Homework 1

Karthik Ragunath Ananda Kumar
KXA200005

Warmup

1) a) Given :-

A function $f: \mathbb{R}^n \to \mathbb{R}$ is convex if for all $x, y \in \mathbb{R}^n$ and $\lambda \in [0,1]$

$$\lambda f(x) + (1-\lambda) f(y) \geq f(\lambda x + (1-\lambda) y) \rightarrow (1)$$

Multiplying (1) by w, where $w \geq 0$

$$w[\lambda f(x) + (1-\lambda) f(y)] \geq w f(\lambda x + (1-\lambda) y)$$

(since $w \geq 0$, inequality is valid)

$$\lambda (w f(x)) + (1-\lambda) w f(y) \geq w f(\lambda x + (1-\lambda) y)$$

It is given that

$f(x) = w f(x)$ where $w \geq 0$

$$\therefore \lambda f(x) + (1-\lambda) f(y) \geq f(\lambda x + (1-\lambda) y) \rightarrow (2)$$

$$(1) = (2)$$

$\therefore$ Hence proved that $f(x) = w f(x)$ is a convex function

③

ⓑ To prove:-

$f(x) = f_1(x) + f_2(x)$ is a convex function

Given

$f_1 : \mathbb{R}^n \to \mathbb{R}$ and $f_2 : \mathbb{R}^n \to \mathbb{R}$ are convex functions.

Proof:-

Since f_1 and f_2 are convex functions

$$\lambda f_1(x) + (1-\lambda) f_1(y) \geq f_1(\lambda x + (1-\lambda)y) \rightarrow ①$$

and also

$$\cancel{\lambda f_2(y)}\ \lambda f_2(x) + (1-\lambda) f_2(y) \geq f_2(\lambda x + (1-\lambda) y) \rightarrow ②$$

Adding ① and ②

$$\lambda f_1(x) + \lambda f_2(x) + (1-\lambda) f_1(y) + (1-\lambda) f_2(y) \geq f_1(\lambda x + (1-\lambda) y) + f_2(\lambda x + (1-\lambda) y)$$

Scanned with CamScanner

$$\lambda(f_1(x)+f_2(x))+(1-\lambda)(f_1(y)+f_2(y)) \geq f_1(\lambda x+(1-\lambda)y)+f_2(\lambda x+(1-\lambda)y) \rightarrow ③$$

It is given that

$f(x)=f_1(x)+f_2(x)$

$\therefore \lambda f(x)+(1-\lambda)f(y) \geq f(\lambda x+(1-\lambda)y)$.

Thus, we have proved that,
$f(x)=f_1(x)+f_2(x)$ is a convex function

(c) To prove :-
$f(x)=\max\{f_1(x), f_2(x)\}$ is a convex function in the domain $f: \mathbb{R}^n \rightarrow \mathbb{R}$

Given :-
f_1 and f_2 are convex functions in the domain $\mathbb{R}^n \rightarrow \mathbb{R}$ and $\lambda \in [0,1]$

Proof: $f(x)=\max\{f_1(x), f_2(x)\}$

$\therefore$ ~~$f(\lambda x+(1-\lambda)$~~

$$f(\lambda x+(1-\lambda)y)=\max\{f_1(\lambda x+(1-\lambda)y), f_2(\lambda x+(1-\lambda)y)\} \rightarrow ①$$

④

since f_1 and f_2 are convex functions

$$f_1(\lambda x+(1-\lambda)y) \leq \lambda f_1(x)+(1-\lambda) f_1(y) \rightarrow ②$$

$$f_2(\lambda x+(1-\lambda) y) \leq \lambda f_2(x)+(1-\lambda) f_2(y) \rightarrow ③$$

Substituting ② and ③ in ①:-

$$\therefore f(\lambda x+(1-\lambda)y) \leq \max\{\lambda f_1(x)+(1-\lambda)f_1(y), \lambda f_2(x)+(1-\lambda)f_2(y)\}$$

$$f(\lambda x+(1-\lambda)y) \leq \lambda \max\{f_1(x), f_2(x)\} + (1-\lambda)\max\{f_1(y), f_2(y)\} \rightarrow ①$$

