
Statistical Properties of Support Vector Machines and Related Topics

Yi Lin ^a

University of Wisconsin – Madison

^apartially based on joint work with Yoonkyung Lee and Grace Wahba.

Contents

1. Support vector machines in classification.
2. Support vector machines and the Bayes optimal rule.
3. Margin-based loss functions and Fisher consistency for classification.
4. Multi-category support vector machines.

The binary classification problem

Population distribution: $P(\mathbf{x}, y)$, $\mathbf{x} \in R^d$, $y \in \{-1, 1\}$.

Classification rule: $\phi: R^d \rightarrow \{-1, 1\}$.

Generalization error: $P[Y \neq \phi(\mathbf{X})]$.

Bayes (optimal) rule: $\text{sign}[p(\mathbf{x}) - 1/2]$, where
 $p(\mathbf{x}) = \text{Pr}\{Y = 1 | \mathbf{X} = \mathbf{x}\}$.

Bayes (optimal) risk: generalization error of the Bayes rule.

Training set: (\mathbf{x}_i, y_i) , $i = 1, \dots, n$.

Consistent sequence of classifiers $\hat{\phi}_n$: the generalization error goes to the Bayes optimal risk as $n \rightarrow \infty$.

Traditional methods

Discriminant analysis: through density estimation.

