Communications for Statistical Applications and Methods
2015, Vol. 22, No. 2, 201–207

DOI: http://dx.doi.org/10.5351/CSAM.2015.22.2.201
Print ISSN 2287-7843 / Online ISSN 2383-4757

Kernel-Trick Regression and Classification

Myung-Hoe Huh[1,a]

[a]Department of Statistics, Korea University, Korea

Abstract

Support vector machine (SVM) is a well known kernel-trick supervised learning tool. This study proposes a working scheme for kernel-trick regression and classification (KtRC) as a SVM alternative. KtRC fits the model on a number of random subsamples and selects the best model. Empirical examples and a simulation study indicate that KtRC's performance is comparable to SVM.

Keywords: Kernel trick, support vector machine, subsampling, cross-validation.

1. Background and Aim

Consider $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$, the p-variate $\mathbf{x}$ and numerical y for regression or binary $y\,(= \pm 1)$ for classification. We assume that $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are standardized and that $y_1, \ldots, y_N$ are centered for numerical case. Standard linear model can be stated as

$$y = \mathbf{x}^T \underline{\beta} + \epsilon,$$

where ϵ is an error with the mean 0 and a finite variance. Classic kernel-trick regression and classification can be stated as follows (Schölkopf and Smola, 1998; Hastie *et al.*, 2009; Fukumizu, 2010).

- Transform $\mathbf{x}_1, \ldots, \mathbf{x}_N$ of p-dim Euclidean space to $\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_N)$ of Hilbert space. Explicit form of the transformation is not necessary.

- Project $\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_N)$ on a linear combination $\mathbf{v}$ of $\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_N)$. Write

$$\mathbf{v} = d_1 \Phi(\mathbf{x}_1) + \cdots + d_N \Phi(\mathbf{x}_N).$$

- The projection of $\Phi(\mathbf{x}_i)$ on $\mathbf{v}$ is given by

$$\sum_{i'=1}^{N} < \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_{i'}) > d_{i'} = \sum_{i'=1}^{N} k_{i,i'} d_{i'},$$

  where $k_{i,i'}$ is the $(i, i')^{th}$ element of K, $k_{i,i'} = < \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_{i'}) >$. Here, we use a reproducing kernel $K(\mathbf{x}, \mathbf{x}')$ for $< \Phi(\mathbf{x}), \Phi(\mathbf{x}') >$.

- Hence, the projections of $\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_N)$ on $\mathbf{v}$ are stacked in $K\mathbf{d}$ for given coefficient vector $\mathbf{d}$ of length N.

[1] Department of Statistics, Korea University, Anam-Dong 5-1, Sungbuk-Gu, Seoul 136-701, Korea.
E-mail: stat420@korea.ac.kr

Published 31 March 2015 / journal homepage: http://csam.or.kr

- Match $\mathbf{y}$ and $K\mathbf{d}$, by choosing $\mathbf{d}$ appropriately. Exact match can be made by

$$\mathbf{d} = K^{-1}\mathbf{y},$$

provided that K is of full rank. But, it is desirable to restrict the magnitude of $\mathbf{v}$ for stability. Since $\mathbf{v} = \sum_{i'=1}^{N} \Phi(\mathbf{x}_{i'})d_{i'}$, $\|\mathbf{v}\|^2 = \mathbf{d}^T K\mathbf{d}$. Thus we consider the minimization of

$$\| \mathbf{y} - K\mathbf{d} \|^2 + \lambda\, \mathbf{d}^T K\mathbf{d},$$

for $\lambda > 0$.

- Therefore, we obtain $\mathbf{d} = (K + \lambda I_N)^{-1}\mathbf{y}$. λ is called the "ridge" parameter in the literature. Large λ stabilizes the fit but it induces the bias.

- To predict the response for new unit with $\mathbf{x}^*$, project $\Phi(\mathbf{x}^*)$ on $\mathbf{v}$: For regression,

$$\hat{y}^* = \mathbf{k}^{*T}\mathbf{d},$$

where $\mathbf{k}^*$ of length N is the vector of inner products between $\Phi(\mathbf{x}^*)$ and $\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_N)$.