We know that

$$f(x) = \max\{f_1(x), f_2(x)\} \text{ (Given)} \rightarrow ②$$

Rewriting ① using ②

$$\cancel{f(\lambda x)} \; f(\lambda x+(1-\lambda)y) \leq \lambda f(x)+(1-\lambda) f(y)$$

Hence we have proved that $f(x) = \max\{f_1(x), f_2(x)\}$ is a convex function

2 Computing subgradients :-

a) $f(x) = \max\{x^2 - 2x, |x|\}$ at $x=0$, $x=-2$, $x=1$.

at $x=0$

$f(x) = \max\{0 - 2(0), |0|\}$

$f(x) = \max\{0, 0\}$

Therefore both functions x^2-2x and $|x|$ have the same value at $x=0$

$\therefore$ Lets consider the function $f(x) = x^2-2x$

$\therefore f(x) = x^2 - 2x$

$\frac{d}{dx}(f(x)) = 2x - 2$

at $x=0$

$$\boxed{f'(0) = 2(0) - 2 = -2}$$

$\therefore$ slope, $m = -2$

The tangent at $x=0$ can be written in two point form as: $(y-y_1) = m(x_1 - x_1)$; At $x=0$; $y = 0^2 - 0 = 0$.

$(y - 0) = m(x - 0)$

$$\boxed{y = -2x}$$

(6)

Lets also consider

$f(x) = |x|$ at $x=0$

$$f(x) = \begin{cases} x, & x \geq 0 \\ -x, & x < 0 \end{cases}$$

Since $x=0$.

$f(x) = x$

$f'(x) = 1$

substituting $x=0$,

$$\boxed{f'(0) = 1}$$

The tangent at $x=0$ can be written as

$y - f(0) = 1(x-0)$

$y - 0 = 1(x-0)$

$$\boxed{y = x}$$

Thus, we have given equations of both subgradients at $x=0$

at $x = -2$:-

$f(x) = \max\{x^2 - 2x, |x|\}$

$= \max\{4 - 2(-2), |-2|\}$

$= \max\{8, 2\} = 8$

Thus at $x = -2$,

$f(x) = x^2 - 2x$

$f'(x) = 2x - 2$

$f'(-2) = 2(-2) - 2$

$= -4 - 2$

$\boxed{f'(-2) = -6}$

The tangent at $x = -2$ can be written as

$y - f(-2) = f'(-2)[x - (-2))$

$y - 8 = -6(x + 2)$

$\boxed{y = -6x - 4}$

8

at $x=1$,

$f(x) = \max\{x^2-2x, |x|\}$

$f(1) = \max\{1-2, |1|\}$

$= \max\{-1, 1\}$

$f(1) = 1$

$\therefore$ at $x=1$

$f(x) = |x|$.

