

# Singular Value Decomposition Machine Study

Yongxin Xi

January 12, 2008

## Contents

<b>1</b>	<b>Main Problem</b>	<b>2</b>
<b>2</b>	<b>Backgrounds of SVD and SVM</b>	<b>3</b>
2.1	SVD . . . . .	3
2.2	SVM: Separable Case . . . . .	3
2.3	SVM: Non-Separable Case . . . . .	3
<b>3</b>	<b>Problems Formulations</b>	<b>4</b>
3.1	Problem 1.a . . . . .	4
3.2	Problem 2.a . . . . .	5
3.3	Problem 1.b . . . . .	6
3.4	Problem 2.b . . . . .	6
<b>4</b>	<b>Claims and Proofs</b>	<b>6</b>
4.1	Claim 1 . . . . .	6
4.2	Claim 2 . . . . .	7
4.3	Claim 3 . . . . .	8
4.4	Claim 4 . . . . .	8
4.5	Claim 5 . . . . .	10

# 1 Main Problem

Given high  $m$ -dimensional data with  $n$  (relatively few) training examples, find a good classifier that generalize well to new data.

## 1. Learning Algorithms:

SVM, Naive Bayes, Artificial Neural Network,  $K$  nearest Neighbors, Decision Trees, etc.

## 2. Dimensionality Reduction:

SVD  $\equiv$  PCA, ICA, Compressive Sampling(random projection), FA(factor analysis), PP(projection pursuit), Multidimensional Scaling, Topologically Continuous Maps, etc.

## 3. Interesting Questions:

- How well do these methods work? Measure the performance of classifiers by some function  $f(m, n, SNR)$ , where  $SNR$  is a measure of the data quality.
- Does the following claim hold: None of these methods works well for  $m \gg n$  (reference needed)
- Does the following claim hold: With dimension reduction techniques, classifiers do better than just learning algorithms alone.(reference needed)
- If claim # 3 holds, what are the key things that the learning methods are missing?
- What are the different criteria of these dimension reduction methods?
- Is there a notion of label distance?
- If data  $X$  is linearly separable, and  $XU = \Sigma V$ , when is  $V$  linearly separable? If  $X$  is not, can  $V$  be linearly separable?

## 2 Backgrounds of SVD and SVM

### 2.1 SVD

For  $n$  points data  $X = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix} \in \mathbf{R}^{n \times m}$ , we want to reduce its dimension

down to  $\mathbf{R}^{n \times p}$ , where  $p \ll m$ , by minimizing the projection distance of the points to the span of new bases  $\{\mu_1, \mu_2, \dots, \mu_p\}$ .

The solution to the problem is:  $\mu_k = k^{th}$  max(e-vector of  $X^T X$ )

### 2.2 SVM: Separable Case

For two-category data  $X \in \mathbf{R}^{n \times m}$  with labels  $y \in \mathbf{R}^n$  and  $y_i \in \{-1, 1\}$ , if they are separable, we can always find  $\{q \in \mathbf{R}^n, b\}$  such that:

$$\begin{aligned} \forall y_i = -1, \quad q^T x_i + b &\leq -1 \\ \forall y_i = 1, \quad q^T x_i + b &\geq 1 \end{aligned}$$

The distance of points which lie on the hyperplane  $q^T x + b = 1$  to the separating hyperplane  $q^T x + b = 0$  is  $\frac{1}{\|q\|}$ . Therefore the distances of all the pairwise points from two classes are guaranteed to be larger than  $\frac{2}{\|q\|}$ . Thus we need to find the pair of hyperplanes which gives the maximum margin by minimizing  $\|q\|$ .