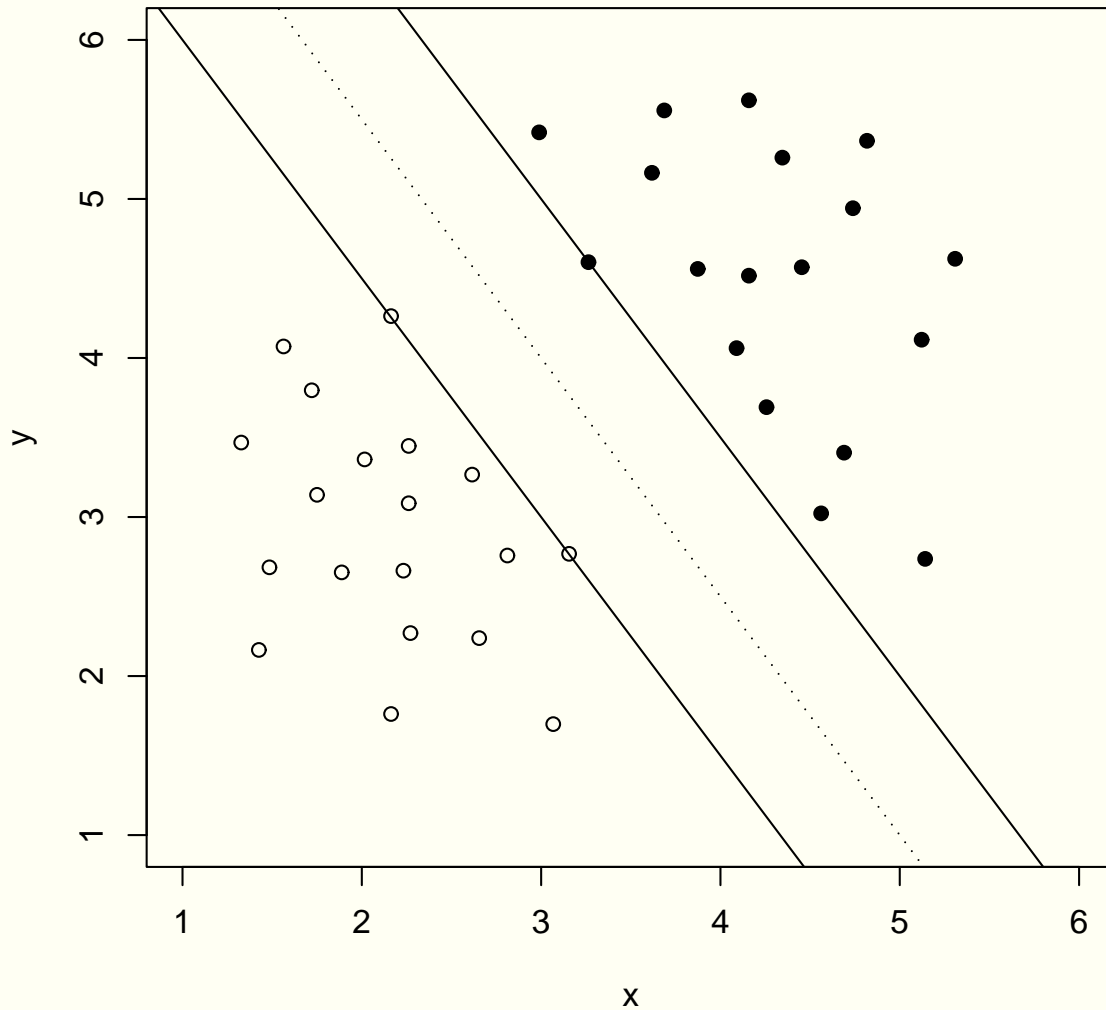
Logistic regression: minimize in H_K

$$\frac{1}{n} \sum_i \log[1 + \exp(-y_i f(\mathbf{x}_i))] + \lambda J(f)$$

The target function is $\log[p(\mathbf{x})/(1 - p(\mathbf{x}))]$, which has the same sign as $p(\mathbf{x}) - 1/2$. Note that $\log[p(\mathbf{x})/(1 - p(\mathbf{x}))]$ is the minimizer of $E \log[1 + \exp(-Y f(\mathbf{X}))]$.

Optimal separating plane

Optimal separating hyperplane



SVMs for linearly separable data

Boser, Guyon, and Vapnik (1992): find the optimal separating hyperplane $\mathbf{x} \cdot \mathbf{w} + b = 0$.

Minimize $\|\mathbf{w}\|^2/2$

subject to

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq +1 \quad \text{for } y_i = +1;$$

$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1 \quad \text{for } y_i = -1;$$

Linear non-separable case

Cortes and Vapnik (1995): Introduce non-negative slack variables.
Modify the constraints to

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq +1 - \xi_i \quad \text{for } y_i = +1;$$

$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1 + \xi_i \quad \text{for } y_i = -1;$$

$$\xi_i \geq 0, \quad \forall i.$$

The objective function is now

$$\|\mathbf{w}\|^2/2 + C\left(\sum_i \xi_i\right)^q.$$

Here C is a parameter to be chosen by the user.

Feature space

Map the input data into high (even infinite) dimensional feature space F with $\Phi : R^d \rightarrow F$:

$$\mathbf{x} \rightarrow \Phi(\mathbf{x}),$$

and then perform the linear algorithm in the feature space.
The linear SVM depends on the feature space only through the (reproducing) kernel

$$(\Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2))_F = K(\mathbf{x}_1, \mathbf{x}_2).$$

The regularization formulation of the SVM

Wahba (1998), Girosi (1998), Smola, Schölkopf, and Müller (1998):
The SVM with kernel K is equivalent to:

$$(1) \quad \min_{f \in H_K} \frac{1}{n} \sum_{i=1}^n [(1 - y_i f(\mathbf{x}_i))_+]^q + \lambda |f|_{H_K}^2.$$

Here

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

Method of regularization

Minimize in H_K

$$1/n \sum_{i=1}^n \ell[y_i, f(\mathbf{x}_i)] + \lambda J(f).$$

Under the assumption that the minimizer of $E\ell[Y, f(\mathbf{X})]$ (the target function) is in H_K , the solution approaches the target function as $n \rightarrow \infty$. [Cox and O'Sullivan (1990)].

The support vector machine

Lin (1999):

Lemma 1 *The minimizer of $E[(1 - Y f(\mathbf{X}))_+]^2$ is $\text{sign}[p(\mathbf{x}) - 1/2]$.*

Lemma 2 *The minimizer of $E[(1 - Y f(\mathbf{X}))_+]^2$ is given by $2p(\mathbf{x}) - 1$.*

Theorem 1 *Assume that $p(\mathbf{x})$ is in $\bigotimes^d H^m([0, 1])$, and $0 < p(\mathbf{x}) < 1, \forall \mathbf{x} \in [0, 1]^d$. Then if $\lambda \rightarrow 0$, and $n^{-1} \lambda^{-(\frac{3}{2m} + \epsilon)} \rightarrow 0$ for some $\epsilon > 0$. Then*

$$\int_{[0,1]^d} [\hat{f} - (2p - 1)]^2 = O_p[\lambda + n^{-1} \lambda^{-\frac{1}{2m}} (\log \frac{1}{\lambda})^{d-1}].$$

Steinwart (2001) studied the consistency of the support vector machine. Lin (2002) studied the rate of convergence in a special case.

