Journal of Machine Learning Research 9 (2008) 521-540 Submitted 3/07; Revised 9/07; Published 3/08

# Estimating the Confidence Interval for Prediction Errors of Support Vector Machine Classifiers

**Bo Jiang** BO-JIANG02@MAILS.TSINGHUA.EDU.CN
**Xuegong Zhang** ZHANGXG@TSINGHUA.EDU.CN
*MOE Key Laboratory of Bioinformatics & Bioinformatics Div., TNLIST and Department of Automation*
*Tsinghua University*
*Beijing 100084, China*

**Tianxi Cai** TCAI@HSPH.HARVARD.EDU
*Department of Biostatistics*
*Harvard University*
*Boston, MA 02115, U.S.A.*

**Editor:** Nicolas Vayatis

## Abstract

Support vector machine (SVM) is one of the most popular and promising classification algorithms. After a classification rule is constructed via the SVM, it is essential to evaluate its prediction accuracy. In this paper, we develop procedures for obtaining both point and interval estimators for the prediction error. Under mild regularity conditions, we derive the consistency and asymptotic normality of the prediction error estimators for SVM with finite-dimensional kernels. A perturbation-resampling procedure is proposed to obtain interval estimates for the prediction error in practice. With numerical studies on simulated data and a benchmark repository, we recommend the use of interval estimates centered at the cross-validated point estimates for the prediction error. Further applications of the proposed procedure in model evaluation and feature selection are illustrated with two examples.

**Keywords:** *k*-fold cross-validation, model evaluation, perturbation-resampling, prediction errors, support vector machine

## 1. Introduction

As a state-of-the-art machine learning algorithm in classifying high-dimensional data, support vector machines (SVMs) developed by Vapnik and his colleagues (1995, 1998) have gained popularity due to many attractive features. The SVM has been used frequently in practice for developing prediction rules. After a prediction rule is constructed, the common practice is to provide a point estimate of the corresponding accuracy without accounting for the sampling variability in the estimated accuracy of the prediction rule. However, to ensure the reproducibility of the reported results, it is crucial to account for such sampling variability and provide interval estimates for the accuracy measures, especially when the sample size is not large relative to the number of unknown model parameters.

Various methods have been available to estimate the prediction error of classifiers based on the cross-validation and bootstrap methods (Efron, 1986; Efron and Tibshirani, 1997; Fu et al., 2005; Molinaro et al., 2005; Shao, 1996; Varma and Simon, 2006). When the sample size is not

sufficiently large, point estimates may be inadequate for choosing the classifier with optimized parameters or features (Reunanen, 2003; Varma and Simon, 2006). For example, in Table 1, we summarize the accuracies of SVM classifiers with different kernels based on on two artificial data sets that are generated as in Section 4.3. It appears that the polynomial kernel outperforms the linear kernel for both data sets with higher accuracy. However, it is unclear whether the difference in the higher accuracy is due to randomness. Due to its high generalization ability, the linear kernel may be preferred unless it results in significantly lower accuracy. As such, the point estimates of the accuracy measures may not provide sufficient evidence for determining which type of kernel should be used.

To adequately assess the accuracy and draw valid conclusions, it is important to account for the sampling variability in the estimated prediction error. Some studies have suggested performing hypothesis testing by considering the variability in the cross-validated estimator (Dietterich, 1998). Bengio and Grandvalet (2004) and Nadeau and Bengio (2003) pointed out that there exists no universally unbiased estimator of the variance of $K$-fold cross-validated estimator that is based only on the results of the cross-validation experiments. Therefore, the estimation of uncertainty around the prediction error estimators remains a theoretical, as well as practical problem.

| Data Type | Sample Size | Linear Kernel Accuracy | Polynomial Kernel Accuracy |
|---|---|---|---|
| 1 | 100 | 94% | 95% |
| 2 | 100 | 92% | 96% |

Table 1: Kernel selection in SVM classifiers based on the cross-validation point estimates for the prediction error.

To assess the predictive performance of SVM derived from data with finite sample size, probabilistic bounds such as VC-based bounds (Vapnik, 1998) and stability-based bounds (Kearns and Ron, 1999; Bousquet and Elisseeff, 2002) have been proposed. However, those theoretical bounds are too conservative to give an accurate estimation. In particular, they do not account for the sampling variability inherent in different types of data. In statistical literature, the bootstrap resampling procedure (Efron, 1979) and its variants (Efron, 1987; Wu, 1986; Liu, 1988; Hall and Mammen, 1994) provide a general framework for ascertaining variances and constructing confidence intervals, but limited effort has been made to study the distributional properties of the estimated prediction error (Efron and Tibshirani, 1995, Section 5).

In this article, we develop procedures to approximate the distribution of the estimated accuracy measures for SVM classifiers and construct confidence intervals for the accuracy measures. The proposed method, which may be linked to the weighted bootstrap resampling (Hall and Mammen, 1994; Hall and Maesono, 2000) and the Bayesian bootstrap method (Rubin, 1981), directly builds on the perturbation-resampling procedure considered in Park and Wei (2003) and Cai et al. (2005). The accuracy measure we consider is the expected absolute difference between the true and predicted responses for future subjects. For SVMs with finite-dimensional kernels, we show that the accuracy measure can be consistently estimated via cross-validation procedures, and the resulting estimators are asymptotically normal. A practical perturbation-resampling procedure is proposed to approximate the sampling distribution of the prediction error. This inference procedure is valid

without having to specify the true association between the response and the predictors. This is particularly appealing when it is difficult, if not impossible, to identify the *true* model under which the data are generated. Numerical studies based on simulated data and a benchmark repository suggest that both the variance estimator and the interval estimator centered at the cross-validated point estimator perform well. The proposed procedure is further illustrated with applications in kernel selection and in the genotypic testing for drug resistance.

## 2. Estimating the Prediction Error of SVM Classifiers

In this section, we provide a brief review on the construction of SVM classifiers and introduce point estimators of the accuracy measure used for evaluating the performance of SVM classifiers.

### 2.1 Basic Notations and Construction of SVM Classifiers

The SVM classifier is derived based on the hinge loss function:

$$L(Y, f(\mathbf{X})) = [1 - Y f(\mathbf{X})]_+ = \begin{cases} 0 & , \quad Y f(\mathbf{X}) > 1 \\ 1 - Y f(\mathbf{X}) & , \quad Y f(\mathbf{X}) \le 1 \end{cases},$$

where $\mathbf{X}$ is the input vector and $Y \in \{-1, 1\}$ is the output label, and $f(\mathbf{X})$ is the prediction function. Here, we first consider the case when $f(\mathbf{X})$ is a linear function, $f(\mathbf{X};\theta) = \mathbf{w}'\mathbf{X} + b$ (we use $\mathbf{V}'$ to denote the transpose of the vector $\mathbf{V}$ hereafter), where $\theta = (\mathbf{w}', b)'$ is the adjustable parameter. Based on $f(\cdot)$, we predict $Y$ by the decision function $\hat{Y}(\mathbf{X},\theta) = \text{sign}\{f(\mathbf{X};\theta)\}$, where $\text{sign}(\cdot)$ denotes the sign of the function value.

To construct an optimal prediction rule, one may consider the prediction function $f(\mathbf{X};\theta)$ that minimizes the SVM risk function

$$Q(\theta) = E\{[1 - Y f(\mathbf{X};\theta)]_+\}.$$

To approximate the expected risk function $Q(\theta)$, one may consider its penalized empirical counterpart,

$$\hat{Q}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}[1 - Y_i f(\mathbf{X}_i;\theta)]_+ + \lambda_n \mathbf{w}'\mathbf{w}, \tag{1}$$

and obtain $\hat{\theta} = \text{argmin}_\theta \hat{Q}_n(\theta)$, where $\{(\mathbf{X}_i, Y_i); i = 1, \ldots, n\}$ are $n$ independent realizations of $(\mathbf{X}, Y)$, and $\lambda_n$ is the regularization parameter that controls the amount of penalty. Subsequently, the prediction of $Y$ may be made based on $f(\mathbf{X};\hat{\theta})$.

In practice, the minimizer $\hat{\theta}$ may be ascertained through quadratic programming techniques since the minimization of $\hat{Q}_n(\theta)$ is equivalent to the minimization of

$$\min_\alpha \{\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i Y_i (\mathbf{X}_i'\mathbf{X}_j) Y_j \alpha_j\}, \tag{2}$$

with linear constraints $0 \le \alpha_i \le C, i = 1, \ldots, n$ and $\sum_{i=1}^{n}\alpha_i Y_i = 0$, where $C = 1/(2\lambda_n n)$. Here, the constraint parameter $C = C(n)$ depends on the sample size $n$ and typically satisfies $nC(n) \to \infty$, or equivalently $\lambda_n \to 0$, under which requirement SVM classifiers are universally consistent (Steinwart, 2002).

