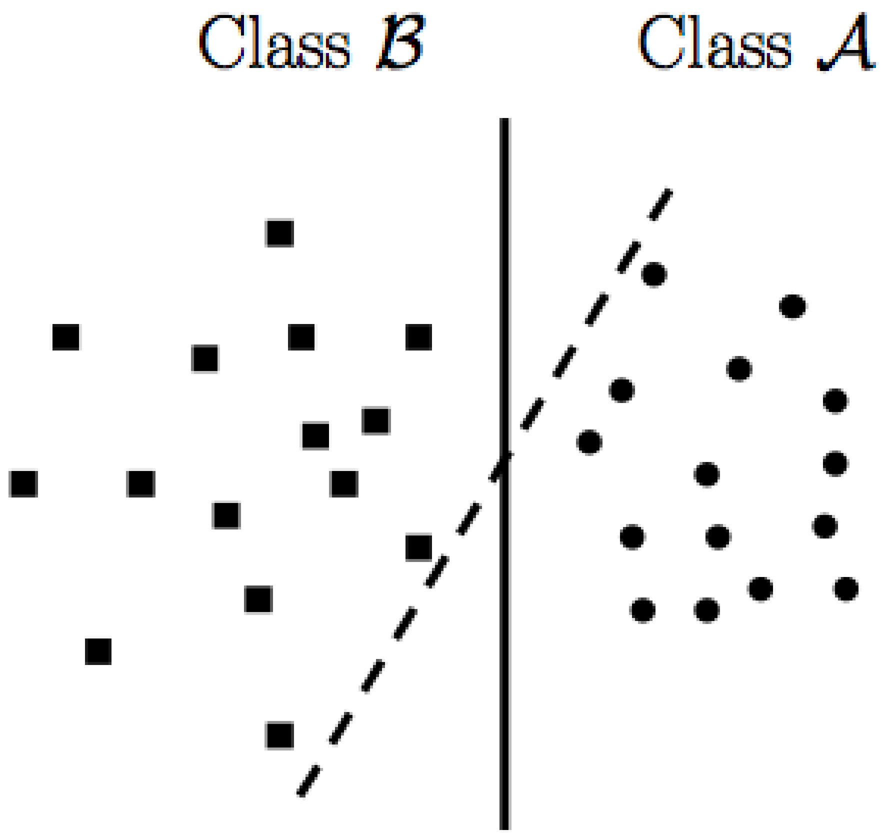
Machine Learning

Lecture # 5
Support Vector Machine

Perceptron



Which plane is best?

Perceptron VS SVM

- The Perceptron does not try to optimize the separation "distance". As long as it finds a hyperplane that separates the two sets, it is good. SVM on the other hand tries to maximize the "support vector", i.e., the distance between two closest opposite sample points.
- The SVM typically tries to use a "kernel function" to project the sample points to high dimension space to make them linearly separable, while the perceptron assumes the sample points are linearly separable.
- SVM Requires more parameters as compared to
  - choice of kernel
  - selection of kernel parameters
  - selection of the value of the margin parameter

Support Vector Machine (SVM)

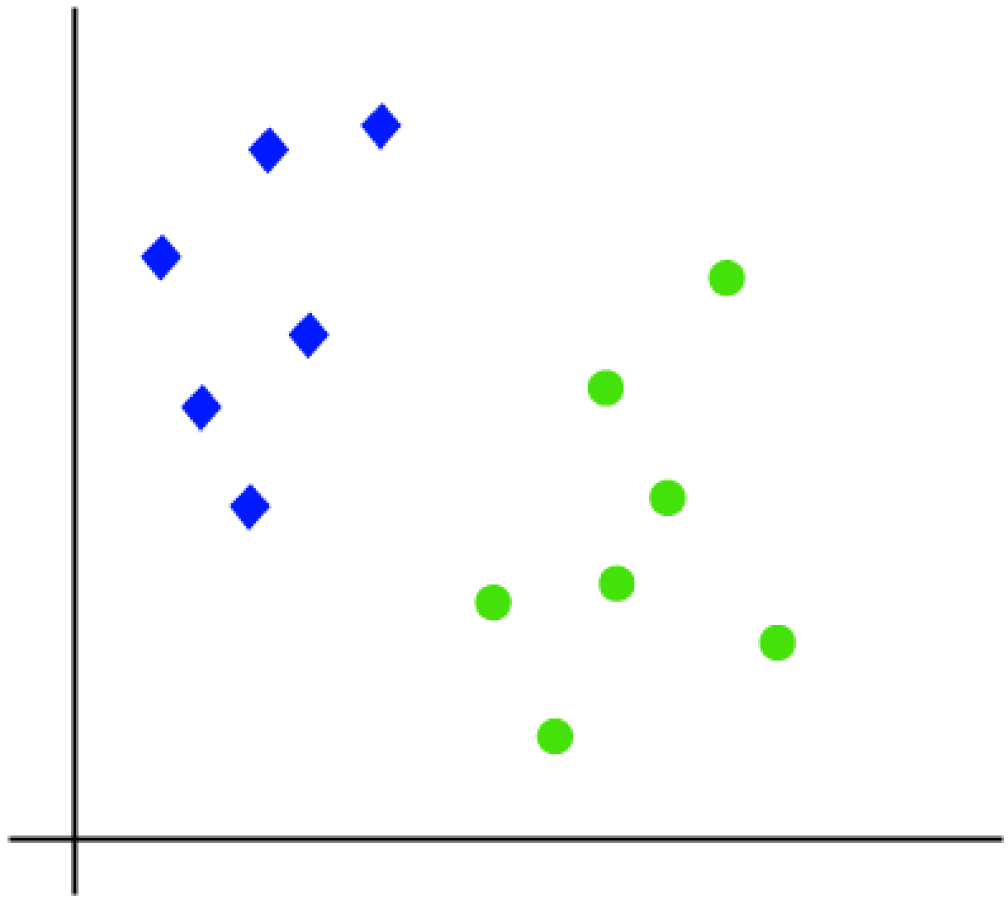
- “Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges.
- However, it is mostly used in classification problems.
- In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate.
- Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at the below snapshot).

How does it work?

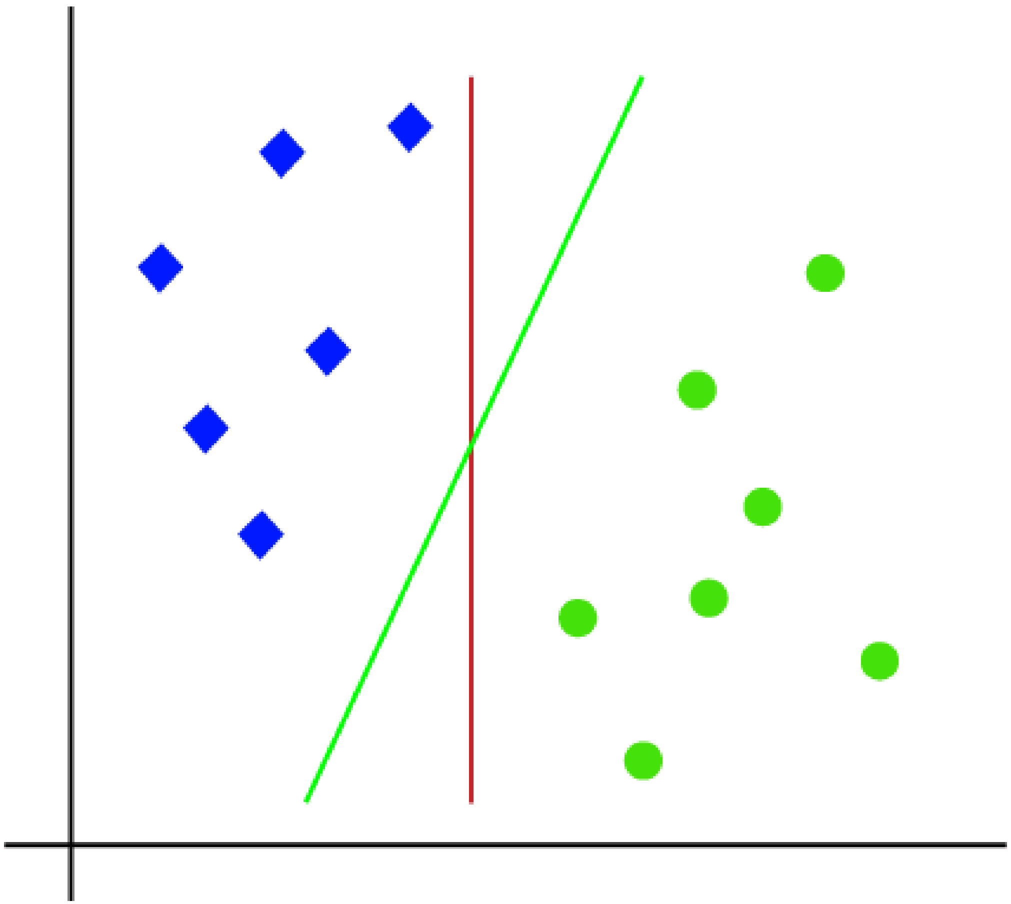
Thumb rule to identify the right hyper-plane

- Select the hyper-plane which segregates the two classes better.
- Maximizing the distances between nearest data point (either class) and hyper-plane. This distance is called as **Margin**.

