Support Vector Machines

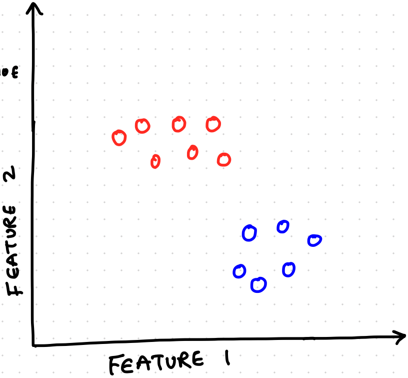
Nipun Batra

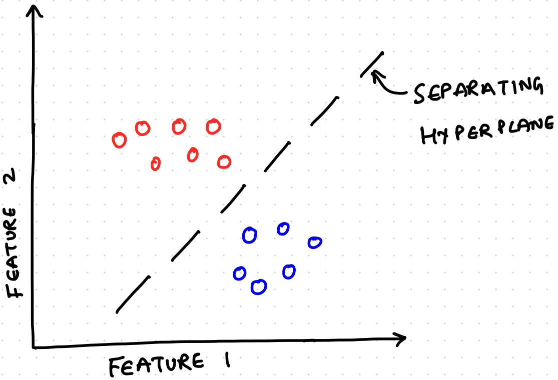
April 23, 2023

IIT Gandhinagar

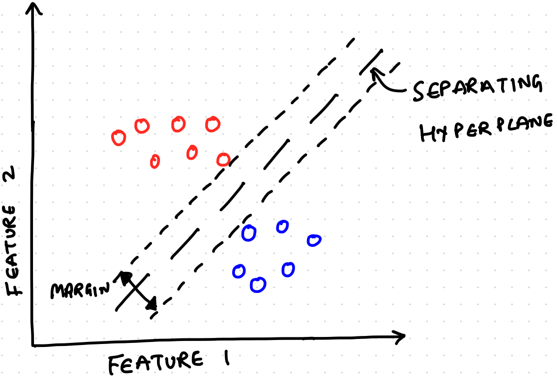
SUPPORT VECTOR MACHINES

POPULAR BINARY CLASSIFICATION TECHNIQUE

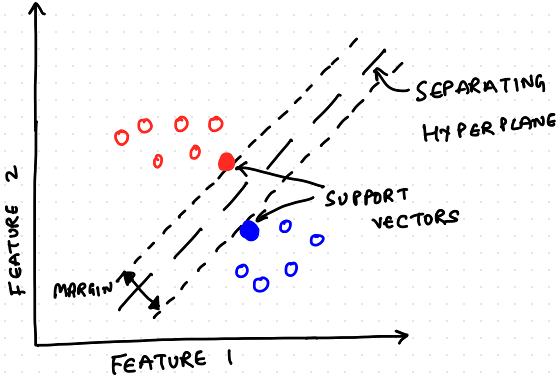




IDEA: DRAW A SEPARATING HYPERPLANE



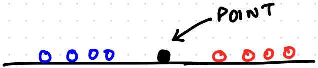
IDEA: MAXIMIZE THE MARGIN



SUPPORT VECTORS: POINTS ON BOUNDARY | MARGIN

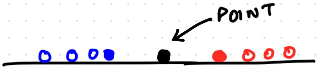
HYPERPLANE V/S # DIMENSIONS

1D

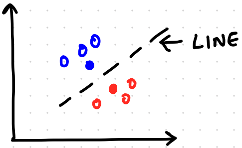


HYPERPLANE V/S # DIMENSIONS

1D

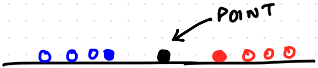


2D

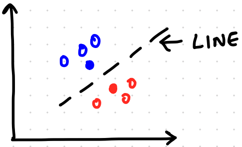


HYPERPLANE V/S # DIMENSIONS

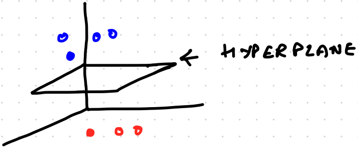
1D



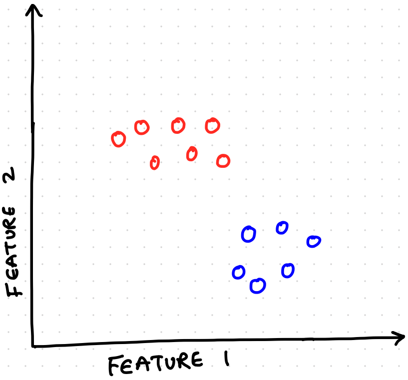
2D



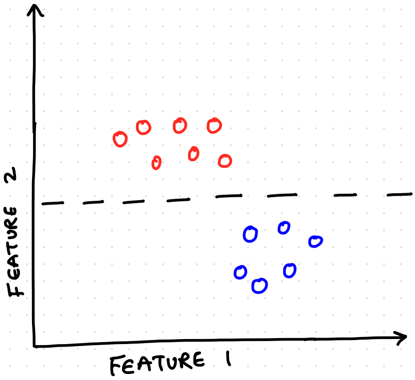
3D
(AND MORE)



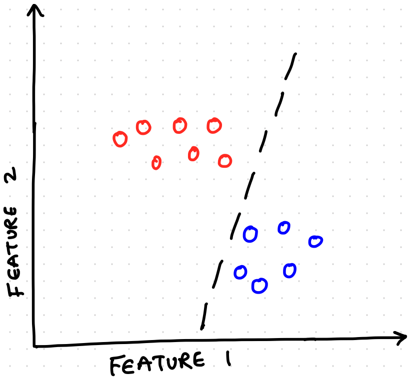
WHICH HYPERPLANE?



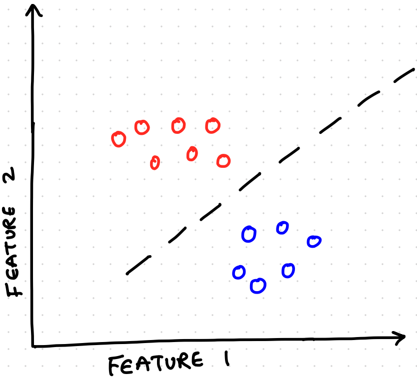
WHICH HYPERPLANE?



WHICH HYPERPLANE?

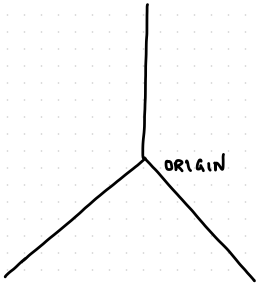


WHICH HYPERPLANE?

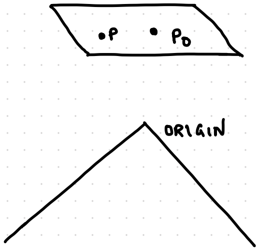


EQUATION OF HYPERPLANE

HOW TO DEFINE?



