On the Bootstrap Approach for Support Vector Machines and Related Kernel Based Methods *

Andreas Christmann[1] and Robert Hable[1]
[1] *University of Bayreuth, Department of Mathematics, Bayreuth, GERMANY*

Abstract

Kernel based methods and in particular support vector Machines (SVMs) based on a general loss function and on a general kernel play an important role in statistical machine learning for many reasons. Such general SVMs can be considered as certain Hilbert-space valued kernel based regularized M-estimators. If some weak assumptions are satisfied, such SVMs are solutions of a well-posed mathematical problem in Hadamard's sense (i.e., there exists a unique solution which continuously depends on the data), are universally consistent with good learning rates, are statistically stable with respect to many notions of statistical robustness, and are attractive from a computational point of view. Last but not least, such kernel based methods have demonstrated their good generalization properties in many large scale applications with an unknown high-dimensional dependency structure.
The question how to compute a good approximation of the finite sample distribution of such kernel based methods is not yet fully addressed in the literature. From an applied point of view, this might be considered as a serious gap, because knowledge of the finite sample distribution or the appropriateness of an asymptotic distribution is the basis to draw statistical decisions like confidence regions, prediction intervals, tolerance intervals or tests. Here we use the empirical bootstrap (Efron, 1979) and show that this approach provides a consistent estimator for the distribution of SVMs under some relatively mild conditions.

Keywords: support vector machine, SVM, kernel based methods, bootstrap, nonparametric statistics, consistency, robustness.

1 Introduction

Support vector machines and related kernel based methods can be considered as very successful methods from modern statistical machine learning theory. They have good statistical and numerical properties under weak assumptions and have demonstrated their often good generalization properties in many applications, see e.g. Vapnik (1995, 1998), Schölkopf and Smola (2002), and Steinwart and Christmann (2008). To our best knowledge, the original SVM approach by Boser *et al.* (1992) was derived from the generalized portrait algorithm invented earlier by Vapnik and Lerner (1963). Throughout the paper, the term SVM will be used in the broad sense, i.e. an SVM is defined by (2.1) and is based on a general convex loss function L and on a general kernel k.

SVMs based on many standard kernels as for example the classical Gaussian RBF kernel are nonparametric methods. The finite sample distribution of many nonparametric methods is unfortunately unknown because the distribution P from which the data were generated is usually completely unknown and because there are often only asymptotical results describing the consistency or the rate of convergence of such methods known so far. Furthermore, there is in general *no* uniform rate of convergence for such

*This research was partially supported by the Deutsche Forschungsgesellschaft (DFG), Grant CH 291/1.

nonparametric methods due to the famous no-free-lunch theorem, see Devroye (1982) and Devroye *et al.* (1996). Informally speaking, the no-free-lunch theorem states that, for sufficiently malign distributions, the average risk of any statistical (classification) method may tend arbitrarily slowly to zero. Theses facts are true for SVMs. SVMs are known to be universally consistent and fast rates of convergence are known for broad *subsets* of all probability distributions. The asymptotic normality of SVMs was recently shown by Hable (2012) under certain conditions. Of course, confidence intervals based on this asymptotic approximation are always symmetric.

Here, we apply a different approach to SVMs, namely Efron's empirical bootstrap. The goal of this paper is to show that empirical bootstrap approximations for SVMs are consistent under mild assumptions, provided these SVMs are based on a general convex and smooth loss function and on a general smooth kernel. More precisely, convergence in outer probability is shown. This result is useful to draw statistical decisions based on SVMs, e.g. confidence intervals, tolerance intervals and so on. Note, that confidence intervals based on this bootstrap approach can be asymmetric.

We mention that both the sequence of SVMs and the sequence of their corresponding risks are qualitatively robust under mild assumptions, see Christmann *et al.* (2013). Hence, Efron's bootstap approach is quite successful for SVMs from several aspects.

The rest of the paper has the following structure. Section 2 gives a brief introduction into SVMs. Section 3 gives the result.

