Analysis of Support Vector Machine Classification

Ding-Xuan Zhou

Department of Mathematics

City University of Hong Kong

E-mail: mazhou@math.cityu.edu.hk

http://www.cityu.edu.hk/ma

Joint work with Qiang Wu

Supported in part by
Research Grants Council of Hong Kong

· **Classification algorithm**

(X, d) is a compact metric space (can be a subset of $\mathbf{R}^n$)

A (binary) **classifier** is a function $f : X \to \{1, -1\} = Y$.

It assigns a label (makes a decision) for each input (vector) $x \in X$:

$f(x) = 1$ (YES) or $f(x) = -1$ (NO)

ρ: (unknown) prob. measure on $Z := X \times Y$, ρ_X marg. distrib.
$\rho(y|x)$ the conditional probability measure at x for a given input x:

$$\begin{cases} P(y = 1|x) & \text{is the probability for the output to be 1 (YES)} \\ P(Y = -1|x) & \text{is the probability for the output to be } -1 \text{ (NO)} \end{cases}$$

Best classifier (Bayes rule) f_c:

$$f_c(x) = \begin{cases} 1, & \text{if } P(y = 1|x) \geq P(Y = -1|x) \\ -1, & \text{if } P(y = 1|x) < P(Y = -1|x) \end{cases}$$

· **Misclassification error** $\mathcal{R}(f)$:

$\text{Prob}\{Y \neq f(X)\} = \int_X P(Y \neq f(x)|x) d\rho_X$

Then $\mathcal{R}(f) \geq \mathcal{R}(f_c)$ for any $f : X \to Y$.

Purpose: find a good approximation $f_{\mathbf{z}}$ of f_c from training data

$\mathbf{z} = (x_i, y_i)_{i=1}^m \subset Z^m$, a set of samples drawn according to ρ

· **Convergence** (a good classification algorithm $f_{\mathbf{z}}$):

$\mathcal{R}(f_{\mathbf{z}}) \to \mathcal{R}(f_c)$ with confidence as $m \to \infty$

or when $f_{\mathbf{z}}$ is found from a function space (not dense) $\mathcal{H}$:

$$\mathcal{R}(f_{\mathbf{z}}) \to \inf_{f \in \mathcal{H}} \mathcal{R}(f) \text{ with confidence as } m \to \infty$$

· Support Vector Machine (SVM) Classification

Let $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ be a space of continuous functions

SVM 1-norm soft margin classifier is $\mathrm{sgn}(f^*_{\mathbf{z}} + b_{\mathbf{z}})$ defined by

$$(f^*_{\mathbf{z}}, b_{\mathbf{z}}) := \arg \min_{f^* \in \mathcal{H}, b \in \mathbf{R}} \quad \tfrac{1}{2}\|f^*\|^2_{\mathcal{H}} + \tfrac{C}{m}\sum_{i=1}^m \xi_i,$$
$$\text{subj. to} \quad y_i(f^*(x_i) + b) \geq 1 - \xi_i,$$
$$\xi_i \geq 0, \qquad i = 1, \ldots, m.$$

The most important SVM classification algorithm: $\mathcal{H} = \mathcal{H}_K$

· **Mercer kernel**

$K : X \times X \to \mathbf{R}$ continuous, symmetric, and positive semidefinite: $(K(x_i, x_j))_{i,j=1}^{\ell}$ is positive semidefinite for any $\ell \in \mathbf{N}, (x_i)_{i=1}^{\ell} \subset X$

Examples of Mercer kernels on $X \subset \mathbf{R}^n$

Vapnik: $K(x, y) = (x \cdot y)^d$ or $(1 + x \cdot y)^d$

Gaussian $K(x, y) = k(x - y) = e^{-\frac{|x-y|^2}{\sigma^2}}$

RBF: $K(x, y) = \int_0^{+\infty} e^{-\rho|x-y|^2} d\beta(\rho)$ with a Borel measure β

Spline kernels

· **Reproducing Kernel Hilbert Space $\mathcal{H}_K$**

closure of the span of the set of functions $\{K_t := K(t, \cdot) : t \in X\}$ with the inner product $< \quad , \quad >_{\mathcal{H}_K}$ satisfying $< K_x, K_y >_{\mathcal{H}_K} = K(x, y)$ and

$$< \sum_i c_i K_{x_i}, \sum_i c_i K_{x_i} >_{\mathcal{H}_K} = \| \sum_i c_i K_{x_i} \|^2_{\mathcal{H}_K} = \sum_{i,j} c_i K(x_i, x_j) c_j \geq 0.$$

$\mathcal{H}_K$ is a subspace of $C(X)$. $\mathcal{H}_K = \Pi_d(\mathbf{R}^n)$ if $K(t, x) = (1 + t \cdot x)^d$.

