GLA UNIVERSITY
MATHURA
Recognised by UGC Under Section 2(f)
Accredited with A Grade by NAAC

MACHINE LEARNING (ML-15)

Dr. NEERAJ GUPTA, Department of CEA, GLA University, Mathura

AGENDA

- Support Vector Machine
- Types of SVM
- Hyperplane
- Support Vectors
- Linear SVM Mathematically
- Examples
- Pros and Cons

INTRODUCTION

- Support Vector Machine abbreviated as SVM
- It can be used for both regression and classification tasks.
- It is widely used in classification objectives.

Logistic Regression

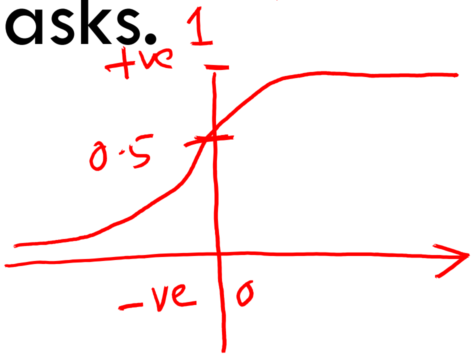
$0 \le h_\theta(x) \le 1$

$h_\theta(x) = g(\theta^T x)$

↑

Sigmoid

$$g(z) = \frac{1}{1+e^{-z}}$$



INTRODUCTION

- SVM algorithm can be used for Face detection, image classification, text categorization, etc.

Facial Expression Classification using SVM

Text Classification using SVM

It's supposed to be automatic, but actually you have to push this button.

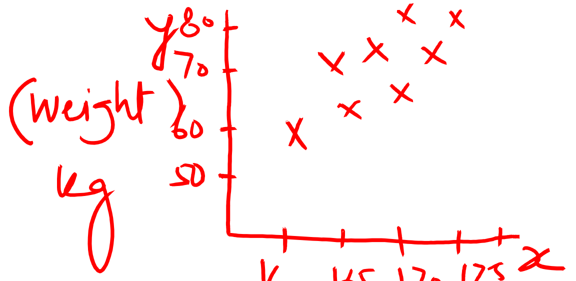
Vs

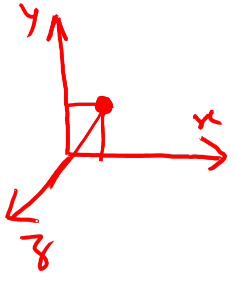
It's supposed to be automatic, but actually you have to push this button.

Human Handwriting

Binary

Computer Alphabets





(height)
cm

WHAT IS SUPPORT VECTOR MACHINE?

Superuised

160 cm
60 kg

- A **support vector machine** is a **machine** learning model that is able to generalize between two different classes if the set of labelled data is provided in the training set to the algorithm.

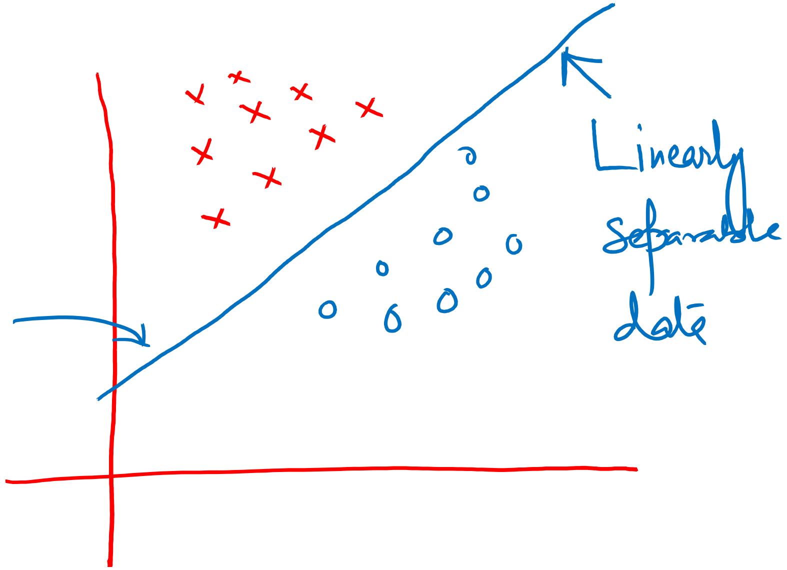
- The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space

- (N—the number of features) that distinctly classifies the data points.

height, weight → 2 dim

n feature → n-dim.

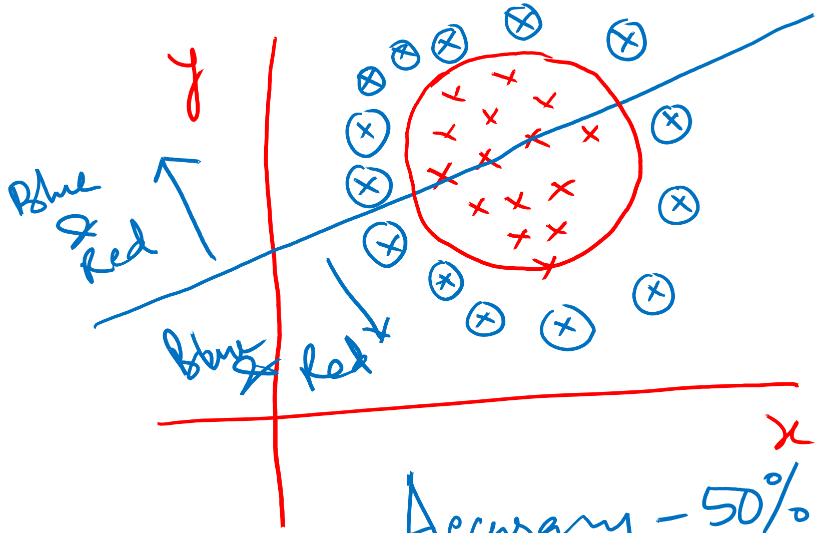
TYPES OF SVM



1. **Linear SVM:**

- Linear SVM is used for linearly separable data.
- It means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data.
- The classifier is used called as Linear SVM classifier.

TYPES OF SVM



Accuracy – 50%

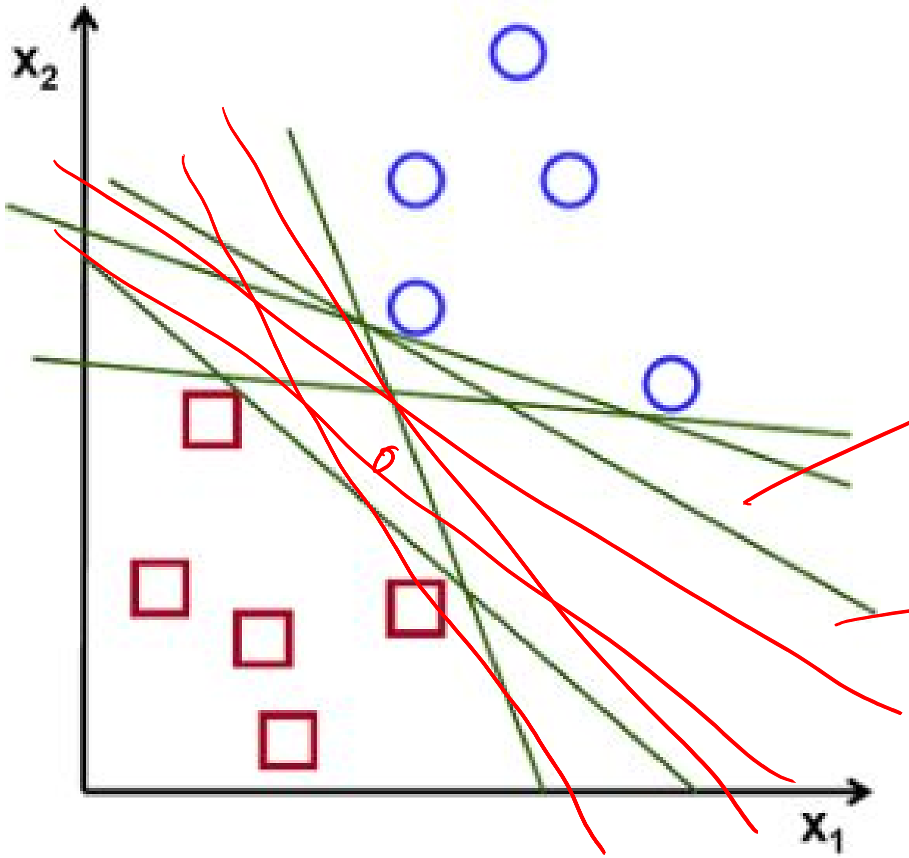
Non Linear SVM

2. Non-linear SVM:

- Non-Linear SVM is used for non-linearly separated data.
- It means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data.
- The classifier used is called as Non-linear SVM classifier.

