Journal of the Korean Statistical Society (2007), **36**: *3, pp 367–372*

WHEN CAN SUPPORT VECTOR MACHINE ACHIEVE FAST RATES OF CONVERGENCE?†

CHANGYI PARK[1]

ABSTRACT

Classification as a tool to extract information from data plays an important role in science and engineering. Among various classification methodologies, support vector machine has recently seen significant developments. The central problem this paper addresses is the accuracy of support vector machine. In particular, we are interested in the situations where fast rates of convergence to the Bayes risk can be achieved by support vector machine. Through learning examples, we illustrate that support vector machine may yield fast rates if the space spanned by an adopted kernel is sufficiently large.

AMS 2000 subject classifications. Primary 68Q32; Secondary 62G20.
Keywords. Classification, empirical process, hinge loss, statistical learning theory.

1. INTRODUCTION

Classification as a tool to extract information from data plays an important role in science and engineering. Among various classification methodologies, support vector machine (SVM), introduced by Cortes and Vapnik (1995), has recently seen significant developments. The central problem this paper addresses is the accuracy of SVM, obtained by minimizing a penalized objective function in binary classifications.

The classification literature based on machine learning is vast and what will be cited below is very brief. Indeed, due to the enormity of the literature, we will only cite those that bear a direct relevance to SVM. Zhang (2004) obtain the Bayes risk consistency for convex margin losses. Steinwart and Scovel (2007) and Blanchard *et al.* (2004) studied the convergence rates to the Bayes risk for SVM using

Received January 2007; accepted March 2007.
†This research was supported by a Korea Research Foundation Grant funded by the Korean government (MOEHRD, Basic Research Promotion Fund) (KRF-2005-070-C00020).
[1]Institute of Statistics, Korea University, Anam-dong, Sungbuk-gu, Seoul 136-701, Korea (e-mail: park463@korea.ac.kr)

Gaussian kernels. The difference of Steinwart and Scovel (2007) and Blanchard *et al.* (2004) lies in the penalty term in the objective function. Blanchard *et al.* (2004) adopt an L_1 penalty whereas Steinwart and Scovel (2007) uses an L_2 penalty. Bartlett *et al.* (2006) obtain rates of convergence for convex losses. Finally Park (2006) provides an unified theory for both convex and nonconvex losses.

To yield fast rates, the low noise assumption in Mammen and Tsybakov (1999) is commonly adopted as in Steinwart and Scovel (2007), Bartlett *et al.* (2006) and Park (2006). The rates in Steinwart and Scovel (2007) seem to be a bit faster than those obtained in Bartlett *et al.* (2006) and Park (2006). The reason may be the fact that Steinwart and Scovel (2007) impose an additional condition called the geometric noise assumption. In this paper, we study the generalization error rates for SVM through learning examples based on the results in Park (2006). The focus of this study is to identify the situations where SVM can achieve fast rates.

This paper is organized as follows. Section 2 sets out the notation and preliminaries. Section 3 discusses the situations where fast rates can be obtained through learning examples.

2. GENERALIZATION ERRORS

The basic components of classification involve the input space $\mathcal{X} \subset \mathbb{R}^d$, an output space $\mathcal{Y} = \{-1, +1\}$, a (measurable) decision function $f : \mathcal{X} \mapsto \mathbb{R}$, and a training sample $\{(X_i, Y_i)\}_{i=1}^n$, consisting of a random sample on the joint probability space $(\mathcal{X} \times \mathcal{Y}, \sigma(\mathcal{X}) \times 2^{\mathcal{Y}}, \mathbb{P}(\cdot, \cdot))$ with $\sigma(\mathcal{X})$ a σ-field on $\mathcal{X}$.

Classification is performed using the training sample to construct f such that the sign, $\text{sign}(f)$, the classifier, decides the class assignment of an input $x \in \mathcal{X}$. The performance is determined by the margin, $yf(x)$, where $(x, y) \in \mathcal{X} \times \mathcal{Y}$, with a correct classification being determined by $yf(x) > 0$. Consequently, the overall performance of a classifier is determined by the margin.

