CSE 4553
Machine Learning
Lecture 6: Support Vector Machine

Winter 2022

Hasan Mahmud | hasan@iut-dhaka.edu

Contents

- Introduction
- What is a Support Vector Machine?
- How does it work?
- Derivation of SVM Equations
- Pros and Cons of SVMs

Introduction

- Support Vector Machines (SVMs) are widely applied in the field of pattern classifications and nonlinear regressions.
- The original form of the SVM algorithm was introduced by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963.
- Since then, SVMs have been transformed tremendously to be used successfully in many real-world problems such as text (and hypertext) categorization, image classification, bioinformatics (Protein classification, Cancer classification), handwritten character recognition, etc.

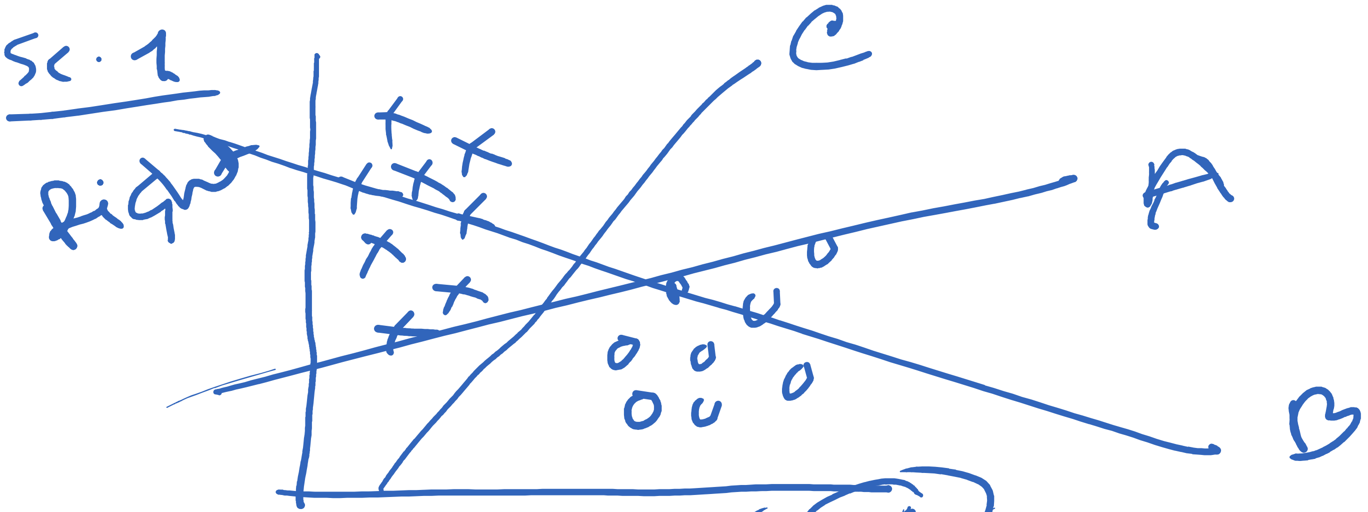
What is SVM?

- A Support Vector Machine is a supervised machine learning algorithm which can be used for both classification and regression problems.
- It follows a technique called the **kernel trick** to transform the data and based on these transformations, it finds an optimal boundary between the possible outputs.
- In simple words, it does some extremely complex data transformations to figure out how to separate the data based on the labels or outputs defined.

How does it work?

- The main idea is to identify the optimal separating hyperplane which maximizes the margin of the training data. Let us understand this objective term by term.

Sc. 1

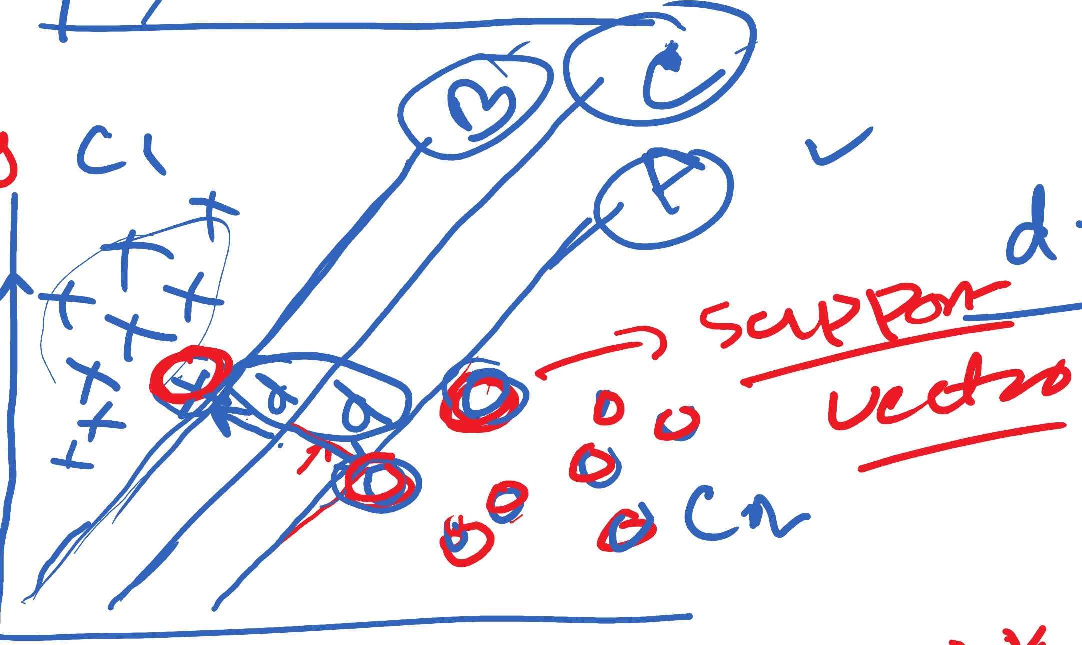


Right hyperplane =?

