Communications for Statistical Applications and Methods
2013, Vol. 20, No. 3, 225–233

DOI: http://dx.doi.org/10.5351/CSAM.2013.20.3.225

A Note on Linear SVM in Gaussian Classes

Yongho Jeon[1,a]

[a]Department of Applied Statistics, Yonsei University

Abstract

The linear support vector machine(SVM) is motivated by the maximal margin separating hyperplane and is a popular tool for binary classification tasks. Many studies exist on the consistency properties of SVM; however, it is unknown whether the linear SVM is consistent for estimating the optimal classification boundary even in the simple case of two Gaussian classes with a common covariance, where the optimal classification boundary is linear. In this paper we show that the linear SVM can be inconsistent in the univariate Gaussian classification problem with a common variance, even when the best tuning parameter is used.

Keywords: Consistency for classification, Fisher consistency, Gaussian linear discriminant analysis, support vector machines.

1. Introduction

We consider the binary classification problem commonly studied in statistics and pattern recognition. Let $X \in \mathbb{R}^d$ be the random input vector, and $Y \in \{-1, 1\}$ be the class label. Let the prior probabilities of the two classes be $\pi_1 = \text{pr}(Y = 1)$ and $\pi_0 = \text{pr}(Y = -1)$, and the nondegenerate class densities be denoted by $g_1(x)$ and $g_0(x)$. Then the density function of X is

$$f(x) = \pi_1 g_1(x) + \pi_0 g_0(x),$$

and the conditional probability of the positive class is

$$p(x) = \text{pr}(Y = 1|X = x) = \frac{\pi_1 g_1(x)}{\pi_1 g_1(x) + \pi_0 g_0(x)}.$$

For a classification rule η: $\mathbb{R}^d \to \{-1, 1\}$, the generalization error is the expected misclassification rate $R(\eta) = \text{pr}\{\eta(X) \neq Y\}$. If the conditional class probability $p(x) = \text{pr}(Y = 1|X = x)$ is available, it is well known that the optimal classification rule that minimizes the generalization error is $\eta_B(x) = \text{sign}\{p(x) - 1/2\}$. This optimal rule is usually called the Bayes (optimal) rule. The corresponding generalization error $R_B = R(\eta_B)$ is called the Bayes (optimal) risk. This is a lower bound of the generalization error of any classification rule.

In practice we do not know the underlying probability distribution of (X, Y), and need to learn a classification rule from a training sample. Denote the training sample by $D_n = \{(x_i, y_i), i = 1, \ldots, n\}$, where (x_i, y_i) are independent realizations of (X, Y). A sequence of classifiers ϕ_n based on the sample D_n is consistent if its generalization error $R(\phi_n)$ converges to the Bayes optimal risk R_B as $n \to \infty$.

This work was supported by the Basic Science Research Fund of the Department of Applied Statistics at Yonsei University.

[1] Assistant Professor, Department of Applied Statistics, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 120-749, Korea. E-mail: yjeon@yonsei.ac.kr

The support vector machine(SVM) is a classification method developed in the machine learning literature. It has been shown to give excellent performance in a number of practical studies. The hard margin linear SVM (Boser *et al.*, 1992) is motivated by the geometric consideration of maximizing the classification margin when the two classes of points in the training set can be separated by a linear hyperplane. This amounts to a quadratic programming problem: find $w \in \mathbb{R}^d$, $b \in \mathbb{R}$, to minimize $\|w\|^2/2$, subject to

$$w'x_i + b \geq +1, \quad \text{for } y_i = +1; \tag{1.1}$$
$$w'x_i + b \leq -1, \quad \text{for } y_i = -1. \tag{1.2}$$

That is, to maximize the distance between the two hyperplanes $w'x + b = +1$ and $w'x + b = -1$ under the condition that these two planes completely separate the positive and negative classes. Once such w and b are found, the SVM classification rule is $\text{sign}(w'x + b)$. Most often the two classes in the training set are not linearly separable, then constraints (1.1) and (1.2) cannot be satisfied simultaneously. The commonly used soft margin SVM (Cortes and Vspnik, 1995) deals with a nonseparable case by incorporating nonnegative slack variables into the hard margin SVM, resulting in a quadratic programming problem: Find $w \in \mathbb{R}^d, b \in \mathbb{R}$, and ξ_i, $i = 1, \ldots, n$, to minimize

$$\lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i,$$

under the constraints

$$w'x_i + b \geq +1 - \xi_i, \quad \text{for } y_i = +1; \tag{1.3}$$
$$w'x_i + b \leq -1 + \xi_i, \quad \text{for } y_i = -1; \tag{1.4}$$
$$\xi_i \geq 0, \quad \text{for all } i,$$

where $\lambda \geq 0$ is a control parameter to be chosen by the user. Often only a small fraction of the training points enter the final solution. Such sparsity enables fast implementation of the SVM.