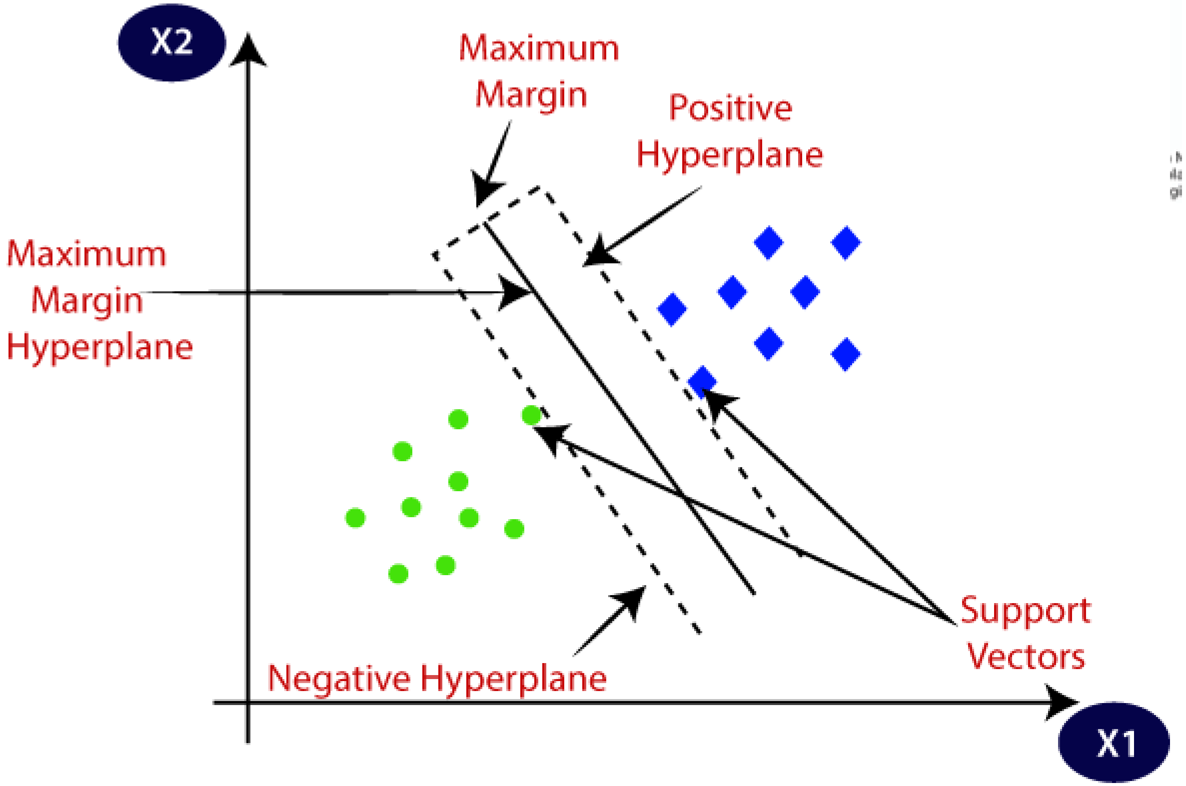
SVM and Margins



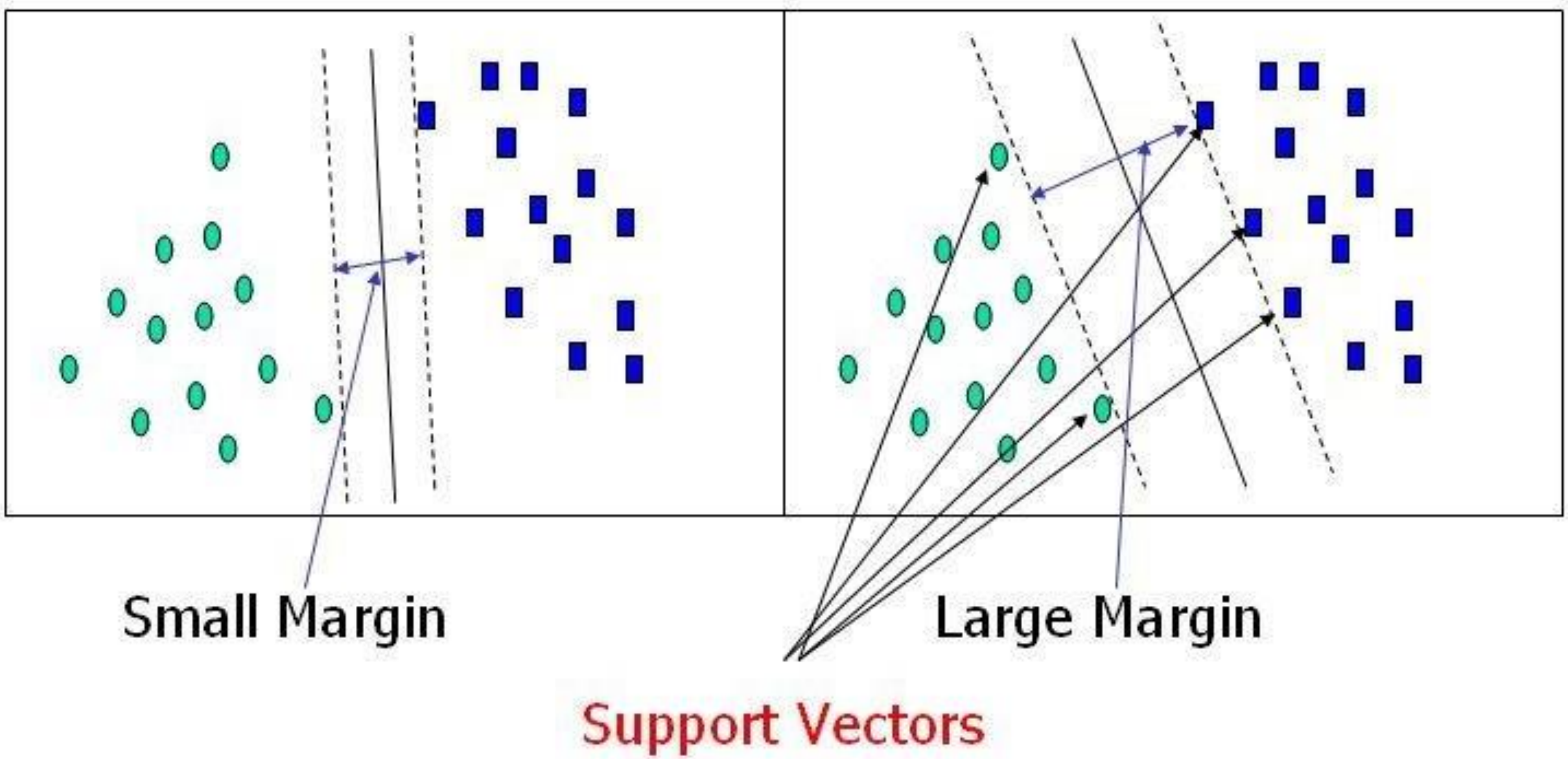
SVM and Margins



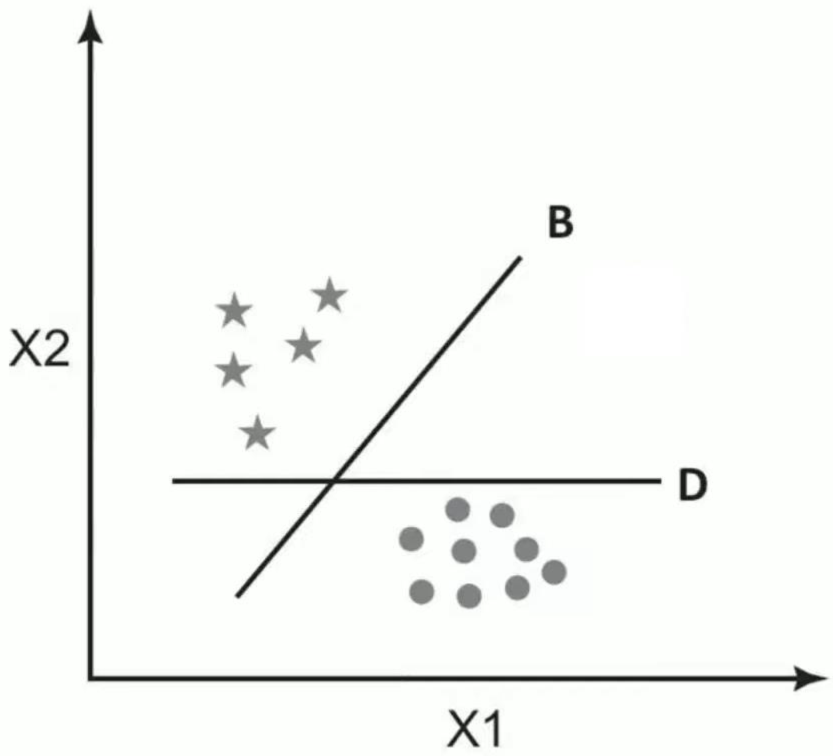
SVM and Margins

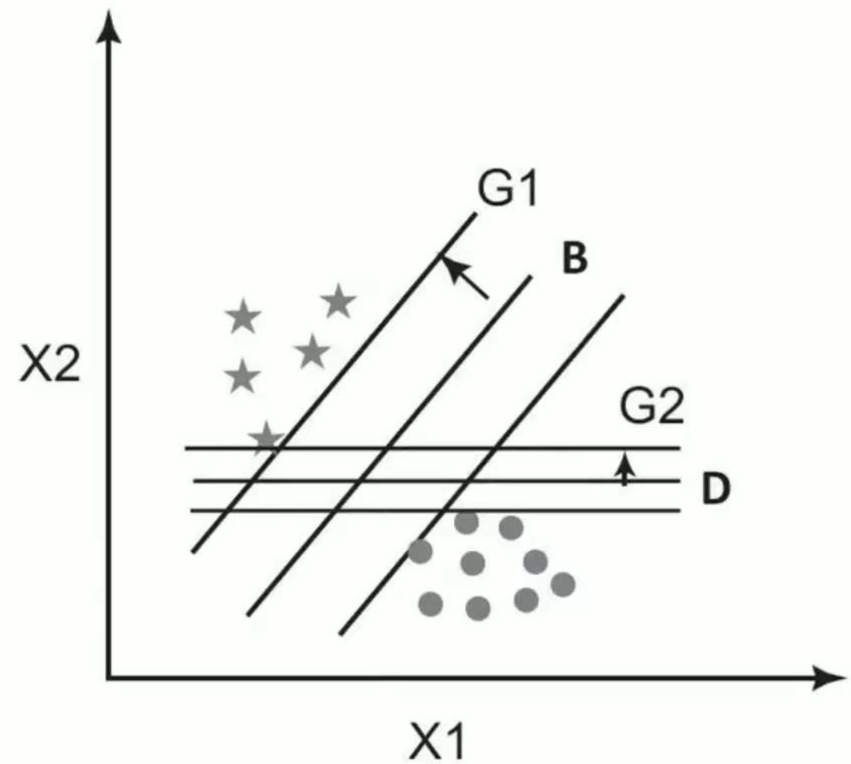


SVM and Margins

