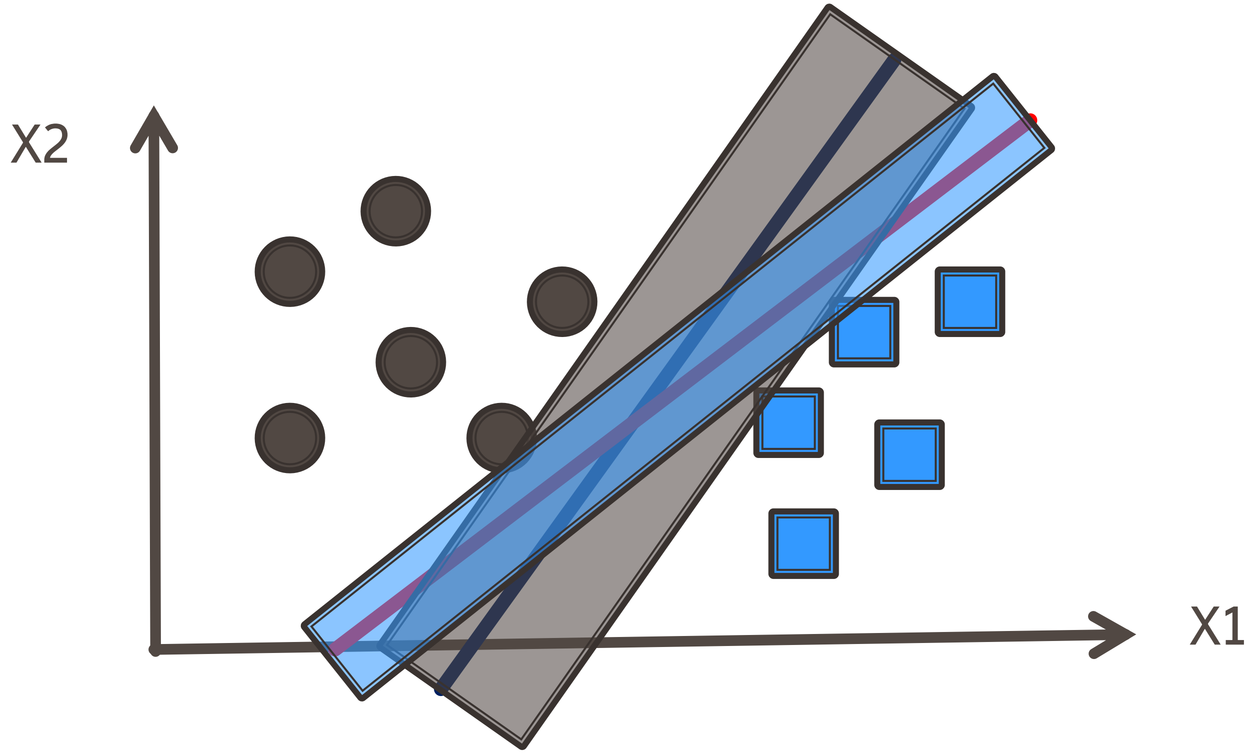
Machine learning

Presented by : Dr. Hanaa Bayomi

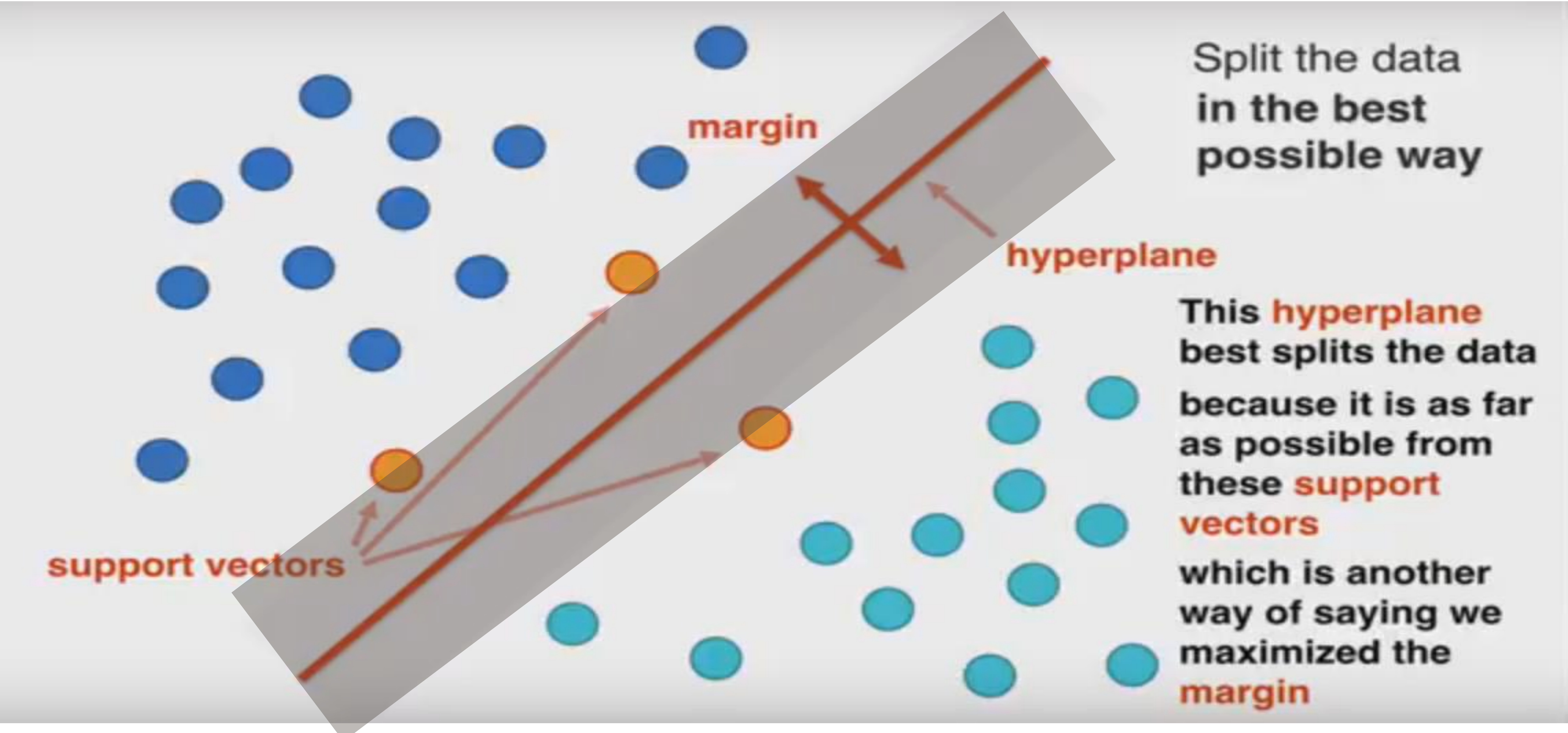
Lecture 8: SVM

overview

- SVM for **linearly separable** binary set
- **Main Goal** to design a hyper plane that classify all training vectors into two classes
- *The best model* **that leaves the maximum margin** from both classes
- the two classes labels **+1** (positive examples and **-1** (negative examples)

