Constrained Optimization in Machine Learning

Dr. Pritam Anand.

Assistant Professor,

DA-IICT, Gandhinagar.

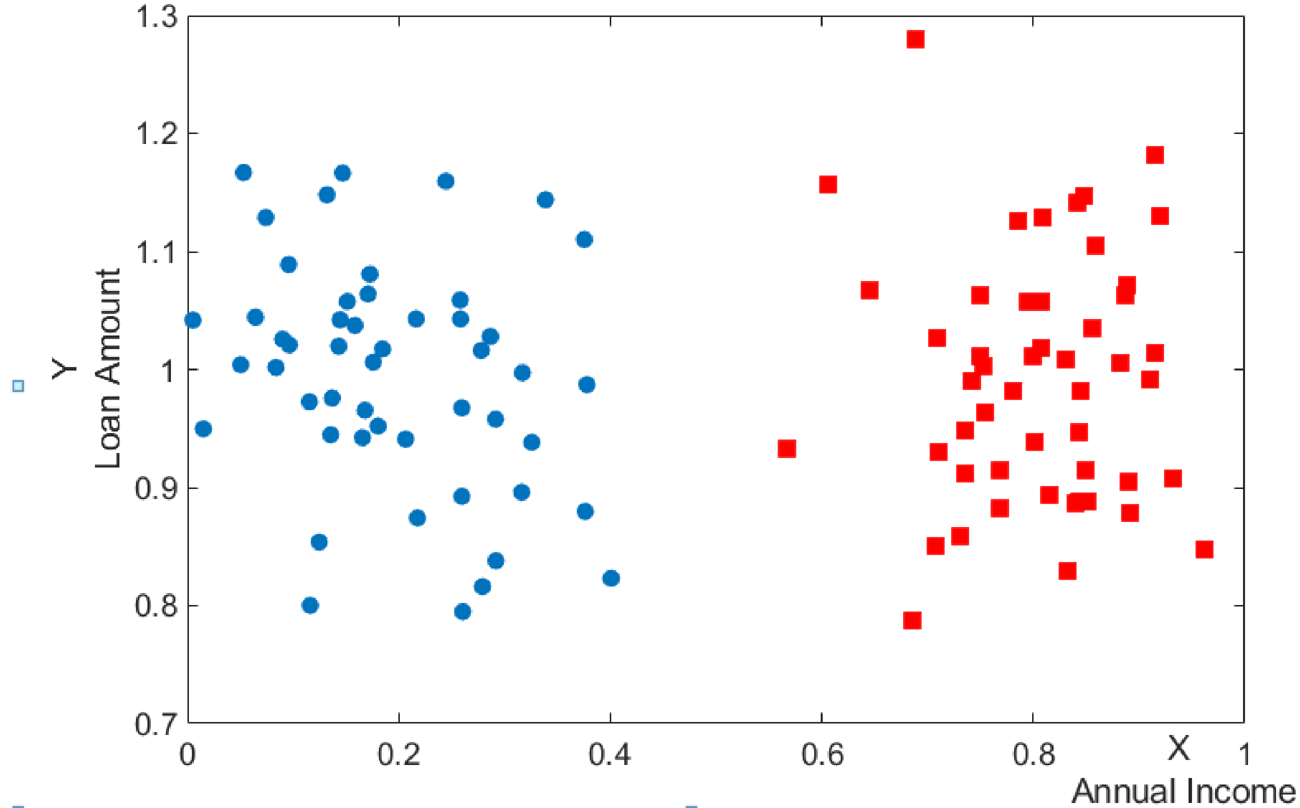
Takeaways

- *Develop the understanding of the Convex Programming Problems.*
- *Aware with the constrained optimization methods used in the solution of Support Vector Machine.*
- *Develop the detailed understanding of the Support Vector Machine.*
- *Understand the dual and primal relationship in context of Support Vector Machine problem.*
- *Ability to use the Support Vector Machine in different domain of applications.*

Load Defaulter Dataset

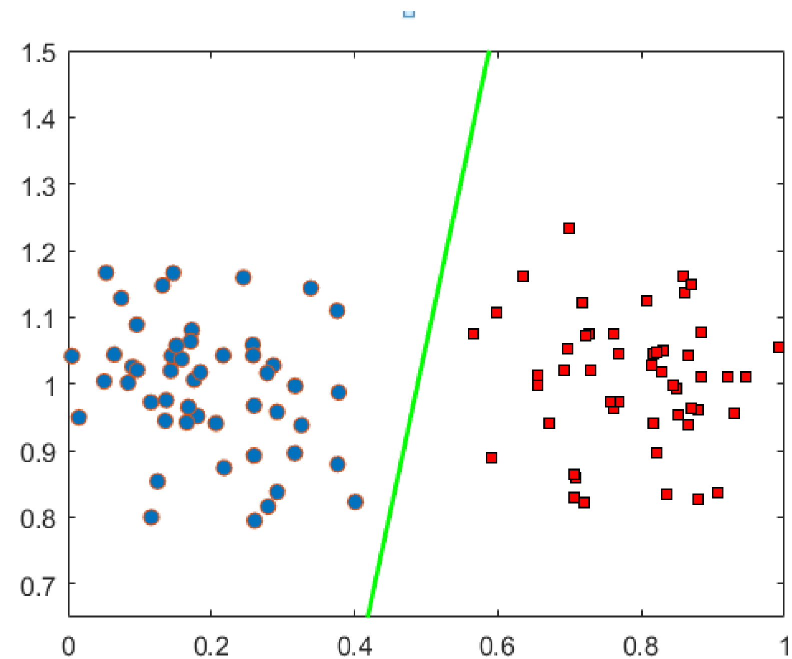
Index	Employed	Bank Balance (in thousands rupees)	Annual Salary (in million rupees)	Loan Amount (in thousands rupees)	Defaulted
1	1	0.4721	0.1358	0.9448	-1
2	0	0.8412	0.3169	0.9972	-1
3	0	0.3687	0.1249	0.8539	-1
4	0	0.2547	0.8416	1.1406	1
5	1	0.3111	0.7502	1.014	1

A random classifier



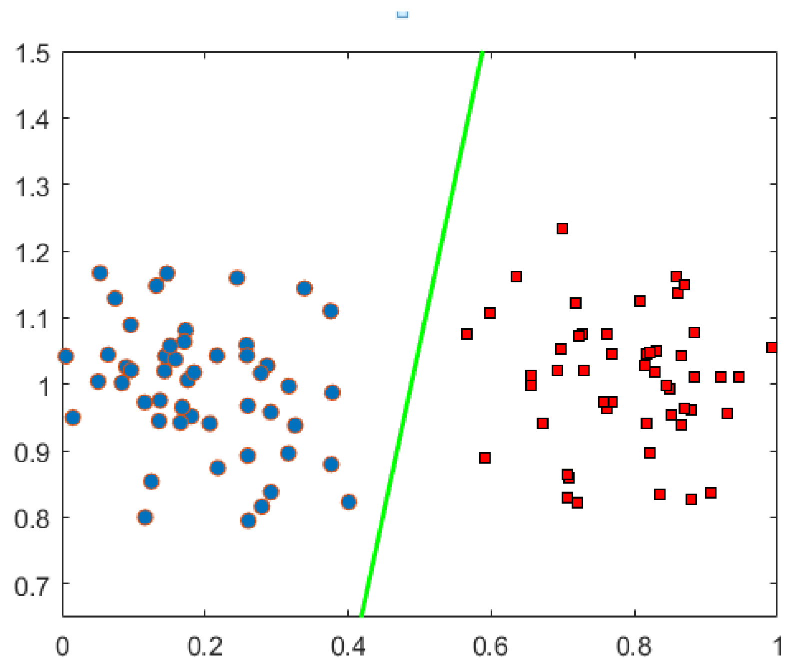
A random Classifier

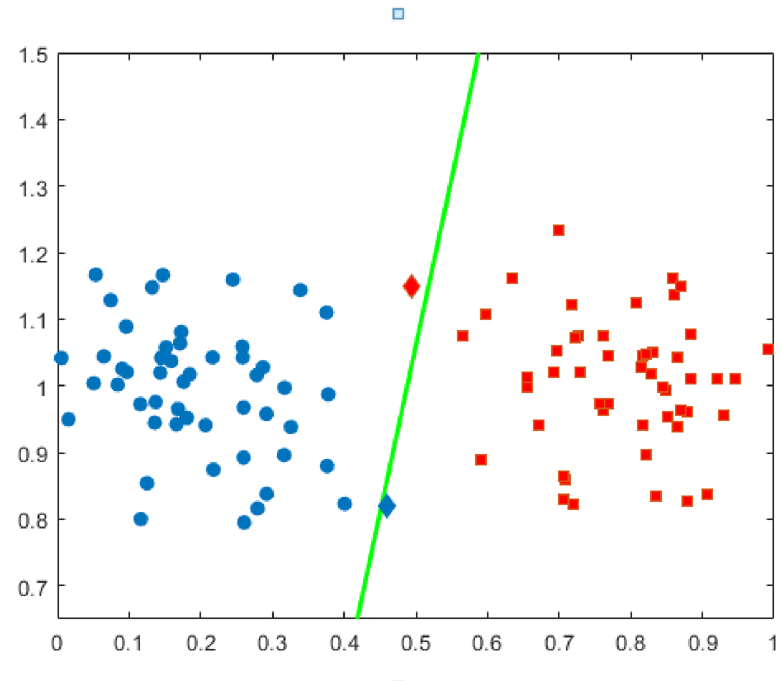
$$\beta_1 x_1 + \beta_2 x_2 + \beta_0 = 0$$



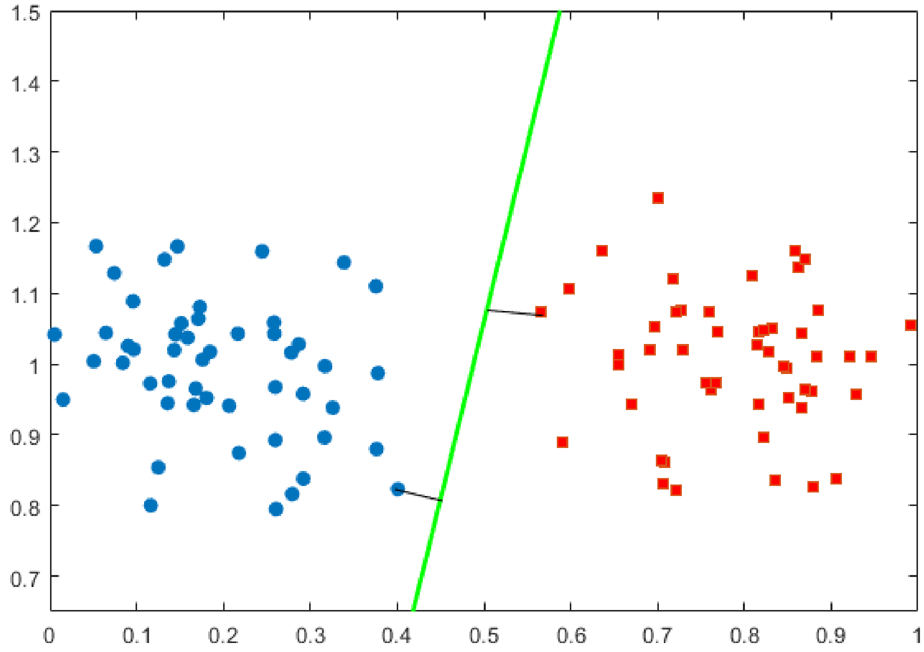
What is Problem ??

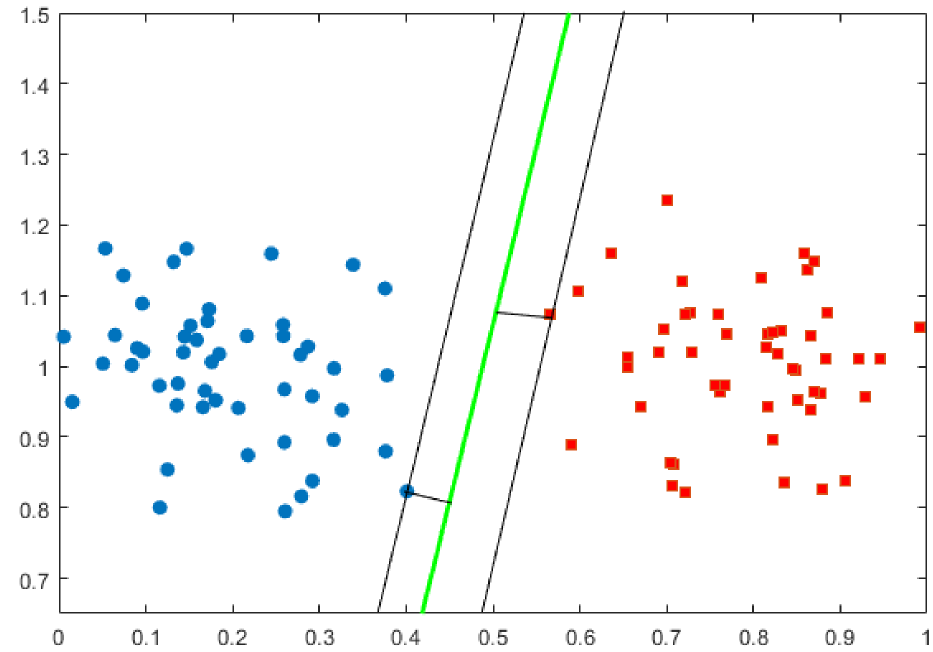
More Susceptible to misclassification

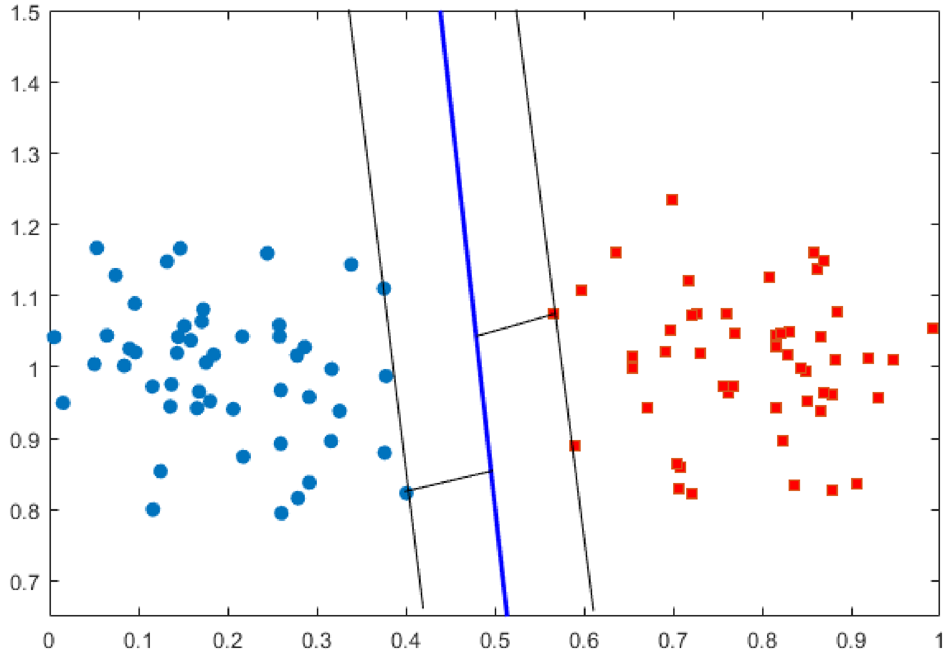




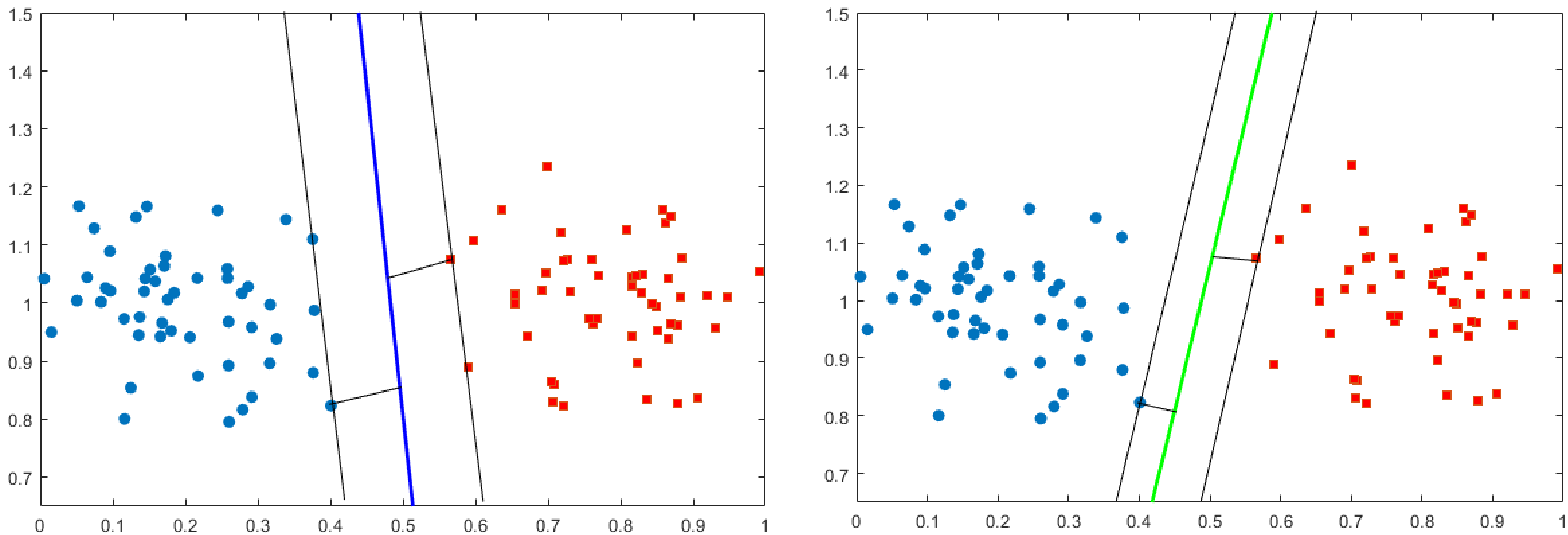
Margin of random classifier



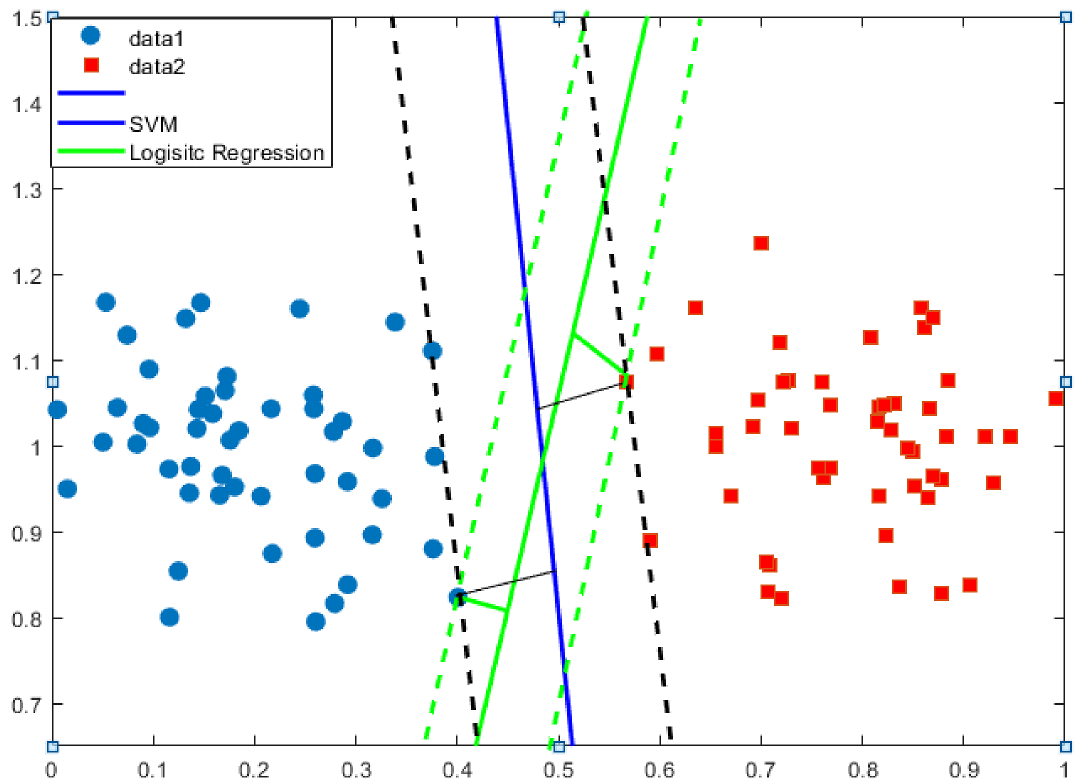



1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
0.7
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