C →

Sc-2

Max Margin



$d + d = 2d$

$x = [1]$

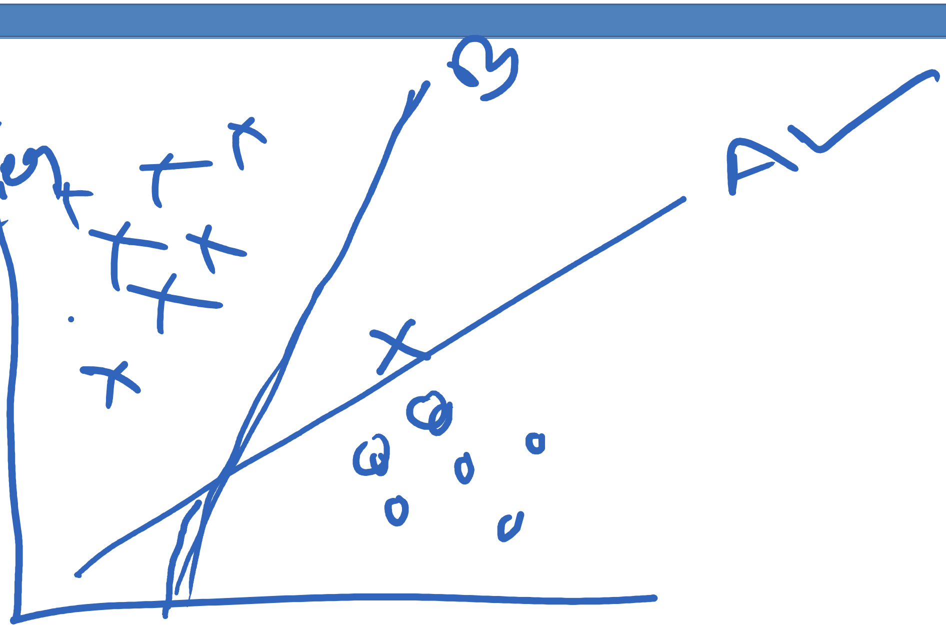
$y = [2]$

→ x

→ $S_1 = [1, 2]$

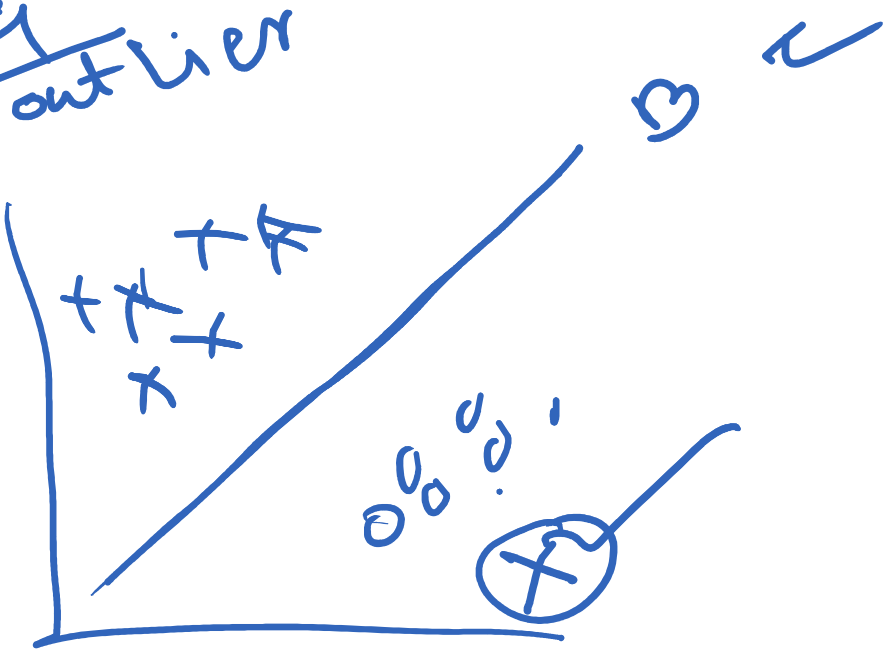
→ $S_2 = [2, 3]$

Sc3
accuracy



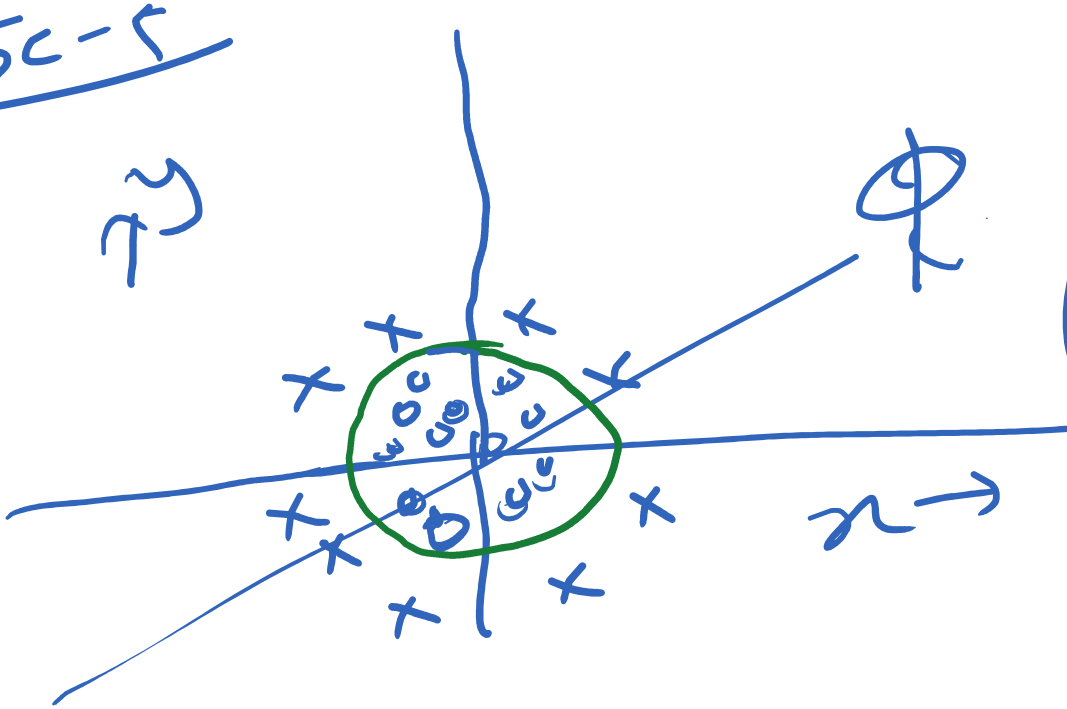
Accuracy ✓
vs
Max Margin

Sc-4
outlier

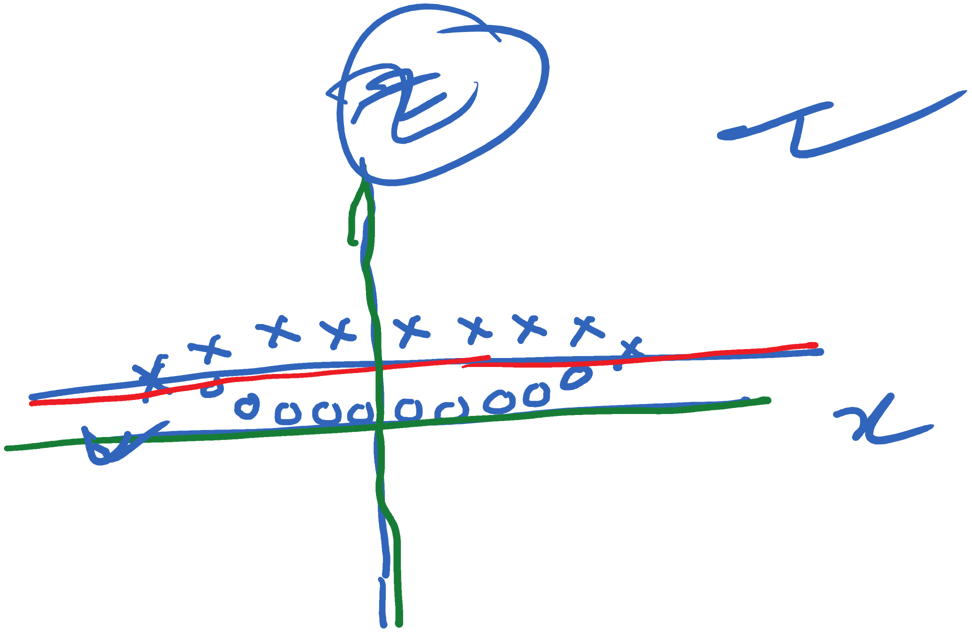


SC-5

Kernel Trick

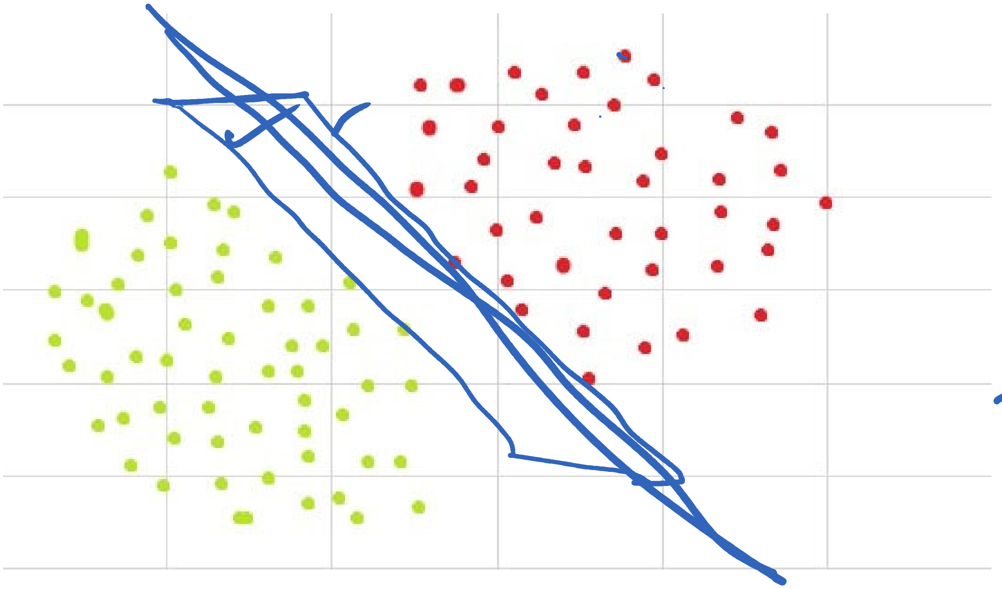


$$z = x^2 + y^2$$



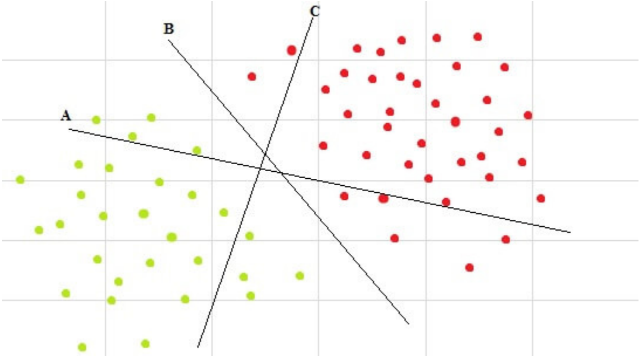
What is separating hyperplane?

- It is an n-1 dimensional subspace of an n-dimensional Euclidean space. So for a