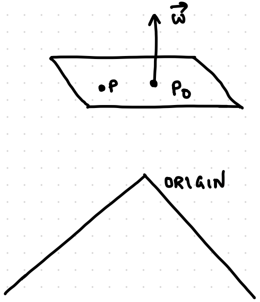
EQUATION OF HYPERPLANE



P : Any point on plane

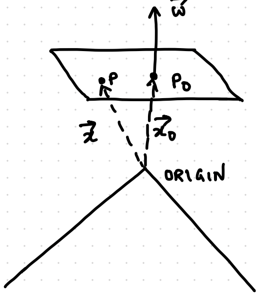
P_0 : One point on plane

EQUATION OF HYPERPLANE



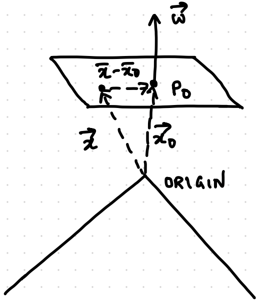
$\vec{w}$: $\perp$ vector to plane at P_0

EQUATION OF HYPERPLANE



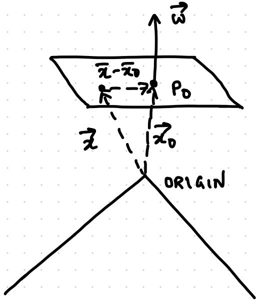
P and P_0 lie on plane

EQUATION OF HYPERPLANE



$\overrightarrow{P P_0} = \vec{x} - \vec{x}_0$ lies on plane

EQUATION OF HYPERPLANE



$\overrightarrow{P P_0} = \vec{x} - \vec{x}_0$ lies on plane

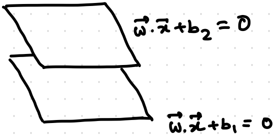
$\Rightarrow \vec{w} \perp (\vec{x} - \vec{x}_0)$

or, $\vec{w} \cdot (\vec{x} - \vec{x}_0) = 0$

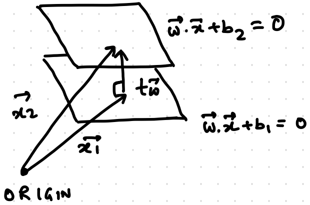
or, $\vec{w} \cdot \vec{x} - \vec{w} \cdot \vec{x}_0 = 0$

or, $\boxed{\vec{w} \cdot \vec{x} + b = 0}$

DISTANCE B/W || HYPERPLANES



DISTANCE B/W || HYPERPLANES



Distance between 2 parallel hyperplanes

Equation of two planes is:

$$\vec{w} \cdot \vec{x} + b_1 = 0$$

$$\vec{w} \cdot \vec{x} + b_2 = 0$$

Distance between 2 parallel hyperplanes

Equation of two planes is:

$$\vec{w} \cdot \vec{x} + b_1 = 0$$

$$\vec{w} \cdot \vec{x} + b_2 = 0$$

For a point $\vec{x_1}$ on plane 1 and $\vec{x_2}$ on plane 2, we have:

Distance between 2 parallel hyperplanes

Equation of two planes is:

$$\vec{w} \cdot \vec{x} + b_1 = 0$$

$$\vec{w} \cdot \vec{x} + b_2 = 0$$

For a point $\vec{x_1}$ on plane 1 and $\vec{x_2}$ on plane 2, we have:

$$\overrightarrow{x_2} = \overrightarrow{x_1} + t\vec{w}$$
$$D = |t\vec{w}| = |t|||\vec{w}||$$

Distance between 2 parallel hyperplanes

Equation of two planes is:

$$\vec{w} \cdot \vec{x} + b_1 = 0$$

$$\vec{w} \cdot \vec{x} + b_2 = 0$$

For a point $\vec{x_1}$ on plane 1 and $\vec{x_2}$ on plane 2, we have:

$$\overrightarrow{x_2} = \overrightarrow{x_1} + t\vec{w}$$
$$D = |t\vec{w}| = |t|||\vec{w}||$$

We can rewrite as follows:

Distance between 2 parallel hyperplanes

Equation of two planes is:

$$\vec{w} \cdot \vec{x} + b_1 = 0$$

$$\vec{w} \cdot \vec{x} + b_2 = 0$$

For a point $\vec{x_1}$ on plane 1 and $\vec{x_2}$ on plane 2, we have:

$$\overrightarrow{x_2} = \overrightarrow{x_1} + t\vec{w}$$
$$D = |t\vec{w}| = |t|||\vec{w}||$$

We can rewrite as follows:

$$\vec{w} \cdot \vec{x}_2 + b_2 = 0$$
$$\Rightarrow \vec{w} \cdot (\vec{x}_1 + t\vec{w}) + b_2 = 0$$

Distance between 2 parallel hyperplanes

Equation of two planes is:

$$\vec{w} \cdot \vec{x} + b_1 = 0$$

$$\vec{w} \cdot \vec{x} + b_2 = 0$$

For a point $\vec{x_1}$ on plane 1 and $\vec{x_2}$ on plane 2, we have:

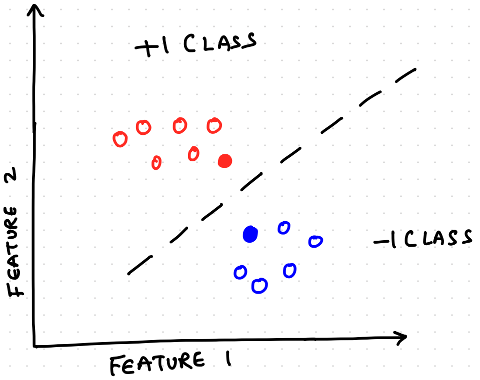
$$\overrightarrow{x_2} = \overrightarrow{x_1} + t\vec{w}$$
$$D = |t\vec{w}| = |t|||\vec{w}||$$

We can rewrite as follows:

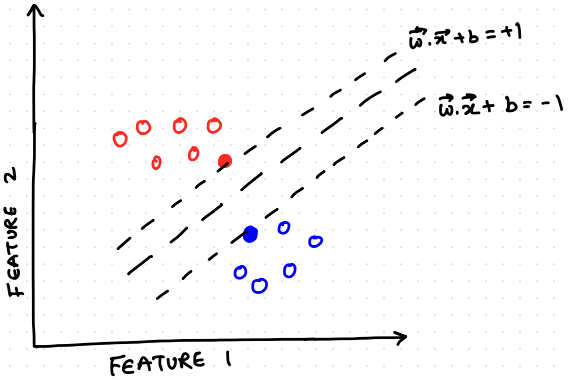
$$\vec{w} \cdot \vec{x}_2 + b_2 = 0$$
$$\Rightarrow \vec{w} \cdot (\vec{x}_1 + t\vec{w}) + b_2 = 0$$

$$\Rightarrow \vec{w} \cdot \vec{x}_1 + t\|\vec{w}\|^2 + b_1 - b_1 + b_2 = 0 \Rightarrow t = \frac{b_1 - b_2}{\|\vec{w}\|^2} \Rightarrow D = t\|\vec{w}\| = \frac{b_1 - b_2}{\|\vec{w}\|}$$