Fisher consistency in classification

The infinite population space view (Duan and Li, 1989; Breiman, 2000):

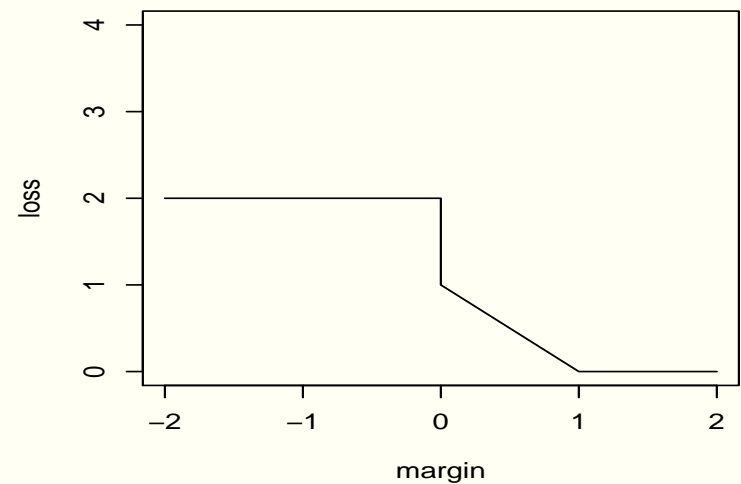
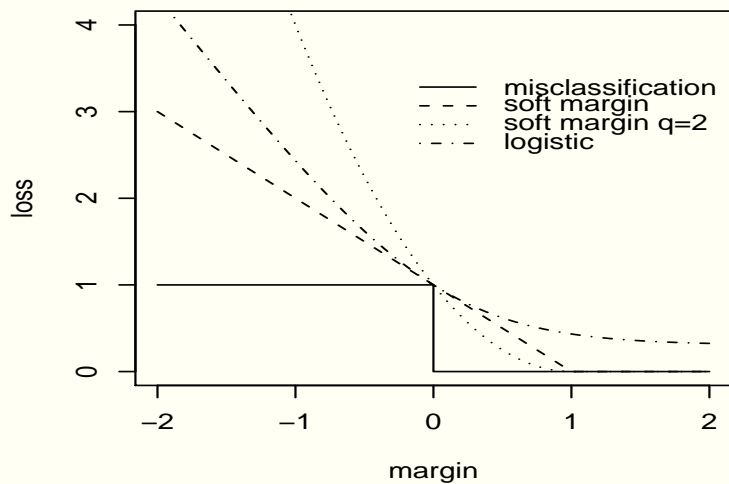
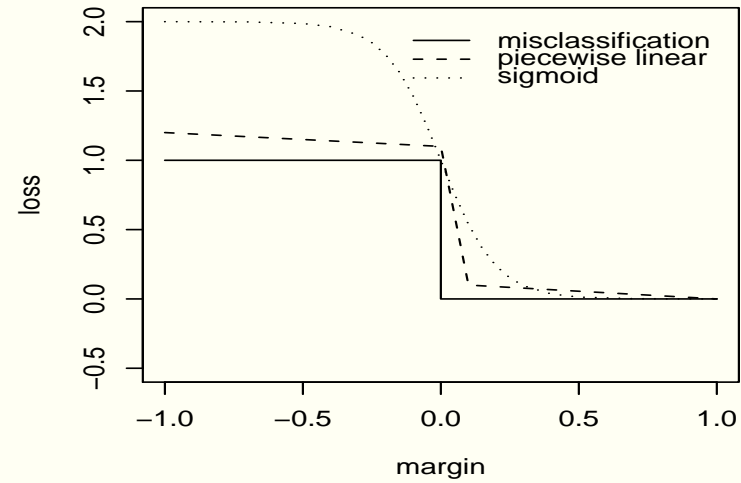
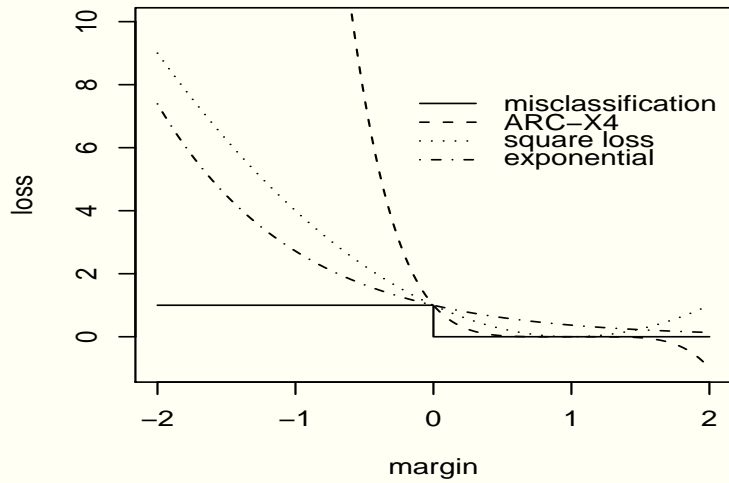
1. In M-estimation: $\arg \min_f \frac{1}{n} \sum_i \ell(y_i, f(\mathbf{x}_i))$, Fisher consistency means that the minimizer of $E\ell(Y, f(\mathbf{X}))$ is f_0 , the true function.
2. In classification, one can define the Fisher consistency to be that the minimizer of $E\ell(Y, f(\mathbf{X}))$ has the same sign as $(p(\mathbf{x}) - 1/2)$.