  - 1D dataset, a **single point** represents the hyperplane.
  - 2D dataset, a **line** is a hyperplane.
  - 3D dataset, a **plane** is a hyperplane.
  - And in the higher dimension, it is called a **hyperplane**.

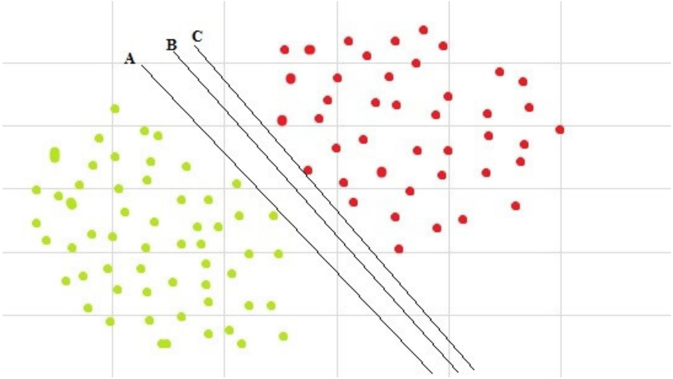


Optimal Hyperplane

- The objective of an SVM is to find the optimal separating hyperplane.
- When is a separating hyperplane said to be optimal?
- The fact that there exists a hyperplane separating the dataset doesn't mean that it is the best one.