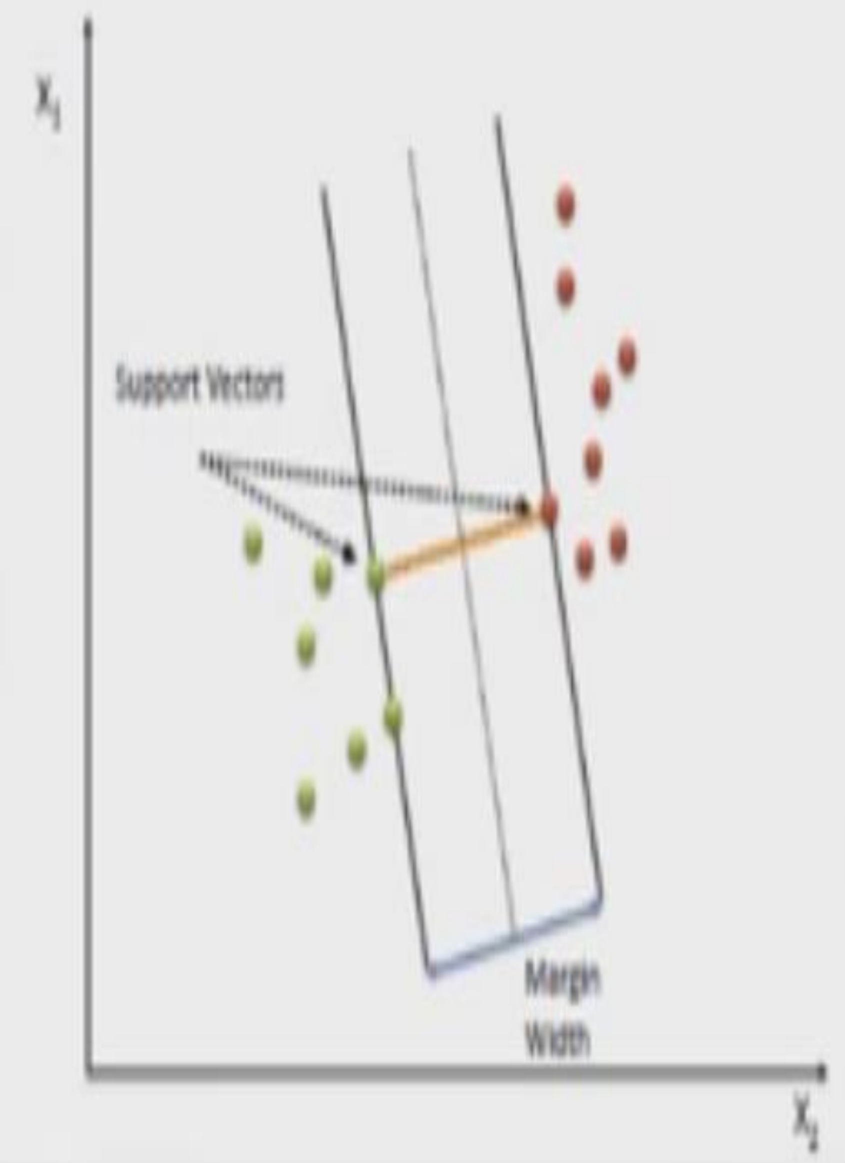


This is a constrained optimization problem



Intuition behind SVM

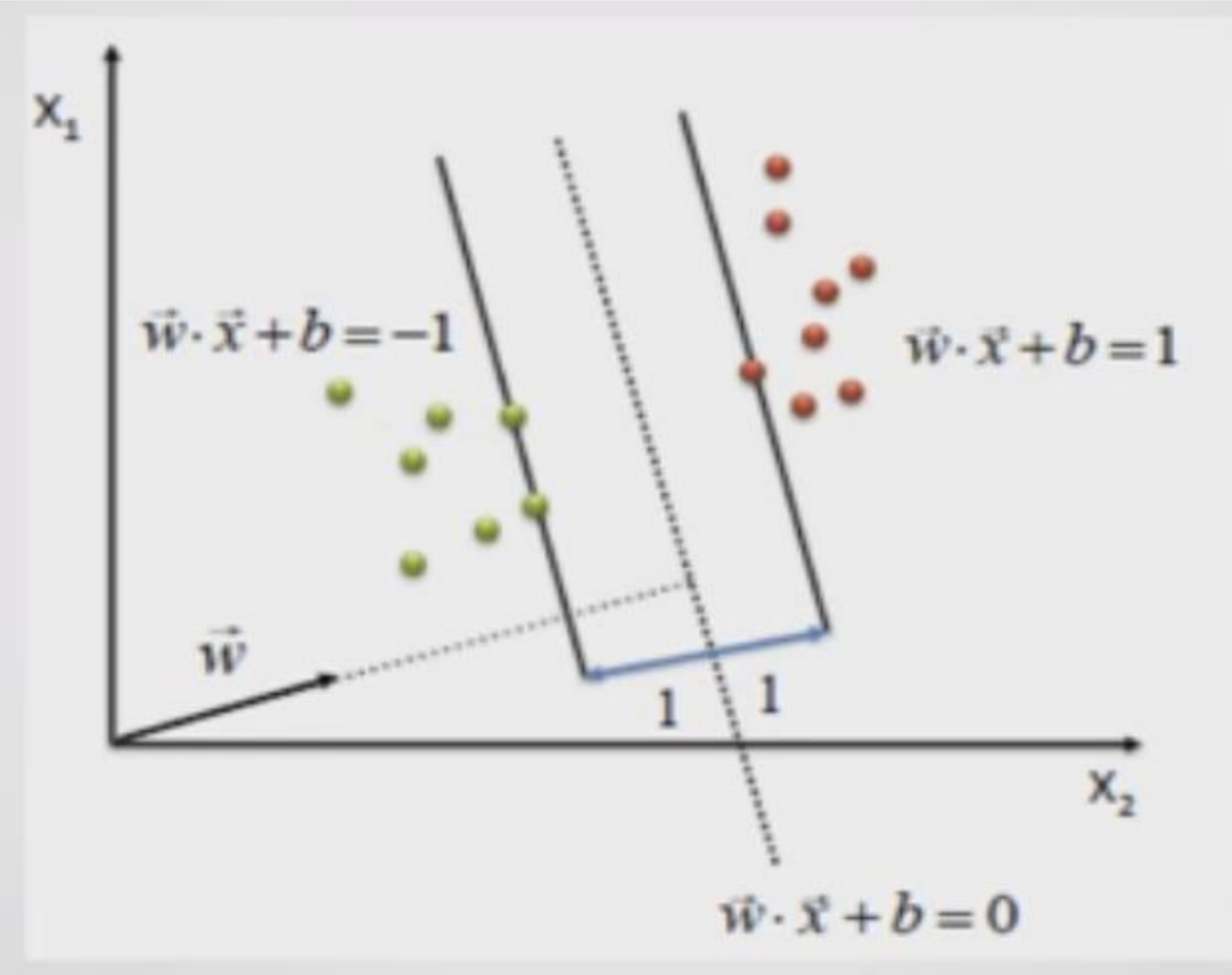
- Points (instances) are like vectors p = (x1,x2,...,xn)
- SVM finds the closest two points from the two classes (see figure), that support (define) the best separating line|plane
- Then SVM draws a line connecting them (the orange line in the figure)
- After that, SVM decides that the best separating line is the line that bisects, and is perpendicular to, the connecting line