SUPPORT VECTOR MACHINE

LSVM

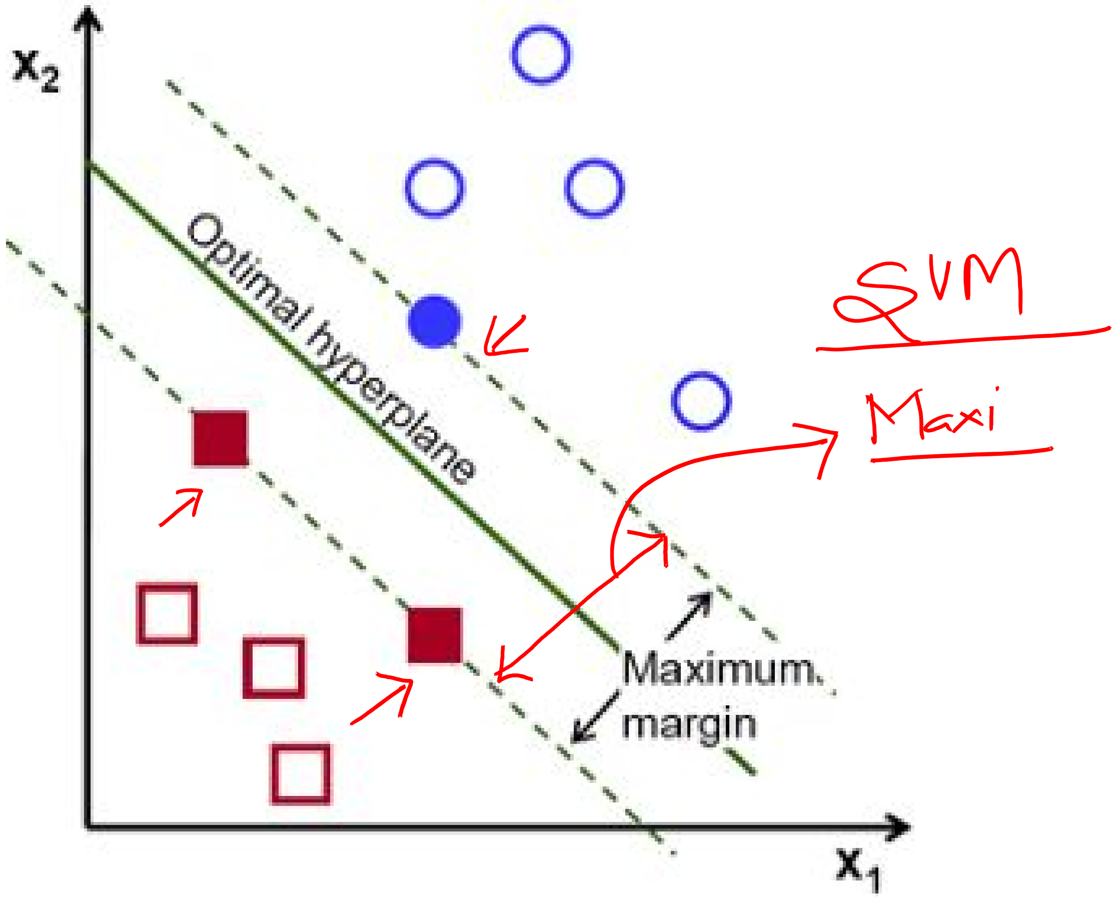


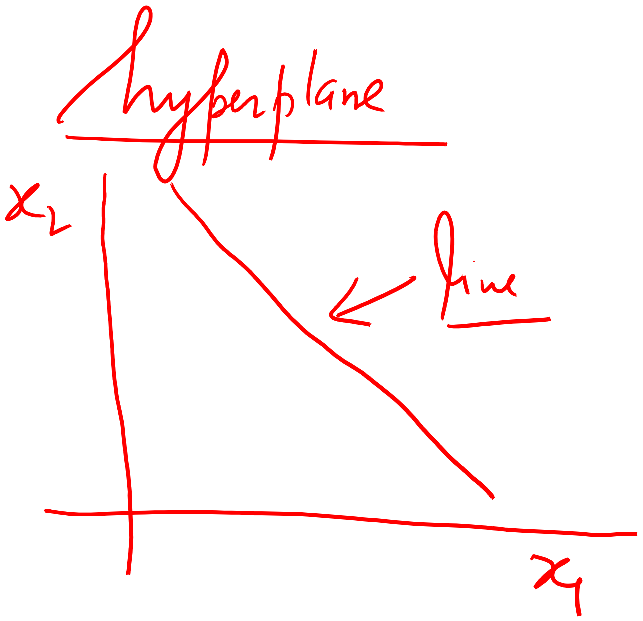
Linear
or
Non Linear

optimal

Infinite no. of lines

SUPPORT VECTOR MACHINE





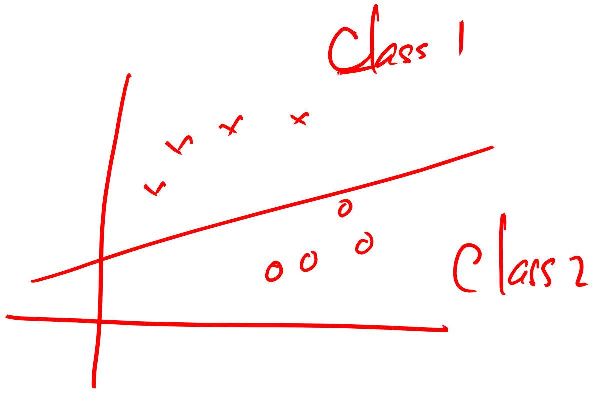
SUPPORT VECTOR MACHINE

- To separate the two classes of data points, there are many possible hyperplanes that could be chosen.

- ✓ The objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes.

- Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

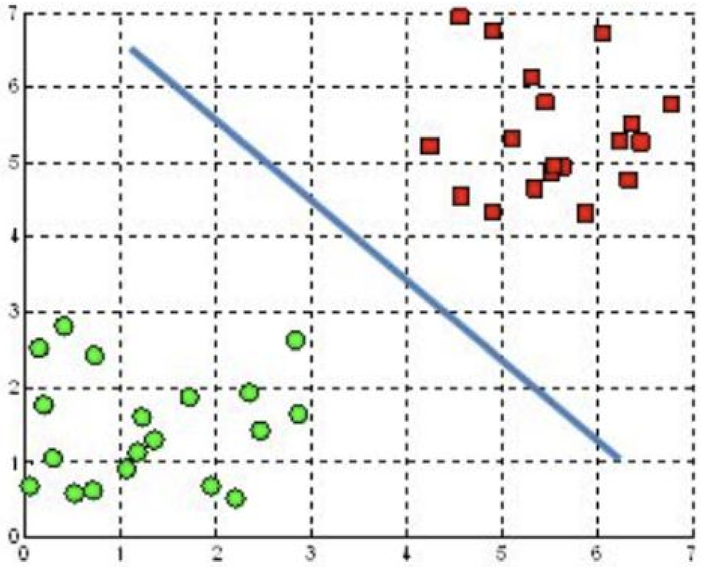
HYPERPLANES



- Hyperplanes are decision boundaries that help classify the data points.
- Data points falling on either side of the hyperplane can be attributed to different classes.
- The dimension of the hyperplane depends upon the number of features.
- If the number of input features is 2, then the hyperplane is just a line.
- If the number of input features is 3, then the hyperplane becomes a two-dimensional plane.
- It becomes difficult to imagine when the number of features exceeds 3.

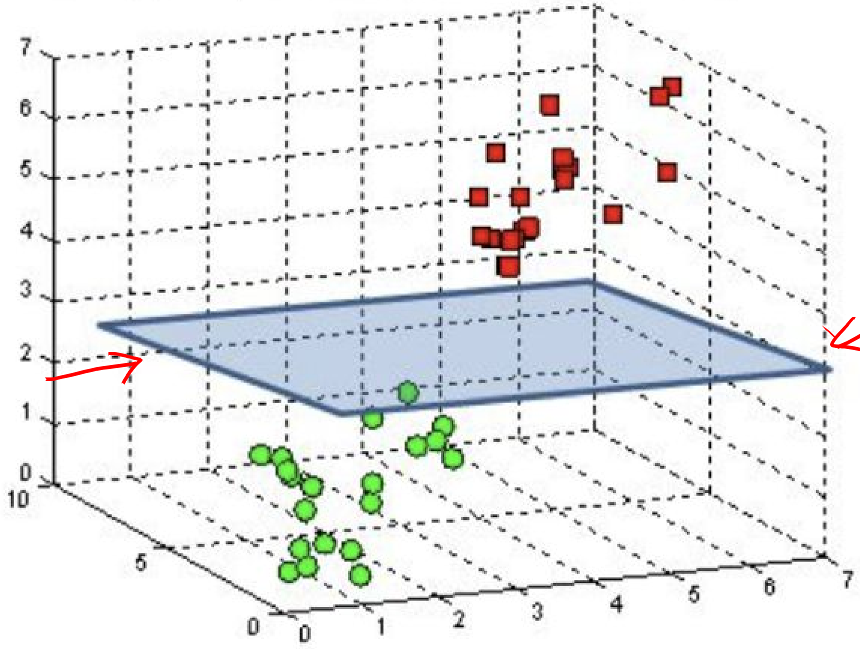
HYPERPLANES

A hyperplane in $\mathbb{R}^2$ is a line



$$\frac{R^n}{n\text{-dim}}$$

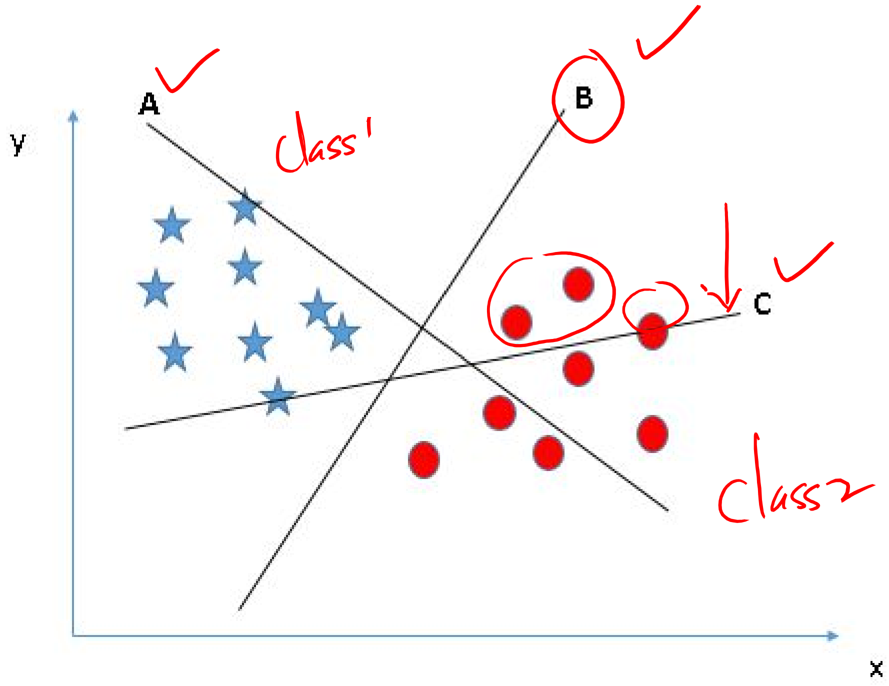
A hyperplane in $\mathbb{R}^3$ is a plane



plane

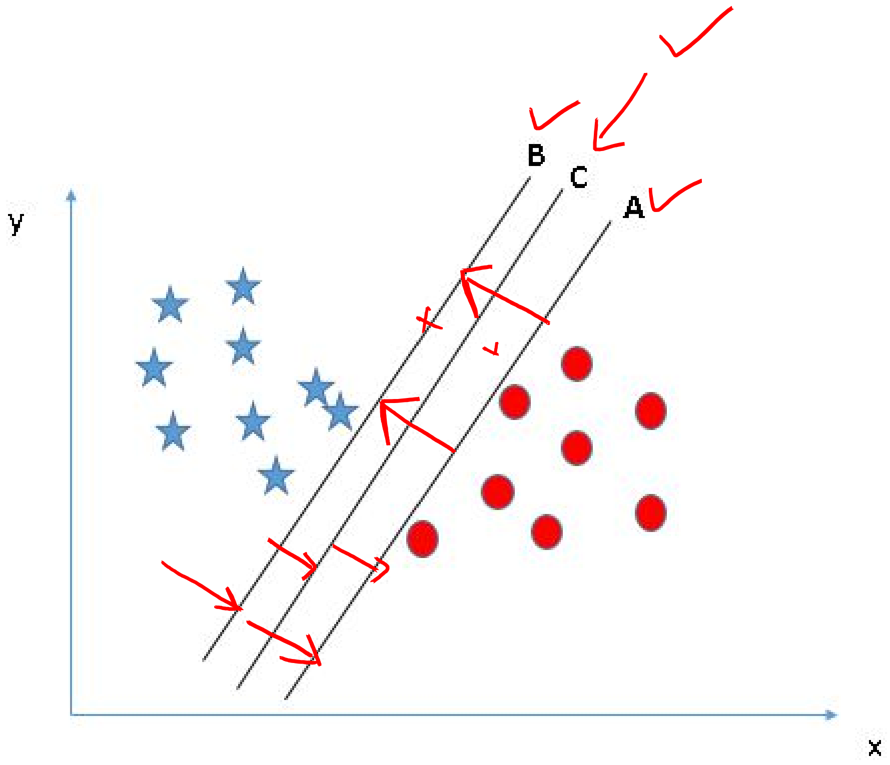
HYPERPLANES

- **Identify the right hyper-plane (Scenario-1):**



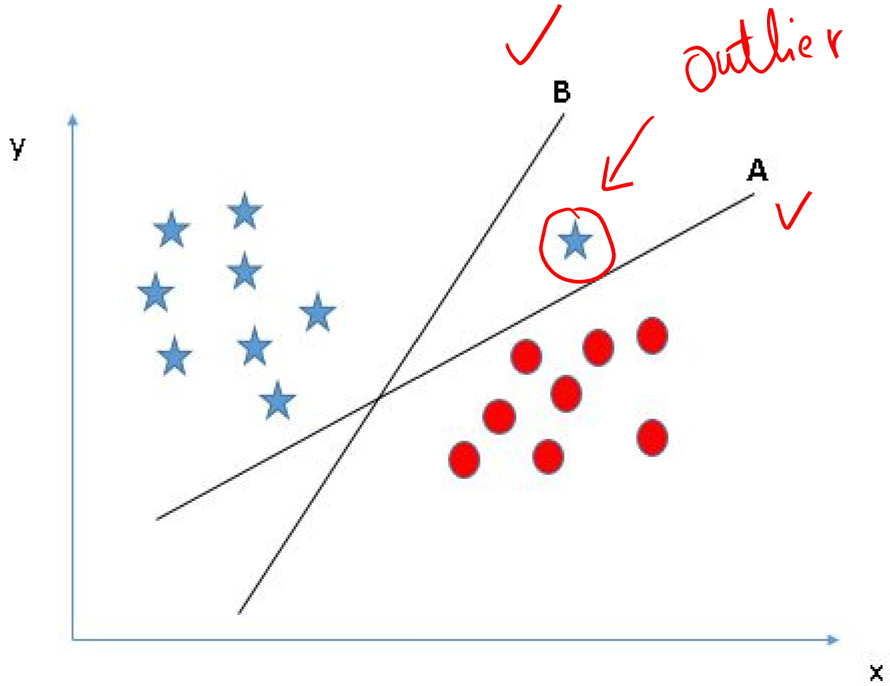
HYPERPLANES

Identify the right hyper-plane (Scenario-2):