The constraints (1.3) and (1.4) can be combined as $\xi_i \geq 1 - y_i(w'x_i + b)$, $i = 1, \ldots, n$, and the linear SVM is equivalent to

$$\frac{1}{n}\sum_{i=1}^{n}\{1 - y_i(w'x_i + b)\}_+ + \lambda\|w\|^2,$$

where the function $(\cdot)_+$ is defined as

$$\tau_+ = \begin{cases} \tau, & \tau > 0, \\ 0, & \tau < 0, \end{cases}$$

which is called the hinge loss function. Therefore it is easy to see that in the population space the linear SVM is to minimize

$$E[\{1 - Yf(X)\}_+] + \lambda_\infty J(f),$$

where $f(X) = w'X + b$ and $J(f) = \|w\|^2$, and λ_∞ is usually set to be zero.

For early references to theoretical results on the SVM, see Vapnik (1999), Cristianini and Shawe-Taylor (2000), and Schölkopf and Smola (2001). Such results typically bound the generalization

error of the SVM with empirical quantities related to the margins of the training sample points, or the empirical misclassification error. In most practical situations the Bayes optimal risk R_B is not zero, therefore any upper bound of the generalization error cannot go to zero. It is more appropriate to study the difference between the generalization error and the Bayes optimal risk, rather than the generalization error itself. The consistency properties of the SVM and its rate of convergence to the Bayes optimal risk have been well studied in Lin (2000, 2004); Zhang (2004); Steinwart (2005); Bartlett *et al.* (2006); Steinwart and Scovel (2007); Xu *et al.* (2009) and the risk function of the SVM associated with the hinge loss is studied in Blanchard *et al.* (2008). Particularly, Lin (2004) showed that, when the specification of the target function is flexible enough, the SVM procedure achieves the same decision boundary as the Bayes optimal rule in the population space, thus the Bayes optimal risk. This property is referred to as the Fisher consistency of a classification procedure. In a different line of research, Koo *et al.* (2008) studied asymptotic properties of the coefficients of variables in the linear SVM solution around the population minimizer.

However, the investigation of the consistency properties of the linear SVM is still limited. In the simple case of two normal class densities with a common covariance where the Bayes optimal decision boundary is linear, it has been unclear whether the linear SVM achieves the optimal boundary. Answering to this question is of natural interest since the SVM is known to be Fisher consistent and the linear SVM models the decision boundary as a linear function. In this paper, we explore this problem and show that the linear SVM can be inconsistent even in standard Gaussian classification problems. Section 2 considers a simple one-dimensional classification problem with two Gaussian classes, and investigate the conditions under which the linear SVM can achieve the Bayes optimal rule as well as the conditions it cannot. A discussion is given in Section 3.

2. Consistency of Linear SVM

In this section, we consider the linear SVM in one-dimensional Gaussian classification problem with a common variance. We show that the linear SVM is consistent to estimate the optimal classification boundary if the prior class probabilities are equal and we have the freedom to pick the best tuning parameter λ for λ_∞. We also show that, surprisingly, the linear SVM is inconsistent under some conditions even when the best possible tuning parameter is used.

Without loss of generality, suppose that the prior class probabilities π_0 and π_1 are positive and the class densities are $X|\{Y=-1\} \sim N(0,1)$, $X|\{Y=+1\} \sim N(\theta,1)$, with $\theta > 0$. Denote the pdf and the cdf of $N(0,1)$ by $\phi(\cdot)$ and $\Phi(\cdot)$ respectively. The population version of the SVM is to minimize

$$\begin{aligned} R(\alpha,\beta) &= \lambda\alpha^2 + E[\{1 - Y(\alpha X+\beta)\}_+] \\ &= \lambda\alpha^2 + \pi_1 E[\{1-(\alpha X+\beta)\}_+|Y=1] + \pi_0 E[\{1+(\alpha X+\beta)\}_+|Y=-1] \\ &= \lambda\alpha^2 + \pi_1 \int_{-\infty}^{\infty} \{1-(\alpha x+\beta)\}_+ \phi(x-\theta)dx + \pi_0 \int_{-\infty}^{\infty} \{1+(\alpha x+\beta)\}_+ \phi(x)dx. \end{aligned} \tag{2.1}$$

Since convexity is preserved under expectations, $R(\alpha,\beta)$ is convex in (α,β). The existence of a global minimizer is established in the following proposition.