2 Support Vector Machines

Current statistical applications are often characterized by a wealth of large and high-dimensional data sets with complicated but unknown dependency structures. In classification and in regression problems there is a variable of main interest, often called "output values" or "response", and a number of potential explanatory variables, which are often called "input values". These input values are used to model the observed output values or to predict future output values. The data set consists of n pairs (x_1, y_1), ..., (x_n, y_n), which will be assumed to be realizations of independent and identically distributed random pairs (X_i, Y_i) with values in some space $\mathcal{X} \times \mathcal{Y}$. We denote the joint distribution of (X_i, Y_i) by P. We are interested in minimizing the risk or to obtain a function $f : \mathcal{X} \to \mathcal{Y}$ such that $f(x)$ is a good predictor for the response y, if $X = x$ is observed. The prediction should be made in an automatic way. We refer to this process of determining a prediction method as "statistical machine learning", see e.g. Vapnik (1995, 1998); Schölkopf and Smola (2002); Cucker and Zhou (2007); Smale and Zhou (2007) for details. Here, by "good predictor" we mean that f minimizes the expected loss, i.e. the risk,

$$\mathcal{R}_{L,\mathrm{P}}(f) = \mathbb{E}_{\mathrm{P}} \left[L\left(X, Y, f(X)\right) \right],$$

where P denotes the unknown joint distribution of the random pair (X, Y) and $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, +\infty)$ is a fixed loss function. As a simple example, the least squares loss

$$L(X, Y, f(X)) = (Y - f(X))^2$$

yields the optimal predictor $f(x) = \mathbb{E}_{\mathrm{P}}(Y|X = x)$, $x \in \mathcal{X}$. Because P is unknown, we can neither compute nor minimize the risk $\mathcal{R}_{L,\mathrm{P}}(f)$ directly.

Support vector machines, see Vapnik and Lerner (1963), Boser *et al.* (1992), Vapnik (1995, 1998), provide a highly versatile framework to perform statistical machine learning in a wide variety of setups. The minimization of regularized empirical risks over reproducing kernel Hilbert spaces was already considered e.g. by Poggio and Girosi (1990). Given a (measurable real-valued) kernel

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

(of course, complex-valued kernels can also be used) we consider predictors $f \in H$, where H denotes the corresponding reproducing kernel Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}$. The space H includes, for example, all functions of the form

$$f(x) = \sum_{j=1}^{m} \alpha_j \, k(x, x_j)$$

where x_j are arbitrary elements in $\mathcal{X}$ and $\alpha_j \in \mathbb{R}$, $1 \leq j \leq m$. To avoid overfitting, a support vector machine $f_{L,\mathrm{P},\lambda}$ is usually defined as the solution of a regularized risk minimization problem. More precisely, given a loss function L and kernel k, a *support vector machine* is defined by

$$f_{L,\mathrm{P},\lambda} \;=\; \arg\inf_{f \in H} \; \mathbb{E}_{\mathrm{P}} L\left(X, Y, f(X)\right) \;+\; \lambda \, \|f\|_H^2 \,, \tag{2.1}$$

where $\lambda \in (0, \infty)$ is the regularization parameter. For a given data set $D = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, the corresponding estimated function is given by

$$f_{L,\mathrm{D}_n,\lambda} \;=\; \arg\inf_{f \in H} \; \frac{1}{n} \sum_{i=1}^{n} L\left(x_i, y_i, f(x_i)\right) \;+\; \lambda \, \|f\|_H^2 \,, \tag{2.2}$$

where D_n denotes the empirical distribution based on D. Note that the optimization problem (2.2) corresponds to (2.1) when using D_n instead of P.

Efficient algorithms to compute $\hat{f}_n := f_{L,\mathrm{D}_n,\lambda}$ exist for a number of convex loss functions and there are often good reasons to consider other convex loss functions than the classical least squares loss function. Examples are the hinge loss

$$L(X, Y, f(X)) = \max\{0, 1 - Y \cdot f(X)\}$$

for binary classification purposes or the ϵ-insensitive loss

$$L(X, Y, f(X)) = \max\{0, |Y - f(X)| - \epsilon\}$$

for regression purposes, where $\epsilon > 0$. As these loss functions are not differentiable, the logistic loss for classification

$$L(X, Y, f(X)) = \ln(1 + \exp(-Y \cdot f(X)))$$

and for regression

$$L(X, Y, f(X)) = -\ln(4e^{Y - f(X)}/(1 + e^{Y - f(X)})^2)$$

and Huber-type loss functions are also used in practice. These loss functions are with respect to the last argument convex and Lipschitz continuous and can be considered as smoothed versions of the hinge and the ϵ-insensitive loss functions.