Separable linear SVM problem:

$$\begin{aligned} \min \quad & \frac{1}{2} \|q\|^2 \\ \text{S.t.} \quad & y_i(q^T x_i + b) \geq 1 \end{aligned}$$

### 2.3 SVM: Non-Separable Case

If the data is not separable, we can relax the constraints by introducing positive variables  $\xi_i, i = 1, \dots, n$ , which then become:

$$\begin{aligned} \forall y_i = -1, \quad q^T x_i + b &\leq -1 + \xi_i \\ \forall y_i = 1, \quad q^T x_i + b &\geq 1 - \xi_i \\ \xi_i &\geq 0 \end{aligned}$$

So we can transform the problem into:

$$\begin{aligned} \min \quad & \frac{1}{2} \|q\|^2 + C \sum \xi_i \\ \text{S.t.} \quad & y_i(q^T x_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned}$$

$\xi_i$  is a measure of the classification error of each point by using the separating hyperplane  $q^T x + b = 0$ . If  $\xi_i = 0$ , the point  $x_i$  is outside the margin of the correct class. If  $0 < \xi_i < 1$ , the point is still classified correctly but it stays between the margin and the separating hyperplane. If  $\xi_i > 1$ , it turns out to be a classification error. From this we know that the number of errors is bounded above by  $\sum \xi_i$ . Therefore, to reduce the classification errors, this term is also added to the objective function.

However, note that  $\sum \xi_i$  is not the error distance in the original space. The actual error distance is  $\frac{\sum \xi_i}{\|q\|}$ .

### 3 Problems Formulations

**The core idea of SVDM:**

For data with extremely high dimension, it is likely that it is noisy and redundancy-existed. The phenomena of classification might be guided by relatively few features what we call latent variables here. In such case, dimension reduction beforehand might do good to the classification. It is even better if we introduce criteria for performance of classifier into the dimension reduction procedure, that is, to do dimension reduction and classification concurrently. SVDM (Support Vector Decomposition Machine) belongs to this case.

In the following sections we formulate different problems related with SVDM, and analyze these classifiers.

#### 3.1 Problem 1.a

**Given:**

Data  $X = [x_1 \ x_2 \ \cdots \ x_n]$ ,  $x_i \in \mathbf{R}^m$ , and label vector  $y \in \mathbf{R}^n$

**Find:**

1.  $Z_{pxn} = [z_1 \ z_2 \ \cdots \ z_n] \quad z_i \in \mathbf{R}^p$

2.  $W_{m \times p} = [w_1 \ w_2 \ \cdots \ w_p]$  with  $\{w_i\}$  an ON basis
3.  $q \in \mathbf{R}^p$  and  $b$

Such that:

$$\arg \min_{q,b} E = \frac{1}{2} \|q\|^2 + C \sum_{i=1}^n h_i$$

S.t.

$$\begin{aligned} Z &= W^T X \\ h_i &\geq 0 \\ y_i(q^T z_i + b) &\geq 1 - h_i, \quad i = 1, 2, \dots, n \end{aligned}$$

### 3.2 Problem 2.a

Given the same problem as Problem 1, define:

$$E = \frac{1}{2} \|q\|^2 + C \sum_{i=1}^n h_i$$

Now find: 1)  $Z$ ; 2)  $W$ ; 3)  $q, b$ ; 4)  $h_i$  as in Problem 1  
Such that:

1.  $Z = W^T X$
2.  $h_i \geq 0$
3.  $y_i(q^T z_i + b) \geq 1 - h_i, \quad i = 1, 2, \dots, n$
4.  $\text{Min} \|X - \hat{X}\|_F^2 + \lambda E$ , where  $\hat{X} = W_{m \times p} Z_{p \times n}$

Note:

- Problem 1.a is a limiting case ( $\lambda \rightarrow \infty$ ) of Problem 2.
- Problem 2.b reduces to finding  $p$  principle components of SVD when  $\lambda \rightarrow 0$ .

### 3.3 Problem 1.b

Under two modifications of  $\|q\|^2$  in Problem 1.a, we have Problem 1.b:

$$\begin{aligned} \min \quad & \sum h_i \\ \text{S.t.} \quad & y_i(q^T z_i + b) \geq 1 - h_i \\ & h_i \geq 0 \\ & \|q\|^2 \leq 1 \end{aligned}$$

### 3.4 Problem 2.b

Under two modifications of  $\|q\|^2$  in Problem 2.a, we have Problem 2.b:

$$\begin{aligned} \min \quad & \|X - WZ\|_F^2 + \lambda \sum h_i \\ \text{S.t.} \quad & y_i(q^T z_i + b) \geq 1 - h_i \\ & h_i \geq 0 \\ & \|q\|^2 \leq 1 \end{aligned}$$

## 4 Claims and Proofs

### 4.1 Claim 1

Problem 1 has a solution.