Proposition 1. *There exists a global minimizer to the population version of the SVM (2.1) for any* $\lambda \geq 0$.

Proof: First consider the case $\alpha > 0$. Then,

$$\begin{aligned} R(\alpha,\beta) &= \lambda\alpha^2 + \pi_1 \int_{-\infty}^{\frac{1-\beta}{\alpha}} (1-\beta-\alpha x)\phi(x-\theta)dx + \pi_0 \int_{-\frac{1+\beta}{\alpha}}^{\infty} (1+\beta+\alpha x)\phi(x)dx \\ &= \lambda\alpha^2 + \pi_1 \int_{-\infty}^{\frac{1-\beta}{\alpha}} (1-\beta-\alpha x)\phi(x-\theta)dx + \pi_0 \int_{-\infty}^{\frac{1+\beta}{\alpha}} (1+\beta-\alpha x)\phi(x)dx \qquad (2.2) \\ &= \lambda\alpha^2 + \pi_1\alpha \int_{-\infty}^{\frac{1-\beta}{\alpha}} \Phi(x-\theta)dx + \pi_0\alpha \int_{-\infty}^{\frac{1+\beta}{\alpha}} \Phi(x)dx \qquad (2.3) \\ &= \lambda\alpha^2 + \pi_1\alpha \int_{-\infty}^{\frac{1-\beta}{\alpha}-\theta} \Phi(x)dx + \pi_0\alpha \int_{-\infty}^{\frac{1+\beta}{\alpha}} \Phi(x)dx. \qquad (2.4) \end{aligned}$$

The third step to obtain (2.3) uses integration by parts. Similar calculation gives that for $\alpha < 0$,

$$R(\alpha,\beta) = \lambda\alpha^2 - \pi_1\alpha \int_{-\infty}^{-\frac{1-\beta}{\alpha}+\theta} \Phi(x)dx - \pi_0\alpha \int_{-\infty}^{-\frac{1+\beta}{\alpha}} \Phi(x)dx. \qquad (2.5)$$

Thus for any $\alpha \neq 0$,

$$\begin{aligned} R(\alpha,\beta) &\geq \pi_1|\alpha| \int_{-\infty}^{\frac{1-\beta}{|\alpha|}-\theta} \Phi(x)dx + \pi_0|\alpha| \int_{-\infty}^{\frac{1+\beta}{|\alpha|}} \Phi(x)dx \\ &\geq |\alpha| \left\{ \int_{-\infty}^{\frac{1-\beta}{|\alpha|}-\theta} \Phi(x)dx + \int_{-\infty}^{\frac{1+\beta}{|\alpha|}} \Phi(x)dx \right\} \min(\pi_0,\pi_1) \\ &\geq |\alpha| \left\{ 2\int_{-\infty}^{\frac{1}{|\alpha|}-\frac{\theta}{2}} \Phi(x)dx \right\} \min(\pi_0,\pi_1). \end{aligned}$$

The last step uses that the function $\int_{-\infty}^{t} \Phi(s)ds$ is strictly convex in t, since $\phi(t) > 0$ for any t. The last expression involves only α and goes to infinity as $|\alpha| \to \infty$. One can choose $C_1 > 0$ independent of β such that for any $|\alpha| > C_1$, $R(\alpha,\beta) > R(0,0)$. However, for any $|\alpha| \leq C_1$, we have

$$\begin{aligned} R(\alpha,\beta) &= \lambda\alpha^2 + \pi_1 E[\{1-(\alpha X+\beta)\}_+ | Y=1] + \pi_0 E[\{1+(\alpha X+\beta)\}_+ | Y=-1] \\ &\geq \pi_1 E\{1-(\alpha X+\beta) | Y=1\} \\ &= \pi_1\{1-(\alpha\theta+\beta)\} \\ &\geq \pi_1(-\beta - C_1\theta). \end{aligned}$$