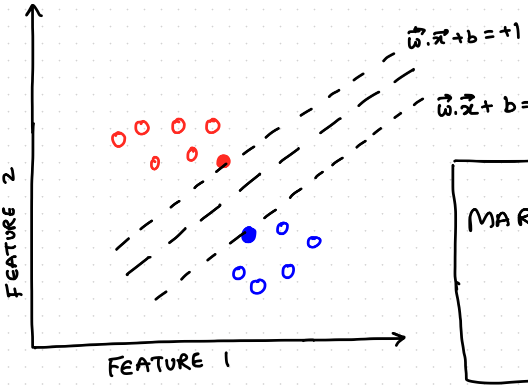
FORMULATION



FORMULATION



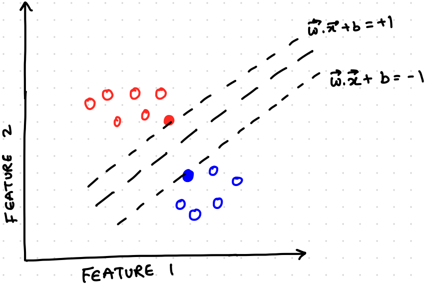
FORMULATION



$$\text{MARGIN} = \frac{(b+1)-(b-1)}{\|\vec{w}\|}$$

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

FORMULATION



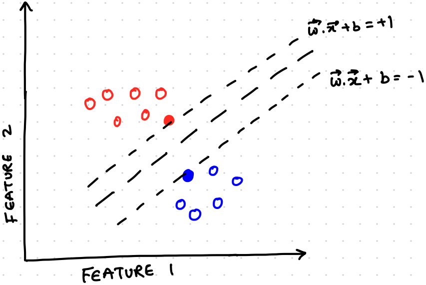
GOAL: MAXIMIZE MARGIN

$\Rightarrow$ MAXIMIZE $\dfrac{2}{||\vec{w}||}$

$\Rightarrow$ MINIMIZE $||\vec{w}||$

S.T. Correctly label points

FORMULATION



GOAL: MAXIMIZE MARGIN

$\Rightarrow$ MAXIMIZE $\dfrac{2}{||\vec{w}||}$

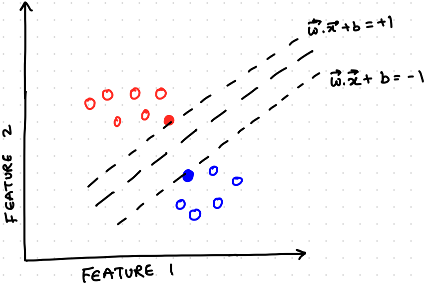
$\Rightarrow$ MINIMIZE $||\vec{w}||$

S.T. Correctly label points

i.e. if $y_i = -1$
$\vec{w}.\vec{x}+b \le -1$

if $y_i = +1$
$\vec{w}.\vec{x}+b \ge 1$

FORMULATION



GOAL: MAXIMIZE MARGIN

$\Rightarrow$ MAXIMIZE $\dfrac{2}{||\vec{w}||}$

$\Rightarrow$ $\boxed{\text{MINIMIZE } ||\vec{w}||}$

S.T. Correctly label points

i.e. if $y_i = -1$

$\vec{w}.\vec{x}+b \leq -1$

if $y_i = +1$

$\vec{w}.\vec{x}+b \geq 1$

$$\boxed{y_i(\vec{w}.\vec{x}+b) \geq 1}$$

Primal Formulation

Objective

$$\text{Minimize } \frac{1}{2}||w||^2$$
$$\text{s.t. } y_i(w.x_i + b) \geq 1 \ \ \forall i$$

Primal Formulation

Objective

$$\text{Minimize } \frac{1}{2}||w||^2$$

$$\text{s.t. } y_i(w.x_i + b) \geq 1 \ \ \forall i$$

Q) What is $||w||$?

Primal Formulation

Objective

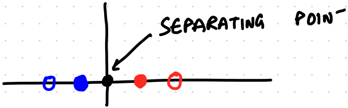
$$\text{Minimize } \frac{1}{2}||w||^2$$
$$\text{s.t. } y_i(w.x_i + b) \geq 1 \;\; \forall i$$

Q) What is $||w||$?

$$w = \begin{bmatrix} w_1 \\ w_2 \\ ... \\ w_n \end{bmatrix}$$

$$||w|| = \sqrt{w^T w}$$
$$= \sqrt{\begin{bmatrix} w_1, w_2, ... w_n \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ ... \\ w_n \end{bmatrix}}$$

EXAMPLE (IN 1D)



Simple Exercise

$$\begin{bmatrix} x & y \\ 1 & 1 \\ 2 & 1 \\ -1 & -1 \\ -2 & -1 \end{bmatrix}$$

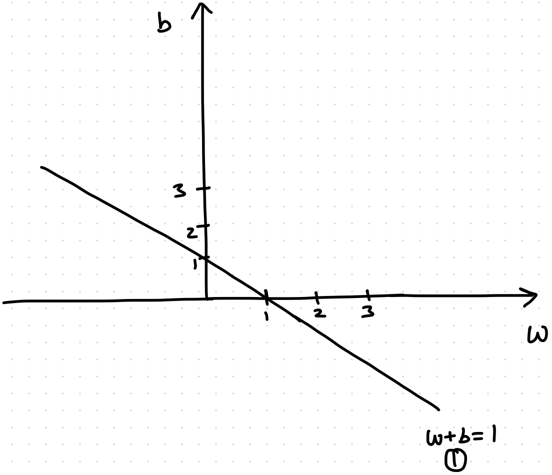
Separating Hyperplane: $wx + b = 0$

Simple Exercise

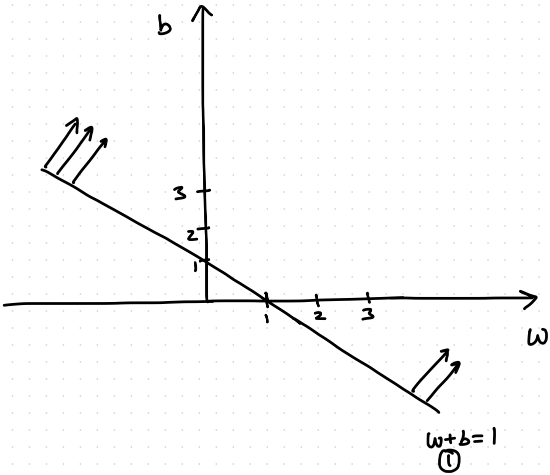
$$y_i(w_i x_i + b) \geq 1$$

$$\begin{bmatrix} x_1 & y \\ 1 & 1 \\ 2 & 1 \\ -1 & -1 \\ -2 & -1 \end{bmatrix}$$