- For binary classification, prediction for new input $\mathbf{x}^*$ is given by

$$\hat{y}^* = \text{sign}\left(\mathbf{k}^{*T}\mathbf{d}\right).$$

As demonstrations, KtRC (kernel-trick regression and classification) is applied to three Monte-Carlo datasets simulated as follows ($N = 100$):

1) $x_1, x_2, \ldots, x_n \sim \text{Uniform}(-2, 2)$; $\quad y_i \,|\, x_i \sim N\left(0.5x_i, 0.5^2\right)$, $\quad i = 1, \ldots, N$.

2) $x_1, x_2, \ldots, x_n \sim \text{Uniform}(-2, 2)$; $\quad y_i \,|\, x_i \sim N\left(x_i + 0.5x_i^2, 0.5^2\right)$, $\quad i = 1, \ldots, N$.

3) $x_1, x_2, \ldots, x_n \sim \text{Uniform}(-2, 2)$; $\quad y_i \,|\, x_i \sim N\left(0.5x_i^3, 0.5^2\right)$, $\quad i = 1, \ldots, N$.

Figure 1 shows fitted KtRC curves for the linear and quadratic cases and Figure 2 shows fitted KtRC curves for the cubic case. True signals are represented by dotted lines in the figures. For model construction, Gaussian kernel with $\sigma = 0.1$ and ridge parameter $\lambda = 0.2$ are used.

Both plots of Figure 1 are fine, but the left plot of Figure 2 shows substantial lack-of-fit. So, the ridge parameter λ is reset to 0.01. Then, the right plot of Figure 2 looks all right. One lesson learned through the Monte-Carlo cases is that the choice of ridge parameter λ is critical for the KtRC performance.

Section 2 proposes a working scheme for kernel-trick regression and classification (KtRC) that handles the optimal choice of λ. In Section 3, the proposed method is applied to two real datasets and model performance measures are computed. In Section 4, a simulation study is reported.