Multiple hyperplanes



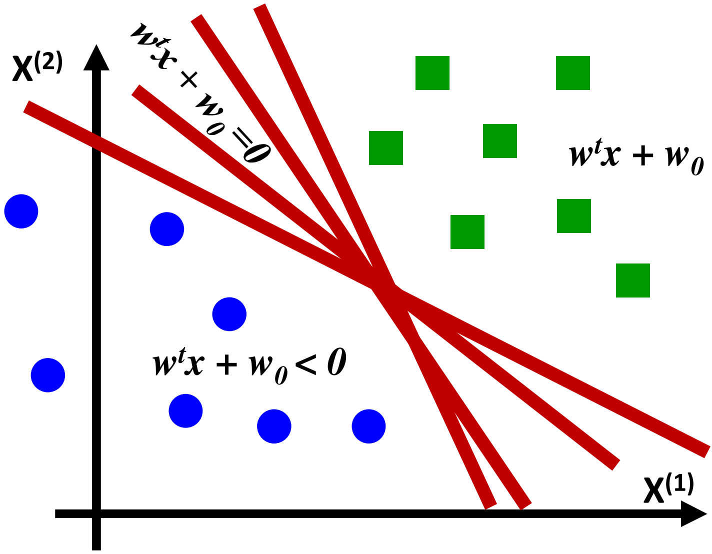
Multiple separating hyperplanes

SVM: Linearly separable case

- Let us start with a two-class problem where the classes are linearly separable with a discriminating function:

$$g(x) = w^t x + w_0$$

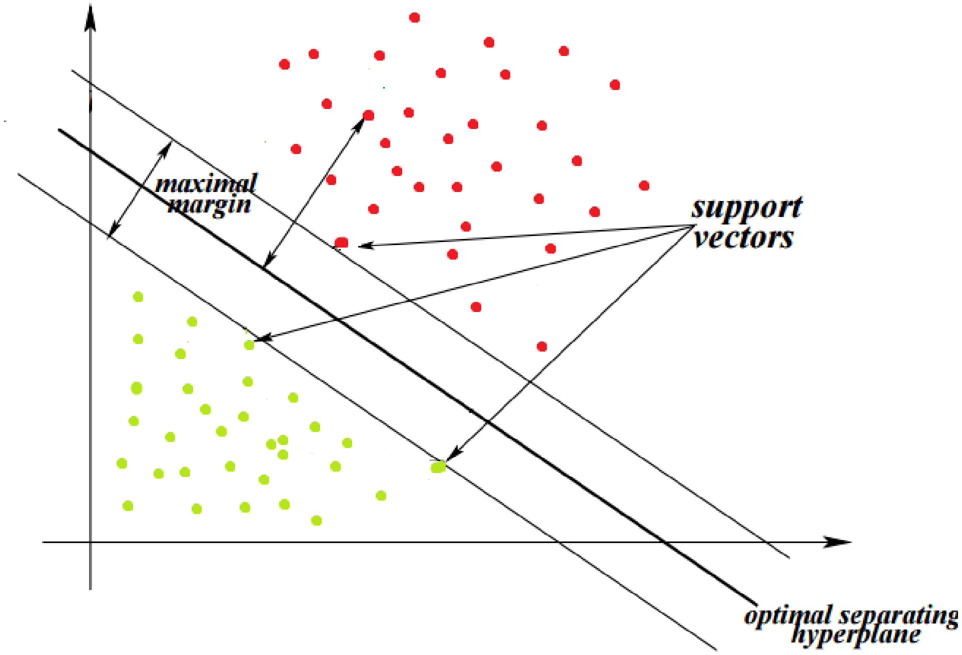
$$\boxed{\begin{array}{l} g(x) > 0 \Rightarrow x \in class\,1 \\ g(x) < 0 \Rightarrow x \in class\,2 \end{array}}$$



- There are infinite number of separating lines that could be drawn.
- Which separating hyperplane should we choose?
- We want to find the best one which will have the minimum classification error on previously unseen samples.

Margin

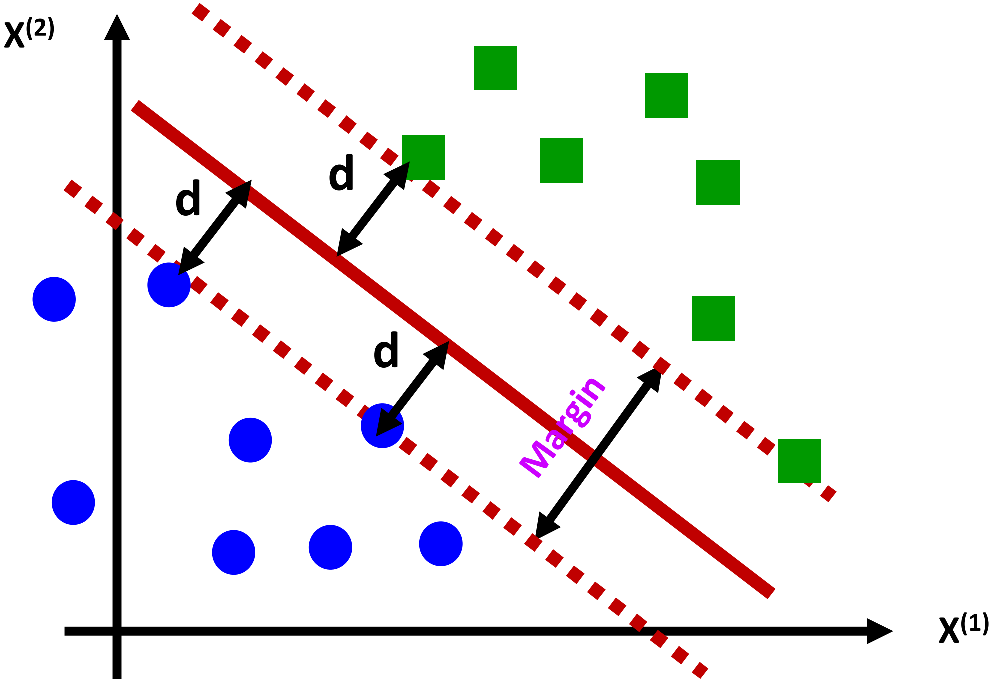
- Therefore, maximizing the distance between the nearest points of each class and the hyperplane would result in an optimal separating hyperplane. This distance is called the margin.