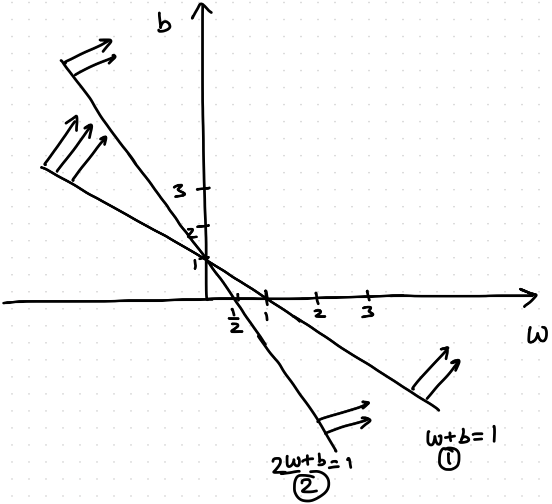
$$\Rightarrow y_i(w_i x_i + b) \geq 1$$
$$\Rightarrow 1(w_1 + b) \geq 1$$
$$\Rightarrow 1(2w_1 + b) \geq 1$$
$$\Rightarrow -1(-w_1 + b) \geq 1$$
$$\Rightarrow -1(-2w_1 + b) \geq 1$$

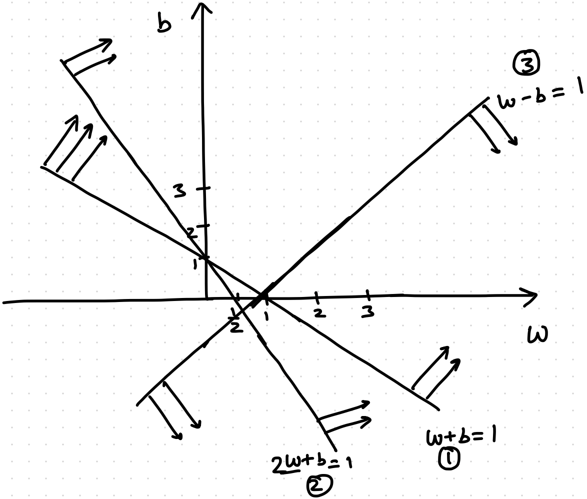


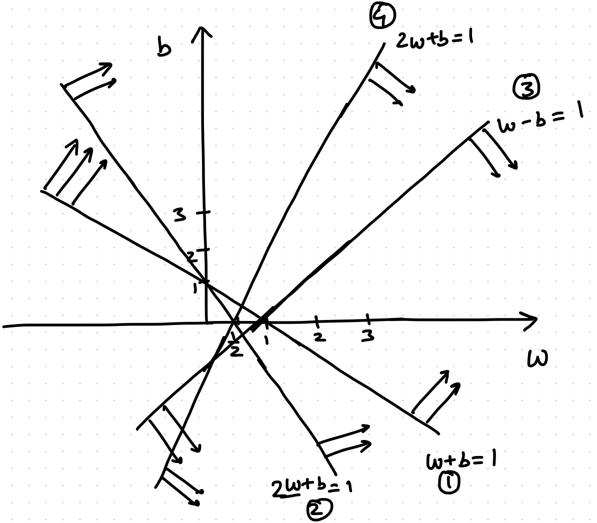
b
3
2
1
1
2
3
w
w+b=1
1



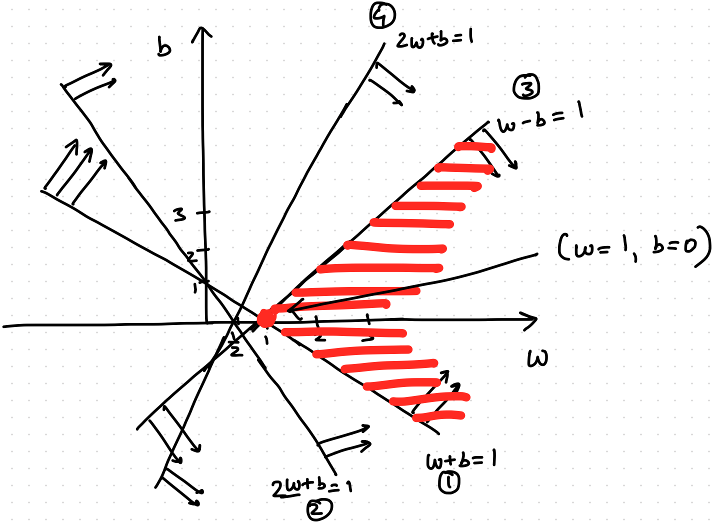
b
3
2
1
1
2
3
w
w+b=1
1


b
3
2
1
1/2
1
2
3
w
w+b=1
1
2w+b=1
2


b
w
3
2
1
1/2
1
2
3
③
w - b = 1
w+b=1
①
2w+b=1
②



b
w
1
2
3
$\frac{1}{2}$
1
2
3
④ $2w+b=1$
③ $w-b=1$
$w+b=1$ ①
$2w+b=1$ ②


b
w
1
2
3
1
2
3
$\frac{1}{2}$
1
④
$2w+b=1$
③
$w-b=1$
$(w=1, b=0)$
$w+b=1$
①
$2w+b=1$
②

Simple Exercise

$$w_{min} = 1, b = 0$$
$$w.x + b = 0$$
$$x = 0$$

Simple Exercise

Minimum values satisfying constraints $\Rightarrow w = 1$ and $b = 0$
$\therefore$ Max margin classifier $\Rightarrow x = 0$

Primal Formulation is a Quadratic Program

Generally;

$\Rightarrow$ Minimize Quadratic(x)

$\Rightarrow$ such that, Linear(x)

Question

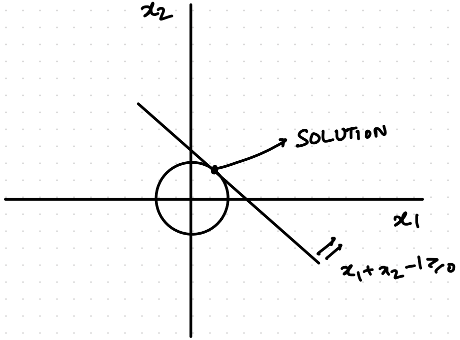
$$x = (x_1, x_2)$$

$$\text{minimize} \quad \frac{1}{2}||x||^2$$

$$: x_1 + x_2 - 1 \geq 0$$

MINIMIZE QUADRATIC

S.t. LINEAR



Converting to Dual Problem

Primal $\Rightarrow$ Dual Conversion using Lagrangian multipliers

$$\text{Minimize } \frac{1}{2}||\bar{w}||^2$$
$$\text{s.t. } y_i(\bar{w}.x_i + b) \geq 1$$
$$\forall i$$

$$L(\bar{w}, b, \alpha_1, \alpha_2, ...\alpha_n) = \frac{1}{2}\sum_{i=1}^{d} w_i^2 - \sum_{i=1}^{N} \alpha_i(y_i(\bar{w}.\bar{x_i} + b) - 1) \ \forall \ \alpha_i \geq 0$$

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

Converting to Dual Problem

$$\frac{\partial L}{\partial w} = 0 \Rightarrow \bar{w} - \sum_{i=1}^{n} \alpha_i y_i \bar{x_i} = 0$$

