数理解析研究所講究録 第2091巻 2018年 14-21

A general framework of SVM in HDLSS settings

Yugo Nakayama
Graduate School of Pure and Applied Sciences
University of Tsukuba

Kazuyoshi Yata
Institute of Mathematics
University of Tsukuba

Makoto Aoshima
Institute of Mathematics
University of Tsukuba

1 Introduction

High-dimension, low-sample-size (HDLSS) data situations occur in many areas of modern science such as genetic microarrays, medical imaging, text recognition, finance, chemometrics, and so on. Suppose we have independent and d-variate two populations, $\Pi_i,\ i=1,2$, having an unknown mean vector $\boldsymbol{\mu}_i$ and unknown covariance matrix $\boldsymbol{\Sigma}_i$ for each i. We have independent and identically distributed (i.i.d.) observations, $\boldsymbol{x}_{i1},\dots,\boldsymbol{x}_{in_i}$, from each Π_i. We assume $n_i\geq 2,\ i=1,2$. Let $\boldsymbol{x}_0$ be an observation vector of an individual belonging to one of the two populations. Let $N=n_1+n_2$. We assume $\boldsymbol{x}_0$ and $\boldsymbol{x}_{ij}$s are independent.

In this paper, we consider classification in the HDLSS context such as $d\to\infty$ while N is fixed. In the HDLSS context, Hall et al. [6], Marron et al. [8] and Qiao et al. [12] considered distance weighted classifiers. Hall et al. [7], Chan and Hall [5] and Aoshima and Yata [2] considered distance-based classifiers. In particular, Aoshima and Yata [2] gave the misclassification rate adjusted classifier for multiclass, high-dimensional data in which misclassification rates are no more than specified thresholds. On the other hand, Aoshima and Yata [1, 3] considered geometric classifiers based on a geometric representation of HDLSS data. Aoshima and Yata [4] considered quadratic classifiers in general and discussed asymptotic properties and optimality of the classifiers under high-dimension, non-sparse settings. For linear SVM in HDLSS settings, Hall et al. [6], Chan and Hall [5] and Qiao and Zhang [13] showed that the misclassification rates tend to zero as $d\to\infty$ under certain severe conditions. Nakayama et al. [9] investigated asymptotic properties of linear SVM for HDLSS data. They proposed a bias-corrected linear SVM and showed that it gives preferable performances compared to linear SVM. On the other hand, Nakayama et al. [10] investigated asymptotic properties of SVM with the Gaussian kernel for HDLSS data.

In this paper, we consider a general framework of SVM in the HDLSS context where $d \to \infty$ while N is fixed. In Section 2, we investigate asymptotic properties of SVM in the HDLSS. In Section 3, we give asymptotic properties of SVM for both the linear and the Gaussian kernels.

2 A general framework of SVM

In this section, we consider a general framework of SVM.

2.1 Setup of SVM

Since HDLSS data are mostly separable by a hyperplane, we consider the hard-margin SVM as follows:

$$y(\boldsymbol{x}) = \boldsymbol{w}^T\phi(\boldsymbol{x}) + b, \tag{1}$$

where $\phi(\cdot)$ is a feature map, $\boldsymbol{w}$ is a weight vector and b is an intercept term. Let us write that $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = (\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1n_1}, \boldsymbol{x}_{21}, \ldots, \boldsymbol{x}_{2n_2})$. Let $t_j = -1$ for $j = 1, \ldots, n_1$ and $t_j = 1$ for $j = n_1 + 1, \ldots, N$. By differentiating the Lagrangian formulation with respect to $\boldsymbol{w}$ and b, we obtain the following dual form:

$$L(\boldsymbol{\alpha}) = \sum_{j=1}^{N}\alpha_j - \frac{1}{2}\sum_{j=1}^{N}\sum_{j'=1}^{N}\alpha_j\alpha_{j'}t_jt_{j'}k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}),$$

where $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) = \phi(\boldsymbol{x}_j)^T\phi(\boldsymbol{x}_{j'})$ is a kernel function, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)^T$ and α_js are Lagrange multipliers such as $\boldsymbol{w} = \sum_{j=1}^{N}\alpha_j t_j\phi(\boldsymbol{x}_j)$. The optimization problem can be transformed into the following: $\underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\, L(\boldsymbol{\alpha})$ subject to

$$\alpha_j \geq 0, \; j = 1, \ldots, N, \;\text{ and }\; \sum_{j=1}^{N}\alpha_j t_j = 0. \tag{2}$$