SVM more formally

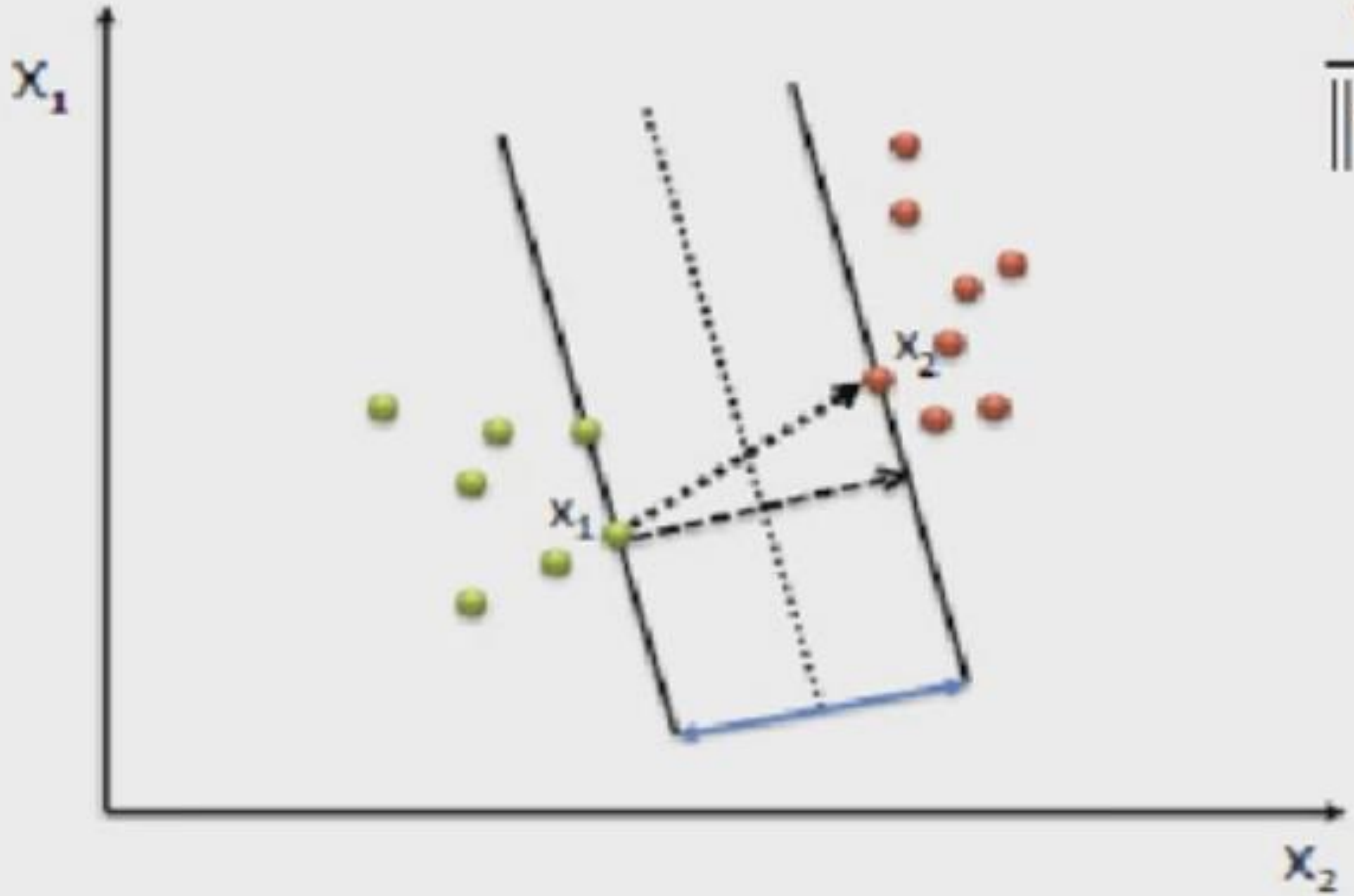
- We can define a separating (decision) hyperplane in terms of an intercept term ***b*** and a normal vector $\vec{w}$ which is perpendicular to the hyperplane (commonly referred to as the weight vector)
- To choose among all the hyperplanes that are perpendicular to the normal vector, we specify the intercept term ***b***
- All points $\vec{x}$ on the hyperplane satisfy $\vec{w}^T\vec{x} = -b$ as the hyperplane is perpendicular to the normal vector
- We represent the training dataset as $\mathbb{D} = \{(\vec{x}_i, y_i)\}$ as a pair of a point and a class label corresponding to it
- Now the linear classifier becomes:

$$f(\vec{x}) = \text{sign}(\vec{w}^T\vec{x} + b)$$



x_1
$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=1$
$\vec{w}$
1
1
x_2
$\vec{w}\cdot\vec{x}+b=0$