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

$$L(\bar{w}, b, \alpha_1, \alpha_2, ...\alpha_n) = \frac{1}{2} \sum_{i=1}^{d} w_i^2 - \sum_{i=1}^{N} \alpha_i (y_i(\bar{w}.\bar{x_i} + b) - 1$$

$$= \frac{1}{2} ||\bar{w}||^2 - \sum_{i=1}^{N} \alpha_i y_i \vec{w}_. \bar{x_i} - \sum_{i=1}^{N} \alpha_i y_i b + \sum_{i=1}^{N} \alpha_i$$

$$= \sum_{=1}^{N} \alpha_i + \frac{\left(\sum_i \alpha_i y_i \bar{x_i}\right) \left(\sum_j \alpha_j y_j \bar{x_j}\right)}{2} - \sum_i \alpha_i y_i \left(\sum_j \alpha_j y_j \bar{x_j} \right) \bar{x_i}$$

Converting to Dual Problem

$$L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \bar{x}_i \cdot \bar{x}_j$$

$$\begin{array}{ll} \text{Minimize } \|\bar{w}\|^2 \Rightarrow & \text{Maximize } L(\alpha) \\ s.t & s.t \\ y_i\left(\bar{w}, x_i + b\right) \geqslant 1 & \sum_{i=1}^{N} \alpha_i y_i = 0 \ \forall \ \alpha_i \geq 0 \end{array}$$

Question

Question:

$\alpha_i \left(y_i \left(\bar{w}, \bar{x}_i + b \right) - 1 \right) = 0 \quad \forall i$ as per KKT slackness

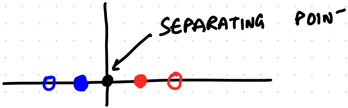
What is α_i for support vector points?

Answer: For support vectors,

$$\bar{w}.\bar{x}_i + b = -1 \text{ (+ve class)}$$
$$\bar{w}.\bar{x}_i + b = +1 \text{ (+ve class)}$$

$y_i \left(\bar{w} \cdot \bar{x}_i + b \right) - 1 \right) = 0 \quad$ for $i = \{\text{support vector points}\}$
$\therefore \alpha_i$ where i $\in \{\text{support vector points}\} \neq 0$
For all non-support vector points $\alpha_i = 0$

EXAMPLE (IN 1D)



Revisiting the Simple Example

$$\begin{bmatrix} x_1 & y \\ 1 & 1 \\ 2 & 1 \\ -1 & -1 \\ -2 & -1 \end{bmatrix}$$

$$L(\alpha) = \sum_{i=1}^{4} \alpha_i - \frac{1}{2} \sum_{i=1}^{4} \sum_{j=1}^{4} \alpha_i \alpha_j y_i y_j \bar{x_i} \bar{x_j} \qquad \alpha_i \geq 0$$

$$\sum \alpha_i y_i = 0 \qquad \alpha_i (y_i(\bar{w}.\bar{x_i} + b - 1) = 0$$

Revisiting the Simple Example

$$
\begin{aligned}
L(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = &\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 \\
&- \frac{1}{2} \{\alpha_1\alpha_1 \times (1 * 1) \times (1 * 1) \\
&\qquad + \\
&\alpha_1\alpha_2 \times (1 * 1) \times (1 * 2) \\
&\qquad + \\
&\alpha_1\alpha_3 \times (1 * -1) \times (1 * 1) \\
&\qquad ... \\
&\alpha_4\alpha_4 \times (-1 * -1) \times (-2 * -2)\}
\end{aligned}
$$

How to Solve? $\Rightarrow$ Use the QP Solver!!

Revisiting the Simple Example

For the trivial example,

We know that only x $= \pm 1$ will take part in the constraint actively.

Thus, $\alpha_2, \alpha_4 = 0$

By symmetry, $\alpha_1 = \alpha_3 = \alpha$ (say)

& $\sum y_i \alpha_i = 0$

$$
\begin{aligned}
L(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = & 2\alpha \\
& - \frac{1}{2} \{ \alpha^2 (1)(-1)(1)(-1) \\
& \quad + \alpha^2 (-1)(1)(-1)(1) \\
& \quad + \alpha^2 (1)(1)(1)(1) + \alpha^2 (-1)(-1)(-1)(-1) \\
& \}
\end{aligned}
$$

$$\underset{\alpha}{Maximize} \quad 2\alpha - \tfrac{1}{2}(4\alpha^2)$$

Revisiting the Simple Example

$$
\begin{aligned}
\frac{\partial}{\partial \alpha}\left(2\alpha - 2\alpha^2\right) = 0 \Rightarrow 2 - 4\alpha = 0 \\
\Rightarrow \ \alpha = 1/2
\end{aligned}
$$

$$
\begin{aligned}
\therefore \alpha_1 = 1/2 \ \ \alpha_2 = 0; \ \ \alpha_3 = 1/2 \ \ \alpha_4 = 0 \\
\vec{w} = \sum_{i=1}^{N} \alpha_i y_i \bar{x}_i = 1/2 \times 1 \times 1 + 0 \times 1 \times 2 \\
+1/2 \times -1 \times -1 + 0 \times -1 \times -2 \\
= 1/2 + 1/2 = 1
\end{aligned}
$$

Revisiting the Simple Example

Finding b:

For the support vectors we have,

$y_i(\vec{w} \cdot \overrightarrow{x_i} + b) - 1 = 0$

or, $y_i\ (\bar{w} \cdot \bar{x}_1 + b) = 1$

or, $y_i^2\ (\bar{w} \cdot \bar{x}_i + b) = y_i$

or, $\quad \bar{w}, \bar{x}_i + b = y_i \ \ (\because y_i^2 = 1)$

or, $b = y_i - w \cdot x_i$

In practice, $b = \frac{1}{N_{SV}} \sum_{i=1}^{N_{SV}} (y_i - \bar{w}\bar{x}_i)$

Obtaining the Solution

$$
\begin{aligned}
b &= \frac{1}{2}\{(1-(1)(1)) + (-1-(1)(-1))\\
&= \frac{1}{2}\{0+0\} = 0\\
&= 0\\
&\therefore w = 1 \;\; \& \;\; b = 0
\end{aligned}
$$

Making Predictions

Making Predictions

$$\hat{y}(x_i) = \text{SIGN}(w \cdot x_i + b)$$

For $x_{test} = 3$; $\hat{y}(3) = \text{SIGN}(1 \times 3 + 0) = +$ve class

Making Predictions

Alternatively,

$$\hat{y}\left(x_{TEST}\right) = \text{SIGN}\left(\bar{w} \cdot \bar{x}_{TEST} + b\right)$$
$$= \text{SIGN}\left(\sum_{i=1}^{N_S} \alpha_j y_j x_j \cdot x_{test} + b\right)$$

In our example,

$\alpha_1 = 1/2; \alpha_2 = 0; \quad \alpha_3 = 1/2; \alpha_4 = 0$

$$\hat{y}(3) = \text{SIGN}\left(\frac{1}{2} \times 1 \times (1 \times 3) + 0 + \frac{1}{2} \times (-1) \times (-1 \times 3) + 0\right)$$
$$= \text{SIGN}\left(\frac{6}{2}\right) = \text{SIGN}(3) = +1$$

ORIGINAL DATA
IN R

Non-Linearly Separable Data

Non-Linearly Separable Data

Data not separable in $\mathbb{R}$

Non-Linearly Separable Data

Data not separable in $\mathbb{R}$

Can we still use SVM?

