Mathematical Programming Applications in Machine Learning

E V N Sai Bharadwaja
Machine Learning and Computing
IIST

Q. What is the goal of any Machine Learning algorithm?

- Training Data
- Testing Data

Three basic basic problems that are often Studied in Machine Learning are :

- Regression
- Classification
- Clustering

There are so many approaches to solve above problem

- But Mathematical Programming has now become very popular in Machine Learning Community and Mathematical Programming Community

Q. What is the Advantage of Studying Mathematical Programming?

- Most of the Machine Learning Models result into Linear Programming , Quadratic Programming , Convex Programming.

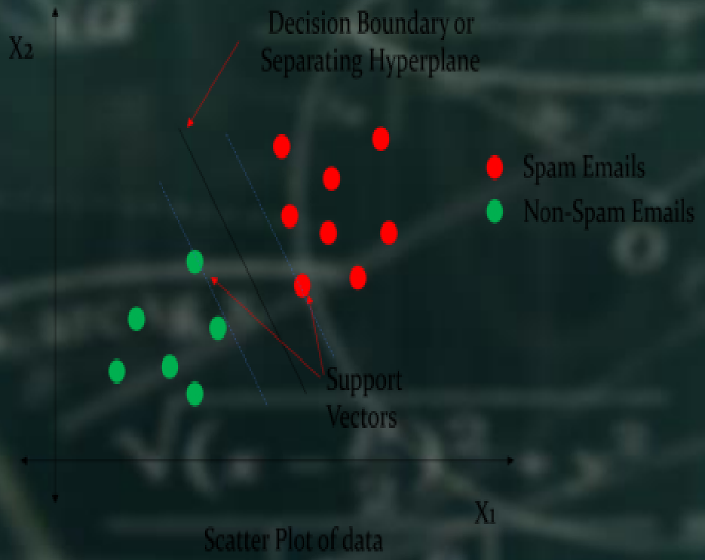
- But Mathematical Programming has
  - Theoretical results(KKT conditions and Duality Theory)
  - Efficient algorithms

Machine Learning Techniques

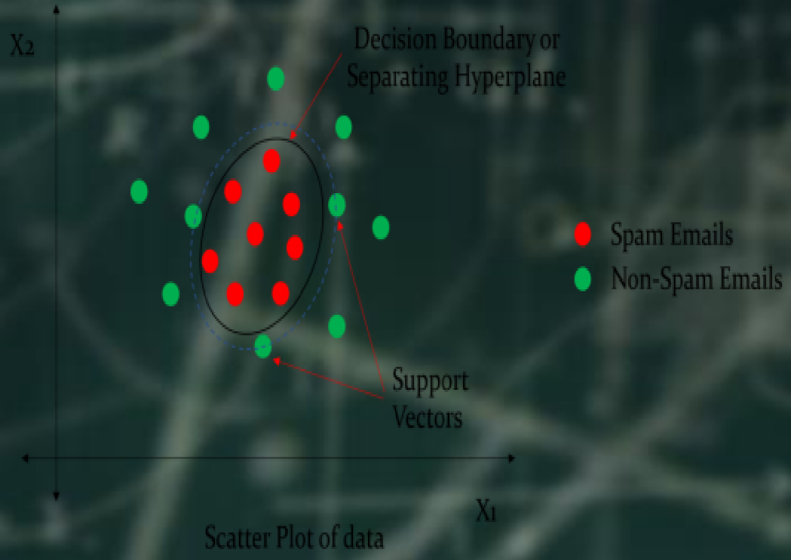
- Supervised Learning (Ex : Classification and Regression)
- Unsupervised Learning (Ex : Clustering)

