SVM versus Least Squares SVM

Jieping Ye
Department of Computer Science
and Engineering
Arizona State University
Tempe, AZ 85287

Tao Xiong
Department of Electrical and
Computer Engineering
University of Minnesota
Minneapolis, MN 55455

Abstract

We study the relationship between Support Vector Machines (SVM) and Least Squares SVM (LS-SVM). Our main result shows that under mild conditions, LS-SVM for binary-class classifications is equivalent to the hard margin SVM based on the well-known Mahalanobis distance measure. We further study the asymptotics of the hard margin SVM when the data dimensionality tends to infinity with a fixed sample size. Using recently developed theory on the asymptotics of the distribution of the eigenvalues of the covariance matrix, we show that under mild conditions, the equivalence result holds for the traditional Euclidean distance measure. These equivalence results are further extended to the multi-class case. Experimental results confirm the presented theoretical analysis.

1 Introduction

Support Vector Machines (SVM) [3, 4, 15, 17] have been shown to be effective for many classification problems. For binary-class classifications, SVM constructs an optimal separating hyperplane between the positive and negative classes with the maximal margin. It can be formulated as a quadratic programming problem involving inequality constraints. The Least Squares formulation of SVM, called LS-SVM was recently proposed [5, 16], which involves the equality constraints only. Hence, the solution is obtained by solving a system of linear equations. Efficient and scalable algorithms, such as those based on conjugate gradient can be applied to solve LS-SVM. Extensive empirical studies [5, 20] have shown that LS-SVM is comparable to SVM in terms of generalization performance. However, the underlying reason for their similarity is not well understood yet.

In this paper, we study the intrinsic relationship between linear SVM and linear LS-SVM under a specific circumstance. More specifically, we show that for binary-class classifications, LS-SVM with no regularization is equivalent to hard margin SVM based on the well-known Mahalanobis distance measure [10], called *Hard M-SVM*, under mild conditions. Mahalanobis distance is based on correlations between variables by which different patterns can be identified and analyzed. It differs from Euclidean distance in that it takes into account the correlations of the dataset and is scale-invariant, i.e., not dependent on the scale of measurements. The well-known Linear Discriminant Analysis (LDA) [10] is based on the Mahalanobis distance measure and is optimal when each class is Gaussian and has a common covariance matrix. The Mahalanobis distance measure has also been used in the Maxi-Min Margin Machine algorithm (M^4) [11] to improve the generalization performance of SVM and form a unified framework for SVM, LDA, and Minimax Probability Machine (MPM) [12]. More general Mahalanobis distance measures can be learned from the data (see [6, 18]).

We further study the asymptotics of hard margin SVM when the data dimensionality tends to infinity with a fixed sample size. Using recently developed asymptotics theory on the distribution of the eigenvalues of the covariance matrix [1, 9], we show that under mild conditions, the equivalence result between LS-SVM and the hard margin SVM holds for the traditional Euclidean distance measure. This implies that the hard margin SVM based on the Euclidean distance measure, called *Hard E-SVM*, may be comparable to LS-SVM for high-dimensional small sample size data.

We also consider the multi-class classification problems. For the multi-class case, the one-against-rest approach [14] is commonly applied, which combines k binary classifications, where k is the number of classes in the training dataset. We extend our equivalence result to the multi-class case. Specifically, we show

that using the one-against-rest approach, multi-class LS-SVM is equivalent to the multi-class Hard M-SVM under mild conditions. We have conducted our experimental studies using a collection of high-dimensional datasets, including microarray gene expression data and text documents. Experimental results confirm our theoretical analysis.

Notation $||\cdot||_2$ denotes the L_2 norm; $||\cdot||_S$ denotes the norm under the Mahalanobis distance measure; Hard E-SVM and Hard M-SVM refer to the hard margin SVM under the Euclidean and Mahalanobis distance measure, respectively.

1.1 An overview of SVM and LS-SVM

We are given a set of n training samples $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ is drawn from a domain $\mathcal{X}$ and each of the label y_i is an integer from $\mathcal{Y} = \{-1, 1\}$. The goal of the binary-class classification in SVM or LS-SVM is to learn a model that assigns the correct label to an unseen test sample. This can be thought of as learning a function $f : \mathcal{X} \to \mathcal{Y}$ which maps each instance x to an element y of $\mathcal{Y}$. Let S be the covariance matrix defined as follows:

$$S = \frac{1}{n}(X - ce^T)(X - ce^T)^T, \tag{1}$$

where $X = [x_1, x_2, \cdots, x_n]$ is the data matrix, c is the centroid of X and e is the vector of all ones.