SVM minimizes the objective function $\sum_{i=1}^n V(Y_i f(X_i))$ over a class of candidate decision functions $f \in \mathcal{F}$, where $V(z) = [1 - z]_+ = \max\{0, 1 - z\}$ is the hinge loss and $\mathcal{F}$ is a class of functions, the parameter space. To prevent overfitting, a nonnegative penalty functional $J(f)$ is added to yield the constrained optimization problem of minimizing

$$l(f) = \sum_{i=1}^{n} V(Y_i f(X_i)) + \lambda J(f) \tag{2.1}$$

over $\mathcal{F}$ where $\lambda > 0$ is a penalization constant for the penalty functional J. Similar to other penalization procedures, see for example, Wahba (1990), λ controls the trade-off between the training error and the penalty. The minimizer of (2.1) with respect to $f \in \mathcal{F}$ yields an estimated decision function $\hat{f}$, and hence the classifier $\text{sign}(\hat{f})$.

In machine learning, the penalty functional is usually the inverse of the geometric margin. In particular, for the linear case, the geometric margin with respect to a linear decision function f is defined to be $2/\|w\|^2$, where $f(x) = \langle w, x\rangle + b$ is a hyperplane with $\langle \cdot, \cdot \rangle$ the usual inner product on $\mathbb{R}^d$ and $b \in \mathbb{R}$. In the nonlinear case, the geometric margin is $2/\|g\|_K^2 = 2/\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(x_i, x_j)$, where f has the representation $g(x)+b \equiv \sum_{i=1}^n \alpha_i K(x, x_i)+b$ and $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a proper kernel assumed to satisfy Mercer's condition (Mercer, 1909). This ensures that $\|g\|_K^2$ is a proper norm.

Minimizing (2.1) is the empirical version of minimizing $\mathbb{E}V(Yf(X))$. In particular, denote by f_V, the minimizer of $\mathbb{E}V(Yf(X))$. Then

$$e_V(f, f_V) = \mathbb{E}V(Yf(X)) - \mathbb{E}V(Yf_V(X)) \tag{2.2}$$

is known as the excess surrogate risk, where f_V is a measurable function.

Misclassification loss is defined by

$$L(z) = \frac{1}{2}(1 - \text{sign}(z)).$$

The Bayes classifier is defined as $\bar{f} = \text{sign}(f^*)$ where $f^*(x) = \mathbb{P}(Y = 1|X = x) - 1/2$ is the Bayes rule obtained by minimizing the generalization error, $\mathbb{E}L(Yf(X))$, over all measurable f. The excess risk is defined as

$$e(f, \bar{f}) = \mathbb{E}L(Yf(X)) - \mathbb{E}L(Y\bar{f}(X)) \geq 0. \tag{2.3}$$

Note that $\bar{f}$ can be taken as f_V for the hinge loss.

Finally, we introduce a complexity measure, called the L_2-metric entropy with bracketing, of a function class $\mathcal{F}$. Given any $\varepsilon > 0$, the set $\{(f_j^l, f_j^u)\}_{j=1}^N$ is called an ε-bracketing function of $\mathcal{F}$ if for any $f \in \mathcal{F}$, there is a j such that $f_j^l \leq f \leq f_j^u$ and $\|f_j^u - f_j^l\|_2 \leq \varepsilon$ for all $j = 1, \ldots, N$ where $\|\cdot\|_2$ is the L_2-norm. The L_2-metric entropy $H_B(\varepsilon, \mathcal{F})$ of $\mathcal{F}$ with bracketing, is defined as the logarithm of the cardinality of ε-bracketing function of $\mathcal{F}$ of the smallest size. For example, let $\mathcal{F}$ be a class of monotone functions $f : \mathbb{R} \to [0, 1]$. Then $H_B(\varepsilon, \mathcal{F}) \leq O(1/\varepsilon)$. Heuristically, we see that $[0, 1]$ can be covered by $C(1/\varepsilon)$ balls with radius ε, where C is a positive constant independent of ε. See van der Vaart and Wellner (1996) for the proof.

3. Fast Rates of Convergence

In this section, we study the convergence of the excess risk of $\hat{f}$ obtained from minimizing the objective function (2.1). Particularly, we are interested in those situations where fast rates can be achieved. It is believed that SVM may yield fast rates with the low noise assumption in place, if the space spanned by an adopted kernel is sufficiently large.