Note that the only way in which the input vectors appear in the minimizing problem (2) is in the form of inner products, $\mathbf{X}_i'\mathbf{X}_j$. If the input vectors are mapped to a so called "feature space" $\mathcal{H}$ via a mapping denoted by $\Phi$, then the minimizing algorithm would only depend on the data through inner products in $\mathcal{H}$, that is, functions of the form $\Phi(\mathbf{X}_i)'\Phi(\mathbf{X}_j)$. Hence, if there is a *kernel function* $K(\cdot,\cdot)$ such that $K(\mathbf{X}_i,\mathbf{X}_j)=\Phi(\mathbf{X}_i)'\Phi(\mathbf{X}_j)$, one may carry out the minimization based on kernel function $K(\cdot,\cdot)$ only. For the simplest case when $K(\mathbf{X}_i,\mathbf{X}_j)=\mathbf{X}_i'\mathbf{X}_j$, we will refer to function $K(\cdot,\cdot)$ as the *linear kernel*. Other examples include the *polynomial kernel* $K(\mathbf{X}_i,\mathbf{X}_j)=(\gamma\mathbf{X}_i'\mathbf{X}_j+b)^d$, and the *RBF kernel* $K(\mathbf{X}_i,\mathbf{X}_j)=\exp\{-\|\mathbf{X}_i-\mathbf{X}_j\|^2/2\sigma^2\}$, with specified hyper-parameters $\gamma$, $b$, $d$ and $\sigma$.

### 2.2 Point Estimators for the Prediction Error

To evaluate how well the trained SVM performs on a future, independent subject $(\mathbf{X}_0,Y_0)$ from the same population of $(\mathbf{X},Y)$, we consider the absolute prediction error $D_0$:

$$D_0 = E\left|Y_0-\hat{Y}(\mathbf{X}_0,\hat{\theta})\right|, \tag{3}$$

where $\hat{\theta}$ is the solution to minimizing function (1), and $\hat{Y}(\mathbf{X},\theta)$ is the decision function introduced in Section 2.1. Note that $\hat{\theta}$ is a function of random variables $\{(\mathbf{X}_i,Y_i);i=1,\ldots,n\}$, and the expectation $E$ in (3) is with respect to $\{(\mathbf{X}_i,Y_i);i=1,\ldots,n\}$ and $(\mathbf{X}_0,Y_0)$. Thus, $D_0$ depends on sample size $n$ and is sometimes referred to as the generalization error (see Nadeau and Bengio, 2003). To estimate $D_0$, we first consider the training error, which is also called apparent or re-substitution error in statistical literature, $\hat{D}=\hat{D}(\hat{\theta})$, where

$$\hat{D}(\theta) = n^{-1}\sum_{i=1}^{n}\left|Y_i-\hat{Y}(\mathbf{X}_i,\theta)\right|. \tag{4}$$

When the sample size $n$ is small or moderate relative to the dimension of parameter $\theta$, training error $\hat{D}(\hat{\theta})$ tends to be biased downward as an estimate of $D_0$. One remedy to reduce such a bias is to use the cross-validation procedure. Here we consider the commonly used $K$-fold cross-validation. Specifically, we randomly split the data into $K$ disjoint subsets of about equal size and label them as $I_k, k=1,\ldots,K$. For each $k$, we use all observations which are not in $I_k$ to obtain an estimate $\hat{\theta}_{(-k)}$ for $\theta$ via (1), and then compute the prediction error estimate $\hat{D}_{(k)}(\theta)$ via (4) based on observations in $I_k$. Then, the cross-validated prediction error estimator for $D_0$ is

$$\hat{\mathcal{D}} = K^{-1}\sum_{k=1}^{K}\hat{D}_{(k)}(\hat{\theta}_{(-k)}). \tag{5}$$

We show in the next section that the cross-validation estimator $\hat{\mathcal{D}}$ is consistent for estimating the prediction error of SVM classifiers under certain conditions. However, as we have mentioned above, point estimates are not adequate in drawing valid conclusions, and we need to further study the distributional properties of the estimated prediction error.

## 3. Interval Estimators for the Prediction Error

In this section, we provide large sample properties of the estimated prediction error. In particular, we discuss the consistency and asymptotic normality of the estimators. Based on these theoretical results, we present a simple perturbation-resampling procedure to obtain interval estimates for the prediction error. In addition, we provide inference procedures for comparing two competing models.

### 3.1 Large Sample Properties of Point Estimators

Suppose that the parameter $\theta$ belongs to a compact set $\Theta$, and both the expectation $E(\mathbf{X})$ and the covariance matrix var$(\mathbf{X})$ of the input vector $\mathbf{X}$ are finite. To derive the asymptotic properties for $\hat{D}$, we first need to establish that $\hat{\theta}$ "stabilizes" as $n$ increases, that is, $\hat{\theta}$ converges to a constant vector in probability, as $n \to \infty$. In **Theorem 1** of Appendix A, we show that under some regularity conditions, the limiting objective function $Q(\theta)$ is strictly convex with a unique minimizer $\theta_0$, and thus for large $n$, there exists a unique minimizer, $\hat{\theta}$, of $\hat{Q}_n(\theta)$. Furthermore, as $n \to \infty$, $\hat{\theta} \to \theta_0$ and $\hat{D}(\hat{\theta}) \to D_0$ in probability.

To further study the large sample property of $\hat{D}$, we explore the distribution of

$$W = n^{1/2}\{\hat{D}(\hat{\theta}) - D_0\} \, . \tag{6}$$

Note that although $\hat{D}(\theta)$ is not differentiable with respect to $\theta$, $E[\hat{D}(\theta)]$ is continuously differentiable at $\theta_0$. In **Theorem 2** of Appendix B, we show that $W$ is asymptotically equivalent to $n^{-1/2}\sum_{i=1}^{n}\eta_i$, and converges in distribution to a zero mean normal with variance $E(\eta_i^2)$, where $\eta_i$ is defined in (14) of Appendix B. The variance of $W$ can be approximated by

$$n^{-1}\sum_{i=1}^{n}\hat{\eta}_i^2 \, , \tag{7}$$

where $\hat{\eta}_i$ is obtained by replacing all the theoretical quantities in $\eta_i$ by their empirical counterparts.

It is commonly known that the training error $\hat{D}$ is biased downward as an estimate of $D_0$ and hence should not be used without correction. To reduce such a bias, we consider the $K$-fold cross-validated estimator given in (5), where $K$ is fixed and relatively small with respect to $n$. Using similar arguments as for the convergence of $\hat{D}(\hat{\theta})$, one may show that $\hat{\mathcal{D}}$ converges to $D_0$ in probability. Furthermore, we show in **Theorem 3** of Appendix C that

$$\mathcal{W} = n^{1/2}\{\hat{\mathcal{D}} - D_0\} \tag{8}$$

is asymptotically equivalent to $W$ in (6) and thus $\mathcal{W}$ also converges in distribution to a zero mean normal with variance $E(\eta_i^2)$. This implies that the cross-validated estimator $\hat{\mathcal{D}}$, while potentially has less bias compared to the training error $\hat{D}$, is expected to have the same magnitude of variability as that of $\hat{D}$. Thus, we recommend to construct confidence intervals for $D_0$ by centering at $\hat{\mathcal{D}}$ with width determined by the variability in $W$. Although the proposed procedure is derived through large sample approximations, the results of numerical studies given below indicate that the distributions of $W$ and $\mathcal{W}$ are reasonably close in finite samples.

### 3.2 Perturbation-Resampling Procedure for Estimating the Confidence Interval

Estimating the variance of $\mathcal{W}$ based on (7) may be difficult in practice with high-dimensional $\theta$ since it requires the estimation of the gradient of an unknown non-parametric function. To overcome such difficulties, we propose a computationally efficient perturbation-resampling procedure to approximate the distribution of $\mathcal{W}$. To be specific, let $\{G_i; i = 1, \ldots, n\}$ be a vector of independent and identically distributed *positive* random variables with unit mean and unit variance that are generated independent of the data. In practice, one may generate $\{G_i; i = 1, \ldots, n\}$ from an exponential

distribution. For any given set of $\{G_i; i = 1, \ldots, n\}$, we define

$$\hat{Q}_n^*(\theta) = \frac{1}{n}\sum_{i=1}^{n} G_i\{[1 - Y_i f(\mathbf{X}_i;\theta)]_+ + \lambda_n \mathbf{w}'\mathbf{w}\}\ , \tag{9}$$

and let $\theta^*$ be the minimizer of $\hat{Q}_n^*(\theta)$. Note that conditionally on the observed data, the only random quantities in $\hat{Q}_n^*(\theta)$ are the $G$'s. Next, let

$$W^* = n^{-1/2}\sum_{i=1}^{n}\{|Y_i - \hat{Y}(\mathbf{X}_i, \theta^*)| - \hat{D}(\hat{\theta})\}G_i\ . \tag{10}$$

It follows from the arguments given in Appendix D that the distribution of $\mathcal{W}$ in (8) can be approximated well by the conditional distribution of $W^*$ in (10) given the data $\{(\mathbf{X}_i, Y_i); i = 1, \ldots, n\}$. The random variables $G_i$ used in (10) may be linked to the Bayesian bootstrap method (Rubin, 1981) with $G_i/(n^{-1}\sum_{i=1}^n G_i)$ being the weights instead.