Non-Linearly Separable Data

Data not separable in $\mathbb{R}$

Can we still use SVM?

Yes!

Non-Linearly Separable Data

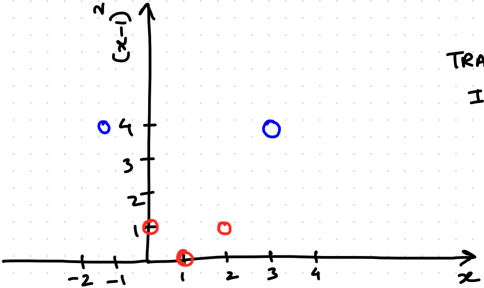
Data not separable in $\mathbb{R}$

Can we still use SVM?

Yes!

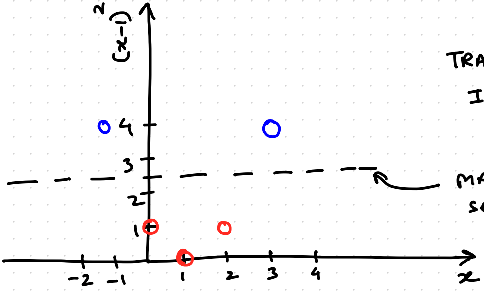
How? Project data to a higher dimensional space.

ORIGINAL DATA
IN R



TRANSFORMED DATA
IN R^2

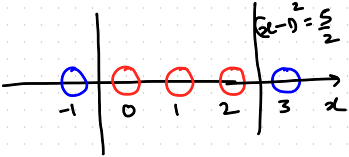
ORIGINAL DATA IN $\mathbb{R}$



TRANSFORMED DATA IN $\mathbb{R}^2$

MAX MARGIN SEPARATING HYPERPLANE

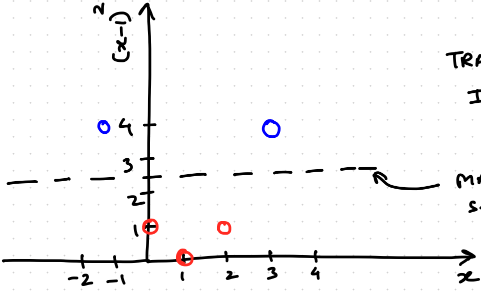
$(x-1)^2 = \frac{5}{2}$



[DECISION BOUNDARY IN R]

ORIGINAL DATA IN R

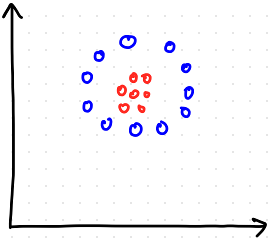
TRANSFORMED DATA IN R^2



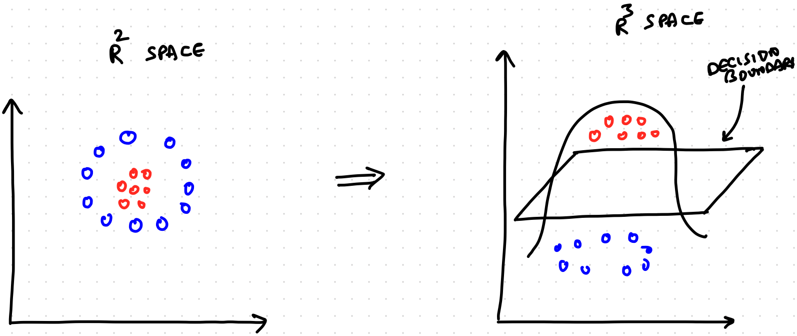
MAX MARGIN SEPARATING HYPERPLANE

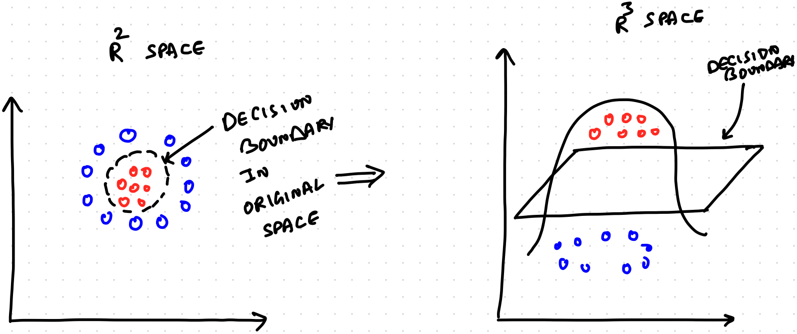
$(x-1)^2 = \frac{5}{2}$

R^2 SPACE



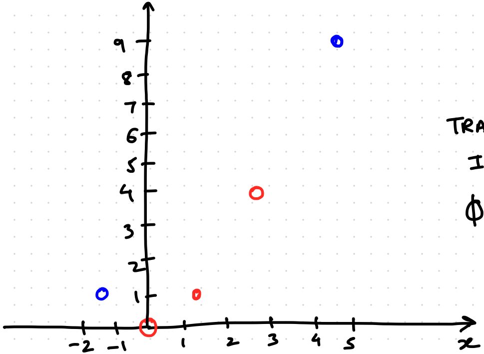
$\Longrightarrow$


R² SPACE
R³ SPACE
DECISION BOUNDARY



R^2 SPACE
DECISION BOUNDARY IN ORIGINAL SPACE
R^3 SPACE
DECISION BOUNDARY

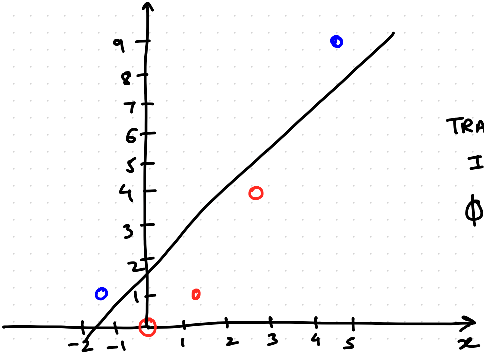
ORIGINAL DATA IN R



TRANSFORMED DATA IN R^2

$$\phi(x) = \begin{bmatrix} \sqrt{2}\, x \\ x^2 \end{bmatrix}$$

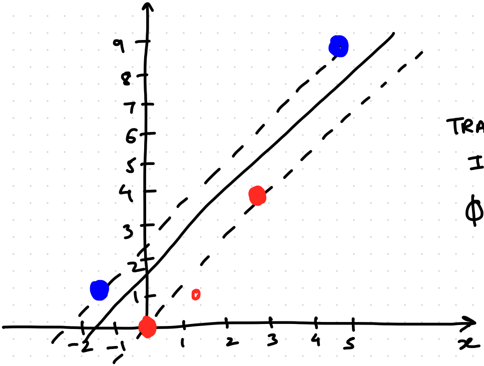
ORIGINAL DATA IN R



TRANSFORMED DATA IN R^2

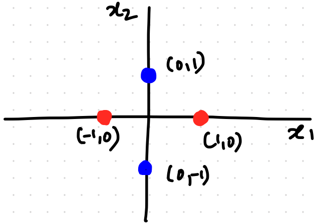
$$\phi(x) = \begin{bmatrix} \sqrt{2}\, x \\ x^2 \end{bmatrix}$$

