

A Statistical View of the Support Vector Machine

Yi Lin

Department of Statistics
University of Wisconsin, Madison
yilin@stat.wisc.edu

Abstract

We establish the relationship between the support vector machine and the Bayes rule. The Bayes rule is the optimal classification rule if the underlying distribution of the data is known. Therefore the Bayes rule is not directly available in practice, but can be used as an ideal benchmark of any classification procedure. We show that the support vector machine approaches the Bayes rule in an asymptotic sense. The results are established under very mild condition, allowing arbitrary number of discontinuity in the underlying conditional probability function. This is in contrast with most other asymptotic results in the statistical literature, where the underlying conditional probability functions are assumed to be smooth to a given order. The results clarify the mechanism beneath the support vector machine, and highlight the advantage and limitation of the support vector machine methodology.

1 Introduction

Consider a training set of n iid observations from an (unknown) distribution $P(x, y)$. For each observation i , $i = 1, 2, \dots, n$, we observe an explanatory vector $x_i \in R^d$, and a class label y_i (1 or -1). A classification rule is a mapping from R^d to $\{-1, 1\}$ used to assign class labels to new subjects.

If the underlying distribution $P(x, y)$ is known, the optimal classification rule with respect to a given loss function can be derived, and is usually called the Bayes rule with respect to the given loss function. In practice, however, we need to get the necessary information from the training samples to approximate the Bayes rule. The most commonly used criterion for classification is the misclassification rate. In this paper we restrict our attention to this measure. Let $p_0(x) = Pr\{Y = 1|X = x\}$, where (X, Y) is a generic pair of random variables with distribution $P(x, y)$. The Bayes rule for minimizing the expected misclassification rate is

$$\phi^*(x) = \text{sign}[p_0(x) - 1/2] \quad (1)$$

2 Support vector machines with reproducing kernel

The support vector machine methodology was introduced in Boser, Guyon, and Vapnik (1992). See also Cortes and Vapnik (1995), Vapnik (1995). The linear SVMs are motivated by the geometric interpretation of maximizing the margin, and the nonlinear SVMs are characterized by the use of reproducing kernels.

It has been shown that the SVM with reproducing kernel K is equivalent to a regularization problem in the reproducing kernel Hilbert space \mathcal{H}_K . See Wahba (1999), Poggio and Girosi (1998). The SVM with reproducing kernel K finds the minimizer \hat{f}_λ of

$$\frac{1}{n} \sum_{i=1}^n [1 - y_i f(x_i)]_+ + \lambda \|h\|_{\mathcal{H}_K}^2 \quad (2)$$

over all functions of the form $f(x) = h(x) + b$, and $h \in \mathcal{H}_K$. Here

$$(\tau)_+ = \begin{cases} \tau & \text{if } \tau > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Once the minimizer \hat{f} is found, the SVM classifies according to the sign of \hat{f} .

An equivalent setup of the support vector machine is to find the minimizer \hat{f}_M of

$$\frac{1}{n} \sum_{i=1}^n [1 - y_i f(x_i)]_+ \quad (3)$$

over all functions of the form $f(x) = h(x) + b$, $h \in \mathcal{H}_K$, and $\|h\|_{\mathcal{H}_K}^2 \leq M$.

From these we can see that the SVM methodology is a regularization method with a particular loss function $[1 - yf(x)]_+$. The regularization methods have been studied extensively in the statistics literature. See Wahba (1990) and the reference therein. Cox and O'Sullivan (1990) provided a general framework for studying regularization methods. The method of regularization has two components: a data fit functional component and a regularization penalty component. The data fit functional component ensures that the estimate should fit the training data well, whereas the regularization penalty component is used to guard against overfitting. The data fit component usually approaches a limiting functional as $n \rightarrow \infty$. The target function of the regularization method is the minimizer of this limiting functional.

3 Support vector machines and the Bayes rule

Several authors have studied the generalization performance of SVMs, See Vapnik (1995), and Shawe-Taylor and Cristianini (1998). These authors established bounds on generalization error based on VC dimension, fat shattering dimension, and the proportion of the training data achieving certain margin. However, SVMs often have very large, even infinite, VC dimension or fat shattering dimension. Hence the bounds established are often very loose, and do not provide a satisfactory explanation as to why SVMs often have good generalization performance.