An important component of statistical analyses concerns quantifying and incorporating uncertainty (e.g. sampling variability) in the reported estimates. For example, one may want to include confidence bounds along the individual predicted values $\hat{f}_n(x_i)$ obtained from (2.2). Unfortunately, the sampling distribution of the estimated function $\hat{f}_n$ is unknown. Recently, Hable (2012) derived the asymptotic distribution of SVMs under some mild conditions. Asymptotic confidence intervals based on those general results are always symmetric.

Here, we are interested in approximating the finite sample distribution of SVMs by Efron's empirical bootstrap approach (Efron, 1979), because confidence intervals based on this approach can be asymmetric. To fix ideas, consider an operator $S : \mathcal{M} \to \mathcal{W}$, where $\mathcal{M}$ is a set of probability measures on some metric space $\mathcal{Z}$ and let $\mathcal{W}$ denote a metric space. Many estimators can be included in this framework. Simple examples include the sample mean (with functional $S(\mathrm{P}) = \int Z \, d\mathrm{P}$) and M-estimators (with operator defined implicitly as the solution to the equation $\mathbb{E}_{\mathrm{P}} \Psi(Z, S(\mathrm{P})) = 0$). Let $\mathcal{B}(\mathcal{Z})$ be the Borel σ-algebra

on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and denote the set of all Borel probability measures on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ by $\mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$. Then, (2.1) specifies the *SVM operator*

$$S : \mathcal{M} \to H, \qquad S(\mathrm{P}) = f_{L,\mathrm{P},\lambda}\,, \tag{2.3}$$

where $\mathcal{M} \subset \mathcal{M}_1(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$. Following van der Vaart and Wellner (1996, p. 345), let

$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{Z_i}$$

be the empirical measure of an independent and identically distributed (i.i.d.) random sample $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n$, from a probability measure P. Given the sample values, let $\hat{Z}_1, \ldots, \hat{Z}_n$ be an i.i.d. random sample from $\mathbb{P}_n$. The *bootstrap empirical distribution* is the empirical measure

$$\hat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\hat{Z}_i}\,,$$

and the *bootstrap empirical process* $\hat{\mathbb{G}}_n$ is

$$\hat{\mathbb{G}}_n = \sqrt{n}\,(\hat{\mathbb{P}}_n - \mathbb{P}_n)\,.$$

Let $S_n(Z_1, \ldots, Z_n) = S(\mathbb{P}_n)$ be the corresponding estimator and denote the distribution of $S(\mathbb{P}_n)$ by

$$\mathscr{L}_n(S; \mathrm{P}) = \mathscr{L}(S(\mathbb{P}_n)).$$

If P was known to us, we could estimate this sampling distribution by drawing a large number of random samples from P and evaluating our estimator on them. The basic idea of Efron's bootstrap approach is to replace the unknown distribution P by the natural non-parametric estimator $\mathbb{P}_n$. In other words, we estimate the distribution of our estimator of interest by its sampling distribution when the data are generated by $\mathbb{P}_n$. In symbols, the empirical bootstrap proposes to use

$$\widehat{\mathscr{L}_n(S; \mathrm{P})} = \mathscr{L}_n(S; \mathbb{P}_n).$$

Since this distribution is generally unknown, in practice one uses Monte Carlo simulation to estimate it by repeatedly evaluating the estimator on random samples drawn with replacement from $\mathbb{P}_n$.