Binary (Two-Class) Pattern Classification Problems

- $X \in R^n$
- m - patterns $\in$ Class +1
- k - patterns $\in$ Class -1



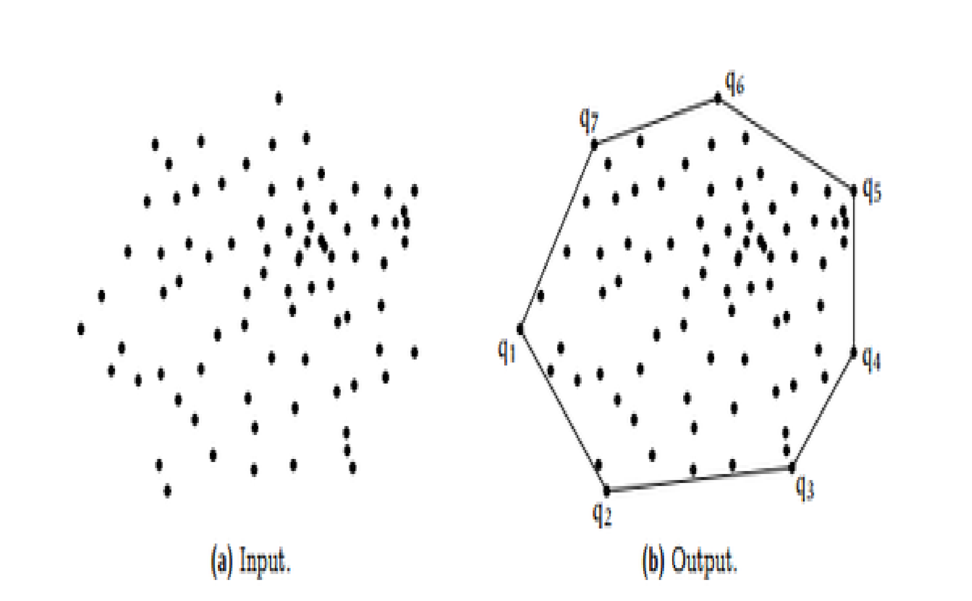
Linear Decision Boundary

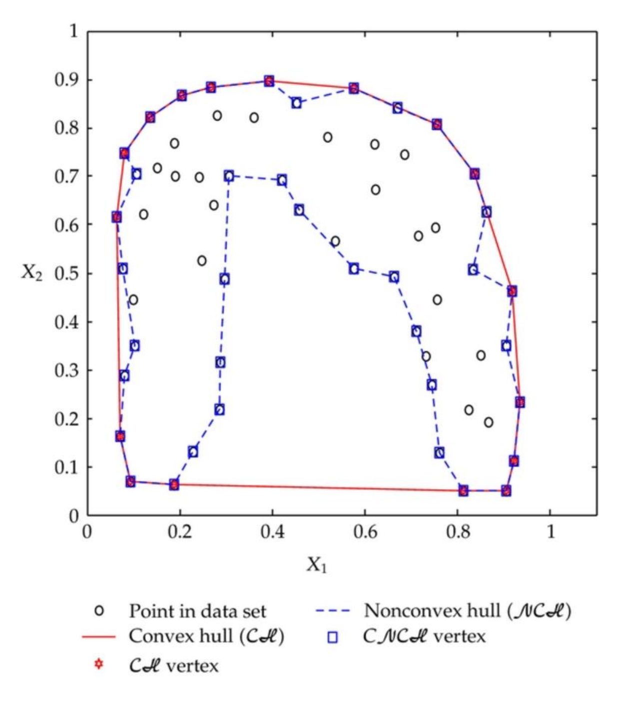


Non- Linear Decision Boundary

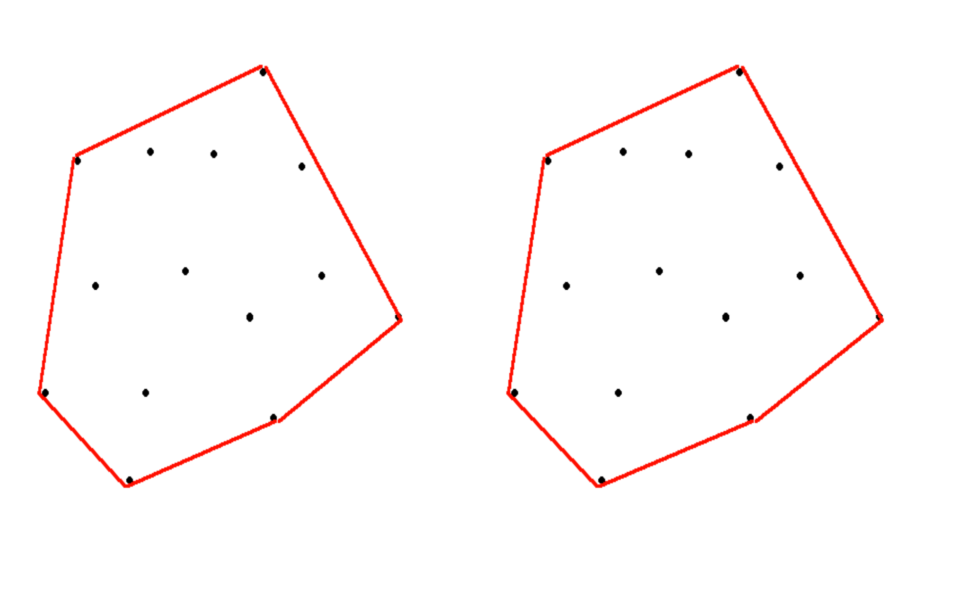
Convex Hull :

The Convex hull of a set of points X in euclidean space is the smallest convex set containing X.