To obtain $\theta^*$ numerically, one may solve the dual problem of (9),

$$\min_{\alpha}\{\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i Y_i(\mathbf{X}_i'\mathbf{X}_j)Y_j\alpha_j\}\ , \tag{11}$$

under the constraints $\sum_{i=1}^n \alpha_i Y_i = 0$ and $0 \le \alpha_i \le CG_i$ for $i = 1, \ldots, n$. The solution of $\mathbf{w}$ is given by $\mathbf{w}^* = (\sum_{i=1}^n Y_i\alpha_i\mathbf{X}_i)/(n^{-1}\sum_{i=1}^n G_i)$. Note that the only difference between (2) and (11) is that there is a random multiplier on the upper bound of $\alpha_i$ in (11). For each generated set of $\{G_i; i = 1, \ldots, n\}$, we compute the corresponding $W^*$ via (10). By repeatedly generating $\{G_i; i = 1, \ldots, n\}$, we may obtain a large number of realizations of $W^*$ which may be used to approximate the distribution of $\mathcal{W}$ and construct confidence intervals for $D_0$. For example, a $100(1-\alpha)\%$ confidence interval for $D_0$ may be obtained as

$$[\hat{\mathcal{D}} - n^{-1/2}\hat{\xi}_{1-\alpha/2}, \hat{\mathcal{D}} - n^{-1/2}\hat{\xi}_{\alpha/2}],$$

where $\hat{\xi}_\alpha$ is the $\alpha^{th}$ percentile of $W^*$. The integrated procedure of perturbation-resampling is given in **Algorithm 1**, where $N$ is the number of perturbations.

**Algorithm 1** Perturbation-Resampling Procedure

1: Given data $\{(\mathbf{X}_i, Y_i); i = 1, \ldots, n\}$, a classifier is trained based on the SVM algorithm
2: Estimate the cross-validation error of the classifier by using (5)
3: **for** $r = 1 \to N$ **do**
4: Generate independent positive random variables $\{G_i; i = 1, \ldots, n\}$ from an exponential distribution with unit mean and unit variance
5: Solve the quadratic programming problem (11), and calculate $W_r^*$ by using (10)
6: **end for**
7: Estimate the resampling distribution of $W^*$ based on $\{W_r^*; r = 1, \ldots, N\}$, which approximates the distribution of $W$ in (6), or asymptotically the distribution of $\mathcal{W}$ in (8)
8: Use the resampling distribution to estimate the confidence interval of the prediction error centered at the cross-validation error estimate
9: Statistical evaluation of different models can be further made based on the resampling distribution (see Section 3.3)

### 3.3 Comparing Models Based on Interval Estimates

Suppose there are two competing models, say, $f_j(\mathbf{X};\hat{\theta}_j), j = 1,2$, where the functions $f_1$ and $f_2$ could be different in the kernels or features used, and $\hat{\theta}_j$ is the solution via (1) with the function $f_j$ and the data $\{(\mathbf{X}_i, Y_i); i = 1,\ldots,n\}$. The theoretical and empirical prediction errors $D_{0j}$ and $\hat{D}_j(\theta_j)$ are defined by (3) and (4) accordingly, $j = 1,2$. We are interested in making inference about, for example, $\Delta = D_{02} - D_{01}$ to assess how much improvement Model 2 is over Model 1.

A consistent estimator for $\Delta$ is $\hat{\Delta} = \hat{D}_2(\hat{\theta}_2) - \hat{D}_1(\hat{\theta}_1)$. It follows from the argument presented in Section 3.1 that

$$W_\Delta = n^{1/2}\{\hat{\Delta} - \Delta\}$$

is asymptotically normal with mean zero. To approximate this normal distribution, one may use the perturbation-resampling technique discussed in Section 3.2. Specifically, let $\theta_j^*$ be the minimizer of $n^{-1}\sum_{i=1}^n G_i\{[1 - Y_i f_j(\mathbf{X}_i;\theta)]_+ + \lambda_n \mathbf{w}'\mathbf{w}\}$, $j = 1,2$. Also, let

$$W_j^* = n^{-1/2}\sum_{i=1}^n \{|Y_i - \hat{Y}_j(\mathbf{X}_i,\theta_j^*)| - \hat{D}_j(\hat{\theta}_j)\}G_i\ ,$$

where $\hat{Y}_j(\mathbf{X},\theta) = \text{sign}\{f_j(\mathbf{X};\theta)\}$. Then, the distribution of $W_\Delta$ can be approximated by the conditional distribution of $W_\Delta^* = W_2^* - W_1^*$. Confidence intervals for $\Delta$ can then be constructed.

Note that even if $\hat{\Delta}$ is a consistent estimator for the prediction gain $\Delta$, it represents the fitting gain of using Model 2 and may lead to a wrong comparison between models with a large probability. By applying the cross-validation procedure, the overfitted model is likely to have a larger prediction error and one would choose the more parsimonious model. Thus, the $K$-fold cross-validated estimator $\hat{\mathcal{D}}_2 - \hat{\mathcal{D}}_1$, where $\hat{\mathcal{D}}_j$ is defined by (5) for Model $j, j = 1,2$, may be less biased than $\hat{\Delta}$ particularly in non-asymptotic situations. Let $\mathcal{W}_j$ be defined by (8) based on Model $j$. Again, the resampling distribution of $\mathcal{W}_2 - \mathcal{W}_1$ can be asymptotically approximated by $W_\Delta^*$. Based on the results of our simulated experiments, this approximation performs quite well even with limited number of samples.

## 4. Numerical Studies and Examples

In this section, we examine the finite-sample performance of the proposed inference procedure via extensive numerical studies based on both simulated data and a benchmark repository. Furthermore, we illustrate the new procedure with examples in kernel and biomarker selections.

### 4.1 Simulation Studies

We first conduct simulation studies to examine how well the proposed inference procedure performs in finite samples. The data are generated as follows: (1) the response $Y$ is generated from $\{-1,1\}$ with equal probabilities; (2) given $Y$, the input vector $\mathbf{X}$ are generated from $d$-dimensional multivariate normal with mean $\mathbf{1}_{d\times 1}I(Y=1) + (-\mathbf{1})_{d\times 1}I(Y=-1)$, where $\mathbf{1}_{d\times 1}$ is a $d$-dimensional vector of ones. We consider sample sizes $n = 50$ and 100, and dimensions $d = 10, 20$, and 30. For each configuration, we generate $1,000$ independent data sets. For each simulated data set, SVM classifiers are trained by using the LIBSVM program (Chang and Lin, 2001) with a linear kernel. For simplicity, we set the penalty parameter $C$ equal to 1 here. We estimate the empirical absolute prediction

error via 5-fold cross-validation. The distribution of the empirical absolute prediction is obtained by using perturbation-resampling procedure with $1,000$ times of perturbations ($N = 1,000$ in **Algorithm 1**). Confidence interval with nominal level of 95% is then constructed based on empirical percentiles of the resampling distribution. To evaluate normal approximation and cross-validation procedures, we also construct confidence intervals based on normal assumption with both the estimated variance and the true variance calculated from the simulation parameters of the samples. For comparison, VC-based bounds (Vapnik, 1998) and stability-based bounds (Bousquet and Elisseeff, 2002) on the prediction error are also obtained with the same nominal level of 95%.

To evaluate these interval estimates, the true prediction errors of the trained SVM classifiers are calculated according to $10,000$ replications of simulated data sets for each setting. Confidence intervals are compared with the true prediction error, and their coverage accuracies are obtained by averaging on $1,000$ data sets. Coverage accuracy is defined as the frequency for true value to fall inside the estimated confidence interval, which measures the accuracy of interval estimates. In the ideal case, the coverage accuracy of an estimated interval should be equal or close to its level of confidence, and with its length as small as possible. In Table 2, we report the coverage accuracies and average lengths of 95% confidence intervals centered at 5-fold cross-validation errors for different procedures.

| Sample Size | Dimension | Empirical Percentiles[1] | | Normal Estimated[2] | | Normal True[3] | | VC Bound | Stability Bound |
|---|---|---|---|---|---|---|---|---|---|
| | | CA | AL | CA | AL | CA | AL | CA | CA |
| 50 | 10 | 94.7 | 0.20 | 93.9 | 0.19 | 94.8 | 0.20 | 100.0 | 100.0 |
| | 20 | 94.4 | 0.16 | 92.5 | 0.15 | 94.5 | 0.20 | 100.0 | 100.0 |
| | 30 | 93.8 | 0.12 | 90.4 | 0.14 | 94.2 | 0.17 | 100.0 | 100.0 |
| 100 | 10 | 95.1 | 0.15 | 94.8 | 0.14 | 95.2 | 0.16 | 100.0 | 100.0 |
| | 20 | 95.2 | 0.15 | 94.5 | 0.13 | 95.1 | 0.16 | 100.0 | 100.0 |
| | 30 | 94.6 | 0.12 | 93.2 | 0.12 | 95.1 | 0.15 | 100.0 | 100.0 |

Table 2: Coverage accuracies (CA) and average lengths (AL) of 95% confidence intervals obtained by using different procedures on simulated data.