3 Consistency of Bootstrap SVMs

In this section it will be shown that the general functional delta-method for bootstrap in probability, see van der Vaart and Wellner (1996, Thm. 3.9.11, p. 378) is applicable for SVMs under appropriate conditions. I.e., the weak consistency of bootstrap estimators carries over to the Hadamard-differentiable SVM operator S given in (2.3) in the sense that the sequence of "conditional random laws" (given $(X_1, Y_1), (X_2, Y_2), \ldots$) of

$$\sqrt{n}\big(S(\hat{\mathbb{P}}_n) - S(\mathbb{P}_n)\big) = \mathscr{L}\big(\sqrt{n}(f_{L,\hat{\mathbb{P}}_n,\lambda} - f_{L,\mathbb{P}_n,\lambda})\big) \tag{3.4}$$

is asymptotically consistent in probability for estimating the laws of the random elements

$$\sqrt{n}\big(S(\mathbb{P}_n) - S(\mathrm{P})\big) = \mathscr{L}\big(\sqrt{n}(f_{L,\mathbb{P}_n,\lambda} - f_{L,\mathrm{P},\lambda})\big). \tag{3.5}$$

In other words, if n is large, the "random distribution"

$$\mathscr{L}(\sqrt{n}(f_{L,\hat{\mathbb{P}}_n,\lambda} - f_{L,\mathbb{P}_n,\lambda}))$$

based on bootstrapping an SVM can be considered as a valid approximation of the unknown distribution

$$\mathscr{L}(\sqrt{n}(f_{L,\mathbb{P}_n,\lambda} - f_{L,\mathrm{P},\lambda})).$$

Assumption 3.1 *Let $\mathcal{X} \subset \mathbb{R}^d$ be closed and bounded and let $\mathcal{Y} \subset \mathbb{R}$ be closed. Assume that $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the restriction of an m-times continuously differentiable kernel $\tilde{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that $m > d/2$ and $k \neq 0$. Let H be the RKHS of k and let P be a probability distribution on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X} \times \mathcal{Y}))$. Let $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex, P-square-integrable Nemitski loss function of order $p \in [1, \infty)$ such that the partial derivatives*

$$L'(x, y, t) := \frac{\partial L}{\partial t}(x, y, t) \qquad \text{and} \qquad L''(x, y, t) := \frac{\partial^2 L}{\partial^2 t}(x, y, t)$$

exist for every $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$. Assume that the maps

$$(x, y, t) \mapsto L'(x, y, t) \qquad \text{and} \qquad (x, y, t) \mapsto L''(x, y, t)$$

are continuous. Furthermore, assume that for every $a \in (0, \infty)$, there is a $b'_a \in L_2(\mathrm{P})$ and a constant $b''_a \in [0, \infty)$ such that, for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\sup_{t \in [-a,a]} |L'(x, y, t)| \leq b'_a(x, y) \qquad \text{and} \qquad \sup_{t \in [-a,a]} |L''(x, y, t)| \leq b'' .$$

The conditions on the kernel k in Assumption 3.1 are satisfied for many kernels used in practice, e.g., Gaussian RBF kernels, exponential kernels, polynomial kernels, and linear kernel, but also for Wendland's compactly supported RBF kernels $k_{d,\ell}$ based on certain univariate polynomials $p_{d,\ell}$ of degree $\lfloor d/2 \rfloor + 3\ell + 1$ for $\ell \in \mathbb{N}$ such that $\ell > d/4$, see Wendland (2005, Thm. 9.13).

The conditions on the loss function L in Assumption 3.1 are satisfied, e.g., for the logistic loss for classification or for regression. The popular non-smooth loss functions hinge, ε-insensitive, and pinball are not covered. However, Hable (2012, Remark 3.5) described an analytical method to approximate such non-smooth loss functions up to an arbitrarily good precision $\epsilon > 0$ by a convex P-square integrable Nemitski loss function of order $p \in [1, \infty)$, see Steinwart and Christmann (2008, Def. 2.16).

We can now state our result on the consistency of the bootstrap approach for SVMs.