Margin in terms of W



$$\frac{w}{\|w\|}\cdot(x_2 - x_1) = \text{width} = \frac{2}{\|w\|}$$

$$w \cdot x_2 + b = 1$$

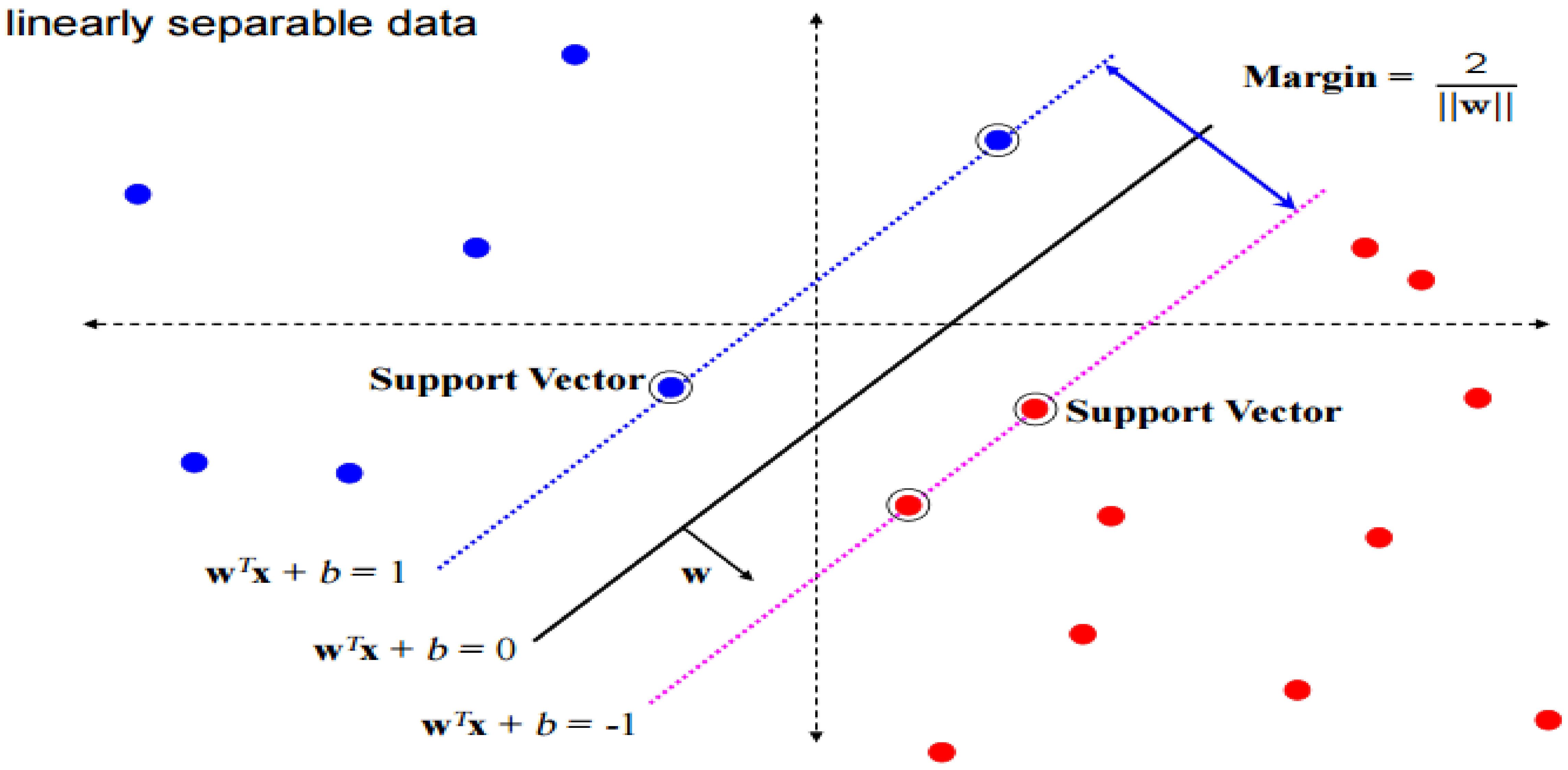
$$w \cdot x_1 + b = -1$$

$$w \cdot x_2 + b - w \cdot x_1 - b = 1 - (-1)$$

$$w \cdot x_2 - w \cdot x_1 = 2$$

$$\frac{w}{\|w\|}(x_2 - x_1) = \frac{2}{\|w\|}$$

Support Vector Machine



SVM as a minimization problem

- Maximizing $2/|\vec{w}|$ is the same as minimizing $|\vec{w}|/2$
- Hence SVM becomes a minimization problem:

Quadratic problem

$$\min_{w,b} \ \frac{1}{2}||w||^2$$

$$s.t. \ y_n(w^T x_n - b) - 1 \geq 0 \quad \forall n$$

Linear constrain

- We are now optimizing a quadratic function subject to linear constraints
- Quadratic optimization problems are a standard, well-known class of mathematical optimization problems, and many algorithms exist for solving them

$$\min \frac{1}{2}\|\mathbf{w}\|^2 \qquad \text{s.t.} \qquad y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0 \;\; \forall_i$$

In order to cater for the constraints in this minimization, we need to allocate them Lagrange multipliers $\boldsymbol{\alpha}$, where $\alpha_i \geq 0 \; \forall_i$:

$$\begin{aligned} L_P &\equiv \frac{1}{2}\|\mathbf{w}\|^2 - \boldsymbol{\alpha}\left[y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \;\forall_i\right] \\ &\equiv \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{L} \alpha_i \left[y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1\right] \\ &\equiv \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{L} \alpha_i y_i(\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^{L} \alpha_i \end{aligned}$$

We wish to find the <u>**w**</u> and <u>**b**</u> which minimizes, and the <u>**α**</u> which maximizes LP(whilst keeping $\alpha_i \geq 0 \; \forall i$). We can do this by differentiating LP with respect to w and b and setting the derivatives to zero:

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{L} \alpha_i y_i \mathbf{x}_i$$

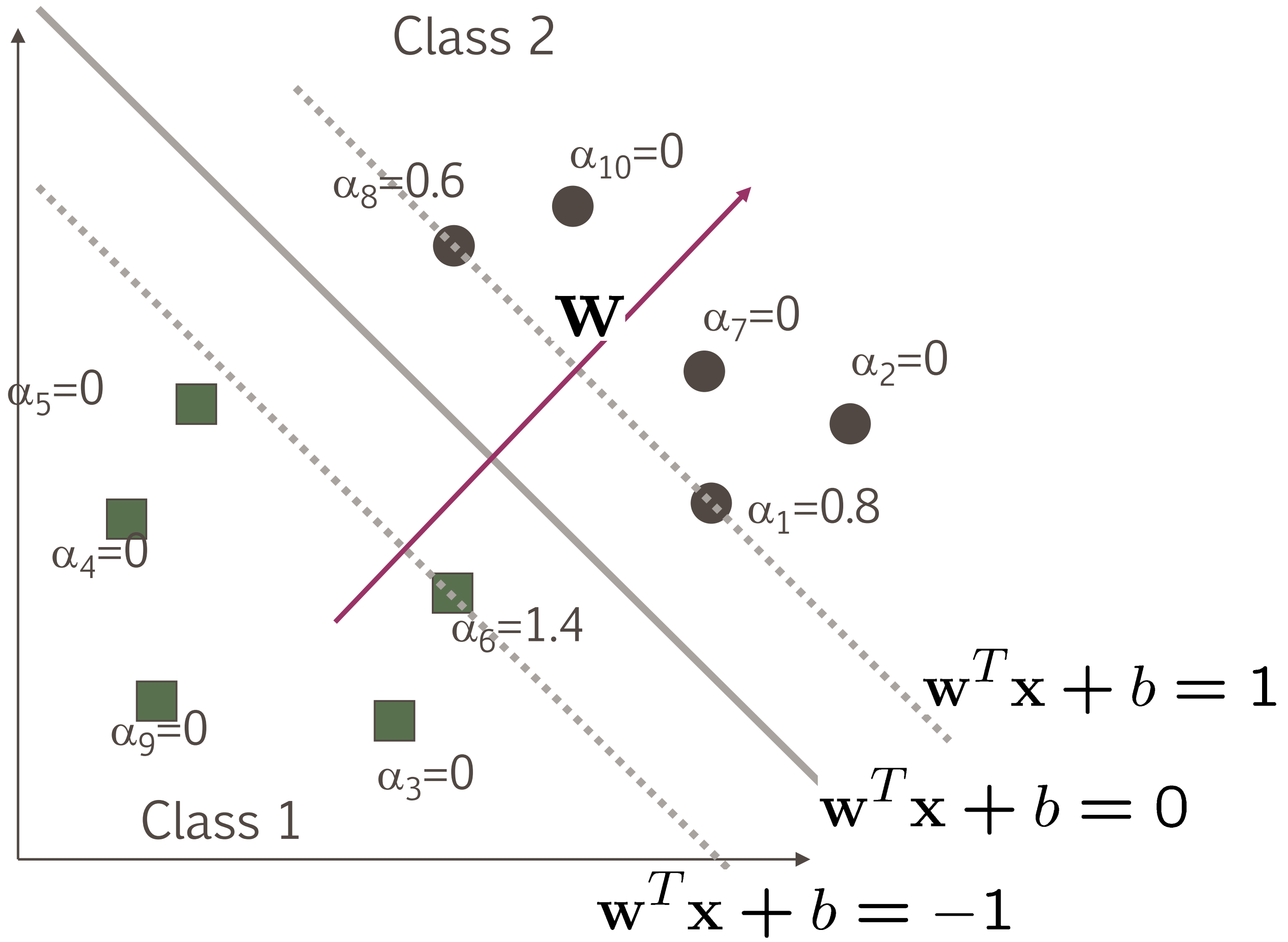
$$\frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_{i=1}^{L} \alpha_i y_i = 0$$

Characteristics of the Solution

- Many of the α_i are zero (see next page for example)
  - $\mathbf{w}$ is a linear combination of a small number of data points
  - This "sparse" representation can be viewed as data compression as in the construction of knn classifier