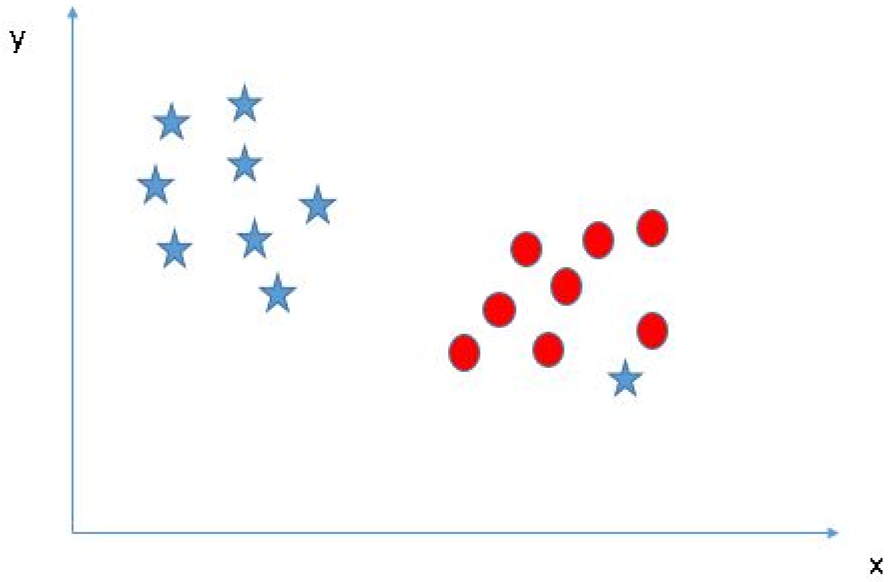
HYPERPLANES

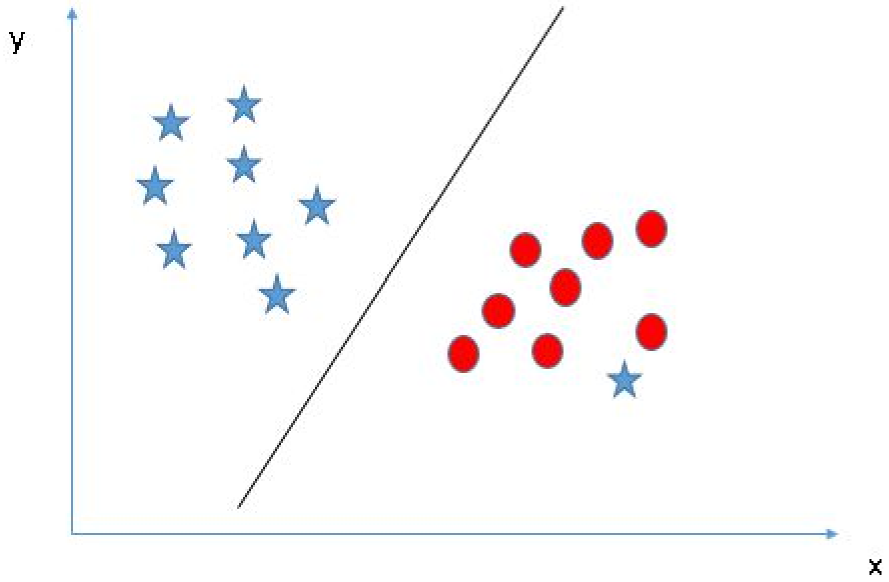
Identify the right hyper-plane (Scenario-3):



HYPERPLANES

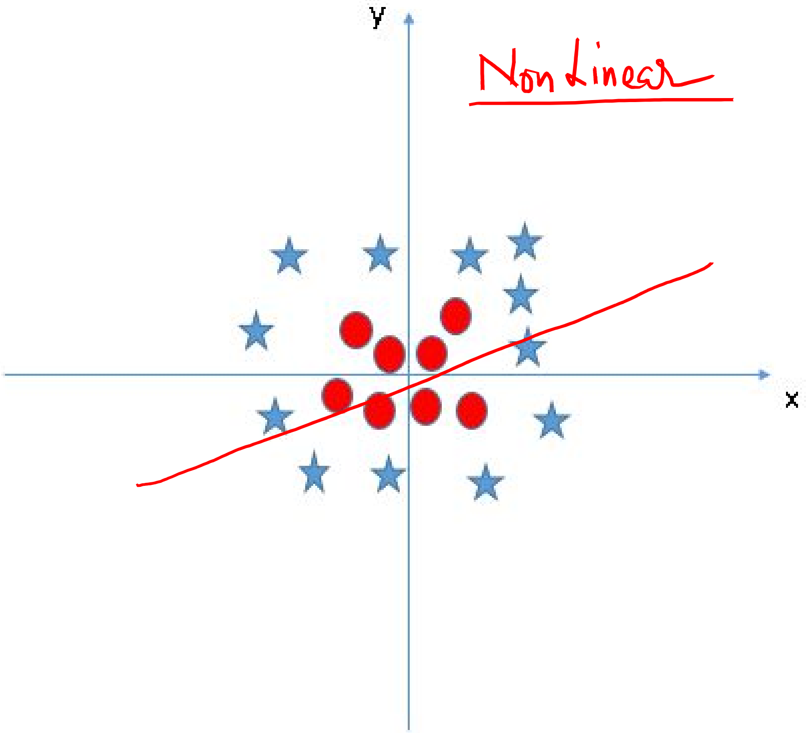
Can we classify two classes (Scenario-4)?





HYPERPLANES

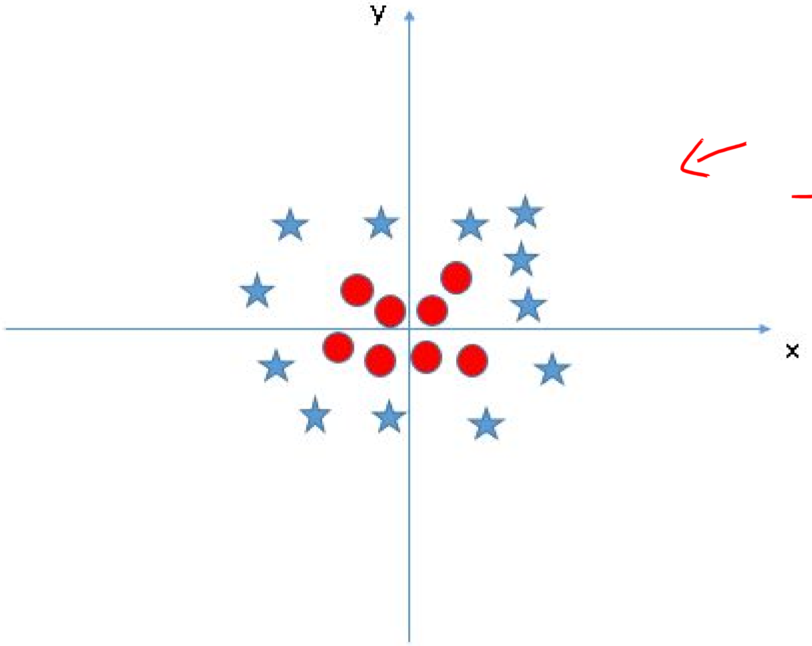
Find the hyper-plane to segregate to classes (Scenario-5):



Non Linear

HYPERPLANES

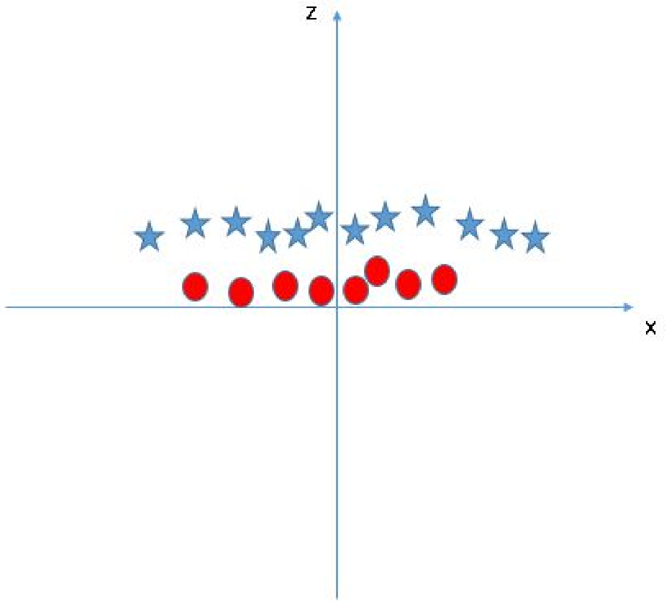
Find the hyper-plane to segregate to classes (Scenario-5): In the scenario below, we can’t have linear hyper-plane between the two classes, so how does SVM classify these two classes?



← 2-d

HYPERPLANES

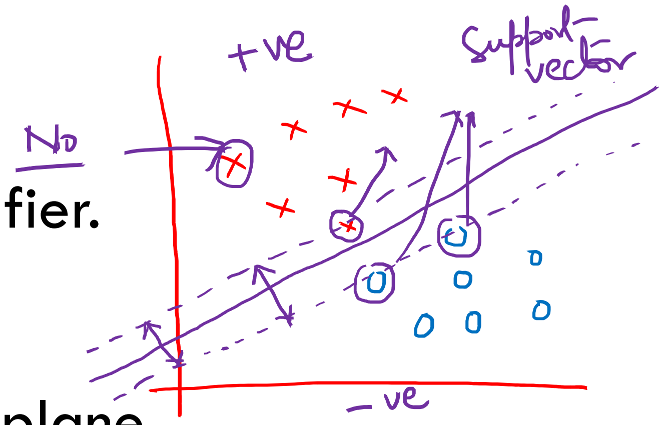
The SVM algorithm has a technique called the **kernel trick.** The SVM kernel is a function that takes low dimensional input space and transforms it to a higher dimensional space i.e. it converts not separable problem to separable problem.



SUPPORT VECTORS

- Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane.

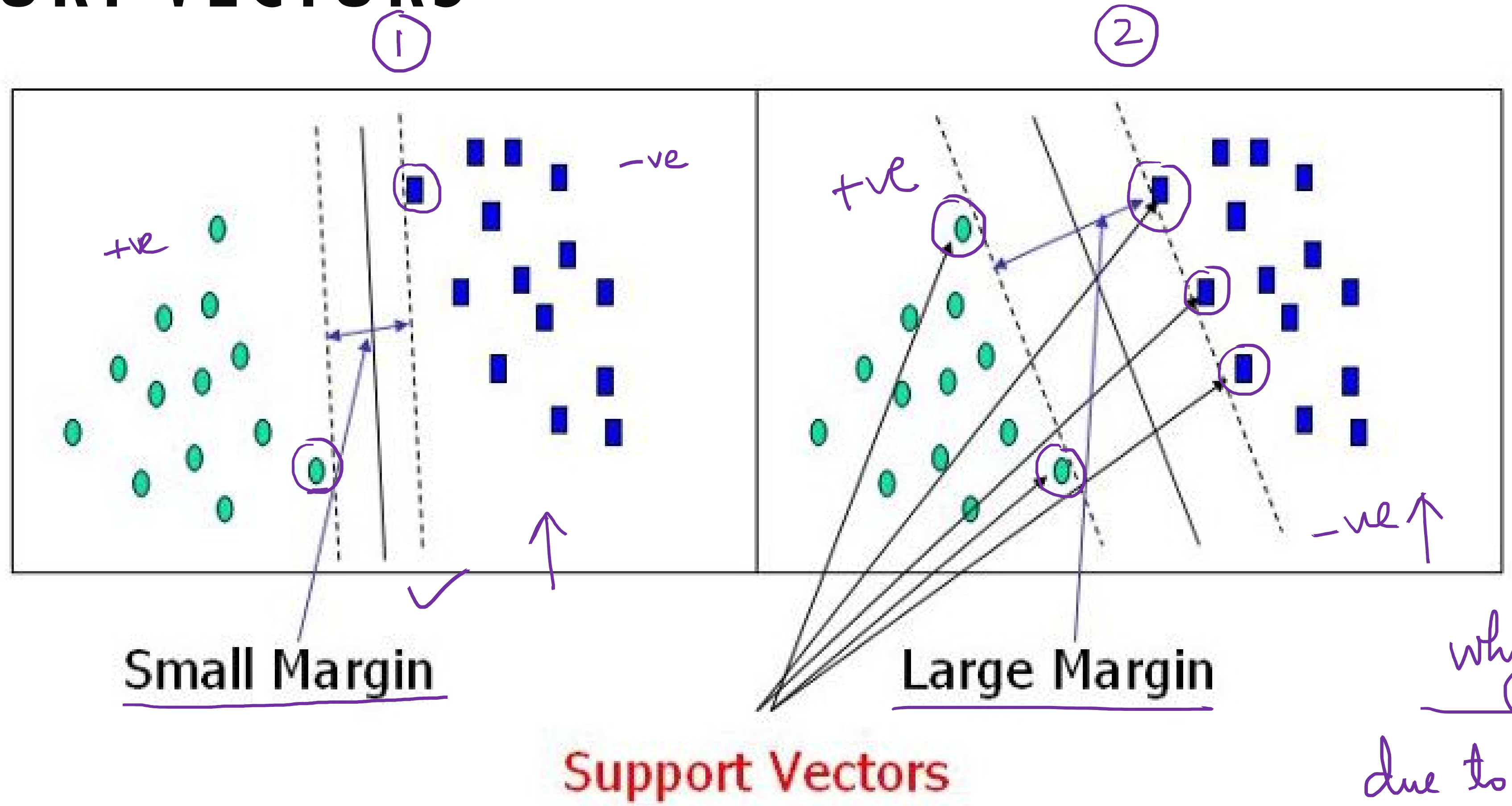
- Using these support vectors, we maximize the margin of the classifier.