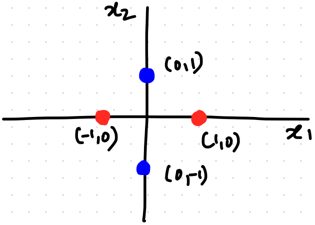
ORIGINAL DATA
IN $\mathbb{R}$



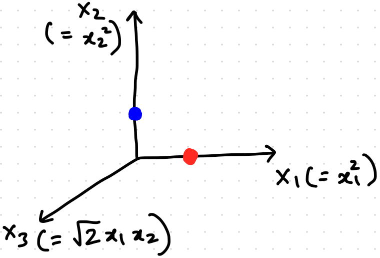
TRANSFORMED DATA
IN $\mathbb{R}^2$

$$\phi(x) = \begin{bmatrix} \sqrt{2}\,x \\ x^2 \end{bmatrix}$$

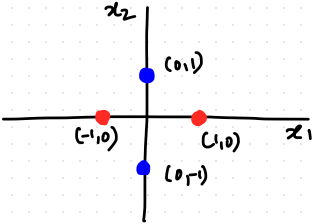

x_2
(0,1)
(-1,0)
(1,0)
x_1
(0,-1)



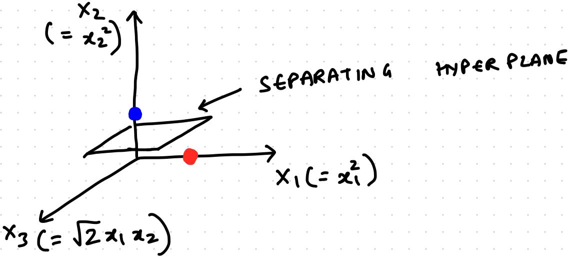
x_2
$(0,1)$
$(-1,0)$
$(1,0)$
x_1
$(0,-1)$



X_2
$(= x_2^2)$
$X_1 (= x_1^2)$
$X_3 (= \sqrt{2} x_1 x_2)$



x_2
(0,1)
(-1,0)
(1,0)
x_1
(0,-1)



X_2 $(= x_2^2)$
SEPARATING HYPER PLANE
$X_1 (= x_1^2)$
$X_3 (= \sqrt{2} x_1 x_2)$

Linear SVMs in higher dimensions

Linear SVM:

Maximize

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

such that constriants are satisfied.

$$\downarrow$$

Transformation (ϕ)

$$\downarrow$$

$$L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \phi(\overline{x_i}).\phi(\overline{x_j})$$

Linear SVMs in higher dimensions: Steps

1. Compute $\phi(x)$ for each point

$$\phi : \mathbb{R}^d \rightarrow \mathbb{R}^D$$

2. Compute dot products over $\mathbb{R}^D$ space

Q. If $D >> d$

Both steps are expensive!

Kernel Trick

Kernel Trick

- Can we compute $K(\bar{x}_i, \bar{x}_j)$, such that

Kernel Trick

- Can we compute $K(\bar{x}_i, \bar{x}_j)$, such that
- $K(\bar{x}_i, \bar{x}_j) = \phi(\bar{x}_i).\phi(\bar{x}_j)$, where

Kernel Trick

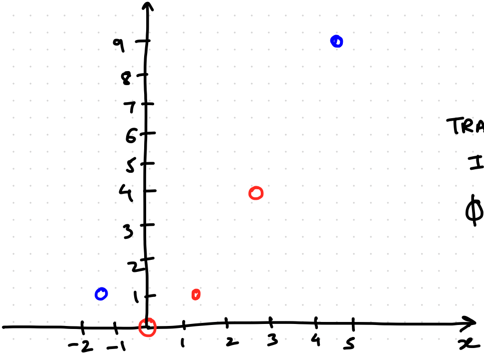
- Can we compute $\mathsf{K}(\bar{x}_i, \bar{x}_j)$, such that
- $\mathsf{K}(\bar{x}_i, \bar{x}_j) = \phi(\bar{x}_i).\phi(\bar{x}_j)$, where
- $\mathsf{K}(\bar{x}_i, \bar{x}_j)$ is some function of dot product in original dimension

Kernel Trick

- Can we compute $\mathrm{K}(\bar{x}_i, \bar{x}_j)$, such that
- $\mathrm{K}(\bar{x}_i, \bar{x}_j) = \phi(\bar{x}_i).\phi(\bar{x}_j)$, where
- $\mathrm{K}(\bar{x}_i, \bar{x}_j)$ is some function of dot product in original dimension
- $\phi(\bar{x}_i).\phi(\bar{x}_j)$ is dot product in high dimensions (after transformation)

KERNEL TRICK

ORIGINAL DATA IN R



TRANSFORMED DATA IN R^2

$$\phi(x) = \begin{bmatrix} \sqrt{2}\,x \\ x^2 \end{bmatrix}$$

KERNEL TRICK

$$\phi(x) = \begin{bmatrix} \sqrt{2}\,x \\ x^2 \end{bmatrix}$$

$$K(x_i, x_j) = \ ?$$

KERNEL TRICK

$$\phi(x) = \begin{bmatrix} \sqrt{2}\, x \\ x^2 \end{bmatrix}$$

$$K(x_i, x_j) = (1 + x_i \cdot x_j)^2 - 1$$

KERNEL TRICK

$$\phi(x) = \begin{bmatrix} \sqrt{2}\, x \\ x^2 \end{bmatrix}$$

$$K(x_i, x_j) = (1 + x_i \cdot x_j)^2 - 1$$

$$(1 + x_i \cdot x_j)^2 - 1 = 1 + 2 x_i \cdot x_j + x_i^2 x_j^2 - 1$$

$$= 2 x_i \cdot x_j + x_i^2 x_j^2$$

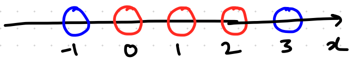
$$= (\sqrt{2} x_i \cdot \sqrt{2} x_j + x_i^2 \cdot x_j^2)$$

$$= \langle \sqrt{2} x_i, x_i^2 \rangle \cdot \langle \sqrt{2} x_j, x_j^2 \rangle$$

$$= \phi(x_i) \cdot \phi(x_j)$$

ORIGINAL DATASET

x	y
-1	-1
0	1
1	1
2	1
3	-1

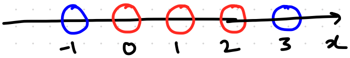


ORIGINAL DATASET

x	y
-1	-1
0	1
1	1
2	1
3	-1

TRANSFORMED DATASET

x	$\sqrt{2}x$	x^2	y
-1	$-\sqrt{2}$	1	-1
0	0	0	1
1	$\sqrt{2}$	1	1
2	$2\sqrt{2}$	4	1
3	$3\sqrt{2}$	9	-1