**Law 4.1.1** *Continuous function on a compact set always achieve its minimum and maximum value on the set.*

**Proof:**

In Problem 1, a set of new basis vectors  $[w_1 w_2 \cdots w_p]$  is to be found such that the projected data will be best to solve the SVM problem. Once  $W$  is decided, by the convexity of the linear SVM problem, the solution to the problem is usually unique. So the problem can be expressed by the function  $E(W)$ . By law 4.1.1, we can show that problem 1 has a solution by showing:

1. Function  $E(W)$  is continuous.  
 $W$  is required to be orthonormal bases, so it is a subset of the Euclidean space. Therefore, it is continuous.

2. The domain of the function is bounded.

We need to show the set  $W, q, b, h$  is bounded. First,  $W$  is bounded by  $\|w_i\| = 1$ . Second, the data  $\{X, y\}$  is finite hence bounded, so  $Z$  is bounded and hence  $\{q, b\}$  is bounded. Lastly, since the data is finite, we can always find  $h_i$  large enough to satisfy the corresponding constraint.

3. The domain of the function is closed.

Suppose  $\{w_j, q_j, b, h_j\} \in F$  is a sequence of the feasible set  $F$ , and it achieves its limit:

$$\lim_{j \rightarrow \infty} (w_j, q_j, b, h_j) \rightarrow (w, q, b, h)$$

Then  $\{w, q, b, h\} \in F$  because the limit of  $W$  is also orthonormal.

4. The feasible set of the problem is nonempty.

Given the finite data and any  $\{W, q, b\}$ , we can always find a set of  $\{h_i\}$ s large enough to make the problem feasible. Therefore, the feasible set of the problem is nonempty.

## 4.2 Claim 2

Let  $\{W, q, b, h\}$  be a solution of Problem 1. Let  $R$  be a  $p \times p$  ON matrix, then  $\{WR, R^T q, b, h\}$  is also a solution of Problem 1.

**Proof:**

Let  $E_1 = \frac{1}{2}\|q\|^2 + C \sum h_i$  be the objective value of the solution.

Then,

$$\begin{aligned} E_2 &= \frac{1}{2}\|R^T q\|^2 + C \sum h_i \\ &= \frac{1}{2}q^T R R^T q + C \sum h_i \\ &= \frac{1}{2}\|q\|^2 + C \sum h_i \\ &= E_1 \end{aligned}$$

Therefore the proposed solution  $\{WR, R^T q, b, h\}$  would achieve the same objective value as the solution  $\{W, q, b, h\}$ . Now let us check the constraints of the new solution:

1.  $h_i \geq 0$
2. Let  $\tilde{Z} = (WR)^T X$ , then  $\tilde{z}_i = R^T W^T x_i$ . So:

$$\begin{aligned}
& y_i((R^T q)^T \tilde{z}_i + b) \\
&= y_i(q^T R \tilde{z}_i + b) \\
&= y_i(q^T R R^T W^T x_i + b) \\
&= y_i(q^T W^T x_i + b) \\
&= y_i(q^T z_i + b) \\
&\geq 1 - h_i
\end{aligned}$$

3.  $\{WR\}$  is also orthonormal.  
 $(WR)^T(WR) = R^T W^T W R = R^T R = I$

### 4.3 Claim 3

Subspace spanned by the columns of  $WR$  is the same as by columns of  $W$ , i.e.  $\text{spancols}(WR) = \text{spancols}(W)$

**Proof:**

We show  $\text{spancols}(WR) \subseteq \text{spancols}(W)$  and  $\text{spancols}(W) \subseteq \text{spancols}(WR)$ .