- Deleting the support vectors will change the position of the hyperplane.

- These are the points that help us build our SVM.

SUPPORT VECTORS



why ?

due to max margin

SVM

MAXIMUM MARGIN: FORMALIZATION

w: decision hyperplane normal vector

Binary SVM (2 classes)

+ve -ve

$\mathbf{x}_i$: data point *i*

y_i: class of data point *i* (+1 or -1) NB: Not 1/0

Classifier is: $f(\mathbf{x}_i) = \text{sign}(\mathbf{w}^T\mathbf{x}_i + b)$ ✓

$\boxed{y = mx + c}$

Functional margin of $\mathbf{x}_i$ is: $y_i\,(\mathbf{w}^T\mathbf{x}_i + b)$

- But note that we can increase this margin simply by scaling **w**, **b**….

Functional margin of dataset is twice the minimum functional margin for any point

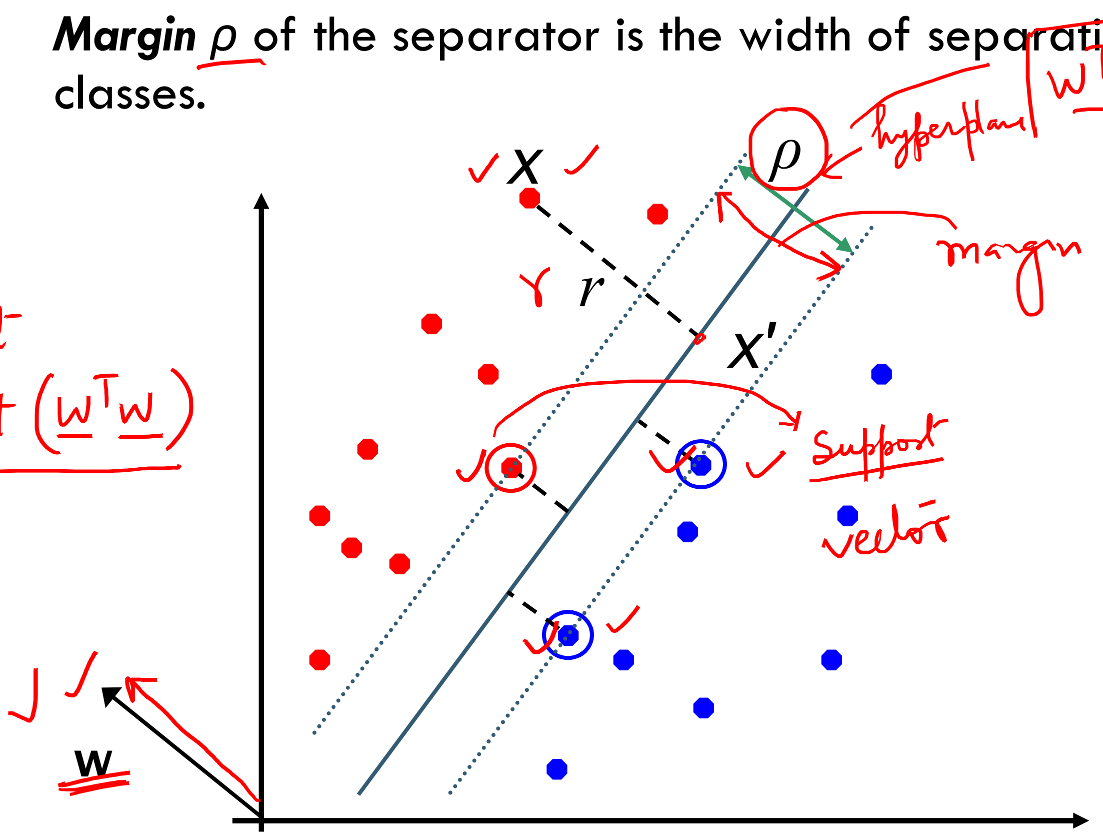
- The factor of 2 comes from measuring the whole width of the margin

GEOMETRIC MARGIN

Distance from example to the separator is $r = y\frac{\mathbf{w}^T\mathbf{x}+b}{\|\mathbf{w}\|}$

Examples closest to the hyperplane are ***support vectors.***

Margin ρ of the separator is the width of separation between support vectors of classes.



Recall that
$|W| = \text{Sqrt}(W^TW)$

Derivation of find γ

Dotted line $x' - x$ ⊥ to decision boundary and it is ∥ to w.

Unit vector is $\frac{w}{|w|}$, so line is $\frac{\gamma w}{|w|}$

$x' = x - y\frac{\gamma w}{|w|}$

x' satifies $\boxed{w^Tx' + b = 0}$

So, $w^T\left(x - y\frac{\gamma w}{|w|}\right) + b = 0$

So, $w^Tx - y\gamma(w) + b = 0$

$\boxed{\gamma = y(w^Tx+b)/|w|}$

LINEAR SVM MATHEMATICALLY

Assume that all data is at least distance 1 from the hyperplane, then the following two constraints follow for a training set $\{(\mathbf{x}_i, y_i)\}$

data points

class < +ve, -ve

$$\mathbf{w}^{\mathbf{T}}\mathbf{x}_{\mathbf{i}} + b \geq 1 \quad \text{if } y_i = 1$$

$$\mathbf{w}^{\mathbf{T}}\mathbf{x}_{\mathbf{i}} + b \leq -1 \quad \text{if } y_i = -1$$

$$\boxed{\begin{matrix} w^T x_i + b \geq 1 \\ w^T x_i + b \leq -1 \end{matrix}}$$

✓ $y_i = +ve$, 1

✓ $y_i = -ve$, -1

For support vectors, the inequality becomes an equality

Then, since each example's distance from the hyperplane is $r = y\dfrac{\mathbf{w}^T\mathbf{x} + b}{\|\mathbf{w}\|}$

$W^T x$

The margin is: $\boxed{\rho = \dfrac{2}{\|\mathbf{w}\|}}$

Sec. 15.1

LINEAR SUPPORT VECTOR MACHINE (SVM)

Hyperplane

$$\boxed{\mathbf{w}^T \mathbf{x} + b = 0}$$

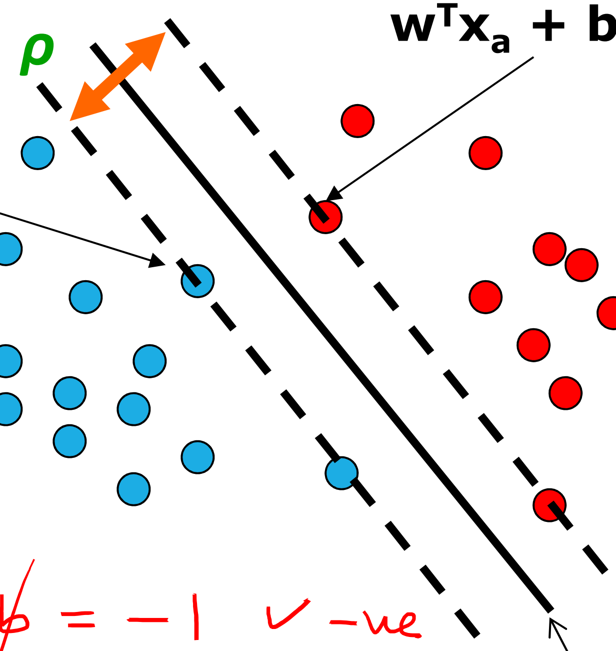
Extra scale constraint:

$$\min_{i=1,\dots,n} |\mathbf{w}^T \mathbf{x}_i + b| = 1$$

This implies:

$$w^T(x_a - x_b) = 2$$

$$\rho = ||x_a - x_b||_2 = 2/||w||_2$$



$\mathbf{w}^T\mathbf{x}_a + \mathbf{b} = \mathbf{1}$

$\mathbf{w}^T\mathbf{x}_b + \mathbf{b} = \mathbf{-1}$

ρ

$\mathbf{w}^T \mathbf{x} + \mathbf{b} = 0$

ρ ↑ Maxi

margin

$$\boxed{\rho = ||x_a - x_b|| = \frac{2}{||w||}}$$

$w^T x_b \cancel{+b} = -1$ ✓ −ve

$w^T x_a \cancel{+b} = 1$ ✓ +ve

$(-) \quad (-) \quad (-)$

$w^T(x_b - x_a) = -2$

$w^T(x_a - x_b) = 2$

$w^T(x_a - x_b) = 2$

$(x_a - x_b) = \frac{2}{w^T}$

LINEAR SVMS MATHEMATICALLY (CONT.)

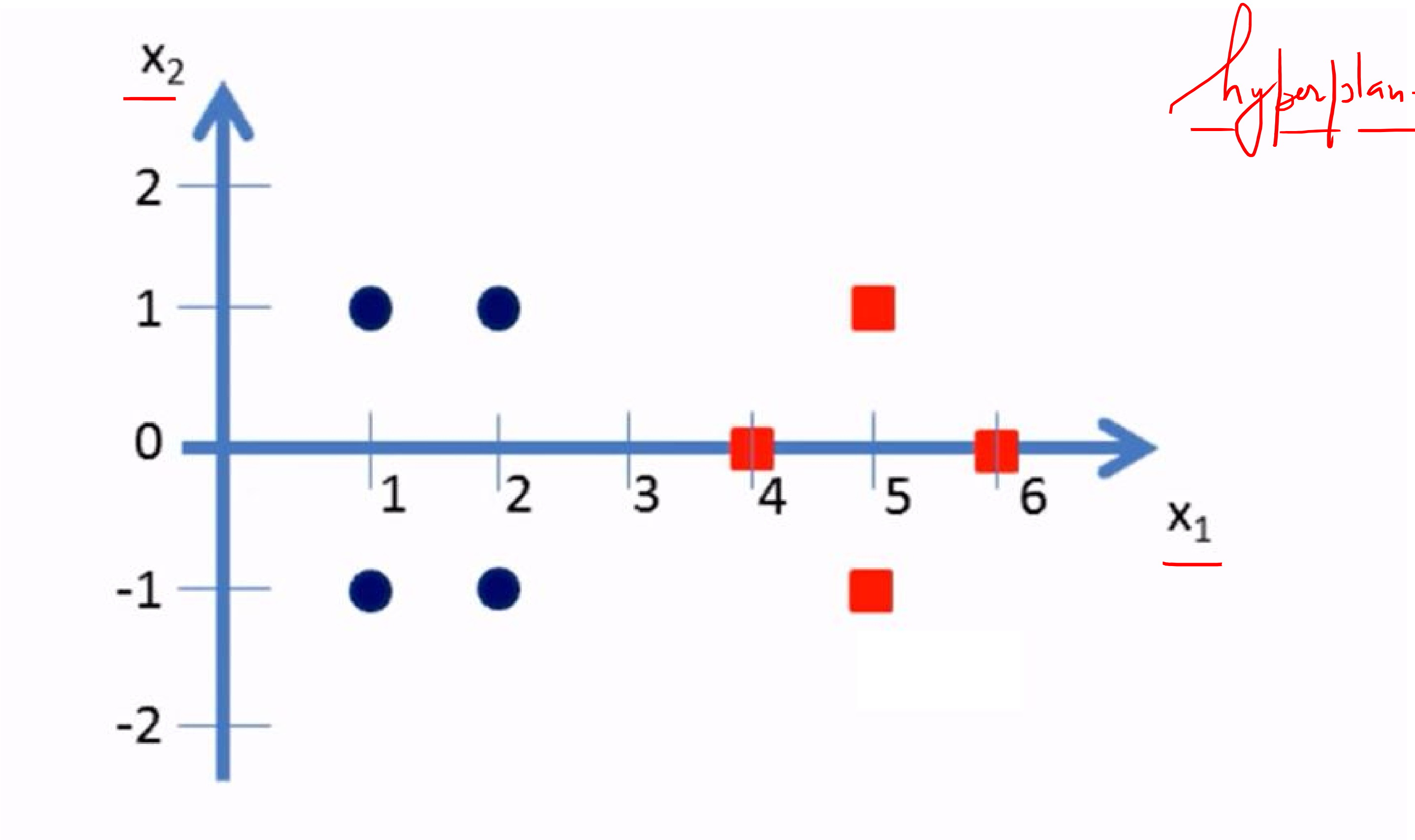
Then we can formulate the *quadratic optimization problem*:

