ or } \alpha_i^* > 0} \{-\varepsilon + y_i - \langle w, x_i \rangle\} \leq b \leq \min_{\alpha_i^* < C} \{-\varepsilon + y_i - \langle w, x_i \rangle\} \tag{13}$$

After the learning process, the number of support vectors is equal to the number of nonzero α_i and α_i^*. Moreover, since w can be described as a linear combination of the training patterns x_i, the complexity of a function's representation is independent of the dimensionality of the input space χ, and depends only on the number of support vectors. This independency is a strength of ε-SVR in dealing with high dimensional input. It is also useful for the formulation of a non-linear extension, in which the kernel approach is employed and $\langle x_i, x \rangle$ is replaced by $K(x_i, x)$ in the equation (12).

2 Related work

The same dataset has been used in multiple machine learning studies, including, but not limited to, Active Set Support Vector Regression[5] and Gaussian Process Networks[6]. However, the comparison in prediction performance among ε-SVR, MLR and regularized regression models ridge and lasso has not been done before. Therefore, this study helps to shed some light on which of these models is most suitable for the task of housing price prediction.

3 Methods

The whole dataset were randomly partitioned into two parts: $\frac{2}{3}$ of it is for model training and the rest $\frac{1}{3}$ for testing. The R package 'e1071'[7], 'MASS' and 'glmnet' were used to perform ε-SVR analysis, ridge regression and lasso regression, respectively.

For ε-SVR, the Radial Basis Function (RBF) kernel ($exp\{-\gamma|u-v|^2\}$) and the linear kernel ($u^T v$) were adopted for non-linear and linear regression, respectively. Data were scaled internally (both x and y variables) to zero mean and unit variance to improve the results. A grid search was conducted to find the optimal combination of hyperparameters (C and ε for the linear kernel; C, ε and γ for the RBF kernel), which minimize 10-fold cross-validation errors using the training set.

For ridge and lasso regression, their optimal penalizing parameter λ were chosen such that they minimize generalized cross-validation (GCV) error and Mallows's C_p value, respectively. The standard MLR was performed as well for benchmarking.

For method validation, prediction accuracy (mean squared error, or MSE) and R^2 (squared correlation coefficient) were measured in testing stage. The whole process from data division to model training and testing were repeated 10 times and the results of the 10 replications were summarized.

4 Results

By cross-validation-based grid search, the best values for the cost C and the margin width ε in ε-SVR using the linear kernel were found to be 7.4 and 0.1, respectively. Also by grid search, the optimal combination of hyperparameters for ε-SVR using the RBF kernel were identified as $C = 1000$, $\gamma = 0.001$ and $\varepsilon = 0.1$. These hyperparameter values were fixed

throughout the ε-SVR training of different subsets of the original data. The same subsets of data were also used to train MLR, ridge and lasso regression. Table 1 summarizes the average MSE and R^2 of these trained models on the same testing sets (10 replications).

Table 1: Average MSE and R^2 for different models in housing price prediction (10 replications)

Model	$\mathbf{MSE}_{test}$	$\mathbf{R}^2_{test}$
MLR	22.696	0.738
Ridge	22.607	0.739
Lasso	41.919	0.554
ε-SVR (linear)	22.930	0.740
ε-SVR (RBF)	12.233	0.859

Figure 1 and 2 are side-by-side boxplots of MSE and R^2 that vividly reflect testing accuracy by different models. As shown in Table 1, Figure 1 and 2, the ridge regression displays a similar performance as the standard MLR. ε-SVR using the linear kernel does not differ much from either standard MLR or ridge regression model, in terms of MSE and R^2 for predictions. In contrast, ε-SVR using the RBF kernel significantly improves the performance, which produces the lowest average MSE (12.233, see Table 1) and the highest average R^2 (0.859, see Table 1) among all the models tested here. Interestingly, the predictions by the lasso regression turn out to be less accurate and less stable than any other model, which has the sample mean and standard deviation of its MSE being 41.919 (Table 1) and 10.780, respectively, and the sample mean and standard deviation of its R^2 being 0.554 (Table 1) and 0.099, respectively.