In Bartlett and Shawe-Taylor (1998), it is indicated that the rate for linear SVM is $n^{-1/2}$ in nonseparable cases and n^{-1} in separable cases. We illustrate that fast rates can be obtained in separable cases for SVM with a polynomial kernel. Note that the linear kernel is a special case of a polynomial kernel. For nonseparable cases, we also illustrate that fast rates can also be obtained by SVM if the adopted kernel K is sufficiently smooth that the spanned space $\mathcal{F}$ can approximate the true decision function closely.

Throughout this section, it is assumed that $\mathcal{X} = \{x \in \mathbb{R}^d : x_1^1 + \cdots + x_d^2 \leq 1\}$ is the unit ball in $\mathbb{R}^d$ for $d \geq 1$ and the underlying marginal distribution on $\mathcal{X}$ is uniform. C denotes a positive generic constant throughout. To apply the results in Corollary 4.3 of Park (2006), we check the conditions A3, A5, and A6. We will assume that A2 is met, *i.e.*, $\mathcal{F}$ is uniformly bounded.

3.1. A separable case

Suppose that the true decision function $f_t(x)$ is a polynomial of degree $p_t \geq 1$. The positive class label $Y = +1$ is assigned if $x_1 \geq 0$ and the negative class label $Y = -1$ is assigned otherwise for any $x \in \mathcal{X}$. Then the classification problem is separable.

Let $K(x, y) = (\langle x, y \rangle + 1)^p$ for $x, y \in \mathcal{X}$ be a polynomial kernel of order $p \geq p_t$. This kernel induces $\mathcal{F}$ consisting of all polynomials of order at most p. From (83) and (84) in Kolmogorov and Tikhomirov (1959), it follows that $H_B(\varepsilon, \mathcal{F}) = O((1/\varepsilon)^{d/p})$.

Since the classification problem is separable, the low noise assumption A6 is satisfied with $\alpha = +\infty$. With the choice of $f_n = nf_t$, $e_V(f_n, \bar{f}) = O(n^{-1})$, implying A3 with $s_V = 1$. By A5, we have $\varepsilon_n = n^{-p/(2p+d)}$ when $(n\lambda J_n)^{-1} \sim n^{-d/(2p+d)}$. By Corollary 4.3 in Park (2006), we have

$$e(\hat{f}, \bar{f}) = O\left(n^{-\frac{2p}{2p+d}} \log\left(\frac{1}{\delta}\right)\right)$$

except for a set of probability less than some small $\delta > 0$ and $\mathbb{E}e(\hat{f}, \bar{f}) =$

$O(n^{-2p/(2p+d)})$. Hence the rate is faster than $n^{-1/2}$ for $p \geq d/2$.

3.2. A nonseparable case

Let us consider a nonseparable classification problem with a mixture distribution as the underlying distribution. Assume that the underlying joint distribution $\mathbb{P}(\cdot,\cdot)$ of (X,Y) is the mixture distribution of two normal distributions with mean vector μ_i; $i = 1,2$ and common covariance matrix I, where $\mu_1 = (+1,0,\ldots,0)'$ and $\mu_2 = (-1,0,\ldots,0)'$. Suppose that the priors for classes are equal. By Bayes' Theorem, $p^*(x) = (1+\exp(-2x_1))^{-1}$. Denote the true decision function as $f_t(x) = x_1$.

Consider the Gaussian kernel

$$K(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right),$$

where $\sigma > 0$. Let $\mathcal{F}$ be the space of functions induced by this kernel. The metric entropy of $\mathcal{F}$ in sup-norm is given by $H_\infty(\varepsilon,\mathcal{F}) = O((\log(1/\varepsilon))^{d+1})$ by (4.8) of Zhou (2002). It is easy to show that $H_B(\varepsilon,\mathcal{F}) = O((\log(1/\varepsilon))^{d+1})$.