$\therefore$ $f(x) = \begin{cases} x, & x \geq 0 \\ -x, & x < 0 \end{cases}$

since $x=1$

$f(x) = x$

$f'(x) = 1$

$\therefore \boxed{f'(1) = 1}$

The tangent at $x=1$ is given by

$(y - f(1)) = f'(1)(x-1)$

$y-1 = 1(x-1)$

$\boxed{y = x}$

(b) $g(x) = \max\{(x-1)^2, (x-2)^2\}$ at $x = 1.5$ and $x = 0$

at $x = 1.5$

$$g(x) = \max\{(1.5-1)^2, (1.5-2)^2\}$$

$$g(x) = \max\{(+0.5)^2, (-0.5)^2\}$$

$$= \max\{0.25, 0.25\} = 0.25$$

$\therefore$ Both $(x-1)^2$ and $(x-2)^2$ have same value at $x = 1.5$

Let us consider

$$g(x) = (x-1)^2 \text{ at } x = 1.5$$

$$g'(x) = 2(x-1)$$

$$g'(1.5) = 2(1.5-1)$$

$$= 2(0.5)$$

$$\boxed{g'(1.5) = 1}$$

The tangent at $x = 1.5$ at $g(x) = (x-1)^2$ can be written as :-

$$y - g(1.5) = g'(1.5)(x - 1.5)$$

$$y - 0.25 = 1(x - 1.5)$$

$$\boxed{y = x - 1.25}$$

Let us also consider

$g(x) = (x-2)^2$ at $x = 1.5$

$g'(x) = 2(x-2)$

$g'(1.5) = 2(1.5-2)$

$g'(1.5) = -1$

The tangent at $x = 1.5$ for $g(x) = (x-2)^2$ can be written as

$y - g(1.5) = g'(1.5)(x-1.5)$

~~for (or)~~ $y - 0.25 = -1(x-1.5)$

$\boxed{y = -x + 1.75}$

at $x=0$:

$g(x) = \max\{(-1)^2, (-2)^2\}$

$= \max\{1, 4\} = 4$

$\therefore$ at $x=0$, $g(x) = (x-2)^2$

$g'(x) = 2(x-2)$

$g'(0) = 2(0-2)$

$\boxed{g'(0) = -4}$

The tangent at $x=0$ for $g(x) = (x-2)^2$ can be written as

$(y - g(0)) = g'(0)(x - 0)$

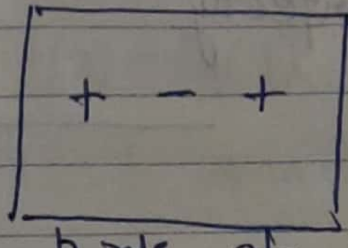
$y - 4 = (-4)(x - 0)$

$\boxed{y = -4x + 4}$

Problem 1:

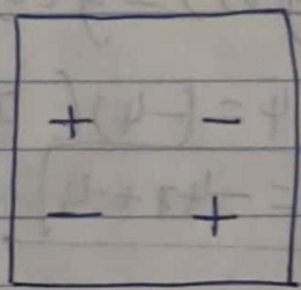
(1.) The smallest size interms of No. of data points, two dimensional data set containing both positive and negative classes will fail to converge is 3.

This is because, considering the case when all the data points lie on the same line in the 2D plane

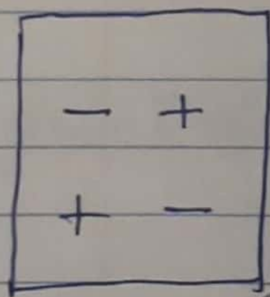


Points of two classes, there exist no linear separator to separate the data

This can be generalized to the case when number of data points is greater than 3,
eg: Consider



(or)



No linear separator can separate the data points of two class labels.

This can be generalized to k dimensions as:
Whenever there are (k+1) data points of two classes to be separated in a k-dimensional space, a linear separator may not be able to separate the classes perfectly.

(2) (a) The loss function is given by

$$L = \frac{1}{M}\left(\sum_{m=1}^{M} \max\{0, 1-y^{(m)}(w^T x^{(m)}+b)\}^2\right) \rightarrow (1)$$

We know that a function f given by
$f(x) = \max\{f_1(x), f_2(x)\}$ where $\rightarrow (2)$
f_1 and f_2 are convex functions

is a convex function.
Comparing (1) and (2)
In the given loss function,
we can assume that

$f_1(x) = 0$

$f_2(x) = 1 - y^{(m)}(w^T x^{(m)} + b)$

A constant is a convex function (and a concave fn too)
$\therefore f_1(x) = 0$ is convex function
Considering

$f_2(x) = 1 - y^{(m)}(w^T x^{(m)} + b) \rightarrow (3)$

We can consider $f_2(x)$ again as a combination of 2 functions

$f_2(x) = f_3(x) + f_4(x)$

where $f_3(x) = 1$ and $f_4(x) = -y^{(m)}(w^T x^{(m)} + b)$

$f_3(x)$ and $f_4(x)$ are both convex functions since both constants and linear functions are convex in nature.
We know that sum of two convex functions is a convex function and hence $f_2(x)$ defined in 3 is a convex function

Now, since $f_1(x)$ and $f_2(x)$ are convex functions.
~~$f(x)$ $f(x)=$~~
$f(x) = \max\{f_1(x), f_2(x)\}$ is a convex functions since we know that max of two convex functions is a convex function.