- The goal of SVMs is to find the optimal hyperplane because it not only classifies the existing dataset but also helps predict the class of the unseen data. And the optimal hyperplane is the one which has the biggest margin.

SVM: Linearly separable case…

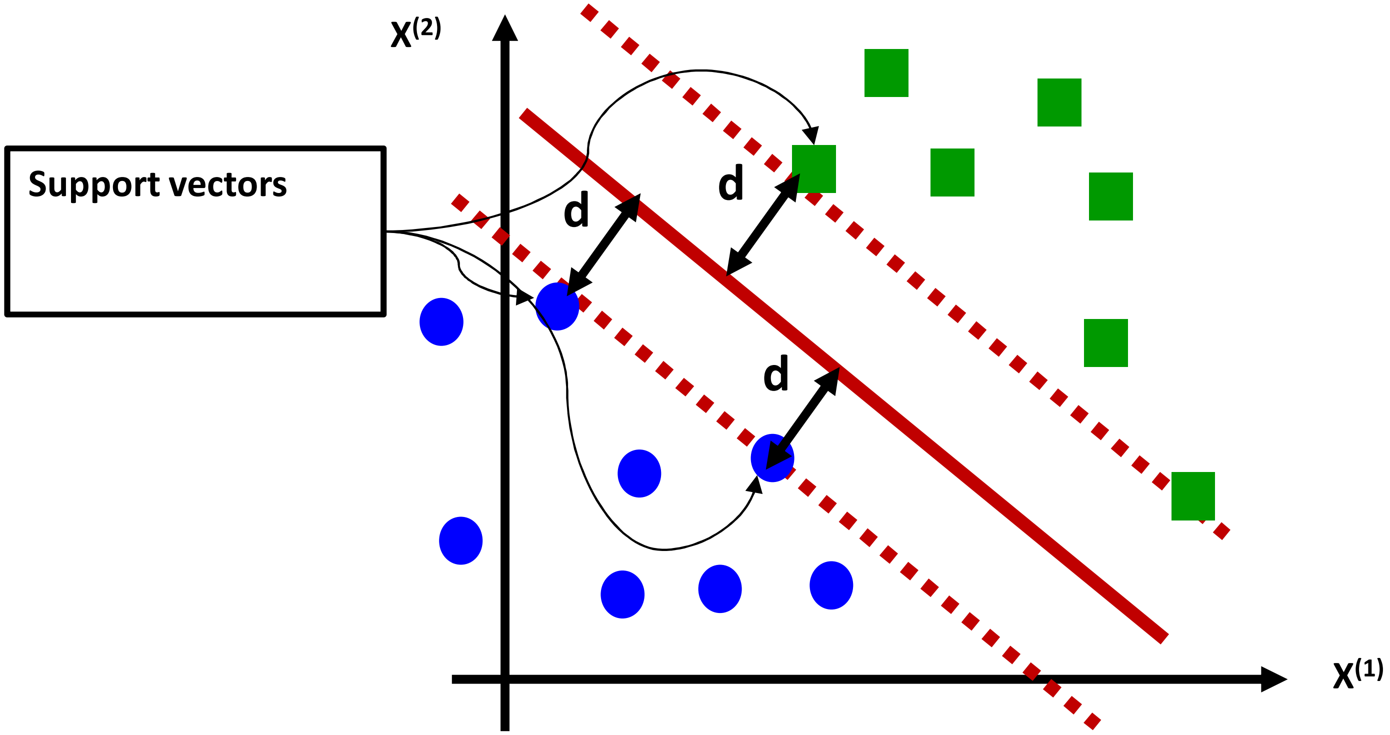
- Margin: SVM maximizes the margin
- Margin of a linear classifier is defined as the width that the boundary could be increased by before hitting a data point.



- Margin is twice the absolute value of distance d of the closest example to the separating hyperplane

SVM: Linearly separable case…

- Support vectors are the samples closest to the separating hyperplane.
- Support vectors are those datapoints that the margin pushes up against.
- They are the most difficult pattern to classify hence are important than other training examples.



- Optical hyperplane is completely defined by support vectors
  - Of course we do not know which samples are support vectors without finding the optimal hyperplane.