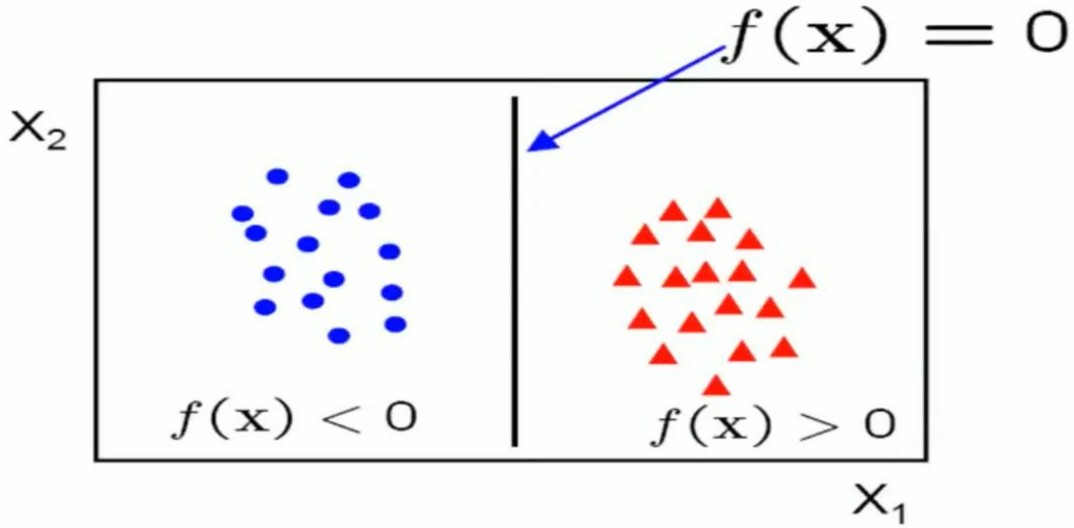


How does it work?



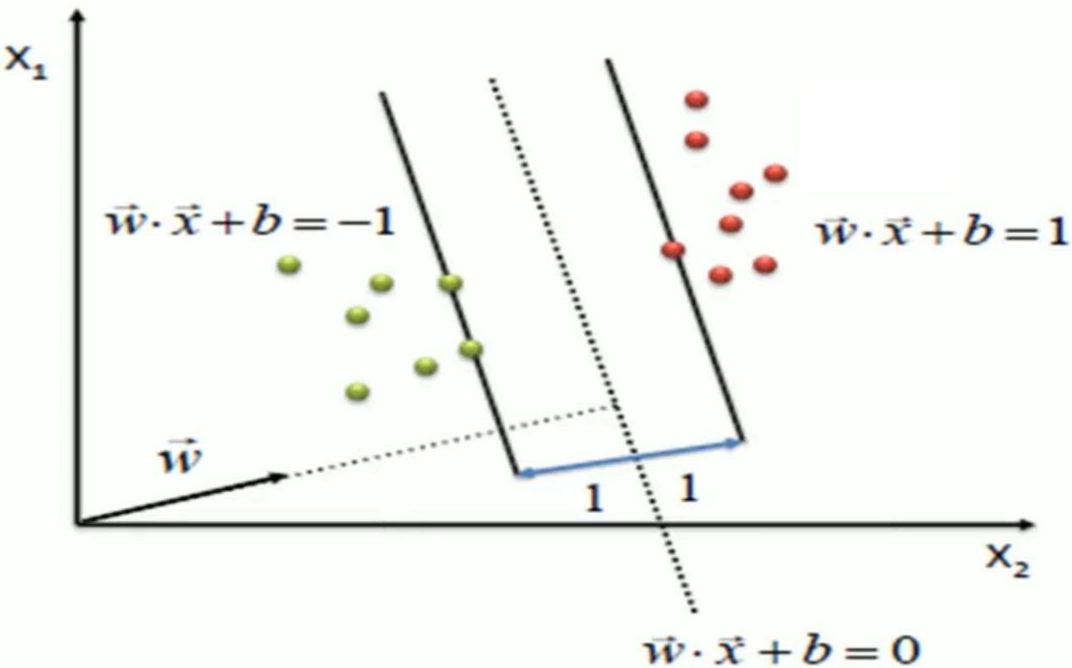


How does it work?



- f(x) = W.X + b
- W is the normal to the line, X is input vector and b the bias
- W is known as the weight vector

How does it work?



