

# Multisurface Proximal Support Vector Machine Classification via Generalized Eigenvalues

O. L. Mangasarian and E. W. Wild

Presented by: Jun Fang

# Outline

1. Conventional Support Vector Machine
2. Proximal Support Vector Machine
3. Multisurface Proximal Support Vector Machine
4. Simulation Results

# 1. Conventional SVM

We consider the problem of classifying  $m$  points  $\{\mathbf{x}_i\}$  in the  $n$  dimensional real space  $R^n$ , represented by the  $m \times n$  matrix  $A$ .

- Let  $A_+$  denote the points whose membership belongs to class 1;  $A_-$  denotes the points whose membership belongs to class 0.  $D$  is a diagonal matrix whose  $i^{\text{th}}$  diagonal entry is the label  $y_i$ .
- The conventional SVM with a linear kernel is given by the following quadratic programming problem with parameter  $c > 0$

$$\begin{aligned} & \text{minimize} && \frac{1}{2} \|\mathbf{w}\|^2 + c \mathbf{e}^T \rho && (1) \\ \text{s.t.} &&& D(A\mathbf{w} - b\mathbf{e}) \geq \mathbf{e} - \rho \\ &&& \rho_i \geq 0, \quad i = 1, \dots, m \end{aligned}$$

# 1. Conventional SVM (Cont.)

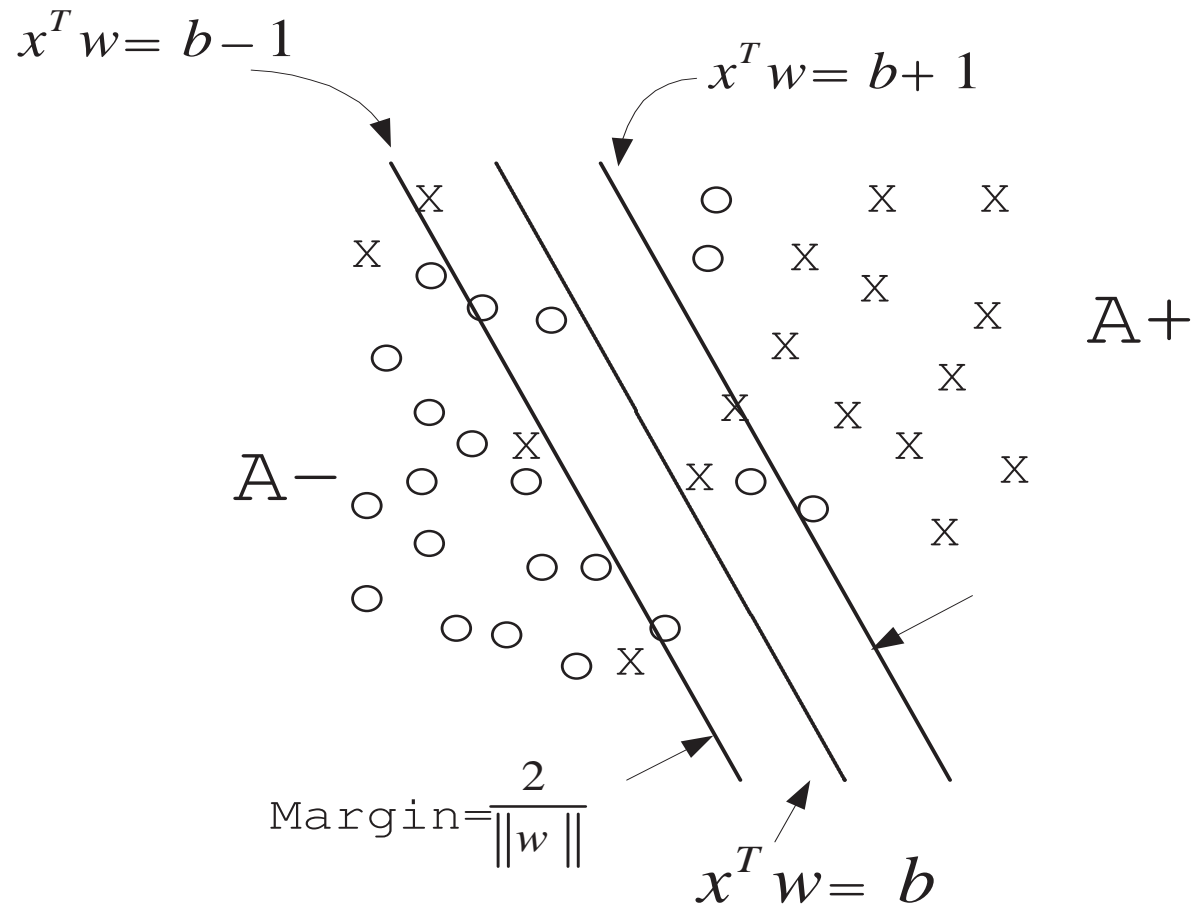


Figure 1: The conventional SVM classifier: two parallel bounding planes are devised to approximately separating  $A -$  from  $A +$ .

## 2. Proximal SVM

- A simple and fundamental change to the optimization problem (1) results in the proximal SVM. We replace the inequality constraint by an equality and replace  $\rho$  with  $\|\rho\|^2$ .
- The proximal SVM is given by the following optimization problem:

$$\begin{aligned} & \text{minimize} && \frac{1}{2} \|\mathbf{w}\|^2 + c \|\rho\|^2 && (2) \\ & \text{s.t.} && D(A\mathbf{w} - b\mathbf{e}) = \mathbf{e} - \rho \end{aligned}$$

- The planes are not bounding planes any more, but can be thought of as “proximal planes”, around which the points of each class are clustered, and the planes are pushed apart as far apart as possible.
- The proximal SVM optimization problem offers an analytical solution.