1. For  $\forall v \in \text{spancols}(WR)$ , let  $v = WR\alpha$ , where  $\alpha \in \mathbf{R}^p$ ,  $v = W(R\alpha) = W\beta \in \text{spancols}(W)$ . Therefore  $\text{spancols}(WR) \subseteq \text{spancols}(W)$ .
2. For  $\forall v \in \text{spancols}(W)$ , let  $v = W\alpha$ , where  $\alpha \in \mathbf{R}^p$ ,  $v = W\alpha = WRR^T\beta = WR\beta \in \text{spancols}(WR)$ . Therefore  $\text{spancols}(W) \subseteq \text{spancols}(WR)$ .

### 4.4 Claim 4

Problem 1.a is equivalent to solving a linear SVM problem, for both separable and non-separable case.

**Proof:**

Let us first discuss the separable SVM case. For a typical separable linear

SVM problem, we have  $n$  data points and  $n$  labels  $\{X \in \mathbf{R}^{m \times n}, y \in \mathbf{R}^n\}$ . Solving this problem is to find:

$$\begin{aligned} \min \quad & \frac{1}{2} \|q\|^2 \\ \text{S.t.} \quad & y_i(q^T x_i + b) \geq 1 \end{aligned}$$

Suppose  $\{q, b\}$  is the solution to the problem. We first show  $\{q, b\}$  is in the feasible set of Problem 1.a. Then we show that there does not exist  $\{\hat{q}, \hat{b}\}$  that could outdo  $\{q, b\}$  in problem 1.a. So every  $\{q, b\}$  (if the solution is not unique) turns out to be the solution to problem 1.a. Meanwhile, problem 1.a cannot have more solutions than linear SVM problem. Therefore the two problems are equivalent.

When  $\{q, b\}$  solves the SVM problem, we know that  $\tilde{q} = \frac{q}{\|q\|}$  is the normal to the separating hyperplane, along which the margin of points from different classes is maximized. The margin  $d = \tilde{q}^T(x_1 - x_2)$ , where  $x_1, x_2$  are the support vectors from each class. Now let  $\{W\}$  contain  $\tilde{q}$ , such that the subspace generated by  $\{W\}$  is parallel to  $\tilde{q}$ . The maximized margin of two classes in the subspace is also  $d$ . Therefore  $\{W, W^T q, b\}$  lies in the feasible set of problem 1.a.

To show this, let  $w_j = \tilde{q}$ , and  $\forall i \neq j, w_i \perp \tilde{q}$ .

The distance of the two support vectors in the new subspace is thereby

$$\begin{aligned} \tilde{d} &= (z_1 - z_2)^T (W^T q) \\ &= (x_1 - x_2)^T W W^T q \\ &= (x_1 - x_2)^T \begin{pmatrix} w_1 & \cdots & \tilde{q} & \cdots & w_p \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ \tilde{q} \\ \vdots \\ w_p \end{pmatrix} q \\ &= (x_1 - x_2)^T \begin{pmatrix} w_1 & \cdots & \tilde{q} & \cdots & w_p \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ \|q\| \\ \vdots \\ 0 \end{pmatrix} \\ &= (x_1 - x_2)^T q \end{aligned}$$

$$= d$$

Check the constraints:

$$\begin{aligned} y_i(q^T x_i + b) &= y_i((W^T q)^T z_i + b) \\ &= y_i(x_i^T W W^T q + b) \\ &= y_i(x_i^T q + b) \\ &\geq 1 \end{aligned}$$

Till now we have shown  $\{W, W^T q, b\}$  lies in the feasible set of problem 1.a. The next step is to show that  $\{W, W^T q, b\}$  is the actual solution.

If it is not the solution, there exists other feasible point which achieves a larger margin  $\hat{d} > d$ . However, this violates that assumption that the solution to the SVM problem has the margin  $d$ , because the projected distance can never be larger than the original distance. Therefore such point does not exist and hence  $\{W, W^T q, b\}$  is the actual solution to Problem 1.a.

The non-separable case can be proved by similar argument.

## 4.5 Claim 5

The actual error measure in the SVM problem should be  $\frac{\sum \xi_i}{\|q\|}$ .