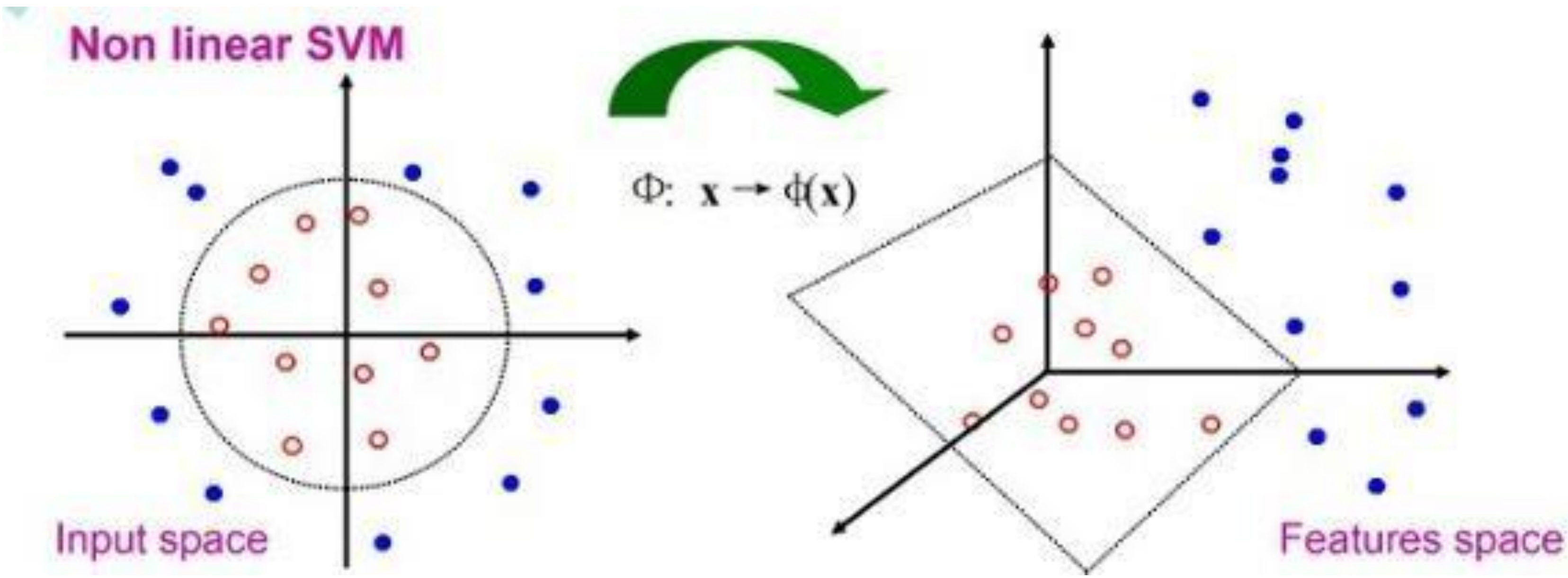
$$\boxed{\max \frac{2}{\|w\|}}$$

s.t.

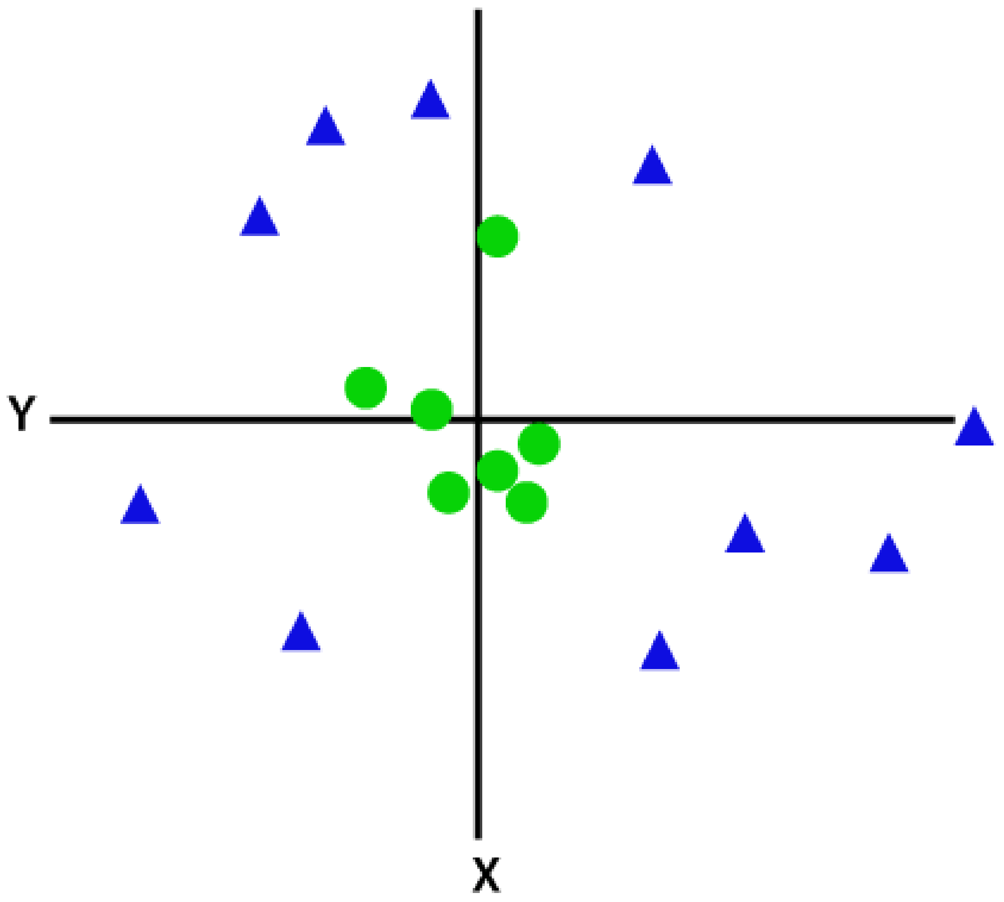
$(w\cdot x+b)\geq 1, \forall x$ of class 1