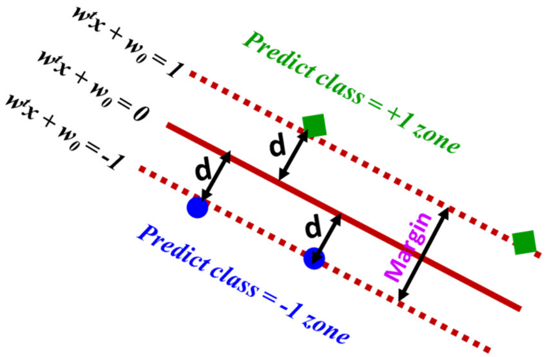
Equation of Hyperplane

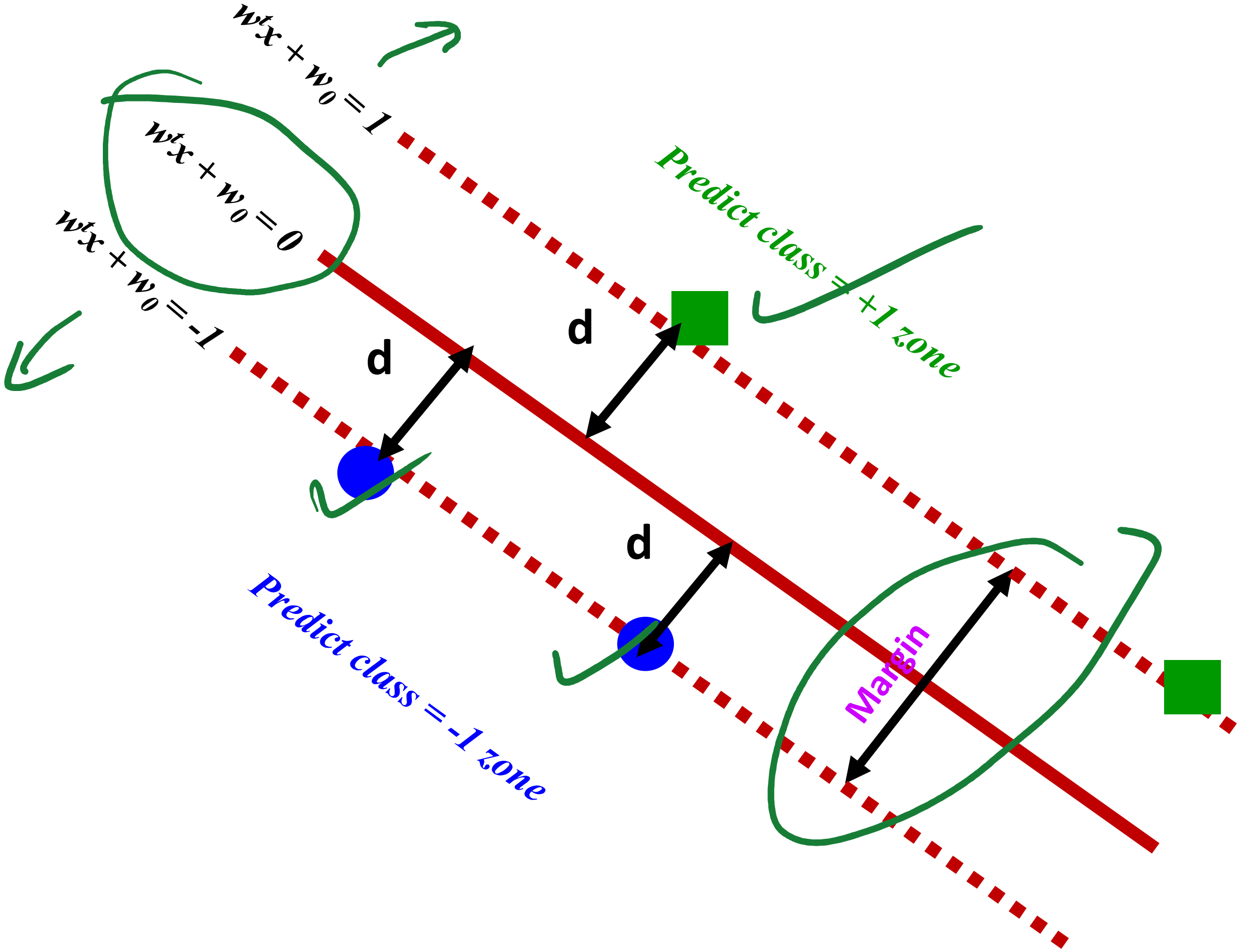
- You must have come across the equation of a straight line as $y = mx + c$, where m is the slope and c is the y-intercept of the line.
- The generalized equation of a hyperplane is as follows:
$$w^T x = 0$$
- Here w and x are the vectors and $w^T x$ represents the dot product of the two vectors. The vector w is often called as the weight vector.

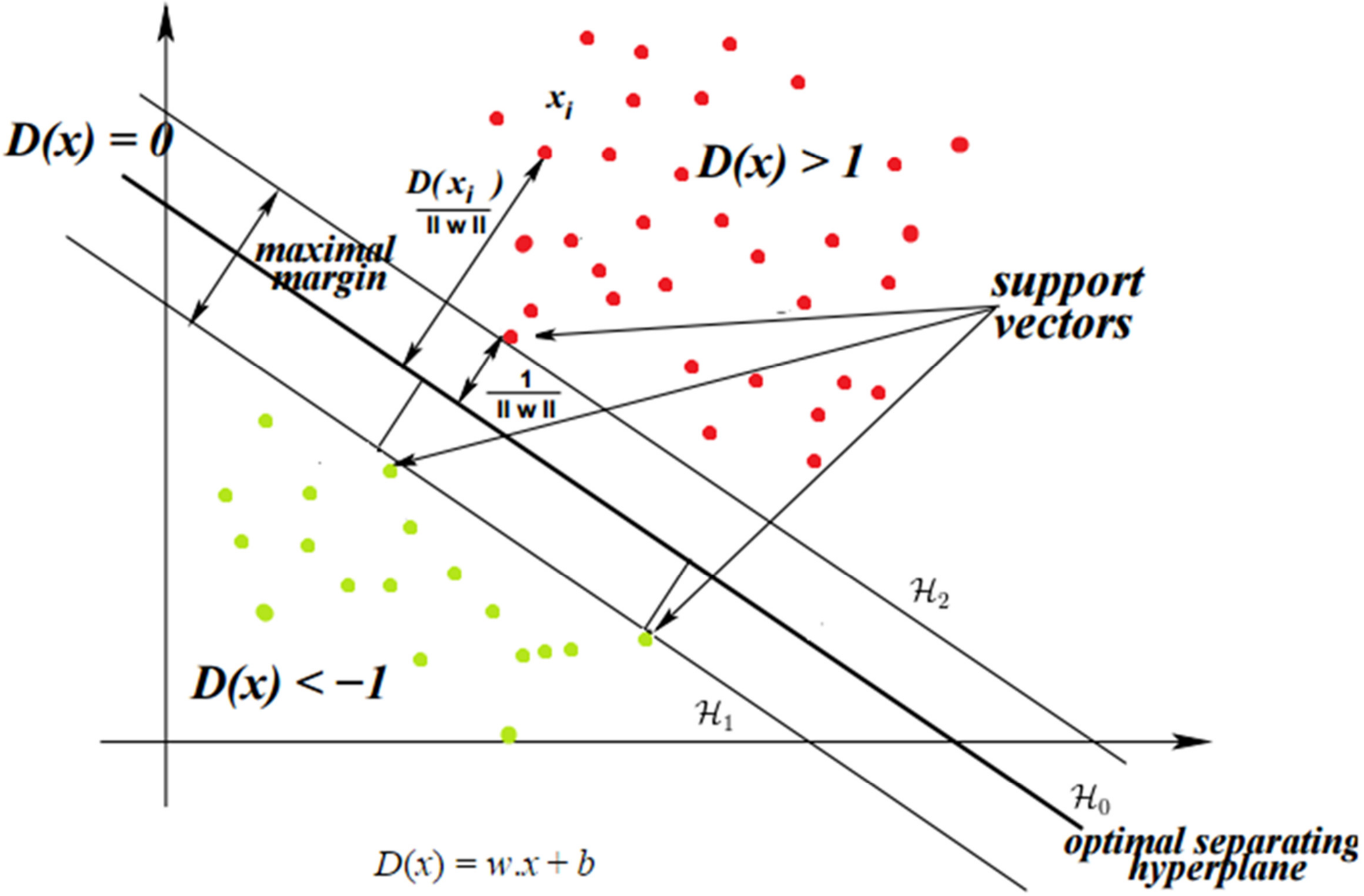
- Consider the equation of the line as $y - mx - c = 0$. In this case,
- $w = \begin{bmatrix} -c \\ -m \\ 1 \end{bmatrix}$ $and\ x = \begin{bmatrix} 1 \\ x \\ y \end{bmatrix}$
- $w^T x = -c \times 1 - m \times x + y = y - mx - c = 0$
- w represents the vector which is normal to the hyperplane. This property will be useful once we start computing the distance from a point to the hyperplane.