Margin based loss functions

For a specific function f , the margin of a subject (\mathbf{x}, y) , if we use the classification rule $\text{sign}[f(\mathbf{x})]$, is defined as $yf(\mathbf{x})$.

- The misclassification loss: $1_{\{-yf(\mathbf{x})\}}$.
- The hinge loss: $[1 - yf(\mathbf{x})]_+$.
- The exponential loss $\exp[-yf(\mathbf{x})]$ (Schapire and Singer, 1998), the negative log-likelihood $\log[1 + e^{-2yf(\mathbf{x})}]$.
- The ARC-X4 loss $[1 - yf(\mathbf{x})]^5$ (Breiman, 1998), the square loss $[1 - yf(\mathbf{x})]^2$ (Lee *et al*, 1996).
- Marginboost losses and their approximations, normalized sigmoid loss $1 - \tanh[\lambda yf(\mathbf{x})]$. (Mason, Bartlett, and Baxter, 1999).
- ψ learning loss. (Shen *et. al.*, 2001).

Margin based loss functions



Fisher consistency of margin based loss functions

Assumptions

- i $V(z) < V(-z), \forall z > 0.$
- ii $V'(0) \neq 0$ exists.

Lemma 3 *Let $\bar{f}(\mathbf{x})$ be the minimizer of $EV[Y f(\mathbf{X})]$ for some function V satisfying assumptions (i) and (ii). Then we have $sign(\bar{f}) = sign(p - 1/2),$ a.s..*

Consequence: all margin based loss functions used in practice are Fisher consistent.

Remark 1 *Zhang (2003) obtained similar results independently.*

Remark 2 *The minimizer \bar{f} may take on values $\pm\infty.$*

Classification and estimation

Assumption iii: $V''(z) > 0, \forall z$.

Lemma 4 *Let \bar{f} be the minimizer of $EV[Y f(\mathbf{X})]$ for some function V satisfying assumptions (i), (ii), and (iii). Then for any function f ,*

$$R(\text{sign}(f)) - R^* \leq cE|f - \bar{f}| \leq c\{E(f - \bar{f})^2\}^{1/2},$$

where c is a constant depending only on V .

Remark 3 *Assumption iii can be relaxed. It can be shown that the SVM loss satisfies the conclusion though does not satisfy Assumption iii.*

Remark 4 *The lemma can be used to establish the rate of convergence to the Bayes optimal risk, but in many cases better rates can be obtained by other approaches.*

Multi-category problems

Population distribution: $P(\mathbf{x}, y), \mathbf{x} \in R^d, y \in \{1, 2, \dots, k\}$.

Classification rule: $\phi: R^d \rightarrow \{1, 2, \dots, k\}$.

Generalization error: $P[Y \neq \phi(\mathbf{X})]$.

Bayes (optimal) rule: $arg \max_{j=1, \dots, k} p_j(\mathbf{x})$, where
 $p_j(\mathbf{x}) = P(Y = j | \mathbf{X} = \mathbf{x})$.

Multi-category classification by SVM

One-vs-rest approach: 1. Solve k binary problems. The j -th problem involves class j and the rest.

2. Assign class label according to the largest $f_j(\mathbf{x})$.

This approach can be inconsistent.

Pairwise approach.