Theorem 3.2 *Let Assumption 3.1 be satisfied. Let $\lambda \in (0, \infty)$. Then*

$$\sup_{h \in \mathrm{BL}_1(H)} \left| \mathbb{E}_M h\big(\sqrt{n}(f_{L,\hat{\mathbb{P}}_n,\lambda} - f_{L,\mathbb{P}_n,\lambda})\big) - \mathbb{E} h(S'_{\mathrm{P}}(\mathbb{G})) \right| \to 0, \tag{3.6}$$

$$\mathbb{E}_M h\big(\sqrt{n}(f_{L,\hat{\mathbb{P}}_n,\lambda} - f_{L,\mathbb{P}_n,\lambda})\big)^* - \mathbb{E}_M h\big(\sqrt{n}(f_{L,\hat{\mathbb{P}}_n,\lambda} - f_{L,\mathbb{P}_n,\lambda})\big)_* \to 0, \tag{3.7}$$

converge in outer probability, where $\mathbb{G}$ is a tight Borel-measurable Gaussian process, S'_{P} is a continuous linear operator with

$$S'_{\mathrm{P}}(\mathrm{Q}) = -K_{\mathrm{P}}^{-1}\big(\mathbb{E}_{\mathrm{Q}}\big(L'(X, Y, f_{L,\mathrm{P},\lambda}(X))\Phi(X)\big)\big), \quad \mathrm{Q} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X} \times \mathcal{Y})) \tag{3.8}$$

and

$$K_{\mathrm{P}} : H \to H, \quad K_{\mathrm{P}}(f) = 2\lambda f + \mathbb{E}_{\mathrm{P}}\big(L''(X, Y, f_{L,\mathrm{P},\lambda}(X))f(X)\Phi(X)\big) \tag{3.9}$$

is a continuous linear operator which is invertible.

For details on S'_{P}, K_{P}, and $\mathbb{G}$ we refer to Hable (2012) and Christmann and Hable (2013). Note that S'_{P} and K_{P} already appeared in the derivation of influence functions for SVMs, see Christmann and Steinwart (2004) and Christmann and Steinwart (2007). The proof of Theorem 3.2 is somewhat lengthy and given in Christmann and Hable (2013). The idea of the proof is to use a functional delta-method proven in van der Vaart and Wellner (1996, Thm. 3.6.2, Thm. 3.9.11) and some empirical process techniques, some results on SVMs given in Steinwart and Christmann (2008) and the recently shown result, that SVMs are based on a map ϕ which is Hadamard differentiable at P tangentially to some appropriate subspace, see Hable (2012). Note that λ is fixed in Theorem 3.2. The cases that λ depends on n or is even data dependent is beyond the scope of this paper.

References

Boser, B. E., Guyon, I., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Computational Learning Theory*, pages 144–152, Madison, WI. ACM.

Christmann, A. and Hable, R. (2013). On the consistency of the bootstrap approach for support vector machines and related kernel based methods. Submitted. Preprint available on `http://arxiv.org/abs/1301.6944`.

Christmann, A. and Steinwart, I. (2004). On robust properties of convex risk minimization methods for pattern recognition. *J. Mach. Learn. Res.*, **5**, 1007–1034.

Christmann, A. and Steinwart, I. (2007). Consistency and robustness of kernel based regression. *Bernoulli*, **13**, 799–819.

Christmann, A., Salibían-Barrera, M., and Aelst, S. V. (2013). Qualitative robustness of bootstrap approximations for kernel based methods. In C. Becker, R. Fried, and S. Kuhnt, editors, *Robustness and Complex Data Structures*. Springer, Heidelberg (to appear, Preprint available on `http://arxiv.org/abs/1111.1876`.

Cucker, F. and Zhou, D. (2007). *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, Cambridge.

Devroye, L. (1982). Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. *IEEE Trans. Pattern Anal. Mach. Intell.*, **4**, 154–157.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.

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

Hable, R. (2012). Asymptotic normality of support vector machine variants and other regularized kernel methods. *Journal of Multivariate Analysis*, **106**, 92–117.

Poggio, T. and Girosi, F. (1990). Networks for approximation and learning. *Proc. IEEE*, **78**, 1481–1497.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA.

Smale, S. and Zhou, D.-X. (2007). Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, **26**, 153–172.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer, New York.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer, New York.

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

Vapnik, V. N. (1998). *Statistical Learning Theory*. John Wiley & Sons, New York.

Vapnik, V. N. and Lerner, A. (1963). Pattern recognition using generalized portrait method. *Autom. Remote Control*, **24**, 774–780.

Wendland, H. (2005). *Scattered Data Approximation*. Cambridge University Press, Cambridge.