Denote $\overline{\mathcal{H}}_K = \mathcal{H}_K + \mathbf{R}$ and for $f = f_1 + b \in \overline{\mathcal{H}}_K$, $f^* = f_1, b_f = b$.

· **SVM 1-norm soft margin classifier** with kernel K : $\mathrm{sgn}(f_{\mathbf{z}})$
(Cortes-Vapnik, 1995)

$$f_{\mathbf{z}} := \arg\min_{f\in\overline{\mathcal{H}}_K} \quad \tfrac{1}{2}\|f^*\|_K^2 + \tfrac{C}{m}\sum_{i=1}^m \xi_i,$$
$$\text{subj. to} \quad y_i(f(x_i)) \geq 1-\xi_i,$$
$$\xi_i \geq 0, \qquad i=1,\ldots,m.$$

Rosenblatt (1962): perceptron $K(x,y) = x\cdot y \quad \overline{\mathcal{H}}_K = \Pi_1$
Boser-Guyon-Vapnik (1992): $K(x,y) = (1+x\cdot y)^d \quad \overline{\mathcal{H}}_K = \Pi_d$

· **Expected convergence**:

$$\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z}})) \to \inf_{f\in\overline{\mathcal{H}}_K} \mathcal{R}(\mathrm{sgn}(f)) \text{ with confidence as } m, C\to\infty$$

· **Negative result**:

WU-Zhou: a counterexample with $K(x,y) = x\cdot y, X = [-1,1]$

· **Positive result**:

I. Steinwart, Y. Lin, T. Zhang ... convergence holds when $\mathcal{H}_K$ is dense, but no convergence rate estimate

Idea: reduce SVM to the empirical risk minimization (ERM)
Vapnik, Evgeniou-Pontil-Poggio, Wahba, Cucker-Smale, Niyogi, ...

Define a loss function V as

$$V(y,f(x)) := |1-yf(x)|\chi_{yf(x)\leq 1} = (1-yf(x))_+,$$

then

$$f_{\mathbf{z}} = \arg\min_{f\in\overline{\mathcal{H}}_K} \frac{1}{m}\sum_{i=1}^m V(y_i,f(x_i)) + \frac{1}{2C}\|f^*\|_K^2.$$

· **Discrepancy Principle (Morozov) in inverse problems**

$$f_{y,\gamma} := \arg\min_{f\in\mathcal{H}_1}\Big\{\|\mathcal{K}f - y\|^2_{\mathcal{H}_2} + \gamma\mathcal{S}(f)\Big\},$$

where $\mathcal{K}:\mathcal{H}_1\to\mathcal{H}_2$ and $\mathcal{S}:\mathcal{H}_1\to\mathbf{R}_+$ are functionals.

Let $\mathcal{S}(f) = \|f\|^2_{\mathcal{H}_1}$ and $\mathcal{K}$ be a bounded linear map with dense range. If $\mathcal{K}f_0 = y_0$ and $\|y_\delta - y_0\|_{\mathcal{H}_2}\le\delta < \|y_\delta\|_{\mathcal{H}_2}$, then we take $\gamma = \gamma_\delta > 0$ such that

$$\|\mathcal{K}f_{y_\delta,\gamma_\delta} - y_\delta\|_{\mathcal{H}_2} = \delta.$$

Conclusion. 1. $f_{y_\delta,\gamma_\delta}\to f_0$ as $\delta\to 0$.

2. If $f_0\in \text{range}\mathcal{K}^*$, then $\|f_{y_\delta,\gamma_\delta} - f_0\|_{\mathcal{H}_1} = O(\delta^{1/2})$.

The Tikhonov regularization scheme with the loss function V:

$\arg\min_{f\in\overline{\mathcal{H}}_K}\Big\{\int_Z V(y,f(x))d\rho + \frac{1}{2C}\|f^*\|^2_K\Big\}$

$= \arg\min_{f\in\overline{\mathcal{H}}_K}\Big\{\int_X\int_Y |y-f(x)|\chi_{\{yf(x)\le\}}d\rho(y|x)d\rho_X + \frac{1}{2C}\|f^*\|^2_K\Big\}.$

For the square loss, the scheme is similar to the scheme in inverse problems:

$$\arg\min_{f\in\mathcal{H}_K}\Big\{\int_X |f(x) - f_\rho(x)|^2 d\rho_X + \gamma\|f^*\|^2_K\Big\}.$$

· **ERM in learning theory**

· **Error** of f is $\mathcal{E}(f) := \int_Z V(y, f(x))d\rho$.