Similarly we have $R(\alpha,\beta) \geq \pi_0(1+\beta)$, and there exists $C_2 > 0$, such that $R(\alpha,\beta) > R(0,0)$ for any $|\beta| > C_2$ and $|\alpha| \leq C_1$. Thus, any (α,β) outside the set $\mathbb{C} = [-C_1, C_1] \times [-C_2, C_2]$ cannot be a global minimizer since $(0,0)$ attains a smaller value. Since $R(\alpha,\beta)$ is convex, if there is a minimizer $(\bar{\alpha},\bar{\beta})$ over $\mathbb{C}$, this must be a global minimizer. The existence of $(\bar{\alpha},\bar{\beta})$ is ensured by that $R(\alpha,\beta)$ is continuous and $\mathbb{C}$ is compact. Therefore, there exists a global minimizer of $R(\alpha,\beta)$ in $\mathbb{C}$. □

In the following, Proposition 2 and Lemma 1 are used to show the main results in Proposition 3.

Proposition 2. *For any $\lambda \geq 0$, $R(\alpha,\beta)$ is strictly convex for $\alpha > 0$.*

Proof: For $\alpha > 0$, continuing with (2.2) and using the fact that $\phi'(t) = -t\phi(t)$, we have

$$\begin{aligned} R(\alpha,\beta) &= \lambda\alpha^2 + \pi_1 \int_{-\infty}^{\frac{1-\beta}{\alpha}} \{1-\beta-\alpha\theta-\alpha(x-\theta)\}\,\phi(x-\theta)dx + \pi_0 \int_{-\infty}^{\frac{1+\beta}{\alpha}} (1+\beta-\alpha x)\phi(x)dx \\ &= \lambda\alpha^2 + \pi_1(1-\beta-\alpha\theta)\Phi\left(\frac{1-\beta}{\alpha}-\theta\right) \\ &\quad + \pi_1\alpha\phi\left(\frac{1-\beta}{\alpha}-\theta\right) + \pi_0(1+\beta)\Phi\left(\frac{1+\beta}{\alpha}\right) + \pi_0\alpha\phi\left(\frac{1+\beta}{\alpha}\right). \end{aligned} \tag{2.6}$$

A straightforward calculation with (2.6) gives A straightforward calculation with (2.6) gives

$$\begin{aligned} \frac{\partial R}{\partial\alpha} &= 2\lambda\alpha - \pi_1\theta\Phi\left(\frac{1-\beta}{\alpha}-\theta\right) + \pi_1\phi\left(\frac{1-\beta}{\alpha}-\theta\right) + \pi_0\phi\left(\frac{1+\beta}{\alpha}\right), \\ \frac{\partial R}{\partial\beta} &= -\pi_1\Phi\left(\frac{1-\beta}{\alpha}-\theta\right) + \pi_0\Phi\left(\frac{1+\beta}{\alpha}\right). \end{aligned}$$

A straightforward calculation with (2.6) gives

$$\begin{aligned} \frac{\partial R}{\partial\alpha} &= 2\lambda\alpha - \pi_1\theta\Phi\left(\frac{1-\beta}{\alpha}-\theta\right) + \pi_1\phi\left(\frac{1-\beta}{\alpha}-\theta\right) + \pi_0\phi\left(\frac{1+\beta}{\alpha}\right), \\ \frac{\partial R}{\partial\beta} &= -\pi_1\Phi\left(\frac{1-\beta}{\alpha}-\theta\right) + \pi_0\Phi\left(\frac{1+\beta}{\alpha}\right), \end{aligned}$$

and

$$\begin{aligned} \frac{\partial^2 R}{\partial\alpha^2} &= 2\lambda + \pi_1(1-\beta)^2\alpha^{-3}\phi\left(\frac{1-\beta}{\alpha}-\theta\right) + \pi_0(1+\beta)^2\alpha^{-3}\phi\left(\frac{1+\beta}{\alpha}\right), \\ \frac{\partial^2 R}{\partial\alpha\partial\beta} &= \pi_1(1-\beta)\alpha^{-2}\phi\left(\frac{1-\beta}{\alpha}-\theta\right) - \pi_0(1+\beta)\alpha^{-2}\phi\left(\frac{1+\beta}{\alpha}\right), \\ \frac{\partial^2 R}{\partial\beta^2} &= \pi_1\alpha^{-1}\phi\left(\frac{1-\beta}{\alpha}-\theta\right) + \pi_0\alpha^{-1}\phi\left(\frac{1+\beta}{\alpha}\right). \end{aligned}$$