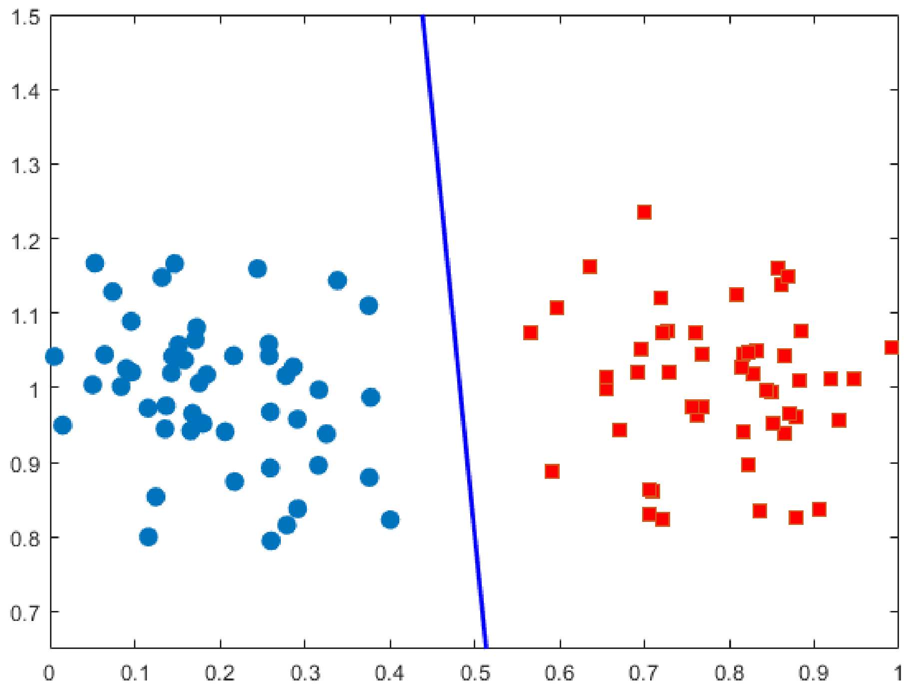
Maximal Margin Classifier

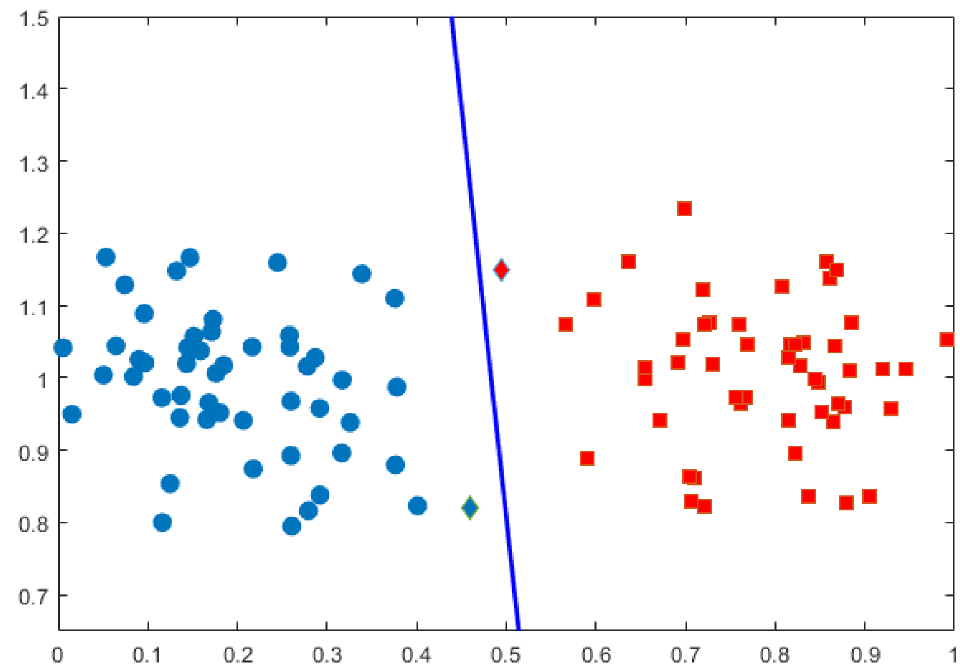


Width of margin

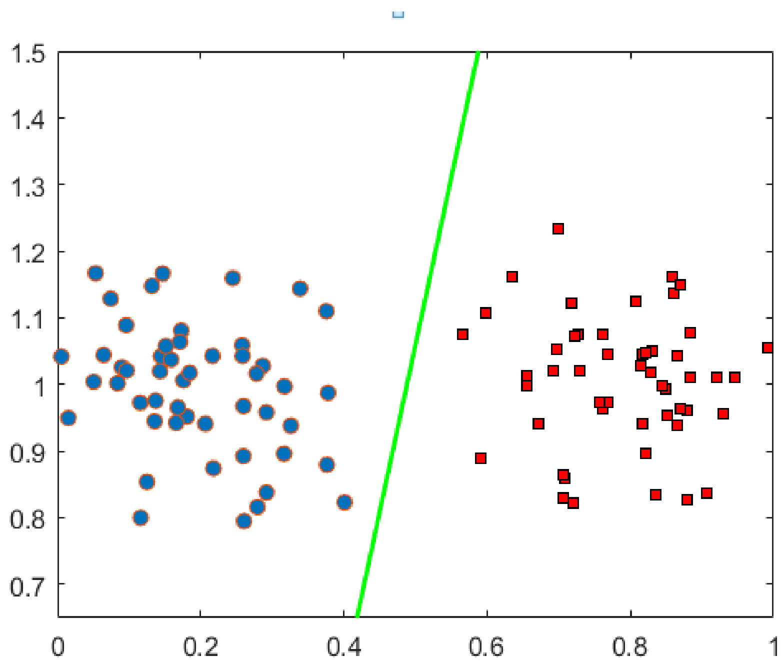


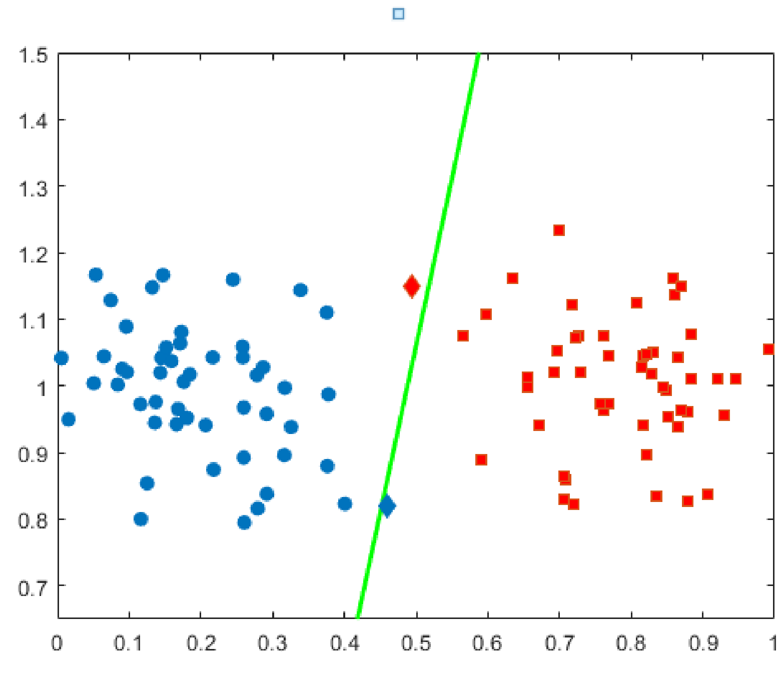
Checking the SVM



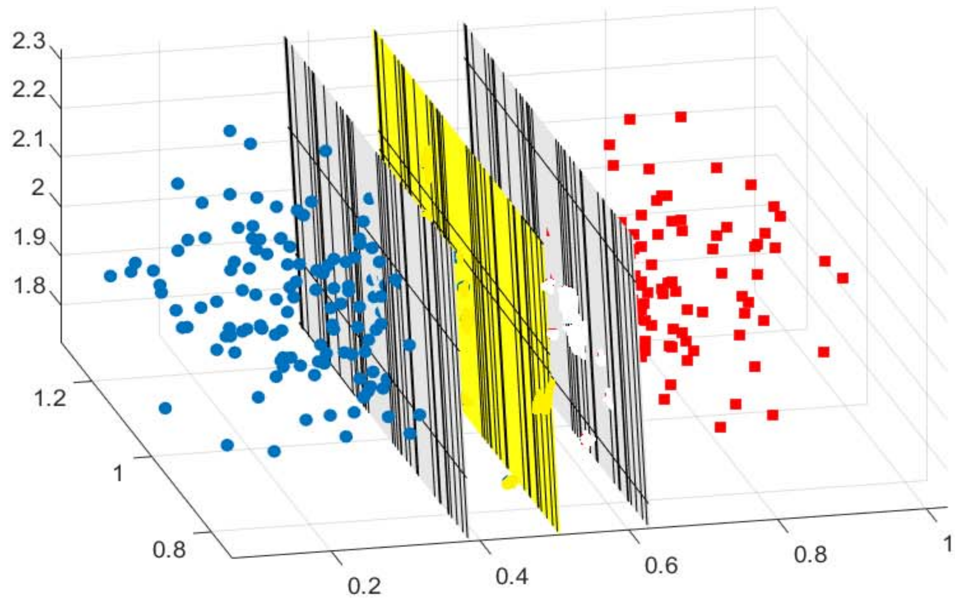


More Susceptible to misclassification





SVM margin



Hyperplane

$\hat{\omega}^T x = c$

$\hat{\omega}^T x - c = 0$

$\left(\frac{\omega}{||\omega||}\right)^T x - c = 0$

$\omega^T x \boxed{- c||\omega||} = 0$

$\omega^T x + b = 0$

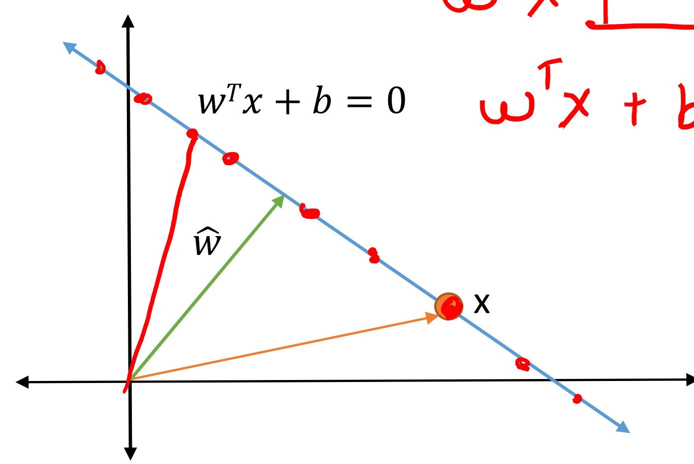
- Hyperplane:- A set of points in R^n satisfying $w^T x + b = 0, w \in Rn, b \in R$.
- For n = 2, it is a line R^2.
- For n = 3, it is a plane in R^3 .

$\hat{\omega}^T x = c$

$\hat{\omega}^T x = - b$

$\hat{\omega} = \frac{\omega}{||\omega||}$

$\hat{\omega}^T x$



$\widehat{w}^T$x= $\frac{-b}{||w||}$, where $\widehat{w} = \frac{w}{||w||}$

Hyperplane , Projection and Distances

$w^T x$

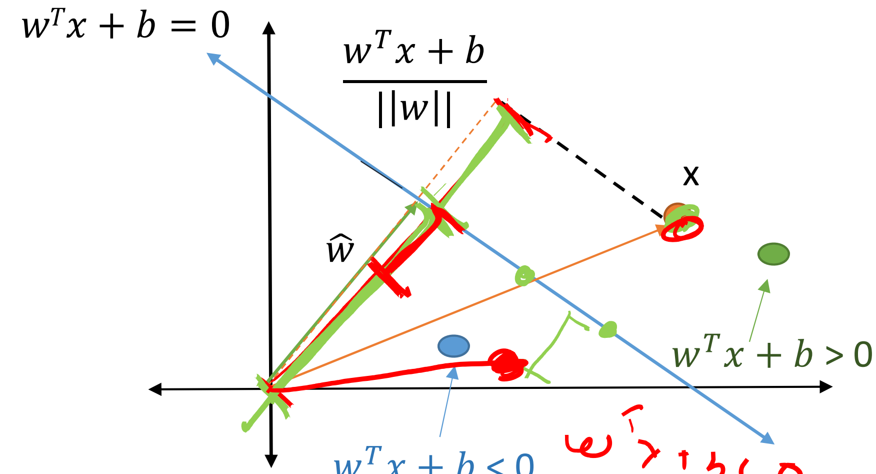
$\frac{w^T x + b}{||w||} > 0$

$w^T x + b > 0$

$w^T x + b = 0$

$\frac{w^T}{\cancel{||w||}} \cancel{||w||} x + b||w|| = 0$

$\Rightarrow w^T x + b||w|| = 0$



$w^T x + b < 0$

$w^T x + b$

$w^T x - C$

$\frac{w^T}{||w||} x - C = \frac{w^T x - C||w||}{||w||} = \frac{w^T x + b}{||w||}$

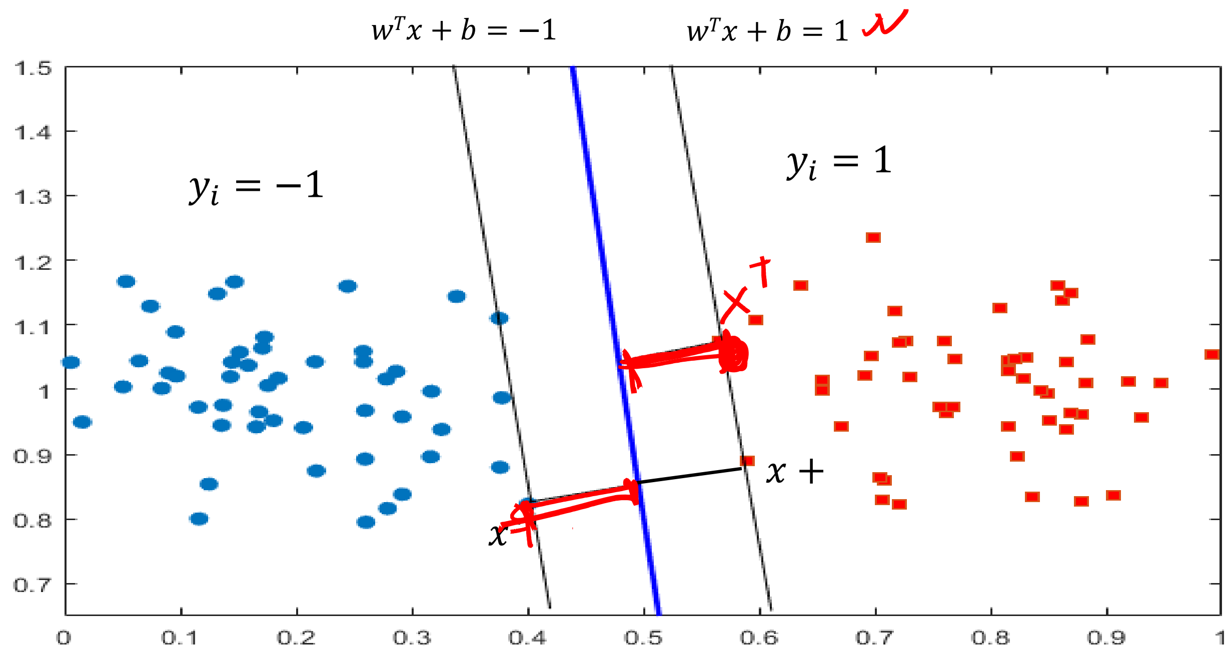
Margin in SVM

$$\frac{(w^T x^+ + b) - (w^T x^- + b)}{||w||} = \frac{1-(-1)}{||w||} = \frac{2}{||w||}$$

Width of the margin:-

$$\frac{(w^T x^{\ +} + b)\ - (w^T x^{\ -} + b)}{||w||}$$

$$= \quad \frac{2}{||w||}$$



$w^T x + b = \delta$

$w^T x + b = -\delta$

$\Rightarrow \left(\frac{w}{\delta}\right)^T x + \frac{b}{\delta} = -1$

Optimization in SVM

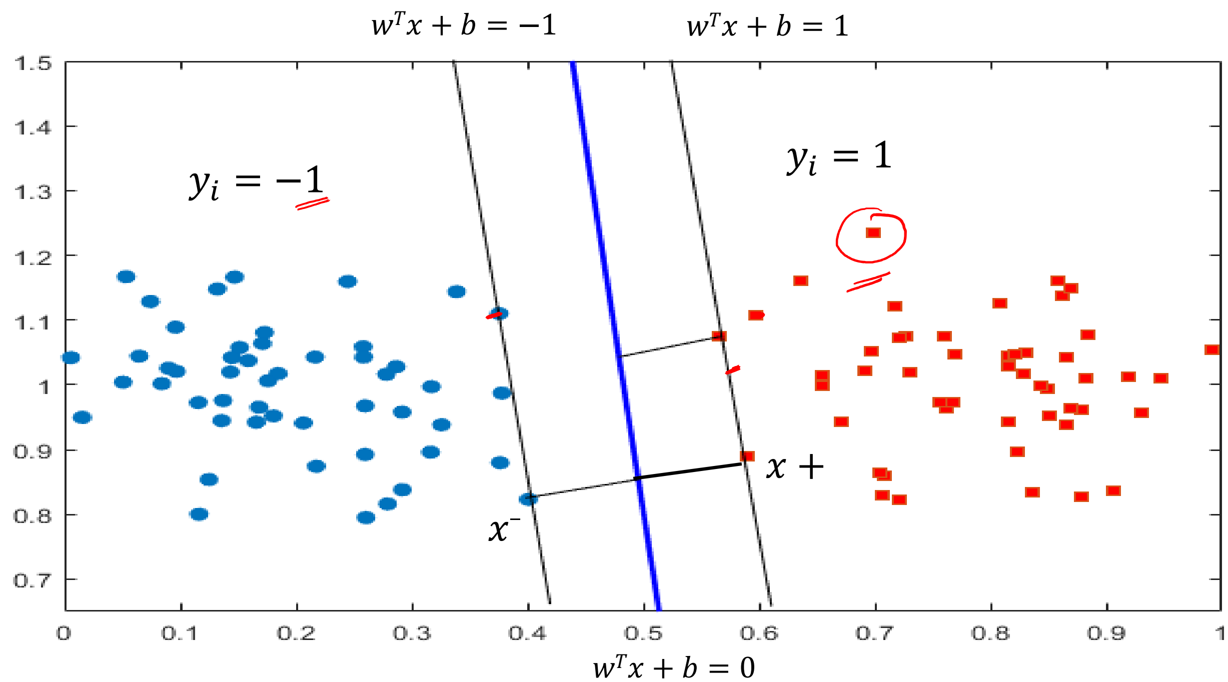
Need to maximize the width of the margin but, subject to certain constraints

$$\max_{(w,b)} \frac{2}{||w||}$$

subject to,

$(w^T x_i + b) \geq 1$, if y_i =1.

$(w^T x_i + b) \leq -1$,if y_i = -1.



Optimization in SVM

$$\max_{(w,b)} \frac{2}{||w||}$$

subject to,

$(w^T x_i + b) \geq 1$, if y_i =1.