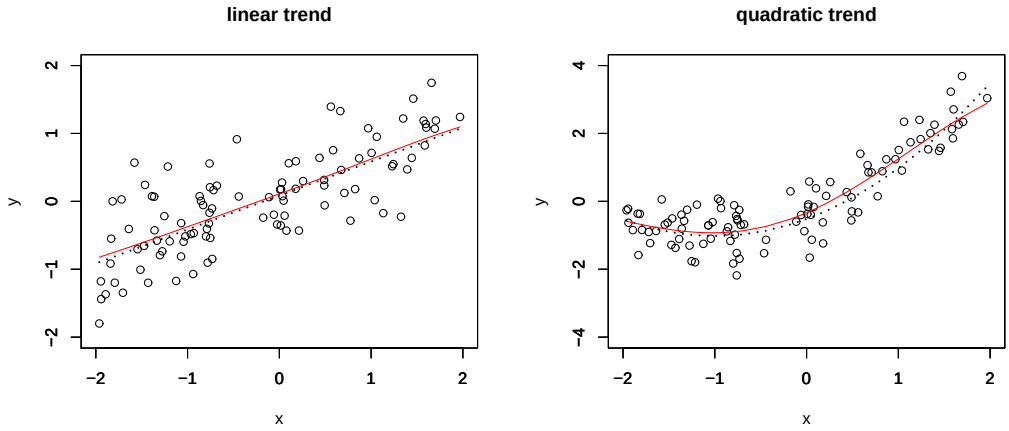

Figure 1: *Dataset with linear [Left] and quadratic trends [Right]*

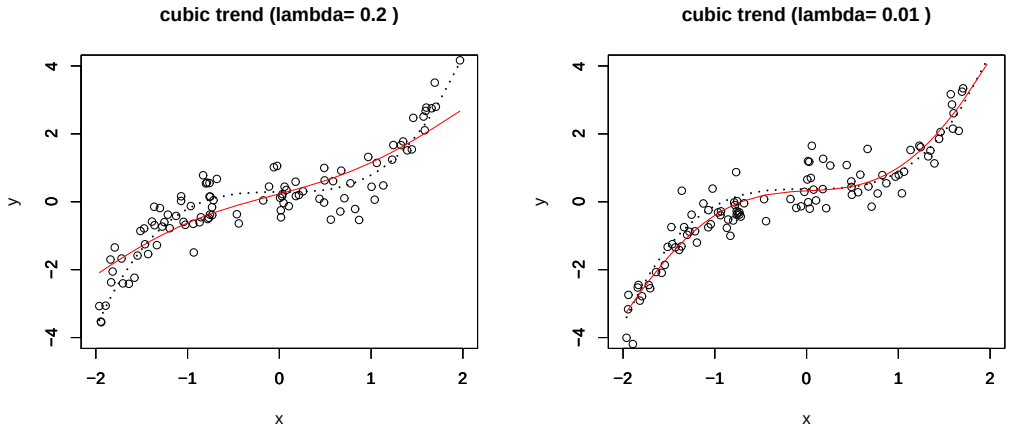

Figure 2: *Dataset with the cubic trend with different λ's*

2. Working Scheme for KtRC

Kernel-trick method of Section 1 has the following difficulties:

- It needs the inversion of $K + \lambda I_N$ and could be expensive for large N.
- Performance of the fitted model depends on the choice of λ.
- In addition, the analyst should specify the kernel type and its parameters. In this paper, we fix the kernel type to Gaussian:

$$< \Phi(\mathbf{x}), \Phi(\mathbf{x}') >= \exp\left(-\sigma \| \mathbf{x} - \mathbf{x}' \|^2\right), \quad \sigma > 0.$$

With the first two problems in mind, we propose the following scheme.

- Draw n units without replacement from the whole sample of N units. Denote n drawn units by $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ and $N - n$ remaining units by $(\mathbf{x}_1^*, y_1^*), \ldots, (\mathbf{x}_{N-n}^*, y_{N-n}^*)$.
- Construct the prediction model.
- Predict the outcomes for $N - n$ remaining units in the sample and evaluate the model performance. For regression, we adopt

$$\text{MAE} = \text{median}\,\{|y_i^* - \hat{y}_i^*|,\ i = 1, \ldots, N - n\}$$

as a performance measure. Alternatively, we also adopt

$$\text{MSE} = \text{mean}\left\{(y_i^* - \hat{y}_i^*)^2,\ i = 1, \ldots, N - n\right\}.$$

For binary classification, we use the number of classification error as a performance measure.

- Repeat the subsampling and fitting/evaluation process R (= 1,000) times. Retain the subsample $(\mathbf{x}_1^0, y_1^0), \ldots, (\mathbf{x}_n^0, y_n^0)$ of size n that recorded the best performance in R repetitions.
- Future predictions are made with KtRC model fitted by $(\mathbf{x}_1^0, y_1^0), \ldots, (\mathbf{x}_n^0, y_n^0)$.

Ridge parameter λ and/or kernel parameter σ could be chosen by comparing the performance for the various choices of the parameter(s). The next question is if the model found is better than the support vector machine (SVM). For fair competition between two models, we need a new sample of significant size which is independently obtained. In the following section, we apply the proposed scheme to two real datasets and evaluate the model performance by the k-fold cross validation (k = 5, 10).

3. Numerical Examples

3.1. Ozone data for regression

The dataset consists of 330(= N) consecutive measurements of the ozone and possibly related 8(= p) atmospheric variables. Thus the model is fitted by 248(= n) randomly selected units (n/N = 0.75), and the remaining 82(= $N - n$) units are used to measure the model performance. Parameter σ of Gaussian kernel is set to 0.125(= p^{-1}). The following results are obtained with 1000(= R) repetitions.