~~Now, we know that~~
~~$L = \frac{1}{M}\sum_{m=}^{M} (f(x))$~~

Now, we can rewrite the Loss function as

$$L = \frac{1}{M}\sum_{m=1}^{M} (f(x))^2.$$

where $f(x) = \max\{0, (1 - y^{(m)}(w^T x^{(m)} + b))\}$

15

We know that composition of two convex function is a convex function.

Let's consider, $g(x) = x^2$ which is a convex function

we already computed that

$f(x) = \max\{0, 1 - y^{(m)}(w^T x^{(m)} + b)\}$

is a convex function

$\therefore\ h(x) = g(f(x)) = (\max\{0, 1 - y^{(m)}(w^T x^{(m)} + b)\})^2$

is a convex function.

$\therefore$ Loss function can be rewritten as

$$L = \frac{1}{M} \sum_{m=1}^{M} h(x)$$

where $h(x)$ is a convex function.

We know that sum of convex functions are convex functions,

hence

$$L = \frac{1}{M} \sum_{m=1}^{M} h(x)$$

where $h(x) = (\max\{0, 1 - y^{(m)}(w^T x^{(m)} + b)\})^2$
is a convex function

$\therefore$ The given Loss function is a convex function.

Therefore, there is only one local optima which is also global optima for this loss function and there are no other local optimas which are not global optima.

(b) (i) Standard subgradient descent with step size $\gamma_t = 1$

The value of weights and biases for the first 3 iterations:-

~~1. loss [~~ ~~2.859, 0.91~~

1. $w_1 = +2.559$; $w_2 = +0.917$; $w_3 = -0.216$; $w_4 = -3.346$; $b = -0.7107$
2. $w_1 = 3.101$; $w_2 = 0.7501$; $w_3 = 0.748$; $w_4 = -1.741$; $b = -1.763$
3. $w_1 = -1.278$; $w_2 = 0.666$; $w_3 = -1.771$; $w_4 = -5.280$; $b = -2.958$

Standard subgradient descent with given loss function doesn't converge for the step size = 1. No. of iterations to get perfect classifier cannot be determined.

The final weights and biases are :-

$w_1 =$ nan (cannot be defined)
$w_2 =$ nan
$w_3 =$ nan
$w_4 =$ nan
$b =$ nan (cannot be defined).

(ii) Stochastic sub gradient descent with step size = 1.

The value of weights and biases for the first 3 iterations :-

1. $w_1 = 9.235$; $w_2 = 4.939$; $w_3 = 3.935$; $w_4 = -3.6267$; $b = -2$
2. $w_1 = 9.235$; $w_2 = 4.939$; $w_3 = 3.935$; $w_4 = -3.6267$; $b = -2$
3. $w_1 = -9.949$; $w_2 = -32.961$; $w_3 = 14.815$; $w_4 = -50.318$; $b = -18.471$

Stochastic subgradient descent with step size of 1 doesn't converge for the given loss function. Therefore, no. of iterations to get perfect classifier cannot be determined.

The final weights and biases are:-
w_1=nan (cannot be determined);
w_2=nan; w_3=nan; w_4= nan;
b=nan (cannot be determined).

(iii) As we reduce the step size, the loss function begins to converge.

Standard sub gradient descent
1) at step size =0.001, converges at 368975th epoch

2) at step size=0.0005, converges at 737952th epoch

3) at step size = 1, ~~Never~~ never converge.

Stochastic subgradient descent
1) at step size = 1, does not converge

2) at step size= 0.01, converges at 97891th iteration.