## 2. Proximal SVM (Cont.)

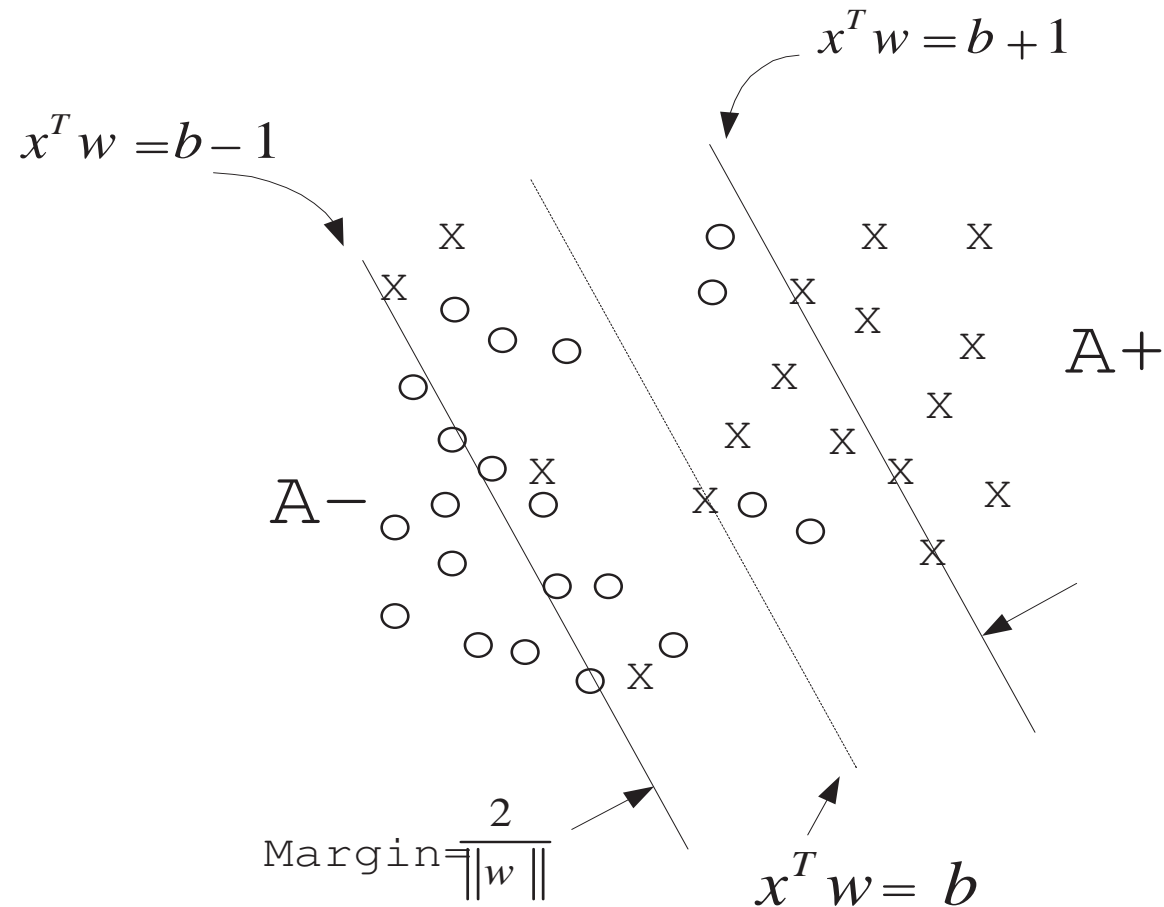


Figure 2: The proximal SVM classifier: the two parallel planes are devised so that around which points of the sets  $A^-$  and  $A^+$  cluster.

# 3. Multisurface Proximal SVM

- The multisurface proximal SVM drops the parallelism condition on the proximal planes and require that each plane be as close as possible to one of the data sets and as far as possible from the other one.
- The problem involves seeking two unparallel planes:  
 $\mathbf{x}^T \mathbf{w}_1 - b_1 = 0$  and  $\mathbf{x}^T \mathbf{w}_2 - b_2 = 0$ .
- To obtain the first plane, minimize the distances between the points in class 0 and the plane, at the same time maximize the distances between the points in class 1 and the plane. This leads to the following optimization problem:

$$\min_{\mathbf{w}_1, b_1} \frac{\|A^- \mathbf{w}_1 + b_1 \mathbf{e}\|^2 + c(\|\mathbf{w}_1\|^2 + b_1^2)}{\|A^+ \mathbf{w}_1 + b_1 \mathbf{e}\|^2} \quad (3)$$

### 3. Multisurface Proximal SVM (Cont.)

- Let  $G \triangleq [A^- \quad -\mathbf{e}]^T [A^- \quad -\mathbf{e}] + c\mathbf{I}$
- Let  $H \triangleq [A^+ \quad -\mathbf{e}]^T [A^- \quad -\mathbf{e}]$ ,  $\mathbf{z} \triangleq \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}$
- The optimization problem (3) becomes:

$$\min_{\mathbf{z} \neq \mathbf{0}} \frac{\mathbf{z}^T G \mathbf{z}}{\mathbf{z}^T H \mathbf{z}} \quad (4)$$

- The above objective function is known as Rayleigh quotient. Under the condition that  $H$  is positive definite, the Rayleigh quotient ranges over the interval  $[\lambda_1, \lambda_{n+1}]$  for normalized  $\mathbf{z}$ , where  $\lambda_1$  and  $\lambda_{n+1}$  are the minimum and maximum eigenvalues of the generalized eigenvalue problem:

$$G\mathbf{z} = \lambda H\mathbf{z} \quad (5)$$

### 3. Multisurface Proximal SVM (Cont.)

- The requirement of  $H$  to be positive definite means that the columns of matrix  $[A^+ \quad -\mathbf{e}]$  are linearly independent.
- The linear independence condition is not restrictive for many classification problems for which  $m_0 \gg n$  and  $m_1 \gg n$ , where  $m_0$  and  $m_1$  denote the number of data points belonging to class 0 and 1, respectively.
- The authors also claim that the linearly independent condition is a sufficient but not necessary condition.
- The second plane  $\mathbf{x}^T \mathbf{w}_2 - b_2 = 0$  can be obtained in a similar way by solving the following optimization problem:

$$\min_{\mathbf{w}_2, b_2} \frac{\|A^+ \mathbf{w}_2 + b_2 \mathbf{e}\|^2 + c(\|\mathbf{w}_2\|^2 + b_2^2)}{\|A^- \mathbf{w}_2 + b_2 \mathbf{e}\|^2} \quad (6)$$