As shown in Table 2, at sample size of $n = 100$, the empirical coverage levels for the 95% confidence intervals under normal approximation with the true variance range from 95.1% to 95.2%, which validates the accuracy of cross-validation and normal approximation. In practice, the true variance of the prediction error estimator is unknown and thus the perturbation-resampling procedure would be used to ascertain the variability of the estimator. From the results in Table 2, we can see that confidence intervals obtained by the empirical percentiles of the perturbed samples perform slightly better than those constructed via normal approximation with estimated variances, in a sense that intervals based on the empirical percentiles have larger coverage accuracies with comparable

1. Interval estimates are constructed by using empirical percentiles of the resampling distribution obtained by perturbation-resampling.
2. Interval estimates are constructed as $\hat{\mathcal{D}} \pm 1.96 n^{-1/2} \hat{\sigma}$ with $\hat{\sigma}^2$ being the conditional variance of $W^*$ estimated by perturbation-resampling.
3. Interval estimates are constructed as $\hat{\mathcal{D}} \pm 1.96 n^{-1/2} \sigma$ with $\sigma^2$ calculated as the true variance of $\mathcal{W}$.

lengths. Although the proposed algorithm may fail when the dimension of the unknown parameters is equal to or larger than the sample size, the simulation results indicate that the procedure derived through large sample approximations performs well even when sample size is moderate relative to the dimension of the parameters. On the contrary, we note that confidence bounds based on VC dimension or stability are too conservative with relative small number of samples in this example. Since these bounds are proposed to provide general guides on the construction of classifiers, they may not be suitable to account for the sampling variability from a specific population.

### 4.2 Variance Estimation on Benchmark Repository

We further validate the ability of the proposed procedure in estimating the variance of the cross-validation estimator on the benchmark repository used in Mika et al. (1999) and Chang and Lin (2001). The benchmark repository consists of 10 artificial and real-world data sets from the UCI, DELVE and STATLOG benchmark repositories. These data sets are collected from a variety of research areas ranging from oncology and disease diagnosis to molecular biology, astronomy, banking and signal processing. Each data set is randomly divided into 100 partitions with equal size (50 partitions for the *flare-solar*, *image* and *titanic* data sets).

To evaluate the variance estimator obtained by the perturbation-resampling procedure, we estimate the standard deviation of 5-fold cross-validation error based *only* on the *first* partition of each data set. We also obtain the 5-fold cross-validation estimates of the SVM classifier on the rest 99 partitions, and the results are used to calculate the sample standard deviation of the cross-validation estimator, which is regarded as the true value. For comparison, we estimate the standard deviation based on two other methods proposed by Nadeau and Bengio (2003) using the first partition of each data set. The first approach is performed by randomly splitting data into two distinct sets (we name it "splitting" method here), and the second approach is based on the approximation of a so-called statistic $\rho$ (we name it "$\rho$-based" method here). The description of the data sets, the standard deviations estimated by different methods, and their computational efficiencies are shown in Table 3. Computational time is tested on a PC with a Pentium 4 running at 2.8GHz and 512MB of RAM.

The results in Table 3 suggest that the perturbation-resampling based estimate of the standard deviation using *only* the first partition of each data set is rather close to the sample standard deviation estimated using the entire data set. To the contrary, the standard deviation estimated by splitting the data set tends to be biased upward, while the $\rho$-based method tends to underestimate the standard deviation of the cross-validation error. In the results shown above, $1,000$ times of randomly splitting or resampling are used in all the three methods, and as a result, the actual computational efficiencies of different methods are comparable. This study demonstrates that the proposed perturbation-resampling procedure can be an accurate and efficient way to estimate the variance of the cross-validation error.

### 4.3 Example in Kernel Selection

To illustrate the application of the proposed procedure in model comparison, we perform kernel selection for SVM classifiers on simulated data. Samples $\{(X_{1i}, X_{2i}); i = 1, \ldots, n\}$ are generated from a uniform distribution on two-dimensional area $[0,1] \times [0,1]$. For data type 1, two classes of samples are separated by the curve corresponding to a linear function, $X_1 + X_2 = 1$, with a few exceptions introduced as "noise". For data type 2, the separating curve corresponds to a cubic function $X_1^3 + X_2^3 = 1$. Intuitively, samples of data type 1 should be classified well by using the simple linear

| Data Set | Sample Size | Dimension | True[4] Sd | Resampling[5] Sd | Resampling[5] Time | Splitting[6] Sd | Splitting[6] Time | ρ-based[7] Sd | ρ-based[7] Time |
|---|---|---|---|---|---|---|---|---|---|
| banana | 53 | 2 | 0.089 | 0.090 | 1.8 | 0.095 | 2.2 | 0.029 | 2.1 |
| covtype | 200 | 54 | 0.058 | 0.060 | 29 | 0.044 | 36 | 0.016 | 29 |
| flare-solar | 21 | 9 | 0.120 | 0.124 | 1.3 | 0.113 | 1.2 | 0.036 | 1.2 |
| ijcnn1 | 283 | 22 | 0.022 | 0.023 | 21 | 0.024 | 25 | 0.009 | 23 |
| image | 46 | 18 | 0.081 | 0.089 | 4.0 | 0.123 | 4.2 | 0.037 | 3.9 |
| ringnorm | 74 | 20 | 0.065 | 0.067 | 17 | 0.107 | 16 | 0.030 | 16 |
| svmguide1 | 70 | 4 | 0.029 | 0.030 | 88 | 0.055 | 71 | 0.017 | 66 |
| titanic | 44 | 3 | 0.078 | 0.082 | 1.3 | 0.123 | 1.3 | 0.034 | 1.2 |
| twonorm | 74 | 20 | 0.027 | 0.026 | 3.9 | 0.084 | 3.9 | 0.021 | 3.8 |
| waveform | 50 | 21 | 0.046 | 0.049 | 3.1 | 0.130 | 3.4 | 0.033 | 3.4 |

Table 3: Estimating the standard deviation of the 5-fold cross-validation error using different methods (computational time is shown in seconds).

kernel, while the cubic polynomial kernel might perform better when classifying samples from data type 2. We generate each type of data with sample size $n$ equal to 100 and 200, respectively.

Polynomial kernels can be generalized as $K(\mathbf{X}_i, \mathbf{X}_j) = (\gamma \mathbf{X}_i' \mathbf{X}_j + b)^d$, where $\mathbf{X}_i$ and $\mathbf{X}_j$, $i, j = 1, \ldots, n$, are input vectors. In our study, we choose the hyper-parameters as $\gamma = 1/n$, $b = 0$, and $d = 3$. Then we apply the SVM algorithm by using the linear kernel and the polynomial kernel with optimal hyper-parameter $C$ chosen by a cross-validation procedure, respectively. To make inference about the performances of different kernels, we use the model comparison procedure introduced in Section 3.3 to obtain 95% confidence intervals for the difference in cross-validation errors when using different kernels. We also compute the true prediction errors, together with the exact confidence intervals on the difference between their cross-validated estimates, based on the prediction results of $1,000$ replications of simulated data sets for each setting. In Table 4, we report the 10-fold cross-validation errors by using linear and polynomial kernels, the 95% confidence intervals on the difference between errors, and their respective true values.

For the first type of data, although the polynomial kernel could potentially lead to slightly lower error rates compared to the linear kernel, 95% confidence intervals for error difference are quite tight around zero. This suggests that the classifiers obtained based on these two types of kernels have similar accuracies as we expect. On the other hand, for the second type of data, 95% confidence intervals for error differences tend to deviate downward from zero, which indicates that the polynomial kernel indeed performs better than the linear kernel. (At the significant level of 0.05, $n = 100$ is not sufficient to conclude this, whereas $n = 200$ allows to make the above statement.) These

4. The true standard deviation (Sd) is calculated based on 5-fold cross-validation errors estimated on the rest 99 partitions of the data.
5. Standard deviation (Sd) and computation time (Time) are obtained by applying perturbation-resampling method on the first partition of the data.
6. Standard deviation (Sd) and computation time (Time) are obtained by applying splitting method (Nadeau and Bengio, 2003) on the first partition of the data.
7. Standard deviation (Sd) and computation time (Time) are obtained by applying ρ-based method (Nadeau and Bengio, 2003) on the first partition of the data.

| Data Type | Sample Size | Linear Kernel Errors | | Polynomial Kernel Errors | | .95 Interval on Difference Between CV Errors | |
|---|---|---|---|---|---|---|---|
| | | CV[8] | True[9] | CV | True | Estimated | Exact[10] |
| 1 | 100 | 0.060 | 0.055 | 0.050 | 0.065 | [-0.111, 0.066] | [-0.050, 0.080] |
| 2 | 100 | 0.080 | 0.078 | 0.040 | 0.037 | [-0.120, 0.004] | [-0.130, 0.001] |
| 1 | 200 | 0.080 | 0.074 | 0.075 | 0.077 | [-0.041, 0.024] | [-0.035, 0.025] |
| 2 | 200 | 0.150 | 0.153 | 0.085 | 0.087 | [-0.106, -0.004] | [-0.105, -0.005] |

Table 4: Kernel selection based on the interval estimates of the difference in cross-validation errors.

conclusions are consistent with the intuitions behind the data generating procedure. In particular, the predicted results are consistent with the true values of both point and interval estimates obtained by simulating a large number of data sets. This study serves as an example to demonstrate how to use the proposed model comparison procedure to choose an appropriate kernel in constructing SVM classifiers.