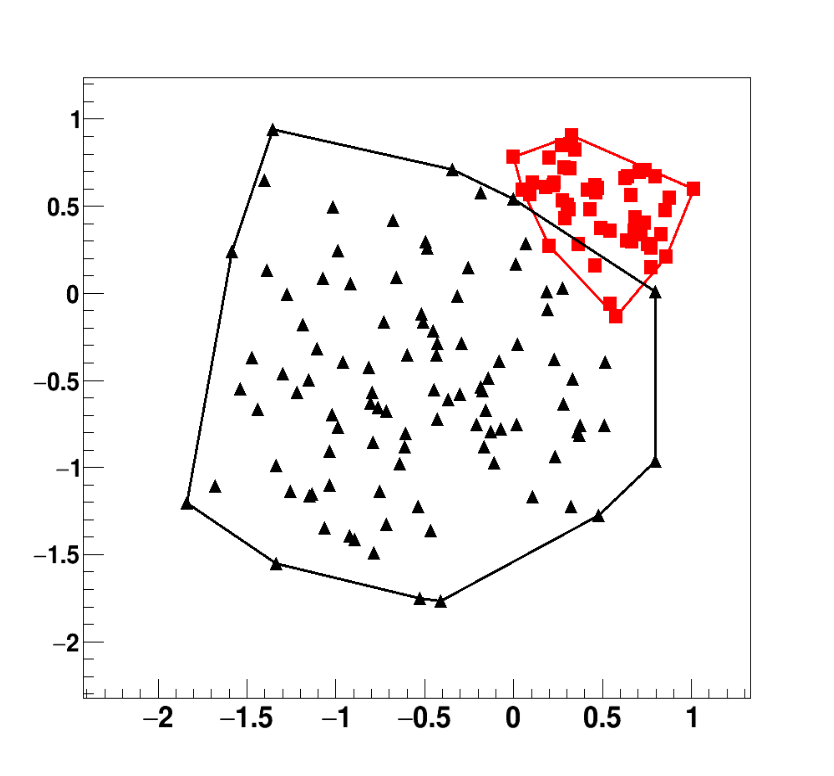




Disjoint Convex hulls
Or Linearly Separable



Joint Convex hulls
Or Linearly non Separable

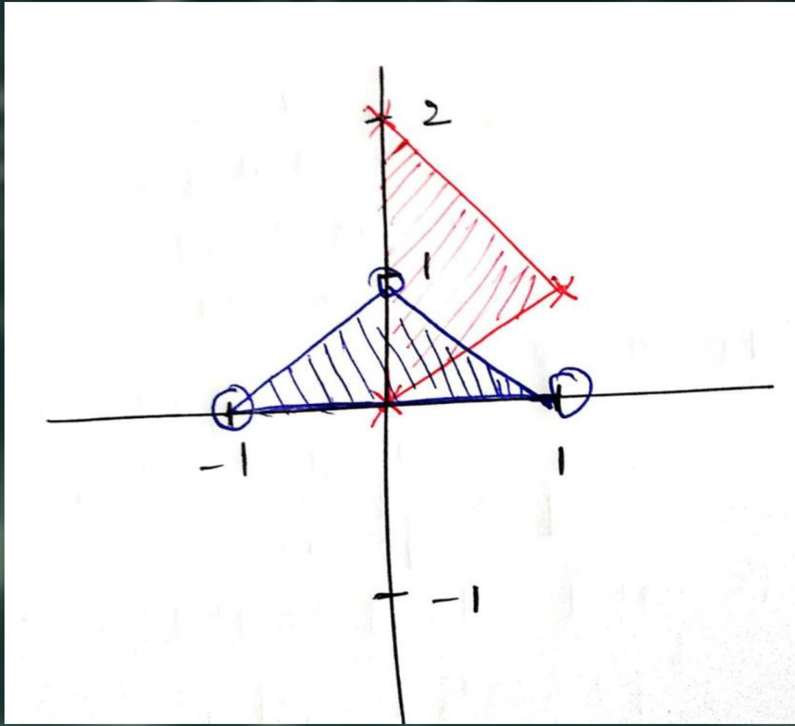


Q. A = { (-1,0),(0,1),(1,0) } ,
B = { (0,0),(1,1),(0,2) }
Are A and B linearly Separable ?

Ans :

- Convex Hulls are NOT Disjoint Hence given Problem is NOT Linearly Separable.

- But Finding convex hulls is not an easy task



So we make use of Linear Programming to decide whether the problem is Linearly Separable or not

Solving the above problem using Linear Programming

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & & \\ x_{N1} & x_{N2} & \cdots & x_{Nn} \end{bmatrix}_{N\times n}$$

m patterns Belongs to class 1.
K patterns Belongs to class -1

→ From the given Dataset x. points those belong to class 1

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & & a_{mn} \end{bmatrix}_{m\times n}$$

$$\sum_{j=1}^{n} a_{ij} w_j \geq b \quad \forall \; i = 1, 2 \cdots m$$

$$\sum_{j=1}^{n} a_{ij} w_j \geq b + \alpha \quad \forall \; i = 1, 2 \cdots m$$

$$\sum_{j=1}^{n} a_{ij} w_j \geq b + \alpha^*$$

where $\alpha^* = \min[\alpha_i]$

$$\boxed{Aw \geq eb + e}$$

→ From the given Data set x. points belongs to class -1

$$B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & & & \\ b_{k1} & b_{k2} & \cdots & b_{kn} \end{bmatrix}$$

$$\sum_{j=1}^{n} b_{rj} w_j \leq b \quad \forall \; r = 1, 2 \cdots k$$

$$\sum_{j=1}^{n} b_{rj} w_j \leq b - \beta_r \quad \forall \; r = 1, 2 \cdots k$$

$$\sum_{j=1}^{n} b_{rj} w_j \leq b - \beta^*$$

where $\beta^* = \min[\beta_r]$

$$\boxed{Bw \leq eb - e}$$

Error Minimizing Linear Programming Problem

→ For a Real number $a_+ = \max(a,0)$.

→ For $x \in R^n$ $\quad x_+ = \{(x_+)_1, (x_+)_2 \cdots (x_+)_n\}^T \in R^n$

→ L-1 norm $\quad \|x\|_1 = \sum_{i=1}^{n} |x_i|$

So the optimization problem.

$$\min_{(w,b)} \frac{1}{m}\|(-Aw+eb+e)_+\|_1 + \frac{1}{k}\|(Bw-eb+e)_+\|_1$$

If A and B are Linearly Separable then Error is zero and Optimal value for the Optimization problem is zero and the converse is also True.

Converse Statement :

If Optimal Value for the optimization problem is zero then A and B are Linearly Separable

otherwise NOT Linearly Separable

Lemma : Transforms into LPP

Lemma: Consider the problems

i] $$\min_{x \in S} \|g(x)_+\|_1 + \|h(x)_+\|_1$$

ii] $$\min_{x \in S} \left\{ e^T y + e^T z : \begin{array}{l} y \geq g(x),\ y \geq 0 \\ z \geq h(x),\ z \geq 0 \end{array} \right\}$$

Where $S \subset R^n$, $g : S \to R^m$, $h : S \to R^k$, $y \in R^m$ and $z \in R^k$.

→ Then both ~~pro~~ problem [i] and [ii] have identical solution sets.

Applying Lemma to our Optimization Problem

Our optimization problem is

$$\min_{(w,b)} \frac{1}{m} \|(-Aw + eb + e)_+\|_1 + \frac{1}{k} \cdot \|(Bw - eb + e)_+\|_1$$

After Applying Lemma.

$$\min_{w,b,y,z} \frac{e^T y}{m} + \frac{e^T z}{k}$$

$$\text{Sub to: } \begin{array}{l} Aw - eb + y \geq e \\ -Bw + eb + z \geq e \\ y, z \geq 0 \end{array}$$

w and b are unrestricted.

LPP.

LPP can be Efficiently Solved by using **Simplex Algorithm**.

Hence we can find whether A and B are Linearly separable or Not

Conclusion :

- A and B are linearly separable iff optimal value for the error minimizing LPP is ZERO . Hence there exists a Hyperplane w.T *x = b which is Linear Separator

- A and B are NOT linearly separable then optimal value for the error minimizing LPP is NOT ZERO . Hence there exists NO Hyperplane w.T *x = b which is Linear Separator

Q. Write error minimizing LPP for AND problem and check if the given Problem is Linearly Separable ?

AND:

x_1	x_2	o/p	Bipolar o/p
0	0	0	−1
0	1	0	−1
1	0	0	−1
1	1	1	1

Class +1

$A = [1 \quad 1]$

$m = 1$

Class −1:

$B = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$

$K = 3$

Optimization problem:

$$\text{Min} \quad y_1 + \frac{1}{3}[z_1 + z_2 + z_3]$$

Sub to:

$$1 \cdot w_1 + 1 \cdot w_2 - b + y_1 \geq 1$$

$$-0 \cdot w_1 - 0 \cdot w_2 + b + z_1 \geq 1$$

$$-0 \cdot w_1 - 1 \cdot w_2 + b + z_2 \geq 1$$

$$-1 \cdot w_1 - 0 \cdot w_2 + b + z_3 \geq 1$$

$$y_1, z_1, z_2, z_3 \geq 0$$

w_1, w_2, b unrestricted.