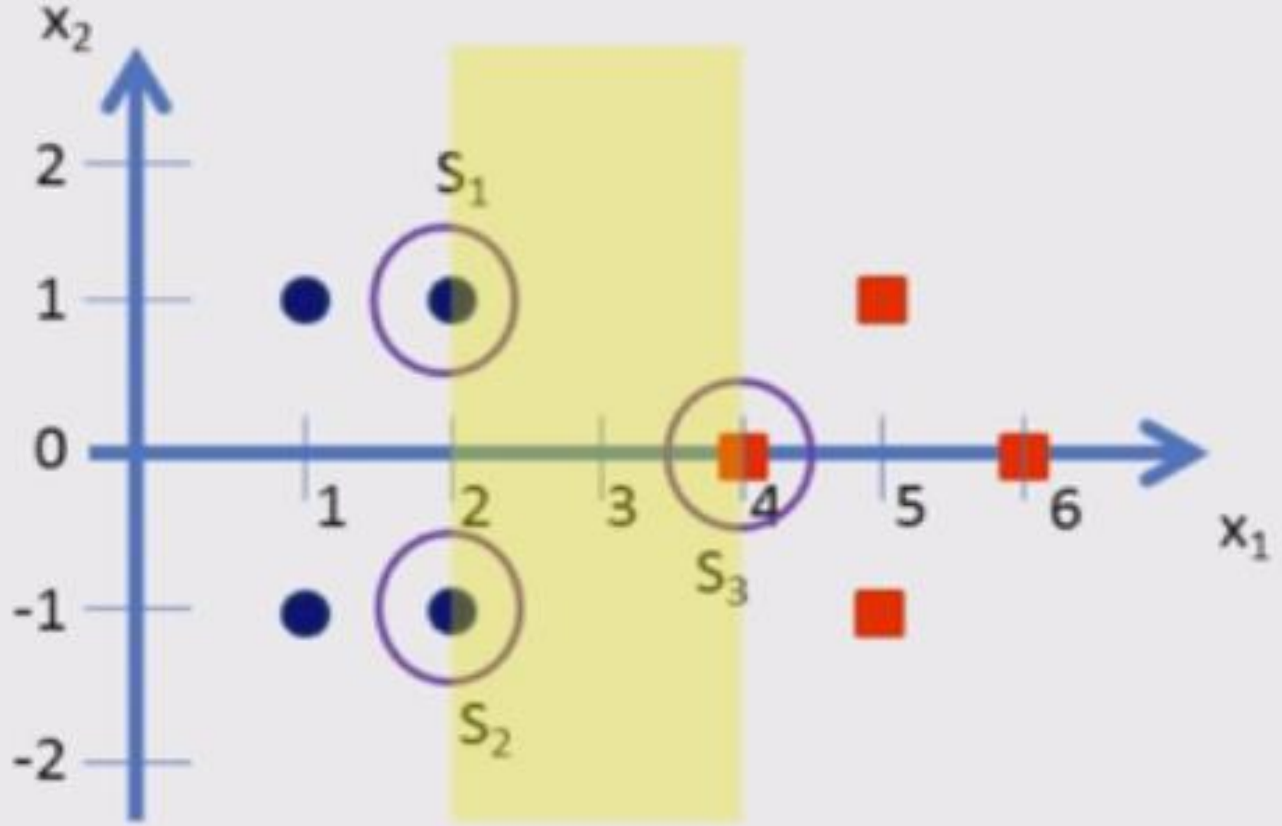
- x_i with non-zero α_i are called support vectors (SV)
  - The decision boundary is determined only by the SV
  - Let t_j (j=1, ..., s) be the indices of the s support vectors. We can write
    $\mathbf{w} = \sum_{j=1}^{s} \alpha_{t_j} y_{t_j} \mathbf{x}_{t_j}$
- For testing with a new data $\mathbf{z}$
  - Compute $\mathbf{w}^T\mathbf{z} + b = \sum_{j=1}^{s} \alpha_{t_j} y_{t_j} (\mathbf{x}_{t_j}^T \mathbf{z}) + b$
    and classify $\mathbf{z}$ as class 1 if the sum is positive, and class 2 otherwise
  - Note: $\mathbf{w}$ need not be formed explicitly

A Geometrical Interpretation



Example

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



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

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

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

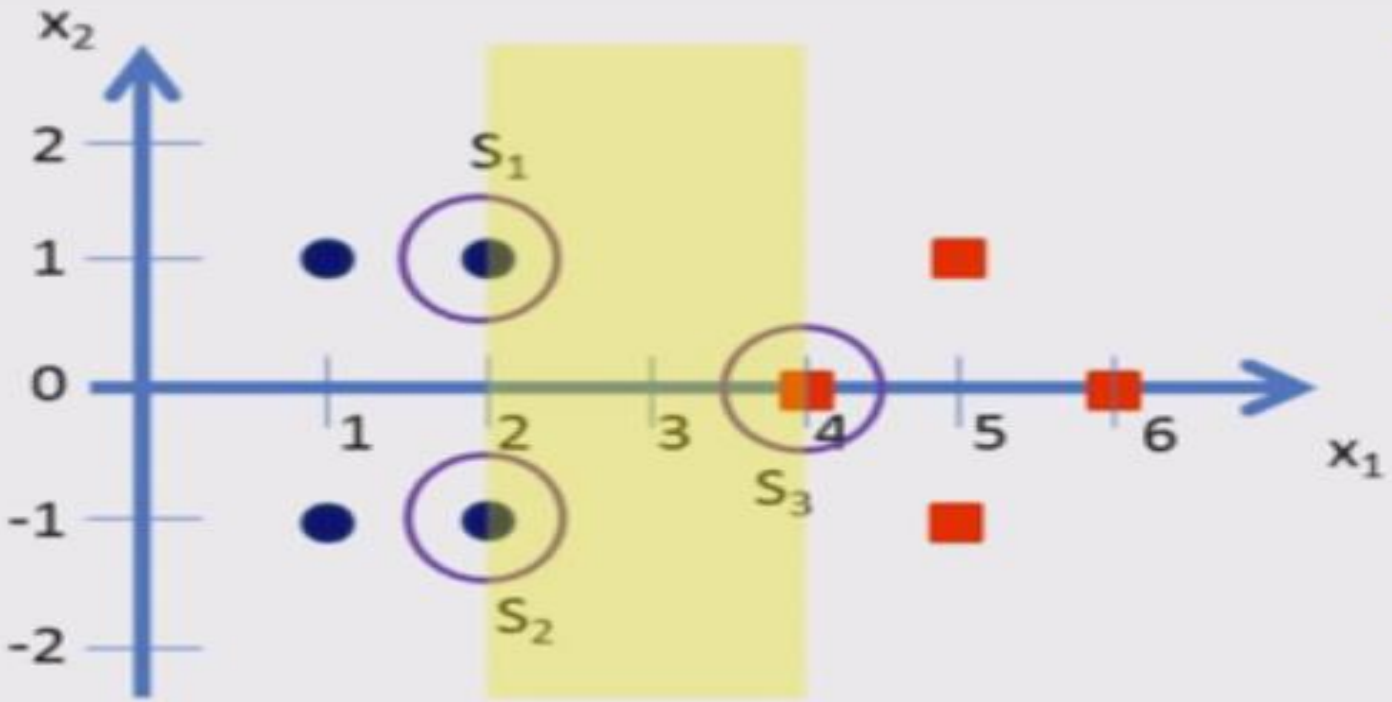
Example

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

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

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

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



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

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

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

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

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

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

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

- Let's substitute the values for $\widetilde{S}_1$, $\widetilde{S}_2$ and $\widetilde{S}_3$ in the above equations.