Assuming the data is separable, the hard margin SVM looks for some hyperplane:

$$f(x) = (x, w) + b = 0,$$

which separates the positive from the negative examples [3], where w is the normal to the hyperplane, $(x, w) = x^T w$ is the inner product between x and w, and $|b|/||w||_2$ is the perpendicular distance from the hyperplane to the origin. For the linearly separable case, the hard margin SVM simply looks for the separating hyperplane with the largest margin. The optimal hyperplane is computed by minimizing $||w||_2$ subject to the constraint that

$$y_i\left((x_i, w) + b\right) \geq 1,$$

for all i. A test point x is assigned to the positive class, if $(w, x) + b > 0$, and to the negative class otherwise. The above formulation can be extended to deal with nonseparable data by introducing the slack variables and a tuning parameter $C > 0$ [3]. This is known as the soft margin SVM. The optimal value of the tuning parameter C is commonly estimated through cross-validation.

Least Squares SVM (LS-SVM) applies the linear model [16]:

$$f(x) = (x, w) + b,$$

where $w \in \mathbb{R}^d$ is the weight vector, and b is the bias of the linear model. w and b are estimated by minimizing the following objective function:

$$L(w, b) = \sum_{i=1}^{n} ||f(x_i) - y_i||_2^2 + C||w||_2^2, \tag{2}$$

where y is the vector of class labels, and $C > 0$ is the regularization parameter. Minimization of $L(w, b)$ leads to

$$\begin{aligned} w &= \frac{2n_1n_2}{n^2}\left(S + \frac{C}{n}I_d\right)^{-1}(c_1 - c_2), \\ b &= \frac{n_1 - n_2}{n} - c^T w, \end{aligned} \tag{3}$$

where c is the global centroid of the data, n_1 and n_2 denote the number of samples from the positive and negative classes, respectively, and c_1 and c_2 are the centroids of the positive and negative classes, respectively. If no regularization is applied, i.e., $C = 0$, the optimal solution to LS-SVM is given by

$$w = \frac{2n_1n_2}{n^2}S^+(c_1 - c_2), \tag{4}$$

where S^+ denotes the pseudo-inverse of S [7]. Similar to SVM, a test point x is assigned to the positive class, if $f(x) = (w, x) + b > 0$, and to the negative class otherwise. Note that the LS-SVM algorithm in [5, 16] is formulated in the dual space using the kernel trick. In this paper, we focus on the primal formulation for LS-SVM.

2 SVM versus LS-SVM for binary-class classifications

We study in this section the relationship between LS-SVM and SVM for binary-class classifications. Our main result in Theorem 2.2 below shows that under mild conditions, LS-SVM is equivalent to Hard M-SVM (hard margin SVM based on the Mahalanobis distance measure) for the binary-class case.

The inner product between two vectors u and v under the Mahalanobis distance measure is defined as $(u, v)_S = u^T S^{-1} v$, assuming S is nonsingular. In the general case where S may be singular, the inner product between u and v is given by

$$(u, v)_S = u^T S^+ v. \tag{5}$$

Note that when S is nonsingular, S^+ equals S^{-1}. The length of u is then given by

$$||u||_S = \sqrt{(u, u)_S} = \sqrt{u^T S^+ u}. \tag{6}$$

u and v are said to be orthogonal to each other, if and only if

$$(u, v)_S = u^T S^+ v = 0. \tag{7}$$

Next, we define two matrices B and W, which are closely related to the between-class and within-class scatter matrices in Linear Discriminant Analysis (LDA) [10]. The matrix B is defined as

$$B = n_1 n_2/n^2 (c_1 - c_2)(c_1 - c_2)^T, \tag{8}$$

where c_1 and c_2 are the centroids of the the positive and negative classes, respectively, as introduced in the last section. Define $W = S - B$. It can be verified that

$$\begin{aligned} W &= \frac{1}{n}\left((X_1 - c_1 e^T)(X_1 - c_1 e^T)^T \right. \\ &+ \left. (X_2 - c_2 e^T)(X_2 - c_2 e^T)^T\right), \end{aligned} \tag{9}$$

where X_1 and X_2 are the data matrices of the positive and negative classes, respectively, and e is the vector of all ones with an appropriate length.

We show in the following lemma that under the condition that $\{x_i\}_{i=1}^n$ are linearly independent, the vector $c_1 - c_2$ is orthogonal to the difference vector between each data point in the training set and its corresponding centroid.

Lemma 2.1 *Let c_1, c_2, X_1, X_2, and S be defined as above. Assume that $\{x_i\}_{i=1}^n$ are linearly independent. Then $(u - c_1, c_1 - c_2)_S = 0$, for any $u \in X_1$, and $(v - c_2, c_1 - c_2)_S = 0$, for any $v \in X_2$. That is, under the Mahalanobis distance measure, $c_1 - c_2$ is orthogonal to $u - c_1$ and $v - c_2$, for any $u \in X_1$ and $v \in X_2$.*