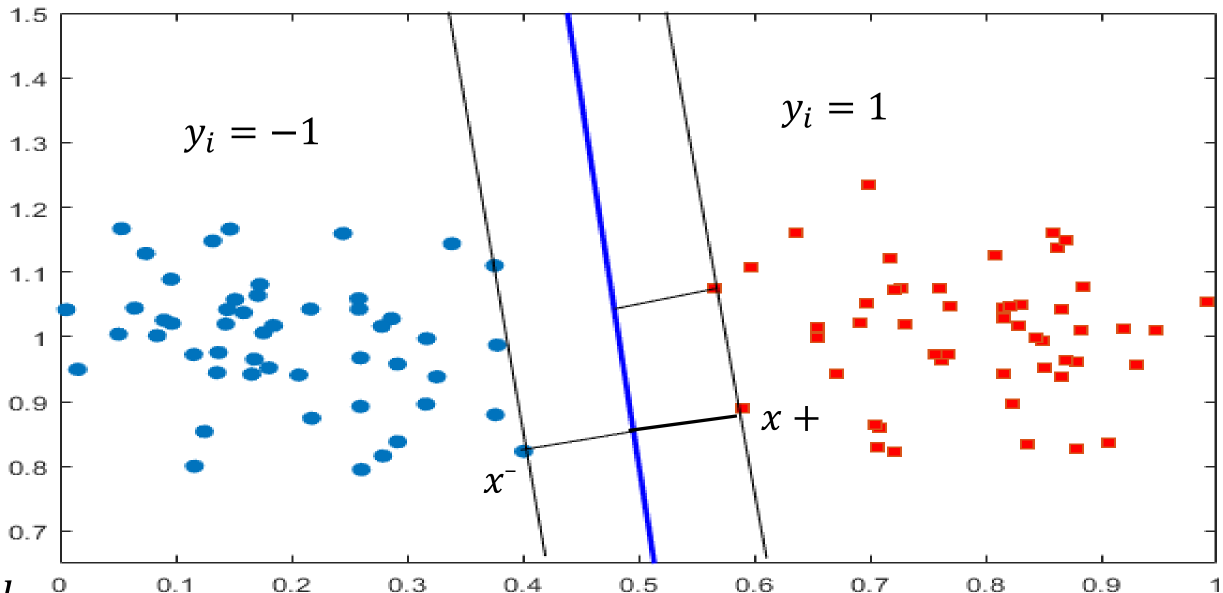
$(w^T x_i + b) \leq -1$,if y_i = -1.

is equivalent to

$$\max_{(w,b)} \frac{2}{||w||}$$

subject to,

$y_i \ (w^T x_i + b) \geq 1 \ , i = 1,2, \dots l$



$(w^T x_i + b) \geq 1$ if $y_i = 1$

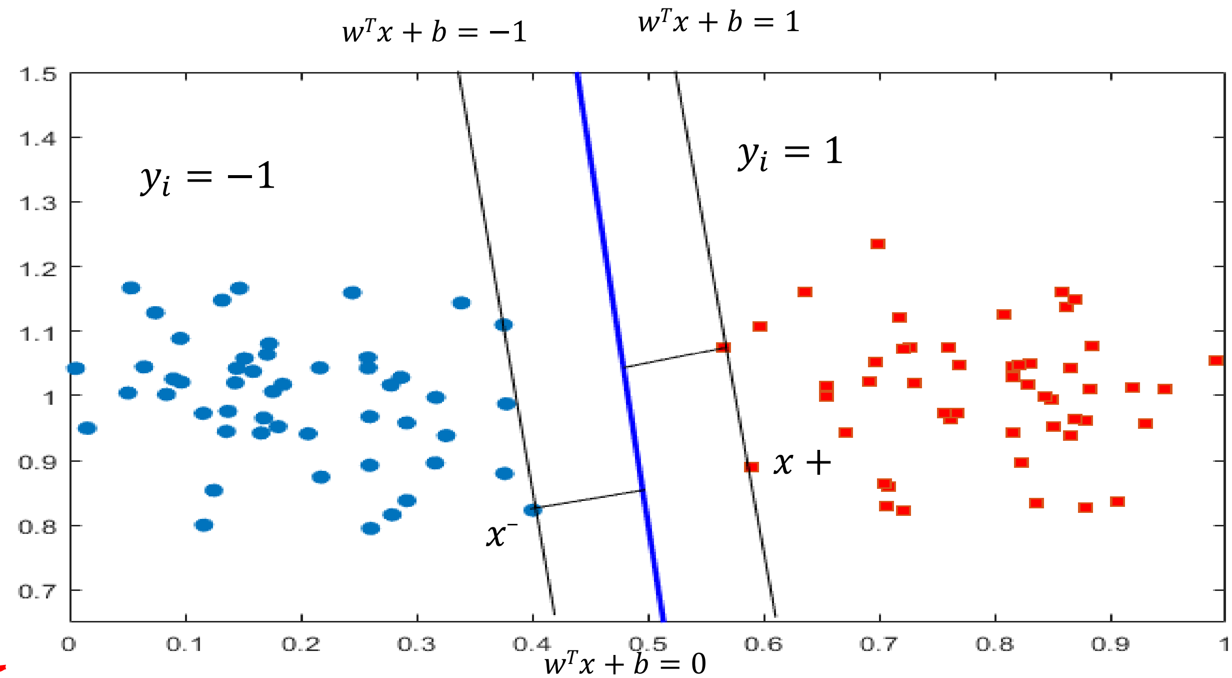
$-w^T x_i + b \leq -1$ $\leftarrow$ $-(w^T x_i + b) \geq -1$ if $y_i = -1$

Max $\frac{2}{||w||}$ Min $\frac{||w||^2}{2}$

Optimization in SVM

- $$\max_{(w,b)} \frac{2}{||w||}$$
  subject to,
  $y_i \ (w^T x_i + b) \geq 1 \ , i = 1,2, \ldots l.$

- $$\min_{(w,b)} \frac{||w||^2}{2}$$
  subject to,
  $y_i \ (w^T x_i + b) \geq 1 \ , i = 1,2, \ldots l.$



$\sqrt{w_1^2 + w_2^2 + \cdots + w_n^2}$

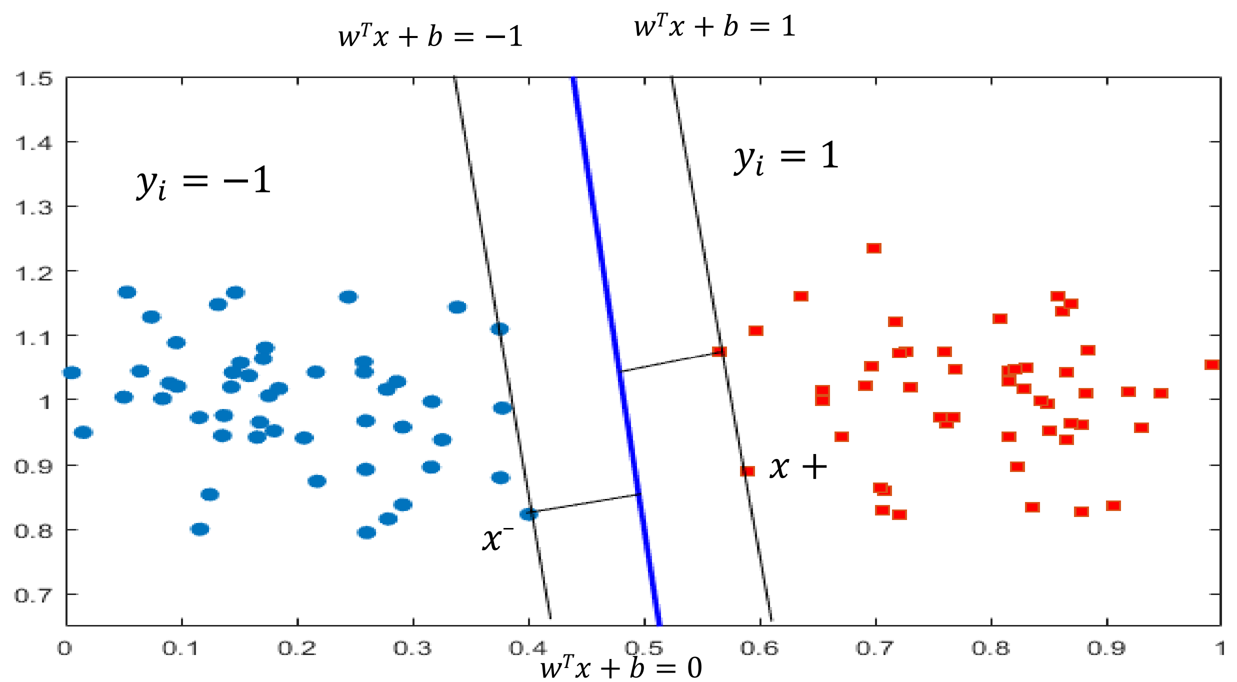
$W^T W$

Optimization in SVM

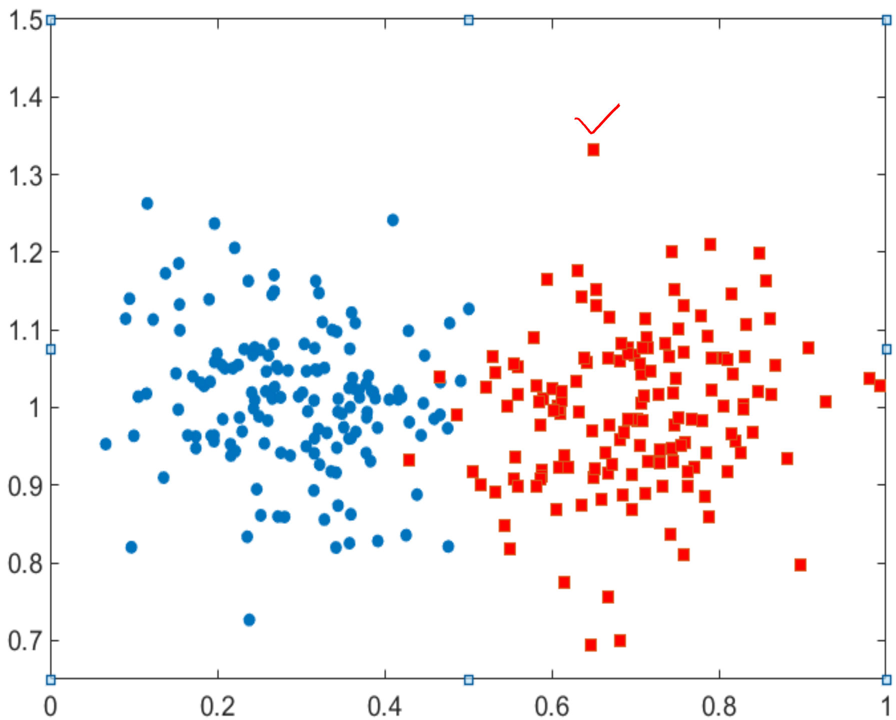
- $$\max_{(w,b)} \frac{2}{||w||}$$
  subject to,
  $$y_i \ (w^T x_i + b) \geq 1 \ , i = 1,2, \dots l.$$

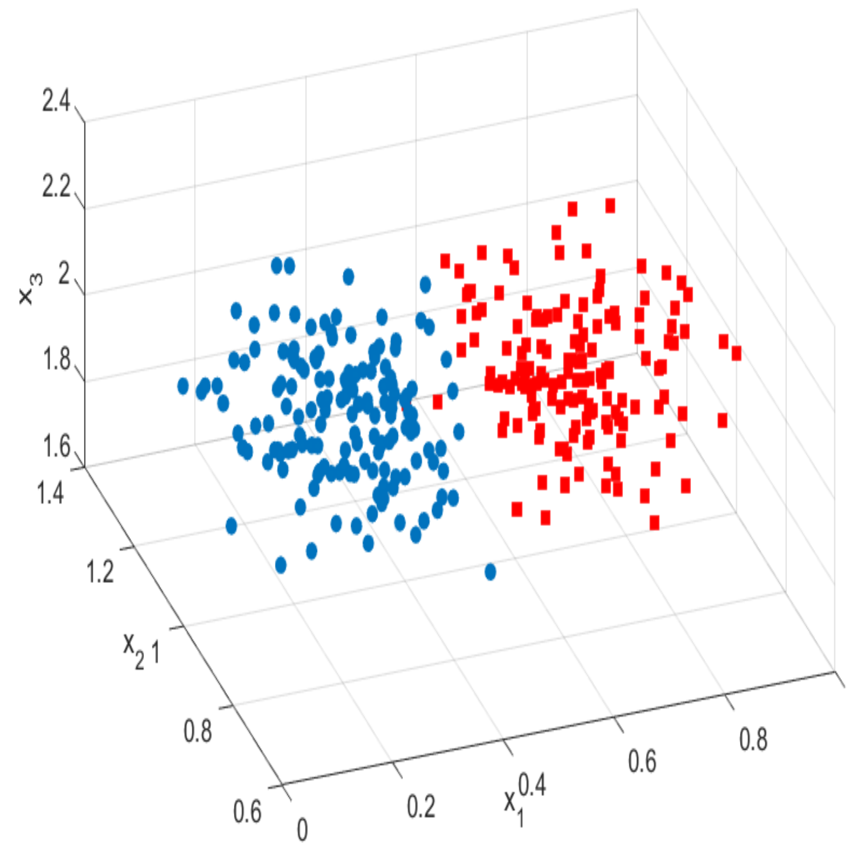
- $$\min_{(w,b)} \frac{||w||^2}{2}$$
  subject to,
  $$y_i \ (w^T x_i + b) \geq 1 \ , i = 1,2, \dots l.$$

Hard-margin SVM problem

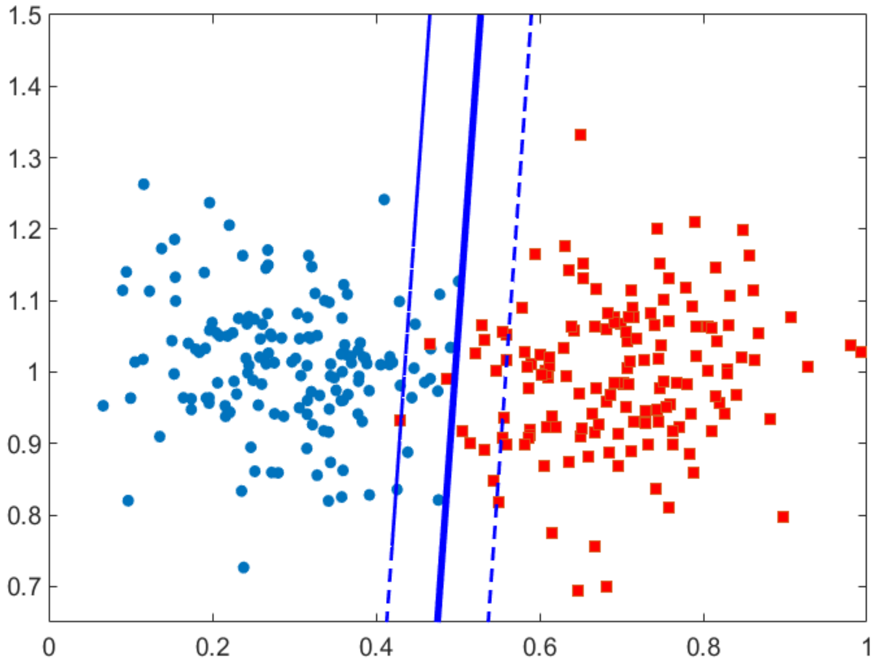


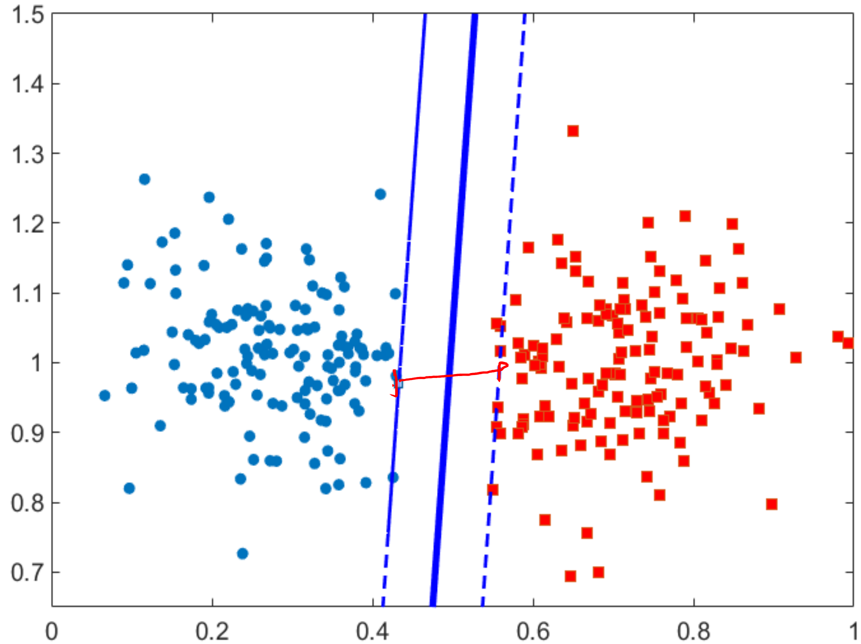
What to do for this





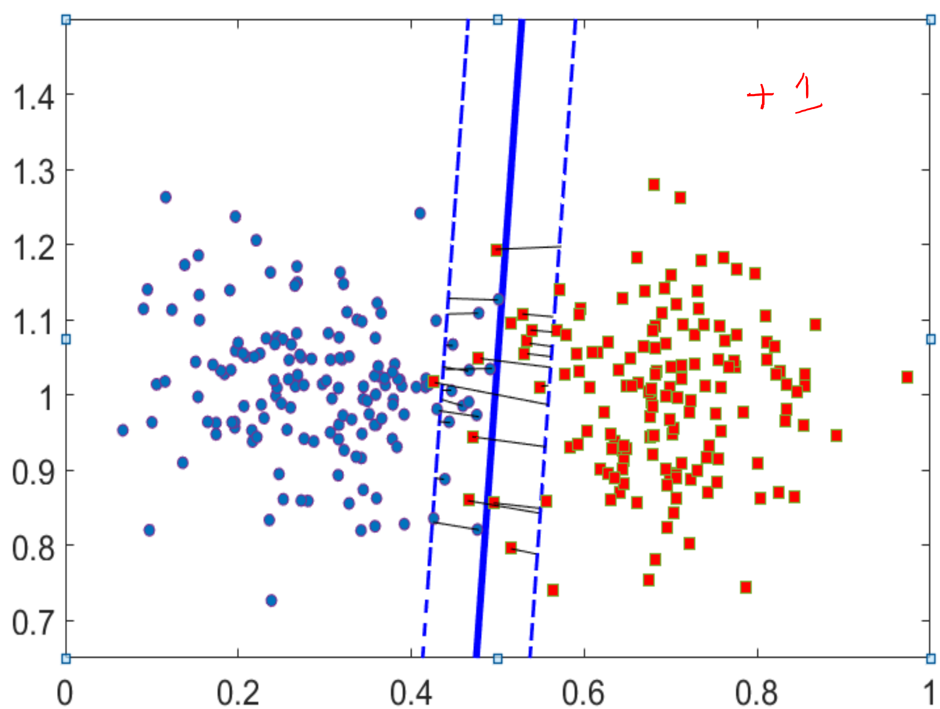
Soft-margin SVM





Soft-margin SVM

$y_i = 1$
$y = -1$



Relax the constraints by introducing slack variable ξ.