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

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

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

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

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

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

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

- 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: $\alpha_1 = \alpha_2 = -3.25$ and $\alpha_3 = 3.5$.

$$\alpha_1 = \alpha_2 = -3.25 \text{ and } \alpha_3 = 3.5$$

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

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

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

- Substituting the values we get:

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

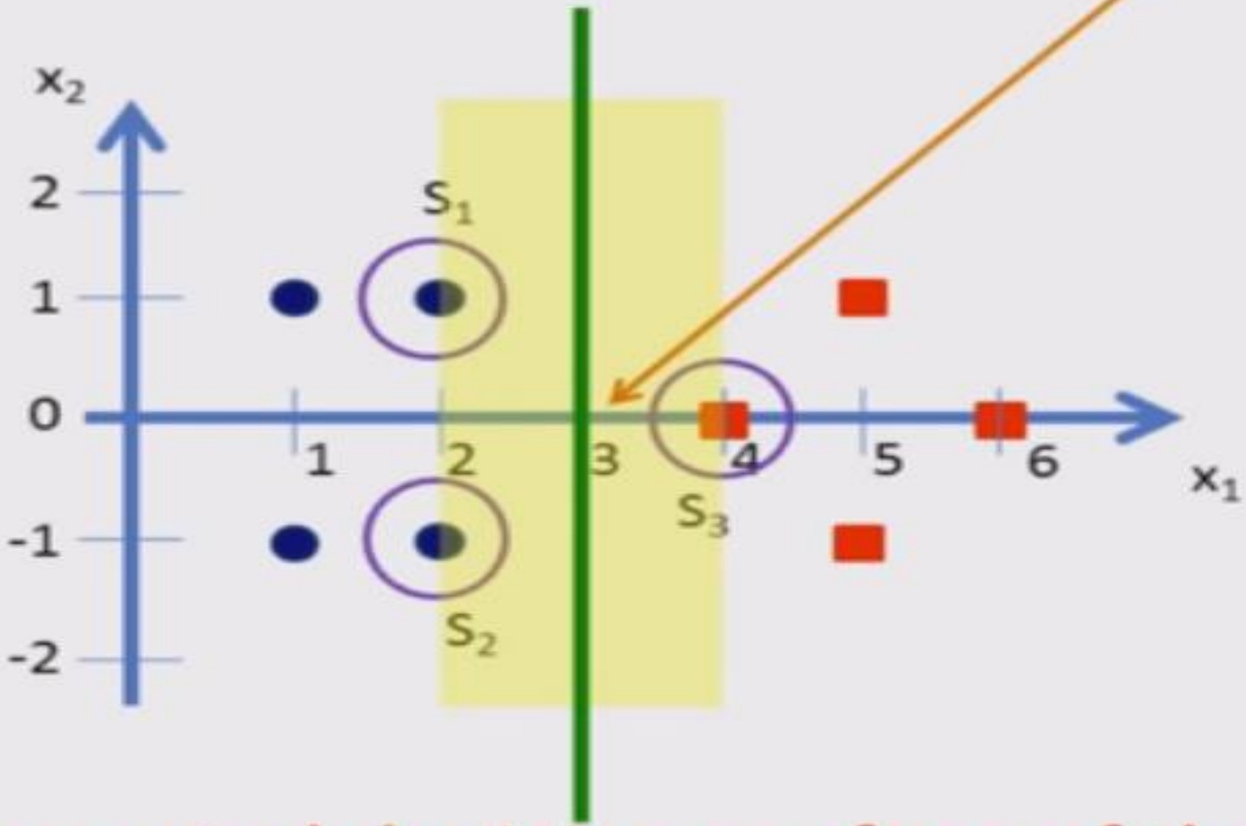
$$\widetilde{w} = (-3.25).\begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} + (-3.25).\begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} + (3.5).\begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ -3 \end{pmatrix}$$

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

- Our vectors are augmented with a bias.
- Hence we can equate the entry in $\widetilde{w}$ 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 $b = -3$.

Support Vector Machines

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



- This is the expected decision surface of the LSVM.

Kernel trick

SVM Algorithm

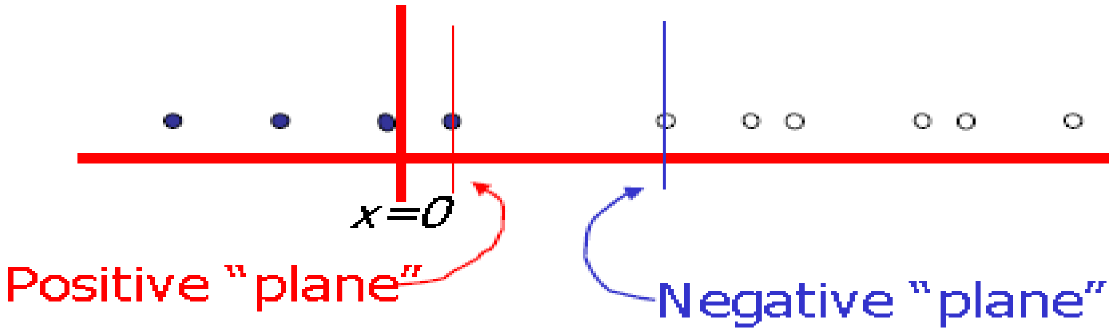
1- Define an optimal hyperplane: maximize margin

2- Extend the above definition for non-linearly separable problems: have a penalty term for misclassifications

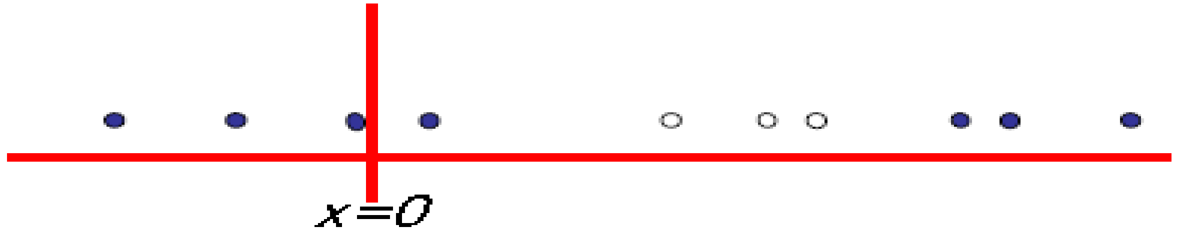
3- Map data to high dimensional space where it is easier to classify with linear decision surfaces: reformulate problem so that data is mapped implicitly to this space