Equation of Margin

- Formula for the margin:
  - We know the equation of the separating hyperplane, $w^t x + w_0 = 0$
  - Any point that lies above the separating hyperplane is, $w^t x + w_0 > 0$
  - Any point that lies below the separating hyperplane is, $w^t x + w_0 < 0$
  - For uniqueness, lets set $|w^t x + w_0| = 1$ for any example x_i closest to the boundary
  - Now, distance from the closest sample x_i to g(x) = 0, is, $\frac{|w^t x + w_0|}{\|w\|} = \frac{1}{\|w\|}$
  - Thus the margin is, $m = \frac{2}{\|w\|}$




$w^t x + w_0 = 1$
$w^t x + w_0 = 0$
$w^t x + w_0 = -1$
Predict class = +1 zone
d
d
d
Margin
Predict class = -1 zone



$D(x) = 0$
x_i
$\frac{D(x_i)}{\|w\|}$
$D(x) > 1$
maximal margin
support vectors
$\frac{1}{\|w\|}$
$\mathcal{H}_2$
$D(x) < -1$
$\mathcal{H}_1$
$\mathcal{H}_0$
optimal separating hyperplane
$D(x) = w.x + b$

SVM: Linearly separable case…

- To have an optimal hyperplane the goal of SVM is:
  - Maximize the margin, $m = \frac{2}{\|w\|}$
    - Subject to constraints, $\begin{cases} w^t x_i + b \geq 1 \;\; if \;\; x_i \; is \; positive \, example \\ w^t x_i + b \leq -1 \;\; if \;\; x_i \; is \, negative \, example \end{cases}$
      - Let, $\begin{cases} y_i = 1 \;\; if \;\; x_i \; is \; positive \, example \\ y_i = -1 \;\; if \;\; x_i \; is \, negative \, example \end{cases}$
    - So , combining the two inequalities we can write, $y_i(w^t x_i + b) \geq 1, \forall i$
  - We can convert the problem to,

$$\boxed{\begin{array}{l} \text{Minimize, } \phi(w) = \frac{1}{2}\|w\|^2 = \frac{1}{2} w^t w \\ \text{Constraint to, } y_i(w^t x_i + b) \geq 1, \forall i \end{array}}$$

  - $\phi(w)$ is a quadratic function, thus there is a single global minimum

Formulation of SVM

- So, our objective is to solve the quadratic optimization problem and solve for w and b,

$$\boxed{\begin{array}{l}\text{Minimize, } \phi(w) = \frac{1}{2}\|w\|^2 = \frac{1}{2}w^t w \\ \text{Constraint to, } y_i(w^t x_i + b) \geq 1, \forall i\end{array}}$$

Formulation of SVM….

- From the SVM theory we know the equations are,

$$\phi(w) = \frac{1}{2}\|w\|^2 = \frac{1}{2}w^t w$$

$$g_i(w,b) = \; y_i(w^t x_i + b) \geq 1, \forall i$$

$$g_i(w,b) = 1 - y_i(w^t x_i + b) \leq 0, \forall i$$

- We can write the constraint function like,
- Now we can put the equation in the form of Lagrangian,

$$L(w,b,\alpha) = \phi(w) + \sum_{i=1}^{N} \alpha_i g_i(w,b) \qquad ....(1)$$

Formulation of SVM…

$$
\begin{aligned}
L(w,b,\alpha) &= \phi(w) + \sum_{i=1}^{N} \alpha_i g_i(w,b) \\
&= \frac{1}{2} w^t w + \sum_{i=1}^{N} \alpha_i (1 - y_i (w^t x_i + b)) \\
&= \frac{1}{2} w^t w + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i y_i w^t x_i - b \sum_{i=1}^{N} \alpha_i y_i \qquad \text{...... (2)}
\end{aligned}
$$

- Now we solve for the gradient (partial derivatives w.r.t. w, b) of the Lagrangian equation (2) to satisfy the KKT (Karush-Kuhn-Tucker) conditions so that we can get an optimal solution.

Formulation of SVM…

- From equation (2),

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{N} \alpha_i y_i x_i = 0 \qquad \Rightarrow w = \sum_{i=1}^{N} \alpha_i y_i x_i \quad ......(3)$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{N} \alpha_i y_i = 0 \;\; ..... \; (4)$$

- In equation (3), we have,

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i$$

$$\therefore \quad w^t w = w^t \sum_{i=1}^{N} \alpha_i y_i x_i \Rightarrow w^t w = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i x_j \;\; ......(5)$$

Formulation of SVM…

- Now combining the equations (3), (4), and (5) with equation (2) we get,

$$
\begin{aligned}
L(w,b,\alpha) &= \frac{1}{2} w^t w + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i y_i w^t x_i - b \sum_{i=1}^{N} \alpha_i y_i \\
&= \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i x_j + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i x_j - b \sum_{i=1}^{N} \alpha_i y_i \\
&= \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i x_j \qquad \text{[using equation(4)]}
\end{aligned}
$$