For ridge parameter $\lambda = 0.1, 0.2, 0.3, 0.4$, MAE is shown in two independent trials for each case as follows.

$$\begin{aligned}
\lambda = 0.1: &\quad \text{MAE} = 1.54, 1.57, &\quad \sqrt{\text{MSE}} = 3.13, 3.16.\\
\lambda = 0.2: &\quad \text{MAE} = 1.35, 1.45, &\quad \sqrt{\text{MSE}} = 3.03, 2.80.\\
\lambda = 0.3: &\quad \text{MAE} = 1.33, 1.49, &\quad \sqrt{\text{MSE}} = 2.90, 3.00.\\
\lambda = 0.4: &\quad \text{MAE} = 1.41, 1.42, &\quad \sqrt{\text{MSE}} = 3.15, 2.92.
\end{aligned}$$

Hence, $\lambda = 0.2$ is selected. For fair evaluation of the model performance, 10-fold cross-validation is executed. Means of 10 MAE's are 2.44 (with standard deviation 0.27) and 2.28 (with s.d. 0.52), while means of 10 $\sqrt{\text{MAE}}$'s are 4.03 (with standard deviation 0.55) and 3.95 (with s.d. 0.64).

In comparison, SVM regression with Gaussian kernel ($\sigma = 0.125, \epsilon = 0.1, C = 1$) produces the average MAE 2.11 (s.d. 0.35) and 2.21 (s.d. 0.41) and the average $\sqrt{\text{MAE}}$ 3.89 (s.d. 0.45) and 3.91 (s.d. 0.76) in the 10-fold cross-validation. Hence SVM regression appears a little better than KtRC. However, the two regression fits are very close (Figure 3).

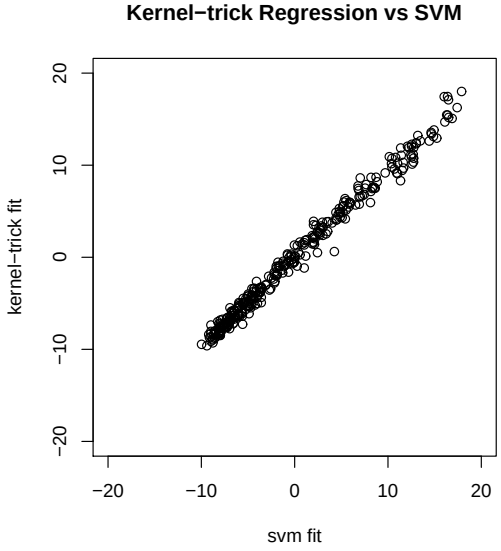

Figure 3: *Kernel-trick regression fit vs SVM regression fit*

3.2. Spam data for classification

The dataset consists of 57 textual characteristics of 4,601(= N) e-mails with a classified tag of either non-spam (–1) or spam (1). We used 1150 (= n (= $0.25N$)) e-mails for fitting the model and $R = 1000$. When Gaussian kernel with $\sigma = 0.1$ is applied to simple kernel-trick classifier, the percentage of incorrect classifications %.errors for leave-out 3,451 (= $N - n$) mails is as follows (Each case was replicated twice).

$$\begin{aligned}
\lambda = 0.1: &\quad \%.\text{errors} = 7.2\%, 7.1\%.\\
\lambda = 0.2: &\quad \%.\text{errors} = 7.0\%, 6.9\%.\\
\lambda = 0.3: &\quad \%.\text{errors} = 6.9\%, 6.6\%.\\
\lambda = 0.4: &\quad \%.\text{errors} = 6.9\%, 6.7\%.
\end{aligned}$$

Thus, λ is set to 0.3 for the subsequent cross-validation. Cross-validation with 5-folds turns out that the average %.errors are 8.13% (s.d. 0.74) and 8.00% (s.d. 0.90). In comparison, SVM with $\sigma = 0.1$ and $C = 10$ which is best tuned among nine combinations of $\sigma = 0.1, 1, 10$ and $C = 0.1, 1, 10$ showed the average %.errors 8.28% (s.d. 0.96) and 8.59% (s.d. 1.1). Hence, in the case of spam dataset, KtRC performs a little better than SVM.

4. Simulation Study

We design a simple simulation study that indicates the performance of Kernel-trick Regression and Classification scheme proposed in Section 2.

- Four-variate N observations (X_1, X_2, X_3, X_4) are generated independently from $N(0, 1)$. Denote the realizations by $(x_{i1}, x_{i2}, x_{i3}, x_{i4})$, $i = 1, \ldots, N$. N is set to 1000.