Let us write that

$$\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_N)^T = \underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\, L(\boldsymbol{\alpha}) \;\text{ subject to (2).}$$

There exist some $\boldsymbol{x}_j$s satisfying that $t_j y(\boldsymbol{x}_j) = 1$ (i.e., $\hat{\alpha}_j \neq 0$). Such $\boldsymbol{x}_j$s are called the support vector. Let $\hat{S} = \{j | \hat{\alpha}_j \neq 0, \; j = 1, \ldots, N\}$ and $N_{\hat{S}} = \#\hat{S}$, where $\#A$ denotes the number of elements in a set A. The intercept term is given by $\hat{b} = N_{\hat{S}}^{-1} \sum_{j\in\hat{S}}\{t_j - \sum_{j'\in\hat{S}} \hat{\alpha}_{j'} t_{j'} k(\boldsymbol{x}_j, \boldsymbol{x}_{j'})\}$. Then, the classifier in (1) is defined by

$$\hat{y}(\boldsymbol{x}) = \sum_{j\in\hat{S}} \hat{\alpha}_j t_j k(\boldsymbol{x}, \boldsymbol{x}_j) + \hat{b}. \tag{3}$$

Finally, in SVM, one classifies $\boldsymbol{x}_0$ into Π_1 if $\hat{y}(\boldsymbol{x}_0) < 0$ and into Π_2 otherwise. See Vapnik [14] for the details. Let $e(i)$ denote the error rate of misclassifying an individual from Π_i into the other class for $i = 1, 2$. We claim that a classifier has consistency if

$$e(i) = o(1) \quad \text{as } d \to \infty \text{ for } i = 1, 2. \tag{4}$$

In this paper, we investigate the following typical kernels.

(I) The linear kernel: $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) = \boldsymbol{x}_j^T \boldsymbol{x}_{j'}$; and
(II) The Gaussian kernel: $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) = \exp(-\|\boldsymbol{x}_j - \boldsymbol{x}_{j'}\|^2/\gamma)$,

where $\gamma(> 0)$ is a scale parameter.

2.2 Asymptotic properties of SVM

First, we assume the following assumption as $d \to \infty$:

(A-i) $k(\boldsymbol{x}_{1j}, \boldsymbol{x}_{1j'}) = \beta_1 + o_P(\Delta)$ for all $1 \le j < j' \le n_1$;
$k(\boldsymbol{x}_{1j}, \boldsymbol{x}_{1j}) = \beta_2 + o_P(\Delta)$ for all $1 \le j \le n_1$;
$k(\boldsymbol{x}_{2j}, \boldsymbol{x}_{2j'}) = \beta_3 + o_P(\Delta)$ for all $1 \le j < j' \le n_2$;
$k(\boldsymbol{x}_{2j}, \boldsymbol{x}_{2j}) = \beta_4 + o_P(\Delta)$ for all $1 \le j \le n_2$; and
$k(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j'}) = \beta_5 + o_P(\Delta)$ for all $1 \le j \le n_1$, $1 \le j' \le n_2$;
$k(\boldsymbol{x}_0, \boldsymbol{x}_{ij}) = \beta_{2i-1} + o_P(\Delta)$ when $\boldsymbol{x}_0 \in \Pi_i$ for all $1 \le j \le n_i$ and $i = 1, 2$;
$k(\boldsymbol{x}_0, \boldsymbol{x}_{i'j}) = \beta_5 + o_P(\Delta)$ when $\boldsymbol{x}_0 \in \Pi_i$ for all $1 \le j \le n_{i'}$ and $i' \ne i$.

