Problem Set 1
CS 6375

Warm Up:

Subgradient & More

(a) $f(x) = \max\{x^2, (x+1)^2\}$ at $x=1$ and $x=0.5$

At $x=1$ ← $\begin{cases} x^2 & 1^2 = 1 \\ (x+1)^2 & (1+1)^2 = 4 \end{cases}$

$\therefore f'(x) = 2(x+1)$

$\therefore$ the subgradient at $x=1$ is $2(x+1) = 4$

At $x=0.5$ ← $\begin{cases} (0.5)^2 & 0.25 \\ (0.5+1)^2 & 2.25 \end{cases}$

$f'(x) = 2(x+1)$

$\therefore$ the subgradient at $x=0.5$ $2(x+1) = 3$

(b) $g(x) = \max\{x^2, \exp(x), 10x\}$ at $x=-1$ and $x=0.0$

At $x=-1$ ← $\begin{cases} x^2 & 1 \\ e^x & e^{-1} \\ 10x & -10 \end{cases}$

$f'(x) = 2x$

∴ the sub gradient of $f(x)$ at $x = -1$ is

$$2x = -2$$

At $x = 0.01$

x^2 $\quad$ 100^{-2}

e^x $\quad$ $e^{0.01}$

$10x$ $\quad$ 0.10

∴ the subgradient, at $x = 0.01$ is $f'(x) = \boxed{10}$

Problem # 1

1. Separability & Feature Vectors

Perceptron loss using subgradient descent

$$w^{(t+1)} = w^{(t)} + y_t \cdot \sum_{i=-y_i t(x^{(i)}) \geq 0} y_i x^{(i)}$$

$$b^{(t+1)} = b^{(t)} + y_t \cdot \sum_{i=-y_i t(x^{(i)}) \geq 0} y_i$$

The w^{t+1} and b^{t+1} will be different from w^t and b^t only when the example/point is misclassified.

If the training set is not linearly seperable

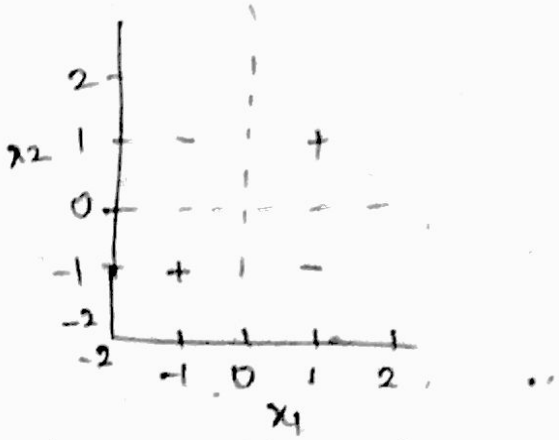
ie no hyperplane can divide it into two seperate classes. we will always have misclassified points/samples that will try to change w^T and b accordingly

Thus we can never find a solution that can seperate the training data.

Therefore, the perceptron loss ~~cost~~ may not converge if the training data is not linearly seperable.

2.

Data set given



As it is not linearly seperable, adding feature vectors will make, the data linerly seperable

Using intution we can see that in the first quater and third quater it is a +ve and other two quater it is negative.

So a hyperbola ie a rectangular hyperbola will transform the data to make it linearly seperable.

∴ $xy = c$ equation of right/rectangular hyperbola

Of the given feature vectors in the question.

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

looks the feature vector.

therefore, now we will find w^T and b. to check if it is converging or not.

$$w^0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad b^0 = 0$$

	x_1	x_2	y
Taking tuple t:	-1	-1	+1

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

$$w^0 \phi + b^0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} +1 \\ +1 \end{bmatrix} + 0 = 0$$

$\therefore$

$$w' = w^0 + y_t \, y_t \, \phi(x_1, x_2) \quad \left[\rightarrow y_t = 1 \right] \text{ for all cases}$$

$$= \begin{bmatrix} 0 \\ 0 \end{bmatrix} + (+1) \begin{bmatrix} +1 \\ +1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$b' = b^0 + y$$

$$= 1$$

Next data input t : $x_1 = -1$, $x_2 = 1$, $y = -1$, $\phi(x_1, x_2) = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

$$w^1 \phi(x_1, x_2) + b^1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} + 1 = 0$$

As it is misclassified.

$$w^2 = w^1 + y_i y \phi(x_1, x_2) \qquad b^2 = b^1 + y_i y$$

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

$$b^2 = 0$$

$$w^2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

Next data input t : $x_1 = 1$, $x_2 = -1$, $y = -1$, $\phi(x_1, x_2) = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

$$w^2 \phi(x_1, x_2) + b^2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} + 0 = -ve$$

As it is correctly classified, there is no need to modify w and b

Next data input t : $x_1 = 1$, $x_2 = 1$, $y = 1$, $\phi(x_1, x_2) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$$w^2 \phi(x_1, x_2) + b^2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + 0 = +ve$$

Again it is correctly classified.