Find $\underline{\mathbf{w}}$ and $\underline{b}$ such that

$\boxed{\rho = \dfrac{2}{\|\mathbf{w}\|}}$ is maximized; and for all $\underline{\{(\mathbf{x_i}\,,\, y_i)\}}$

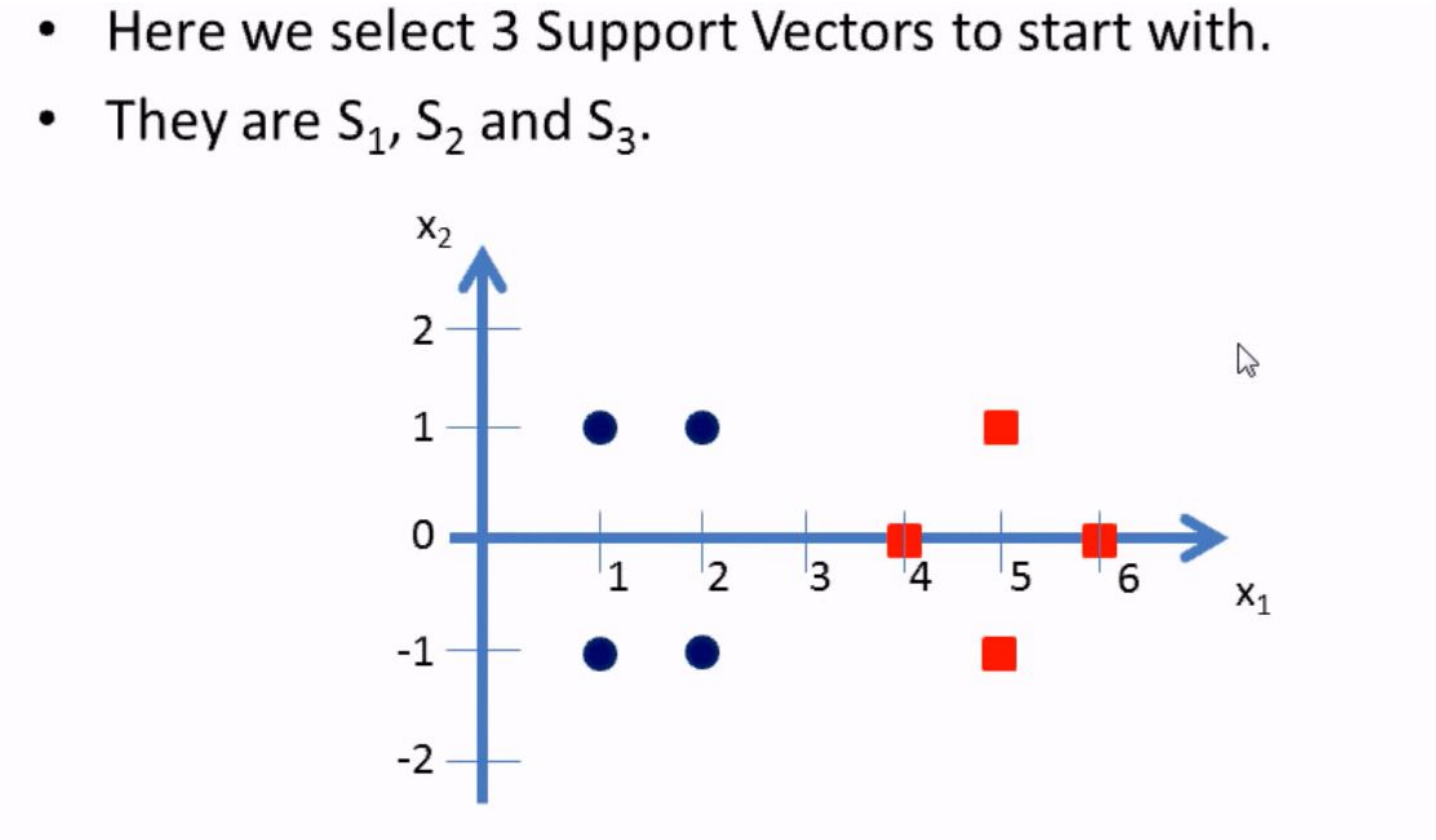
$$\boxed{\mathbf{w}^{\mathbf{T}}\mathbf{x_i} + b \geq 1 \text{ if } y_i=1;} \quad \mathbf{w}^{\mathbf{T}}\mathbf{x_i} + b \leq -1 \quad \text{if } y_i = -1$$

SVM EXAMPLE



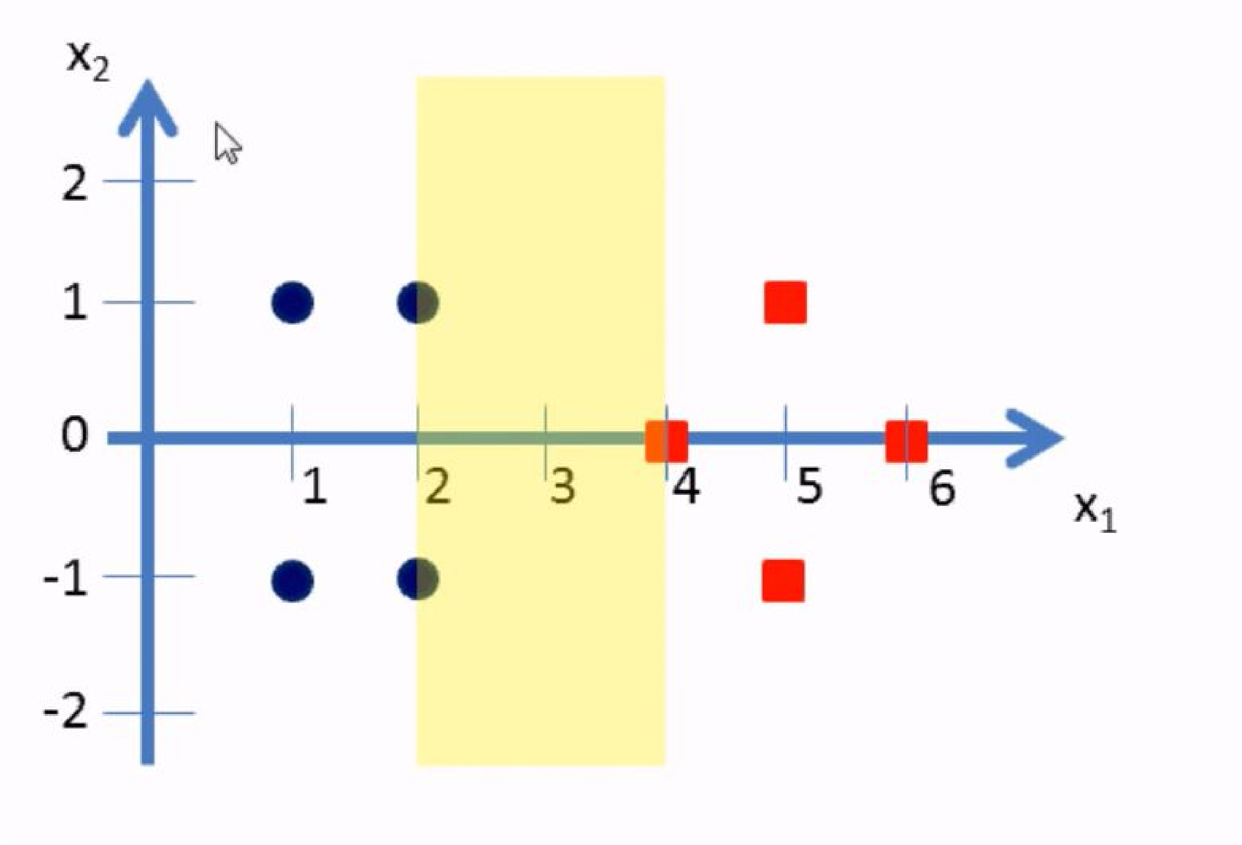
hyperplane

SVM EXAMPLE

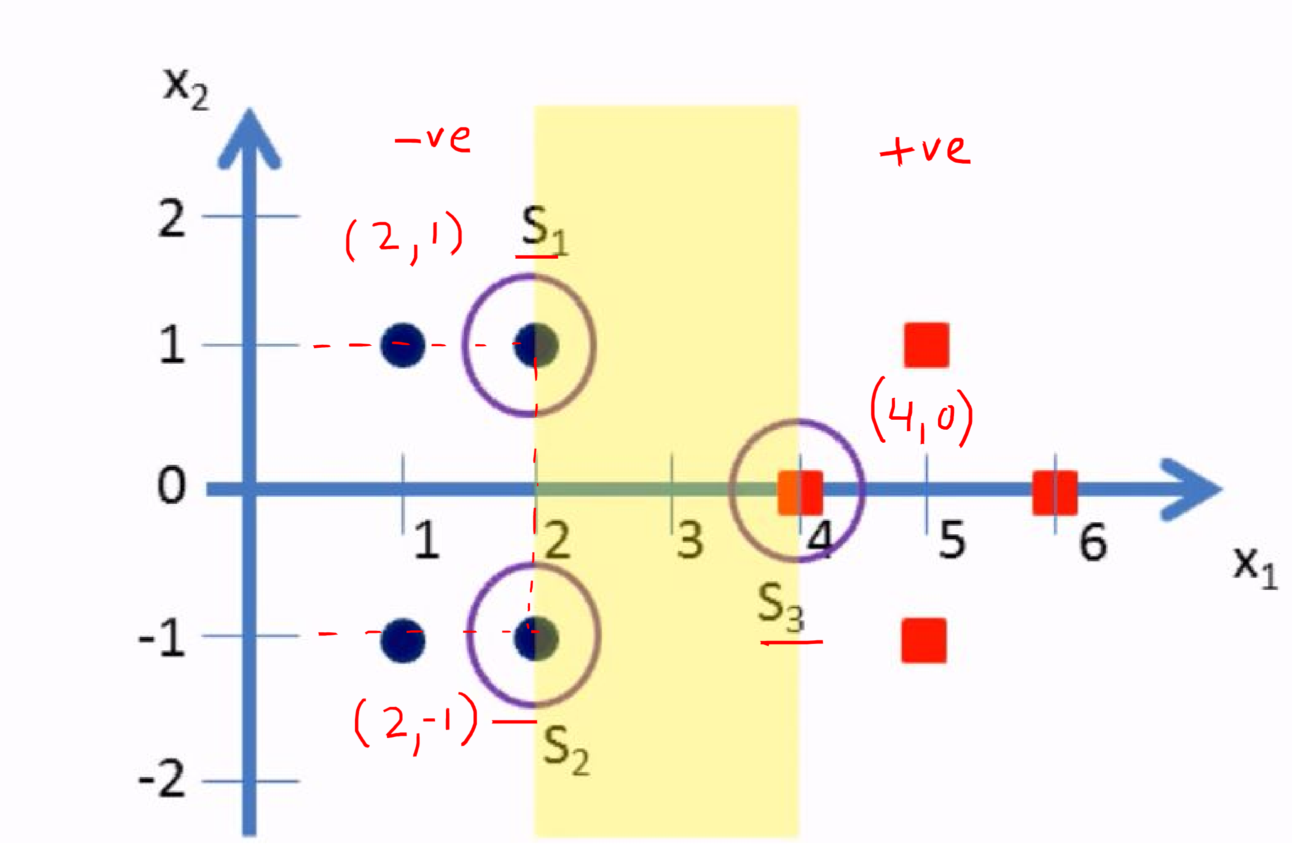


- Here we select 3 Support Vectors to start with.
- They are S_1, S_2 and S_3.