Here, β_l is a variable (which may depend on d) for $l = 1, \ldots, 5$ and $\Delta = \beta_1 + \beta_3 - 2\beta_5$, where $\Delta > 0$, $\beta_2 - \beta_1 \ge 0$ and $\beta_4 - \beta_3 \ge 0$.

We note that Δ is a distance between the two populations. For example, $\Delta = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$ when $k(\cdot, \cdot)$ is the linear kernel. See Section 3.1 for the details. Let $\eta_1 = \beta_2 - \beta_1$ and $\eta_2 = \beta_4 - \beta_3$. We note that $\sum_{j=1}^{n_1} \alpha_j = \sum_{j=n_1+1}^{N} \alpha_j$ $(= \alpha_\star$, say) under (2). Then, from Section 2 of Nakayama et al. [11], we have the following lemma.

Lemma 1 ([11]). *Under (2) and (A-i), it holds that as* $d \to \infty$

$$L(\boldsymbol{\alpha}) = 2\alpha_\star - \frac{\Delta}{2}\alpha_\star^2 - \frac{1}{2}\Big(\eta_1 \sum_{j=1}^{n_1} \alpha_j^2 + \eta_2 \sum_{j=n_1+1}^{N} \alpha_j^2\Big) + o_P(\Delta\alpha_\star^2).$$

We can claim that

$$\max_{\boldsymbol{\alpha}}\Big\{-\frac{1}{2}\Big(\eta_1\sum_{j=1}^{n_1}\alpha_j^2+\eta_2\sum_{j=n_1+1}^{N}\alpha_j^2\Big)\Big\}=-\frac{\alpha_\star^2}{2}(\eta_1/n_1+\eta_2/n_2)$$

when $\alpha_1=\cdots=\alpha_{n_1}=\alpha_\star/n_1$ and $\alpha_{n_1+1}=\cdots=\alpha_N=\alpha_\star/n_2$ under (2). Let $\Delta_*=\Delta+\eta_1/n_1+\eta_2/n_2$. We consider the following condition:

$$\liminf_{d\to\infty}\frac{\eta_i}{\Delta}>0 \quad \text{for } i=1,2. \tag{5}$$

Then, in a way similar to Section 2 of Nakayama et al. [9], from Lemma 1 it holds that

$$\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha})=-\frac{\Delta_*}{2}\Big(\alpha_\star-\frac{2+o_P(1)}{\Delta_*}\Big)^2\{1+o_P(1)\}+\frac{2+o_P(1)}{\Delta_*} \tag{6}$$

under (2), (5) and (A-i), so that $\alpha_\star\approx 2/\Delta_*$. Then, from (6), we have the following result.

Proposition 1 ([11]). *Let $\delta=\eta_1/n_1-\eta_2/n_2$. Assume (A-i) and (5). It holds that as $d\to\infty$*

$$\hat{\alpha}_j=\frac{2}{\Delta_* n_1}\{1+o_P(1)\} \quad \textit{for all } j=1,\dots,n_1; \quad \textit{and}$$
$$\hat{\alpha}_j=\frac{2}{\Delta_* n_2}\{1+o_P(1)\} \quad \textit{for all } j=n_1+1,\dots,N.$$

Furthermore, it holds that as $d\to\infty$

$$\hat{y}(\boldsymbol{x}_0)=\frac{\Delta}{\Delta_*}\Big((-1)^i+\frac{\delta}{\Delta}+o_P(1)\Big) \quad \textit{when } \boldsymbol{x}_0\in\Pi_i \textit{ for } i=1,2.$$

Now, we consider the following condition:

(C-i) $\displaystyle\limsup_{d\to\infty}\frac{|\delta|}{\Delta}<1.$

For the misclassification rates, from Section 2 of Nakayama et al. [11], we have the following results.