ORIGINAL DATASET

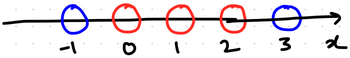
x	y
-1	-1
0	1
1	1
2	1
3	-1

TRANSFORMED DATASET

x	$\sqrt{2}x$	x^2	y
-1	$-\sqrt{2}$	1	-1
0	0	0	1
1	$\sqrt{2}$	1	1
2	$2\sqrt{2}$	4	1
3	$3\sqrt{2}$	9	-1

Calculation w/o Kernel Trick

$$\left.\begin{array}{l} \phi(x_1) = \langle \sqrt{2}x, x^2 \rangle \ : 2 \\ \phi(x_2) = \langle \sqrt{2}x, x^2 \rangle \ : 2 \\ \phi(x_1)\cdot\phi(x_2) = 2 \text{ MULTIPLICATION} + 1 \text{ Addition} \end{array}\right\} 7$$



ORIGINAL DATASET

x	y
-1	-1
0	1
1	1
2	1
3	-1

TRANSFORMED DATASET

x	$\sqrt{2}x$	x^2	y
-1	$-\sqrt{2}$	1	-1
0	0	0	1
1	$\sqrt{2}$	1	1
2	$2\sqrt{2}$	4	1
3	$3\sqrt{2}$	9	-1

Calculation with Kernel Trick

$$K(x_1, x_2) = (1 + x_1 \cdot x_2)^2 - 1$$

$x \cdot x_2 \longrightarrow 1$

$1 + x_1 \cdot x_2 \longrightarrow 1$

$(1 + x_1 \cdot x_2)^2 \longrightarrow 1$

$(1 + x_1 \cdot x_2)^2 - 1 \longrightarrow 1$

$\}\ 4$

Kernel Trick

Q) Why did we use dual form?

Kernel Trick

Q) Why did we use dual form?

Kernels again!!

Kernel Trick

Q) Why did we use dual form?

Kernels again!!

Primal form doesn't allow for the kernel trick $\mathrm{K}\left(\bar{x}_1, \bar{x}_2\right)$ in dual and compute $\phi(x)$ and then dot product in D dimensions

Some Kernels

1. Linear: $\mathsf{K}(\bar{x}_1, \bar{x}_2) = \bar{x}_1 \bar{x}_2$
2. Polynomial: $\mathsf{K}(\bar{x}_1, \bar{x}_2) = (p + \bar{x}_1 \bar{x}_2)^q$
3. Gaussian: $\mathsf{K}(\bar{x}_1, \bar{x}_2) = e^{-\gamma ||\bar{x}_1 - \bar{x}_2||^2}$ where $\gamma = \frac{1}{2\sigma^2}$ - Also called Radial Basis Function (RBF)

Kernels

Q) For $\bar{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ what space does kernel $\mathrm{K}(\bar{x}, \bar{x'}) = (1 + \bar{x}\bar{x'})^3$ belong to?

$\bar{x} \in \mathbb{R}^2$

$\phi(\bar{x}) \in \mathbb{R}^?$

$$\begin{aligned} K(x,z) &= (1 + x_1 z_1 + x_2 z_2)^3 \\ &= \dots \\ &= < 1, x_1, x_2, x_1^2, x_2^2, x_1^2 x_2, x_1 x_2^2, x_1^3, x_2^3, x_1 x_2 > \end{aligned}$$

10 dimensional?

Kernels

Q) For $\bar{x} = x$; what space does RBF kernel lie in?

$$
\begin{aligned}
K(x, z) &= e^{-\gamma ||x-z||^2} \\
&= e^{-\gamma (x-z)^2}
\end{aligned}
$$

Now:

$$e^{\alpha} = \sum_{n=0}^{\infty} \frac{\alpha^n}{n!}$$

$\therefore$ $e^{-\gamma(x-z)^2}$ is ∞ dimensional!!

SVM: Parametric or Non-Parametric

Q) Is SVM parametric or non-parametric?

SVM: Parametric or Non-Parametric

Q) Is SVM parametric or non-parametric?

Yes and No

Yes $\rightarrow$ Linear kernel or polynomial kernel (form fixed)

No $\rightarrow$ RBF (form changes with data)

RBF is Non-Parametric

$$\hat{y}(x_{test}) = sign(\bar{w}\bar{x}_{test} + b)$$
$$= sign(\sum_{j=1}^{N_{SV}} \alpha_j y_j \bar{x}_j \bar{x}_{test} + b)$$
$$\hat{y}(X_{test}) = sign(\sum_{j=1}^{N} \alpha_j y_j K(\bar{x}_j, \bar{x}_{test}) + b)$$

$\alpha_j = 0$ where $j \neq$ S.V.

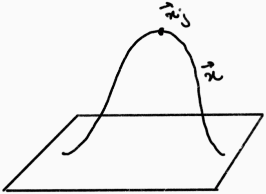
RBF is Non-Parametric

Now $K(\bar{x}_j, \bar{x}_{test})$ for RBF is:

$$e^{-\gamma||\bar{x}_j - \bar{x}_{test}||^2}$$

$\therefore$ Hypothesis is a function of “all” train points

Closer $\bar{x}$ is to $\bar{x}_N$; more is it influencing $\hat{y}(\bar{x})$ - hypothesis function



γ = Low

High influence of $\vec{x}_j$

RBF is Non-Parametric

- Now if we add a point to the dataset

RBF is Non-Parametric

- Now if we add a point to the dataset
- Functional form can adapt (similar to KNN)

RBF is Non-Parametric

- Now if we add a point to the dataset
- Functional form can adapt (similar to KNN)
- $\therefore$ SVM with RBF kernel is non-parametric

Interpretation of RBF

- $\hat{y}(x) = sign(\sum \alpha_i y_i e^{-||x-x_i||^2} + b)$

Interpretation of RBF

- $\hat{y}(x) = sign(\sum \alpha_i y_i e^{-||x-x_i||^2} + b)$
- $-||x - x_i||^2$ corresponds to radial term

Interpretation of RBF

- $\hat{y}(x) = sign(\sum \alpha_i y_i e^{-||x-x_i||^2} + b)$
- $-||x - x_i||^2$ corresponds to radial term
- $\sum \alpha_i y_i$ is the activation component

Interpretation of RBF

- $\hat{y}(x) = sign(\sum \alpha_i y_i e^{-||x-x_i||^2} + b)$
- $-||x - x_i||^2$ corresponds to radial term
- $\sum \alpha_i y_i$ is the activation component
- $e^{-||x-x_i||^2}$ is the basis component

RBF: Effect of γ

γ: How far is the influence of a single training sample