Suppose we're in 1-dimension

Not a big surprise



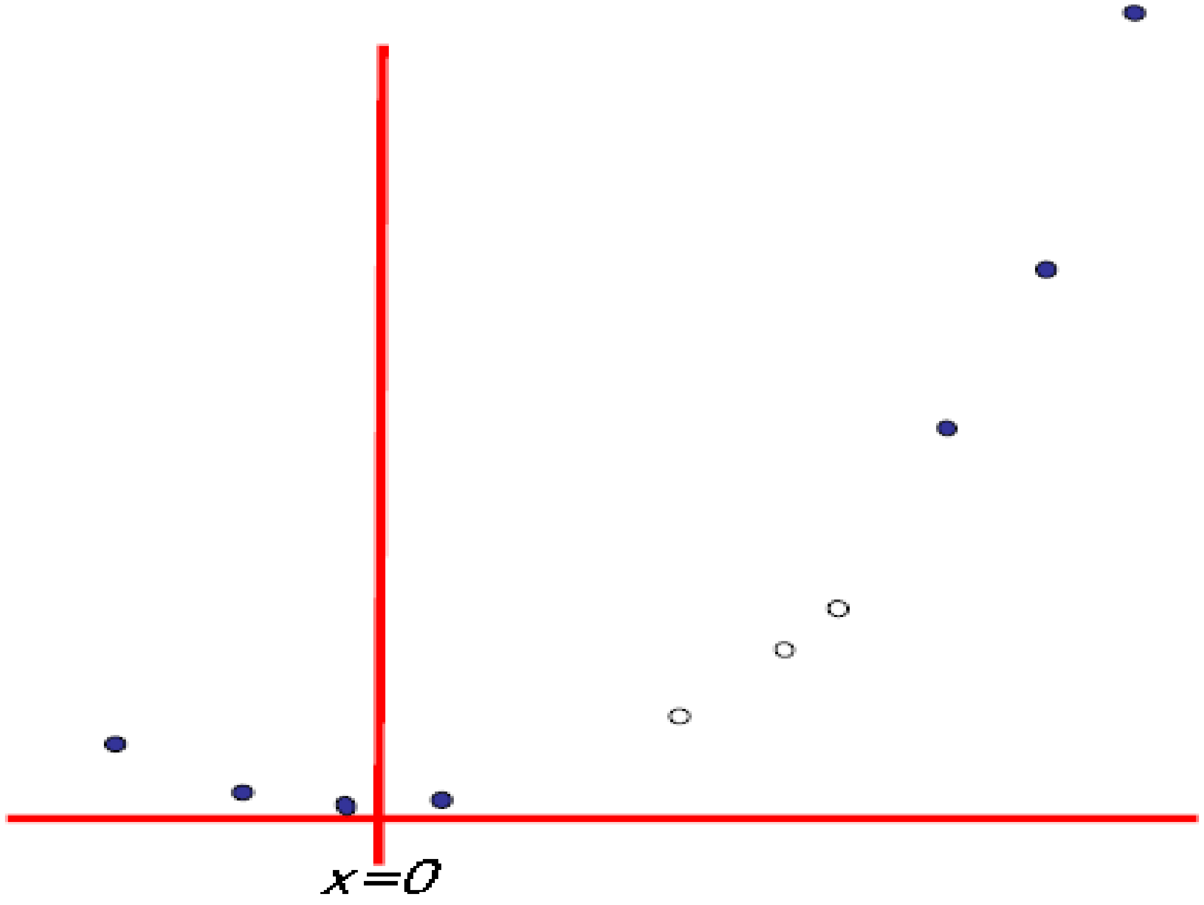
Harder 1-dimensional dataset



That's wiped the smirk off SVM's face.

What can be done about this?

Harder 1-dimensional dataset

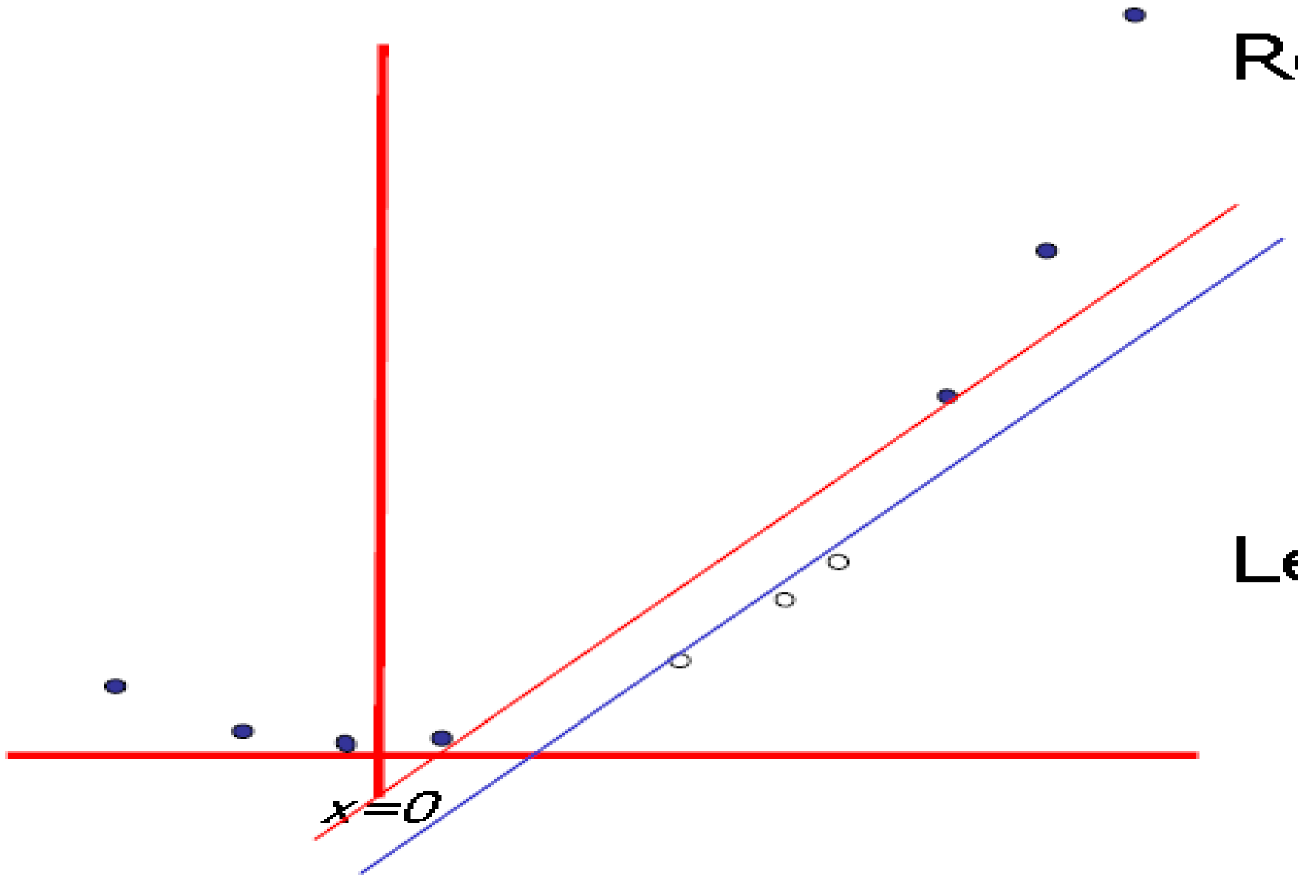


Remember how permitting non-linear basis functions made linear regression so much nicer?

