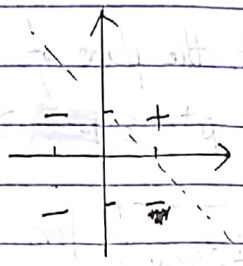
CS M146
Homework 2
Yongqian Li
004997466
1A

1. (a).

x_1	x_2	y
1	1	1
1	-1	-1
-1	1	-1
-1	-1	-1

(AND)



[1] let $W^T = [w_1, w_2]$, b

the prediction is $y_i = \text{sgn}(W^T x + b)$

$W^T x + b = w_1 x_1 + w_2 x_2 + b$

$$\begin{cases} w_1 + w_2 + b \geq 0 \\ w_1 - w_2 + b < 0 \\ -w_1 + w_2 + b < 0 \\ -w_1 - w_2 + b < 0 \end{cases}$$

∴ one solution is $\begin{cases} W^T = [1, 1] \\ b = -1 \end{cases}$

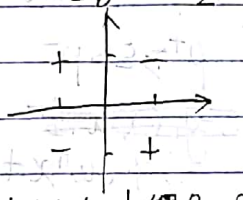
another valid linear model is

$$\begin{cases} W^T = [1, 2] \\ b = -2 \end{cases}$$

(b).

x_1	x_2	y
1	1	-1
1	-1	1
-1	1	1
-1	-1	-1

(XOR)



$$\begin{cases} w_1 + w_2 + b < 0 & ① \\ w_1 - w_2 + b \geq 0 & ② \\ -w_1 + w_2 + b \geq 0 & ③ \\ -w_1 - w_2 + b < 0 & ④ \end{cases}$$

① + ④ $\Rightarrow 2b < 0$

$\therefore b < 0$

② + ③ $\Rightarrow 2b \geq 0 \Rightarrow b \geq 0$

∴ There is no solution.

No such linear model exists because the data is not linearly separable.

2. (a). $\sigma(W^T x_n) = \frac{1}{1+e^{-W^T x_n}}$, $\therefore 1 - \frac{1}{1+e^{-W^T x_n}} = \frac{1+e^{-W^T x_n} - 1}{1+e^{-W^T x_n}} = \frac{e^{-W^T x_n}}{1+e^{-W^T x_n}}$

$$J(w) = -\sum_{n=1}^{N} [y_n (\log(1) - \log(1+e^{-W^T x_n})) + (1-y_n)(\log(e^{-W^T x_n}) - \log(1+e^{-W^T x_n}))]$$

$$\therefore J(w) = -\sum_{n=1}^{N} [y_n (0 - \log(1+e^{-W^T x_n})) + (1-y_n)(-W^T x_n - \log(1+e^{-W^T x_n}))]$$

$$= -\sum_{n=1}^{N} [y_n \log(1+e^{-W^T x_n}) - W^T x_n - \log(1+e^{-W^T x_n}) + y_n W^T x_n + y_n \log(1+e^{-W^T x_n})]$$

$$= -\sum_{n=1}^{N} [-W^T x_n - \log(1+e^{-W^T x_n}) + y_n W^T x_n]$$

(b). $W^T x_n = w_1 x_{n,1} + w_2 x_{n,2} + \cdots$

$$\frac{\partial (W^T x_n)}{\partial w_j} = x_{n,j}$$

$$\therefore \frac{\partial J(w)}{\partial w_j} = -\sum_{n=1}^{N} [-x_{n,j} - \frac{x_{n,j} e^{-W^T x_n}}{1+e^{-W^T x_n}} + y_n x_{n,j}]$$

$$= -\sum_{n=1}^{N} [\frac{-x_{n,j} - x_{n,j} e^{-W^T x_n} + x_{n,j} e^{-W^T x_n}}{1+e^{-W^T x_n}} + y_n x_{n,j}]$$

$$= \sum_{n=1}^{N} [\frac{x_{n,j}}{1+e^{-W^T x_n}} - y_n x_{n,j}]$$

$$= \sum_{n=1}^{N} [\frac{1}{1+e^{-W^T x_n}} - y_n] x_{n,j}$$

3. (a). $\frac{\partial J}{\partial w_0} = \sum_{n=1}^{N} 2\alpha_n (w_0 + w_1 x_{n,1} - y_n) \cdot 1$

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

$$\therefore \nabla J = [\sum_{n=1}^{N} 2\alpha_n (w_0 + w_1 x_{n,1} - y_n), \sum_{n=1}^{N} 2\alpha_n (w_0 + w_1 x_{n,1} - y_n) x_{n,1}]$$