From the above analysis, we can

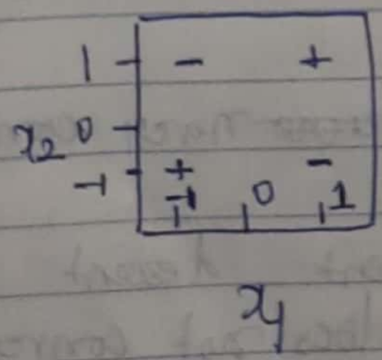
(19a)

infer that: ~~*~~ the convex loss function doesn't converge when step size is too large

* Converges as we decrease the step size below certain value (obtained from experimenting)
* If the step size is too small, it takes a very long time to converge.

(c) Subgradient descent with step size of 1 doesn't always converge for this loss function as we saw from our training. ~~th~~

Problem 2 :-

1. 

(a) $\phi(x_1, x_2) = \begin{bmatrix} x_1 + x_2 \\ x_1 - x_2 \end{bmatrix}$

$\therefore$ In tranformed space.

At point $(+1, +1)$

$$\phi(x_1, x_2) = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

At point $(-1, -1)$

$$\phi(x_1, x_2) = \begin{bmatrix} -2 \\ 0 \end{bmatrix}$$

At point $(+1, -1)$

$$\phi(x_1, x_2) = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

At point $(-1, +1)$

$$\phi(x_1, x_2) = \begin{bmatrix} 0 \\ -2 \end{bmatrix}$$

If there exist a valid linear separator then at $(x_1, x_2) = (1, 1)$

$$w_1 x_1 + w_2 x_2 + b > 0$$

$$\therefore \; w_1 + w_2 + b > 0 \rightarrow 1$$

at $-1, -1$

$$w_1 x_1 + w_2 x_2 + b > 0$$

$$-w_1 - w_2 + b > 0 \rightarrow 2$$

at $+1, -1$

$$w_1 x_1 + w_2 x_2 + b < 0$$

$$w_1 - w_2 + b < 0 \rightarrow 3$$

at $-1, +1$

$$w_1 x_1 + w_2 x_2 + b < 0$$

$$-w_1 + w_2 + b < 0 \rightarrow 4$$

Adding 1 and ~~4~~ 2

$2b > 0$

$\therefore b > 0 \rightarrow 5$

Adding 4 and 3

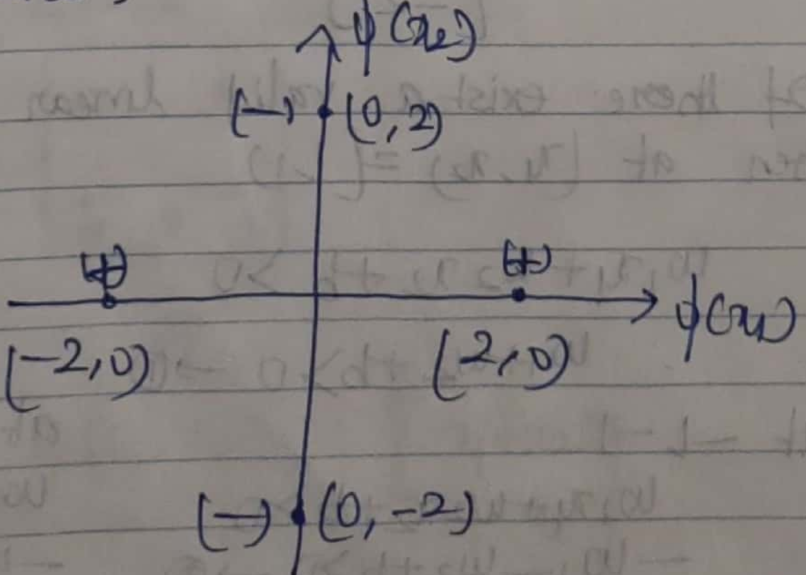
$2b < 0$

$b < 0 \rightarrow 6$

5 and 6 contradicts.

This shows that linear separator is not possible

Also graphically in transformed space.



We can clearly see that no linear separator can perfectly classify the points.