### 4.4 Example in the Genotypic Testing for Drug Resistance

In this section, we give an example to show how the proposed procedure can be used in selecting important markers in the genotypic testing for HIV protease inhibitor (PI) resistance on the HIV RT and Protease Sequence Database (Rhee et al., 2003). First, we divide the sample set into two classes by labeling each protease sequence sample with 99 amino acids as either "resistant" or "susceptible", depending on whether the resistance factor of the sample exceeds a certain drug-specific cutoff value or not (Beerenwinkel et al., 2002). Then, we predict the resistance to seven FDA-approved PIs using 10 sites on the substrate binding cleft or its flap that are reported to cause resistance by reducing the binding affinity between the inhibitor and the mutant protease enzyme. Aside from these mutations, mutation information at site 90, denoted by $X_{(90)}$, on the protease sequence has been reported to either contribute to or directly confer *in vitro* and *in vivo* resistance to each of the seven approved PIs, but the mechanism by which these mutations cause PI resistance is still not known. It is interesting to assess the incremental value of $X_{(90)}$ in predicting HIV drug resistance. To this end, we compare the prediction errors for the models with and without $X_{(90)}$ and evaluate the incremental value of $X_{(90)}$ based on the reduction in the prediction error, denoted by $\Delta_{X_{(90)}}$. We obtain the point and interval estimates of $\Delta_{X_{(90)}}$ based on the model comparison method discussed in Section 3.3 with 10-fold cross-validation. In both cases, the hyper-parameter $C$ is chosen by using a cross-validation procedure, respectively.

The results in Table 5 show that the 95% confidence intervals for $\Delta_{X_{(90)}}$ are tight around zero for drugs APV, ATV, and LPV, which indicates that $X_{(90)}$ adds rather modest value, if any, on top of other variables, for predicting resistance to these drugs. On the other hand, by including information on $X_{(90)}$, the prediction of drug resistance to IDV and RTV can be significantly improved in a sense that

8. 10-fold cross-validation errors are computed.
9. The true errors are estimated based on the prediction results of $1,000$ replications of simulated data sets for each setting.
10. The exact confidence intervals are estimated based on the prediction results of $1,000$ replications of simulated data sets for each setting.

| Drug Name | Sample Size | Resistant Fraction | Without Site 90 Error | Without Site 90 .95 Interval | With Site 90 Error | With Site 90 .95 Interval | .95 Interval for Difference |
|---|---|---|---|---|---|---|---|
| APV | 577 | 38.1% | 0.149 | [0.130,0.202] | 0.147 | [0.129,0.198] | [-0.025,0.025] |
| ATV | 142 | 51.4% | 0.261 | [0.195,0.403] | 0.197 | [0.135,0.270] | [-0.110,0.022] |
| IDV | 579 | 50.6% | 0.123 | [0.108,0.163] | 0.081 | [0.067,0.133] | [-0.063,-0.006] |
| LPV | 253 | 74.7% | 0.119 | [0.090,0.236] | 0.115 | [0.078,0.167] | [-0.032,0.027] |
| NFV | 617 | 64.0% | 0.113 | [0.093,0.147] | 0.092 | [0.076,0.130] | [-0.050,0.001] |
| RTV | 510 | 50.2% | 0.098 | [0.069,0.123] | 0.057 | [0.039,0.090] | [-0.060,0.000] |
| SQV | 598 | 43.6% | 0.172 | [0.146,0.211] | 0.132 | [0.113,0.166] | [-0.080,0.002] |

Table 5: Interval estimates for the prediction errors and their difference in the genotypic testing for HIV drug resistance with or without mutation information at site 90 on the protease sequence.

the 95% confidence intervals for $\Delta_{X_{(90)}}$ tend to locate on the negative side of the zero point. These results are consistent with studies in literature (see Para et al., 2000; Shulman et al., 2002; Campo et al., 2003; Saah et al., 2003). Therefore, $X_{(90)}$ is an important marker for choosing antiretroviral drugs and therapies, and the roles played by $X_{(90)}$ in reducing the susceptibility of these two drugs need to be further studied.

## 5. Discussion

In this paper, we propose procedures for making inference about the prediction error of SVM classifiers based on cross-validated point estimators and their corresponding interval estimators. We establish large sample theory for the cross-validated estimators, and present a perturbation-resampling procedure to construct the confidence interval for prediction errors. The proposed interval estimates are obtained by approximating the spread of $\mathcal{W}$ with that of $W$. Alternatively, one may consider directly perturbing $\mathcal{W}$ to yield potentially better approximations. However, such a perturbation procedure may be computationally intensive since a $K$-fold cross-validation scheme has to be conducted for each realization of the resampling weights. Results from extensive simulation studies suggest that the proposed point and interval estimators perform well in finite samples. Furthermore, through numerical studies, we demonstrate that the interval estimates provide much more information about the true underlying prediction accuracy than the point estimates. Although it is unclear whether similar theoretical results hold for SVM classifier with the RBF kernel (see the discussion in Appendix B), the framework in this article is likely to be applicable to other inductive learning algorithms with different types of loss functions.

The proposed procedures also allow us to tackle the issue of model evaluation and selection by taking the uncertainty of estimators for the prediction error into account. We give several examples to illustrate some direct applications of the method, such as to provide confidence intervals around the estimated prediction error in kernel and biomarker selections. In addition to the examples outlined above, the proposed procedures may have other practical applications in model evaluation or variable selection.

## Acknowledgments

We thank the Editor and the reviewers for many helpful comments and invaluable suggestions. This work is supported in part by NSFC grants 60575014, 30625012, 60721003, the National Basic Research Program (2004CB518605) and Hi-tech Research and Development Program (2006AA02Z325) of China, and NIH grant R01 EB006195 of USA.

## Appendix A. Consistency of $\hat{\theta}$ and $\hat{D}$

In the following theorem, we will show that as $n \to \infty$, $\hat{\theta} \to \theta_0$ and the training error $\hat{D}(\hat{\theta})$ will converge to the absolute prediction error $D_0$ in probability. Without loss of generality, we assume that $g_0(\mathbf{X}) = P(Y = 1 \mid \mathbf{X})$ and the distribution function of $\mathbf{X}$ are continuously differentiable hereafter.

**Theorem 1** Let $\theta_0 = (\mathbf{w}_0', b_0)' = \text{argmin}_{\theta \in \Theta} Q(\theta)$, $\Omega$ be the input vector space, and

$$\Lambda(Y, \theta_1) = \{\mathbf{X} \in \Omega \mid [1 - Y(\mathbf{w}_0'\mathbf{X} + b_0)][1 - Y(\mathbf{w}_1'\mathbf{X} + b_1)] < 0\}$$

for $\theta_1 = (\mathbf{w}_1', b_1)'$. Furthermore, we assume the following regularity condition:

$$P(Y = 1, \mathbf{X} \in \Lambda(1, \theta_1)) + P(Y = -1, \mathbf{X} \in \Lambda(-1, \theta_1)) > 0 \tag{12}$$

for any $\theta_1 \neq \theta_0$. Then, as $n \to \infty$, $\hat{\theta} \to \theta_0$ and $\hat{D}(\hat{\theta}) \to D_0$ in probability.

**Proof**. In view of Theorem 2.1 of Newey and McFadden (1994, Section 2), we can establish the convergence of $\hat{\theta} \to \theta_0$ by showing that (*a*) $Q(\theta)$ has a unique minimizer $\theta_0$; and (*b*) $\hat{Q}_n(\theta)$ converges to $Q(\theta)$ in probability, uniformly in $\theta$.

For (*a*), we note that since $Q(\theta)$ is continuous with respect to $\theta$ and $\Theta$ is compact, it must have a minimum within $\Theta$. Furthermore, it is easy to verify that for any $a, b \in R$,

$$(a + b)_+ \leq a_+ + b_+ , \tag{13}$$

and a strict inequality holds if and only if $ab < 0$. As a result, under condition (12), $Q(\theta)$ is a strictly convex function at $\theta_0$, and thus has a unique minimizer $\theta_0$.

