M146 HW2

Elwyn Chu
104977961

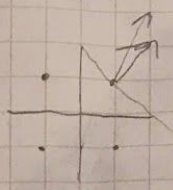
1.a) $y = \vec{w}^T \vec{x} + b$

$x = \begin{bmatrix} +1 \\ +1 \end{bmatrix} \rightarrow y > 0$

$x = \begin{bmatrix} +1 \\ -1 \end{bmatrix} \rightarrow$

$x = \begin{bmatrix} -1 \\ +1 \end{bmatrix} \rightarrow y \leq 0$

$x = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \rightarrow$



$y = \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + b$

choose $b=0$, $w_1 = 1$, $w_2 = +1$

This is not unique!

$b=-1$, $w_1 = 1$, $w_2 = 2$: $y = \begin{bmatrix} 1 & 2 \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - 1$

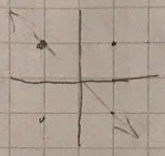
1.b) XOR

$x = \begin{bmatrix} +1 \\ +1 \end{bmatrix} \rightarrow y \leq 0$

$x = \begin{bmatrix} +1 \\ -1 \end{bmatrix} \rightarrow y > 0$

$x = \begin{bmatrix} -1 \\ +1 \end{bmatrix} \rightarrow$

$x = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \rightarrow y > 0$



$y = \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + b$

$w_1 x_1 + w_2 x_2 + b > 0$

~~$w_1 = +1$, $w_2 = -1$, $b = 0$~~

there is no such linear model

(can't seperate w/ a single line)

2.) $\frac{\partial}{\partial w} \frac{1}{1+e^{-w^T x}} = \frac{\partial}{\partial z}(\sigma(z)) = \sigma(z)(1-\sigma(z))$

$z = w^T x \Rightarrow \sigma(w^T x)(1-\sigma(w^T x))$

$y_n \log h_w(x_n) + (1-y_n) \log(1-h_w(x_n))$

$y_n \left(\frac{1}{h_w(x_n)}\right)(\sigma(w^T x)(1-\sigma(w^T x)) + (1-y_n)\left(\frac{1}{1-\sigma(w^T x)}\right)(\sigma(w^T x)(1-\sigma(w^T x))$

~~$\frac{1}{\sigma(w^T x)}$~~ ~~$\sigma(w^T x)(1-\sigma(w^T x))$~~

$= y_n(1-\sigma(w^T x)) - (1-y_n)\sigma(w^T x)$

$= y_n - y_n \sigma(w^T x) - \sigma(w^T x) + y_n \sigma(w^T x)$

$= y_n - \sigma(w^T x)$

$\frac{\partial J}{\partial w_j} = y_n - \frac{1}{1+\exp(w^T x)}$

3a) $\frac{dJ}{dw_1} = \alpha_n (2(w_0 + w_1 x_{n,1} - y_n))(x_{n,1})$

$\frac{\partial J}{\partial w_0} = \alpha_n (2(w_0 + w_1 x_{n,1} - y_n))(1)$

$\nabla J = \left(\alpha_n (2(w_0 + w_1 x_{n,1} - y_n))(x_{n,1}), \alpha_n (2(w_0 + w_1 x_{n,1} - y_n))\right)$

$= \sum 2 \alpha_n (w_0 + w_1 x_{n,1} - y_n) \begin{bmatrix} 1 \\ x_{n,1} \end{bmatrix}$

3 b.) $\sum_{n=1}^{N} 2\alpha_n (w_0 + w_1 x_{n,1} - y_n)[x'_{n,1}] = [0]$

$\sum_{n=1}^{N} \alpha_n (w_0 + w_1 x_{n,1} - y_n) = 0$

i.e. $\sum \alpha_n (w_0 + w_1 x_{n,1} - y_n) = 0$

(1) $\sum \alpha_n w_0 + \sum \alpha_n w_1 x_{n,1} - \sum \alpha_n y_n = 0$

$\sum \alpha_n w_0 + \alpha_n w_1 x_{n,1} = \sum \alpha_n y_n$

$\sum \alpha_n (w_0 + w_1 x_{n,1} - y_n) x_{n,1} = 0$

$\sum \alpha_n x_{n,1} w_0 + \sum \alpha_n x_{n,1} x_{n,1} w_1 - \sum \alpha_n x_{n,1} y_n = 0$