Let's permit them here too

$$\mathbf{z}_k = (x_k, x_k^2)$$

Harder 1-dimensional dataset



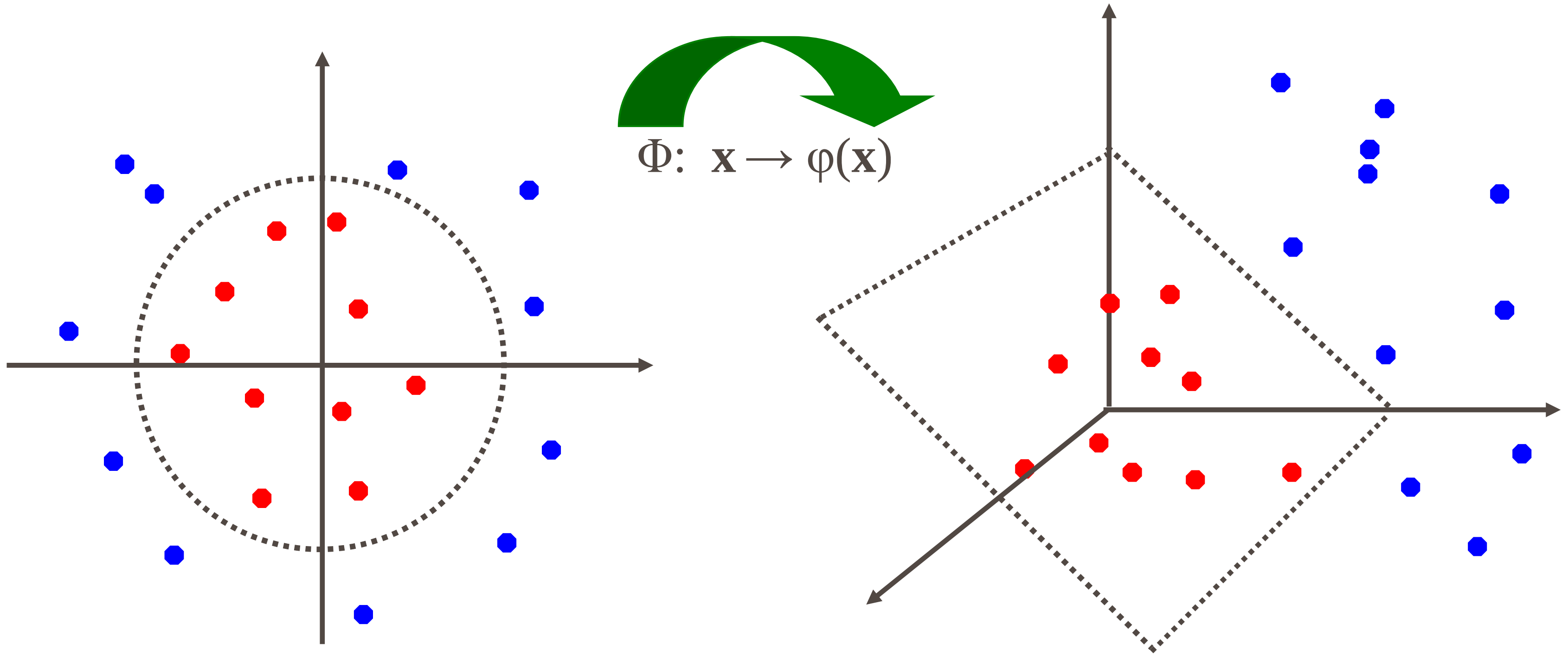
Remember how permitting non-linear basis functions made linear regression so much nicer?

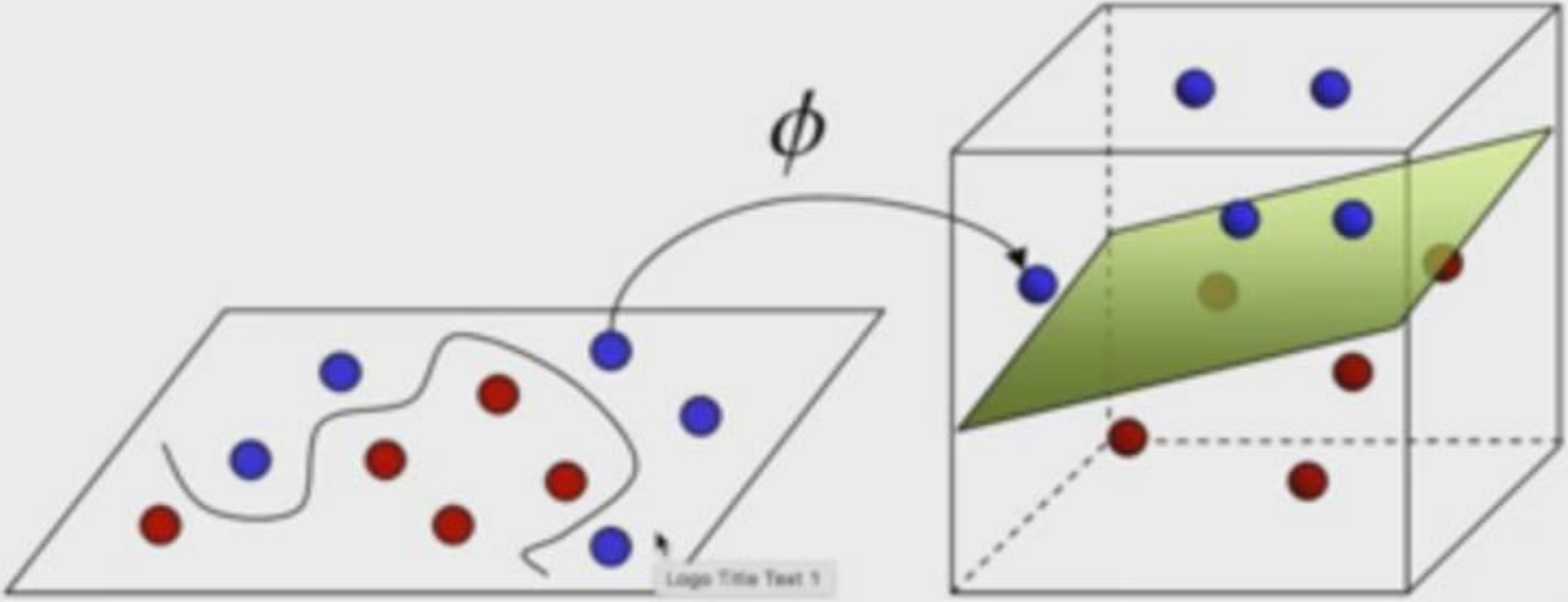
Let's permit them here too

$$\mathbf{z}_k = (x_k, x_k^2)$$

Non-linear SVMs: Feature spaces

- General idea: the original feature space can always be mapped to some higher-dimensional feature space where the training set is separable:





Input Space

Feature Space

SVM for nonlinear reparability

- The simplest way to separate two groups of data is with a straight line, flat plane an N-dimensional hyperplane
- However, there are situations where a nonlinear region can separate the groups more efficiently
- SVM handles this by using a kernel function (nonlinear) to map the data into a *different space* where a hyperplane (linear) cannot be used to do the separation
- It means a non-linear function is learned by a linear learning machine in a high-dimensional feature space while the capacity of the system is controlled by a parameter that does not depend on the dimensionality of the space
- This is called **kernel trick** which means the kernel function transform the data into a higher dimensional feature space to make it possible to perform the linear separation

Kernels

- Why use kernels?
  - Make non-separable problem separable.
  - Map data into better representational space
- Common kernels
  - Linear
  - Polynomial $K(x,z) = (1+x^Tz)^d$
    - Gives feature conjunctions
  - Radial basis function (infinite dimensional space)

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2}$$

- Haven't been very useful in text classification