(b) $\phi(x_1, x_2) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ x_1 x_2 \end{bmatrix}$

At $(+1, +1)$

$\phi(x_1, x_2) = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

At $(-1, -1)$

$\phi(x_1, x_2) = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

At $(+1, -1)$

$\phi(x_1 x_2) = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$

At ~~...~~ $(-1, +1)$

$\psi(x_1, x_2) = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$

Therefore there are only two unique points in transformed space

(1) $\phi(x_1, x_2) = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ for +re class

(2) $\phi(x_1, x_2) = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$ for -re class

23

A valid linear separator will be of the form

at $\phi(x_1, x_2, x_3) = (1, 1, 1)$ for +ve class.

$w_1 x_1 + w_2 x_2 + w_3 x_3 + b > 0$

$\therefore$ ~~$x_1 + x_2 + x$~~

$\therefore w_1 + w_2 + w_3 + b > 0 \rightarrow$ 1

at $\phi(x_1, x_2, x_3) = (1, 1, -1)$ for -ve class

$\therefore w_1 + w_2 - w_3 + b < 0 \rightarrow$ 2

Fitting the transformed points with perceptron loss:

x_1	x_2	x_3	y
1	1	1	+1
1	1	-1	-1

we can get $w_1 = 0$
$w_2 = 0$
$w_3 = 0.02$
$b = 0$

as the parameters of linear separator.

substituting w_1, w_2, w_3, b values in
(1) and (2)
we could see that

$0.02 > 0$

$-0.02 < 0$

Thus, we have found a valid linear separator.

graphically too, since the no of data points having both labels/classes is less than the no of dimensions of the input data, there will always exist a linear separator

---

(c) $\phi_{(x_1,x_2)} = \begin{bmatrix} \exp(x_1) \\ \exp(x_2) \end{bmatrix}$

at point +1,+1

$\phi(x_1, x_2) = \begin{bmatrix} e^1 \\ e^1 \end{bmatrix}$

at point +1,−1

$\phi(x_1, x_2) = \begin{bmatrix} e^1 \\ e^{-1} \end{bmatrix}$

at point −1,−1

$\phi(x_1, x_2) = \begin{pmatrix} e^{-1} \\ e^{-1} \end{pmatrix}$

at point −1,+1

$\phi(x_1, x_2) = \begin{bmatrix} e^{-1} \\ e^1 \end{bmatrix}$

It valid linear separators.
we can form the following equations where valid values for w_1, w_2, b will exist

at (1,1): $w_1 e^1 + w_2 e^1 + b > 0 \rightarrow$ 1
at (-1,-1): $w_1 e^{-1} + w_2 e^{-1} + b > 0 \rightarrow$ 2
at (+1,-1): $w_1 e^1 + w_2 e^{-1} + b < 0 \rightarrow$ 3
at (-1,+1): $w_1 e^{-1} + w_2 e^1 + b < 0 \rightarrow$ 4

Adding 1 and 2 :-

$w_1(e^1 + e^{-1}) + w_2(e^1 + e^{-1}) + b > 0 \rightarrow$ 5

Adding 3 and 4
$w_1(e^1 + e^{-1}) + w_2(e^1 + e^{-1}) + b < 0 \rightarrow$ 6

Equations 5 and 6 contradict with each other.

∴ we cannot find valid w_1, w_2, b for the linear separator.

∴ The given transformed space is not linearly separable.

(d) $\phi(x_1, x_2) = \begin{bmatrix} x_1 \sin(x_2) \\ x_1 \end{bmatrix}$

~~At point +1,+1~~

$\phi(1,1) = \begin{bmatrix} \sin(1) \\ 1 \end{bmatrix}$

$\phi(-1,-1) = \begin{bmatrix} -1 \times \sin(-1) \\ -1 \end{bmatrix} = \begin{bmatrix} \sin(+1) \\ -1 \end{bmatrix}$

$\phi(+1,-1) = \begin{bmatrix} +\sin(-1) \\ 1 \end{bmatrix} = \begin{bmatrix} -\sin(1) \\ 1 \end{bmatrix}$