Again we start from top.

	x_1	x_2	y	$\phi(x_1, x_2)$
West point $t=$	-1	-1	$+1$	$\begin{bmatrix}1\\1\end{bmatrix}$

$$w^2\phi(x_1, x_2) + b^2 = \begin{bmatrix}2\\0\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix} + 0 = +ve$$

Correctly classified.

Finally the last point to check.

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

$$w^2\phi(x_1, x_2) + b^2 = \begin{bmatrix}2\\0\end{bmatrix}\begin{bmatrix}-1\\1\end{bmatrix} + 0 = -ve$$

Correctly classified

∴ The linear seperator using ~~function~~ feature variable is

$$w^T\phi(x_1, x_2) + b = 0$$

where $w^T = \begin{bmatrix}2\\0\end{bmatrix}$, $\phi(x_1, x_2) = \begin{bmatrix}x_1 x_2\\1\end{bmatrix}$, $b = 0$

Problem #2 : Perceptron Learning

$$\sum_{i=1}^{n} \max \{0, -y_i (w^T x^{(i)} + b)\}$$

① Standard Gradient Descent

In standard gradient descent we minimize the function by moving in the direction where the loss is less.

In other words, we modify the function such that the total error/loss is decreased.

Therefore in Perceptron loss algorithm:

To minimize the perceptron loss we use the equations

$$w^{(t+1)} = w^{(t)} + y_t \cdot \sum_{i=-y_i f(x^{(i)}) \geq 0} y_i x^{(i)}$$

$$b^{(t+1)} = b^{(t)} + y_t \cdot \sum_{i=-y_i f(x^{(i)}) \geq 0} y_i$$

We start with $w^0 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$ $b^0 = [0]$

and we take y_t (learning rate) = 1

After first iteration

$w^1 = \begin{bmatrix} 124 \\ -113 \\ 216 \\ 5 \end{bmatrix}$ $b^1 = -6$

After second iterations

$w^2 = \begin{bmatrix} 123 \\ -105 \\ 216 \\ -8 \end{bmatrix}$ $b^2 = -7$

After third iteration

$w^3 = \begin{bmatrix} 121 \\ -106 \\ 218 \\ -17 \end{bmatrix}$ $b^3 = -7$

Find iteration.

$w = \begin{bmatrix} 111 \\ -106 \\ 217 \\ -50 \end{bmatrix}$ $b = -3$

Number of iteration taken = 9

Each iteration includes sumation of errors considering the complete data set and 1 iteration to check the convergence

(2) Stochastic Gradient descent

In stochastic gradient descent, instead of taking the whole sum of indices we take a few indices ~~and~~ randomly and compute the sum and then minimize the function accordingly

In this question we are taking data points one after another and repeat accordingly. Therefore the equation used is

$$\left.\begin{aligned} w^{(t+1)} &= w^{(t)} + y_t\, y_i\, x^{(i)} \\ b^{(t+1)} &= b^{(t)} + y_t\, y_i \end{aligned}\right\} \text{if misclassified}$$

$$\left.\begin{aligned} w^{(t+1)} &= w^{t} \\ b^{(t+1)} &= b^{t} \end{aligned}\right\} \text{otherwise}$$

We start with $w^0 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$ $\quad b^0 = [0]$

and y_t (learning rate) $= 1$

After first iteration

$$w^1 = \begin{bmatrix} -3 \\ 2 \\ 4 \\ 3 \end{bmatrix} \quad b^1 = -1$$

After second iteration

$$w^2 = \begin{bmatrix} 1 \\ 1 \\ 4 \\ 6 \end{bmatrix} \quad b^2 = -1$$

After third iteration

$$w^3 = \begin{bmatrix} 5 \\ -2 \\ 6 \\ 2 \end{bmatrix} \quad b^3 = -1$$

Final iteration

$$w = \begin{bmatrix} 7 \\ -4 \\ 11 \\ 0 \end{bmatrix} \quad b = 2$$

Number of iteration taken = 112

Each iteration is taken into consideration at every data point. Total iteration includes 100 iteration of complete classifications.