- $$\min_{(w,b)} \frac{1}{2}w^T w + C \sum_{i=1}^{l} \xi_i$$

  subject to,

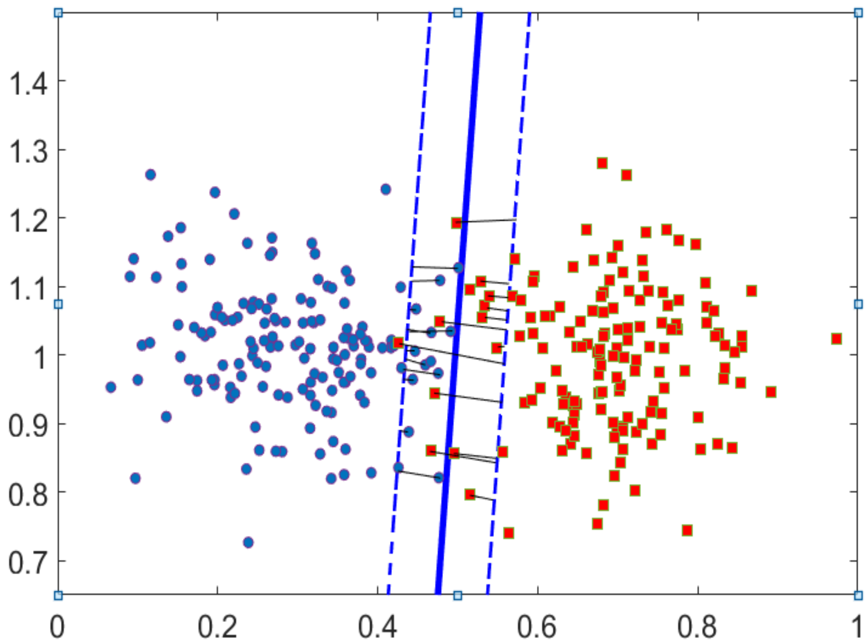
  $$y_i \ (w^T x_i + b) \geq 1 - \xi_i,$$
  $$\xi_i \geq 0, \qquad i = 1,2, \ldots l.$$

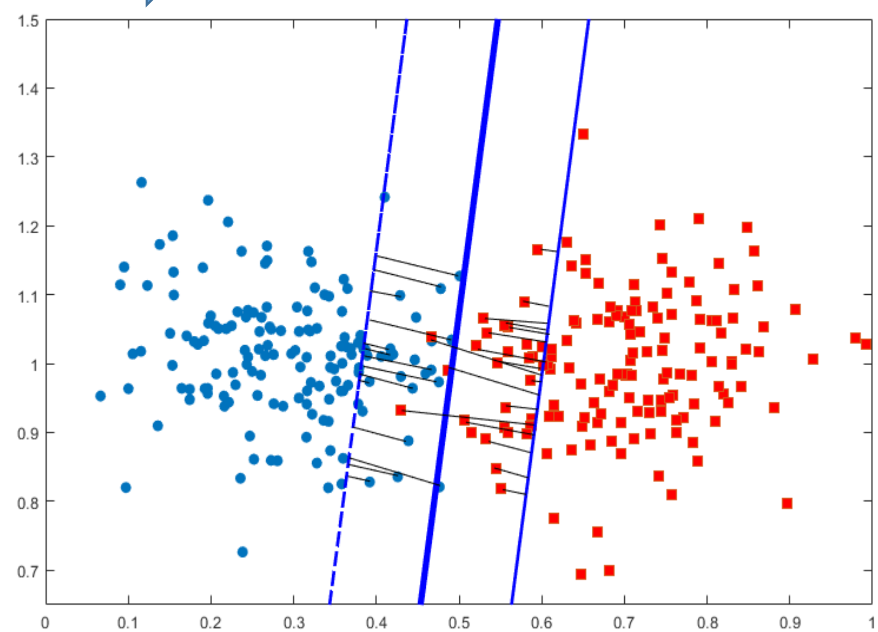
where C > 0 is the user defined parameter

Soft -margin SVM problem

Trading-off the training error and width of margin

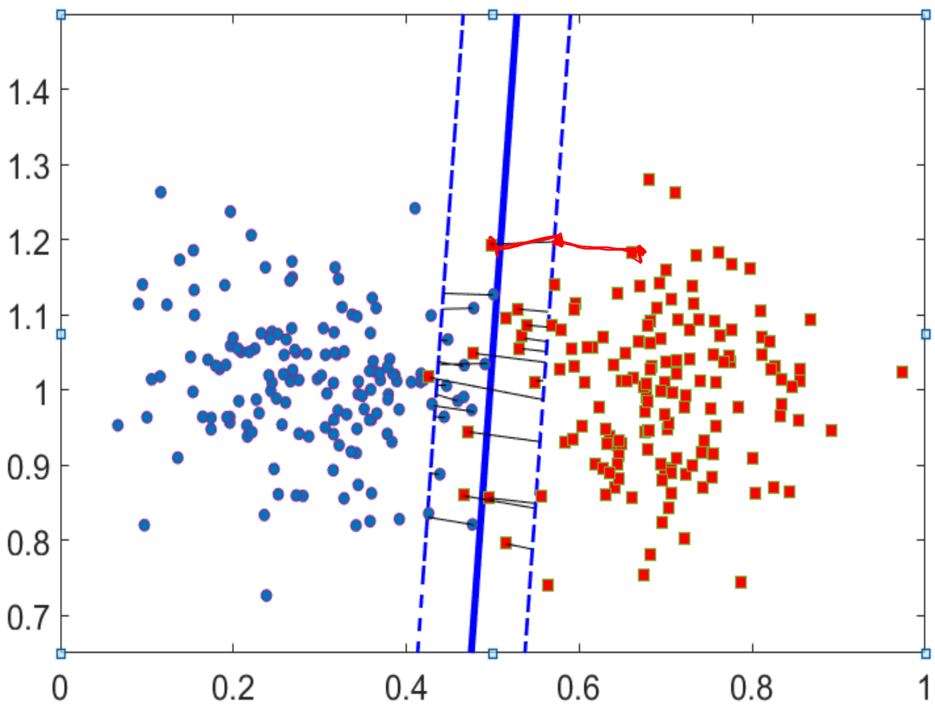
Decrease the values of C





Training Error

Soft-margin SVM optimization problem



- $$\min_{(w,b)} \ \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$
  subject to,
  $$y_i \ (w^T x_i + b) \geq 1 - \xi_i,$$
  $$\xi_i \geq 0, \qquad i = 1,2, \ldots l.$$

  where C > 0 is the user defined parameter.

- Limitation :- The number of constraint becomes huge for large-scale datasets

$$\min_{w,b} \ \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \ldots l$$

$$\xi_i \geq 0, \quad i = 1, 2, \ldots l$$

$\min (\max(a, b))$

$\min c$

$c > a$

$c > b$

— (1)

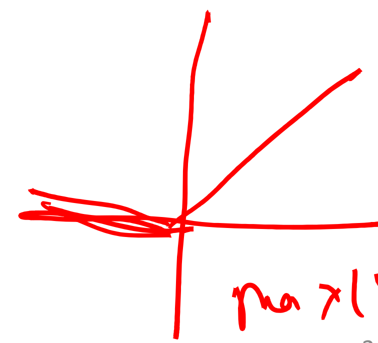
$$\min_{w,b} \ \frac{1}{2} w^T w + C \sum_{i=1}^{l} \max(0, \ 1 - y_i(w^T x_i + b)) \quad \text{— (2)}$$

$\xi_i =$

$$\xi_i = \max(0, 1 - y_i(w^T x_i + b))$$

$$\min \ \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$

$$\xi_i \geq 0, \quad \xi_i \geq 1 - y_i(w^T x_i + b)$$



$\max(u, 0)$

$$\min_{w,b} J(w,b) = \frac{1}{2}w^T w + C\sum_{i=1}^{l} \underbrace{\max(1-y_i(w^T x_i+b), 0)}$$

$\hookrightarrow$ $1-y_i(w^T x_i+b)$ if $y_i(w^T x_i+b)<1$

0 otherwise

$w \in R^n$

$b \in R$

$$\nabla_w J(w,b) = w + C\sum_{i=1}^{l} \begin{cases} -x_i y_i & \text{if } y_i(w^T x_i+b)<1 \\ 0 & \text{otherwise} \end{cases}$$

$$\nabla_b J(w,b) = C\sum_{i=1}^{l} \begin{cases} -y_i & \text{if } y_i(w^T x_i+b)<1 \\ 0 & \text{otherwise} \end{cases}$$

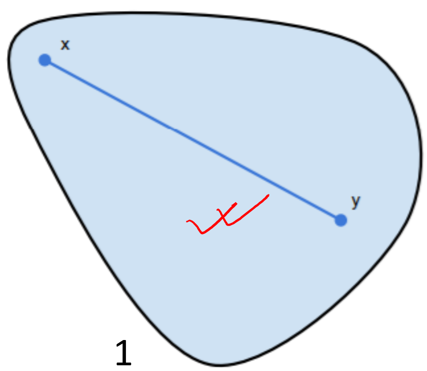
Convex Programming Problem

Convex Sets

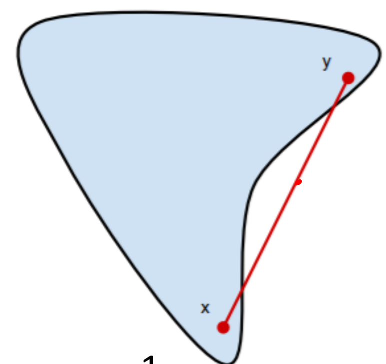
- $C \subseteq R^n$ is convex

  if $\lambda x + (1 - \lambda)y \in C$ for any $x, y \in C$ and $0 \le \lambda \le 1$.

  that is, a set is convex if the line connecting any two points in the set is entirely inside the set.



1



1

1. Nick Henderson et al., Convex set functions and Problems , 2019 .

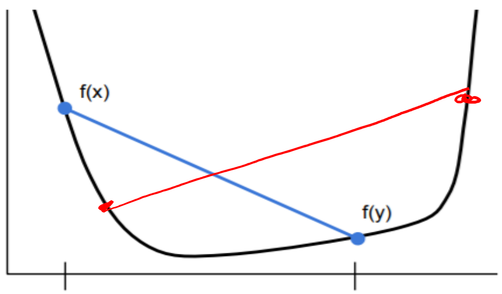
Convex Functions

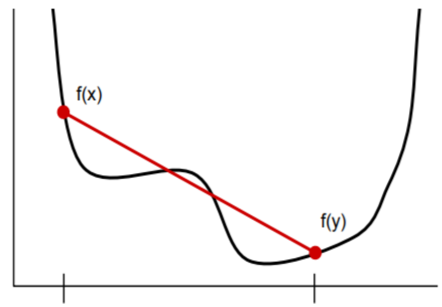
- f : $R^n \rightarrow R$ is convex

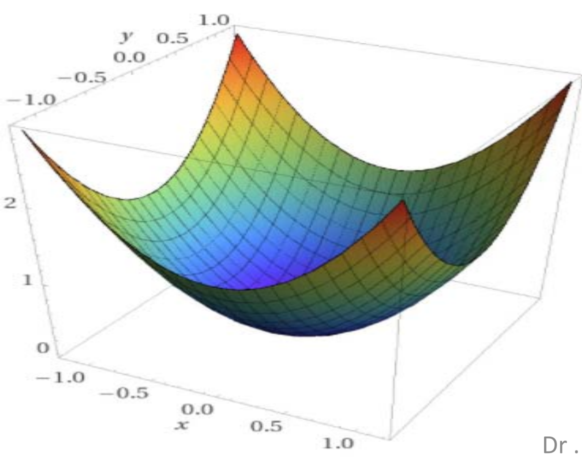
  if dom (f) (the domain of f) is a convex set,

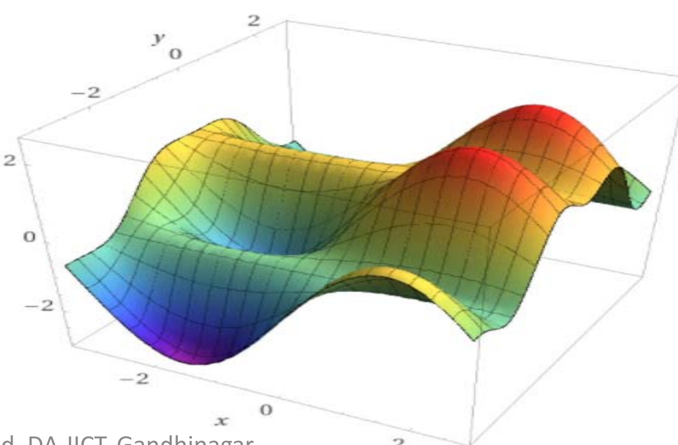
  and if $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ for any $x, y \in dom(f)$ and $0 \leq \lambda \leq 1$.

  that is, the line connecting any two points on the graph of the function stays above the graph.

[1]


[1]






1. Nick Henderson et al., Convex set functions and Problems , 2019 .

Convex Programming Problem

A convex programming problem is an optimization problem in the form

$$\min_{x} \quad f_0(x)$$ → Convex func of x

subject to ,

convex func ← $f_i(x) \leq 0, i = 1, 2, \ldots . m,$

$h_i(x) = 0, i = 1, 2, \ldots . p,$

$f(x)$

$g_i(x) \geq 0$

$-g_i(x) \leq 0$

where $f_0(x)$ and $f_i(x)$, i = 1,··· ,m are continuous convex functions of R^n, and $h_i(x)$, i = 1,··· ,p are linear functions.

Convex Programming Problem

- A Convex Programming Problem (CPP) is an optimization problem in the form

$$\min_{x} \quad f_0(x)$$

subject to ,

$$f_i(x) \leq 0, i = 1, 2, \dots . m,$$

$$h_i(x) = 0, i = 1, 2, \dots . p,$$

where $f_0(x)$ and $f_i(x)$,i = 1,$\cdots$,m are continuous convex functions of R^n, and $h_i(x)$, i = 1,$\cdots$,p are linear functions.

Convex Programming Problem

- A Convex Programming Problem (CPP) is an optimization problem in the form

$$\min_{x} \quad f_0(x)$$

subject to ,

$$f_i(x) \leq 0, i = 1, 2, \ldots . m,$$

$$h_i(x) = 0, i = 1, 2, \ldots . p, \qquad (1)$$

where $f_0(x)$ and $f_i(x)$, $i = 1, \cdots, m$ are continuous convex functions of R^n, and $h_i(x)$, $i = 1, \cdots, p$ are linear functions.

- If x^* is its local solution of CPP (1), then x^* is also its global solution.

Duality Theory

$D = \{ x : f_i(x) < 0, i = 1, 2 \ldots m$ and $h_j(x) = 0$ $j = 1, 2 \ldots p \}$

- A Convex Programming Problem (CPP) is an optimization problem in the form

$$\min_x \quad f_0(x)$$

subject to ,

$$f_i(x) \leq 0, i = 1, 2, \ldots . m,$$

$$h_i(x) = 0, i = 1, 2, \ldots . p, \qquad (1)$$

Min $x^2 + y^2$

subject

$x + y \leq 3$

$f_i(x^*) \leq 0$

$h_i(x^*) = 0$ . . . .

where $f_0(x)$ and $f_i(x)$, $i = 1, \cdots, m$ are continuous convex functions of R^n, and $h_i(x)$, $i = 1, \cdots, p$ are linear functions.

- Let p* is the optimal solution of CPP (1), i,e., p* = $\inf(f_0(x) | x \in D\}$, *where*,

D= $\{ x | f(x_i) \leq 0 , i = 1,2, \ldots m, h_i(x) = 0, i = 1, 2, \ldots . p\}$

Lagrangian Function

$$\max_{\lambda, \nu} \inf_{x\in \mathbb{R}^n} L(x,\lambda,\nu) \le \inf_{x\in D} L(x,\lambda,\nu) \le \inf_{x\in D} f_0(x)$$

The Lagrangian function is given as

$$L(x,\lambda,\nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x),$$

where $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)$ and $\nu = (\nu_1, \nu_2, \ldots, \nu_p)$ are the Lagrangian multipliers.

$[2, 5]$ $5x^2+2x$

Lagrangian Function

The Lagragian function is given as

$f_i(x) \leq 0$ $h_i(x) = 0$

$$L(x,\lambda,\nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x),$$

where λ= ($\lambda_{1,}$ $\lambda_{2,}$ ...., λ_{m}) and ν= ($\nu_{1,}$ $\nu_{2,}$ ...., ν_{p}) are the Lagrangian multipliers.

Obviously, when $x \in D, \lambda \geq 0$, we have

$$L(x,\lambda,\nu) \leq f_0(x).$$

Thus , $\underset{x \in Rn}{\mathrm{Inf}}\, L(x,\lambda,\nu) \leq \underset{x \in D}{\mathrm{Inf}}\, L(x,\lambda,\nu) \leq \underset{x \in D}{\mathrm{Inf}}\ \mathrm{f}_0(\mathrm{x}) = p*$.

Lagrangian dual function

$$\inf_{x \in R^n} L(x,\lambda,\nu) \leq \inf_{x \in D} L(x,\lambda,\nu) \leq \inf_{x \in D} f_0(x) = p* \; .$$

- Consider the dual function

$$g(\lambda,\nu) = \inf_{x \in Rn} L(x,\lambda,\nu)\,, \text{ then}$$
$$g(\lambda,\nu) \leq p^*$$

- The above inequality indicates that, for any $\lambda \geqslant 0$, $g(\lambda,\nu)$ is a lower bound of p^*.

- Among these lower bounds, finding the best one leads to the optimization problem

$$\max g(\lambda,\nu) = \inf_{x \in Rn} L(x,\lambda,\nu), \tag{2}$$
$$\text{subject to, } \lambda \geq 0$$

Weak Duality

- Let d* is solution of the dual problem

$$\max \left(g(\lambda, \nu) = \operatorname*{Inf}_{x \in Rn} L(x, \lambda, \nu) \right), \quad (2)$$
$$\text{subject to, } \lambda \geq 0$$

- Let p* is the optimal solution of the CPP(1)

$$\min_{x} \quad f_0(x)$$
subject to ,
$$f_i(x) \leq 0, i = 1, 2, \ldots . m,$$
$$h_i(x) = 0, i = 1, 2, \ldots . p,$$

then $d^* \leq p*$

Slater's condition

The CPP(1)

$$\min_{x} \quad f_0(x)$$

subject to ,

$$f_i(x) \leq 0, i = 1, 2, \dots . m,$$

$$h_i(x) = 0, i = 1, 2, \dots . p,$$

is said to satisfy the Slater's condition if there exists a feasible point x such that

$$f_{i(x)} < 0, i = 1, 2, \dots . m \text{ and}$$

$$h_i(x) = 0, i = 1, 2, \dots . p.$$

Strong Duality Theorem

- Let p* is the optimal value of the CPP (1) satisfying the Slater's condition and d* is the solution of the dual problem

$$\max \left(g(\lambda, \nu) = \inf_{x \in Rn} L(x, \lambda, \nu) \right),$$

$$\text{subject to, } \lambda \geq 0$$

then p* = d*.

- Furthermore, if p^* is attained, i.e. there exists a solution x^* to the primal CPP(1) , then d^* is also attained, i.e. there exists a global solution (λ^*, ν^*) to the dual problem such that $p^* = f_0(x^*) = g(\lambda^*, \nu^*) = d^* < \infty$.

KKT conditions

For the given CPP (1), the Point x^* is said to satisfy the Karush-Kuhn-Tucker(KKT) conditions, if there exist the multipliers $\lambda^* = (\lambda_1^*, \lambda_2^* \cdots, \lambda_m^*)$ and $\nu = (\nu_1^*, \nu_2^*, \ldots, \nu_p^*)$ corresponding to constraints

$$f_i(x) \le 0, i = 1, 2, \ldots . m,$$

$$h_i(x) = 0, i = 1, 2, \ldots . p,$$

min $f(x)$

Subject to

$f_i(x) \le 0, i = 1, 2, \ldots m$

$h_j(x) = 0, j = 1, 2, \ldots p$

respectively, such that the Lagrangian function

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

satisfies

$$\max_{\lambda, \nu} \left(\inf_{x} L(x, \lambda, \nu) \right)$$

KKT conditions

$$
\begin{cases}
f_i(x^*) \leq 0, i = 1, 2, \ldots . m, \\
h_i(x^*) = 0, i = 1, 2, \ldots . p, \\
\lambda_i^* \geq 0, i = 1, 2, \ldots . m, \\
\lambda_i^* f_i(x^*) = 0 \\
\nabla_x L(x, \lambda, \nu) = \nabla f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) + \sum_{i=1}^{p} \nu_i \nabla h_i(x) = 0
\end{cases}
$$

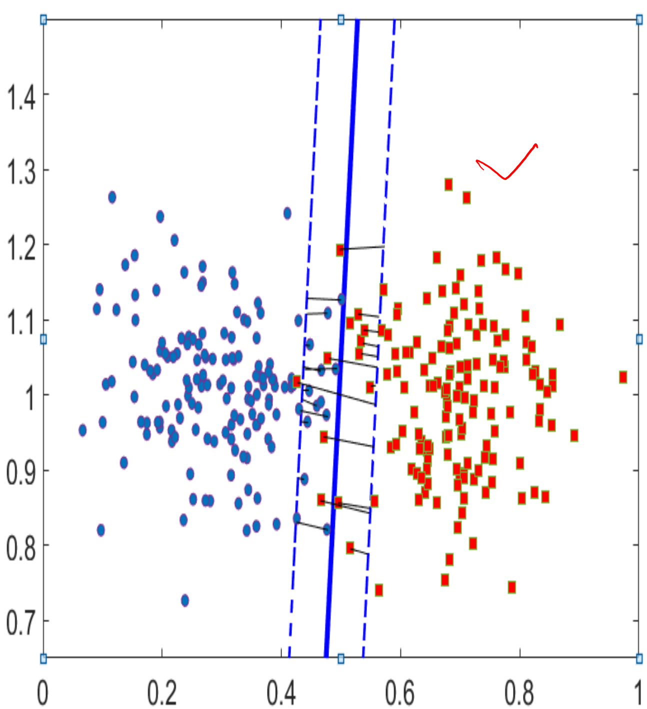
- If x* is the solution of the CPP(1) satisfying the Slater condition , then it must satisfy the KKT condition .
- If x* satisfies the KKT condition, then it is the solution of the CPP(1).