(2) $\sum_{n=1}^{N} \alpha_n x_{n,1} w_0 + \sum_{n=1}^{N} \alpha_n x_{n,1} x_{n,1} w_1 = \sum_{n=1}^{N} \alpha_n x_{n,1} y_n$

matrix:

$$\begin{bmatrix} \sum_{n=1}^{N} \alpha_n & \sum_{n=1}^{N} \alpha_n x_{n,1} \\ \sum_{n=1}^{N} \alpha_n x_{n,1} & \sum_{n=1}^{N} \alpha_n x_{n,1} x_{n,1} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} \sum_{n=1}^{N} \alpha_n y_i \\ \sum_{n=1}^{N} \alpha_n y_n x_{n,1} \end{bmatrix}$$

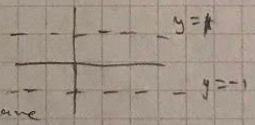
solve using $\vec{w} = A^{-1} B$

$$\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} \sum_{n=1}^{N} \alpha_n & \sum_{n=1}^{N} \alpha_n x_{n,1} \\ \sum_{n=1}^{N} \alpha_n x_{n,1} & \sum_{n=1}^{N} \alpha_n x_{n,1} x_{n,1} \end{bmatrix}^{-1} \begin{bmatrix} \sum_{n=1}^{N} \alpha_n y_n \\ \sum_{n=1}^{N} \alpha_n y_n x_{n,1} \end{bmatrix}$$

4a) given:
D is linearly seperable

Then: $y_i = \begin{cases} 1 & \vec{w}^T \vec{x}_i + b \geq 0 \\ -1 & \vec{w}^T \vec{x}_i + b < 0 \end{cases}$

$x_i \neq 0$



Define subplane $w^T x_i + b$

where $y_i \in \{-1, 1\}$

then, if $b=0$, the hyperplane lies on the datapoint (x_i, y_i)

Then we shift the hyperplane by some $b = \varepsilon$

$w^T x_i + \varepsilon$ defines our new hyperplane

and $y_i (w^T x_i + \varepsilon) = 1 - \delta = \varepsilon$

now we normalize by dividing by ε

$$\frac{y_i (w^T x_i + \varepsilon)}{\varepsilon} = 1$$

and $y_i \left(\frac{w^T x_i}{\varepsilon} + 1 \right) = 1 \geq 1 - \delta$ w/ $\delta = 0$

thus, we have an optimal solution w/ $\delta = 0$

tubing $\vec{w} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\varepsilon = -$

$x_i = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \rightarrow y_i \left(\frac{\quad}{0.1} \right) = 1$

4b.) Given: there is an optimal solution w/ $\delta = 0$

Then we know there is some $\vec{w}$ and $\vec{b}$ such that

$y_i (\vec{w}^T \vec{x}_i + \vec{b}) \geq 1$, then we know that for any quantity $(\vec{w}^T \vec{x}_i + \vec{b}) < 0$ (negative), y_i is also negative and for any $(\vec{w}^T \vec{x}_i + \vec{b}) > 0$, y_i is also positive.