Regression function f_ρ^V the function minimizing the error $\mathcal{E}(f)$

· **Target function** $f_\mathcal{H}$: approximation of f_ρ^V

Let $\mathcal{H}$ be a compact subset of $C(X)$, and $f_\mathcal{H}$ minimizes the error:

$$f_\mathcal{H} = \arg\min_{f\in\mathcal{H}} \mathcal{E}(f)$$

· ρ is unknown! Hence f_ρ^V and $f_\mathcal{H}$ cannot be found.

· **Empirical target function** $f_\mathbf{z}$

Given $\mathbf{z} := (x_i, y_i)_{i=1}^m$, $f_\mathbf{z}$ minimizes the empirical error in $\mathcal{H}$:

$$f_\mathbf{z} := \arg\min_{f\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m V(y_i, f(x_i)).$$

$f_\mathcal{H} \approx f_\rho$ when $\mathcal{H}$ is large, and $f_\mathbf{z} \approx f_\mathcal{H}$ when m is large

· **Empirical target function in the RKHS**

Take $R > 0$, and $B_R := \{f \in \mathcal{H}_K : \|f\|_K \leq R\}$. Then we choose the hypothesis space $\mathcal{H}$ to be $\overline{I_K(B_R)}$, the closure in $C(X)$ of the image $I_K(B_R)$ of the ball B_R under the inclusion map: $I_K : \mathcal{H}_K \subset C(X)$. The empirical target function $f_\mathbf{z}$:

$$f_\mathbf{z} = \arg\min_{f\in\overline{I_K(B_R)}} \frac{1}{m}\sum_{i=1}^m V(y_i, f(x_i)).$$

ERM: $\mathcal{E}(f_\mathbf{z}) \to \mathcal{E}(f_\mathcal{H})$ with confidence as $m \to \infty$

Theorem 1. *For any function* $f : X \to \mathbf{R}$,

(1) $\mathcal{E}(f) \geq \mathcal{E}(f_c)$, i.e., $f_c = f_\rho^V$ (Y. Lin)

(2) $\mathcal{R}(\mathrm{sgn}(f)) - \mathcal{R}(f_c) \leq \mathcal{E}(f) - \mathcal{E}(f_c)$ (T. Zhang)

(3) $\mathcal{E}(f) - \mathcal{E}(f_c) \leq \int_X |f(x) - f_c(x)| d\rho_X$

In particular, $\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c) \leq \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_c)$

· **Main difficulty**: bounding the **offset** b

Lemma. *For any* $C > 0, m \in \mathbf{N}$ *and* $\mathbf{z} \in Z^m$, *one minimizer* $f_{\mathbf{z}}$ *satisfies*

$$\min_{1 \leq i \leq m} |f_{\mathbf{z}}(x_i)| \leq 1.$$

Hence $|b_{\mathbf{z}}| \leq 1 + \sqrt{2\|K\|_\infty C}$ *from* $\|f_{\mathbf{z}}^*\|_K \leq \sqrt{2C}$.

Theorem 2. *(Wu-Zhou) Let* $C > 0, m \in \mathbf{N}$, *and* $\mathbf{z}$ *be samples independently drawn according to* ρ. *Then for every* $\varepsilon > 0$, *with confidence at least*

$$1 - \frac{64(1 + \sqrt{2\|K\|_\infty C})}{\varepsilon} \mathcal{N}\left(\overline{I_K(B_{\sqrt{2C}})}, \frac{\varepsilon}{16}\right) \exp\left\{-\frac{m\varepsilon^2}{128(1 + \sqrt{2\|K\|_\infty C})^2}\right\}$$

we have

$$\begin{aligned} \mathcal{R}(sgn(f_{\mathbf{z}})) - \mathcal{R}(f_c) \leq \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_c) &\leq \varepsilon + \inf_{f \in \overline{\mathcal{H}}_K} \left\{\mathcal{E}(f) - \mathcal{E}(f_c) + \tfrac{1}{2C}\|f^*\|_K^2\right\} \\ &\leq \varepsilon + \inf_{f \in \overline{\mathcal{H}}_K} \left\{\|f - f_c\|_{L^1_{\rho_X}} + \tfrac{1}{2C}\|f^*\|_K^2\right\}. \end{aligned}$$

A K-functional represents the last term:

$$K(f_c, \gamma) := \inf_{f \in \overline{\mathcal{H}}_K} \left\{\|f - f_c\|_{L^1_{\rho_X}} + \gamma\|f^*\|_K^2\right\}.$$

covering number $\mathcal{N}(\overline{I_K(B_R)}, \eta)$: the minimal integer $\ell \in \mathbf{N}$ such that there exist ℓ disks with radius η covering the set $\overline{I_K(B_R)}$ in $C(X)$.