Back to SVM Optimization Problem

$-\alpha_i \left(1 - y_i(w^T x_i + b) - \xi_i \leq 0 \right)$

$-\alpha_i \left(y_i(w^T x_i + b) - 1 + \xi_i \right)$

SVM Optimization problem



SVM primal problem

- $$\min_{(w,b),\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$

  subject to,

  $$\boxed{y_i \; (w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0,} \quad i = 1,2,\ldots l.$$

  $-\xi_i \leq 0 \qquad -\xi_i \leq 0$

  where C > 0 is the user defined parameter

- Convex Programming Problem, hence global optimal solution.
- Quadratic Programming Problem, hence can be solved with QPP solver.
- But, need to handle too many constraints ,which makes the solution computationally expensive.

Lagrangian Function for SVM problem

$\xi_i \geq 0$

$-\xi_i \leq 0$

$\sum_{i=1} \alpha_i \xi_i$

$\beta_i \xi_i$

The Lagragian function is given as $\rightarrow f_0(x)$

$$L(w,b,\alpha,\beta,\xi) = \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i - \sum_{i=1}^{l} \alpha_i (y_i \ (w^T x_i) + b) - 1 + \xi_i, \ ) - \sum_{i=1}^{l} \beta_i \xi_i,$$

where, $\alpha = (\alpha_{1,} \alpha_{2,} \ldots, \alpha_l)$ and $\beta = (\beta_{1,} \beta_{2,} \ldots, \beta_l)$ are the positive Lagrangian multipliers.

Dual Problem:-

$$\max_{\alpha \geq 0, \beta \geq 0} \left(\inf_{(w,b,\xi)} L(w,b,\alpha,\beta,\xi) \right)$$

max

$\alpha \geq 0$

$\beta \geq 0$

$$\frac{1}{2} \left(\sum_{i=1}^{l} \alpha_i y_i x_i \right)^T \left(\sum_{j=1}^{l} \alpha_j y_j x_j \right) - \sum_{i=1}^{l} \alpha_i y_i \left(\sum_{j=1}^{l} \alpha_j y_j x_j \right)^T x_i + \sum_{i=1}^{l} \alpha_i$$

subject to,

$C - \alpha_i - \beta_i = 0$

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

KKT conditions for optimality

$\underline{\alpha_i = C}$ $\quad \beta_i = 0$

$y_i(w^T x_i + b) - 1 + \xi_i = 0$

$y_i(w^T x_i + b) - 1 = -\xi_i$

$1 - y_i(w^T x_i + b) = \xi_i$

$1 - y_i(w^T x_i + b) > 0$

$y_i(w^T x_i + b) < 1$

$$\nabla_{\mathrm{w}} L(w, b, \alpha, \beta, \xi) = 0 \implies \mathrm{w} = \sum_{i=1}^{l} \alpha_i y_i x_i \qquad (1)$$

$$\nabla_{\mathrm{b}} L(w, b, \alpha, \beta, \xi) = 0 \implies \sum_{i=1}^{l} \alpha_i y_i = 0 \qquad (2)$$

$$\nabla_{\xi} L(w, b, \alpha, \beta, \xi) = 0 \implies C - \alpha_i - \beta_i = 0 \qquad (3)$$

$$y_i \ (w^T x_i + b) \geq 1 - \xi_i, \quad i = 1,2, \ldots l. \qquad (4)$$

$$\xi_i \geq 0, \ i = 1,2, \ldots l. \qquad (5)$$

$$\alpha_i (y_i \ (w^T x_i + b) - 1 + \xi_i,) = 0, \ i = 1,2, \ldots l. \qquad (6)$$

$$\beta_i \xi_i = 0, \ i = 1,2, \ldots l. \qquad (7)$$

$$\alpha_i \geq 0, i = 1,2, \ldots l. \qquad (8)$$

$$\beta_i \geq 0, i = 1,2, \ldots l. \qquad (9)$$

$(\alpha_1 y_1 x_1 + \alpha_2 y_2 x_2 + \cdots \alpha_l y_l x_l)(\alpha_1 y_1 x_1 + \cdots$

$$\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i^T x_j)$$

Dual SVM Problem

The Lagragian function is given as

$$L(w,b,\alpha,\beta,\xi) = \frac{1}{2}w^T w + C\ \sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\alpha_i(y_i\ \ (w^T x_i + b) - 1 + \xi_i,\ ) - \sum_{i=1}^{l}\beta_i \xi_i,$$

where, α= $(\alpha_{1,}\ \alpha_{2,} \ldots, \alpha_l)$ and β= $(\beta_{1,}\beta_{2,} \ldots, \beta_l)$ are the positive Lagrangian multipliers.

Dual Problem:-

$$\max_{\alpha \geq 0, \beta \geq 0} (\text{lnf}_{(w,b,\xi)}\ L(w,b,\alpha,\beta,\xi))$$

$$\max_{\alpha \geq 0, \beta \geq 0} \frac{1}{2}\left(\sum_{i=1}^{l}\alpha_i y_i x_i\right)^T \sum_{j=1}^{l}(\alpha_j y_j x_j) + \sum_{i=1}^{l}\alpha_i y_i - \sum_{i=1}^{l}\alpha_i y_i \left(\sum_{j=1}^{l}\alpha_j y_j x_j\right)^T x_i$$

Subject to, $\sum_{i=1}^{l}\alpha_i y_i = 0,$

$C - \alpha_i - \beta_i = 0\ ,\ i = 1, 2, \ldots l$

$$\max_{\alpha, \beta} \; -\frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i^T x_j) + \sum_{i=1}^{l} \alpha_i y_i$$

Subject to

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

$$C - \alpha_i - \beta_i = 0 \Rightarrow 0 \leq \alpha_i \leq C$$

$$\alpha_i, \beta_i \geq 0$$

Dual SVM Problem

Dual Problem:- $\max_{\alpha \geq 0, \beta \geq 0} (\text{Inf}_{(w,b,\xi)} L(w, b, \alpha, \beta, \xi))$

$$where, \quad L(w, b, \alpha, \beta, \xi) = \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i - \sum_{i=1}^{l} \alpha_i (y_i \ (w^T x_i + b) - 1 + \xi_i, \ ) - \sum_{i=1}^{l} \beta_i \xi_i,$$

$$\max_{\alpha \geq 0, \beta \geq 0} \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \ \alpha_j y_i y_j x_i{}^T x_j - \sum_{i=1}^{l} \alpha_i \big(y_i \ \big(\sum_{j=1}^{l} \ \alpha_j y_j x_j\big)^T x_i\big) - 1\big)$$

Subject to, $\sum_{i=1}^{l} \alpha_i y_i \ = 0,$

$C - \alpha_i - \beta_i = 0,$

$\alpha_i \geq 0, \beta_i \geq 0, i = 1,2, \ldots l$

Dual SVM Problem

$k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$

$$\max_{\alpha} \frac{-1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \, \alpha_j y_i y_j {x_i}^T x_j + \sum_{i=1}^{l} \alpha_i$$

Subject to, $\sum_{i=1}^{l} \alpha_i y_i \; = 0,$

$0 \le \alpha_i \le C, i = 1,2, \dots l.$

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \, \alpha_j y_i y_j {x_i}^T x_j - \sum_{i=1}^{l} \alpha_i$$

Subject to, $\sum_{i=1}^{l} \alpha_i y_i \; = 0,$

$0 \le \alpha_i \le C, i = 1,2, \dots l.$

Dual SVM Problem

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \, \alpha_j y_i y_j {x_i}^T x_j - \sum_{i=1}^{l} \alpha_i$$

Subject to, $\sum_{i=1}^{l} \alpha_i y_i \ = 0,$

$0 \le \alpha_i \le C, i = 1,2, \ldots l.$

$$\min_{(w,b)} \ \frac{1}{2} w^T w + C \ \sum_{i=1}^{l} \xi_i$$

subject to,

$y_i \ (w^T x_i + b) \ge 1\text{-}\ \xi_i,$

$\xi_i \ge 0, \qquad i = 1,2, \ldots l.$

- The Dual problem is a Quadratic Programming Problem (QPP) but, less number of constraints.
- Suppose that $\alpha * =(\alpha_1 *, . . ., \alpha_l *)$ is a solution to the dual problem, Then the training point (x_i,y_i), is said to be a support vector if the corresponding component $\alpha_i *$ is non-zero and otherwise it is a non-support vector.

$(x_1, y_1) \rightarrow \alpha_1^*$

$(x_2, y_2) \rightarrow \alpha_2^*$

$(x_l, y_l) \rightarrow \alpha_l^*$

SVM Solution

$$\min_{(w,b)} \quad \frac{1}{2}w^T w + C \sum_{i=1}^{l} \xi_i$$

subject to,

$$y_i \;(w^T x_i + b) \geq 1 - \xi_i,$$
$$\xi_i \geq 0, \qquad i = 1,2, \ldots l.$$

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i\, \alpha_j y_i y_j {x_i}^T x_j - \sum_{i=1}^{l} \alpha_i$$

Subject to, $\sum_{i=1}^{l} \alpha_i y_i \;= 0,$

$0 \leq \alpha_i \leq C, i = 1,2, \ldots l.$

- After obtaining the solution of dual problem $\alpha * =({\alpha^*}_1,\ldots,{\alpha^*}_l)$, the w^* can be obtained by using the KKT condition (1) as

$$w^* = \sum_{i=1}^{l} {\alpha^*}_i y_i x_i$$

$\frac{1}{y_j} = y_j$

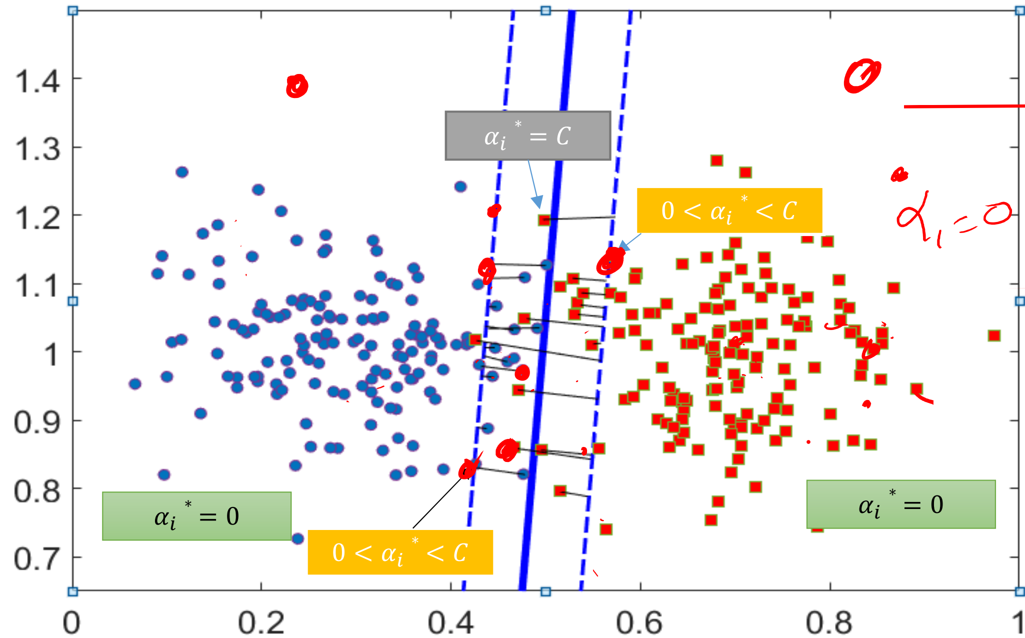
- Further, for a given $0 < {\alpha^*}_j < C$, we have $y_j \left(w^{*T} x_j + b^*\right) = 1$ from KKT condition (6) and (7) which gives,

$T x_j$

$$b^* = y_j - \sum_{i=1}^{l} {\alpha^*}_i y_i {x^T}_i x_j$$

- In practice, we consider all ${\alpha^*}_j$, satisfying $0 < {\alpha^*}_j < C$ and compute the value of b*. The final value of b* is considered as the mean of all computed values.

SVM Solution



$\alpha_i = C$

$y_i(w^T x_i + b) \geq 1 - \xi_i$

$\alpha_i = 0$

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

$\xi_i > 0$

$0 < \alpha_i < C$

- For data point (x_i, y_i), satisfying $y_i(w^{*T} x_i + b^*) < 1$, the corresponding $\alpha^*_i = C$.
- For data point (x_i, y_i), satisfying $y_i(w^{*T} x_i + b^*) = 1$, the corresponding $0 < \alpha^*_i < C$.
- For data point (x_i, y_i), satisfying $y_i(w^{*T} x_i + b^*) > 1$, the corresponding $\alpha^*_i = 0$.

SVM is a sparse classification model

Proof using the KKT conditions

$$\nabla_{\mathrm{w}} L(w, b, \alpha, \beta, \xi) = 0 \quad \Longrightarrow \mathrm{w} = \sum_{i=1}^{l} \alpha_i y_i x_i \qquad (1)$$

$$\nabla_{\mathrm{b}} L(w, b, \alpha, \beta, \xi) = 0 \quad \Longrightarrow \sum_{i=1}^{l} \alpha_i y_i x_i \; = 0 \qquad (2)$$

$$\nabla_{\xi} L(w, b, \alpha, \beta, \xi) = 0 \quad \Longrightarrow C - \alpha_i - \beta_i = 0, \qquad (3)$$

$$y_i \; (w^T x_i + b) \geq 1 - \xi_i \,, \quad i = 1,2, \dots l. \qquad (4)$$

$$\xi_i \geq 0, \;\; i = 1,2, \dots l. \qquad (5)$$

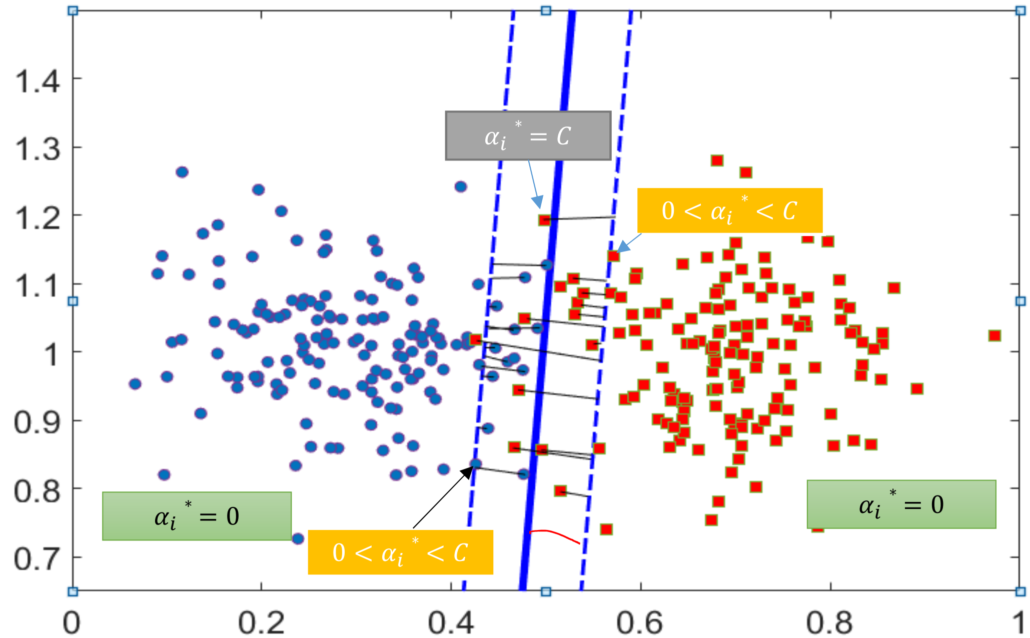
$$\alpha_i (y_i \; (w^T x_i + b) - 1 + \xi_i, \;) = 0, i = 1,2, \dots l. \qquad (6)$$

$$\beta_i \xi_i = 0, i = 1,2, \dots l. \qquad (7)$$

$$\alpha_i \geq 0, i = 1,2, \dots l. \qquad (8)$$

$$\beta_i \geq 0, i = 1,2, \dots l. \qquad (9)$$

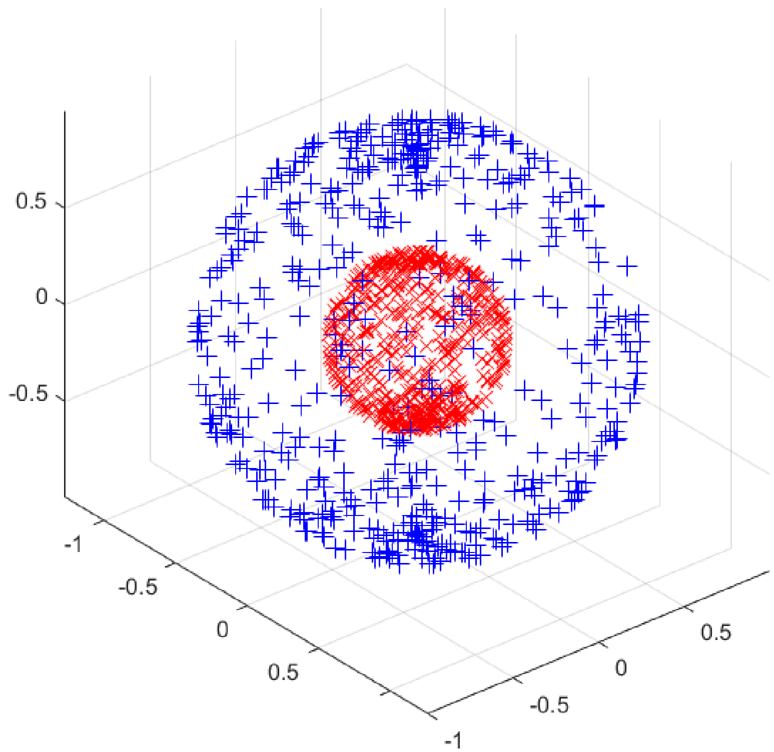
SVM Solution



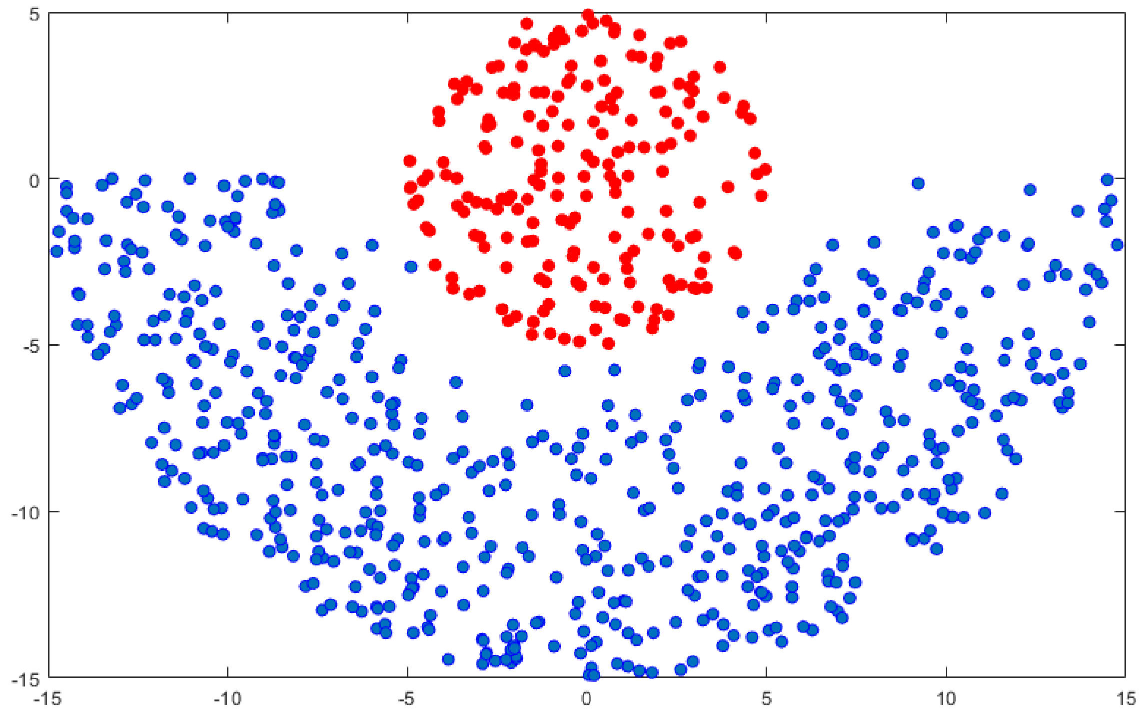
- For data point (x_i, y_i), satisfying $y_i(w^{*T}x_i + b^*) < 1$, the corresponding α^*_i = C.
- For data point (x_i, y_i), satisfying $y_i(w^{*T}x_i + b^*) = 1$, the corresponding $0 < \alpha^*_i$ < C.
- For data point (x_i, y_i), satisfying $y_i(w^{*T}x_i + b^*) > 1$, the corresponding α^*_i = 0.