Theorem 1 ([11]). *Under (A-i) and (C-i), SVM (3) holds consistency (4).*

Corollary 1 ([11]). *Under (A-i), SVM (3) holds the following properties:*

$$e(1)=1+o(1) \quad \textit{and} \quad e(2)=o(1) \quad \textit{as } d\to\infty \tag{7}$$
$$\textit{if} \quad \liminf_{d\to\infty}\frac{\delta}{\Delta}>1; \quad \textit{and}$$
$$e(1)=o(1) \quad \textit{and} \quad e(2)=1+o(1) \quad \textit{as } d\to\infty \tag{8}$$
$$\textit{if} \quad \limsup_{d\to\infty}\frac{\delta}{\Delta}<-1.$$

For linear SVM, Nakayama et al. [9] showed consistency (4) and the results in Corollary 1. From Corollary 1, if $|\delta|$ is larger than Δ, SVM would give a bad performance. Nakayama et al. [11] proposed a robust SVM in HDLSS settings.

3 Asymptotic properties of SVM with kernel functions (I) or (II)

We assume that $\limsup_{d\to\infty}\|\boldsymbol{\mu}_i\|^2/d<\infty$ and $\mathrm{tr}(\boldsymbol{\Sigma}_i)/d\in(0,\infty)$ as $d\to\infty$ for $i=1,2$. Here, for a function, $f(\cdot)$, "$f(d)\in(0,\infty)$ as $d\to\infty$" implies $\liminf_{d\to\infty}f(d)>0$ and $\limsup_{d\to\infty}f(d)<\infty$. Similar to Aoshima and Yata [2], we assume the following assumption for Π_is as necessary:

(A-ii) Let $\boldsymbol{z}_{ij},\ j=1,\dots,n_i$, be i.i.d. random p_i-vectors having $E(\boldsymbol{z}_{ij})=\boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{z}_{ij})=\boldsymbol{I}_{p_i}$ for each $i\ (=1,2)$ and some p_i. Let $\boldsymbol{z}_{ij}=(z_{i1j},\dots,z_{ip_ij})^\top$ whose components satisfy that $\limsup_{d\to\infty}E(z_{irj}^4)<\infty$ for all r and

$$E(z_{irj}^2z_{isj}^2)=E(z_{irj}^2)E(z_{isj}^2)=1 \quad\text{and}\quad E(z_{irj}z_{isj}z_{itj}z_{iuj})=0$$

for all $r\neq s,t,u$. Then, the observations, $\boldsymbol{x}_{ij}$s, from each $\Pi_i\ (i=1,2)$ are given by $\boldsymbol{x}_{ij}=\boldsymbol{\Gamma}_i\boldsymbol{z}_{ij}+\boldsymbol{\mu}_i,\ j=1,\dots,n_i$, where $\boldsymbol{\Gamma}_i$ is a $d\times p_i$ matrix such that $\boldsymbol{\Gamma}_i\boldsymbol{\Gamma}_i^\top=\boldsymbol{\Sigma}_i$.

Note that z_{irj}s are i.i.d. as the standard normal distribution when the Π_is are Gaussian and $\boldsymbol{\Gamma}_i=\boldsymbol{H}_i\boldsymbol{\Lambda}_i^{1/2}$, where $\boldsymbol{\Lambda}_i=\mathrm{diag}(\lambda_{i(1)},\dots,\lambda_{i(d)})$ is a diagonal matrix of eigenvalues, $\lambda_{i(1)}\geq\cdots\geq\lambda_{i(d)}\geq0$, and $\boldsymbol{H}_i$ is an orthogonal matrix of the corresponding eigenvectors. Thus, (A-ii) naturally holds when the Π_is are Gaussian.

