EE-559 - H.W-6

RAKSHITHA PANDURANGA
rpandura@usc.edu
7890-1614-34.

Problem-1

Assumptions:

- → Fixed increment $\eta(i) = \eta = \text{constant} > 0$
- → Sequential gradient descent
- → Data points are linearly separable
- → Use reflected data points $z_n x_n$, $n = 1, 2 \cdots N$

We can set $\eta = 1$ without loss of generality:

Let $z_n \underline{x}_n' = z_n \eta \underline{x}_n$, $\eta > 0$

Then drop primes.

Algorithm:
$$\begin{cases} \underline{w}(0) \rightarrow \text{arbitrary} \\ \underline{w}(i+1) = \underline{w}(i) + z_i \underline{x}_i [\![\underline{w}(i)^T z_i \underline{x}_i \leq b]\!] \end{cases}$$

in which $z_i \underline{x}_i$, $i = 0, 1, 2 \dots$ are the cyclically order set of training data points.

let $z^i \underline{x}^i$ be the misclassified points at each iteration

Algorithm:
$$\begin{cases} \underline{w}(0) = \text{arbitrary} \\ \underline{w}(i+1) = \underline{w}(i) + z^i \underline{x}^i \end{cases}$$

where $\underline{w}(i)^T z^i x^i \leq b \;\forall i$.

Note that if $\hat{\underline{w}}$ is a solution, then $a\hat{\underline{w}}$, $a > 1$ is also a solution:

$$\hat{w}^T z_n \underline{x}_n > b \quad \forall n$$
$$a\hat{\underline{w}}^T z_n \underline{x}_n > b \quad \forall n, \text{ if } a > 1$$

Error measure →

$\mathcal{E}_w(i) = \| \underline{w}(i) - a\hat{\underline{w}} \|_2^2$

Show $\mathcal{E}_{\underline{w}}(i)$ must decrease at each iteration

$\underline{w}(i+1) - a\hat{\underline{w}} = (\underline{w}(i) - a\hat{\underline{w}}) + z^i \underline{x}^i$, $a > 1$

$\| \underline{w}(i+1) - a\hat{w} \|_2^2 = \| \underline{w}(i) - a\hat{\underline{w}} \|_2^2 + 2\, \underline{w}(i)\, z^i \underline{x}^i - 2a\hat{\underline{w}}^T z^i \underline{x}^i$
$+ \| z^i \underline{x}^i \|_2^2$

Comparing 1 & 2,
Error measure $= 2\underline{w}(i)\, z^i \underline{x}^i - 2a\hat{\underline{w}}^T z^i \underline{x}^i + \underbrace{\| z^i \underline{x}^i \|_2^2}$

But

$0 \leq$ Error measure.

$\therefore$

$\lambda^2 = \max \| x_j \|_2^2$

$c = \min \{ \hat{w}^T z_j x_j \} > b.$

Let $a = \dfrac{\lambda^2 + c}{c}$ $(a > 1)$

$\therefore \Delta_{error} = 2b - 2(\lambda^2 + c) + \lambda^2$
$= 2b - 2\lambda^2 - 2c + \lambda^2$
$= 2(b - c) - \lambda^2$

If $b - c \approx -\Delta$ where Δ is small

then $\Delta_{error} = -2\Delta - \lambda^2$ which is minimum amount of decrease in error.