$(w\cdot x+b)\leq -1, \forall x$ of class 2

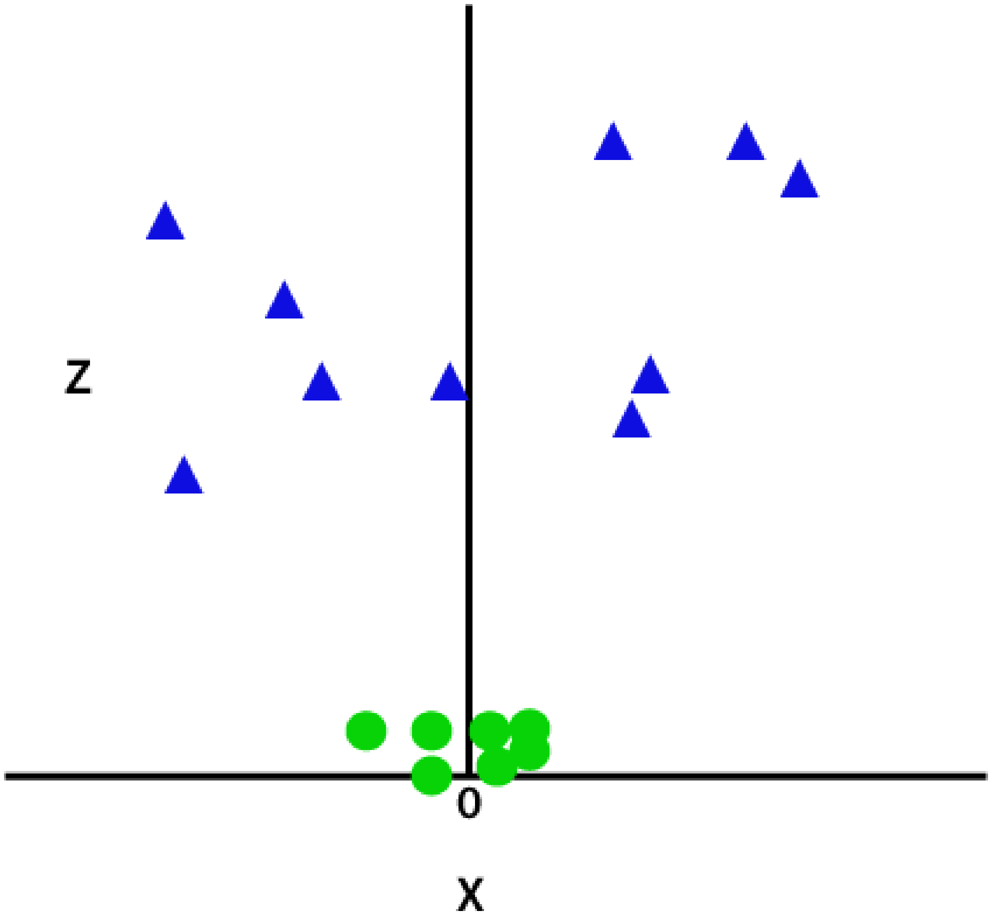
SVM for Nonlinear Data



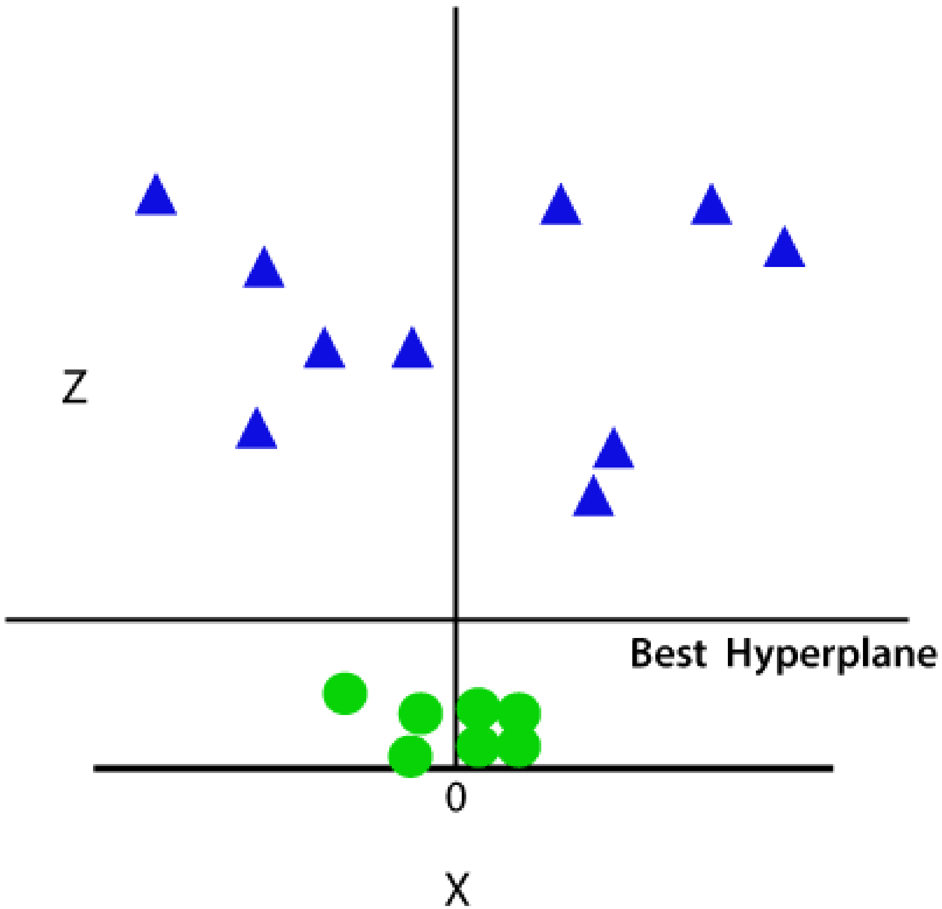
SVM for Nonlinear Data



SVM for Nonlinear Data



SVM for Nonlinear Data



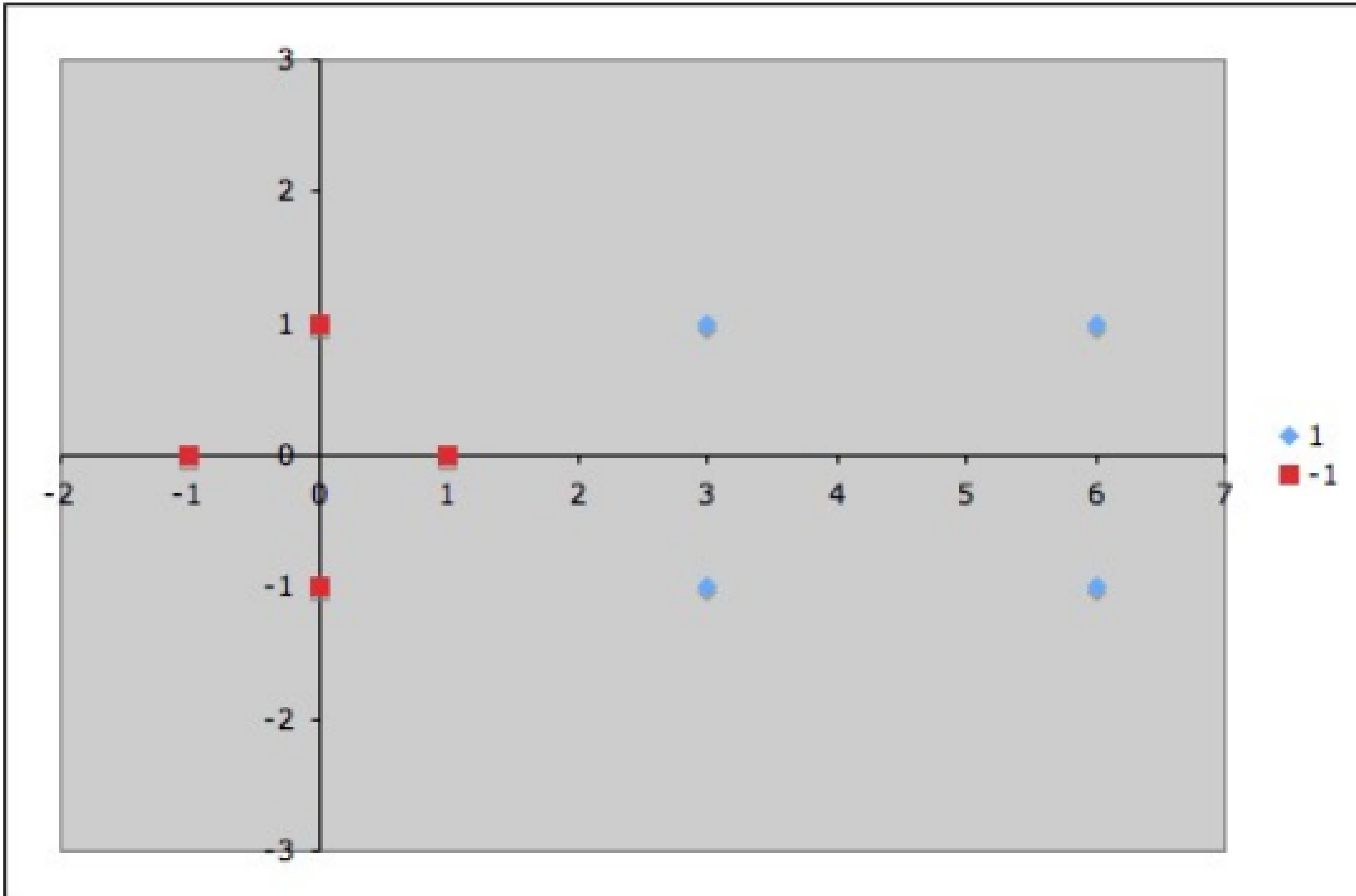
Linear Example

Suppose we are given the following positively labeled data points,

$$\left\{ \begin{pmatrix} 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ -1 \end{pmatrix}, \begin{pmatrix} 6 \\ 1 \end{pmatrix}, \begin{pmatrix} 6 \\ -1 \end{pmatrix} \right\}$$

and the following negatively labeled data points,

$$\left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \end{pmatrix} \right\}$$

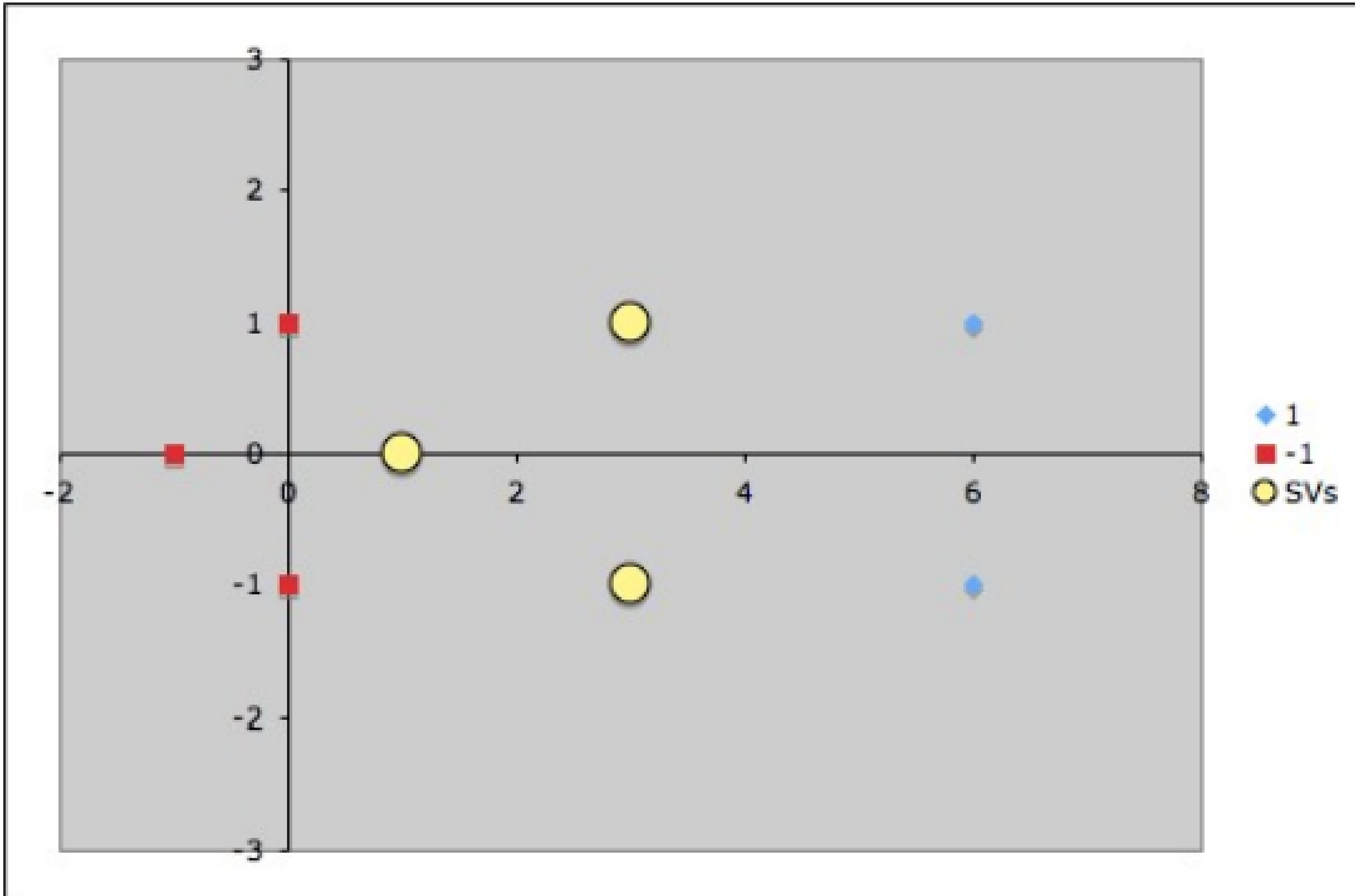


3
2
1
0
-1
-2
-3
-2
-1
0
1
2
3
4
5
6
7
1
-1

Linear Example

- By inspection, it should be obvious that there are **three** support vectors,

$$\left\{ s_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, s_2 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}, s_3 = \begin{pmatrix} 3 \\ -1 \end{pmatrix} \right\}$$