3.1 Linear kernel function (I)

We consider linear SVM (LSVM), that is, the classifier (3) having kernel function (I). We set $\beta_1=\|\boldsymbol{\mu}_1\|^2$, $\beta_2=\|\boldsymbol{\mu}_1\|^2+\mathrm{tr}(\boldsymbol{\Sigma}_1)$, $\beta_3=\|\boldsymbol{\mu}_2\|^2$, $\beta_4=\|\boldsymbol{\mu}_2\|^2+\mathrm{tr}(\boldsymbol{\Sigma}_2)$ and $\beta_5=\boldsymbol{\mu}_1^T\boldsymbol{\mu}_2$, so that

$$\Delta=\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|^2\ (=\Delta_{(I)},\ \text{say})\quad\text{and}\quad \eta_i=\mathrm{tr}(\boldsymbol{\Sigma}_i)\ (=\eta_{i(I)},\ \text{say})\quad\text{for } i=1,2.$$

We note that LSVM is invariant to linear transformations on the data set. Thus, in Section 3.1, we assume $\boldsymbol{\mu}_2=\boldsymbol{0}$ without loss of generality, so that $\beta_3=\beta_5=0$, $\beta_4=\eta_{2(I)}$ and $\Delta_{(I)}=\|\boldsymbol{\mu}_1\|^2$. In addition, we assume the following condition as $d\to\infty$:

(C-ii) $\dfrac{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}{\Delta_{(I)}^2}=o(1)$ for $i=1,2$.

Then, from Section 3 of Nakayama et al. [11], we have the following lemma.

Lemma 2 ([11]). *Assume (A-ii) and (C-ii). Then, the assumption (A-i) is met for kernel function (I).*

By combining Lemma 2 with Theorem 1 and Corollary 1, we have the following results.

Corollary 2. *For LSVM, one can claim that*

$$(4)\ \textit{holds if}\ \limsup_{d\to\infty}\frac{|\delta_{(I)}|}{\Delta_{(I)}}<1;\quad (7)\ \textit{holds if}\ \liminf_{d\to\infty}\frac{\delta_{(I)}}{\Delta_{(I)}}>1;\quad \textit{and}$$
$$(8)\ \textit{holds if}\ \limsup_{d\to\infty}\frac{\delta_{(I)}}{\Delta_{(I)}}<-1$$

under (A-ii) and (C-ii), where $\delta_{(I)}=\eta_{1(I)}/n_1-\eta_{2(I)}/n_2$.

Nakayama et al. [9] provided a bias correction of linear SVM (BC-LSVM). They compared BC-LSVM with LSVM both in numerical simulations and actual data analyses. They concluded that BC-LSVM gives adequate performances for HDLSS settings even when n_is are quite unbalanced.

3.2 Gaussian kernel function (II)

We consider Gaussian kernel SVM (GSVM), that is, the classifier (3) with kernel function (II). We set $\beta_1=\exp\{-2\mathrm{tr}(\boldsymbol{\Sigma}_1)/\gamma\}$ $(=\beta_{1(II)}$, say), $\beta_3=\exp\{-2\mathrm{tr}(\boldsymbol{\Sigma}_2)/\gamma\}$ $(=\beta_{3(II)}$, say), $\beta_2=\beta_4=1$, and $\beta_5=\exp[-\{\mathrm{tr}(\boldsymbol{\Sigma}_1)+\mathrm{tr}(\boldsymbol{\Sigma}_2)+\Delta_{(I)}\}/\gamma]$ $(=\beta_{5(II)}$, say), so that

$$\begin{aligned}\Delta=&\beta_{1(II)}+\beta_{3(II)}-2\beta_{5(II)}\ (=\Delta_{(II)},\ \text{say})\ \text{ and}\\ \eta_i=&1-\exp\big(-2\mathrm{tr}(\boldsymbol{\Sigma}_i)/\gamma\big)\ (=\eta_{i(II)},\ \text{say})\ \text{ for } i=1,2.\end{aligned}$$

We note that $\Delta_{(II)}>0$ when $\boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2$ or $\mathrm{tr}(\boldsymbol{\Sigma}_1)\neq\mathrm{tr}(\boldsymbol{\Sigma}_2)$. Let $\mathrm{tr}(\boldsymbol{\Sigma}_{\min})=\min_{i=1,2}\mathrm{tr}(\boldsymbol{\Sigma}_i)$ and $\psi=\exp\{-2\mathrm{tr}(\boldsymbol{\Sigma}_{\min})/\gamma\}$. We assume the following condition as $d\to\infty$:

$$\textbf{(C-iii)}\quad \frac{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)+\Delta_{(I)}\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\}^{1/2}}{\min\{\gamma^2\Delta_{(II)}^2/\psi^2,\ \gamma^2\}}=o(1)\ \text{for } i=1,2.$$

Then, from Section 3 of Nakayama et al. [11], we have the following lemma.