(b). $\begin{cases} \frac{\partial J}{\partial w_0} = 0 \\ \frac{\partial J}{\partial w_1} = 0 \end{cases} \Leftrightarrow \begin{cases} \sum_{n=1}^{N} 2\alpha_n (w_0 + w_1 x_{n,1} - y_n) = 0 & ① \\ \sum_{n=1}^{N} 2\alpha_n (w_0 + w_1 x_{n,1} - y_n) x_{n,1} = 0 & ② \end{cases}$

$$① \Leftrightarrow \sum_{n=1}^{N} 2\alpha_n w_0 + \sum_{n=1}^{N} 2\alpha_n w_1 x_{n,1} - \sum_{n=1}^{N} 2\alpha_n y_n = 0$$

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

$$\therefore w_0 = \frac{-\sum_{n=1}^{N} \alpha_n w_1 x_{n,1} + \sum_{n=1}^{N} \alpha_n y_n}{\sum_{n=1}^{N} \alpha_n}$$

$$② \Leftrightarrow \sum_{n=1}^{N} 2\alpha_n w_0 x_{n,1} + \sum_{n=1}^{N} 2\alpha_n w_1 x_{n,1}^2 - \sum_{n=1}^{N} 2\alpha_n y_n x_{n,1} = 0$$

$$\therefore \sum_{n=1}^{N} \alpha_n w_0 x_{n,1} = w_0 \sum_{n=1}^{N} \alpha_n x_{n,1}$$

$\therefore -\frac{\sum_{n=1}^{N}\alpha_n x_{n,1}}{\sum_{n=1}^{N}\alpha_n} w_1 + \frac{\sum_{n=1}^{N}\alpha_n y_n}{\sum_{n=1}^{N}\alpha_n} \cdot \sum_{n=1}^{N}\alpha_n x_{n,1} + w_1 \sum_{n=1}^{N}\alpha_n x_{n,1}^2$

$= \sum_{n=1}^{N} 2\alpha_n y_n x_{n,1}$

$-w_1 \frac{\sum_{n=1}^{N}\alpha_n x_{n,1}}{\sum_{n=1}^{N}\alpha_n} \cdot \frac{\sum_{n=1}^{N}\alpha_n x_{n,1}}{\sum_{n=1}^{N}\alpha_n} + w_1 \sum_{n=1}^{N}\alpha_n x_{n,1}^2 = \sum_{n=1}^{N}\alpha_n y_n x_{n,1} + \frac{\sum_{n=1}^{N}\alpha_n y_n}{\sum_{n=1}^{N}\alpha_n} \cdot \sum_{n=1}^{N}\alpha_n x_{n,1}$

$-w_1\left(\frac{(\sum_{n=1}^{N}\alpha_n x_{n,1})^2}{\sum_{n=1}^{N}\alpha_n} + \sum_{n=1}^{N}\alpha_n x_{n,1}^2\right) = RHS$

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

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

shift the plane of $(w^T x + b)$ by ε:

$\rho^+ - \varepsilon > \rho^- - \varepsilon$

for $(w^T x + (b-\varepsilon))$;

$\rho^+_{new} = \min_{i \in \text{positive } y_i} (w^T x + (b-\varepsilon))$

$= \rho^+ - \varepsilon$

$= \frac{\rho^+ - \rho^-}{2}$

$\rho^-_{new} = \max_{i \in \text{negative } y_i} (w^T x + (b-\varepsilon))$

$= \rho^- - \varepsilon$

$= \frac{\rho^- + \rho^+}{2}$

$\therefore \rho^+_{new} \geq 0, \ \rho^-_{new} < 0$

$\therefore \rho^+ - \varepsilon \geq 0 > \rho^- - \varepsilon$

$\rho^+ \geq \varepsilon > \rho^-$

$\therefore y_i (w^T x + b) \geq \varepsilon$

$\frac{y_i (w^T x + b)}{\varepsilon} \geq 1$

define $w_1^T = \frac{w^T}{\varepsilon}$, $b_1 = \frac{b}{\varepsilon}$

then $y_i (w_1^T x + b_1) \geq 1 - 0$

$\therefore$ optimal $\delta = 0$

4. (a). define $\rho^+ = \min_{i \in \text{positive } y_i} (w^T x_i + b)$,

$\rho^- = \max_{j \in \text{negative } y_j} (w^T x_j + b)$,

($\because D$ is linear separable and satisfies (1))

by (1), $\rho^+ \geq 0 > \rho^-$

let $\varepsilon > 0$ be the solution of $\frac{|\rho^+ - \varepsilon|}{\|w\|} = \frac{|\rho^- - \varepsilon|}{\|w\|}$

$\Leftrightarrow \frac{\rho^+ - \varepsilon}{\|w\|} = \frac{-\rho^- + \varepsilon}{\|w\|}$

$\Rightarrow \varepsilon = \frac{\rho^+ + \rho^-}{2}$

(b). If there is an optimal solution with $\delta = 0$,

$y_i (w^T x_i + b) \geq 1 - 0 = 1$

$\therefore$ when $y_i = 1$, $w^T x_i + b \geq 1 > 0$

when $y_i = -1$, $w^T x_i + b \leq -1 < 0$

$\therefore y_i = \begin{cases} 1 & \text{if } w^T x_i + b \geq 0 \\ -1 & \text{if } w^T x_i + b < 0 \end{cases}$