Problem 3

Support Vector Machines

In SVM we have to

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

such that

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

∴ Using Lagrangian, we get

$$L(w, b, \lambda) = \frac{1}{2} w^T w + \sum_i \lambda_i \left(1 - y_i \left(w^T x^{(i)} + b\right)\right)$$

$$\therefore \frac{\partial L}{\partial w_k} = w_k + \sum_i -\lambda_i y_i x_k^{(i)} = 0$$

$$\therefore \frac{\partial L}{\partial b} = \sum_i -\lambda_i y_i = 0$$

$$\therefore w = \sum_i \lambda_i y_i x_k^{(i)}$$

$$\sum_i \lambda_i y_i = 0$$

∴ substituting we get the dual as

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

such that $$\sum_i \lambda_i y_i = 0$$

Using the kernal trick, we have to take dot product now.

Data set

	x_1	x_2	x_3	y
s_1	0	0	1	1
s_2	-1	-1	-1	1
s_2	1	1	-2	-1

We will use this vectors as they are assumed as support vectors.

Now we augment 1 as a bias input ∴

∴ $s_1 = (0\ 0\ 1\ 1)$

$s_2 = (-1\ -1\ -1\ 1)$

$S_3 = (1\ 1\ -2\ 1)$

$$\therefore \quad \lambda_1 s_1 \cdot s_1 + \lambda_2 s_2 \cdot s_1 + \lambda_3 s_3 \cdot s_1 = 1$$
$$\lambda_1 s_1 \cdot s_2 + \lambda_2 s_2 \cdot s_2 + \lambda_3 s_3 \cdot s_2 = 1$$
$$\lambda_1 s_1 \cdot s_3 + \lambda_2 s_2 \cdot s_3 + \lambda_3 s_3 \cdot s_3 = -1$$

$$\Rightarrow \quad \lambda_1 \cdot 2 + \lambda_2 \cdot 0 + \lambda_3 \cdot (-1) = 1$$
$$\lambda_1 \cdot 0 + \lambda_2 \cdot 4 + \lambda_3 \cdot 1 = 1$$
$$\lambda_1 \cdot (-1) + \lambda_2 \cdot 1 + \lambda_3 \cdot 7 = -1$$

$$\Rightarrow \quad 2\lambda_1 - \lambda_3 = 1$$
$$4\lambda_2 + \lambda_3 = 1$$
$$-\lambda_1 + \lambda_2 + 7\lambda_3 = -1$$

$$\lambda_3 = 2\lambda_1 - 1$$
$$4\lambda_2 + 2\lambda_1 - 1 = 1 \Rightarrow 4\lambda_2 + 2\lambda_1 = 2$$
$$\Rightarrow 2\lambda_2 + \lambda_1 = 1$$
$$\Rightarrow \lambda_2 = \frac{1 - \lambda_1}{2}$$

$$-\lambda_1 + \lambda_2 + 7\lambda_3 = -1$$
$$-\lambda_1 + \frac{(1-\lambda_1)}{2} + 7(2\lambda_1 - 1) = -1 \quad \Big] \times 2$$

$$-2\lambda_1 + (1-\lambda_1) + 14(2\lambda_1 - 1) = -2$$
$$-2\lambda_1 + 1 - \lambda_1 + 28\lambda_1 - 14 = -2$$
$$25\lambda_1 = 11$$
$$\lambda_1 = \frac{11}{25} \qquad \lambda_2 = \frac{14}{250} \qquad \lambda_3 = -\frac{3}{25}$$

As $\vec{w} = \sum_i \lambda_i \vec{s}$

$$= \frac{11}{25}\begin{pmatrix}0\\0\\1\\1\end{pmatrix} + \frac{14}{50}\begin{pmatrix}-1\\-1\\-1\\1\end{pmatrix} + \frac{-3}{25}\begin{pmatrix}1\\1\\-2\\1\end{pmatrix}$$

$$= \begin{pmatrix}\frac{-14-6}{50}\\ \frac{-14-6}{50}\\ \frac{22-14+12}{50}\\ \frac{22+14-6}{50}\end{pmatrix} = \begin{pmatrix}-20/50\\ -20/50\\ 20/50\\ 30/50\end{pmatrix}$$