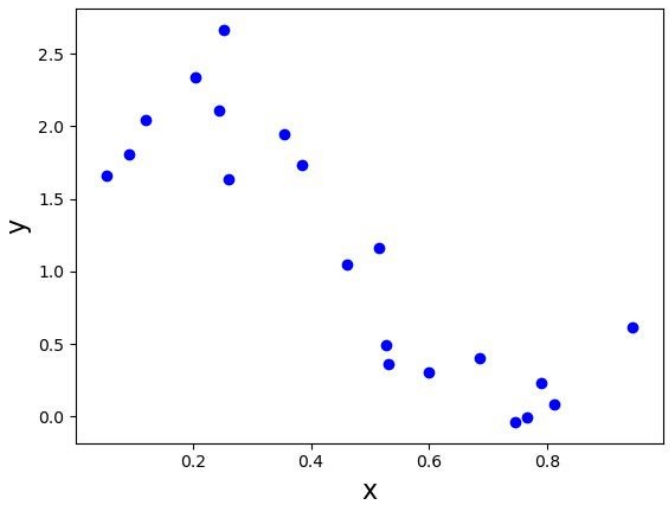
Thus, D is linearly seperable since y_i can be written as $y_i = \begin{cases} 1 & w^T x_i + b \geq 0 \\ -1 & w^T x_i + b < 0 \end{cases}$

4c.) If we have a hyperplane that satisfies condition a w/ $\delta > 0$, then we can say that there also exists a hyper plane w/ $\delta = 0$, through normalization and thus that the data is linearly seperable.

4d.) Optimal solution independent of D is $\vec{w} = \vec{0}$, $\theta = 0$, $\delta = 0$.
This does not give us a hyper plane at all, so this is not a good formulation.

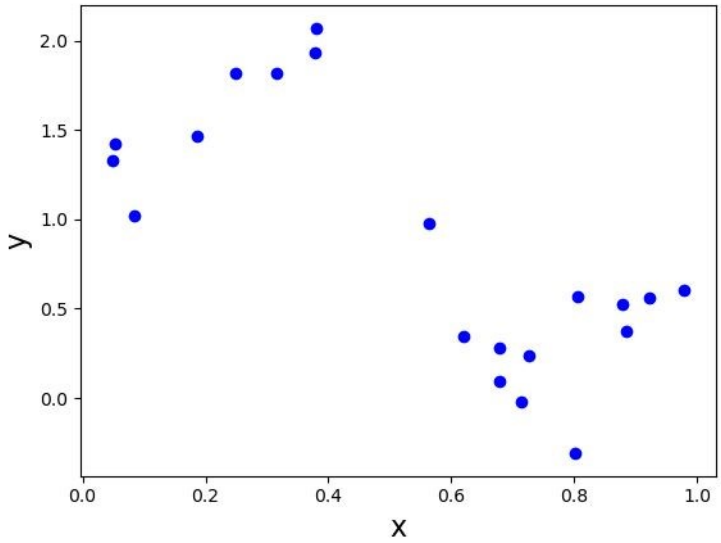
4e.) Given D has only 2 points, we know it is linearly seperable. Possible optimal solutions is $w_1 + w_2 \cdots w_n \geq 1 + \theta$

5a.) Train data plot:



We can see that the train data looks somewhat linear, so linear regression seems like it would be fairly effective in predicting the data.

Test data plot:



This data looks somewhat linear, although less so than the train data plot. It would be fair to say that linear regression would be less effective in predicting the data.

5d.)

Coefficient	Iterations to Convergence	Cost	Coefficients
0.01	765	3.9125	(2.4464, -2.8163)
0.001	7021	3.9125	(2.4464, -2.8163)
0.0001	10000 (does not converge)	4.0864	(2.27045, -2.46065)
0.0407	10000 (does not converge)	2.7109e39	(-9.4047e39, -4.6523e18)

5e.) The closed form solution is (2.44640709, -2.81635359). The closed form solution runs almost 3x as fast as the gradient descent. The coefficients are nearly the same for when gradient descent converges. The costs are the same since they both

5f.) It takes the algorithm 1679 iterations to converge with the proposed learning rate.

5h.) We might prefer RMSE as a metric over the cost because while an overfit model may seem accurate on the training data and the cost will continuously decrease, the RMSE will reach a minimum then as the model becomes more overfit, increase again. Thus minimizing RMSE will allow one to minimize cost while not overfitting the model..