SVM EXAMPLE



SVM EXAMPLE

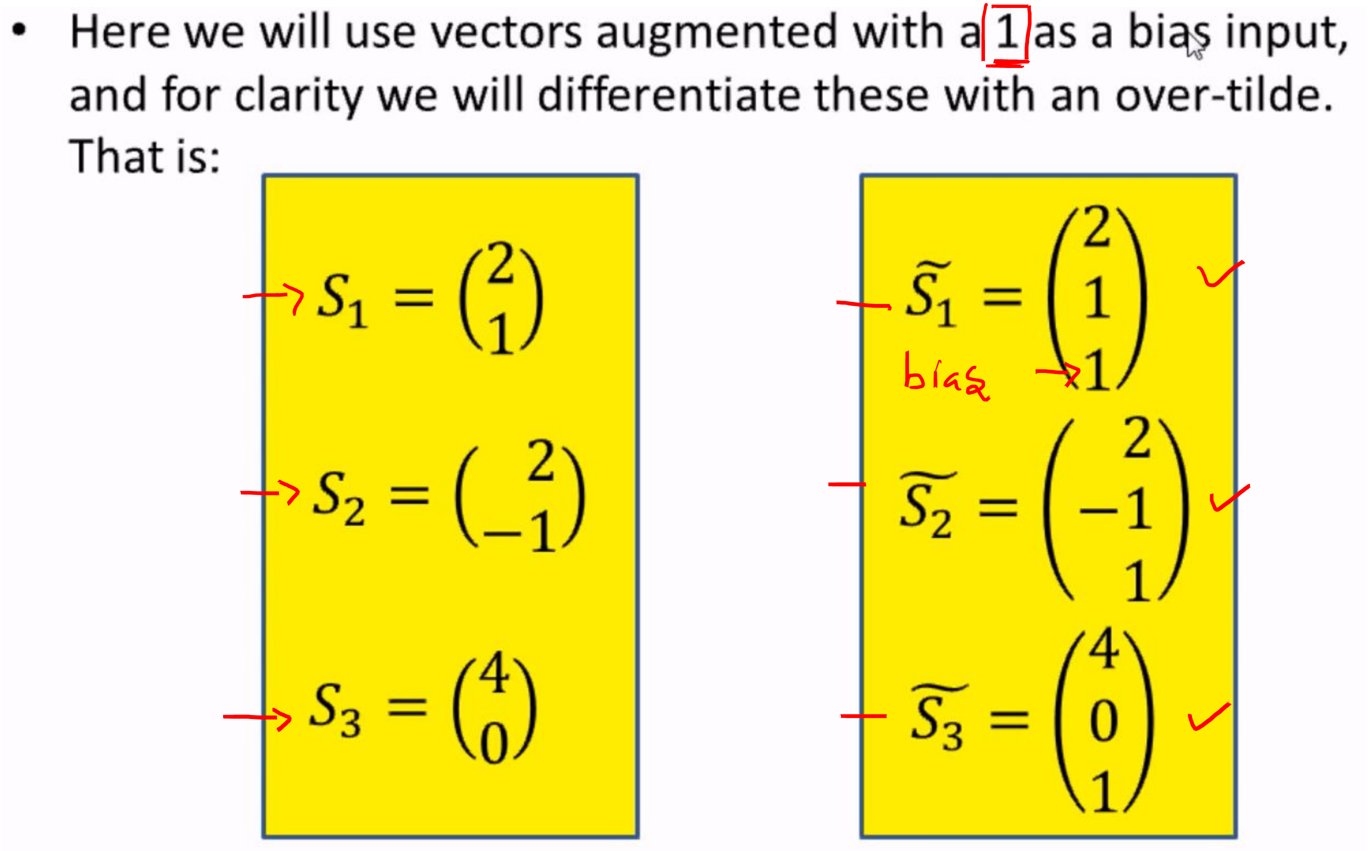


$$\rightarrow S_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \checkmark$$

$$\rightarrow S_2 = \begin{pmatrix} 2 \\ -1 \end{pmatrix} \checkmark$$

$$\rightarrow S_3 = \begin{pmatrix} 4 \\ 0 \end{pmatrix} \checkmark$$

SVM EXAMPLE



- Here we will use vectors augmented with a $\boxed{1}$ as a bias input, and for clarity we will differentiate these with an over-tilde. That is:

$$S_1 = \begin{pmatrix}2\\1\end{pmatrix}$$

$$S_2 = \begin{pmatrix}2\\-1\end{pmatrix}$$

$$S_3 = \begin{pmatrix}4\\0\end{pmatrix}$$

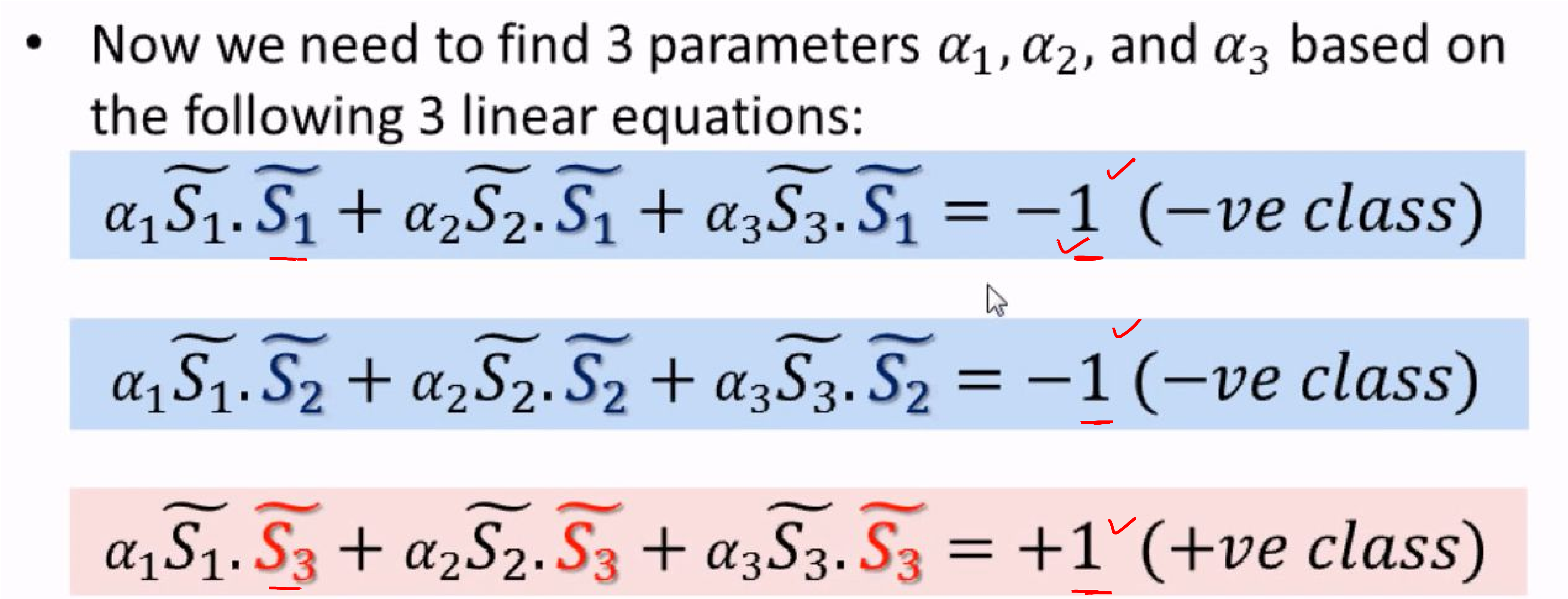
$$\widetilde{S_1} = \begin{pmatrix}2\\1\\1\end{pmatrix} \checkmark$$

bias

$$\widetilde{S_2} = \begin{pmatrix}2\\-1\\1\end{pmatrix} \checkmark$$

$$\widetilde{S_3} = \begin{pmatrix}4\\0\\1\end{pmatrix} \checkmark$$

SVM EXAMPLE



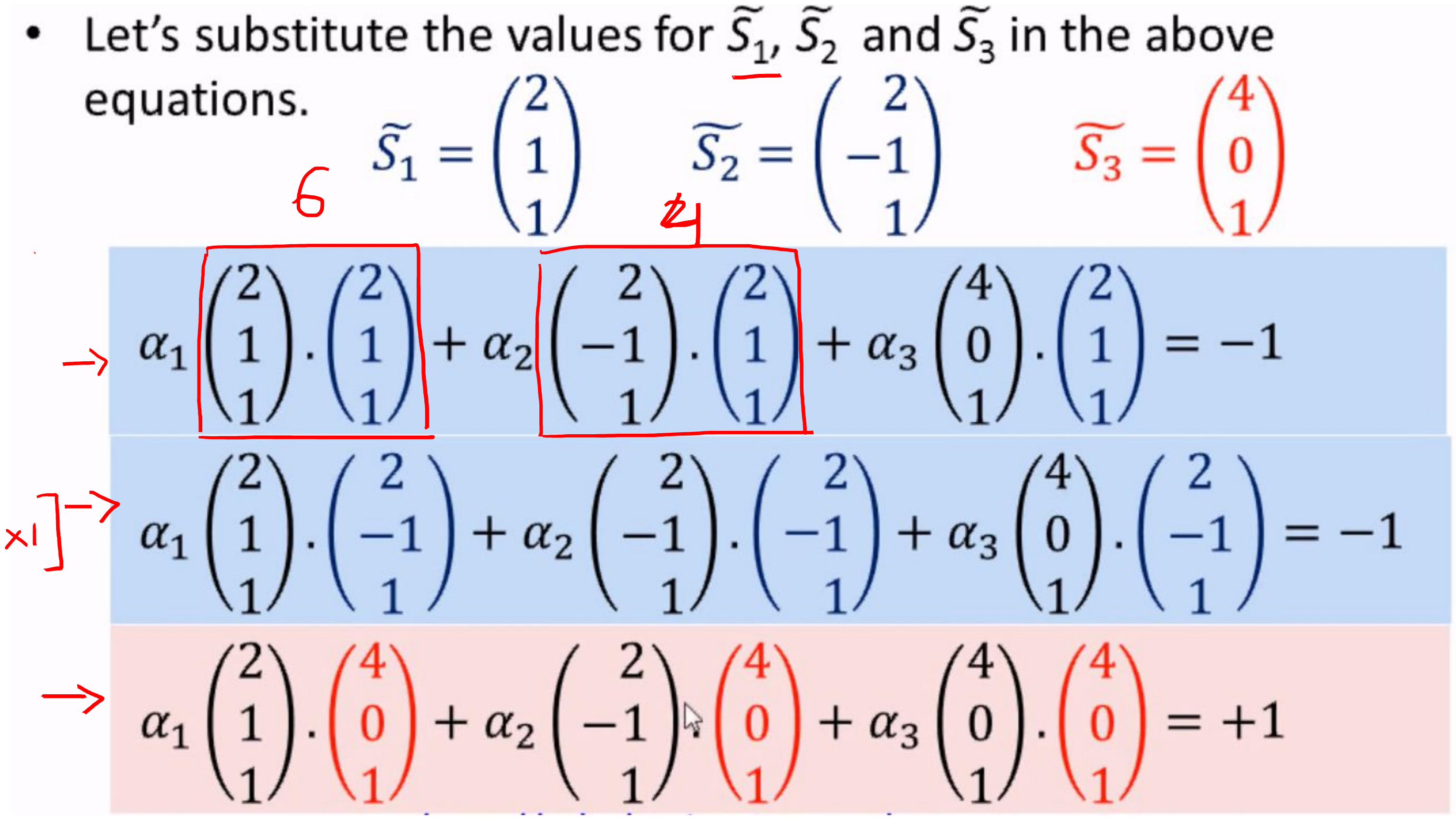
- Now we need to find 3 parameters α_1, α_2, and α_3 based on the following 3 linear equations:

$$\alpha_1 \widetilde{S_1} . \underline{\widetilde{S_1}} + \alpha_2 \widetilde{S_2} . \widetilde{S_1} + \alpha_3 \widetilde{S_3} . \widetilde{S_1} = -\underset{✓}{\underline{1}}^{✓} \; (-ve\ class)$$

$$\alpha_1 \widetilde{S_1} . \widetilde{S_2} + \alpha_2 \widetilde{S_2} . \widetilde{S_2} + \alpha_3 \widetilde{S_3} . \widetilde{S_2} = -\underline{1}^{✓} \; (-ve\ class)$$

$$\alpha_1 \widetilde{S_1} . \underline{\widetilde{S_3}} + \alpha_2 \widetilde{S_2} . \widetilde{S_3} + \alpha_3 \widetilde{S_3} . \widetilde{S_3} = +\underline{1}^{✓} \; (+ve\ class)$$