For any sufficiently small $\delta > 0$,

$$\begin{aligned}\mathbb{P}(x \in \mathcal{X} : |f^*(x)| \leq \delta) &\leq \mathbb{P}\left(x \in \mathcal{X} : |x_1 - x_1^*| \leq 2\ln\left(\frac{1+2\delta}{1-2\delta}\right)\right)\\ &\leq \mathbb{P}\left(x \in \mathcal{X} : |x_1 - x_1^*| \leq C\delta\right)\\ &= O(\delta)\end{aligned}$$

using Taylor series expansion. Hence A6 is satisfied with $\alpha = 1$. We can take a sequence of bounded functions $\{\tanh(nf_t)\}$ in $C^\infty(\mathcal{X})$ converging to $\bar{f}$ in sup-norm. Hence A3 is satisfied with some $0 < s_V \leq 1$ because $C^\infty(\mathcal{X})$ is a subset of the space spanned by $\mathcal{F}$. From the metric entropy equation in A5, $\varepsilon_n = n^{-1/3}(\log n)^{(d+1)/3}$. By Corollary 4.3 in Park (2006), we have $e(\hat{f},\bar{f}) = O\left(n^{-2/3}(\log n)^{2(d+1)/3}\log(1/\delta)\right)$ except for a set whose probability tends to zero and $\mathbb{E}e(\hat{f},\bar{f}) = O\left(n^{-2/3}(\log n)^{2(d+1)/3}\right)$. Due to the approximation error, the rate is at best $n^{-2/3}(\log n)^{2(d+1)/3}$. Note that, in nonseparable cases with $\alpha = +\infty$, the best possible rate is n^{-1} up to some factor of $\log n$ under the condition that the approximation error rate n^{-s_V} does not impede the estimation error rate.

From the examples, one can see that SVM may yield faster rates than $n^{-1/2}$ under the low noise assumption. To be more precise, if the function space spanned by a specific kernel is sufficiently large so that the function space can approximate

the true decision function closely, then SVM may be able to yield fast rates. In this sense, the choice of an appropriate kernel for SVM is important in optimizing the predictive performance of SVM. If the spanned space by the chosen kernel is not sufficiently large, then SVM may not achieve its best predictive performance. On the other hand, if the spanned space is too large, then the convergence can be slowed down due to the increased complexity of the function space over which the optimization of (2.1) is carried out.

REFERENCES

BARTLETT, P. AND SHAWE-TAYLOR, J. (1998). "Generalization performance of support vector machines and other pattern classifiers", In *Advances in Kernel Methods: Support Vector Learning* (Schölkopf, B., Burges, C. J. C. and Smola, A. J., eds.), 43–54, MIT Press, Cambridge, USA.

BARTLETT, P. L., JORDAN, M. I. AND MCAULIFFE, J. D. (2006). "Convexity, classification and risk bounds", *Journal of the American Statistical Association*, **101**, 138–156.

BLANCHARD, G., BOUSQUET, O. AND MASSART, P. (2004). "Statistical performance of support vector machines", preprint.

CORTES, C. AND VAPNIK, V. (1995). "Support-vector networks", *Machine Learning*, **20**, 273–297.

KOLMOGOROV, A. N. AND TIKHOMIROV, V. M. (1959). "ε-entropy and ε-capacity of sets in a functional spaces", *Uspekhi Mat. Nauk*, **14**, 3–86. In Russian. English Translations in *American Society Translations*, **17**, 277–364 (1961).

MAMMEN, E. AND TSYBAKOV, A. B. (1999). "Smooth discrimination analysis", *The Annals of Statistics*, **27**, 1808–1829.

MERCER, J. (1909). "Functions of positive and negative type, and their connection with the theory of integral equations", *Philosophical Transactions of the Royal Society of London*, Ser. A, **209**, 415–446.

PARK, C. (2006). "Convergence rates of generalization errors for margin-based classification", preprint.

STEINWART, I. AND SCOVEL, C. (2007). "Fast rates for support vector machines using Gaussian kernels", *The Annals of Statistics*, **35**, 575–607.

VAN DER VAART, A. W. AND WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: with Applications to Statistics*, Springer-Verlag, New York.

WAHBA, G. (1990). *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics, Philadelphia.

ZHANG, T. (2004). "Statistical behavior and consistency of classification methods based on convex risk minimization", *The Annals of Statistics*, **32**, 56–85.

ZHOU, D.-X. (2002). "The covering number in learning theory", *Journal of Complexity*, **18**, 739–767.