### 3. Multisurface Proximal SVM (Cont.)

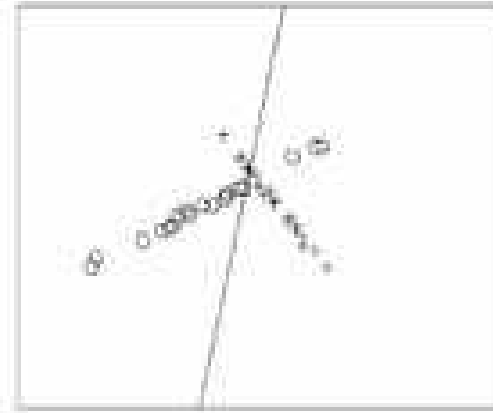
- The above results can be extended to nonlinear multisurface classifiers using kernels. A kernel-based nonlinear surface can be obtained by generalizing (3) to the following:

$$\min_{\mathbf{u}_1, b_1} \frac{\|K(A^-, A^T)\mathbf{u}_1 + b_1\mathbf{e}\|^2 + c(\|\mathbf{u}_1\|^2 + b_1^2)}{\|K(A^+, A^T)\mathbf{u}_1 + b_1\mathbf{e}\|^2} \quad (7)$$

- The optimization problem (7) can also be solved as (3).
- The multisurface proximal SVM has the following two advantages: first, it is more effective in solving the “cross-plane” problem (Fig. 3); second, for the linear kernel classifier, very large data sets can be handled by multisurface proximal SVM provided the input space dimension  $n$  is moderate in size.



GEP SVM: 100% correct



Linear PSVM: 80% correct

Figure 3: The “cross plane” learned by multisurface proximal SVM and linear proximal SVM respectively.

**TABLE 1**  
**Linear Kernel GEP-SVM, PSVM [7], and SVM-Light [9]**  
**10-Fold Testing Correctness and p-Values**

Data Set $m \times n$	GEP-SVM Correctness	PSVM Correctness p-value	SVM-Light Correctness p-value
Cross Planes $300 \times 7$	<b>98.0%</b>	55.3%* 5.24671e-07	45.7%* 1.4941e-08
NDC $300 \times 7$	86.7%	88.3% 0.244333	<b>89.0%</b> 0.241866
Cleveland Heart $297 \times 13$	81.8%	<b>85.2%</b> 0.112809	83.6% 0.485725
Cylinder Bands $540 \times 35$	71.3%	71.7% 0.930192	<b>76.1%</b> 0.229676
Pima Indians $768 \times 8$	73.6%	<b>75.9%</b> 0.274187	75.7% 0.380633
Spambase $4601 \times 57$	76.8%	<b>77.1%</b> 0.0654478	<b>77.1%</b> 0.0654478
Galaxy Bright $2462 \times 14$	<b>98.6%</b>	97.3%* 0.031226	98.3% 0.506412
Mushroom $8124 \times 22$	81.1%	80.9% 0.722754	<b>81.5%</b> 0.356003

*The p-values are from a t-test comparing each algorithm to GEP-SVM. Best correctness results are in bold. An asterisk (\*) denotes a significant difference from GEP-SVM based on p-values less than 0.05.*

**TABLE 2**  
**Nonlinear Kernel GEPSVM, PSVM [7], and SVM-Light [9]**  
**10-Fold Testing Correctness and p-Values**

Data Set $m \times n$	GEPSVM Correctness	PSVM Correctness p-value	SVM-Light Correctness p-value
Cross Planes $300 \times 7$	<b>99.0%</b>	73.7%* 0.00025868	79.3%* 8.74044e-06
WPBC (60 mo.) $110 \times 32$	62.7%	<b>64.5%</b> 0.735302	63.6% 0.840228
BUPA Liver $345 \times 6$	63.8%	67.9% 0.190774	<b>69.9%</b> 0.119676
Votes $435 \times 16$	94.2%	94.7% 0.443332	<b>95.6%</b> 0.115748
Haberman's Survival $306 \times 3$	75.4%	<b>75.8%</b> 0.845761	71.7% 0.0571092

*The p-values were calculated using a t-test comparing each algorithm to GEPSVM. Best results are in bold. An asterisk (\*) denotes a significant difference from GEPSVM based on p-values less than 0.05.*

**TABLE 3**  
**Average Time to Learn One Linear Kernel GEP SVM, PSVM [7],  
and SVM-Light [9] on the Cylinder Bands Data Set [19]**

GEP SVM Time (seconds)	PSVM Time (seconds)	SVM-Light Time (seconds)
0.96	0.08	75.4