SVM EXAMPLE



- Let's substitute the values for $\tilde{S_1}$, $\tilde{S_2}$ and $\tilde{S_3}$ in the above equations.

$$\tilde{S_1} = \begin{pmatrix}2\\1\\1\end{pmatrix} \qquad \tilde{S_2} = \begin{pmatrix}2\\-1\\1\end{pmatrix} \qquad \tilde{S_3} = \begin{pmatrix}4\\0\\1\end{pmatrix}$$

$$\rightarrow \alpha_1 \overset{6}{\boxed{\begin{pmatrix}2\\1\\1\end{pmatrix} \cdot \begin{pmatrix}2\\1\\1\end{pmatrix}}} + \alpha_2 \overset{4}{\boxed{\begin{pmatrix}2\\-1\\1\end{pmatrix} \cdot \begin{pmatrix}2\\1\\1\end{pmatrix}}} + \alpha_3 \begin{pmatrix}4\\0\\1\end{pmatrix} \cdot \begin{pmatrix}2\\1\\1\end{pmatrix} = -1$$

$$\rightarrow \alpha_1 \begin{pmatrix}2\\1\\1\end{pmatrix} \cdot \begin{pmatrix}2\\-1\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}2\\-1\\1\end{pmatrix} \cdot \begin{pmatrix}2\\-1\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}4\\0\\1\end{pmatrix} \cdot \begin{pmatrix}2\\-1\\1\end{pmatrix} = -1$$

$$\rightarrow \alpha_1 \begin{pmatrix}2\\1\\1\end{pmatrix} \cdot \begin{pmatrix}4\\0\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}2\\-1\\1\end{pmatrix} \cdot \begin{pmatrix}4\\0\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}4\\0\\1\end{pmatrix} \cdot \begin{pmatrix}4\\0\\1\end{pmatrix} = +1$$

$$\begin{bmatrix}2\\1\\1\end{bmatrix} \cdot \begin{bmatrix}2\\1\\1\end{bmatrix}$$

$$= \left[2 \times 2 + 1 \times 1 + 1 \times 1\right]$$

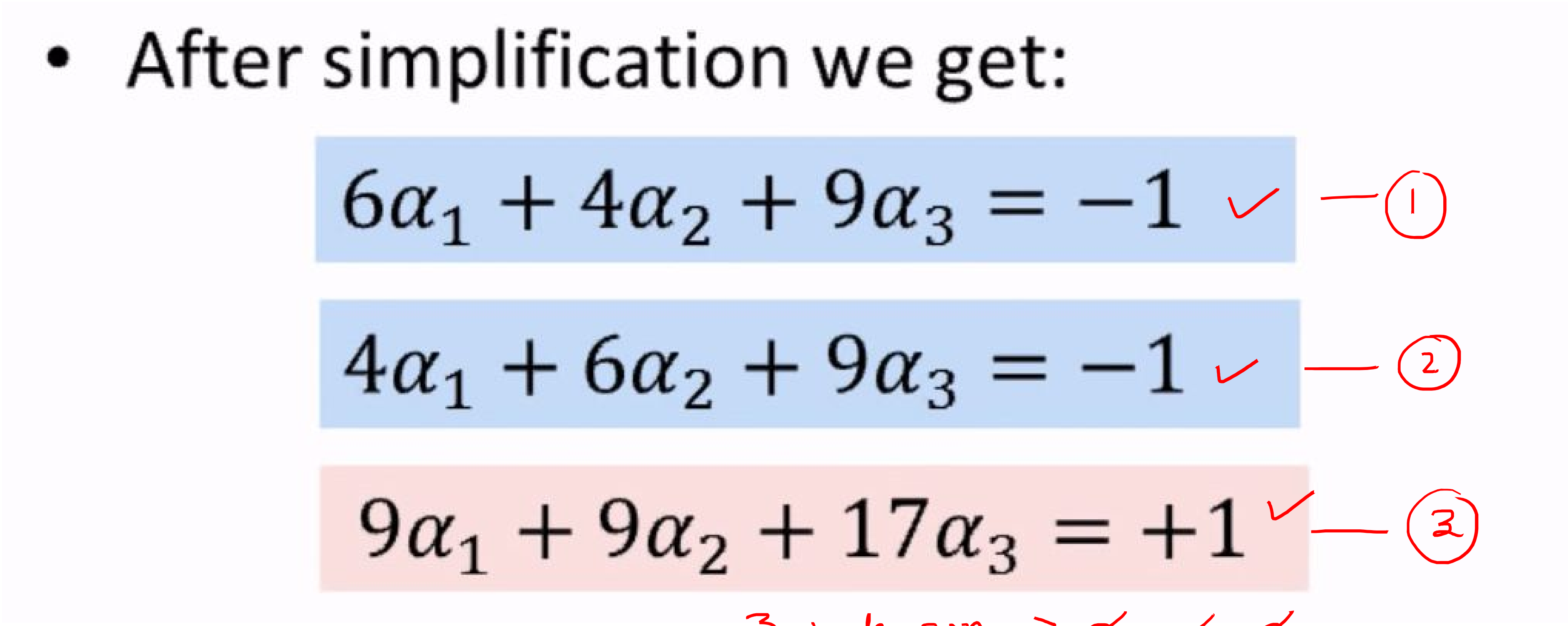
$$= \left[4 + 1 + 1\right]$$

$$= 6$$

$$\begin{bmatrix}2\\-1\\1\end{bmatrix} \begin{bmatrix}2\\1\\1\end{bmatrix}$$

$$= 2 \times 2 - 1 + 1 = 4$$

EXAMPLE



- After simplification we get:

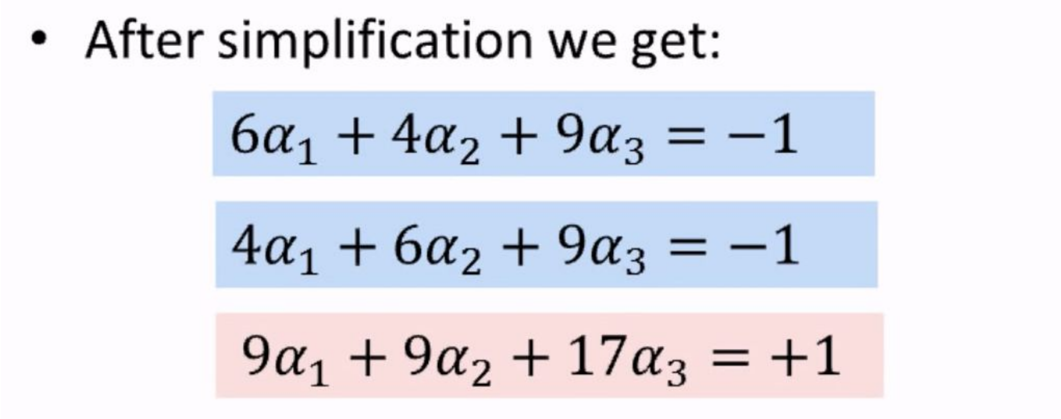
$$6\alpha_1 + 4\alpha_2 + 9\alpha_3 = -1 \quad \checkmark \quad — \textcircled{1}$$

$$4\alpha_1 + 6\alpha_2 + 9\alpha_3 = -1 \quad \checkmark \quad — \textcircled{2}$$

$$9\alpha_1 + 9\alpha_2 + 17\alpha_3 = +1 \quad \checkmark \quad — \textcircled{3}$$

3 unknown → $\underline{\alpha_1, \alpha_2, \alpha_3}$

SVM EXAMPLE

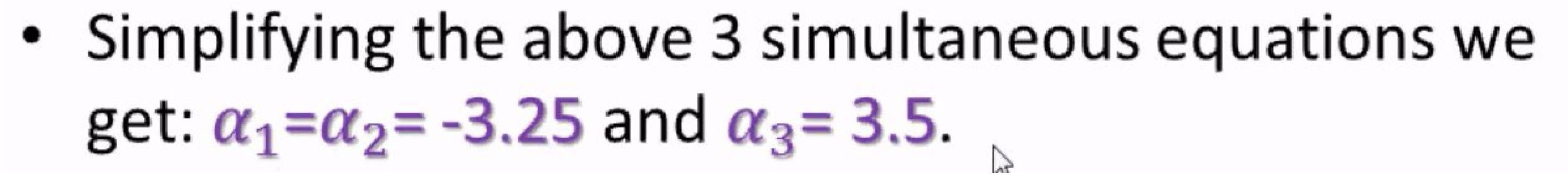


- After simplification we get:

$$6\alpha_1 + 4\alpha_2 + 9\alpha_3 = -1$$

$$4\alpha_1 + 6\alpha_2 + 9\alpha_3 = -1$$

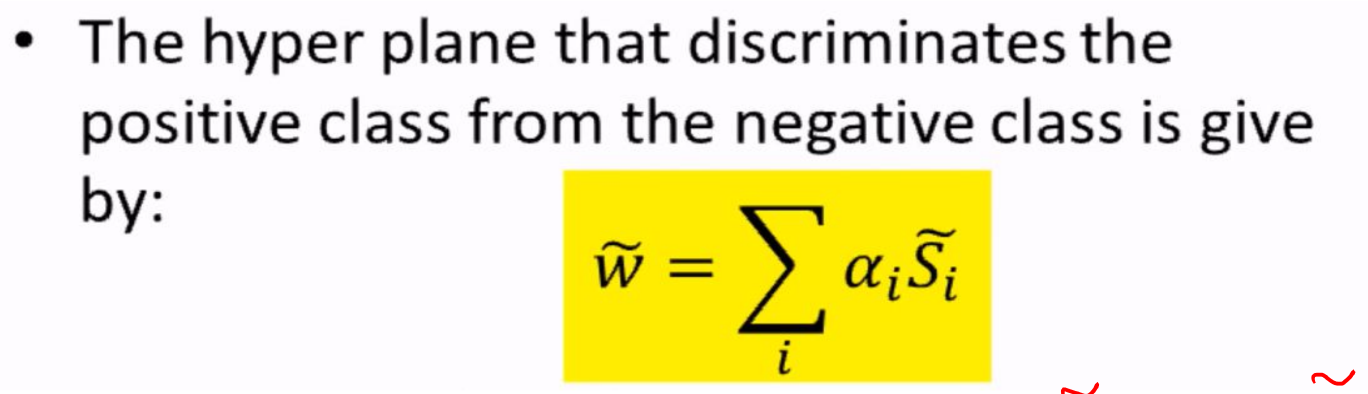
$$9\alpha_1 + 9\alpha_2 + 17\alpha_3 = +1$$



- Simplifying the above 3 simultaneous equations we get: α_1=α_2= -3.25 and α_3= 3.5.