Proof Since $\{x_i\}_{i=1}^n$ are linearly independent, $\mathrm{rank}(S) = n-1$. Let $S = UDU^T$ be the Singular Value Decomposition (SVD) [7] of S, where $U \in \mathbb{R}^{d\times(n-1)}$ has orthonormal columns, and $D \in \mathbb{R}^{(n-1)\times(n-1)}$ is diagonal with positive diagonal entries. Then $S^+ = UD^{-1}U^T$ and $(S^+)^{1/2} = UD^{-1/2}U^T$. It follows from $S = B + W$ that

$$\begin{aligned} I_{n-1} &= U^T (S^+)^{1/2} S (S^+)^{1/2} U \\ &= U^T (S^+)^{1/2} (B + W)(S^+)^{1/2} U \\ &= \tilde{B} + \tilde{W}, \end{aligned} \tag{10}$$

where

$$\begin{aligned} \tilde{B} &= U^T (S^+)^{1/2} B (S^+)^{1/2} U = b_1 b_1^T, \\ b_1 &= \sqrt{\frac{n_1 n_2}{n^2}} U^T (S^+)^{1/2} (c_1 - c_2) \\ &= \sqrt{\frac{n_1 n_2}{n^2}} D^{-1/2} U^T (c_1 - c_2), \\ \tilde{W} &= U^T (S^+)^{1/2} W (S^+)^{1/2} U \\ &= D^{-1/2} U^T W U D^{-1/2}. \end{aligned}$$

Thus,

$$b_1^T \tilde{W} b_1 = \frac{n_1 n_2}{n^2} (c_1 - c_2)^T S^+ W S^+ (c_1 - c_2). \tag{11}$$

Construct $\hat{Q} \in \mathbb{R}^{(n-1)\times(n-2)}$ so that

$$Q = \left[b_1/||b_1||_2, \hat{Q}\right]$$

is orthogonal. That is, $\hat{Q}^T b_1 = 0$. From Eq. (10), we have

$$\begin{aligned} I_{n-1} &= Q^T I_{n-1} Q = Q^T (\tilde{B} + \tilde{W}) Q \\ &= \mathrm{diag}(||b_1||_2^2, 0, \cdots, 0) + Q^T \tilde{W} Q. \end{aligned}$$

That is,

$$Q^T \tilde{W} Q = \mathrm{diag}(1 - ||b_1||_2^2, 1, \cdots, 1). \tag{12}$$

Since $\{x_i\}_{i=1}^n$ are linearly independent, we have $\mathrm{rank}(S) = n-1$ and $\mathrm{rank}(B) = 1$. From the definition of W, $\mathrm{rank}(W) \le n-2$. Since $S = B + W$ and both B and W are positive semi-definite, we have $\mathrm{rank}(W) \ge \mathrm{rank}(S) - \mathrm{rank}(B) = n-2$. Thus, $\mathrm{rank}(W) = n-2$. It follows that $\mathrm{rank}(Q^T \tilde{W} Q) = \mathrm{rank}(W) = n-2$. From Eq. (12), we have $1 - ||b_1||_2^2 = 0$, since $Q^T \tilde{W} Q \in \mathbb{R}^{(n-1)\times(n-1)}$. From Eq. (11), the first diagonal entry of $Q^T \tilde{W} Q$ in Eq. (12) is given by

$$\begin{aligned} 0 &= b_1^T/||b_1||_2 \tilde{W} b_1/||b_1||_2 = b_1^T \tilde{W} b_1 \\ &= \frac{n_1 n_2}{n^2} (c_1 - c_2)^T S^+ W S^+ (c_1 - c_2). \end{aligned}$$

It follows from the definition of W in Eq. (9) that $(u - c_1, c_1 - c_2)_S = 0$, for any $u \in X_1$, and $(v - c_2, c_1 - c_2)_S = 0$, for any $v \in X_2$. □

We will show in the following theorem that the optimal normal vector to the maximal margin hyperplane in Hard M-SVM is in the direction of $c_1 - c_2$. It is based on the idea [2] that the optimal normal vector of the hard margin SVM is identical to that of the hyperplane bisecting closest points in the two convex hulls with vertices consisting of data points from X_1 and X_2, respectively.

Theorem 2.1 *Assume x and y lie in the convex hull of $\{x_i\}_{x_i \in X_1}$ and $\{x_j\}_{x_j \in X_2}$, respectively. That is, $x = \sum_{x_i \in X_1} \alpha_i x_i$ and $y = \sum_{x_j \in X_2} \beta_j x_j$, where $\alpha_i \ge 0$, for all i, $\beta_j \ge 0$, for all j, $\sum_i \alpha_i = 1$, and $\sum_j \beta_j = 1$. Assume that $\{x_i\}_{i=1}^n$ are linearly independent. Then $||x - y||_S \ge ||c_1 - c_2||_S$, and $||c_1 - c_2||_S$ is the largest margin between X_1 and X_2. Furthermore, the optimal normal vector w of Hard M-SVM is in the direction of $c_1 - c_2$.*