- Conditional on each $(x_{i1}, x_{i2}, x_{i3}, x_{i4})$, Y assumes $+1$ with probability θ, or -1 with probability $1 - \theta$,

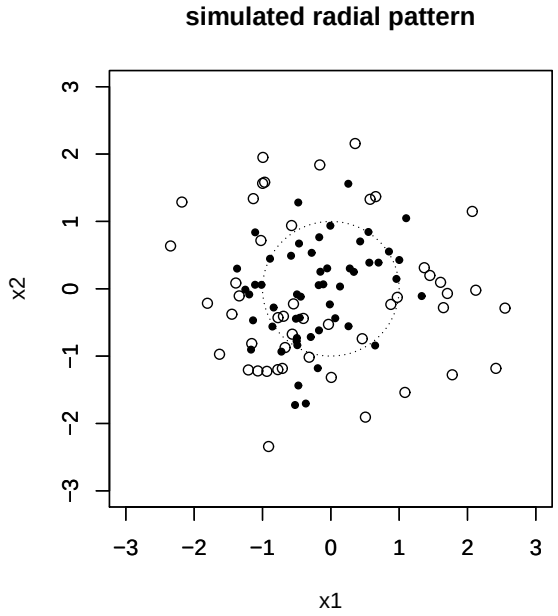

Figure 4: *Simulated observations on the* (x_1, x_2)*-plane: Filled circles for* $Y = 1$ *and unfilled circles for* $Y = -1$.

Table 1: Misclassification percentages of KtRC and SVM classifiers.

	KtRC ($f = 0.5$)	KtRC ($f = 0.75$)	SVM
mean	26.9	26.3	26.5
sd	14.0	14.1	14.2

where

$$\theta = \theta(x_1, x_2, x_3, x_4) = \exp\left(-\frac{1}{2}\left(x_1^2 + x_2^2\right)\right).$$

Hence, for the group variable Y, X_1 and X_2 are information carriers while the other two variables are not. Marginal hitting probability $P(Y = 1)$ is approximately equal to 0.5. Figure 4 shows a typical case.

- Test dataset of size N is generated from the same postulated stochastic mechanism as the training data.

We apply KtRC with the kernel parameter $\sigma = 0.1$ and the ridge parameter $\lambda = 0.2$, which are selected as the best among nine combinations of $\sigma = 0.01, 0.1, 1$ and $\lambda = 0.1, 0.2, 0.4$. Fraction $f = n/N$ of the subsample for fitting the model is set to either 0.5 and 0.75. The number R of repetitions is set to 500.

KtRC is to be compared with SVM classifier with kernel parameter $\sigma = 0.1$ and the unit cost $C = 10$, which are selected as the best among nine combinations of $\sigma = 0.01, 0.1, 1$ and $C = 1, 10, 100$.

Table 1 summarizes the results that are obtained from KtRC and SVM with 400 training datasets and 400 test datasets. Numbers represent percentages of misclassification. This simulation study indicates that KtRC's performance is similar to the SVM classifier.

5. Remarks

The kernel-trick regression and classification (KtRC) proposed in this study shows comparable performance with SVM. Currently, SVM computes the model faster than KtRC which simply relies on Monte-Carlo repetitions.

However, there are two potential strong points of KtRC: 1) KtRC fits the model with a smaller number n of training units, compared to N units for SVM. Hence for the case of "large" N, KtRC can be a viable choice compared to SVM. 2) KtRC computing can be distributed easily into parallel machines, so that it could be scalable for the datasets of "big" N.

Many questions remain unanswered such as the choice of the fraction f of the subsample size for fitting. The author personally prefers $f = 0.75$ for $N = 100$, $f = 0.5$ for $N = 1000$, and $f = 0.25$ for $N = 10000$.

References

Fukumizu, K. (2010). *Introduction to Kernel Methods*, (written in Japanese), Asakura Publishing, 8–9.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*, Second Edition, Springer, 436–437.

Schölkopf, B. and Smola, A. (1998). *Learning with Kernels*, MIT Press, 118–120.

Received February 25, 2015; Revised March 23, 2015; Accepted March 25, 2015