The determinant of the Hessian of $R(\alpha,\beta)$ is

$$2\lambda\alpha^{-1}\left\{\pi_1\phi\left(\frac{1-\beta}{\alpha}-\theta\right) + \pi_0\phi\left(\frac{1+\beta}{\alpha}\right)\right\} + 4\pi_0\pi_1\alpha^{-4}\phi\left(\frac{1-\beta}{\alpha}-\theta\right)\phi\left(\frac{1+\beta}{\alpha}\right) > 0,$$

and $\partial^2 R/\partial\alpha^2 > 0$. Therefore, $R(\alpha,\beta)$ is strictly convex for $\alpha > 0$. □

Lemma 1. *For all $t \in \mathbb{R}$,*

$$t\Phi(t) + \phi(t) > t_+.$$

The function $h(t) = t\Phi(t) + \phi(t)$ is convex, and $\lim_{t\to-\infty} h(t) = 0$ and $\lim_{t\to\infty} h(t) - t = 0$.

Proof: $-t\Phi(t) = \int_{-\infty}^{t}\{-s\phi(s) - \Phi(s)\}ds < \int_{-\infty}^{t} -s\phi(s)ds = \phi(t)$. Therefore $t\Phi(t)+\phi(t) > 0$, for all t. By plugging $-t$ in the place of t, we get $t\Phi(t) + \phi(t) > t$, for all t. The rest of the proof is straightforward and omitted. □

Proposition 3.

(a) Any solution $(\bar{\alpha},\bar{\beta})$ to (2.1) should satisfy $\bar{\alpha} \geq 0$.

(b) Suppose $\pi_1 \neq \pi_0$. Any solution $(\bar{\alpha},\bar{\beta})$ should satisfy $\bar{\alpha} = 0$ if and only if π_1, π_0 and θ satisfy that for all $t \in \mathbb{R}$,

$$\min\left\{\frac{\pi_1}{\pi_0}, \frac{\pi_0}{\pi_1}\right\}(t+\theta) \leq t\Phi(t) + \phi(t). \tag{2.7}$$

(c) If $\pi_1 = \pi_0$, the condition (2.7) does not hold and any solution $(\bar{\alpha},\bar{\beta})$ should satisfy $\bar{\alpha} > 0$.

(d) The solution is unique.

Proof:

(a) From (2.4) and (2.5), for any $\alpha > 0$ and $\beta \in \mathbb{R}$, $R(-\alpha,\beta) > R(\alpha,\beta)$. Therefore, $\bar{\alpha}$ cannot be negative.

(b) Consider $R(0,\beta) = E\{1 - Y\beta\}_+ = \pi_1(1-\beta)_+ + \pi_0(1+\beta)_+$ which is piecewise linear in β. If $\pi_1 < \pi_0$, $R(0,\beta)$ is uniquely minimized at $\beta = -1$ with minimum $2\pi_1$. If $\pi_1 > \pi_0$, $R(0,\beta)$ is uniquely minimized at $\beta = +1$ with minimum $2\pi_0$. If $\pi_1 = \pi_0$, $R(0,\beta)$ is minimized at any point in $[-1,1]$ with minimum 1.

($\Leftarrow$) In the case $\pi_1 < \pi_0$, for any $\alpha > 0, \beta \in \mathbb{R}$,

$$\begin{aligned}
R(\alpha,\beta) &= \lambda\alpha^2 + \pi_1\alpha\left\{\left(\frac{1-\beta}{\alpha} - \theta\right)\Phi\left(\frac{1-\beta}{\alpha} - \theta\right) + \phi\left(\frac{1-\beta}{\alpha} - \theta\right)\right\} \\
&\quad + \pi_0\alpha\left\{\left(\frac{1+\beta}{\alpha}\right)\Phi\left(\frac{1+\beta}{\alpha}\right) + \phi\left(\frac{1+\beta}{\alpha}\right)\right\} \\
&> \lambda\alpha^2 + \pi_1\alpha\left(\frac{1-\beta}{\alpha} - \theta\right) + \pi_0\alpha\left\{\left(\frac{1+\beta}{\alpha}\right)\Phi\left(\frac{1+\beta}{\alpha}\right) + \phi\left(\frac{1+\beta}{\alpha}\right)\right\} \\
&\geq \lambda\alpha^2 + \pi_1(1-\beta-\alpha\theta) + \pi_1(1+\beta+\alpha\theta) \\
&= \lambda\alpha^2 + 2\pi_1 \\
&\geq 2\pi_1.
\end{aligned}$$