For iteration i_0, $\mathcal{E}_{\underline{w}}(i_0) < (2\Delta + \lambda^2)$ then,
$\mathcal{E}_{\underline{w}}(i_0) < 0$ which is not reached with the condition

$\therefore$ Convergence is reached at $(i_0 - 1)^{th}$ iteration

Scanned by CamScanner

Problem-2 a)

$g_1(x) = -x_1 - x_2 - x_3 - x_4 - x_5$

$g_2(x) = x_1 + x_2 + x_3 + x_4 + x_5$

$g_3(x) = 0$

Datapoints	$g_1(x^1)$	$g_2(x^1)$	$g_3(x^1)$
[1 0 1 -1 2]	$-1-1+1-2$ $-4+1$ $=-3$	$+1+1-1+2$ $=3$	0

It is misclassified since $g_1(x^1) < g_2(x^1)$

updated weights

$$w^{(1)}_{(i+1)} = w^{(1)}(i) + \eta(i)x^{(1)}$$

$$= \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 1 \\ -1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 0 \\ -2 \\ 1 \end{bmatrix}$$

$$w^{(2)}_{(i+1)} = w^{(2)}(i) - \eta(i)x^{(2)}$$

$$= \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \\ 1 \\ -1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 2 \\ -1 \end{bmatrix}$$

updated $\rightarrow$ $g_1(x) = -x_2 - 2x_4 + x_5$

$g_2(x) = x_2 - 2x_4 + x_5$

$g_3(x) = 0$

Datapoints	g_1	g_2	g_3
[1 1 1 1 1]	$g_1(x^2)$ $-1-2+1=-2$	$g_2(x^2)$ $1+2-1=2$	$g_3(x^2)$ 0
[1 2 1 1 1]	$g_1(x^3)$ $-2-2+1=-3$	$g_2(x^3)$ $2+2-1=3$	$g_3(x^3)$ 0
[1 -1 1 0 -1]	$g_1(x^4)$ $1+0-1=0$	$g_2(x^4)$ $-1+0+1=0$	$g_3(x^4)$ 0

updated weights

$$w^{(3)}_{(i+1)} = w^{(3)}(i) + \eta(i)\, x^{(4)}$$

$$= \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ -1 \\ 1 \\ 0 \\ -1 \end{bmatrix}$$

$$w^{(3)} = \begin{bmatrix} 1 \\ -1 \\ 1 \\ 0 \\ -1 \end{bmatrix}$$

$$w^{(1)} = \begin{bmatrix} 0 \\ -1 \\ 0 \\ -2 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ -1 \\ 1 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ -1 \\ -2 \\ 2 \end{bmatrix}$$

updated $g_1(x) = -x_1 - x_3 - 2x_4 + 2x_5$

$g_2(x) = x_2 + 2x_4 - x_5$

$g_3(x) = x_1 - x_2 + x_3 - x_5$

II[nd] iteration

Datapoints

Datapoints	$g_1(x^1)$	$g_2(x^1)$	$g_3(x^1)$
[1 0 1 -1 2]	$-1 - 1 + 2 + 4$ $= 4$	$0 - 2 - 2$ $= -4$	$1 - 0 + 1 - 2$ $= 0$

$g_1(x^1) > g_2(x^1)$ & $g_1(x^1) > g_3(x^1)$

$\boxed{x^1 \in S_1}$

Datapoints	g_1	g_2	g_3
[1 1 1 1 1]	$-1 - 1 - 2 + 2$ $= -2$	$1 + 2 - 1$ 2	$1 - 1 + 1 - 1$ $= 0$

$g_2(x^2) > g_1(x^2)$ $g_2(x^2) > g_3(x^2)$

$\boxed{x^2 \in S_2}$

Datapoints	g_1	g_2	g_3
[1 2 1 1 1]	$-1 - 1 - 2 + 2$ $= -2$	$2 + 4 - 1$ $= 5$	$1 - 4 + 1 - 1$ -3

$g_2(x^3) > g_1(x^3)$ $g_2(x^3) > g_3(x^3)$

$\boxed{x^3 \in S_2}$

Datapoints	g_1	g_2	g_3
[1 -1 1 0 -1]	$-1 - 1 + 0 - 2 = -4$	$-1 + 0 + 1 = 0$	$1 + 1 + 1 + 1 = 4$