$\phi(-1,+1) = \begin{bmatrix} -1 \times \sin(+1) \\ -1 \end{bmatrix} = \begin{bmatrix} -\sin(1) \\ -1 \end{bmatrix}$

linear separator equations can be written as :

At $\phi(1,1) \Rightarrow w_1 \sin(1) + w_2 + b > 0 \rightarrow (1)$
At $\phi(-1,-1) \Rightarrow w_1 \sin(1) - w_2 + b > 0 \rightarrow (2)$
At $\phi(+1,-1) \Rightarrow -w_1 \sin(1) + w_2 + b < 0 \rightarrow (3)$
At $\phi(-1,+1) \Rightarrow -w_1 \sin(1) - w_2 + b < 0 \rightarrow (4)$

If we try to fit a linear separator for the transformed data space:

$\phi(x_1)$	$\phi(x_2)$	y
$\sin(1)$	1	1
$\sin(1)$	-1	1
$-\sin(1)$	1	-1
$-\sin(1)$	-1	-1

On fitting the above data with perceptron loss, we found a valid linear separator with parameters

$w_1 = 0.03365$

$w_2 = 0$

$b = 0$

On substituting the values of w_1, w_2 and b in ①, ②, ③, we could see that:

① $\Rightarrow 0.033 \times \sin(1) > 0$

② $\Rightarrow 0.033 \times \sin(1) > 0$

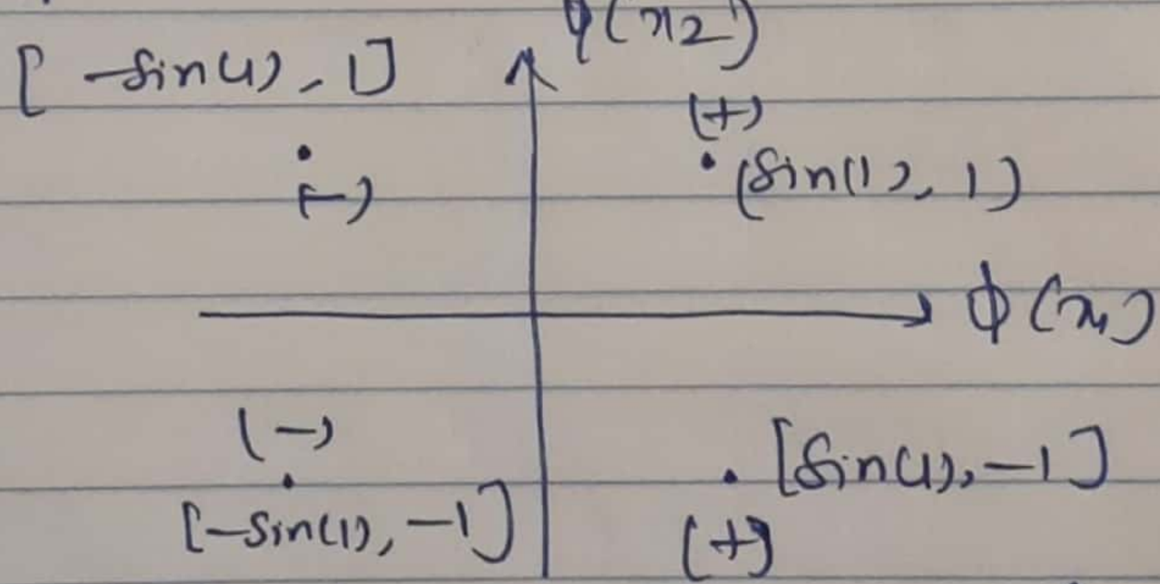
③ $\Rightarrow -0.033 \sin(1) < 0$

④ $\Rightarrow -0.033 \sin(1) < 0$.

$\therefore$ All inequalities hold correctly.

Thus, we found a valid linear separator

Graphically, the transformed space could be mapped as



Graphically too, we can see that linear separator can separate the data points of different labels.