Formulation of SVM…

- So we get the final problem,

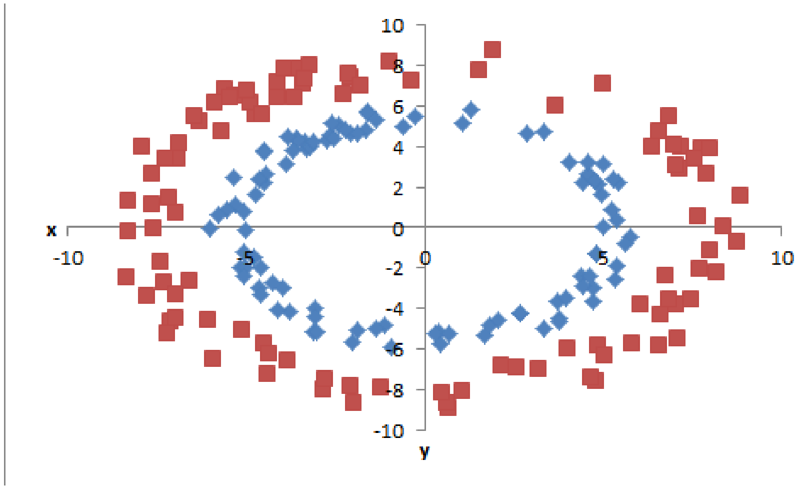
Maximize $$L(w,b,\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i x_j$$

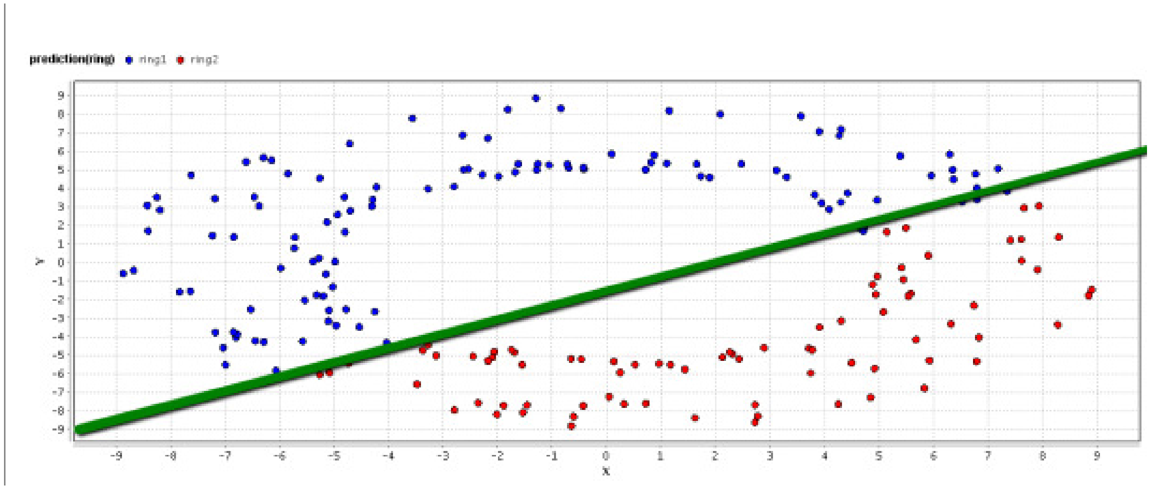
Subject to $$\alpha_i \geq 0$$

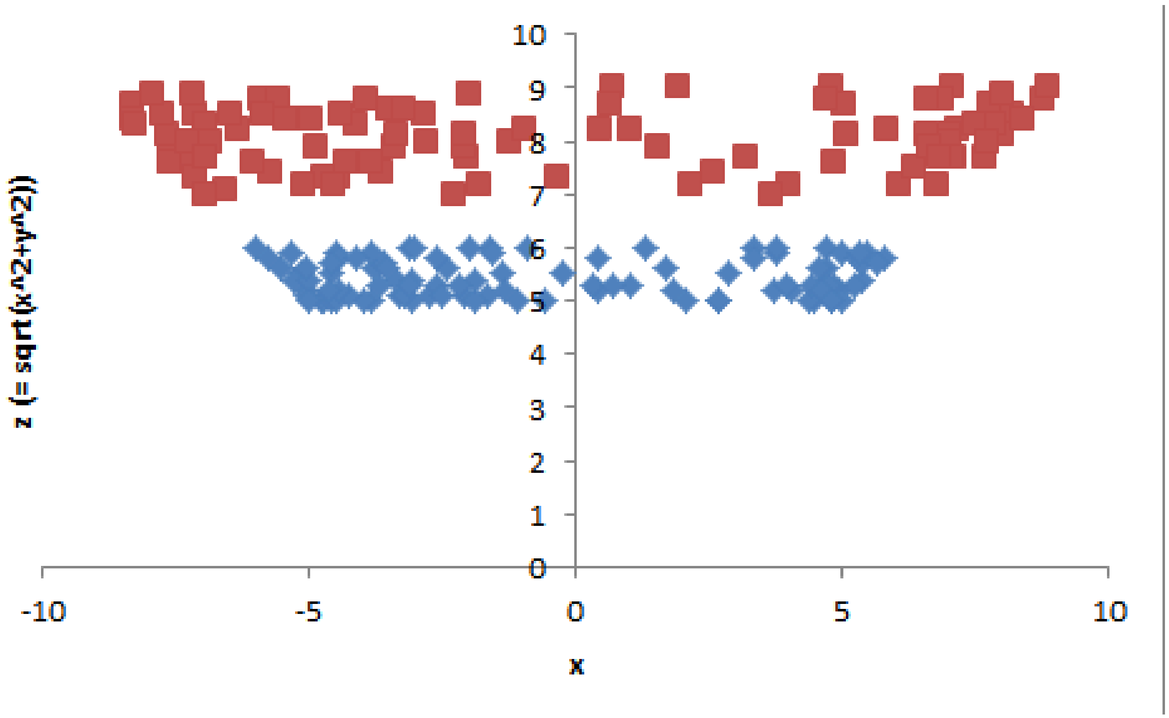
$$\sum_{i=1}^{N}\alpha_i y_i = 0$$

What is Kernel function?

- Kernel function or basis function transforms the data from non-linear feature space to linear feature space.






10
9
8
7
6
5
4
3
2
1
0
-10
-5
0
5
10
x
z (= sqrt(x^2+y^2))