The second step uses Lemma 1 and the third step uses (2.7). Therefore, $R(\alpha,\beta) > 2\pi_1 = R(0,-1)$ for any $\alpha > 0$ and $\beta \in \mathbb{R}$, thus $(\bar{\alpha},\bar{\beta})$ with $\bar{\alpha} > 0$ cannot be a solution. When $\pi_1 > \pi_0$, a similar argument gives that $R(\alpha,\beta) > 2\pi_0 = R(0,+1)$ for any $\alpha > 0$ and $\beta \in \mathbb{R}$, and $(\bar{\alpha},\bar{\beta})$ with $\bar{\alpha} > 0$ cannot be a solution.

($\Rightarrow$) We prove this by negation. Consider the case $\pi_1 > \pi_0$ and suppose that (2.7) does not hold, then one can take k such that $(k-\theta)\Phi(k-\theta) + \phi(k-\theta) < k\pi_0/\pi_1$. For $\alpha > 0$, if we choose a path $\beta(\alpha) = 1 - k\alpha$, then

$$\frac{\partial R(\alpha,\beta(\alpha))}{\partial\alpha} = 2\lambda\alpha - \pi_1\theta\Phi(k-\theta) + \pi_1\phi(k-\theta) + \pi_0\phi\left(\frac{2}{\alpha} - k\right) - k\left\{\pi_0\Phi\left(\frac{2}{\alpha} - k\right) - \pi_1\Phi(k-\theta)\right\}.$$

$R(\alpha, 1-k\alpha)$ is decreasing at $\alpha = 0$, since $\lim_{\alpha\to+0} \partial R(\alpha,\beta(\alpha))/\partial\alpha = \pi_1\{(k-\theta)\Phi(k-\theta) + \phi(k-\theta) - k\pi_0/\pi_1\} < 0$. Therefore, there exists a positive $\bar{\alpha}_*$ such that $R(\bar{\alpha}_*, 1-k\bar{\alpha}_*) < R(0,1) = \min_{\beta\in\mathbb{R}} R(0,\beta)$.

When $\pi_1 < \pi_0$, supposing that (2.7) does not hold, one can take k such that $k\Phi(k) + \phi(k) < (k+\theta)\pi_1/\pi_0$. For $\alpha > 0$, if we choose a path $\beta(\alpha) = -1 + k\alpha$, then

$$\frac{\partial R(\alpha,\beta(\alpha))}{\partial\alpha} = 2\lambda\alpha - \pi_1\theta\Phi\left(\frac{2}{\alpha} - k - \theta\right) + \pi_1\phi\left(\frac{2}{\alpha} - k - \theta\right) + \pi_0\phi(k) + k\left\{\pi_0\Phi(k) - \pi_1\Phi\left(\frac{2}{\alpha} - k - \theta\right)\right\}.$$

$R(\alpha, -1+k\alpha)$ is decreasing at $\alpha = 0$, since $\lim_{\alpha\to+0} \partial R(\alpha,\beta(\alpha))/\partial\alpha = \pi_0\{k\Phi(k) + \phi(k) - (k+\theta)\pi_1/\pi_0\} < 0$. Therefore there exists a positive $\bar{\alpha}_*$ such that $R(\bar{\alpha}_*, -1+k\bar{\alpha}_*) < R(0,-1) = \min_{\beta\in\mathbb{R}} R(0,\beta)$.

(c) If the condition (2.7) holds, π_1 and π_0 cannot be the same since $\theta > 0$. Since $R(0,\beta)$ is minimized at any point in $[-1,1]$ with minimum 1 when $\pi_1 = \pi_0$, we have $\min_{\beta\in\mathbb{R}} R(0,\beta) = R(0,0)$. On the other hand $R(\alpha,0)$ is continuous in α and $\lim_{\alpha\to+0} \partial R(\alpha,0)/\partial\alpha = -\pi_1\theta < 0$ from

$$\frac{\partial R(\alpha,0)}{\partial\alpha} = 2\lambda\alpha - \pi_1\theta\Phi\left(\frac{1}{\alpha} - \theta\right) + \pi_1\phi\left(\frac{1}{\alpha} - \theta\right) + \pi_0\phi\left(\frac{1}{\alpha}\right).$$

$R(\alpha,0)$ is decreasing at $\alpha = 0$ and there exists a positive $\bar{\alpha}_*$ such that $R(\bar{\alpha}_*, 0) < R(0,0)$. Therefore $\bar{\alpha}$ cannot be zero.