For a test point say $\bar{x}$, the SVM prediction is $sign(w^{*T}\bar{x} + b^*)$

Non-linear separable data

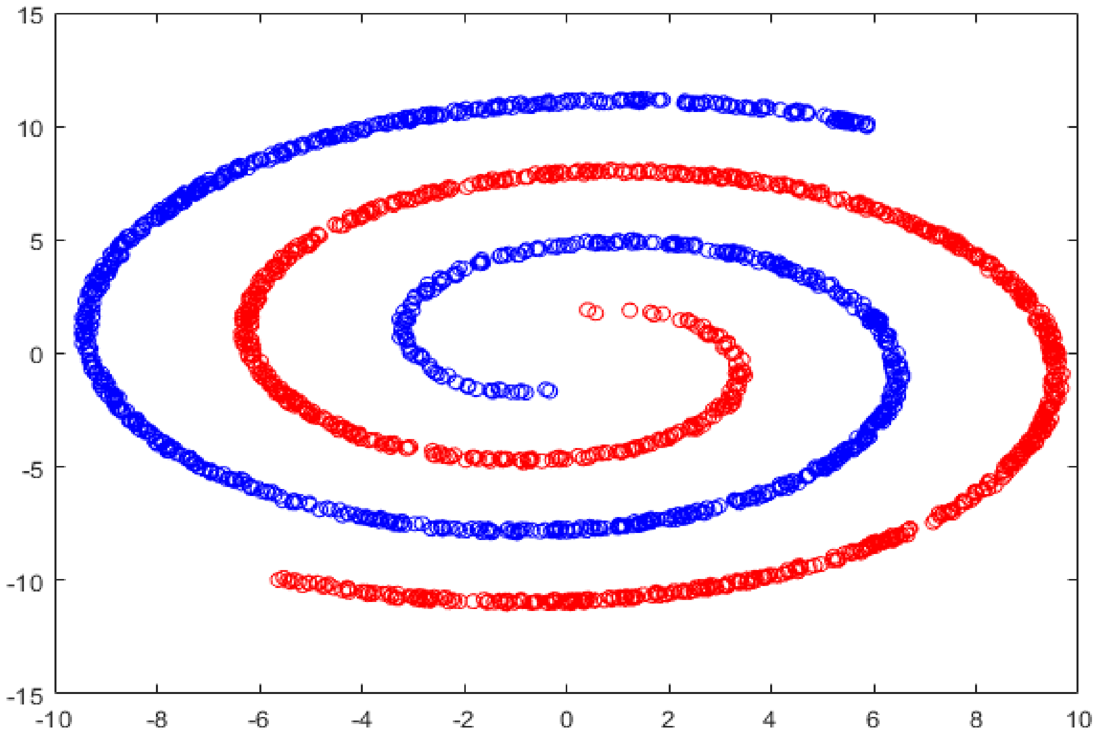


Non-linear separable data



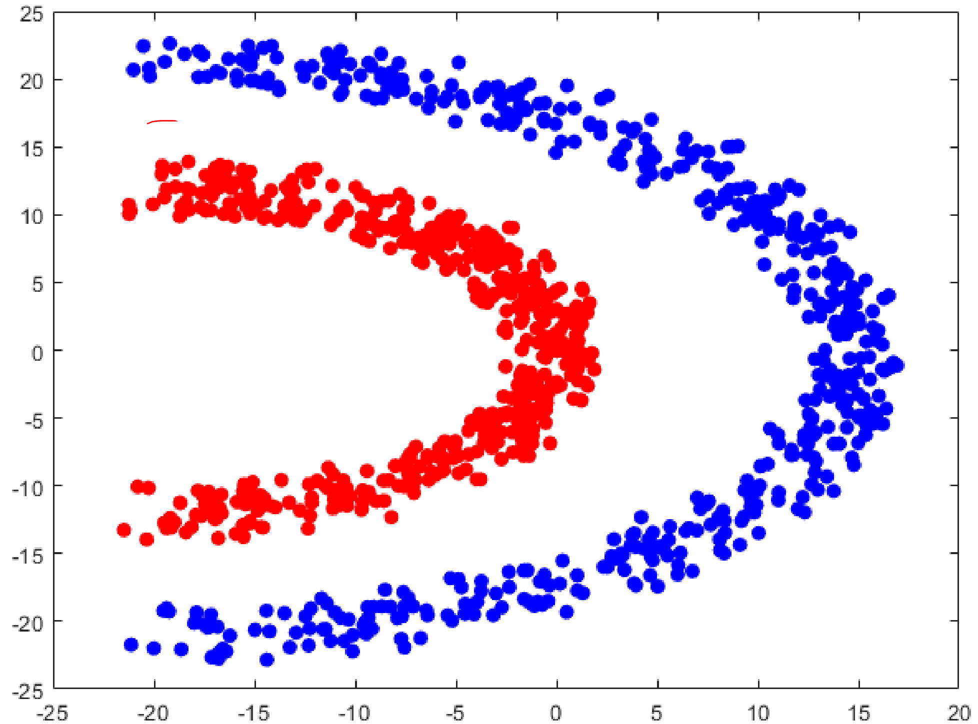
Crescent full moon dataset

Non-linear separable data



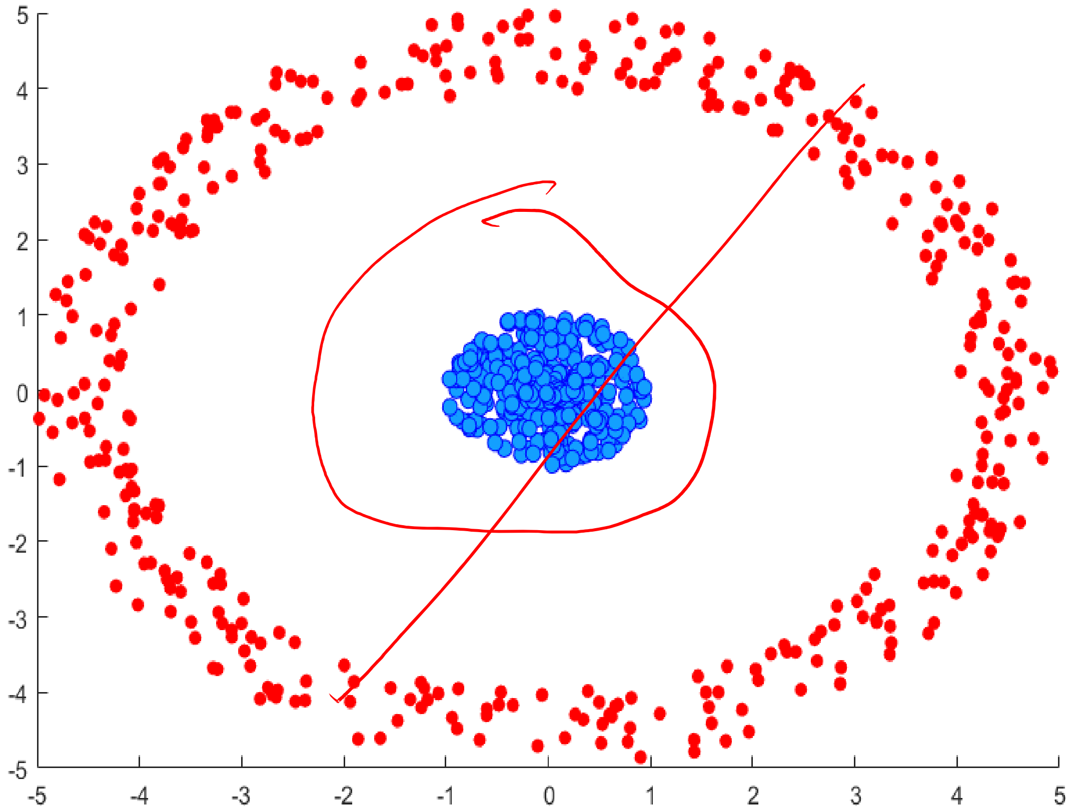
Spiral Dataset

Non-linear separable data



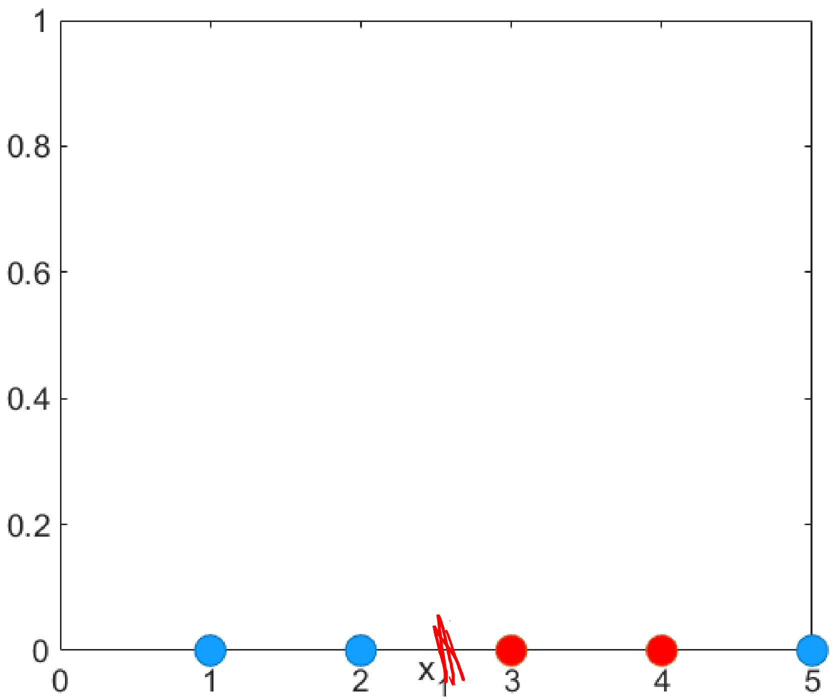
Half kernel Dataset

Non-linear separable data

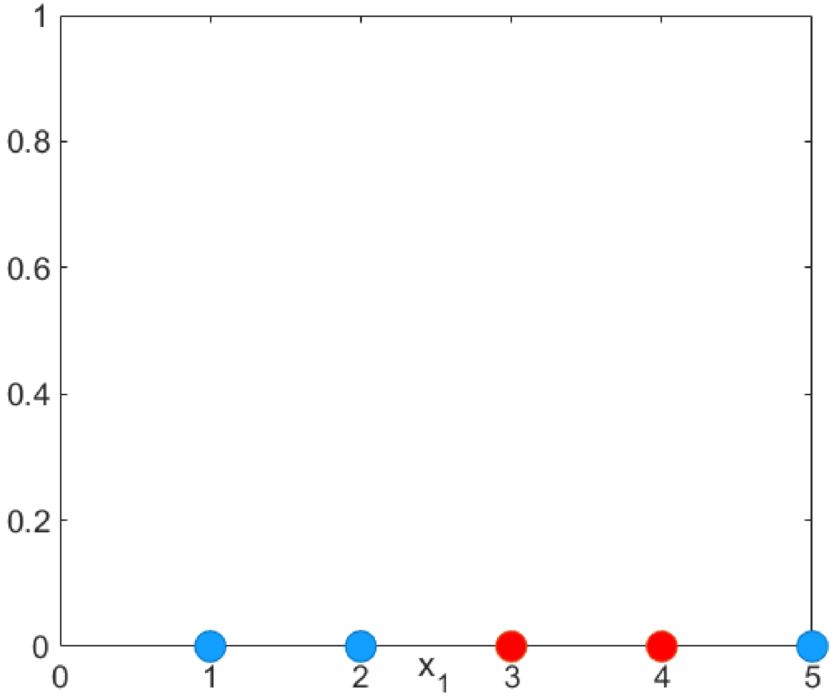


Co-centric Circle Dataset

Kernel Trick



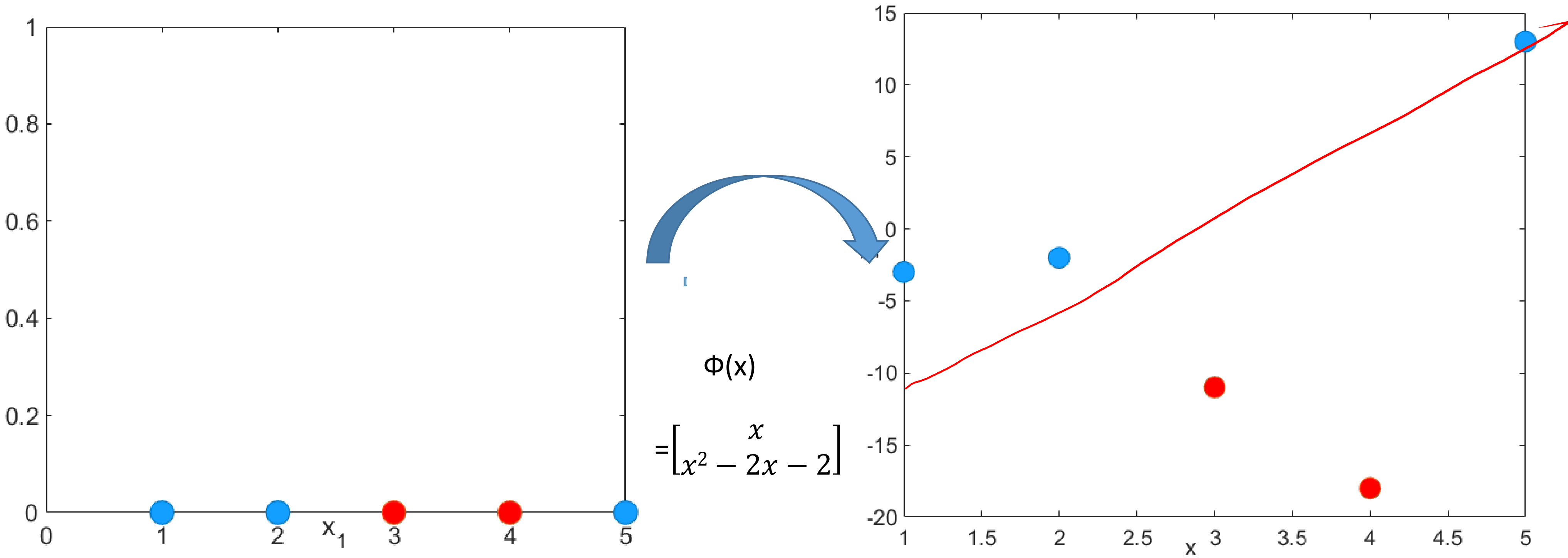
Kernel Trick



Φ(x)

$$=\begin{bmatrix} x \\ x^2 - 2x - 2 \end{bmatrix}$$

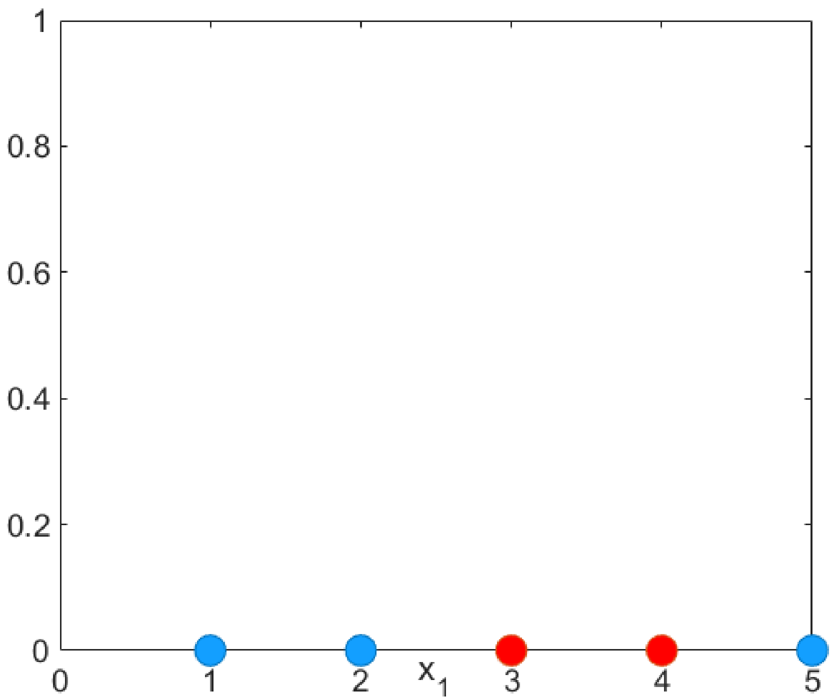
Kernel Trick



Φ(x)

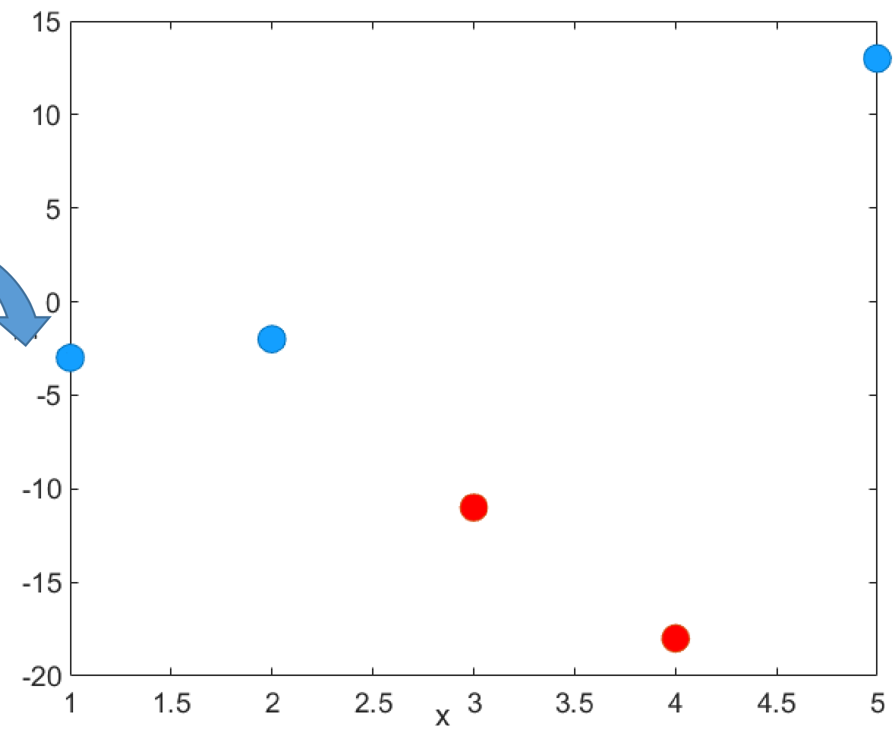
$$= \begin{bmatrix} x \\ x^2 - 2x - 2 \end{bmatrix}$$