· **Corollaries**

Corollary 1. *Let ρ be an arbitrary Borel probability measure on $X \times Y$. Then for any $\varepsilon > 0, \eta > 0$, we can find $d_0 \in \mathbf{N}$ such that for every $d \geq d_0$ there exist $C_d > 0$ and $m_d \in \mathbf{N}$ satisfying*

$$Prob_{\mathbf{z} \in Z^m}\{\mathcal{R}(sgn(f_{\mathbf{z}})) - \mathcal{R}(f_c) \leq \varepsilon\} \geq 1 - \eta, \qquad \forall\ m \geq m_d,$$

where $f_{\mathbf{z}}$ is defined with $K_d(x, y) = (1 + x \cdot y)^d$ and $C = C_d$.

Corollary 2. *If f_c lies in the closure of $\overline{\mathcal{H}}_K$ in $L^1_{\rho_X}$, then for every $\varepsilon > 0$ there is some $C_0 > 0$ such that for each $C > C_0$ there holds*

$$\lim_{m \to \infty} Prob_{\mathbf{z} \in Z^m}\{\mathcal{R}(sgn(f_{\mathbf{z}})) - \mathcal{R}(f_c) \leq \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_c) \leq \varepsilon\} = 1.$$

Example 1. *Let K be a Mercer kernel on $X = [0, 1]$:*

$$K(x, y) = \sum_{j \in J} a_j (x \cdot y)^j,$$

where J is a subset of $\mathbf{N}$, $a_j > 0$ for each $j \in J$, and $\sum_{j \in J} a_j < \infty$. Then K is not universal in the sense that $\mathcal{H}_K$ is not dense in $C(X)$. But if $\sum_{j \in J} \frac{1}{j} = \infty$, then $\overline{\mathcal{H}}_K$ is dense in $C(X)$. Hence for an arbitrary Borel probability measure ρ on $X \times Y$, and any $\varepsilon > 0, \eta > 0$, there exist $C > 0$ and $m_0 \in \mathbf{N}$ satisfying

$$Prob_{\mathbf{z} \in Z^m}\{\mathcal{R}(sgn(f_{\mathbf{z}})) - \mathcal{R}(f_c) \leq \varepsilon\} \geq 1 - \eta, \qquad \forall\ m \geq m_0.$$

· **The separable case:** reducing ε^2 to ε

Definition. *The probability distribution ρ is* **strictly separable** *with margin $\gamma > 0$ by $\overline{\mathcal{H}}_K$ if there is a function $f_p = f_p^* + b_p \in \overline{\mathcal{H}}_K$ such that*

$$\|f_p^*\|_K = 1, \quad \textit{and} \quad y f_p(x) \geq \gamma \quad \textit{almost everywhere.}$$

Then $V(y, \frac{1}{\gamma} f_p(x)) = 0$, hence $\mathcal{E}(f_c) \leq \mathcal{E}(\frac{1}{\gamma} f_p) = 0$, and $\mathcal{R}(f_c) = 0$.

Theorem 3. *Let $C > 0, m \in \mathbf{N}$. If ρ is strictly separable with margin $\gamma > 0$ by $\overline{\mathcal{H}}_K$, then for every $\varepsilon > 0$, with confidence at least*

$$1 - \frac{32(1 + \sqrt{\|K\|_\infty}/\gamma)}{\varepsilon} \mathcal{N}\Big(\overline{I_K(B_{1/\gamma})}, \frac{\varepsilon}{16}\Big) \exp\Big\{-\frac{m\varepsilon}{172(1 + \sqrt{\|K\|_\infty}/\gamma)}\Big\}$$

we have

$$\mathcal{R}(sgn(f_{\mathbf{z}})) \leq \mathcal{E}(f_{\mathbf{z}}) \leq \varepsilon + \frac{1}{\gamma C}.$$

One may take $C = \infty$ in the separable case

Proof follows the idea of Cucker-Smale (Theorem C*), Barron, Lee-Bartlett-Williamson for square loss with convex hypothesis space