As are vectors are augmented with a bias.

$$\therefore \quad w = \begin{bmatrix}-2/5\\-2/5\\2/5\end{bmatrix} \qquad b = \begin{bmatrix}+3/5\end{bmatrix}$$

the size of the margin is $\frac{2}{\|w\|}$

$$= \frac{2}{\sqrt{(2/5)^2 + (-2/5)^2 + (2/5)^2}}$$

$$= \frac{2}{\sqrt{3 \times 4/25}} = \frac{2}{2/5\sqrt{3}}$$

$$= \frac{5}{\sqrt{3}} \approx 2.88$$

$\therefore$ the size of the margin is $2.88 \approx 5/\sqrt{3}$

Problem 4:

Lagrange Multipliers and Duality

Primal problem

$$\min_{p_1,\ldots,p_n} \sum_{i=1}^{n} p_i \log p_i$$

such that $p_1, \ldots, p_n \geq 0$ and $\sum_{i=1}^{n} p_i = 1$

Now we have to construct a dual optimization problem.

Lets first take the Lagrangian:

$$L(p, v) = \sum_{i=1}^{n} p_i \log p_i + v\left[\sum_{i=1}^{n} p_i - 1\right]$$

$$\frac{\partial L}{\partial p_k} = \log p_k + 1 + v = 0 \Rightarrow v = -1 - \log p_k$$

$$\Rightarrow p_k = e^{-(1+v)}$$

$$\frac{\partial L}{\partial v} = \sum_{i=1}^{n} p_i - 1 = 0 \Rightarrow \sum_{i=1}^{n} p_i = 1$$

as $v = -1 - \log p_k$

For $k = 1, 2, \ldots n$ v will be same

$\therefore$ $p_1, p_2, p_3, \ldots p_n$ all have to be same

$\therefore p_1 = p_2 = p_3 \ldots = p_n = p$ ← assumed

As we also know

$$\sum_{i=1}^{n} p_i = 1 \Rightarrow p \cdot n = 1$$

$$\Rightarrow p = \frac{1}{n}$$

$$\Rightarrow p_1 = p_2 = \ldots = p_n = p = \frac{1}{n}$$

$\therefore v = -1 - \log \frac{1}{n}$

$= -1 + \log n = (\log n + 1)$

$\Rightarrow v = (\log n + 1)$

$$g(v) = \inf_p L(p, v)$$

$$= \sum_{i=1}^{n} \left[e^{-(1+v)} \log e^{-(1+v)} \right] + e^{-(1+v)} \left[\sum_{i=1}^{n} e^{(-1+v)} - 1 \right]$$

$$= n \left[e^{-(1+v)} \cdot (-1)(1+v) \right] + e^{-(1+v)} \left[n \cdot e^{-(1+v)} - 1 \right]$$

$$g(v) = -n e^{-(1+v)} - vne^{-(1+v)} + n e^{-2(1+v)} - e^{-(1+v)}$$

To solve in $g(v)$ substitute $v = (\log n + 1)$

$$\therefore \cancel{-n e^{-(1+\log n+1)}} \cancel{- (\log n+1) n e^{-(1+v)}}$$

First solve $e^{-(1+v)} = e^{-(1+\log n+1)}$

$$= e^{-(2+\log n)}$$

$$= e^{+(\log (e^2 n)^{-1})}$$

$$= (e^2 n)^{-1}$$

$$\therefore -\cancel{n} \cdot e^{-2} \cancel{n^{-1}} - (\log n + 1) \cdot e^{-2} \cancel{n^{-1}} \cdot \cancel{n} + e^{-2} \cdot n^{-2} \cdot n$$

$$- e^2 n^{-1}$$

$$= -e^{-2} - e^{-2} \log n - e^{-2} + e^{-4} n^{-1} - e^2 n^{-1}$$

$$= -e^{-2} \left[2 + \log n + e^{-2} n^{-1} + \cancel{0} n^{-2} \right]$$

Thus the solution of the primal problem is

$$-e^{-2}\left[2+\log n - e^{-2}n^{-1} + n^{-2}\right]$$