The connection between the SVM classification rule and the Bayes rule is of special interest to us. The following lemma provides some insight to this question.

Lemma 3.1 [*Lin (1999)*] *The minimizer of $E[(1 - Yf(X))_+]$ is $\text{sign}[p_0(x) - 1/2]$.*

As a regularization method, the limiting functional of the data fit component of the SVM is $E[(1 - Yf(X))_+]$. Lemma 3.1 thus identifies the target function of the SVM as $\text{sign}[p_0(x) - 1/2]$. Notice this is exactly the Bayes rule. Lemma (3.1) suggests that the SVM with kernel tries to implement the Bayes rule by approximating the function $\text{sign}[p_0(x) - 1/2]$ directly. If the RKHS is rich enough to contain functions that approximate $\text{sign}[p_0(x) - 1/2]$, the solution to the SVM problem (2) approaches the function $\text{sign}[p_0(x) - 1/2]$ as $n \rightarrow \infty$ when the smoothing parameter is chosen appropriately.

A variety of reproducing kernels have been used successfully in practical applications, including polynomial kernels, Gaussian kernels, and spline kernels. Roughly speaking, the reproducing kernel Hilbert space induced by the spline kernel of order m contains all the functions that have m th order derivatives. The requirement that the RKHS is rich enough to contain functions that approximate the sign function is important for our result. For example, our result does not cover the linear SVM.

Now we make the connection between the SVM and the Bayes rule more precise in the case of a simple reproducing kernel: spline kernel with $m = 1$ and $d = 1$. This simple reproducing kernel under consideration facilitates the proofs. However, in principle the same line of argument can be applied to the SVM with other commonly used reproducing kernels. We show that under very general conditions without any smoothness assumption, the solution to (2) converges to $\text{sign}[p_0(x) - 1/2]$. We further show that under very mild boundary conditions on $p_0(x)$, the generalization error rate of the SVM converges to the Bayes rate at a certain rate. These conditions are much weaker than the usual smoothness conditions imposed in regression and density estimation, and can easily be satisfied by nonsmooth, even discontinuous functions. For more details on the results, see Lin (2000).

Assumption 3.1 *The density $d(x)$ of X is supported on $[-1, 1]$, and it is bounded away from zero and infinity in this interval. That is, there exists constants $D_2 > D_1 > 0$, such that $D_1 \leq d(x) \leq D_2$ for all $x \in [-1, 1]$.*

Theorem 3.1 [*Lin (2000)*] *Under Assumption 3.1, suppose $p_0(x)$ is bounded away from $1/2$ from below (or above) by some positive constant D_3 in an interval $[x_0 - \delta, x_0 + \delta]$. Then there exists a positive number Λ depending only on D_3 and δ , such that for any fixed $\lambda < \Lambda$, or for any fixed sequence $\lambda_{(n)}$ going to zero, we have*

$$|\text{sign}[p_0(x_0) - 1/2] - \hat{f}_\lambda(x_0)| = O_p(n^{-1/3}\lambda^{-2/3}).$$

Before we can state our second result, we need to characterize the behavior of $p_0(x)$ at its cross points with $1/2$. We say a point r is a positive cross point if there exists a positive number $a > 0$, such that $p_0(x) > 0$ in $(r, r + a]$ and $p_0(x) < 0$ in $[r - a, r)$. Negative cross points are defined likewise.

Assumption 3.2 *The function $p_0(x)$ crosses $1/2$ finite many (k) times, and there exists $\zeta > 0$ and $D_4 > 0$ such that for any cross point r_j , $j = 1, 2, \dots, k$, there exists $\alpha_j \geq 0$, and $D_{6j} > D_{5j} > 0$, satisfying*

$$D_{5j}|x - r_j|^{\alpha_j} \leq |p_0(x) - 1/2| \leq D_{6j}|x - r_j|^{\alpha_j}, \quad \forall x \in (r_j - \zeta, r_j + \zeta), \quad (4)$$

and $p_0(x)$ is bounded away from $1/2$ by D_4 when x is more than ζ away from all the cross points. Denote $\max_j D_{6j} = \bar{D}$, $\min_j D_{5j} = \underline{D}$, $\max_j \alpha_j = \bar{\alpha}$, and $\min_j \alpha_j = \underline{\alpha}$.