Solution:

An optimal solution for the LPP

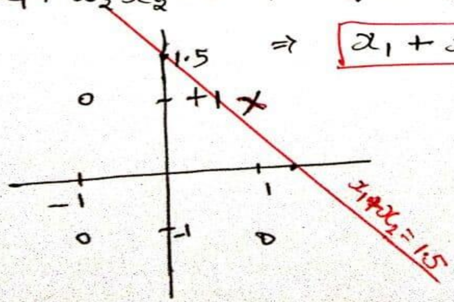
$w_1 = 2 \quad w_2 = 2 \quad b = 3$

$y_1 = 0 \quad z_1 = 0 \quad z_2 = 0$
$z_3 = 0$

→ The optimal value of the error minimizing LPP is zero, given AND problem is Linearly separable. with Hyperplane.

$$w_1 x_1 + w_2 x_2 = b \Rightarrow 2x_1 + 2x_2 = 3$$

$$\Rightarrow \boxed{x_1 + x_2 = 1.5}$$



Q. Write error minimizing LPP for XOR problem and check if the given Problem is Linearly Separable ?

XOR problem:

x_1	x_2	o/p	Bipolar o/p
0	0	0	-1
0	1	1	1
1	0	1	1
1	1	0	-1

Class +1   Class -1

$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ $B = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}$

$m = 2$ $k = 2$

Optimization problem:

$$\text{Min} \quad \frac{1}{2}[y_1 + y_2] + \frac{1}{2}[z_1 + z_2]$$

Sub to:

$$0 \cdot w_1 + 1 \cdot w_2 - b + y_1 \geq 1$$
$$1 \cdot w_1 + 0 \cdot w_2 - b + y_2 \geq 1$$
$$-0 \cdot w_1 - 0 w_2 + b + z_1 \geq 1$$
$$-1 \cdot w_1 - 1 \cdot w_2 + b + z_2 \geq 1$$
$$y_1, y_2, z_1, z_2 \geq 0$$

w_1, w_2, b are unrestricted.

The solution for LPP.

$w_1 = 2 \quad w_2 = 2 \quad b = 1$

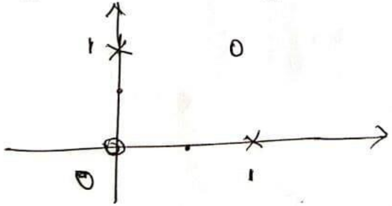
$y_1 = 0 \quad y_2 = 0$

$z_1 = 0 \quad z_2 = 4$

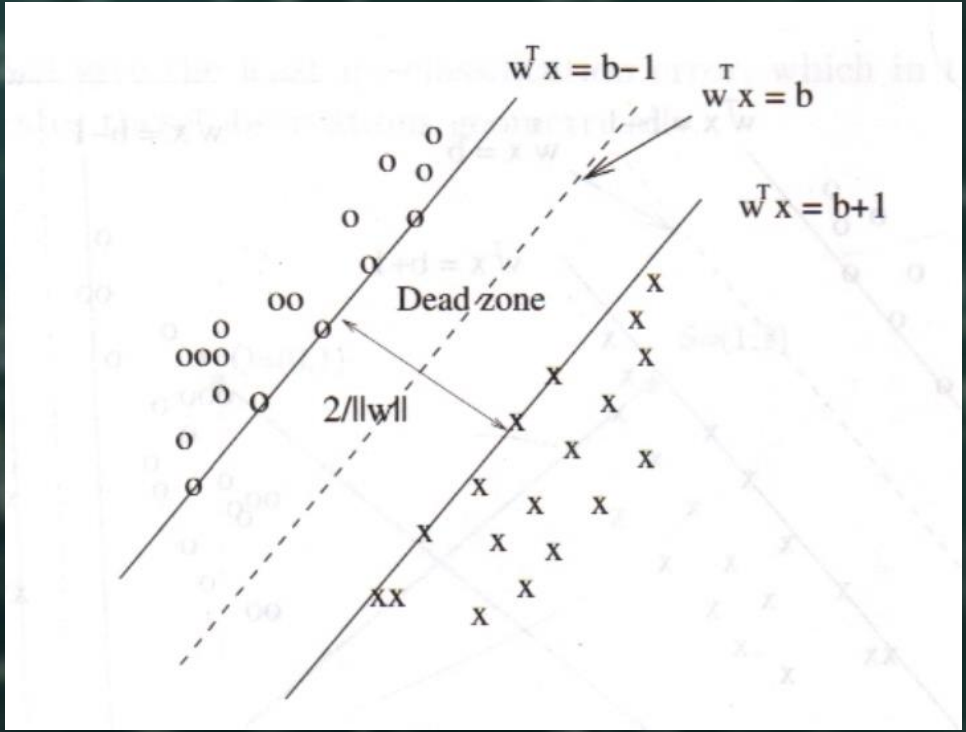
The optimal value for error minimizing LPP is 2.

optimal value $\neq 0$. [Hence There exits No hyperplane]