Proof Denote $\hat{x} = c_2 - c_1 + x$. We have $x - \hat{x} = c_1 - c_2$ and $y - \hat{x} = y - c_2 - (x - c_1)$. From Lemma 2.1, $(x_i - c_1, c_1 - c_2)_S = 0$ for all $x_i \in X_1$ and $(x_j - c_2, c_1 - c_2)_S = 0$ for all $x_j \in X_2$. Hence,

$$\begin{aligned} (x - c_1, c_1 - c_2)_S &= \sum_{x_i \in X_1} \alpha_i (x_i - c_1, c_1 - c_2)_S = 0, \\ (y - c_2, c_1 - c_2)_S &= \sum_{x_j \in X_2} \beta_j (x_j - c_2, c_1 - c_2)_S = 0. \end{aligned}$$

It follows that $(y-\hat{x}, x-\hat{x})_S = (y-c_2, c_1-c_2)_S - (x-c_1, c_1-c_2)_S = 0$. Thus

$$\begin{aligned} ||x-y||_S^2 &= (x-\hat{x}+\hat{x}-y, x-\hat{x}+\hat{x}-y)_S \\ &= (x-\hat{x}, x-\hat{x})_S + (\hat{x}-y, \hat{x}-y)_S \\ &\geq ||x-\hat{x}||_S^2 = ||c_1-c_2||_S^2. \end{aligned}$$

On the other hand, the maximal margin between X_1 and X_2 is no larger than the distance between c_1 and c_2, i.e., $||c_1-c_2||_S$, as c_1 and c_2 are the centroids of X_1 and X_2. Thus, under the Mahalanobis distance measure, the maximal margin is $||c_1-c_2||_S$. Consider the projection of all data points onto the direction c_1-c_2. The distance between the two projection points of x and y is given by $(x-y, c_1-c_2)_S/||c_1-c_2||_S = (x-c_1-y+c_2+c_1-c_2, c_1-c_2)_S/||c_1-c_2||_S = ||c_1-c_2||_S$. Thus, the optimal normal vector w is in the direction of c_1-c_2. □

Note that the Mahalanobis distance measure is essentially equivalent to the normalization of the data by $(S^+)^{1/2}$, that is, $x_i \rightarrow (S^+)^{1/2}x_i$. The projection of x_i onto the normal vector w under the Mahalanobis distance measure is given by

$$(x_i, w)_S = ((S^+)^{1/2}x_i, (S^+)^{1/2}w) = (x_i, S^+w).$$

Thus the projection of x_i onto the normal vector w under the Mahalanobis distance measure is equivalent to the projection of x_i onto the direction S^+w under the Euclidean distance measure. Since w is in the direction of c_1-c_2, S^+w is in the direction of $S^+(c_1-c_2)$, which coincides with the weight vector of LS-SVM in Eq. (4). We thus have the following main result:

Theorem 2.2 *Let c_1, c_2, X_1, X_2, and X be defined as above. Assume that the data points in X are linearly independent. Then LS-SVM under the Euclidean distance measure is equivalent to Hard M-SVM.*

Proof Let w^M be the normal vector of Hard M-SVM. From Theorem 2.1, the distance between any two data points from X_1 and X_2 after the projection onto the direction c_1-c_2 is $||c_1-c_2||_S$, which is the largest margin between X_1 and X_2. Thus, under the Mahalanobis distance measure,

$$y_i\left((x_i, w^M)_S + b^M\right) = 1,$$

for all i, that is, all data points lie on these two hyperplanes:

$$(x, w^M)_S + b^M = \pm 1.$$

Here both w^M and b^M need to be estimated. It follows that

$$(x_i, w^M)_S + b^M = 1,$$

for all $x_i \in X_1$, and

$$(x_j, w^M)_S + b^M = -1,$$

for all $x_j \in X_2$. Summing over all $x_i \in X_1 \bigcup X_2$, we have

$$\left(\sum_{i=1}^{n} x_i, w^M\right)_S + nb^M = n(c, w^M)_S + nb^M = n_1 - n_2$$

and

$$b^M = (n_1-n_2)/n - (c, w^M)_S.$$

Summing over all $x_i \in X_1$, we have

$$(c_1, w^M)_S + b^M = 1.$$

Similarly, we have

$$(c_2, w^M)_S + b^M = -1.$$

It follows that $(c_1-c_2, w^M)_S = 2$. From Theorem 2.1, w^M is in the direction of c_1-c_2, that is $w^M = (c_1-c_2)\alpha$, for some α. We have

$$\alpha = 2/||c_1-c_2||_S^2,$$

and

$$w^M = 2(c_1-c_2)/||c_1-c_2||_S^2.$$

Let w^{LS} be the weight vector of LS-SVM, which applies the decision function:

$$f(x) = x \cdot w^{LS} + b^{LS}.$$

From Section 1.1, the bias term b^{LS} in LS-SVM is

$$b^{LS} = (n_1-n_2)/n - c^T w^{LS},$$

and the normal vector is given by

$$w^{LS} = 2n_1n_2/n^2 S^+(c_1-c_2).$$

From Lemma 2.1, $||b_1||_2 = 1$, where

$$b_1 = \sqrt{\frac{n_1n_2}{n^2}} D^{-1/2}U^T(c_1-c_2).$$

Thus,

$$\begin{aligned} 1 &= ||b_1||_2^2 = b_1^T b_1 = \frac{n_1n_2}{n^2}(c_1-c_2)^T S^+(c_1-c_2) \\ &= \frac{n_1n_2}{n^2}||c_1-c_2||_S^2. \end{aligned}$$