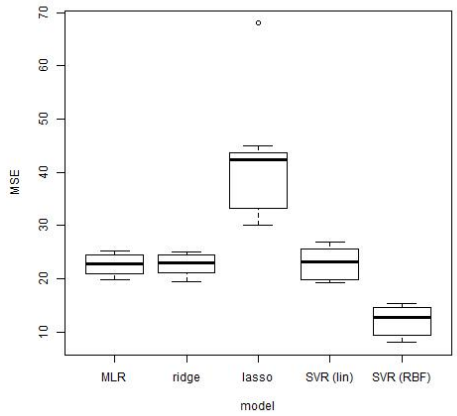

Figure 1: Boxplot of MSE in predictions by different models for 10 replications. SVR (lin) and SVR (RBF) are ε-SVR models using linear and RBF kernel, respectively.

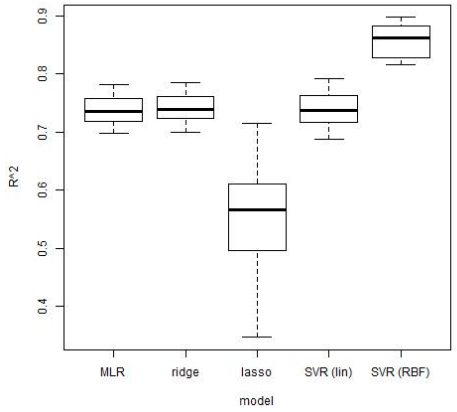

Figure 2: Boxplot of R^2 in predictions by different models for 10 replications. SVR (lin) and SVR (RBF) are ε-SVR models using linear and RBF kernel, respectively.

5 Discussion

This study compares implementations of MLR, ridge regression, lasso regression, linear ε-SVR and ε-SVR with the RBF kernel in terms of Boston housing price prediction accuracy. Using MSE as the test error, ε-SVR with the RBF kernel does the best in prediction. This is not quite surprising, because the RBF kernel is a nonlinear kernel function which can capture nonlinear relationships in the data, whereas the other four linear regression models fail to do so. Also, it suggests that there exists nonlinear relationships in this data set, which is pretty common in reality. Indeed, the RBF kernel is recommended as a general-purpose kernel used when there is no prior knowledge about the data[8], whereas the linear kernel is

useful when dealing with large sparse data vectors[4], which is obviously not the case here (506 data vectors vs. 13 features).

Slightly worse than the best model, linear ε-SVR, ridge regression and the standard MLR give similar performances. This is probably because the data space has a much higher dimension than the feature space and there is no high correlations among covariates, such that the underlying assumptions of MLR largely hold. Surprisingly, among the five regression models, lasso regression produces the worst prediction accuracy. One possibility is that there is no collinearity between covariates in this case, so the increased bias in lasso regression dominates the MSE and leads to poor predictions.

6 Conclusion

In Boston house price prediction, ε-SVR with the RBF kernel works better than linear ε-SVR, MLR, ridge and lasso regression, revealing the involvement of nonlinear relationship in the Boston housing data set and thus the advantage of RBF kernel over the linear kernel in modeling the complex data. This analysis also illustrates the power of ε-SVR in nonlinear regression by introduction of kernels. Last but not least, linear ε-SVR, ridge and lasso regression do not necessarily perform better than the standard MLR when there are far more observations than covariates and not much multicollinearity is involved.

References

[1] Brereton, R.G. & Lloyd, G.R. (2010) Support vector machines for classification and regression. *The Analyst*, **135**(2):230-67.

[2] Bellman, R. (1961) *Adaptive Control Processes.* Princeton University Press.

[3] Bache, K., & Lichman, M. (2013) UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. *Journal of Neuroscience* Irvine, CA: University of California, School of Information and Computer Science.

[4] Gunn, S. R. (1998) Support Vector Machines for Classification and Regression. *Technical Report*, University of Southampton.

[5] Musicant, D.R. & Feinberg, A. (2004) Active set support vector regression. *IEEE Transactions on Neural Networks*, **15**(2):268 - 275.

[6] Friedman, N. & Nachman, I. (2000) Gaussian Process Networks. *UAI*, pp. 211

[7] Meyer, D., et al (2014) Support Vector Machinesthe Interface to libsvm in package e1071. [http://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf]. FH Technikum Wien, Austria

[8] Hsu, C.-W., et al (2010) A Practical Guide to Support Vector Classification. [http://www.csie.ntu.edu.tw/ cjlin/papers/guide/guide.pdf]. Department of Computer Science, National Taiwan University, Taiwan.