For (*b*), since $\hat{Q}_n(\theta)$ is also a convex function of $\theta$ because of (13), and $\hat{Q}_n(\theta)$ converges in probability to $Q(\theta)$ for each $\theta \in \Theta$, we have $\sup_{\theta \in \Theta} |\hat{Q}_n(\theta) - Q(\theta)|$ goes to zero in probability, a uniform convergence property for convex functions proved by Pollard (1991, Section 6). This concludes the proof for the convergence of $\hat{\theta}$ to $\theta_0$ in probability.

It remains to show the consistency of $\hat{D}(\hat{\theta})$ for $D_0$. Since $g_0(\mathbf{X})$ is continuously differentiable, $E|Y_0 - \hat{Y}(\mathbf{X}_0, \theta)|$ is continuously differentiable in $\theta$ with bounded derivatives. Moreover, since $0 \leq E|Y_0 - \hat{Y}(\mathbf{X}_0, \theta)| \leq 2$, it follows from a uniform law of large numbers (Pollard, 1990, Chapter 8) that $\sup_{\theta \in \Theta} |\hat{D}(\theta) - E|Y_0 - \hat{Y}(\mathbf{X}_0, \theta)|| \to 0$ in probability. This, coupled with the convergence of $\hat{\theta}$ to $\theta_0$, implies that $\hat{D}(\hat{\theta}) - D_0 \to 0$ in probability. ■

The regularity condition in (12) guarantees the existence and uniqueness of the minimizer to the objective function. This condition states that any deviation of the parameter $\theta$ from the minimizer $\theta_0$ will always result in the change of output labels of certain samples. Given the continuous differentiability of both $g_0(\mathbf{X}) = P(Y = 1 \mid \mathbf{X})$ and the distribution function of $\mathbf{X}$, the condition can be satisfied if the probability density function of the input vector $\mathbf{X}$ is not equal to zero in some neighboring area of the optimal separating hyperplane.

## Appendix B. Large Sample Distribution for $\hat{D}$

With the assumption that $g_0(\mathbf{X}) = P(Y = 1 \mid \mathbf{X})$ and the distribution function of $\mathbf{X}$ are continuously differentiable, we have $\nabla_{\theta=\theta_0} Q(\theta) = -E\{YI(Yf(\mathbf{X};\theta_0) < 1)(\mathbf{X}', 1)'\}$, which is also differentiable, almost everywhere $\theta \in \Theta$. Thus, $Q(\theta)$ is twice differentiable almost everywhere $\theta \in \Theta$. Let $\mathbf{H}$ be the Hessian matrix of $Q(\theta)$ at $\theta_0$, $d(\theta) = E\{\hat{D}(\theta)\}$ and $\dot{d}(\theta) = \nabla d(\theta)$, we prove the following theorem:

**Theorem 2** Under the regularity condition (12) in Theorem 1, the distribution of $W$ is asymptotically equivalent to $n^{-1/2}\sum_{i=1}^n \eta_i$ and converges to a zero mean normal with variance $E(\eta_i^2)$, where

$$\eta_i = |Y_i - \hat{Y}(\mathbf{X}_i, \theta_0)| - D_0 - \dot{d}(\theta_0)\mathbf{H}^{-1}\mathbf{M}_i(\theta_0) \,, \tag{14}$$

and $\mathbf{M}_i(\theta_0) = -Y_i I(Y_i f(\mathbf{X}_i;\theta_0) < 1)(\mathbf{X}_i', 1)' + 2\lambda_n(\mathbf{w}_0', 0)'$.

**Proof**. Under the regularity condition, the limiting objective function $Q(\theta)$ is strictly convex at $\theta_0$, and thus its Hessian matrix $\mathbf{H}$ at $\theta_0$ is positive definite. To derive the asymptotic distribution theory for $W$, we first show that

$$\sqrt{n}(\hat{\theta} - \theta_0) = -n^{-1/2}\mathbf{H}^{-1}\sum_{i=1}^n \mathbf{M}_i(\theta_0) + o_p(1) \,, \tag{15}$$

where $\mathbf{M}_i(\theta_0) = -Y_i I(Y_i f(\mathbf{X}_i;\theta_0) < 1)(\mathbf{X}_i', 1)' + 2\lambda_n(\mathbf{w}_0', 0)'$.

To this end, let $\mathbf{Z} = (\mathbf{X}', Y)'$, $\mathbf{t} = (\mathbf{w}_t', b_t)'$, and write

$$[1 - Yf(\mathbf{X};\theta_0 + \mathbf{t})]_+ - [1 - Yf(\mathbf{X};\theta_0)]_+ = \mathbf{B}(\mathbf{Z})'\mathbf{t} + R(\mathbf{Z}, \mathbf{t}) \,, \tag{16}$$

where $\mathbf{B}(\mathbf{Z}) = -YI(Yf(\mathbf{X};\theta_0) < 1)(\mathbf{X}', 1)'$, and

$$R(\mathbf{Z}, \mathbf{t}) \;=\; \{1 - Yf(\mathbf{X};\theta_0 + \mathbf{t})\}[I\{Yf(\mathbf{X};\theta_0 + \mathbf{t}) < 1\} - I\{Yf(\mathbf{X};\theta_0) < 1\}] \,.$$

Noting that $R(\mathbf{Z}, \mathbf{0}) = 0$, and that the distribution function of $\mathbf{X}$ and the conditional probability mass function of $Y$ given $\mathbf{X}$ are continuous differentiable, it is easy to verify that

$$ER(\mathbf{Z}, \mathbf{t}) = \frac{1}{2}\mathbf{t}'\mathbf{H}\mathbf{t} + o(\|\mathbf{t}\|^2) \text{ and } ER(\mathbf{Z}, \mathbf{t})^2 = O(\|\mathbf{t}\|^3) \,.$$

Furthermore, $E\mathbf{B}(\mathbf{Z})$ is just the first order derivative of $Q(\theta)$ at $\theta_0$, thus $E\mathbf{B}(\mathbf{Z}) = 0$. Let $\mathbf{Z}_i = (\mathbf{X}_i', Y_i)'$, $\mathbf{s} = (\mathbf{w}_s', b_s)'$, and $A_n(\mathbf{s}) = \sum_{i=1}^n \{[1 - Y_i f(\mathbf{X}_i;\theta_0 + \mathbf{s}/\sqrt{n})]_+ + \lambda_n(\mathbf{w}_0 + \mathbf{w}_s/\sqrt{n})'(\mathbf{w}_0 + \mathbf{w}_s/\sqrt{n}) - [1 - Y_i f(\mathbf{X}_i;\theta_0)]_+ - \lambda_n \mathbf{w}_0'\mathbf{w}_0\}$. $A_n(\mathbf{s})$ is convex with respect to $\mathbf{s}$ because of (13), and it is minimized by $\sqrt{n}(\hat{\theta}_n - \theta_0)$. Note first that $nER(\mathbf{Z}, \mathbf{s}/\sqrt{n}) = \frac{1}{2}\mathbf{s}'\mathbf{H}\mathbf{s} + r_{n,0}(\mathbf{s})$, where $r_{n,0}(\mathbf{s}) = o(\|\mathbf{s}\|^2) \to 0$ for fixed $\mathbf{s}$. Accordingly, using (16),

$$\begin{aligned} A_n(\mathbf{s}) \;&=\; \sum_{i=1}^n \{[\mathbf{B}(\mathbf{Z}_i) + 2\lambda_n(\mathbf{w}_0', 0)']'\mathbf{s}/\sqrt{n} + R(\mathbf{Z}_i, \mathbf{s}/\sqrt{n}) - ER(\mathbf{Z}_i, \mathbf{s}/\sqrt{n})\} \\ &\quad + nER(\mathbf{Z}, \mathbf{s}/\sqrt{n}) + \lambda_n \mathbf{w}_s'\mathbf{w}_s \\ &=\; U_n'\mathbf{s} + \frac{1}{2}\mathbf{s}'\mathbf{H}\mathbf{s} + r_{n,0}(\mathbf{s}) + r_{n,1}(\mathbf{s}) + r_{n,2}(\mathbf{s}) \,, \end{aligned}$$

where $U_n = n^{-1/2}\sum_{i=1}^n \{\mathbf{B}(\mathbf{Z}_i) + 2\lambda_n(\mathbf{w}_0', 0)'\} = n^{-1/2}\sum_{i=1}^n \mathbf{M}_i(\theta_0)$, $r_{n,2}(\mathbf{s}) = \lambda_n \mathbf{w}_s'\mathbf{w}_s$, and $r_{n,1}(\mathbf{s}) = \sum_{i=1}^n \{R(\mathbf{Z}_i, \mathbf{s}/\sqrt{n}) - ER(\mathbf{Z}_i, \mathbf{s}/\sqrt{n})\}$. Now $r_{n,1}(\mathbf{s})$ tends to be zero in probability for each $\mathbf{s}$, since its

mean is zero and its variance is $\sum_{i=1}^{n} \mathrm{var}\{R(\mathbf{Z}_i, \mathbf{s}/\sqrt{n})\} = o(\|\mathbf{s}\|^2)$. Moreover, $r_{n,2}(\mathbf{s})$ also tends to be zero in probability when $\lambda_n \to 0$. Since $A_n(\mathbf{s})$ is a convex function, $\mathbf{H}$ is positive definite, and the covariance matrix $\mathrm{var}(\mathbf{X})$ is finite, it follows from the Basic Corollary of Hjort and Pollard (1993) that (15) holds.