Kernel Trick

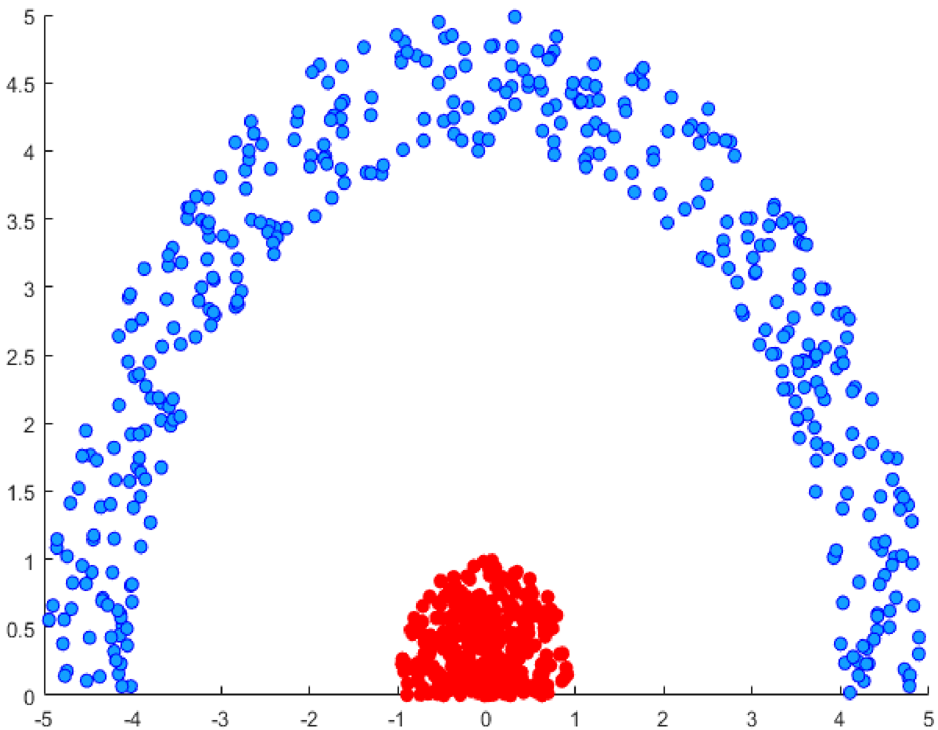


$\Phi(x)$

$$=\begin{bmatrix} x \\ x^2 - 2x - 2 \end{bmatrix}$$

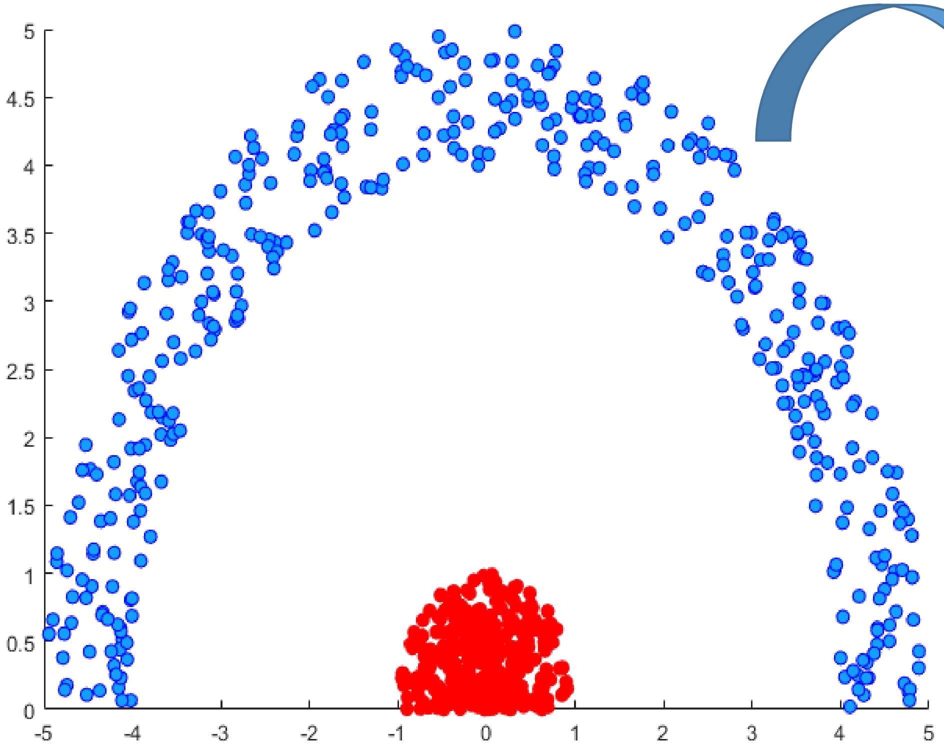


Kernel Trick



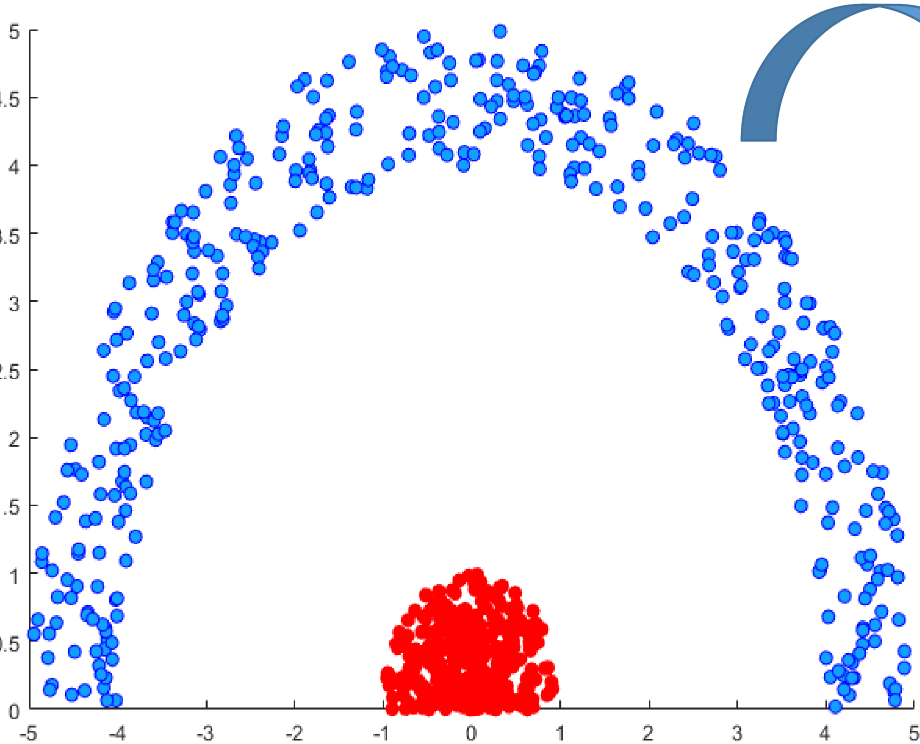
Kernel Trick

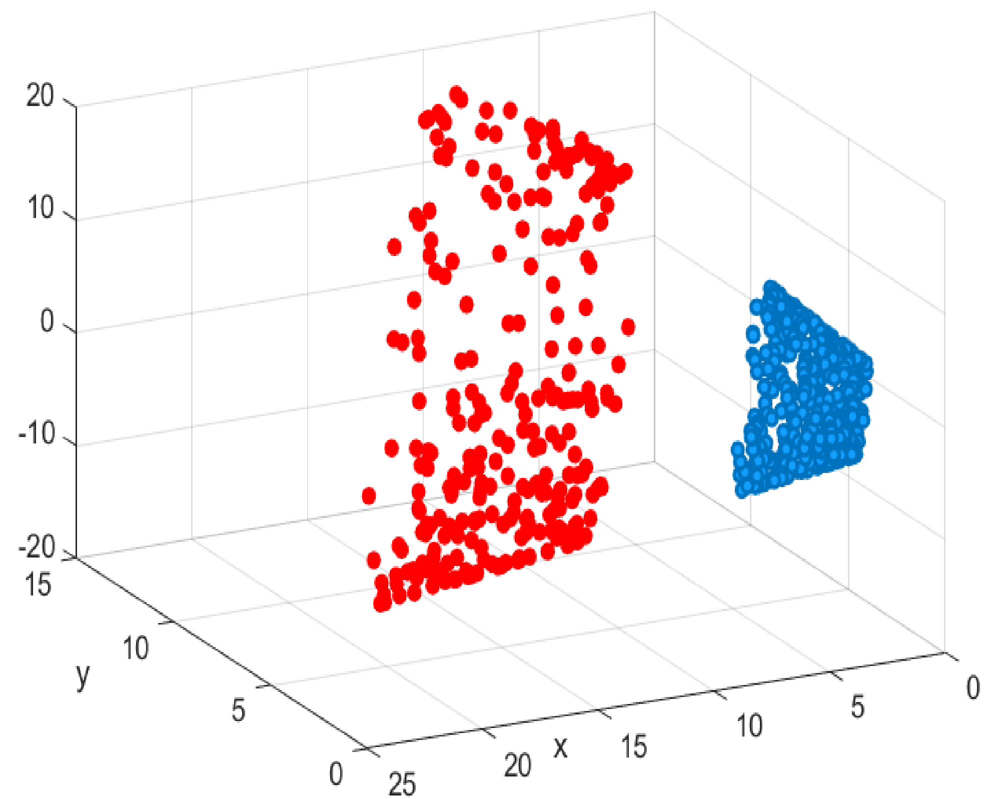
$$\Phi(x) = \Phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} {x_1}^2 \\ {x_2}^2 \\ \sqrt{2}x_2x_1 \end{bmatrix}$$



Kernel Trick

$$\Phi(x) = \Phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} {x_1}^2 \\ {x_2}^2 \\ \sqrt{2}x_2x_1 \end{bmatrix}$$

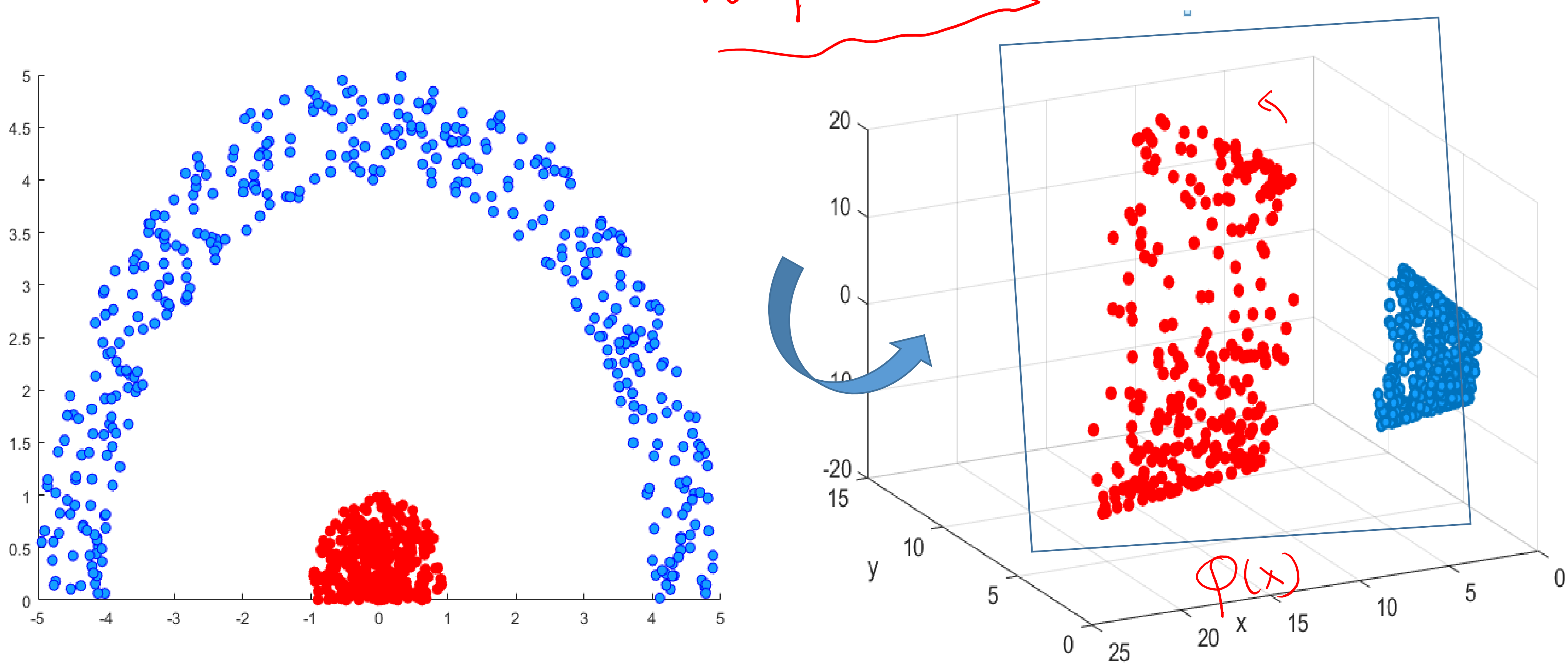




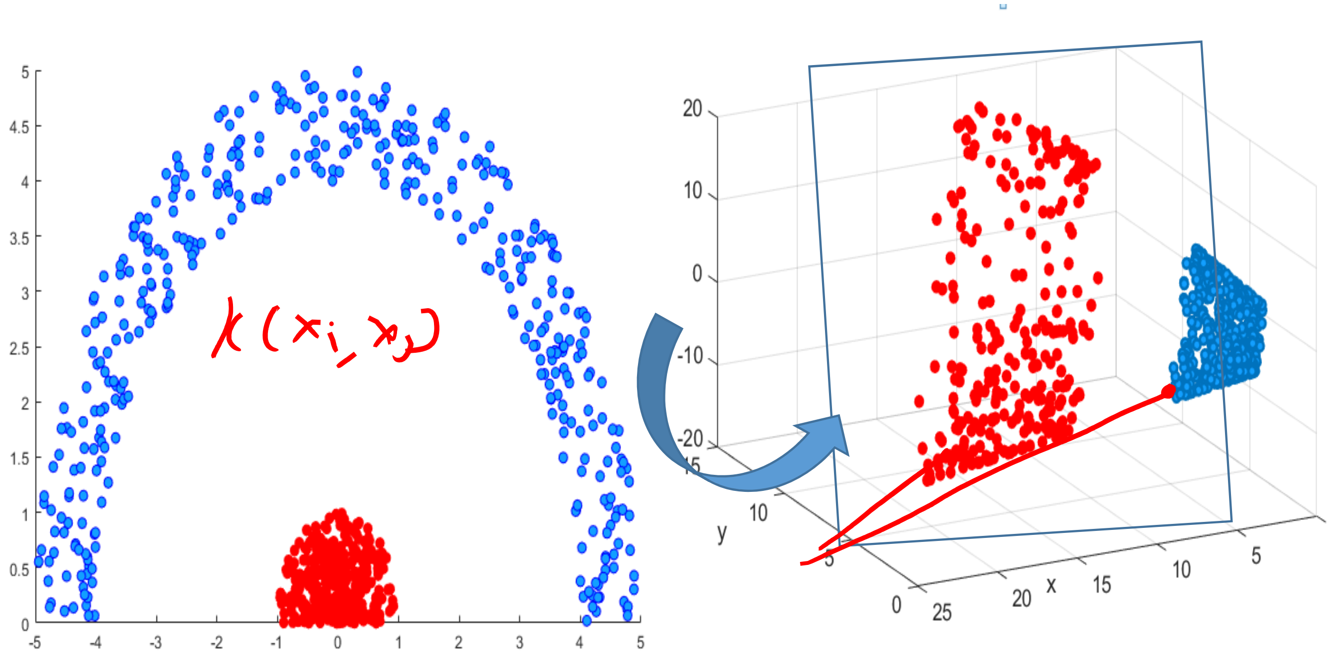
Kernel Trick

$$\Phi(x) = \Phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_1^{\,2} \\ x_2^{\,2} \\ \sqrt{2}x_2x_1 \end{bmatrix}$$

$w^T\phi(x) + b = 0$



Non-linear SVM



Primal Problem

$$\min_{(w,b)} \ \frac{1}{2} w^T w + C \ \sum_{i=1}^{l} \xi_i$$

subject to,

$$y_i \ \ (w^T \phi(x_i) + b) \geq 1 - \xi_i,$$

$$\xi_i \geq 0, \qquad i = 1,2, \ldots l.$$

Dual Problem

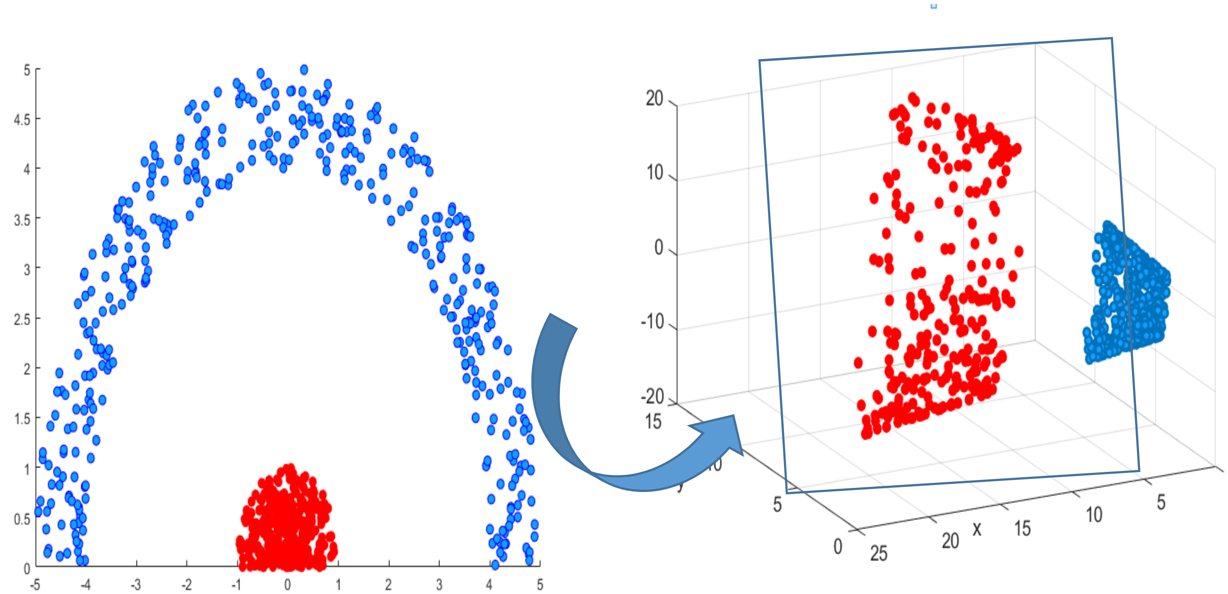
$\phi(x_i)^T \phi(x_j)$

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \, \alpha_j y_i y_j \phi(x_i)^T \phi(x_j) - \sum_{i=1}^{l} \alpha_i$$

Subject to, $\sum_{i=1}^{l} \alpha_i y_i \ = 0,$

$0 \leq \alpha_i \leq C, i = 1,2, \ldots l.$

Non-linear SVM



- The dual problem requires the knowledge of only $\phi(x_i)^T\phi(x_j)$.
- We can use a kernel function such that $k(x_i, x_j) = \phi(x_i)^T\phi(x_j)$.
- Dual Problem can be solved without the explicit knowledge of mapping ϕ .