We have $1/||c_1-c_2||_S^2 = \frac{n_1n_2}{n^2}$. It follows that

$$\begin{aligned} w^{LS} &= \frac{2n_1n_2}{n^2} S^+(c_1-c_2) \\ &= S^+ 2(c_1-c_2)/||c_1-c_2||_S^2 \\ &= S^+ w^M, \end{aligned}$$

and

$$\begin{aligned} b^{LS} &= (n_1 - n_2)/n - c^T w^{LS} \\ &= (n_1 - n_2)/n - c^T S^+ 2(c_1 - c_2)\frac{n_1 n_2}{n^2} \\ &= (n_1 - n_2)/n - (c, w^M)_{S^+} = b^M. \end{aligned}$$

Hence, the decision functon in LS-SVM: $f^{LS} = (x, w^{LS}) + b^{LS}$ is identical to the one in Hard M-SVM: $f^M = (x, w^M)_S + b^M$. Thus, LS-SVM is equivalent to Hard M-SVM. □

Theorem 2.2 above shows the equivalence relationship between LS-SVM and Hard M-SVM. However, the traditional SVM formulation is based on the Euclidean distance measure. We show in the following theorem that under a certain condition on S, the normal vector w of Hard E-SVM (hard margin SVM based on the Euclidean distance measure) is equivalent to that of Hard M-SVM (hard margin SVM based on the Mahalanobis distance measure), as summarized below.

Theorem 2.3 *Assume that the data points in X are linearly independent and that all nonzero eigenvalues of S are λ. Then the normal vector w to the maximal margin hyperplane in Hard E-SVM is in the direction of $c_1 - c_2$. Furthermore, Hard E-SVM is equivalent to Hard M-SVM.*

Proof Let

$$S = U\mathrm{diag}(\lambda, \cdots, \lambda)U^T = \lambda UU^T$$

be the SVD of S and let $\hat{U} \in \mathbb{R}^{d\times(d-n+1)}$ be the orthogonal complement of U. Since $\hat{U}$ lies in the null space of S and $S = B + W$, $\hat{U}$ also lies in the null space of B and W. That is,

$$\begin{aligned} \hat{U}^T(c_1 - c_2) &= 0, \quad \hat{U}^T(u - c_1) = 0, \\ \hat{U}^T(v - c_2) &= 0, \quad \hat{U}^T(u - v) = 0, \end{aligned}$$

for any $u \in X_1$ and $v \in X_2$. Since $S^+ = \lambda^{-1}UU^T$, we have

$$\begin{aligned} ||c_1 - c_2||_S^2 &= \lambda^{-1}(c_1 - c_2)^T UU^T(c_1 - c_2) \\ &= \lambda^{-1}||U^T(c_1 - c_2)||_2^2 \\ &= \lambda^{-1}||[U, \hat{U}]^T(c_1 - c_2)||_2^2 \\ &= \lambda^{-1}||c_1 - c_2||_2^2, \\ ||u - v||_S^2 &= \lambda^{-1}(u - v)^T UU^T(u - v) \\ &= \lambda^{-1}||U^T(u - v)||_2^2 \\ &= \lambda^{-1}||[U, \hat{U}]^T(u - v)||_2^2 \\ &= \lambda^{-1}||u - v||_2^2. \end{aligned}$$

From Theorem 2.1, we have $||u - v||_S \geq ||c_1 - c_2||_S$, for any $u \in X_1$ and $v \in X_2$. It follows that $||u - v||_2 \geq ||c_1 - c_2||_2$. Following similar arguments in Theorem 2.1, the maximal margin between X_1 and X_2 under the Euclidean distance measure is $||c_1 - c_2||_2$.

Consider the projection of all data points onto the direction, $c_1 - c_2$, under the Euclidean distance. Similar to the case of the Mahalanobis distance measure as in Lemma 2.1, the following orthogonality condition holds:

$$\begin{aligned} (u - c_1, c_1 - c_2) &= \Big([U, \hat{U}]^T(u - c_1), [U, \hat{U}]^T(c_1 - c_2)\Big) \\ &= \big(U^T(u - c_1), U^T(c_1 - c_2)\big) \\ &= (c_1 - c_2)^T UU^T(u - c_1) \\ &= \lambda(u - c_1, c_1 - c_2)_S = 0, \\ (v - c_2, c_1 - c_2) &= \Big([U, \hat{U}]^T(v - c_2), [U, \hat{U}]^T(c_1 - c_2)\Big) \\ &= \big(U^T(v - c_2), U^T(c_1 - c_2)\big) \\ &= (c_1 - c_2)^T UU^T(v - c_2) \\ &= \lambda(v - c_2, c_1 - c_2)_S = 0. \end{aligned}$$