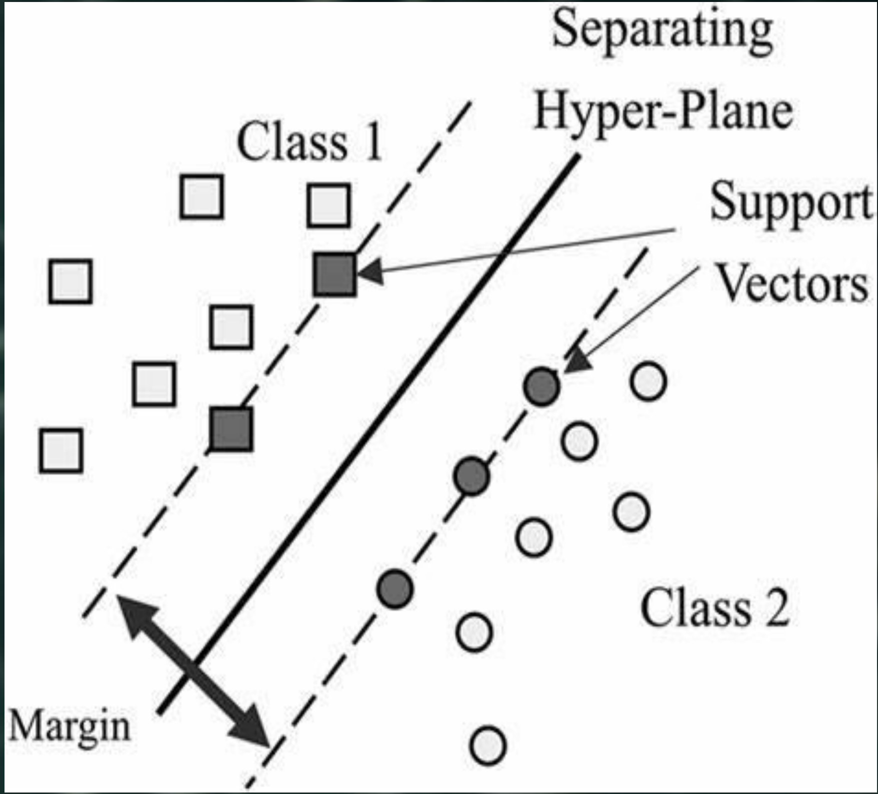
$2x_1 + 2x_2 = 1 \Rightarrow x_1 + x_2 = 0.5$



- If A and B are linearly separable then there exists infinitely many hyperplanes
- So how should we choose a Hyperplane ?
- Is there any optimal Hyperplane ?



$w^T x = b-1$
$w^T x = b$
$w^T x = b+1$
Dead zone
2/||w||



Separating
Hyper-Plane
Class 1
Support
Vectors
Class 2
Margin

Canonical Hyperplane

Dead Zone

→ **Canonical Hyperplane:**

Let $A, B \subset \mathbb{R}^n$ be linearly separable.

Then the seperating hyperplane $\boxed{w^T \cdot x = b}$ is called canonical hyperplane if it satisfies

$$\boxed{\begin{aligned} Aw &\geq eb + e \\ Bw &\leq eb - e \end{aligned}}$$

→ **Dead zone:**

Let $A, B \subset \mathbb{R}^n$ be linearly separable and $w^T x = b$ be a canonical separating hyperplane. Then there exists a Region $\boxed{\{x : (b-1) < w^T x < (b+1)\}} \subset \mathbb{R}^n$ surrounding the seperating hyperplane $w^T x = b$, which is $\boxed{\text{void of points}}$ sets A and B.

This Region is called Dead zone for the separating Hyperplane $w^T x = b$.

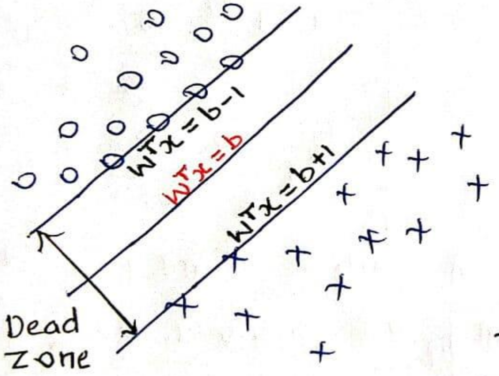
Margin

Margin:

Let $A, B \subset R^n$ be linearly separable and $w^T x = b$ be canonical seperating hypersplane.

Let $w^T x = b-1$ and $w^T x = b+1$ be the bounding hyperplanes which define the dead zone.

The $\boxed{\text{distance between}}$ the bounds $\boxed{w^T x = b-1}$ and $\boxed{w^T x = b+1}$ is called the margin.



→ The perpendicular distance between $w^Tx=b$ and $w^Tx=b-1$ is $\frac{1}{\|w\|}$

→ So the perpendicular distance between $w^Tx=b-1$ and $w^Tx=b+1$ is

$$\boxed{\text{Margin} = \frac{2}{\|w\|}}$$

where $\|w\| = \sqrt{w_1^2 + \cdots w_n^2}$

[*The distance of a point* (x', y') *to the plane* $ax + by = 0$ *is given by* $\frac{|ax' + by'|}{\sqrt{a^2 + b^2}}$]

The distance of a point $x_i = (x_{i1}, x_{i2}, \ldots x_{in}\}$ to the hyperplane $w^T x + b = 0$ is given by

$$\begin{aligned} dist(x_i, \tilde{f}(x)) &= \frac{|w_1 x_{i1} + w_2 x_{i2} + \cdots + w_n x_{in} + b|}{\sqrt{w_1^2 + w_2^2 + \ldots w_n^2}} \\ &= \frac{|w^T x_i + b|}{||w||} \end{aligned}$$

$\boxed{w^T x + b \geq 1}$ for all x in the positive class. Therefore $\boxed{dist(x, \tilde{f}(x)) \geq \frac{1}{||w||}}$, for all x in the positive class. The points in the positive class, that lie closest to the separating hyperplane $w^T x + b = 0$ lies in the hyperplane

$$\boxed{H_1 : w^T x + b = 1}$$

Therefore $\boxed{d+ = \frac{1}{||w||}.}$

In the negative class, for all points $\boxed{w^T x + b \leq -1.}$ Therefore $\boxed{dist(x, \tilde{f}(x)) \geq \frac{1}{||w||}}$ for al x in the negative class. The points in the negative class, that lie closest to the separating hyperplane $w^T x + b = 0$ lies in the hyperplane