(d) If the condition (2.7) holds, then $\bar{\alpha}$ cannot be positive. So, the solution can be found by minimizing $R(0,\beta)$, and is unique since $\pi_1 \neq \pi_0$. If the condition (2.7) does not hold, then uniqueness follows from that one cannot have a minimizer with $\bar{\alpha} = 0$ and $R(\alpha,\beta)$ is strictly convex when $\alpha > 0$ by Proposition 2. □

Proposition 3 (a) states that the SVM classifier in the population space, $\text{sign}\{\bar{\alpha}x + \bar{\beta}\}$ cannot have $\bar{\alpha} < 0$. This is natural as we assume the class mean θ for class +1 is positive.

For Proposition 3 (b), the condition (2.7) can be written in a different form. The tangent line to $y = t\Phi(t) + \phi(t)$ at $t = t_0$ is $y = \Phi(t_0)t + \phi(t_0)$. We have $at + b \le t\Phi(t) + \phi(t), \forall t$ if and only if there exists t_0 such that $at + b \le \Phi(t_0)t + \phi(t_0), \forall t$, i.e., $a - \Phi(t_0) = 0$ and $b - \phi(t_0) \le 0$. Therefore, the condition is equivalent to $b \le \phi(\Phi^{-1}(a))$. Letting $\rho = \log(\pi_1/\pi_0)$ and taking $a = e^{-|\rho|}$ and $b = e^{-|\rho|}\theta$, we obtain the following.

Remark 1. The condition (2.7) is equivalent to $\theta \le e^{|\rho|}\phi\{\Phi^{-1}(e^{-|\rho|})\}$.

Figure 1 shows the boundary for the condition (2.7) and we have $\bar{\alpha} = 0$ under the curve. In this area, the classifier is $\text{sign}\{\bar{\alpha}x + \bar{\beta}\} = \text{sign}\{\bar{\beta}\}$, thus the classifier does not consider the x value and all the cases are classified into the same class. Since $R(0,\beta) = \pi_1(1-\beta)_+ + \pi_0(1+\beta)_+$ is uniquely minimized at $\beta = -1$ when $\pi_1 < \pi_0$, all the cases are classified into -1 in the left side under the curve. Likewise, all the cases are classified into +1 in the right side under the curve.

The following theorem is from Li and Duan (1989), Theorem 5.1, and used to show that the one dimensional linear SVM can be consistent when $\pi_0 = \pi_1 = 1/2$.

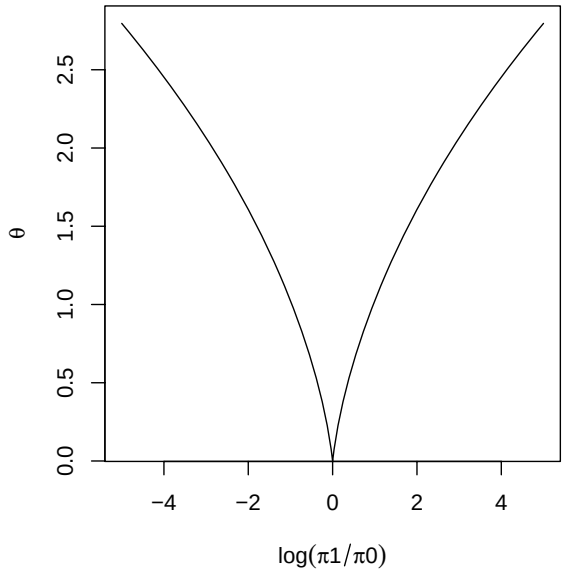

Figure 1: *Boundary for the condition (2.7). The curve corresponds to* $\theta = e^{|\rho|}\phi\{\Phi^{-1}(e^{-|\rho|})\}$*, where* $\rho = \log(\pi_1/\pi_0)$*. The solution is* $(\bar{\alpha}, \bar{\beta}) = (0, -1)$ *in the left side under the curve and* $(\bar{\alpha}, \bar{\beta}) = (0, +1)$ *in the right side under the curve.*