We will consider the setup (3) and its solution \hat{f}_M in our second result for technical convenience. For any $\theta > 0$, denote $\rho(\theta) = \min(\underline{\alpha} + 1 - \theta, \theta/\bar{\alpha}, (\underline{\alpha} + 2)/(\bar{\alpha} + 2))$. ($\theta/\bar{\alpha} = +\infty$ if $\bar{\alpha} = 0$.) For any real valued function g , let $R(g)$ be the misclassification rate of the classification rule $\text{sign}(g)$.

Theorem 3.2 [*Lin (2000)*] *Under Assumption 3.1 and Assumption 3.2, for any fixed $\theta > 0$, suppose $M_{(n)} \sim n^t$ for some $0 < t \leq 2/[3(1 + \rho(\theta))]$, then for any fixed $s > 0$, there exists finite constant $D(s)$ depending on s , and $N > 0$, such that for any $n > N$,*

$$n^{\gamma s} E[R(\hat{f}_M) - R(\phi^*)]^s \leq D(s); \quad (5)$$

where $\gamma = \min\{[t(\rho(\theta) + \theta)], 2/3 - [t(1 + \theta)]/3\}$. The constants $D(s)$ and N depend on $p_0(x)$ only through ζ , \bar{D} , \underline{D} , D_4 , and $\bar{\alpha}$.

4 Discussion

The SVM implement the Bayes rule in an interesting way: Instead of estimating $p_0(x)$, it estimates $\text{sign}[p_0(x) - 1/2]$. Our results show that this gives the SVM some advantages when our goal is binary classification with minimal expected misclassification rate. However, in some other situations the SVM needs to be modified, and should not be used as is.

In the case of unequal costs of false positive and false negative, or the case where sampling bias exists, the Bayes rule is $\text{sign}[p_0(x) - c]$, with $c \neq 1/2$. Hence the SVM as is will not perform optimally in this situation, and there is no direct way of getting $\text{sign}[p_0(x) - c]$ from $\text{sign}[p_0(x) - 1/2]$. Lin, Lee, and Wahba (2000) contains some extension of the SVM to such nonstandard situations.

Multi-class (N -class) classification problem arises naturally in practice. The Bayes rule in this case assigns the class label corresponding to the largest conditional class probability. Some authors suggested training N one-versus-rest SVMs and taking the class for a test subject to be that corresponding to the largest value of the classification functions. Our results show that this approach should work well when one of the conditional class probabilities is larger than $1/2$, (there is a majority class), but will not approach the Bayes rule when there is no majority class.

References

- [1] Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In D. Haussler (Ed.), *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*. Pittsburgh, PA: ACM Press.
- [2] Cox, D. D., & O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimates. *The Annals of Statistics*, 18:4, 1676-1695.
- [3] Cortes, C., & Vapnik, V. N. (1995). Support vector networks. *Machine Learning*, 20, 273 - 297.

- [4] Lin, Y. (1999). Support vector machines and the Bayes rule in classification. Technical Report 1014. Department of Statistics, University of Wisconsin, Madison. Accepted by *Datamining and Knowledge Discovery*.
 Lin, Y. (2000). On support vector machines. Technical Report 1029. Department of Statistics, University of Wisconsin, Madison. Submitted.
- [5] Lin, Y., Lee, Y., & Wahba, G. (2000). Support vector machines for classification in nonstandard situations. Technical Report 1016. Department of Statistics, University of Wisconsin, Madison. To appear in *Machine Learning*.
- [6] Poggio, T., & Girosi, F. (1998). A sparse representation for function approximation. *Neural Computation*, 10, 1445 - 1454.
- [7] Shawe-Taylor, J. & Cristianini, N. (1998). Robust Bounds on the Generalization from the Margin Distribution. Neuro COLT Technical Report TR-1998-029.
- [8] Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. New York. Springer Verlag.
- [9] Wahba, G. (1990). Spline Models for Observational Data. Philadelphia, PA: Society for Industrial and Applied Mathematics.
- [10] Wahba, G. (1999). Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In B. Scholkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in Kernel Methods - Support Vector Learning*. MIT Press.