$$\left.\begin{aligned} &\alpha_1 = \alpha_2 = -3.25, \\ &\alpha_3 = 35 \end{aligned}\right\}$$

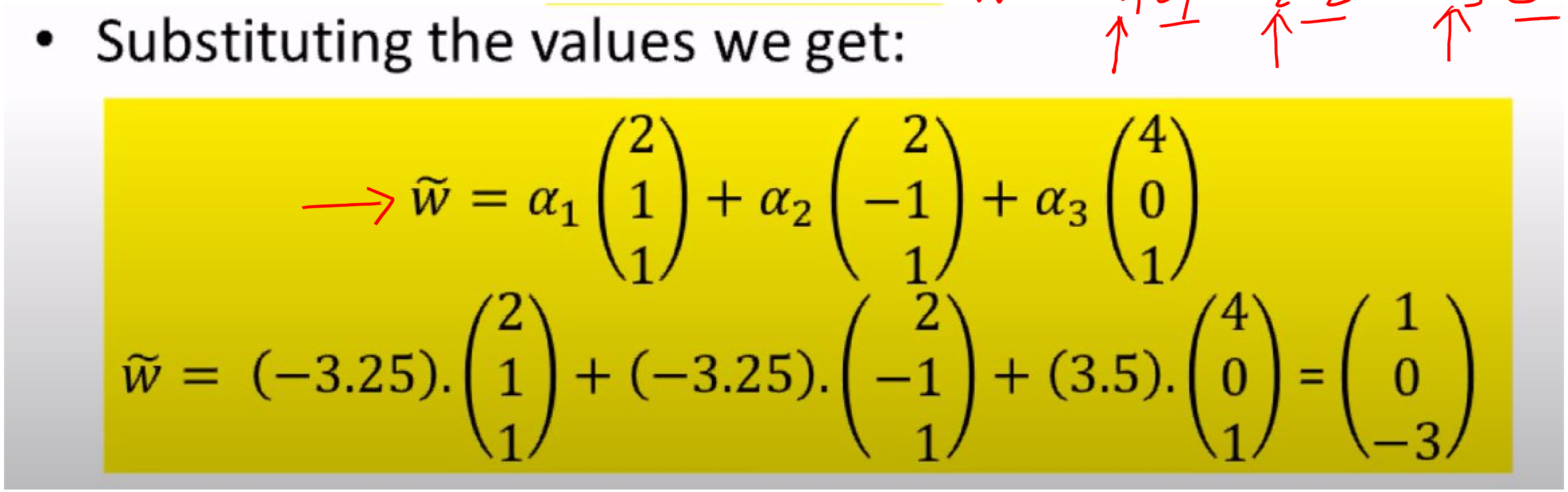
SVM EXAMPLE



- The hyper plane that discriminates the positive class from the negative class is give by:

$$\tilde{w} = \sum_i \alpha_i \tilde{S}_i$$

$$\tilde{w} = \underset{\uparrow}{\alpha_1} \underline{\tilde{S}_1} + \underset{\uparrow}{\alpha_2} \underline{\tilde{S}_2} + \underset{\uparrow}{\alpha_3} \underline{\tilde{S}_3}$$



- Substituting the values we get:

$$\rightarrow \tilde{w} = \alpha_1 \begin{pmatrix}2\\1\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}2\\-1\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}4\\0\\1\end{pmatrix}$$

$$\tilde{w} = (-3.25).\begin{pmatrix}2\\1\\1\end{pmatrix} + (-3.25).\begin{pmatrix}2\\-1\\1\end{pmatrix} + (3.5).\begin{pmatrix}4\\0\\1\end{pmatrix} = \begin{pmatrix}1\\0\\-3\end{pmatrix}$$

$$\underline{\tilde{w} = \begin{pmatrix}1\\0\\-3\end{pmatrix}}$$

SVM EXAMPLE

$y = wx + b$

$= \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

$\boxed{b = -3}$

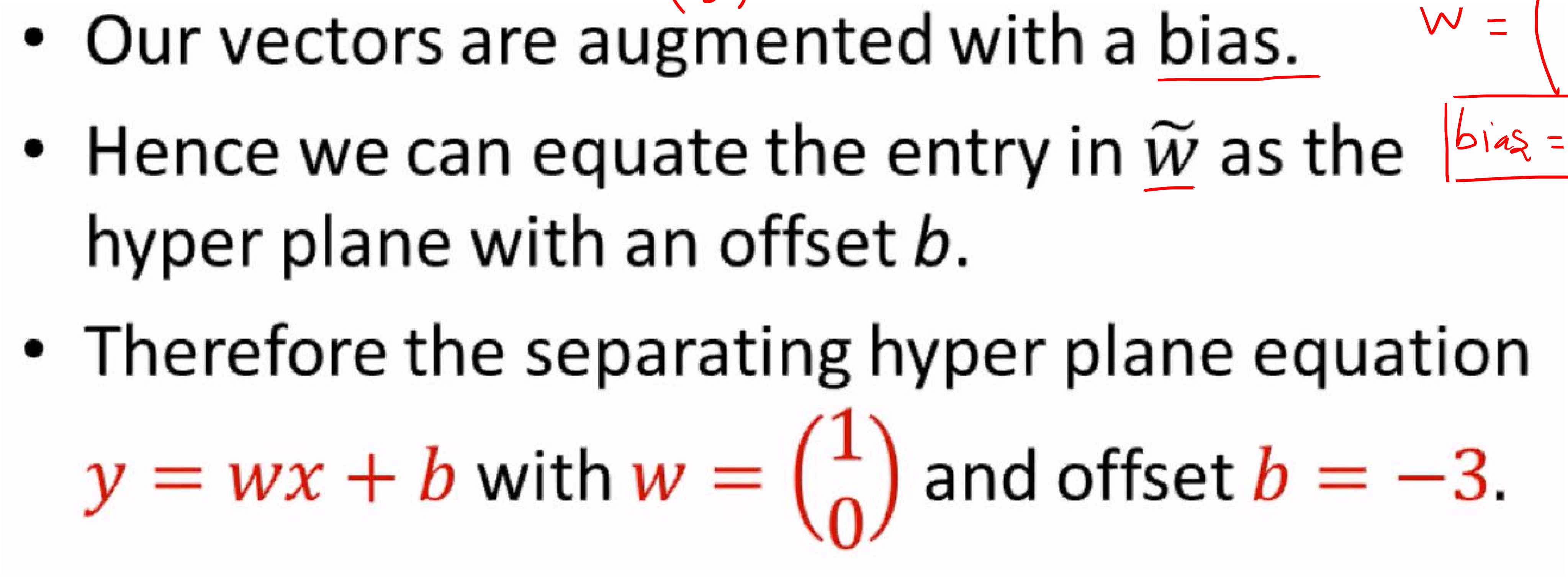
$S_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$

$\underline{\tilde{S}_1} = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} \leftarrow$ bias

↖ bias

$\tilde{W} = \begin{pmatrix} 1 \\ 0 \\ -3 \end{pmatrix}$

$\boxed{\text{bias} = -3}$

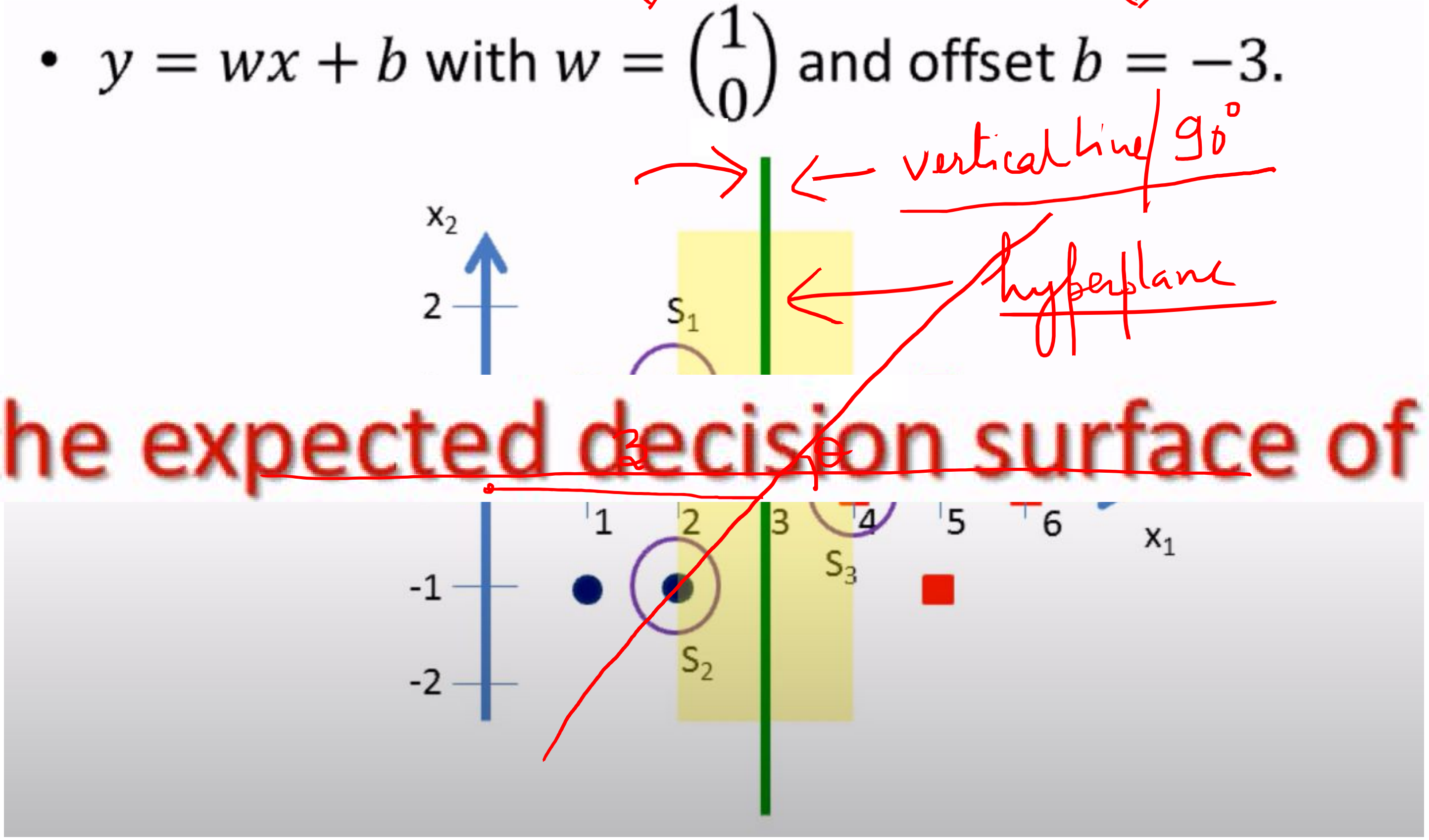


- Our vectors are augmented with a bias.
- Hence we can equate the entry in $\tilde{w}$ as the hyper plane with an offset b.
- Therefore the separating hyper plane equation $y = wx + b$ with $w = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and offset $b = -3$.

SVM EXAMPLE

$\rightarrow \begin{pmatrix}1\\0\end{pmatrix} \rightarrow$ vertical line

$\rightarrow \begin{pmatrix}0\\1\end{pmatrix} \rightarrow$ horizontal line



- $y = wx + b$ with $w = \begin{pmatrix}1\\0\end{pmatrix}$ and offset $b = -3$.

- This is the expected decision surface of the LSVM.

PROS AND CONS ASSOCIATED WITH SVM

$y = \underline{m}x + \underline{c}$

↓

$\underline{S_1, S_2}$ $\underline{S_3}$

$-ve$ $\underline{+ve}$

Pros:

- It works really well with a clear margin of separation
- It is effective in high dimensional spaces. ←
- It is effective in cases where the number of dimensions is greater than the number of samples.
- It uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

PROS AND CONS ASSOCIATED WITH SVM

Cons:

- It doesn't perform well when we have large data set because the required training time is higher.
- It doesn't perform very well, when the data set has more noise i.e. target classes are overlapping.

THANKS

Dr. Neeraj Gupta
Assistant Professor, Dept. of CEA
neeraj.gupta@gla.ac.in