$g_3(x^4) > g_2(x^4)$ $\qquad$ $g_3(x^4) > g_1(x^4)$

$$\boxed{x^4 \in S_3}$$

correctly classified with weights

$$w^{(1)} = \begin{bmatrix} -1 \\ 0 \\ -1 \\ -2 \\ 2 \end{bmatrix} \qquad w^{(2)} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 2 \\ -1 \end{bmatrix} \qquad w^{(3)} = \begin{bmatrix} 1 \\ -1 \\ 1 \\ 0 \\ -1 \end{bmatrix}$$

linear discriminant functions,

$$g_1(x) = -x_1 - x_3 - 2x_4 + 2x_5$$

$$g_2(x) = x_2 + 2x_4 - x_5$$

$$g_3(x) = x_1 - x_2 + x_3 - x_5$$

Problem 2b)

$$\underline{x} = (1, x_1, x_2, 0, 0).$$

$$g_1(x) = -1 - x_2$$

$$g_2(x) = x_1$$

$$g_3(x) = 1 - x_1 + x_2$$

H_{12}

$$g_1(x) = g_2(x)$$

$$-1 - x_2 = x_1$$

$$x_2 = -(x_1 + 1)$$

$$\boxed{x_2 = -x_1 - 1}$$

x_1	-2	-1	0	1	2
x_2	+1	0	-1	-2	-3

H_{13}

$$g_1(x) = g_3(x)$$

$$-1 - x_2 = 1 - x_1 + x_2$$

$$-2x_2 = 1 - x_1 \cancel{+x_2} + 1$$

$$-2x_2 = 2 - x_1$$

$$\boxed{x_2 = \frac{1}{2}x_1 - 1}$$

x_1	-2	-1	0	1	2
x_2	-2	-3/2	-1	-1/2	0

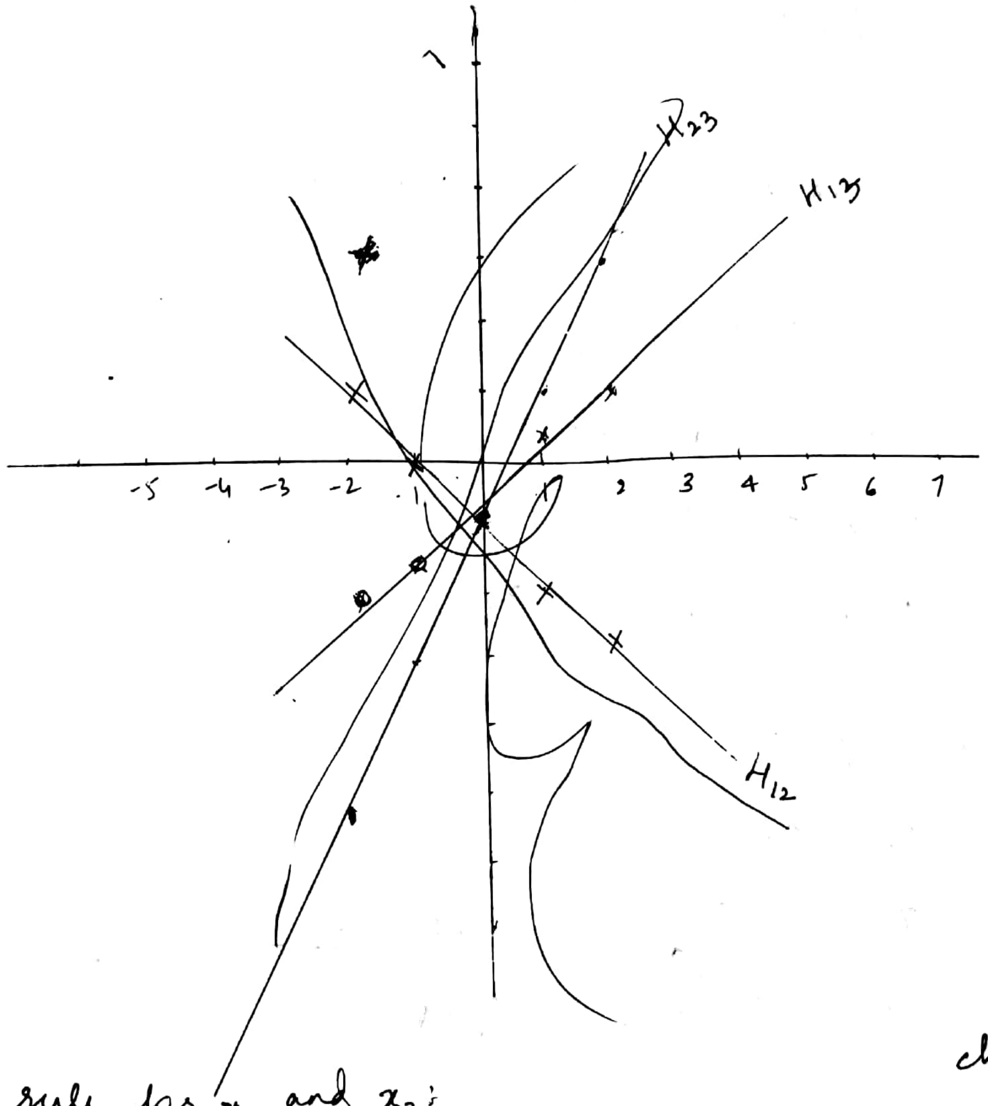
H_{23}

~~$x_1 - x_2 - x_3 = x_1 - x_2 + x_3 - x_5$~~

$x_1 = 1 - x_1 + x_2$

$\boxed{x_2 = 2x_1 - 1}$

x_1	-2	-1	0	1	2
x_2	-5	-3	-1	1	3



Decision rule for x_1 and x_2:

checking for origin (0,0)

h_{12} $g_1(x_1) - g_2(x) > 0 \rightarrow S_1 \rightarrow -1 - x_2 - x_1$ $\quad -1$