Secondly, we show that the class of functions indexed by $\theta$, $\Im = \{|Y - \hat{Y}(\mathbf{X}, \theta)| : \|\theta - \theta_0\| \le \delta\}$ is a Donsker class, where $\delta$ is a given positive number and $\hat{Y}(\mathbf{X}, \theta) = \mathrm{sign}(\mathbf{w}'\mathbf{X} + b)$. Both of the classes of function $\{1 + \hat{Y}(\mathbf{X}, \theta) : \|\theta - \theta_0\| \le \delta\}$ and $\{1 - \hat{Y}(\mathbf{X}, \theta) : \|\theta - \theta_0\| \le \delta\}$ are VC classes (van der Vaart and Wellner, 1996, Lemma 2.6.15), and hence are Donsker. Note that in this case $\Im = \{I(Y = -1)[1 + \hat{Y}(\mathbf{X}, \theta)] + I(Y = 1)[1 - \hat{Y}(\mathbf{X}, \theta)] : \|\theta - \theta_0\| \le \delta\}$, and therefore is a Donsker class. It follows that $n^{1/2}[\hat{D}(\theta) - d(\theta)]$, a process in $\theta$, converges weakly to a zero mean Gaussian process and thus is stochastic equicontinuous at $\theta_0$. This, coupled with (15), implies that $n^{1/2}\{\hat{D}(\hat{\theta}) - D_0\} = n^{1/2}[\hat{D}(\hat{\theta}) - \hat{D}(\theta_0)] + n^{1/2}[\hat{D}(\theta_0) - D_0]$ is asymptotically equivalent to

$$n^{1/2}\{\hat{D}(\theta_0) - D_0\} + \dot{d}(\theta_0) n^{1/2}(\hat{\theta} - \theta_0) \simeq n^{-1/2} \sum_{i=1}^{n} \eta_i \,,$$

where

$$\eta_i = |Y_i - \hat{Y}(\mathbf{X}_i, \theta_0)| - D_0 - \dot{d}(\theta_0)\mathbf{H}^{-1}\mathbf{M}_i(\theta_0) \,.$$

Here and in the sequel, we use the notation $a \simeq b$ to denote that $a = b + o_p(1)$. Thus, $W = n^{1/2}\{\hat{D}(\hat{\theta}) - D_0\}$ converges in distribution to a zero mean normal random variable. ■

Since the limiting variance of $W$ is $\sigma^2 = E(\eta_i^2)$ and $n^{-1}\sum_{i=1}^{n} \eta_i^2$ converges to $\sigma^2$ in probability (based on the law of large numbers), one may estimate $\sigma^2$ by $n^{-1}\sum_{i=1}^{n} \eta_i^2$. Furthermore, it is not difficult to show that $n^{-1}\sum_{i=1}^{n}(\eta_i^2 - \hat{\eta}_i^2) \to 0$ since we expect that all the empirical estimates of the theoretical quantities in $\eta_i$ are consistent. We note that although $n^{-1}\sum_{i=1}^{n} \hat{\eta}_i^2$ is a consistent estimator of $\sigma^2$, we approximate $\sigma^2$ based on the resampling method, not $n^{-1}\sum_{i=1}^{n} \hat{\eta}_i^2$.

For a more general case, when the prediction function $f(\mathbf{X})$ in (1) is not a linear function of the input vector $\mathbf{X}$, one can rewrite the prediction function in the form of $f(\mathbf{X}) = \mathbf{w}'\Phi(\mathbf{X}) + b$, where $\Phi$ is a mapping from the input vector space to a "feature space" $\mathcal{H}$. Given that the expectation $E\Phi(\mathbf{X})$, the covariance matrix $\mathrm{var}(\Phi(\mathbf{X}))$, and the VC dimension of $f(\mathbf{X})$ are all finite (for example, when $\Phi$ is a mapping corresponding to a polynomial kernel function), and coupled with the fact that a $\{0,1\}$-valued class of functions is a uniform Donsker class if and only if its VC dimension is finite (Dudley, 1999), one can prove all the results given above using similar arguments. Note that when it comes to construct SVM classifier using the RBF kernel, however, these conditions cannot be satisfied because of the infinite-dimensional feature space of the RBF kernel.

## Appendix C. Large Sample Property of $\hat{\mathcal{D}}$

In this appendix, we will show that the distribution of $\mathcal{W}$ is asymptotically equivalent to that of $W$ based on training error.

**Theorem 3** $\mathcal{W}$ in (8) is asymptotically equivalent to $W$ in (6).

**Proof**. For each partition $I_k$, $n^{-1/2}\{\hat{D}_{(k)}(\hat{\theta}_{(-k)}) - D_0\}$ is asymptotically equivalent to $n^{-1/2}K\sum_{i=1}^{n} I(\xi_i = k)\{|Y_i - \hat{Y}(\mathbf{X}_i, \hat{\theta}_{(-k)})| - D_0\}$, where $\{\xi_i; i = 1, \ldots, n\}$ are $n$ exchangeable discrete random

variables uniformly distributed over $\{1,\ldots,K\}$, independent of the data, which satisfy $\sum_{i=1}^{n} I(\xi_i = k) \simeq n/K, k = 1,\ldots,K$. Conditionally on $\{\xi_i; i = 1,\ldots,n\}$, it follows from similar arguments in Appendix B that

$$\hat{\theta}_{(-k)} - \theta_0 = -\frac{K}{n(K-1)}\mathbf{H}^{-1}\sum_{i=1}^{n} I(\xi_i \neq k)\mathbf{M}_i(\theta_0) + o_p(n^{-1/2}) .$$

Then using the same argument as given for $n^{1/2}\{\hat{D}(\hat{\theta}) - D_0\}$, one can show that $n^{1/2}\{\hat{D}_{(k)}(\hat{\theta}_{(-k)}) - D_0\}$ is asymptotically equivalent to

$$n^{1/2}\{\hat{D}_{(k)}(\theta_0) - D_0\} + \dot{d}(\theta_0)n^{1/2}(\hat{\theta}_{(-k)} - \theta_0) \simeq n^{1/2}\sum_{i=1}^{n}\eta_{ki} ,$$

where

$$\eta_{ki} = I(\xi_i = k)K\{|Y_i - \hat{Y}(\mathbf{X}_i,\theta_0)| - D_0\} + I(\xi_i \neq k)\frac{K}{K-1}\dot{d}(\theta_0)\mathbf{H}^{-1}\mathbf{M}_i(\theta_0) .$$

It follows that $n^{1/2}(\hat{\mathcal{D}} - D_0) \simeq n^{-1/2}\sum_{i=1}^{n}(\sum_{k=1}^{K} K^{-1}\eta_{ki})$. Since $\sum_{k=1}^{K} I(\xi_i = k) = 1$ and $\sum_{k=1}^{K} I(\xi_i \neq k) = K-1$, it is straightforward to show that

$$\begin{aligned} n^{-1/2}\sum_{i=1}^{n}(\sum_{k=1}^{K} K^{-1}\eta_{ki}) &= n^{-1/2}\sum_{i=1}^{n}\left\{|Y_i - \hat{Y}(\mathbf{X}_i,\theta_0)| - D_0 + \dot{d}(\theta_0)\mathbf{H}^{-1}\mathbf{M}_i(\theta_0)\right\} \\ &= n^{-1/2}\sum_{i=1}^{n}\eta_i . \end{aligned}$$

■

## Appendix D. Justification for the Perturbation-Resampling Procedure

Here, we give a brief justification for the perturbation-resampling approach presented in Section 3.2. For formal justification of the approach, please see similar but more rigorous derivations given in Park and Wei (2003) and Cai et al. (2005).

To justify the resampling method, we first note that it follows from the arguments in Appendix B that

$$\sqrt{n}(\hat{\theta} - \theta_0) = -n^{-1/2}\mathbf{H}^{-1}\sum_{i=1}^{n}\mathbf{M}_i(\theta_0) + o_p(1) ,$$

and that

$$\sqrt{n}(\theta^* - \theta_0) = -n^{-1/2}\mathbf{H}^{-1}\sum_{i=1}^{n} G_i\mathbf{M}_i(\theta_0) + o_p(1) ,$$

where $\mathbf{M}_i(\theta_0) = -Y_i I(Y_i f(\mathbf{X}_i;\theta_0) < 1)(\mathbf{X}_i', 1)' + 2\lambda_n(\mathbf{w}_0', 0)'$.