It follows that the projection of $u - v$, for any $u \in X_1$ and $v \in X_2$, onto the direction, $c_1 - c_2$, under the Euclidean distance measure is

$$(u - v, c_1 - c_2)/||c_1 - c_2||_2 = ||c_1 - c_2||_2.$$

Thus, under the Euclidean distance measure, the optimal normal vector w of the hard margin SVM is in the direction of $c_1 - c_2$. Recall that the projection onto the normal vector in Hard M-SVM is equivalent to the projection onto the direction $S^+(c_1 - c_2)$ under the Euclidean distance measure. To show the equivalence between Hard E-SVM and Hard M-SVM, we need to show that $S^+(c_1 - c_2)$ is in the same direction as $c_1 - c_2$. Since $S^+ = \lambda^{-1}UU^T$, we have

$$\begin{aligned} S^+(c_1 - c_2) &= \lambda^{-1}UU^T(c_1 - c_2) \\ &= \lambda^{-1}(I_d - \hat{U}\hat{U}^T)(c_1 - c_2) \\ &= \lambda^{-1}(c_1 - c_2), \end{aligned}$$

where the second equality follows since $\hat{U}$ is the orthogonal complement of U, and the last equality follows since $\hat{U}^T(c_1 - c_2) = 0$. Thus $S^+(c_1 - c_2)$ is in the same direction as $c_1 - c_2$. □

The assumption in Theorem 2.3 above is unlikely to hold exactly for most real-world datasets. However, recent studies on the geometric representation of high-dimensional small sample size data [1, 9] show that under mild conditions, S approaches to a scaled identity matrix, when the data dimension d tends to infinity with a fixed sample size n. This makes all the eigenvalues of S have a common value. In other words, the data behave as if the underlying distribution is spherical. Thus, for high-dimensional small sample size data,

the condition in Theorem 2.3 is likely to be approximately satisfied. It is thus expected that the normal vector of Hard E-SVM is close to that of Hard M-SVM, which has been shown in Theorem 2.2 to be equivalent to the weight vector of LS-SVM. (see Section 4 for detailed empirical studies)

3 SVM versus LS-SVM for multi-class classifications

We study in this section the relationship between LS-SVM and SVM for multi-class classifications. In the multi-class case, each of the label y_i is an integer from $\mathcal{Y} = \{1, 2, \cdots, k\}$ with $k \geq 3$. Many different approaches have been proposed to solve multi-class SVM [14] and multi-class LS-SVM [5]. A common way to solve the multi-class problems in the context of SVM is to first build a set of k one-versus-rest binary classification models $\{f_1, \cdots, f_k\}$, use all of them to predict an instance x, and then based on the prediction of these classifiers, assign x to y^* given by

$$y^* = \text{argmax}_{i=1,\cdots,k}\{f_i(x)\}. \tag{13}$$

In learning the i-th binary classifier f_i, the i-th class is assigned to the positive class, while the rest of classes is assigned to the negative class. Let X_i be the data matrix of the i-th class and c_i be its mean, and $\hat{X}_i$ be the data matrix from all classes except the i-th class and $\hat{c}_i$ be its mean. Define

$$B_i = \frac{n_i \hat{n}_i}{n^2}(c_i - \hat{c}_i)(c_i - \hat{c}_i)^T. \tag{14}$$

It can be verified that

$$\begin{aligned} W_i &= S - B_i & (15)\\ &= \frac{1}{n}\Big((X_i - c_i(e^{(i)})^T)(X_i - c_i(e^{(i)})^T)^T + \\ &\quad (\hat{X}_i - \hat{c}_i(\hat{e}^{(i)})^T)(\hat{X}_i - \hat{c}_i(\hat{e}^{(i)})^T)^T\Big), & (16)\end{aligned}$$

where $e^{(i)}$ and $\hat{e}^{(i)}$ are vectors of all ones.

Assume that $\{x_i\}_{i=1}^n$ are linearly independent. It follows from Lemma 2.1 and Theorem 2.1 that $||c_i - \hat{c}_i||_S$ is the largest margin between X_i and $\hat{X}_i$, under the Mahalanobis distance measure with the optimal normal vector w given in the direction of $c_i - \hat{c}_i$, as summarized in the following theorem:

Theorem 3.1 *Let X_i, c_i, $\hat{X}_i$, and $\hat{c}_i$ be defined as above. Assume that $\{x_i\}_{i=1}^n$ are linearly independent. Then the normal vector w_i to the maximal margin hyperplane in Hard M-SVM between X_i and $\hat{X}_i$ is in the direction of $c_i - \hat{c}_i$. Furthermore, the multi-class LS-SVM is equivalent to the multi-class Hard M-SVM, when one-versus-rest approach is applied for the classification in both cases.*

Note that the key to the equivalence result in the above theorem is the use of a common covariance matrix for all pairs $(X_i, \hat{X}_i)$, as $X_i \bigcup \hat{X}_i = X$, for all i. The equivalence result may not hold when other approaches [13, 14] for multi-class classifications such as the one-against-one method are applied.