$g_1(x_1) - g_2(x) < 0 \rightarrow S_2 \rightarrow -1 - x_2 - x_1$

$\therefore$ $\boxed{\text{Origin i.e. } (0,0) \text{ on } S_2 \text{ side}}$

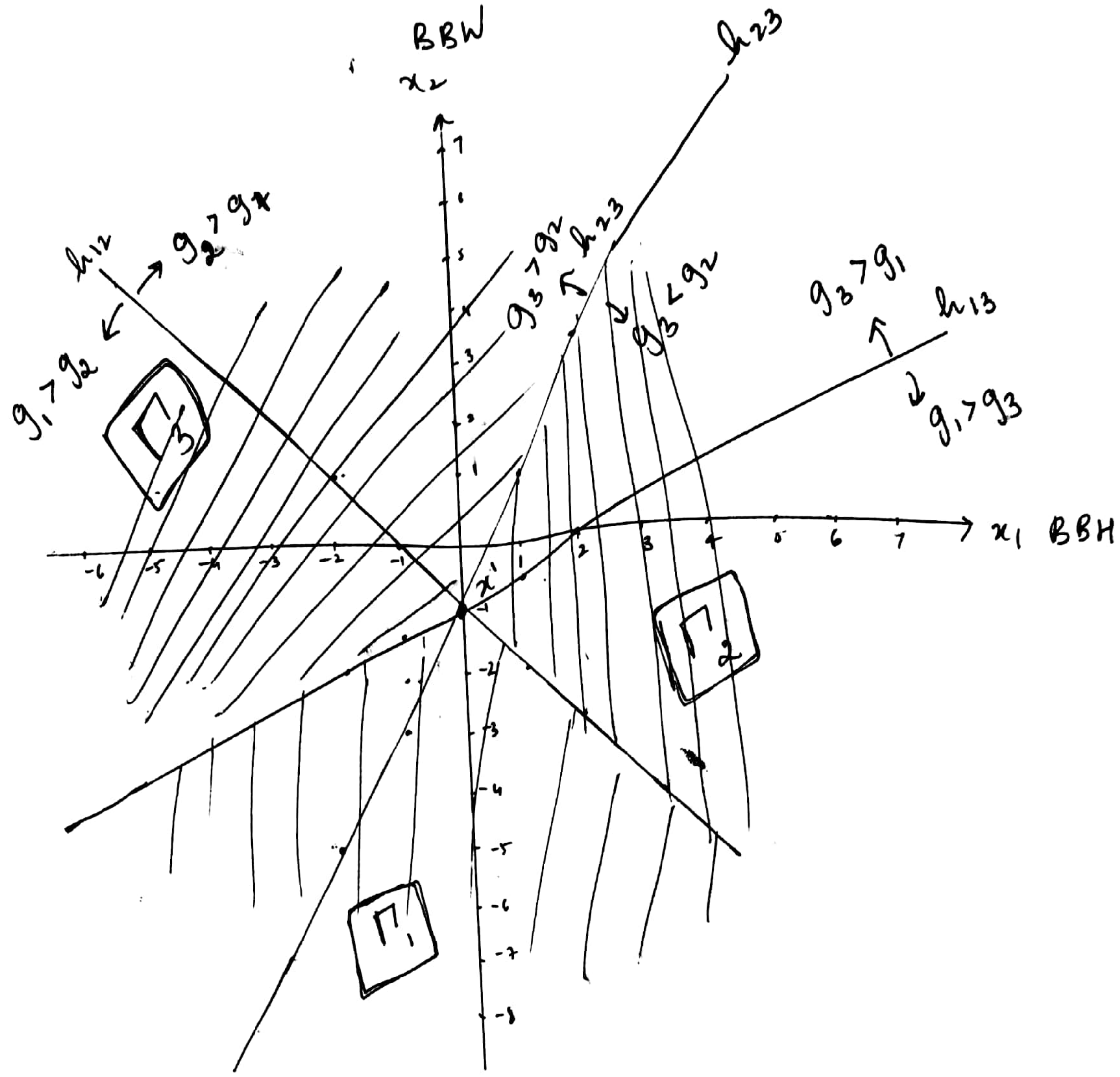
h_{13} $g_1(x) - g_3(x) > 0 \rightarrow S_1$ $\quad -1 - x_2 - 1 + x_1 - x_2$ $\quad (0,0)$ -2

$g_1(x_1) - g_3(x) < 0 \rightarrow S_3$

Origion of S_3 side

h_{23} $g_2(x) - g_3(x) > 0 \rightarrow S_2$ $x_1 - 1 - x_1 + x_2 = -1$

$g_2(x) - g_3(x) < 0 \rightarrow S_3$ Origin on S_3 side



(3) $J(\underline{w}) = \frac{1}{N}\|\underline{\underline{X}}\,\underline{w} - \underline{b}\|_2^2 + \lambda\|\underline{w}\|_2^2$

$$\nabla J(\underline{w}) = \frac{1}{N}\left[(\underline{X}\,\underline{w} - \underline{b})^T(\underline{X}\,\underline{w} - \underline{b})\right] + \lambda\left[w^T w\right]$$

$$\nabla J(\underline{w}) = \frac{1}{N}\left[\underline{w}^T\underline{\underline{X}}^T\underline{\underline{X}}\,\underline{w} - \underline{w}^T\underline{\underline{X}}^T\underline{b} - \underline{b}^T\underline{\underline{X}}\,\underline{w} + \underline{b}^T\underline{b}\right] + \lambda\left[w^T w\right]$$

$$= \frac{1}{N}\left[\underline{w}^T\underline{\underline{X}}^T\underline{\underline{X}}\,\underline{w} - 2\,\underline{w}^T\underline{\underline{X}}^T b + \underline{b}^T\underline{b}\right] + \lambda\left[w^T w\right]$$