Lemma 3 ([11]). *Assume (A-ii) and (C-iii). Then, the assumption (A-i) is met for kernel function (II).*

By combining Lemma 3 with Theorem 1 and Corollary 1, we have the following results.

Corollary 3. *For GSVM, one can claim that*

$$(4)\ \textit{holds if}\ \limsup_{d\to\infty}\frac{|\delta_{(II)}|}{\Delta_{(II)}}<1;\quad (7)\ \textit{holds if}\ \liminf_{d\to\infty}\frac{\delta_{(II)}}{\Delta_{(II)}}>1;\quad \textit{and}$$
$$(8)\ \textit{holds if}\ \limsup_{d\to\infty}\frac{\delta_{(II)}}{\Delta_{(II)}}<-1$$

under (A-ii) and (C-iii), where $\delta_{(II)}=\eta_{1(II)}/n_1-\eta_{2(II)}/n_2$.

Nakayama et al. [11] provided a bias correction of GSVM (BC-GSVM). They compared BC-GSVM with GSVM both in numerical simulations and actual data analyses. They also discussed the choice of γ.

Acknowledgements
The research of the second author was partially supported by Grant-in-Aid for Young Scientists (B), Japan Society for the Promotion of Science (JSPS), under Contract Number 26800078. The research of the third author was partially supported by Grants-in-Aid for Scientific Research (A) and Challenging Research (Exploratory), JSPS, under Contract Numbers 15H01678 and 17K19956.

References

[1] M. Aoshima, K. Yata, Two-stage procedures for high-dimensional data, Sequential Anal. 30 (2011) 356–399.

[2] M. Aoshima, K. Yata, A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data, Ann. Inst. Statist. Math. 66 (2014) 983–1010.

[3] M. Aoshima, K. Yata, Geometric classifier for multiclass, high-dimensional data, Sequential Anal. 34 (2015) 279–294.

[4] M. Aoshima, K. Yata, High-dimensional quadratic classifiers in non-sparse settings, arXiv:1503.04549 (2015).

[5] Y.-B. Chan, P. Hall, Scale adjustments for classifiers in high-dimensional, low sample size settings, Biometrika 96 (2009) 469–478.

[6] P. Hall, J.S. Marron, A. Neeman, Geometric representation of high dimension, low sample size data, J. R. Statist. Soc. B 67 (2005) 427–444.

[7] P. Hall, Y. Pittelkow, M. Ghosh, Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes, J. R. Statist. Soc. B 70 (2008) 159–173.

[8] J.S. Marron, M.J. Todd, J, Ahn, Distance-weighted discrimination, J. Amer. Statist. Assoc. 102 (2007) 1267–1271.

[9] Y. Nakayama, K. Yata, M. Aoshima, Support vector machine and its bias correction in high-dimension, low-sample-size settings, J. Stat. Plan. Infer. 191 (2017) 88–100.

[10] Y. Nakayama, K. Yata, M. Aoshima, Asymptotic properties of support vector machines in HDLSS settings, RIMS Koukyuroku, 2047 (2017) 10–18.

[11] Y. Nakayama, K. Yata, M. Aoshima, Bias-corrected support vector machine with Gaussian kernel in high-dimension, low-sample-size settings, submitted (2018).

[12] X. Qiao, H.H. Zhang, Y. Liu, M.J. Todd, J.S. Marron, Weighted distance weighted discrimination and its asymptotic properties, J. Amer. Statist. Assoc. 105 (2010) 401–414.

[13] X. Qiao, L. Zhang, Flexible high-dimensional classification machines and their asymptotic properties, J. Mach. Learn. Res. 16 (2015) 1547–1572.

[14] V.N. Vapnik, The Nature of Statistical Learning Theory (second ed.), Springer, New York, 2000.