3
2
1
0
-1
-2
-3
-2
0
2
4
6
8
1
-1
SVs

Linear Example

- Each vector is augmented with a 1 as a bias input

- So, $s_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, then $\widetilde{s_1} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$

- Similarly,

- $s_2 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$, then $\widetilde{s_2} = \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix}$ and $s_3 = \begin{pmatrix} 3 \\ -1 \end{pmatrix}$, then $\widetilde{s_3} = \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}$

Linear Example

$$\begin{aligned}
\alpha_1 \tilde{s_1} \cdot \tilde{s_1} + \alpha_2 \tilde{s_2} \cdot \tilde{s_1} + \alpha_3 \tilde{s_3} \cdot \tilde{s_1} &= -1 \\
\alpha_1 \tilde{s_1} \cdot \tilde{s_2} + \alpha_2 \tilde{s_2} \cdot \tilde{s_2} + \alpha_3 \tilde{s_3} \cdot \tilde{s_2} &= +1 \\
\alpha_1 \tilde{s_1} \cdot \tilde{s_3} + \alpha_2 \tilde{s_2} \cdot \tilde{s_3} + \alpha_3 \tilde{s_3} \cdot \tilde{s_3} &= +1
\end{aligned}$$

$$\alpha_1(1+0+1) + \alpha_2(3+0+1) + \alpha_3(3+0+1) = -1$$

$$\alpha_1(3+0+1) + \alpha_2(9+1+1) + \alpha_3(9-1+1) = 1$$

$$\alpha_1(3+0+1) + \alpha_2(9-1+1) + \alpha_3(9+1+1) = 1$$

$$\alpha_1 \begin{pmatrix}1\\0\\1\end{pmatrix}\begin{pmatrix}1\\0\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}3\\1\\1\end{pmatrix}\begin{pmatrix}1\\0\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}3\\-1\\1\end{pmatrix}\begin{pmatrix}1\\0\\1\end{pmatrix} = -1$$

$$\alpha_1 \begin{pmatrix}1\\0\\1\end{pmatrix}\begin{pmatrix}3\\1\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}3\\1\\1\end{pmatrix}\begin{pmatrix}3\\1\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}3\\-1\\1\end{pmatrix}\begin{pmatrix}3\\1\\1\end{pmatrix} = 1$$

$$\alpha_1 \begin{pmatrix}1\\0\\1\end{pmatrix}\begin{pmatrix}3\\-1\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}3\\1\\1\end{pmatrix}\begin{pmatrix}3\\-1\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}3\\-1\\1\end{pmatrix}\begin{pmatrix}3\\-1\\1\end{pmatrix} = 1$$

$$2\alpha_1 + 4\alpha_2 + 4\alpha_3 = -1$$

$$4\alpha_1 + 11\alpha_2 + 9\alpha_3 = 1$$

$$4\alpha_1 + 9\alpha_2 + 11\alpha_3 = 1$$

$$\alpha_1 = -3.5$$

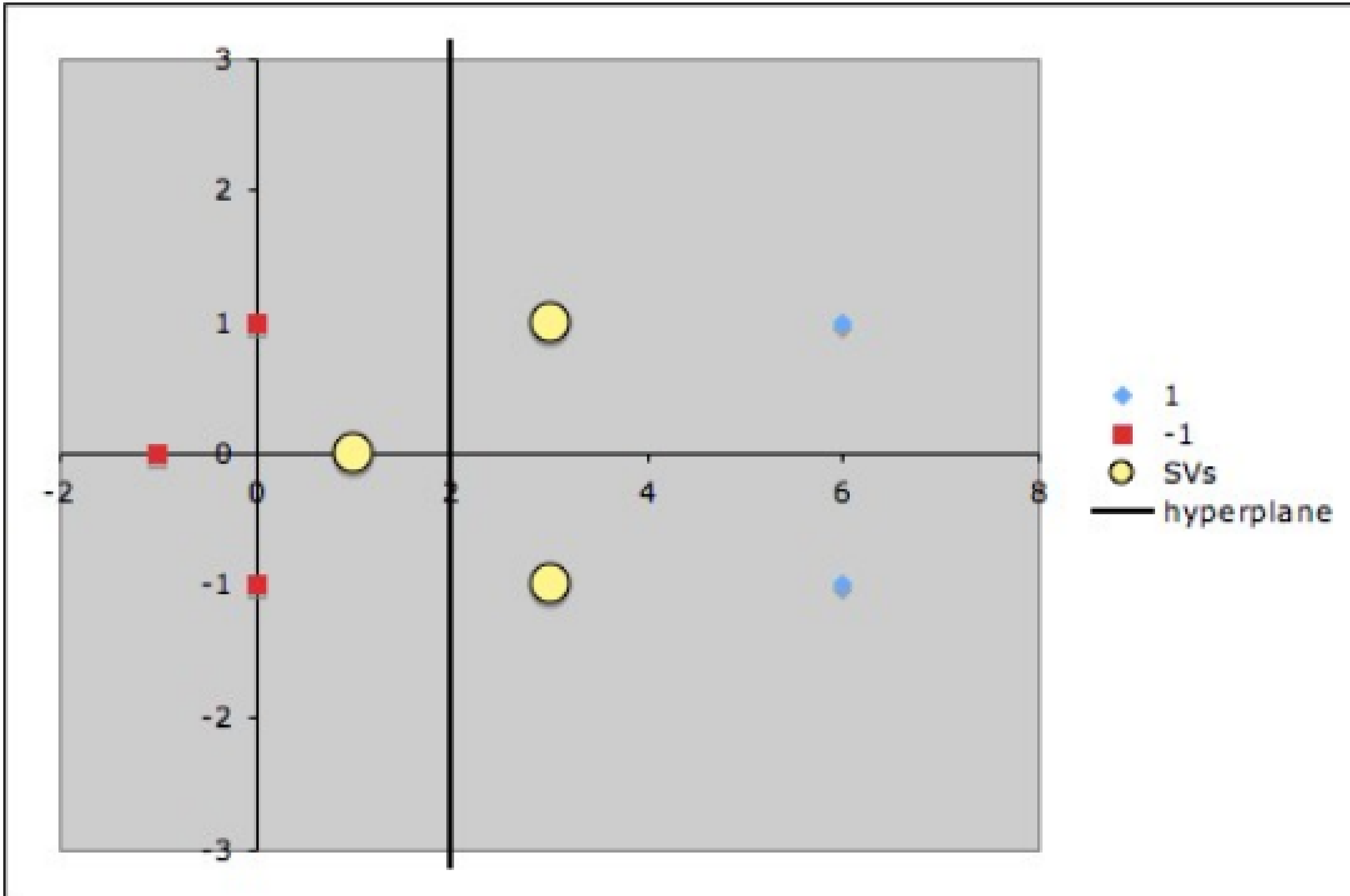
$$\alpha_2 = 0.75$$

$$\alpha_3 = 0.75$$

Linear Example

$$\begin{aligned}\tilde{w} &= \sum_i \alpha_i \tilde{s}_i \\ &= -3.5\begin{pmatrix}1\\0\\1\end{pmatrix} + 0.75\begin{pmatrix}3\\1\\1\end{pmatrix} + 0.75\begin{pmatrix}3\\-1\\1\end{pmatrix} \\ &= \begin{pmatrix}1\\0\\-2\end{pmatrix}\end{aligned}$$

- Finally, remembering that our vectors are augmented with a bias.
- We can equate the last entry in $\tilde{w}$ as the hyperplane offset b and write the separating
- Hyperplane equation $\boldsymbol{y} = \boldsymbol{wx} + \boldsymbol{b}$
- with $\boldsymbol{w} = \binom{\mathbf{1}}{\mathbf{0}}$ and $\boldsymbol{b} = -\mathbf{2}$.