Then, consider the unconditional version of $W^*$. Let $D^*(\theta) = n^{-1}\sum_{i=1}^{n}\{|Y_i - \hat{Y}(\mathbf{X}_i,\theta)|G_i\}$, $\hat{D}(\theta) = n^{-1}\sum_{i=1}^{n}|Y_i - \hat{Y}(\mathbf{X}_i,\theta)|$, and $D_0 = E|Y_0 - \hat{Y}(\mathbf{X}_0,\hat{\theta})|$, where $(\mathbf{X}_0, Y_0)$ is an independent sample

from the same population of $(\mathbf{X}, Y)$, and the expectation $E$ is with respect to $\{(\mathbf{X}_i, Y_i); i = 1, \ldots, n\}$ and $(\mathbf{X}_0, Y_0)$. As $\hat{\theta}$ converges to $\theta_0$ and $\hat{D}(\hat{\theta})$ converges to $D_0$, it is straight forward to show that

$$\begin{aligned}
W^* &= n^{1/2}(D^*(\theta^*) - D^*(\theta_0)) - n^{1/2}(\hat{D}(\hat{\theta}) - \hat{D}(\theta_0)) \\
&\quad + n^{1/2}(D^*(\theta_0) - \hat{D}(\theta_0)) - n^{-1/2}\sum_{i=1}^{n} \hat{D}(\hat{\theta})(G_i - 1) \\
&\simeq \dot{d}(\theta_0)n^{1/2}(\theta^* - \theta_0) - \dot{d}(\theta_0)n^{1/2}(\hat{\theta} - \theta_0) \\
&\quad + n^{-1/2}\sum_{i=1}^{n} |Y_i - \hat{Y}(\mathbf{X}_i, \theta_0)|(G_i - 1) - n^{-1/2}\sum_{i=1}^{n} D_0(G_i - 1) \\
&\simeq n^{-1/2}\sum_{i=1}^{n} \{|Y_i - \hat{Y}(\mathbf{X}_i, \theta_0)| - D_0\}(G_i - 1) - \dot{d}(\theta_0)n^{-1/2}\mathbf{H}^{-1}\sum_{i=1}^{n} \mathbf{M}_i(\theta_0)(G_i - 1) \\
&= n^{-1/2}\sum_{i=1}^{n} \eta_i(G_i - 1)\ ,
\end{aligned}$$

where $\eta_i = |Y_i - \hat{Y}(\mathbf{X}_i, \theta_0)| - D_0 - \dot{d}(\theta_0)\mathbf{H}^{-1}\mathbf{M}_i(\theta_0)$.

Conditionally on the data, it follows from the Multiplier Central Limit Theorem (van der Vaart and Wellner, 1996, Chapter 2.9) that the conditional distribution of $W^*$ converges to a normal with mean 0 and variance $n^{-1}\sum_{i=1}^{n} \eta_i^2$, which are the same as the unconditional distribution of W (or its cross-validation counterpart $\mathcal{W}$). This implies that for $\varepsilon > 0$, there exists an $N_0$ such that when $n > N_0$, the probability, with respect to samples $\mathcal{S} = \{(\mathbf{X}_i, Y_i); i = 1, \ldots, n\}$ , of the event

$$\sup_{u \in \mathcal{R}} |P(W^* \leq u|\mathcal{S}) - P(W \leq u)| < \varepsilon\ ,$$

is at least $1 - \varepsilon$.

## References


L. Beerenwinkel, B. Schmidt, H. Walter, R. Kaiser, T. Lengauer, D. Hoffmann, K. Korn, and J. Selbig. Diversity and complexity of HIV-1 drug resistance: A bioinformatics approach to predicting phenotype from genotype. *Proceedings of the National Academy of Sciences U.S.A.*, 99:8271–8276, 2002.

Y. Bengio and Y. Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5:1089–1105, 2004.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

T. Cai, L. Tian, and L. J. Wei. Semiparametric Box-Cox power transformation models for censored survival observations. *Biometrika*, 92:619–632, 2005.

R. E. Campo, J. N. Moreno, G. Suarez, N. Miller, M. A. Kolber, D. J. Holder, M. Shivaprakash, D. M. DeAngelis, J. L. Wright, W. A. Schleif, E. A. Emini, and J. H. Condra. Efficacy of indinavir-ritonavir-based regimens in HIV-1-infected patients with prior protease inhibitor failures. *AIDS*, 17:1933–1939, 2003.

C. C. Chang and C. H. Lin. Libsvm—a library for support vector machine. *Available at http://www.csie.ntu.edu.tw/cjlin/libsvm/*, 2001.

T. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923, 1998.

R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge, U.K.: Cambridge University Press, 1999.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7:1–26, 1979.

B. Efron. How biased is the apparent error rate of a prediction rule. *Journal of the American Statistical Association*, 81:461–470, 1986.

B. Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82: 171–185, 1987.

B. Efron and R. Tibshirani. Cross-validation and the bootstrap: Estimating the error rate of a prediction rule. Technical report, Stanford University, 1995.

B. Efron and R. Tibshirani. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92:548–560, 1997.

W. J. Fu, R. J. Carroll, and S. Wang. Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics*, 21:1979–1986, 2005.

P. Hall and Y. Maesono. A weighted bootstrap approach to bootstrap iteration. *Journal of the Royal Statistical Society Series B*, 62:137–144, 2000.

P. Hall and E. Mammen. On general resampling algorithms and their performance in distribution estimation. *The Annals of Statistics*, 22:2011–2030, 1994.

N. L. Hjort and D. Pollard. Asymptotics for minimizers of convex processes. Technical report, Yale University, 1993.

M. Kearns and D. Ron. Algorithmic stability and sanity check bounds for leave-one-out cross-validation. *Neural Computation*, 11:1427–1453, 1999.

R. Y. Liu. Bootstrap procedures under some non-i.i.d. models. *The Annals of Statistics*, 16:1696–1708, 1988.

S. Mika, G. R*ä*tsch, J. Weston, B. Sch*ö*lkopf, and K. R. M*ü*ller. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX*, pages 41–48, 1999.

A. Molinaro, R. Simon, and R. Pfeiffer. Prediction error estimation: A comparison of resampling methods. *Bioinformatics*, 21:3301–3307, 2005.

C. Nadeau and Y. Bengio. Inference for the generalization error. *Machine Learning*, 52:239–281, 2003.

W. Newey and D. McFadden. Large sample estimation and hypothesis testing. In D. McFadden and R. Engler, editors, *Handbook of Econometrics IV*, pages 2113–2245. Amsterdam: North Holland, 1994.

M. F. Para, D. V. Glidden, R. Coombs, A. Collier, J. Condra, C. Craig, R. Bassett, R. Leavitt, S. Snyder, V. J. McAuliffe, and C. Boucher. Baseline human immunodeficiency virus type I phenotype, genotype, and RNA response after switching from long-term hard-capsule saquinavir to indinavir or soft-gel-capsule saquinavir in AIDS clinical trials group protocol 333. *Journal of Infectious Diseases*, 182:733–743, 2000.

Y. Park and L. J. Wei. Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika*, 90:717–723, 2003.

D. Pollard. *Empirical Process: Theory and Applications*. Hayward, CA: Institute of Mathematical Statistics, 1990.

D. Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7: 186–199, 1991.

J. Reunanen. Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, 3:1371–1382, 2003.

S. Y. Rhee, M. J. Gonzales, R. Kantor, B. J. Betts, J. Ravela, and R. W. Shafer. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Research*, 31: 298–303, 2003.

D. Rubin. The bayesian bootstrap. *The Annals of Statistics*, 9:130–134, 1981.

A. J. Saah, D. W. Haas, M. J. DiNubile, J. Chen, D. J. Holder, R. R. Rhodes, M. Shivaprakash, K. K. Bakshi, R. M. Danovich, D. J. Graham, and J. H. Condra. Treatment with indinavir, efavirenz, and adefovir after failure of nelfinavir therapy. *Journal of Infectious Diseases*, 187:1157–1162, 2003.

J. Shao. Bootstrap model selection. *Journal of the American Statistical Association*, 91:655–665, 1996.

N. Shulman, A. Zolopa, D. Havlir, A. Hsu, C. Renz, S. Boller, P. Jiang, R. Rode, J. Gallant, E. Race, D. J. Kempf, and E. Sun. Virtual inhibitory quotient predicts response to ritonavir boosting of indinavir-based therapy in human immunodeficiency virus-infected patients with ongoing viremia. *Antimicrobial Agents and Chemotherapy*, 146:3907–3916, 2002.

I. Steinwart. Support vector machines are universally consistent. *Journal of Complexity*, 18:768–779, 2002.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. New York: Springer-Verlag Inc., 1996.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. New York: Springer, 1995.

V. N. Vapnik. *Statistical Learning Theory*. New York: John Wiley and Sons Inc., 1998.

S. Varma and R. Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, 2006.

C. F. J. Wu. Jackknife, bootstrap, and other resampling methods in regression analysis (with discussion). *The Annals of Statistics*, 14:1261–1295, 1986.