4 Experiments and discussions

We use two types of data in our empirical study: Microarray gene expression data (ALL, LEUKEMIA, and ALLAML3), and text documents (re0, re1, and tr41).

- The ALL dataset [19] is one that covers six subtypes of acute lymphoblastic leukemia (248 samples with 12558 dimensions). The LEUKEMIA dataset (72 samples with 7129 dimensions) comes from a study of gene expression in two types of acute leukemias: acute lymphoblastic leukemia (ALL) and acute myeloblastic leukemia (AML). This dataset was fist studied in the seminal paper of Golub et al. [8]. Golub et al. studied this problem to address the binary classification problem between the AML samples and the ALL samples. The ALL part of the dataset comes from two sample types, *B-cell* and *T-cell*. Thus it can also be treated as a three-class dataset, named ALLAML3.
- re0 and re1 are from *Reuters-21578* text categorization test collection Distribution 1.0[1]. re0 contains 320 documents with 2887 dimensions from 4 classes, and re1 contains 490 documents with 3759 dimensions from 5 classes. tr41 is derived from the TREC-5, TREC-6, and TREC-7 collections[2]. It contains 210 documents with 7455 dimensions from 7 classes.

We perform our comparative study by randomly splitting the data into training and test sets. The data is randomly partitioned into a training set consisting of two-thirds of the whole set and a test set consisting of one-third of the whole set. To give better estimation of accuracy, the splitting is repeated 30 times and the resulting accuracies are averaged. We compare the classification performance of Hard E-SVM, Hard M-SVM, and LS-SVM. The soft margin SVM based on the Euclidean distance, called Soft SVM and the least squares SVM with regularization, called rLS-SVM are also reported with both regularization parameters estimated through cross-validation. We also compare the weight vector in LS-SVM with the normals to the largest margin hyperplane in Hard E-SVM and Hard

[1] http://www.research.att.com/∼lewis
[2] http://trec.nist.gov

Table 1: Comparison of classification accuracy (%) and standard deviation (in parenthesis) of different algorithms.

	Datasets					
	ALL	LEUKEMIA	ALLAML3	re0	re1	tr41
Soft SVM[a]	96.98	97.69	95.97	85.77	94.34	96.25
	(1.74)	(2.58)	(4.60)	(2.68)	(1.56)	(2.43)
Hard E-SVM[a]	96.98	97.46	95.50	85.59	94.42	96.05
	(1.74)	(2.73)	(4.93)	(2.83)	(1.50)	(2.45)
Hard M-SVM[a]	97.33	97.00	93.83	85.31	94.28	96.19
	(1.27)	(3.15)	(5.13)	(2.87)	(1.65)	(2.44)
LS-SVM[b]	97.33	97.00	93.83	85.26	94.28	96.19
	(1.27)	(3.15)	(5.13)	(2.93)	(1.65)	(2.44)
rLS-SVM[b]	97.53	97.00	94.53	85.59	94.53	96.22
	(1.18)	(3.15)	(5.04)	(2.83)	(1.54)	(2.38)
Corr_1[c]	0.933	0.956	0.970	0.968	0.956	0.997
Corr_2[c]	1.000	1.000	1.000	0.998	1.000	1.000

[a]Soft SVM, Hard E-SVM, and Hard M-SVM refer to soft margin SVM based on the Euclidean distance measure, hard margin SVM based on the Euclidean distance measure, and hard margin SVM based on the Mahalanobis distance measure, respectively.

[b]LS-SVM and rLS-SVM refer to least squares SVM with and without regularization, respectively.

[c]Corr_1 refers to the average correlation between the normal vectors of Hard E-SVM and LS-SVM, while Corr_2 refers to the average correlation between the normal vectors of Hard M-SVM and LS-SVM.

M-SVM. Let w_i^E, w_i^M, and w_i^{LS} denote the normal vectors of Hard E-SVM and Hard M-SVM, and the weight vector of LS-SVM, respectively, when the i-th class, X_i, and the rest of classes, $\hat{X}_i$, are assigned to the positive and negative classes, respectively. For comparison, we define

$$\text{Corr}_1 = \frac{1}{k}\sum_{i=1}^{k}\frac{w_i^E \cdot w_i^{LS}}{||w_i^E||_2||w_i^{LS}||_2}, \quad (17)$$

$$\text{Corr}_2 = \frac{1}{k}\sum_{i=1}^{k}\frac{w_i^M \cdot w_i^{LS}}{||w_i^M||_2||w_i^{LS}||_2}. \quad (18)$$

That is, Corr_1 denotes the average correlation between the normals of Hard E-SVM and the weight vectors of LS-SVM, while Corr_2 denotes the average correlation between the normals of Hard M-SVM and the weight vectors of LS-SVM.