2) Let's consider the two input dimensions as x_1 and x_2.

for fitting polynomial of degree 1:

A valid equation with the given 2 dimensional input space is:

$$x_1 + x_2 + b = 0$$

where b is a constant

∴ There are 2 terms of degree 1: (x_1, x_2) and a constant

for degree 2:-

$$x_1^2 + x_2^2 + x_1x_2 + x_1 + x_2 + b = 0$$ → 3 terms of degree 2, 2 terms of deg 1, 1 constant

for degree 3:

$$x_1^3 + x_2^3 + x_1^2 x_2 + x_2^2 x_1 + x_1^2 + x_2^2 + x_1 x_2 + x_1 + x_2 + b = 0.$$

∴ In a polynomial of degree k:

there will be $(k+1)$ terms with degree k

k terms with degree $(k-1)$

⋮

3 terms with degree 2

2 terms with degree 1

1 constant term.

③

∴ In total, there will be

$(k+1) + k + (k-1) + \cdots + 1$
parameters to update for 1 iteration of gradient.

we know that
$$1+2+\cdots+k+(k+1) = \frac{(k+1)(k+2)}{2}$$

$$= \frac{1}{2}(k^2+3k+2) \text{ parameters for 1 iteration}$$

Also there are m samples.
Inorder to find Loss function we must ~~for~~ compute linear equation $wx+b$ for all samples interms of parameter multiplications is
∴ Total complexity
$$= \frac{1}{2}(k^2+3k+2)(M) + c$$
where M is the no of samples.

The big O notation of ~~the~~ time complexity ⇒ $O(k^2 M)$ [for one iteration]
where $k \rightarrow$ degree of polynomial
$M \rightarrow$ No. of samples.

for stochastic gradient descent, we will only consider one sample at a time for one iteration.

$\therefore$ Time complexity $\}$ $O(k^2)$ (for stochastic gradient descent)

where $k \rightarrow$ degree of polynomial

Problem 3:

The optimization problem solved is

$$\min_{w,b} \frac{1}{2} \|w\|^2$$

Such that

$$y_i (w^T x^{(i)} + b) \geq 1 \text{ for all } x_i, y_i$$

By using duality principle on Lagrangian over primal variables, the dual objective function can be represented as

$$\max_{\lambda \geq 0, \sum \lambda_i y_i = 0} -\frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j k(x^{(i)}, x^{(j)}) + \sum \lambda_i$$

By Kernal trick,
polynomial of degree d can be written as:

$$K(x,z) = (x^T z)^d$$

$\therefore$, Optimization eq given to quadratic programming solver is

$$\max_{\lambda} \sum_{i=1}^{M} \lambda_i - \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} \lambda_i \lambda_j y_i y_j K(x^{(i)}, x^{(j)})$$

here $M \rightarrow$ no. of data points

such that

(1) $\sum_{i=1}^{M} \lambda_i y_i = 0$

(2) $\lambda_i \geq 0 \ \forall i$

CVXOPT expects optimization problem in the form of

$$\min \frac{1}{2} x^T P x + q^T x$$

subject to $Ax = b$

$Gx \leq h$

∴ we multiply our constraint by −1 to turn it into minimization task.

$$\min_{\lambda} (1/2)\, \lambda^T H \lambda - 1^T \lambda$$

such that $y^T \lambda = 0$

$-\lambda_i \leq 0 \ \forall i$

where H is defined as

$H_{ij} = y y^T (x x^T)^d$

where d is degree of the ~~kernel~~ polynomial we are fitting.

we used $\boxed{d=3}$ for this problem.

(i) The learned parameters are :-

$w = [-1.86439, -8.03600, 16.5629, -8.00807]$

$b = -4.0478$

(ii) The optimal margin is $\frac{2}{\|w\|} \Rightarrow 0.099195$.

~~(iii) The support vector points are:- [−2.157069255, 0.068605112, 0.369604117, 0.44147098] for +1 class~~

(iii) The support vector points are :

(a) $[-1.714642, 4.435944, -4.851051, 4.532657]$ for +1 class

(b) $[1.028716, -4.743268, 4.908330, -4.957036] \rightarrow -1$ class