Dual Problem

$K(x_i, x_j)$

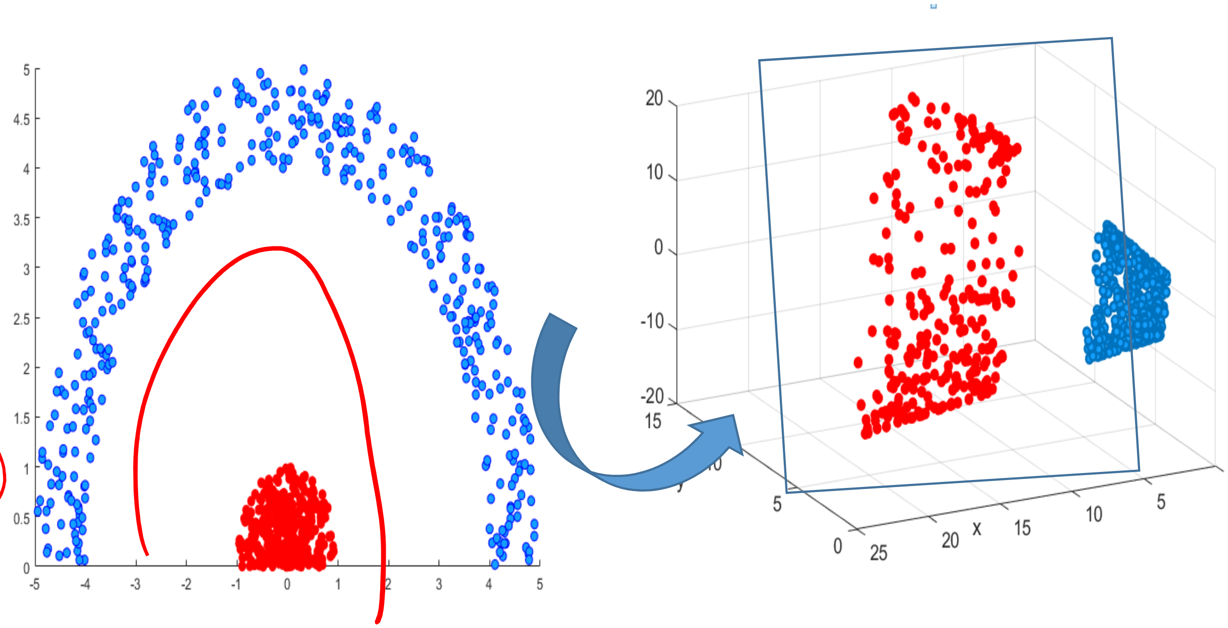
$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\,\alpha_j y_i y_j \phi(x_i)^T\phi(x_j) - \sum_{i=1}^{l}\alpha_i$$

Subject to, $\sum_{i=1}^{l}\alpha_i y_i = 0,$
$0 \le \alpha_i \le C, i = 1,2, \ldots l.$

Non-linear SVM solution

$$sign\left(\sum_{i=1}^{l} \alpha_i y_i \phi(x_i)\right)^T \phi(x)$$

$$\sum_{i=1}^{l} y_i \alpha_i \left(\underbrace{\phi(x_i)^T \phi(x)}_{k(x_i,\, x)}\right)$$



Dual Problem

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i\, \alpha_j y_i y_j k(x_i, x_j) - \sum_{i=1}^{l} \alpha_i$$

Subject to, $\sum_{i=1}^{l} \alpha_i y_i \ = 0,$

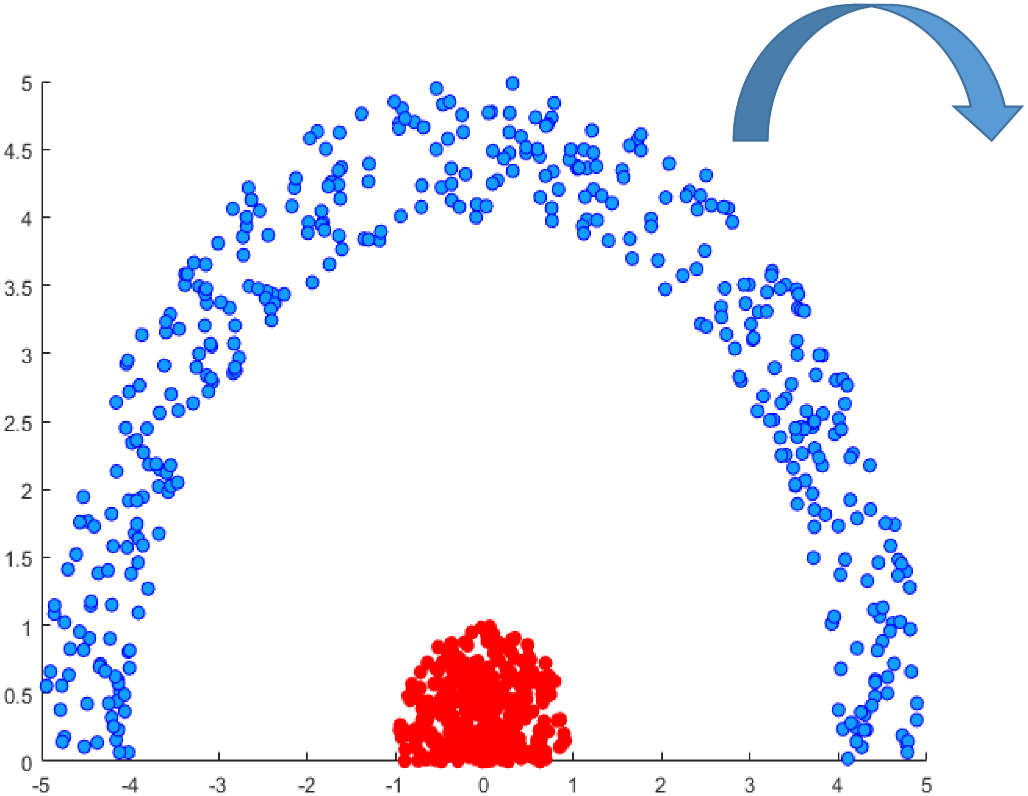
$0 \leq \alpha_i \leq C, i = 1,2, \dots l.$

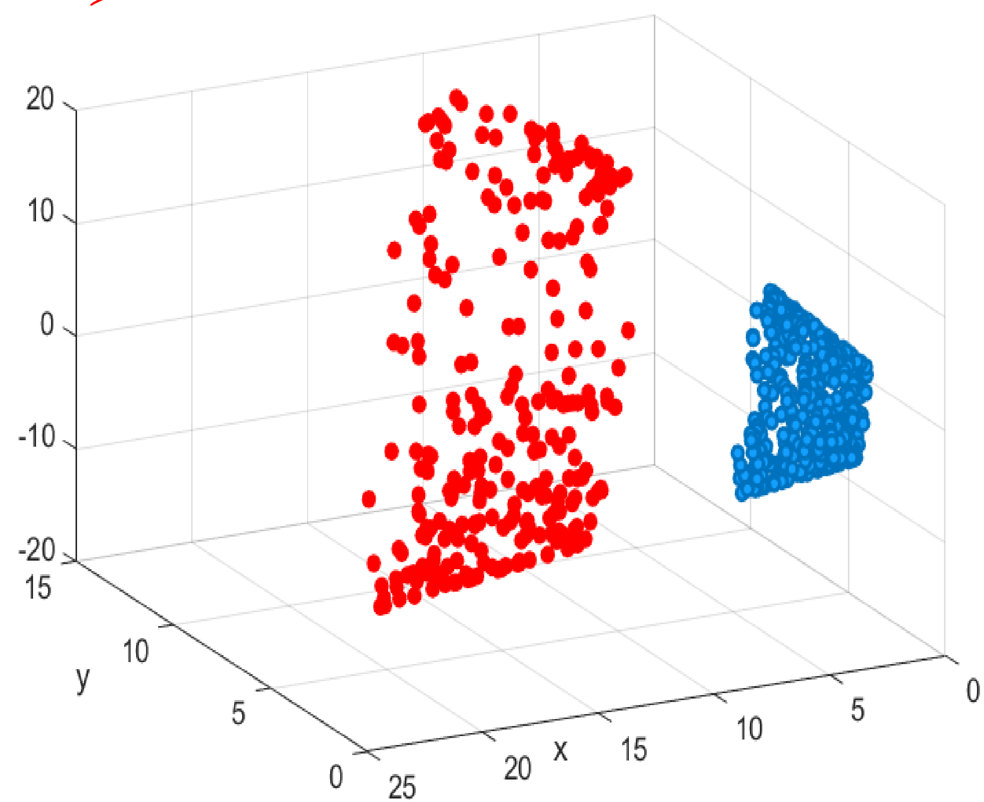
After obtaining the optimal solution of the dual problem $\alpha * =(\alpha^*_1,. . .,\alpha^*_l)$, the decision function can be obtained as

$$f(x) = sign(w^{*T}\, \phi(x) + b^*)$$
$$= sign\left(\sum_{i=1}^{l} \alpha^*_i\, y_i \phi(x_i)^T \phi(x) + b^*\right)$$
$$= sign\left(\sum_{i=1}^{l} \alpha^*_i y_i\, k(x_i, x) + b^*\right)$$

Kernel Trick

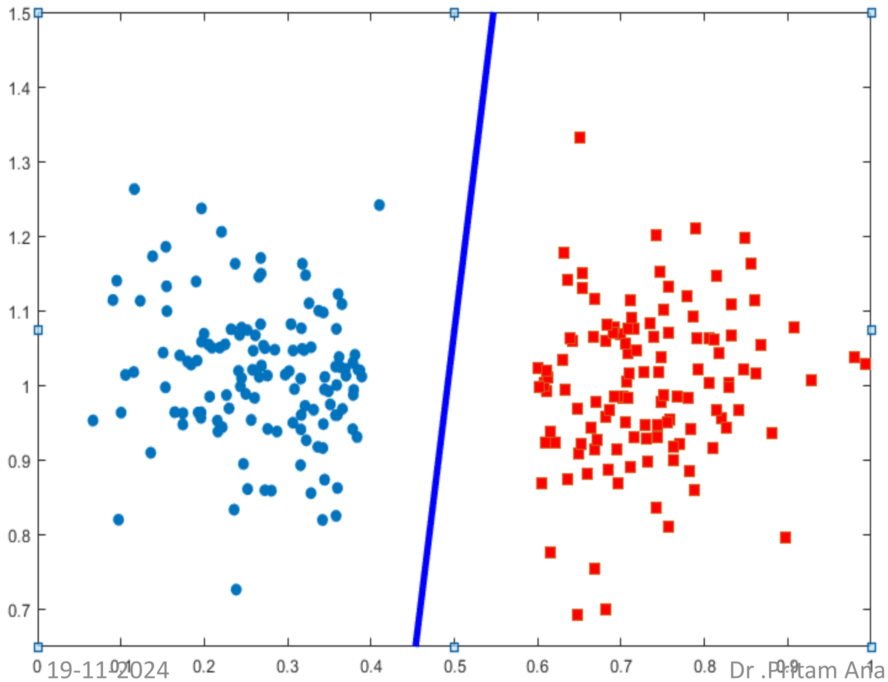
$$k(x,y) = \Phi(x)^T\Phi(y) = (x^Ty)^2$$





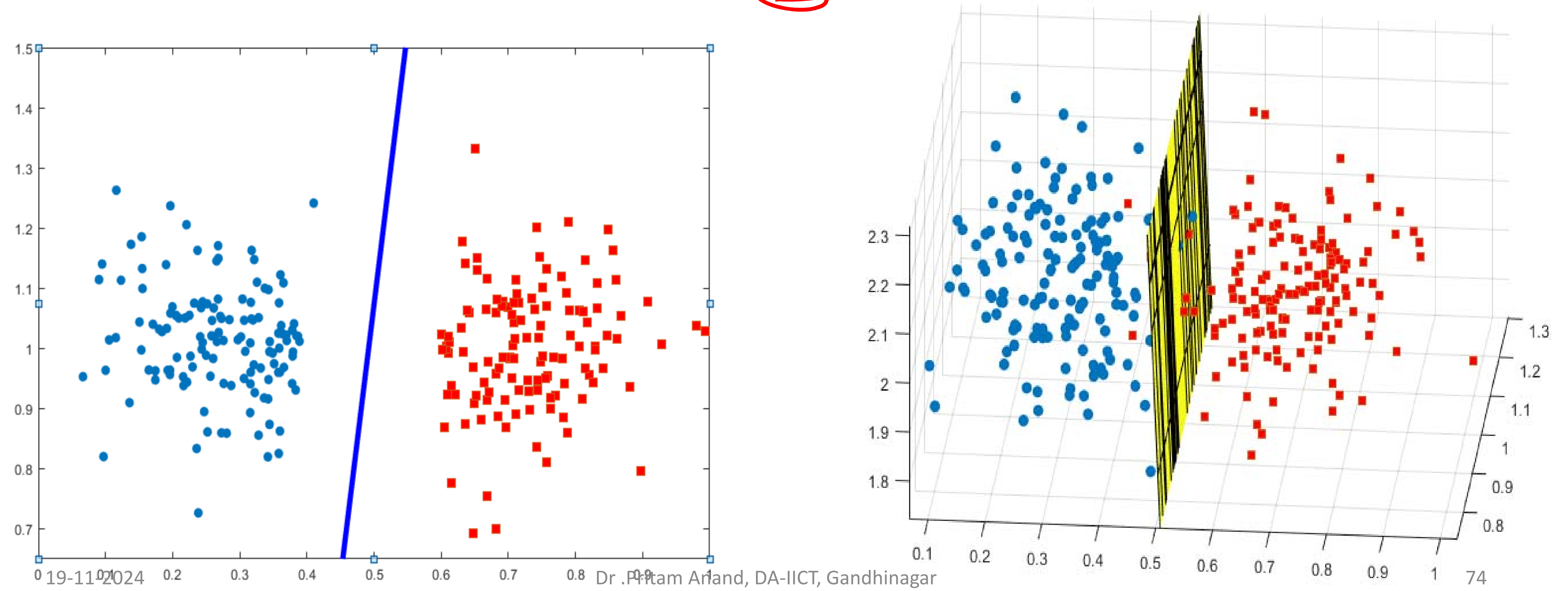
Kernel Types.

- Linear Kernel: $k(x,y) = \Phi(x)^T\Phi(y) = (x^Ty)$



Kernel Types.

- Linear Kernel: $k(x,y) = \Phi(x)^T\Phi(y) = (x^Ty)$

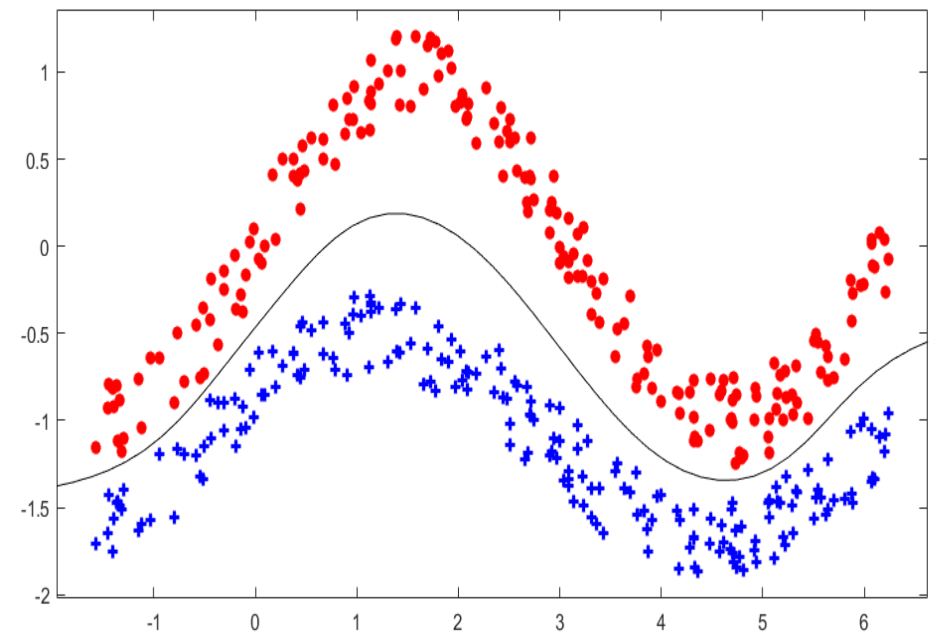


Kernel Types

- Quadratic Kernel:

$$k(x,y) = \Phi(x)^T\Phi(y) = (x^Ty+c)^2$$

,where c is the user-defined parameter.

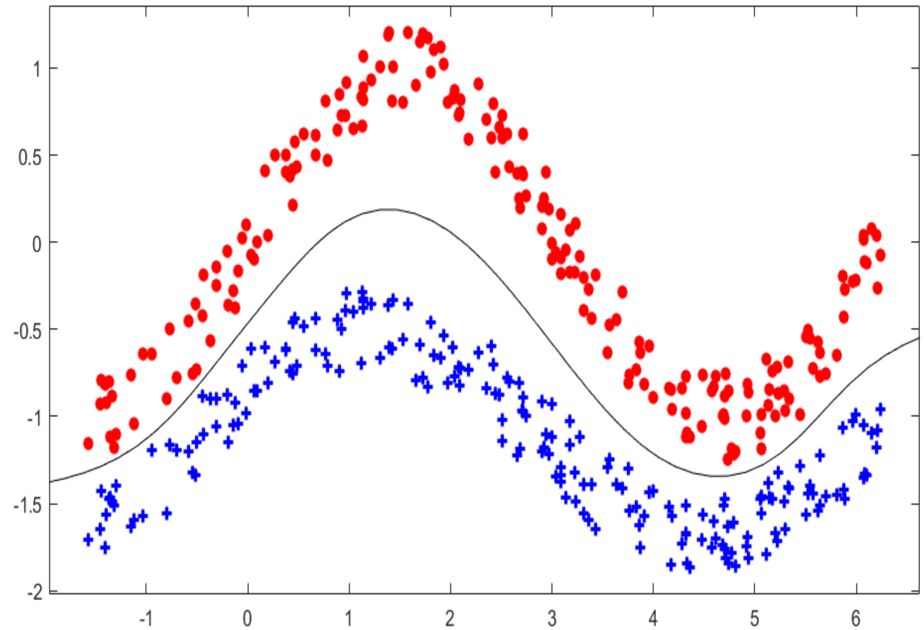


Kernel Types

- Polynomial Kernel:

$$k(x,y)= \Phi(x)^T\Phi(y)= (x^Ty+c)^p$$

,where c is the user-defined parameter.



- It can generate any type of polynomial surfaces.

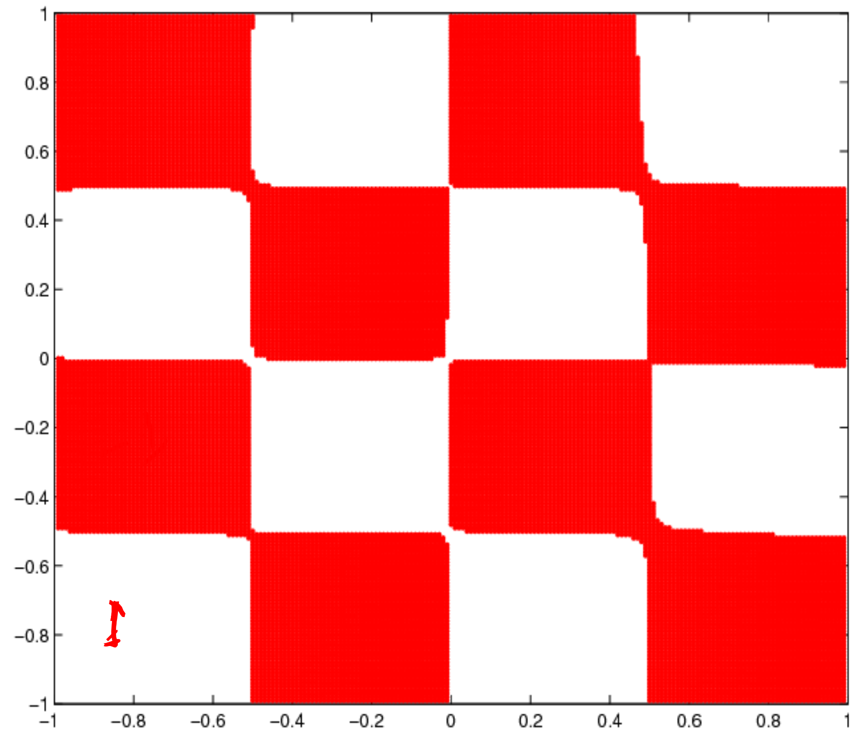
Kernel Types.

- RBF kernel:

$$k(x,y) = \Phi(x)^T\Phi(y)$$

$$= e^{-q||x-y||^2},$$

where q is the user-defined parameter.

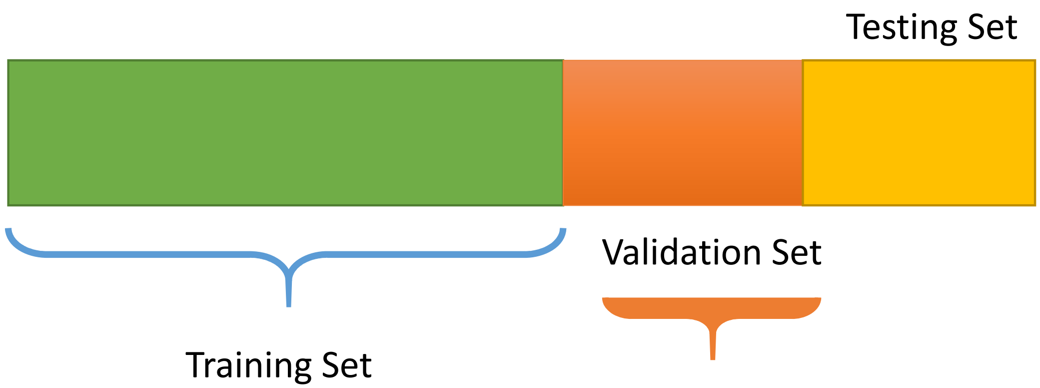


- It can generate any type of continuous surfaces.

TOY SVM

https://greitemann.dev/svm-demo

Model selection in SVM



Model selection in SVM

2^{-5}	2^{-4}	2^{-3}	2^{-2}	2^{-1}	2^{0}	2^{1}	2^{2}	2^{3}	2^{4}	2^{5}	C
2^{-5}	2^{-4}	2^{-3}	2^{-2}	2^{-1}	2^{0}	2^{1}	2^{2}	2^{3}	2^{4}	2^{5}	q

Implementation

- A number of libraries in MATLAB and Python is available.
- The *fitcsvm* in MATLAB provides an efficient implementation.
- The LIBSVM library was once popular for SVM . Link-https://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html
- You can also code the SVM from scratch. You need to solve the dual problem ,which is a QPP. You can use the *quadprog* function for solving the QPP. For coding the SVM from scratch, you can refer the Appendices of the https://svms.org/tutorials/Gunn1998.pdf

Popular Tutorial

- Gunn, Steve R. "Support vector machines for classification and regression." *ISIS technical report* 14.1 (1998): 5-16.
- Burges, Christopher JC. "A tutorial on support vector machines for pattern recognition." *Data mining and knowledge discovery* 2.2 (1998): 121-167.

Popular Books

- Deng, Naiyang, Yingjie Tian, and Chunhua Zhang. *Support vector machines: optimization based theory, algorithms, and extensions*. CRC press, 2012. (Optimization Prespective)
- Scholkopf, Bernhard, and Alexander J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2018.

Questions ??

Thanks