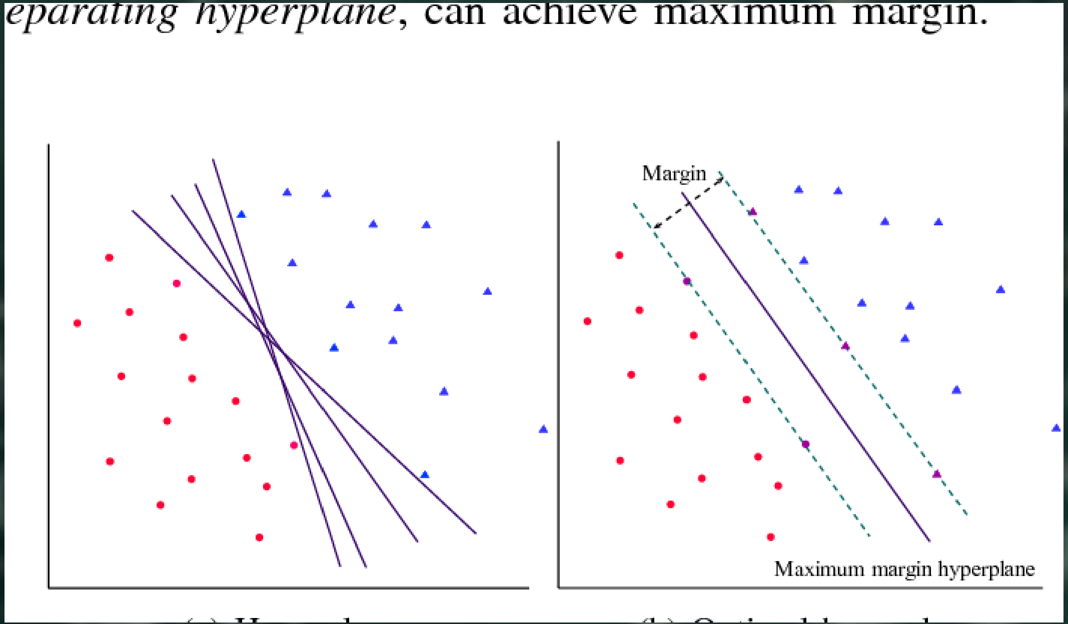
$$\boxed{H_2 : w^T x + b = -1}$$

Therefore $\boxed{d^- = \frac{1}{||w||}}.$

Hence $\boxed{\text{margin } \gamma = d^+ + d^- = \frac{2}{||w||}}$

Optimal Separating Hyperplanes:

- Let sets A , B be linearly separable and w.T * x = b be a canonical hyperplane .
- Then w.T * x = b is called Optimal Separating Hyperplane if it’s Margin is Maximum among all the Canonical Hyperplanes



- In Classification , our aim is not only to classify the available to classify the training data in accordance with the class labels
- But also to have good generalization on Unseen Data or Test Data
- The Larger the dead Zone , the smaller the misclassification Error will be.
- Therefore we would like to Maximize the dead Zone , which implies we would like to Maximize the Margin
- From the above Point of View , for the linearly Separable patterns , our aim is not only just finding the Hyperplane , But also for which the Margin is Maximum
- Ques : Will Error Minimizing LPP gives Optimal Separating Hyperplane ?
- So we should construct a Optimization problem in which Objective Function should contain Margin
- This Discussion Leads us to Hard Margin Classifier and Support Vector Machines

Hard Margin Classifier :

For the given patterns which are Linearly separable

→ Let $\{(x^i, y_i), i = 1, 2 \cdots p\}$ be a set of finite Training samples (or) patterns where $x^{(i)} \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$.

→ Here $y_i \in \{-1, 1\}$ represents the class labels of the patterns $x^{(i)}$, $(i = 1, 2 \cdots p)$.

→ Let the given patterns be Linearly separable. Then there exists a $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$. $\forall i = 1, 2 \cdots p$

$$w^T x^{(i)} - b > 0 \quad \forall i \text{ having } y = 1$$

$$w^T x^{(i)} - b < 0 \quad \forall i \text{ having } y = -1$$

By suitable scaling we can write above as

$$w^T x^{(i)} - b \geq 1 \quad \forall i \text{ having } y = 1$$

$$w^T x^{(i)} - b \leq -1 \quad \forall i \text{ having } y = -1$$

writting in more compact form.

$$y_i\left(w^T x^{(i)} - b\right) \geq 1 \quad \forall\ i = 1, 2 \cdots p.$$

→ Amongst all separating hyperplanes $w^T x = b$ we have to choose the one for which

margin $\frac{2}{\|w\|}$ is maximum, (or)

Equivalently $\frac{\|w\|}{2}$ is minimum.

→ However minimizing $\frac{\|w\|}{2}$ is equivalent to minimizing $\frac{\|w\|^2}{2}$ because $\|w\|$ is monotonically increasing for $w > 0$

∴ optimization problem:

$$\min_{(w, b)} \ \frac{1}{2} w^T w$$

$$\text{sub to: } \ y_i\left(w^T x^{(i)} - b\right) \geq 1 \quad \forall\ i = 1, 2 \cdots p.$$

→ The above problem is in the form of Standard Quadratic Programming Problem [QPP]. Which can be solved by standard QPP algorithm.

→ Once $\boxed{\text{optimal solution of } (w, b)}$ is known, The maximum margin classifier will be found

$\boxed{w^T x = b}$ → Hard Margin classifier

Difficulties in Solving Quadratic Programming Problem

- QPP has as many constraints as the number of patterns and hence extremely hard to solve for Large Data sets
- This QPP will make use of Kernel Methods, which again is difficult to Store kernel matrices.
- In Machine Learning there are special algorithms so that Not all Constraints are included at once
  - Chunking algorithm
  - Decomposition algorithm
    - SMO algorithm (Special Case of Decomposition)
- The above Techniques are also called as Active Set or Working Set approach

- In Mathematical Programming , Standard approach to handle Large number of constraints is to examine if Dual problem can be solved efficiently

- This analysis requires KKT conditions

- KKT conditions

min 1/2 || w || **2
Sub : yi(<w, xi> + b) >= 1

$$\boxed{L(w, b, \alpha) = \frac{1}{2}\langle w, w\rangle - \sum_{i=1}^{N} \alpha_i [y_i(\langle w, x_i\rangle + b) - 1]}$$

where $\alpha_i \geq 0$ are the Lagrange multipliers.
The corresponding dual is found by differentiating with respe
to w and b, imposing stationarity,

$$\boxed{\frac{\partial L(w, b, \alpha)}{\partial w} = w - \sum_{i=1}^{N} y_i \alpha_i x_i = 0}$$

$$\boxed{\frac{\partial L(w, b, \alpha)}{\partial b} = \sum_{i=1}^{N} y_i \alpha_i = 0}$$

For αi != 0
W exist

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

$$f(x) = \langle w, x \rangle = \sum_{i=1}^{N} y_i \alpha_i \boxed{\langle x_i, x \rangle}$$

Support Vectors:

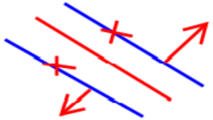
- Only Those points for which Xi is influencing Weights 'W'
- KKT multipliers (Alpha) > 0
- So vector W is know once KKT multipliers are know.