Example 2. *(Gaussian kernel in the separable case)*

$$K(x, y) = \exp\{-\frac{|x - y|^2}{\sigma^2}\}, \qquad x, y \in X = [0, 1]^n.$$

If ρ is strictly separable with margin $\gamma > 0$, then for any $\delta > 0$, with confidence at least $1 - \delta$, we have

$$\mathcal{R}(sgn(f_{\mathbf{z}})) = O\Big(\frac{(\log m)^{n+1}}{m\gamma} + \frac{\delta}{\gamma}(\log \frac{1}{\delta})^{n+1}\Big).$$

· **Gaussian kernel example** (in the general case)

Example 3. *Let $X = [0,1]^n$, $\sigma > 0, n/2 > s > 0$ and K be the Gaussian kernel $K(x,y) = \exp\{-\frac{|x-y|^2}{\sigma^2}\}$. Assume $\frac{d\rho_X(x)}{dx} \le C_0$ for almost every $x \in X$. If f_c is the restriction of some function $\widetilde{f}_c \in H^s(\mathbf{R}^n)$ onto X, then for every $\varepsilon > 0$ and $C \ge \exp\{180n^2/\sigma^2 + 22n + 6\}/512$, we have*

$$\begin{aligned} &Prob_{\mathbf{z}\in Z^m}\Big\{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_c) \le c_{s,n}(1+C_0)(\log C)^{-s/4}\Big\} \\ &\qquad \ge 1 - c_{s,n}C\exp\Big\{(\log C)^{n+1} - \tfrac{m}{C(\log C)^{s/2}}\Big\}, \end{aligned}$$

where $c_{s,n}$ is a constant depending on s and n.

Two tools: Approximation error and covering number

f_c is not continuous in general, and $\int_X |f(x)-f_c(x)|d\rho_X \le \|f-f_c\|_{L^2_{\rho_X}}$

· **Approximation Error**: Poggio-Girosi, Smale-Zhou, ...

Example 4. *Let $\sigma > 0$ and $K(x,y) = \exp\{-\frac{|x-y|^2}{\sigma^2}\}$, $x, y \in X = [0,1]^n$. If $f_\rho \in H^s(\mathbf{R}^n)$, then*

$$I(f_\rho, R) := \|f_\rho - f_{\mathcal{H}}\|_2 = \inf_{\|f\|_{\mathcal{H}_K}\le R} \|f_\rho - f\|_2 \le C_{n,s,\sigma}\|f_\rho\|_{s,2}\Big(\frac{1}{\log R}\Big)^{s/4}.$$

Conversely, if $I(f_\rho, R) = O(R^{-\epsilon})$ for some $\epsilon > 0$, then $f_\rho \in C^\infty$.

· **Covering number** for the Gaussian kernel: (Zhou 2002, 2003)

$$C\Big(\log\frac{R}{\eta}\Big)^{n/2} \le \log\mathcal{N}\left(\overline{I_K(B_R)}, \eta\right) \le C'\left(\log\frac{R}{\eta}\right)^{n+1}$$

· **Uniform stability**: Bousquet-Elisseeff (2002)

The 1-norm soft margin classifier is not uniformly stable, because of the offset.

· **q-norm soft margin classifier** (Chen-Wu-Ying-Zhou)

· **Support vector machine regression** with ϵ-insensitive norm loss function:

converges to the medium function

· **Clustering algorithms**

· **Learning from subspaces** (Smale-Zhou)

$$f_{\mathbf{z},\gamma} := \arg\min_{f\in\mathcal{H}_{K,\overline{t}}} \left\{ \frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2 + \gamma\|f\|_K^2 \right\},$$

where for a discrete subset $\overline{t} = \{t_j\}_{j=1}^d$, $\mathcal{H}_{K,\overline{t}} = \text{span}\{K_{t_j}\}_{j=1}^d$.

Then $f_{\mathbf{z},\gamma} = \sum_{j=1}^d c_j K_{t_j}$ with $\{c_j\}_{j=1}^d$ satisfying

$$[K_{\overline{t},\overline{x}}K_{\overline{x},\overline{t}} + m\gamma K_{\overline{t},\overline{t}}]\ [c_j]_{j=1}^d = K_{\overline{t},\overline{x}}[y_i]_{i=1}^m.$$

The coefficient matrix has size $d\times d$, while the matrix $K_{\overline{x},\overline{x}} + m\gamma I$ has size $m\times m$.

For the general loss function, the scheme is

$$f_{\mathbf{z},\gamma} := \arg\min_{f\in\mathcal{H}_{K,\overline{t}}} \left\{ \frac{1}{m}\sum_{i=1}^{m} V(y_i, f(x_i)) + \gamma\|f\|_K^2 \right\}.$$

· **Linear programming SVM**: Niyogi-Girosi (1996), Wu, ...