$$\boxed{\nabla_{\underline{w}} J(\underline{w}) = \frac{1}{N}\left[\underline{\underline{X}}^T\underline{\underline{X}}\,\underline{w} - 2\,\underline{\underline{X}}^T\underline{b}\right] + 2\lambda\underline{w}}$$

$$= \frac{1}{N}\,2\,\underline{\underline{X}}^T(\underline{\underline{X}}\,\underline{w} - \underline{b})$$

b) $\nabla_{\underline{w}} J(\underline{w}) = 0$

$$0 = \frac{1}{N}\left[2\underline{\underline{X}}^T\underline{\underline{X}}\,\underline{w} - 2\underline{\underline{X}}^T\underline{b}\right] + 2\lambda\underline{w}$$

$$-2\lambda\underline{w} = \frac{1}{N}\left[2\underline{\underline{X}}^T\underline{\underline{X}}\,\underline{w} - 2\underline{X}^T\underline{b}\right]$$

$$-N2\lambda\underline{w} = 2\underline{\underline{X}}^T\underline{\underline{X}}\,\underline{w} - 2\underline{X}^T b$$

$$2\underline{X}^T b = 2\underline{\underline{X}}^T\underline{\underline{X}}\,\underline{w} + 2N\lambda\underline{w}$$

$$2\underline{X}^T b = \underline{w}\left(2\underline{\underline{X}}^T\underline{\underline{X}} + 2N\lambda\right)$$

$$\boxed{\underline{w} = \left(\underline{\underline{X}}^T\underline{\underline{X}}\right)^{-1}\left(\underline{\underline{X}}^T b - N\lambda\right)}$$

Comparing this to the pseudoinverse result, we have an extra $N\lambda$ term ~~added~~ subtracted to the ~~vector~~ ($\underline{\underline{X}}^T\underline{\underline{X}}$) ~~matrix~~ ($\underline{\underline{X}}^T b$) term

4) a)

$$J_n(\underline{w}) = \|\underline{w}^T \underline{x}_n - b_n\|_2^2$$

$$= (\underline{w}^T \underline{x}_n - b_n)^T (\underline{w}^T \underline{x}_n - b_n)$$

$$= x_n^T w w^T x_n - w^T x_n b_n - b_n^T w^T x_n + b_n^T b_n$$

$$\boxed{J_n(\underline{w}) = x_n^T w w^T x_n - 2 w^T x_n b_n + b_n^T b_n}$$

reflected → $J_n(\underline{w}) = x_n^T w w^T x_n z_n - 2 w^T x_n b_n z_n + b_n^T b_n$

$$\nabla_{\underline{w}} J_n(\underline{w}) = 2 x_n^T x_n \underline{w} - 2 x_n b_n$$

$$\nabla_{\underline{w}} J_n(w) = 2 x_n (x_n^T \underline{w} - b_n)$$

$$\boxed{\nabla_{\underline{w}} J_n(w) = 2 \underline{x}_n (\underline{w}^T x_n - b_n)}$$

reflected → $\nabla_w (J_n(w)) = 2 x_n z_n [\underline{w}^T x_n z_n - b_n]$

b)

substituting in

$$w(i+1) = w(i) - \eta(i) \nabla_{\underline{w}} J_n(w)$$

we get,

$$w(i+1) = w(i) - \eta(i) [2 \underline{x}_n (\underline{w}^T(i) \underline{x}_n - b_n]$$

$$= w(i) + \underbrace{2\eta(i)}_{\text{since constant}} \underline{x}_n [\underline{w}^T(i) \underline{x}_n - b_n]$$

$$\boxed{w(i+1) = w(i) + \eta(i)\, \underline{x}_n [b_n - \underline{w}^T(i) \underline{x}_n]}$$

Differences-

1. widrow-Hoff learning algorithm was done for $C=2$, here it is done for $C>2$
2. widrow-hoff was done considering the reflected data points, here it is the actual data points that are considered.