Some multi-category extensions: Vapnik (1998), Weston and Watkins (1999), and Bredensteiner and Bennett (1999).

Multi-category SVM

Lee, Lin, and Wahba (2002):

Class codes: for examples in class j , \mathbf{y} is represented by a k -dimensional vector with 1 in the j -th coordinate and $-\frac{1}{k-1}$ elsewhere. For example, when $k = 3$, $\mathbf{y} = (-\frac{1}{2}, 1, -\frac{1}{2})$ for the second class.

Separating functions: $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_k(\mathbf{x}))$ with $\sum_{j=1}^k f_j(\mathbf{x}) = 0$ for any $\mathbf{x} \in R^d$.

Multi-category SVM (cont.)

Multi-category SVM:

$$\min \frac{1}{n} \sum_{i=1}^n L(\mathbf{y}_i) \cdot (\mathbf{f}(\mathbf{x}_i) - \mathbf{y}_i)_+ + \frac{1}{2} \lambda \sum_{j=1}^k |f_j|_{H_K}^2$$

where $L(\mathbf{y}_i) \cdot (\mathbf{f}(\mathbf{x}_i) - \mathbf{y}_i)_+ = \sum_{j=1}^k L_{cat(i)j} (f_j(\mathbf{x}_i) - y_{ij})_+$,

$cat(i)$: the category of \mathbf{y}_i ;

$L_{jj'}$: the cost of misclassifying j as j' .

When $L_{jj'} = I(j \neq j')$,

$$L(\mathbf{y}_i) \cdot (\mathbf{f}(\mathbf{x}_i) - \mathbf{y}_i)_+ = \sum_{j \neq cat(i)} (f_j(\mathbf{x}_i) + \frac{1}{k-1})_+.$$

The classification rule: $\phi(\mathbf{x}) = \arg \max_j f_j(\mathbf{x})$.

Multi-category SVM (cont.)

1. The multi-category SVM formulation reduces to the binary SVM when $k = 2$.
2. The multi-category SVM retains the sparsity of solution in the same way as binary SVM.
3. Fisher consistency:

Lemma 5 *The minimizer of $E[L(Y) \cdot (\mathbf{f}(X) - Y)_+]$ under the sum-to-zero constraint is $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_k(\mathbf{x}))$ with*

$$f_j(\mathbf{x}) = \begin{cases} 1 & \text{if } j = \arg \max_{l=1, \dots, k} p_l(\mathbf{x}) \\ -\frac{1}{k-1} & \text{otherwise} \end{cases}$$

Some Other Multi-category SVM

Other extensions were given by Vapnik (1998), Weston and Watkins (1999), and Bredensteiner and Bennett (1999).

Guermeur (2000) showed that they are essentially equivalent and amount to using the loss function:

$$l(y_i, \mathbf{f}(\mathbf{x}_i)) = \sum_{j=1, j \neq y_i}^k (f_j(\mathbf{x}_i) - f_{y_i}(\mathbf{x}_i) + 2)_+.$$

The induced classifier is $\phi(\mathbf{x}) = \mathit{arg} \max_j f_j(\mathbf{x})$.

Some Other Multi-category SVM (cont.)

The population version of the loss at \mathbf{x} is given by

$$E [l(Y, \mathbf{f}(X)) | X = \mathbf{x}] = \sum_{j=1}^k \left[\sum_{m \neq j} (f_m(\mathbf{x}) - f_j(\mathbf{x}) + 2)_+ \right] p_j(\mathbf{x}).$$

The following lemma shows that the population minimizer of the loss does not always implement the Bayes decision rule through $\phi(\mathbf{x}) = \mathit{arg} \max_j f_j(\mathbf{x})$.

Lemma 6 *Consider the case of $k = 3$ classes with $p_1 < 1/3 < p_2 < p_3 < 1/2$ at a given point \mathbf{x} . To insure uniqueness, without loss of generality we can fix $f_1(\mathbf{x}) = -1$. Then the unique population minimizer of the above loss, (f_1, f_2, f_3) at \mathbf{x} is $(-1, 1, 1)$.*

Discussion

1. The SVM targets the Bayes rule $\text{sign}[p(\mathbf{x}) - 1/2]$ directly:
 - (a) Impossible to recover probability.
 - (b) Nonstandard situation.
 - (c) Multicategory SVM.
2. In binary classification, a very general class of margin based loss functions are Fisher consistent.