3
2
1
0
-1
-2
-3
-2
0
2
4
6
8
1
-1
SVs
hyperplane

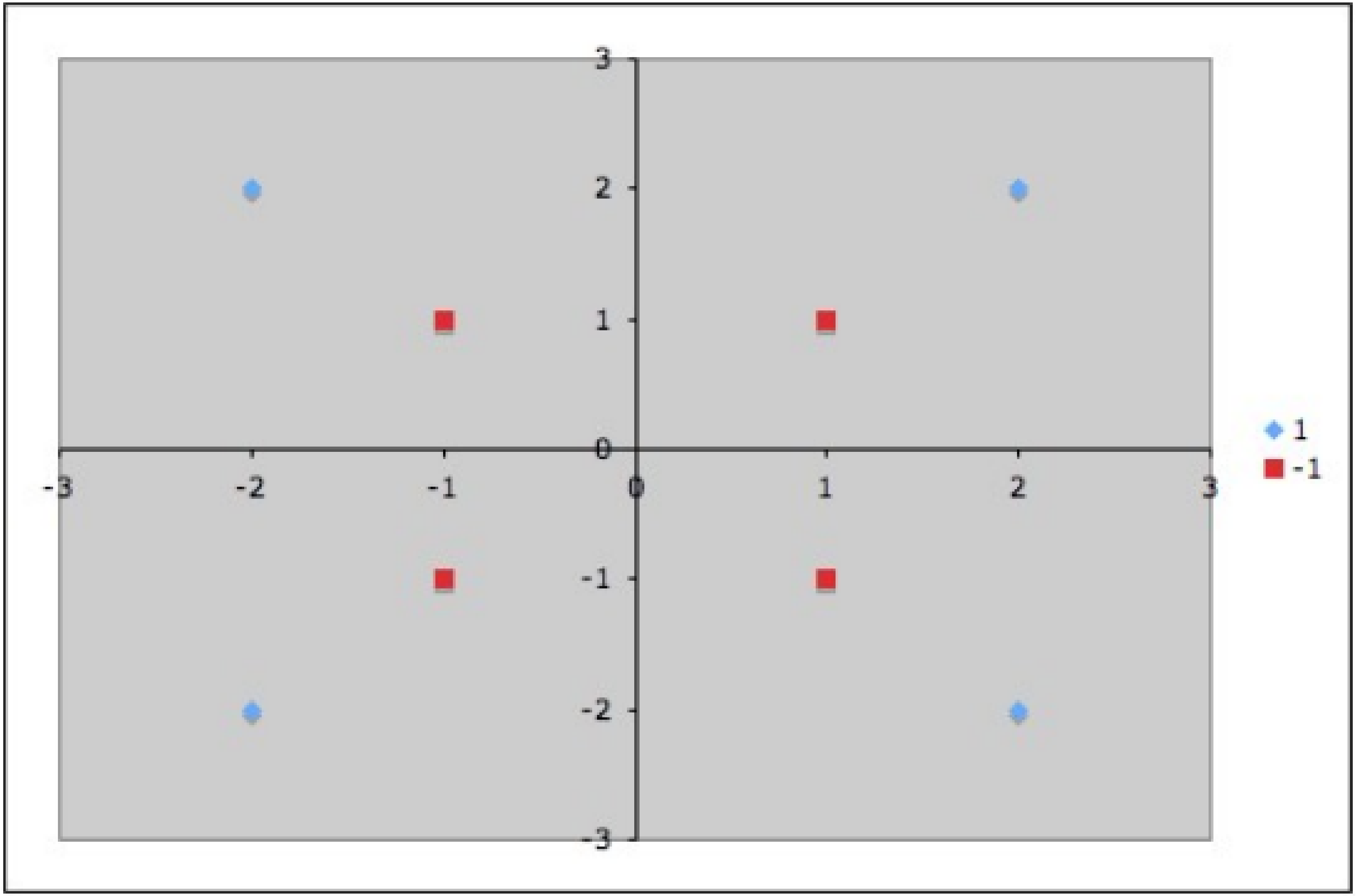
Non-Linear Example

Suppose we are given the following positively labeled data points,

$$\left\{ \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ -2 \end{pmatrix}, \begin{pmatrix} -2 \\ -2 \end{pmatrix}, \begin{pmatrix} -2 \\ 2 \end{pmatrix} \right\}$$

and the following negatively labeled data points,

$$\left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix} \right\}$$


3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
1
-1

Non-Linear Example

- Our goal, again, is to discover a separating hyperplane that accurately discriminates the two classes.
- Of course, it is obvious that no such hyperplane exists in the input space
- Therefore, we must use a nonlinear SVM (that is, we need to convert data from one feature space to another.

$$\Phi_1\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{cases} \begin{pmatrix} 4 - x_2 + |x_1 - x_2| \\ 4 - x_1 + |x_1 - x_2| \end{pmatrix} & \text{if } \sqrt{x_1^2 + x_2^2} > 2 \\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} & \text{otherwise} \end{cases}$$

Non-Linear Example

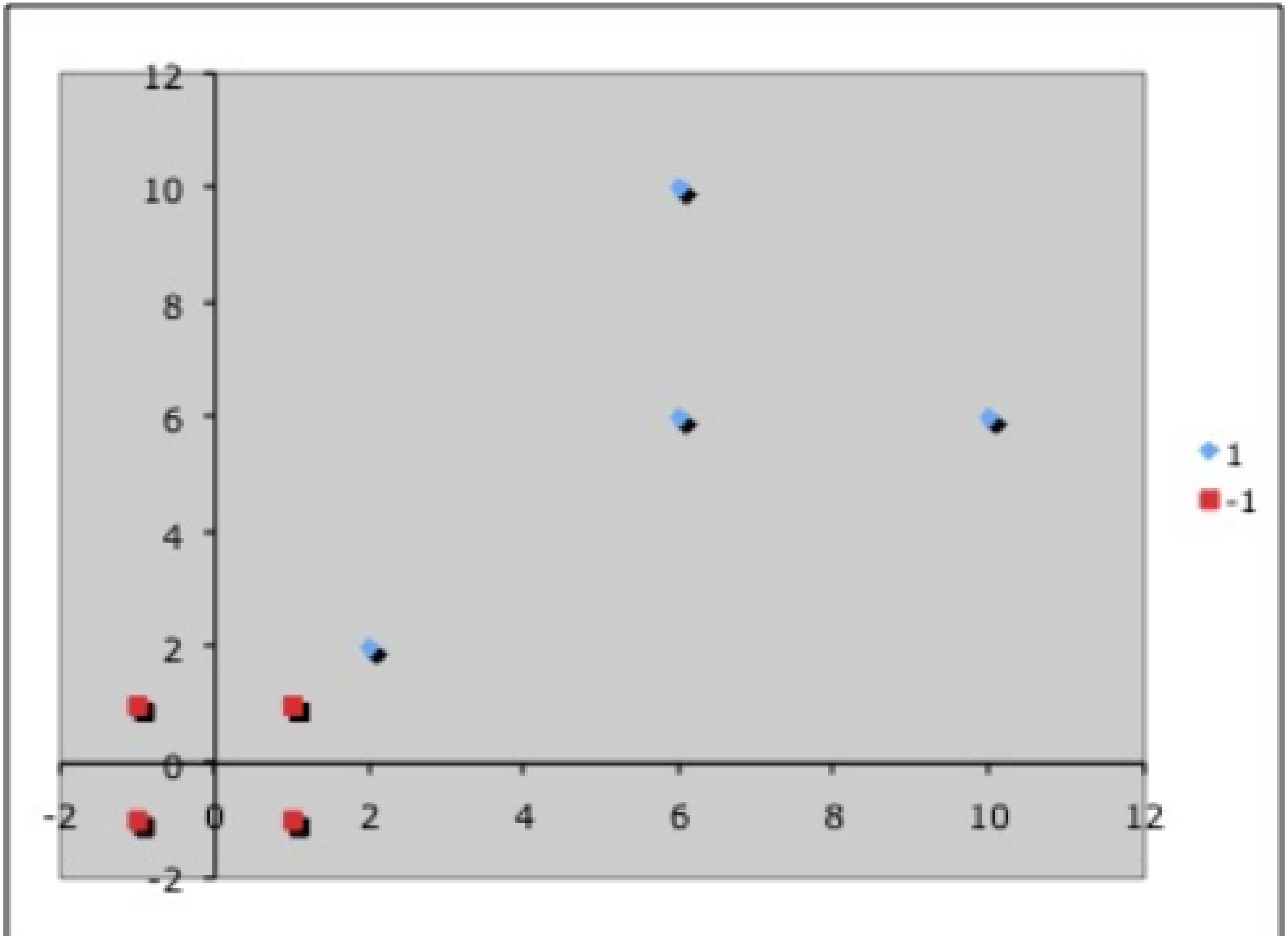
$$\Phi_1\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{cases} \begin{pmatrix} 4 - x_2 + |x_1 - x_2| \\ 4 - x_1 + |x_1 - x_2| \end{pmatrix} & \text{if } \sqrt{x_1^2 + x_2^2} > 2 \\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} & \text{otherwise} \end{cases}$$