KKT Complimentary Condition

Using Karush-Kuhn-Tucker complementarity conditions,

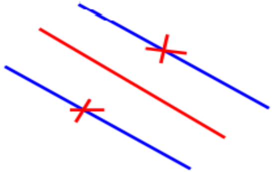
$$\alpha_i^* [y_i(\langle w, x_i \rangle + b) - 1] = 0, i = 1, 2, \dots N$$

Therefore two cases



α >= 0

- $\alpha_i^* = 0$
  - $y_i(\langle w, x_i \rangle + b) - 1 \geq 0$
  - x_i lies either above H_1 or on H_1 or above H_2 or on H_2
- $\alpha_i^* \neq 0$
  - $y_i(\langle w, x_i \rangle + b) - 1 = 0$
  - $\boxed{x_i}$ lies either on H_1 or H_2



Support Vectors :
α != 0 ,
α Contributes in weight W
Hence those Xi for α ! = 0 are called Support Vectors

- The SVM algorithm clearly needs the knowledge of KKT multipliers
- Once KKT multipliers are known Support Vectors

    - Support Vectors:

        - Only Those points for which Xi is influencing Weights ‘W’
        - KKT multipliers (Alpha) > 0
        - So vector W is know once KKT multipliers are know.

- Hence we get to know vector (w) and b
- From the concept of Non Linear Programming Duality , KKT multipliers are nothing but Dual Variables.
- Therefore we go for Wolfe Dual of the Problem to find KKT multipliers

$$
\begin{aligned}
\boxed{\|w\|^2} &= \langle w, w\rangle \\
&= \langle \sum_i \alpha_i y_i x_i, \sum_i \alpha_i y_i x_i \rangle
\end{aligned}
$$

$$
\boxed{\textbf{|| W || **2} = \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle}
$$

$$
\begin{aligned}
\boxed{\sum_{i=1}^{N} \alpha_i [y_i(\langle w, x_i \rangle)]} &= \sum_{i=1}^{N} \alpha_i [y_i(\langle \sum_{j=1}^{N} y_j \alpha_j x_j, x_i \rangle)] \\
&= \boxed{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle}
\end{aligned}
$$

$\boxed{L(w,b,\alpha) = \frac{1}{2}\langle w,w\rangle - \sum_{i=1}^{N} \alpha_i[y_i(\langle w,x_i\rangle + b) - 1]}$

$\boxed{\textbf{|| W || **2} = \sum_i \sum_j \alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle}$

Dual Formulation

$\boxed{b\sum_i \alpha_i y_i = 0}$

$$\boxed{\sum_{i=1}^{N} \alpha_i[y_i(\langle w,x_i\rangle]} = \sum_{i=1}^{N} \alpha_i[y_i(\langle \sum_{j=1}^{N} y_j\alpha_j x_j, x_i\rangle]$$

$$= \boxed{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle}$$

Sub (8) and (9) into (5) to form the dual,

$$
\begin{aligned}
W(\alpha) &= \frac{1}{2}\sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j \langle x_i, x_j\rangle - \sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j \langle x_i, x_j\rangle + \sum_{i=1}^{N} \alpha_i \\
&= \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j \langle x_i, x_j\rangle
\end{aligned}
$$

Wolfe Dual :

- If we clearly observe dual problem , we can notice that there is only one constraint
- The objective Function of the Dual problem is Concave
- Therefore Dual problem is Easier to Solve Than Primal problem
- Once optimal alpha is known [w] and b can be computed.
- Hence separating hyperplane with Maximum Margin can be determined

Primal Problem : Number of Constraints = Number of Data Points
Dual Problem : Number of Constraints = 1 [Easy to Solve]

Hence the dual optimisation problem is

$$\text{maximise } W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle$$

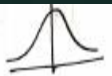
Only One Constraint Easy to Solve

$$\frac{\partial L(w,b,\alpha)}{\partial b} = \sum_{i=1}^{N} y_i \alpha_i = 0 \;\longrightarrow\; \text{subject to} \sum_{i=1}^{N} y_i \alpha_i = 0$$

$$\alpha_i \geq 0, i = 1, 2, \ldots N$$

$L(w, \alpha, b)$ and $W(\alpha)$ arise from the same objective function but with different constraints and the solution is found by minimizing $L(w, \alpha, b)$ or by maximizing $W(\alpha)$.

1. Gaussian function $\phi(x) = e^{-\frac{x^2}{2\sigma^2}}$



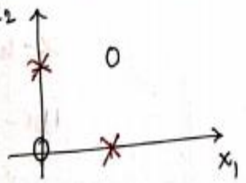
2. Multiquadratic functions $\phi(x) = \sqrt{[x^2+c^2]}$ $c>0$

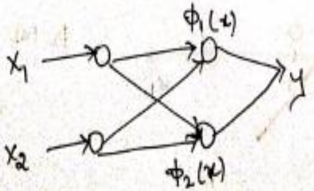
3. Inverse multi quadratic fn $\phi(x) = \dfrac{1}{\sqrt{x^2+c^2}}$ $c>0$.