Theorem 1. *For a loss function* $L(\theta, y)$*, if a linear classification method of the form*

$$\min_{w,b} R_\lambda(w, b) = \min_{w,b} \left[E\{L(w'X + b, Y)\} + \lambda w'w \right].$$

in the population space has a unique solution $(\bar{w}, \bar{b})$*, then the set of estimates* $(\hat{w}, \hat{b})$ *from its empirical version*

$$\min_{w,b} \left\{ n^{-1} \sum_i L\left(w'x_i + b, y_i\right) + \lambda_n w'w \right\}.$$

with $\lambda_n \to \lambda$ *converges almost surely to the solution.*

Proposition 4. *When* $\pi_0 = \pi_1 = 1/2$*, the one dimensional linear SVM leads to consistent classification if we have the freedom to pick any sequence* $\lambda_n \geq 0$*.*

Proof: Let us take any $S > 0$ such that

$$\theta\Phi\left(S - \frac{\theta}{2}\right) - 2\phi\left(S - \frac{\theta}{2}\right) > 0.$$

This is always possible since the left hand side goes to $\theta > 0$ as $S \to \infty$. Take $\lambda = \{\theta\Phi(S - \theta/2) - 2\phi(S - \theta/2)\}S/4$. Then it is easy to check that at $(1/S, -\theta/(2S))$, both $\partial R/\partial\alpha$ and $\partial R/\partial\beta$ are zero, therefore $(1/S, -\theta/(2S))$ is the unique global solution since $1/S > 0$. The classification decision is $\text{sign}\{x/S - \theta/(2S)\} = \text{sign}(x - \theta/2)$, which is the Bayes optimal rule. Therefore by Theorem 1, the one dimensional linear SVM with a choice of λ can be consistent. □

3. Discussion

In the Gaussian classification problem with a common covariance, it is well known that the Bayes optimal classification boundary is linear. The SVM procedure is known to be Fisher consistent with flexible specification, and the linear SVM models the decision boundary as a linear function. Therefore it is of interest whether the linear SVM achieves the optimal classification boundary. In this paper, we consider the univariate Gaussian classification problem with a common variance and show that the linear SVM can be consistent if the prior probabilities are the same but it is not consistent with the condition (2.7). It is still unclear if the linear SVM can achieve the optimal classification boundary in the case where $\pi_0 \neq \pi_1$ but the condition (2.7) does not hold, although we conjecture that the linear SVM does not lead to consistent classification for this case either. This merits further investigations in the future study.

References

Bartlett, P., Jordan, M. and McAuliffe, J. (2006). Convexity, classification, and risk bounds, *Journal of the American Statistical Association*, **101**, 138–156.

Blanchard, G., Bousquet, O. and Massart, P. (2008). Statistical performance of support vector machines, *The Annals of Statistics*, **36**, 489–531.

Boser, B., Guyon, I. and Vapnik, V. (1992). A training algorithm for optimal margin classifiers, In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–152, ACM.

Cortes, C. and Vapnik, V. (1995). Support-vector networks, *Machine Learning*, **20**, 273–297.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University press.

Koo, J., Lee, Y., Kim, Y. and Park, C. (2008). A bahadur representation of the linear support vector machine, *The Journal of Machine Learning Research*, **9**, 1343–1368.

Li, K.-C. and Duan, N. (1989). Regression analysis under link violation, *The Annals of Statistics*, **17**, 1009–1052.

Lin, Y. (2000). Some asymptotic properties of the support vector machine, Technical Report 1029, Department of Statistics, University of Wisconsin-Madison.

Lin, Y. (2004). A note on margin-based loss functions in classification, *Statistics & Probability Letters*, **68**, 73–82.

Schölkopf, B. and Smola, A. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT press.

Steinwart, I. (2005). Consistency of support vector machines and other regularized kernel classifiers, *IEEE Transactions on Information Theory*, **51**, 128–142.

Steinwart, I. and Scovel, C. (2007). Fast rates for support vector machines using Gaussian kernels, *The Annals of Statistics*, **35**, 575–607.

Vapnik, V. (1999). *The Nature of Statistical Learning Theory*, Springer.

Xu, H., Caramanis, C. and Mannor, S. (2009). Robustness and regularization of support vector machines, *The Journal of Machine Learning Research*, **10**, 1485–1510.

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

Received April 7, 2013; Revised April 23, 2013; Accepted April 27, 2013