Positive Examples

$$\left\{\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ -2 \end{pmatrix}, \begin{pmatrix} -2 \\ -2 \end{pmatrix}, \begin{pmatrix} -2 \\ 2 \end{pmatrix}\right\} \rightarrow \left\{\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 10 \\ 6 \end{pmatrix}, \begin{pmatrix} 6 \\ 6 \end{pmatrix}, \begin{pmatrix} 6 \\ 10 \end{pmatrix}\right\}$$

Negative Examples

$$\left\{\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix}\right\} \rightarrow \left\{\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix}\right\}$$


12
10
8
6
4
2
0
-2
-2
0
2
4
6
8
10
12
1
-1

Non-Linear Example

- Now we can easily identify the support vectors,

$$\left\{ s_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, s_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \right\}$$

- Each vector is augmented with a 1 as a bias input

$$\widetilde{s_1} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \qquad \widetilde{s_2} = \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}$$

Non-Linear Example

$$\alpha_1 \tilde{s_1} \cdot \tilde{s_1} + \alpha_2 \tilde{s_2} \cdot \tilde{s_1} = -1$$

$$\alpha_1 \tilde{s_1} \cdot \tilde{s_2} + \alpha_2 \tilde{s_2} \cdot \tilde{s_2} = +1$$

$$\alpha_1(1+1+1) + \alpha_2(2+2+1) = -1$$

$$\alpha_1(2+2+1) + \alpha_2(4+4+1) = 1$$

$$\alpha_1 \begin{pmatrix}1\\1\\1\end{pmatrix}\begin{pmatrix}1\\1\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}2\\2\\1\end{pmatrix}\begin{pmatrix}1\\1\\1\end{pmatrix} = -1$$

$$3\alpha_1 + 5\alpha_2 = -1$$

$$5\alpha_1 + 9\alpha_2 = 1$$

$$\alpha_1 \begin{pmatrix}1\\1\\1\end{pmatrix}\begin{pmatrix}2\\2\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}2\\2\\1\end{pmatrix}\begin{pmatrix}2\\2\\1\end{pmatrix} = 1$$

$$\alpha_1 = -7$$

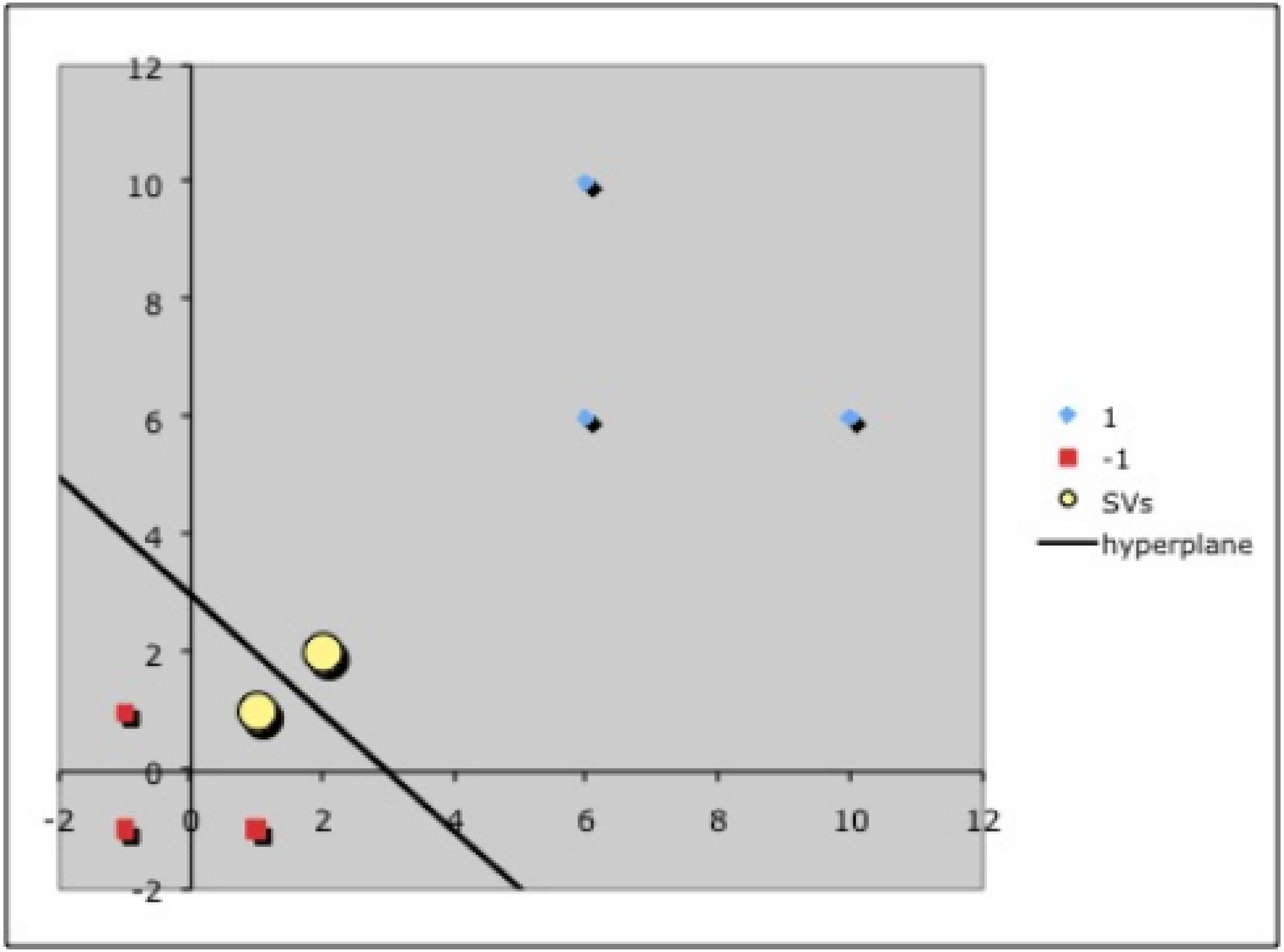
$$\alpha_2 = 4$$

Non-Linear Example

$$\tilde{w} = \sum_i \alpha_i \tilde{s}_i = -7\begin{pmatrix}1\\1\\1\end{pmatrix} + 4\begin{pmatrix}2\\2\\1\end{pmatrix}$$

$$= \begin{pmatrix}1\\1\\-3\end{pmatrix}$$

- Finally, remembering that our vectors are augmented with a bias.
- We can equate the last entry in $\widetilde{w}$ as the hyperplane offset b and write the separating
- Hyperplane equation $\boldsymbol{y} = \boldsymbol{wx} + \boldsymbol{b}$
- with $\boldsymbol{w} = \binom{\mathbf{1}}{\mathbf{1}}$ and $\boldsymbol{b} = -\mathbf{3}$.


12
10
8
6
4
2
0
-2
-2
0
2
4
6
8
10
12
1
-1
SVs
hyperplane

Advantages of SVM

- The main strength of SVM is that they work well even when the number of SVM features is much larger than the number of instances.
- It can work on datasets with huge feature space, such is the case in spam filtering, where a large number of words are the potential signifiers of a message being spam.
- Even when the optimal decision boundary is a nonlinear curve, the SVM transforms the variables to create new dimensions such that the representation of the classifier is a linear function of those transformed dimensions of the data.
- SVMs are conceptually easy to understand. They create an easy-to-understand linear classifier.
- SVMs are now available with almost all data analytics toolsets.

Disadvantages of SVM

- The SVM technique has two major constraints
  - It works well only with real numbers, i.e., all the data points in all the dimensions must be defined by numeric values only,
  - It works only with binary classification problems. One can make a series of cascaded SVMs to get around this constraint.
- Training the SVMs is an inefficient and time consuming process, when the data is large.
- It does not work well when there is much noise in the data, and thus has to compute soft margins.
- The SVMs will also not provide a probability estimate of classification, i.e., the confidence level for classifying an instance.

Applications of SVM

1. Classification
2. Regression analysis
3. Pattern recognition
4. Outliers detection.
5. Relevance based applications

Acknowledgements

- Introduction to Machine Learning, Alphaydin
- Statistical Pattern Recognition: A Review – A.K Jain et al., PAMI (22) 2000
- Pattern Recognition and Analysis Course – A.K. Jain, MSU
- *Pattern Classification"* by Duda et al., John Wiley & Sons.

Material in these slides has been taken from, the following resources