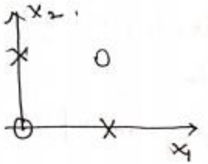
$$\phi_i(x) = e^{-\frac{1}{2\sigma^2}\|x-c_i\|^2} \qquad i=1,2,\dots m.$$

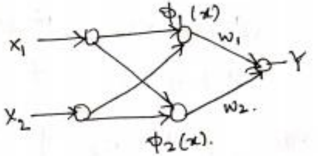
Ex: XOR-Problem. [linearly non seperable]

x_1	x_2	op
0	0	0
0	1	1
1	0	1
1	1	0









$\phi_1(x) = e^{-\|x-c_1\|^2}$, $2\sigma^2=1$

$\phi_2(x) = e^{-\|x-c_2\|^2}$, $2\sigma^2=1$

where c_1 & c_2 are the center of the patter clusters.

let $c_1 = \begin{bmatrix}0\\0\end{bmatrix}$ $c_2 = \begin{bmatrix}1\\1\end{bmatrix}$ $x = \begin{bmatrix}x_1\\x_2\end{bmatrix}$

these centres can be found out in clusturing Techniques.

x_1	x_2	$\phi_1(x)$	$\phi_2(x)$	Y
0	0	1	0.1	0
0	1	0.4	0.4	1
1	0	0.4	0.4	1
1	1	0.1	1	0

$x - c_1 = \begin{bmatrix}0\\0\end{bmatrix} - \begin{bmatrix}0\\0\end{bmatrix}$

$\|x-c_1\| = \|\begin{bmatrix}0\\0\end{bmatrix}\| = 0$

$\phi_1(x) = e^{-0} = 1$

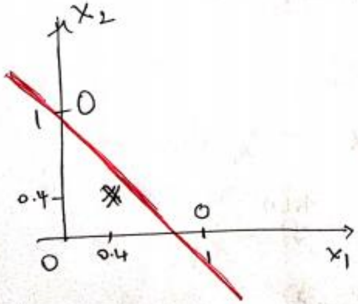
$x - c_2 = \begin{bmatrix}0\\0\end{bmatrix} - \begin{bmatrix}1\\1\end{bmatrix}$

$\|x-c_2\| = \|\begin{bmatrix}-1\\-1\end{bmatrix}\|$

$= \sqrt{2}$

$\phi_2(x) = e^{-(\sqrt{2})^2}$

$= e^{-2}$.



Soft Margin Classifier :

For the given patterns which are NOT Linearly separable

→ let $\{(x^{(i)}, y_i),\ i = 1, 2 \ldots p\}$ be a finite Training sample of patterns where $x^{(i)} \in R^n$ and $y_i \in \{-1, 1\}$.

→ We consider the case where the patterns are NOT linearly separable

→ In this case Error minimizing problem will have a non-zero objective function value.

→ In the optimal solution not all the errors will be zero

→ Let the $\boxed{\text{error variables}}$ will be denoted by $\boxed{\xi_i}\ (i = 1, 2 \ldots p)$.

→ So our aim is to $\boxed{\text{maximize margin } \frac{2}{\|w\|}}$ and $\boxed{\sum_{i=1}^{p} \xi_i \text{ is least}}$.

→ $\therefore$ $\boxed{\text{minimize}}$ both $\boxed{\frac{w^T w}{2}}$ and $\boxed{\sum_{i=1}^{p} \xi_i}$

This will be $\boxed{\text{multi objective optimization}}$

→ Since this is not possible, we consider weighted sum $\boxed{\min \frac{w^T w}{2} + c \cdot \sum_{i=1}^{p} \xi_i}$

Optimisation Problem

$$\min_{f\in\mathcal{F}, b\in\mathbb{R}} \frac{1}{2}||f||^2 + C\sum_{i=1}^{N}\xi_i$$

subject to

$$y_i(\boxed{\langle f, k_{x_i}\rangle} + b) - 1 + \xi_i \geq 0, \; i = 1, \ldots, N$$

$$\xi_i \geq 0.$$

- The objective function is quadratic and all the constrain linear

The primal Lagrangian is,

$$L(f, b, \xi, \alpha) = \frac{||f||^2}{2} + C\sum_{i=1}^{N}\xi_i - \sum_{i}\alpha_i[y_i(\langle f, k_{x_i}\rangle + b) - 1 + \xi_i] - \sum_{i}\mu_i\xi_i$$

$$\frac{\partial L}{\partial f} = f - \sum_{i=1}^{N} \alpha_i y_i k_{x_i} = 0$$

$$\boxed{f = \sum_{i=1}^{N} \alpha_i y_i k_{x_i}}$$

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

For i=1,2,. . . , N,

$$\boxed{\begin{gathered} \frac{\partial L(w, b, \alpha)}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \\ \alpha_i \geq 0 \\ \mu_i \geq 0 \end{gathered}}$$

The KKT complimentary conditions are,

$$\boxed{\begin{aligned}\alpha_i[y_i(\langle f, k_{x_i}\rangle + b) - 1 + \xi_i] &= 0 \\ \mu_i \xi_i &= 0\end{aligned}}$$

$$C - \alpha_i = \mu_i \geq 0$$

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

Three cases

- $\alpha_i = 0$ $\quad y_i(\langle f, k_{x_i}\rangle + b) \geq 1$.
- $0 < \alpha_i < C$ $\quad y_i(\langle f, k_{x_i}\rangle + b) = 1$
- $\alpha_i = C$ $\quad y_i(\langle f, k_{x_i}\rangle + b) \leq 1$.

Support vectors are those points for which $y_i(\langle f, k_{x_i} \rangle + b) \leq 1$.

- $y_i(\langle f, k_{x_i} \rangle + b) = 1$, k_{x_i} on H_1 or on H_2
- $0 \leq y_i(\langle f, k_{x_i} \rangle + b) < 1$ Between H_1 and H or H_2 and H or on H
- $y_i(\langle f, k_{x_i} \rangle + b) < 0$ Incorrect classification

Dual Optimisation

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

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

$$C \geq \alpha_i \geq 0, i = 1, 2, \ldots N.$$

$$\boxed{f = \sum_{i=1}^{\boxed{sv}} y_i \alpha_i^* k_{x_i}}$$

where sv is the number of support vectors. Then

$$\begin{aligned}
\boxed{\tilde{f}(x)} &= \langle f, k_x \rangle + b \\
&= \sum_{i=1}^{sv} y_i \alpha_i^* \langle k_{x_i}, k_x \rangle + b \\
&= \boxed{\sum_{i=1}^{sv} y_i \alpha_i^* k(x_i, x) + b}
\end{aligned}$$

Thank You