We can observe from Table 1 that LS-SVM and Hard M-SVM achieve the same classification performance in all cases except re0. We checked all datasets and found that the training data points are linearly independent in all cases except re0. However, the accuracy difference between LS-SVM and Hard M-SVM is small in re0. The empirical resuts confirm the theoretical analysis presented in Sections 2 and 3. This is further confirmed by the values of Corr_2, since a value of 1 implies that the normals of the hard margin SVM are equivalent to the weight vectors of LS-SVM. Overall, Hard E-SVM is comparable to Hard M-SVM. Note that the values of Corr_1 are about 0.96 in average, which implies that the normals of Hard E-SVM is close to that of Hard M-SVM, as discussed in Section 3. Soft SVM and rLS-SVM do not perform significantly better than Hard E-SVM and LS-SVM with no regularization. Overall, Soft SVM and rLS-SVM are comparable.

5 Conclusions

We examine in this paper the intrinsic relationship between SVM and LS-SVM. Our main result shows that when the data points are linearly independent, LS-SVM is equivalent to Hard M-SVM. We further study the asymptotics of SVM when the data dimensionality tends to infinity with a fixed sample size. Using recently developed theory on the asymptotics of the distribution of the eigenvalues of the covariance matrix, we show that the equivalence result between LS-SVM and hard margin SVM holds for the traditional Euclidean distance measure. These equivalence results can be further extended to the multi-class case, when one-against-rest approach is applied. Experimental results on a collection of high-dimensional datasets confirm the claimed theoretical results. Results also show that for high-dimensional data, soft margin SVM is comparable to the hard margin SVM based on either the Euclidean or the Mahalanobis distance measure. Our theoretical and empirical results give further insights into the nature of these two algorithms as well as their relationship.

The presented analysis can be directly extended to the feature space when the kernel trick is applied. We

have done preliminary studies on low-dimensional data using Gaussian kernels. Unlike the high-dimensional case, the regularization in Soft SVM and rLS-SVM are effective and may significantly improve the classification performance of Hard E-SVM and LS-SVM. Overall, Soft SVM and rLS-SVM are comparable, as has been observed in previous studies. We plan to explore this further in the future.

Acknowledgement

We thank the reviewers for their comments, which helped improve the paper significantly. This research is sponsored by the Center for Evolutionary Functional Genomics of the Biodesign Institute at the Arizona State University and by the National Science Foundation Grant IIS-0612069.

References

[1] J. Ahn, J.S. Marron, K.E. Muller, and Y.Y. Chi. The high dimension, low sample size geometric representation holds under mild conditions. *http://www.stat.uga.edu/∼jyahn/HDLSS_web.pdf*, preprint, 2005.

[2] J. C. Bennett and C. Campbel. Support vector machines: Hype or hallelujah? *SIGKDD Explorations*, 2(2):1–13, 2000.

[3] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[4] N. Cristianini and J.S. Taylor. *Support Vector Machines and other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[5] T.V. Gestel, J.A. K. Suykens, B. Baesens, S. Viaene, J. Vanthienen, G. Dedene, B. De Moor, and J. Vandewalle. Benchmarking least squares support vector machine classifiers. *Machine Learning*, 54(1):5–32, 2004.

[6] A. Globerson and S. Roweis. Metric learning by collapsing classes. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 451–458, 2005.

[7] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, MD, USA, third edition, 1996.

[8] T.R. Golub and et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[9] P. Hall, J.S. Marron, and A. Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society series B*, 67:427–444, 2005.

[10] T. Hastie, R. Tibshirani, and J.H. Friedman. *The elements of statistical learning : data mining, inference, and prediction*. Springer, 2001.

[11] K. Huang, H. Yang, I. King, and M. R. Lyu. Learning large margin classifiers locally and globally. In *ICML*, pages 401–408, 2004.

[12] G.R.G. Lanckriet, L.E. Ghaoui, C. Bhattacharyya, and M.I. Jordan. A robust minimax approach to classification. *Journal of Machine Learning Research*, pages 555–582, 2002.

[13] Y.F. Liu and X. Shen. Multicategory SVM and psi-learning-methodology and theory. *Journal of the American Statistical Association*, 101:500–509, 2006.

[14] R.M. Rifkin and A. Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, 2004.

[15] B. Schökopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.

[16] J. A. K. Suykens, T. V. Gestel, J. D. Brabanter, B. D. Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific Pub. Co., Singapore, 2002.

[17] V.Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[18] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2005.

[19] E.J. Yeoh and et al. Classification, subtype diswcovery, and prediction of outcome in pediatric lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1(2):133–143, 2002.

[20] P. Zhang and J. Peng. SVM vs regularized least squares classification. In *ICPR*, pages 176–179, 2004.