$\therefore$ it satisfies condition (1)

$\therefore D$ is linearly separable

(c). when $0 < \delta < 1$,

$y_i = 1 \Rightarrow w^T x_i + b \geq 1 - \delta > 0$

$y_i = -1 \Rightarrow w^T x_i + b < -(1-\delta) < 0$

condition (1) is still satisfied.

when $\delta \geq 1$, $1-\delta \leq 0$

$y_i = 1 \Rightarrow w^T x_i + b \geq 1-\delta \leq 0$

$y_i = -1 \Rightarrow w^T x_i + b \leq -(1-\delta) \geq 0$

(1) may not be satisfied

∴ It is uncertain.

∴ when $\delta \in (0,1)$, D is still linearly saparable; when $\delta \geq 1$, D may or may not be linearly separable.

(d). The optimal solution is: $\delta = 0$,

$\vec{w}^T = \vec{0}$, $b = 0$

So the problem is that this formulation may be satisfied even when there is no separating hyperplane.

(e). let $w^T = [w_1\ w_2\ w_3]$,

then (2) $\Rightarrow \min \delta$

subject to $y_i(w_1 x_1 + w_2 x_2 + w_3 x_3 + b) \geq 1-\delta$,

$\forall (x_i, y_i) \in D$, $\delta \geq 0$

$\delta > 0$

$$\therefore \begin{cases} w_1 + w_2 + w_3 + b \geq 1-\delta & ① \\ -(-w_1 - w_2 - w_3 + b) \geq 1-\delta & ② \end{cases}$$

② $\Leftrightarrow w_1 + w_2 + w_3 - b \geq 1-\delta$ ③

when $\delta = 0$, ① + ② $\Rightarrow 2(w_1+w_2+w_3) \geq 2$

$w_1 + w_2 + w_3 \geq 1$

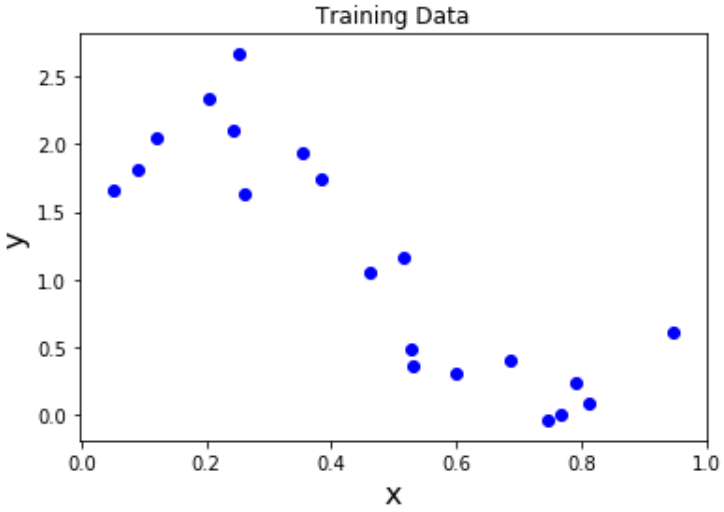
$\therefore w_1 + w_2 + w_3 \pm b \geq 1$

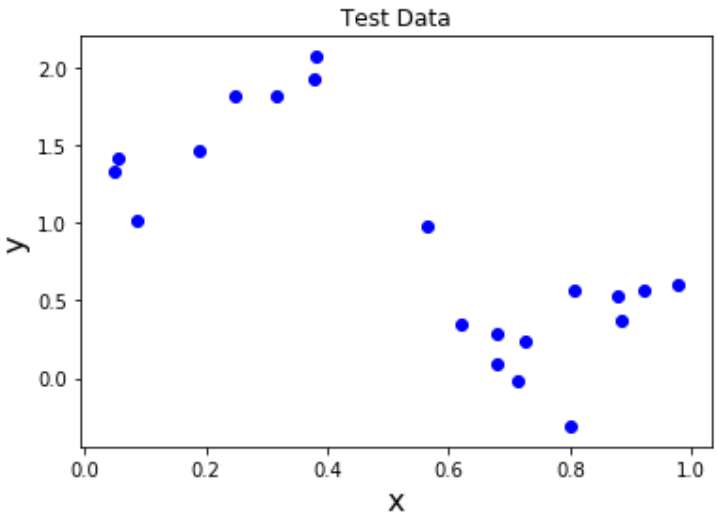
$\therefore w_1 + w_2 + w_3 \geq 1 \pm b$

$\therefore w_1 + w_2 + w_3 \geq 1 + |b|$

The optimal solutions are $w^T = [w_1, w_2, w_3]$ and b satisfying $w_1 + w_2 + w_3 \geq 1 + |b|$ at $\delta = 0$.

5(a).





I observe that the distributions of training data and test data are similar. However, in the training data, there is a clear negative linear correlation, but it is hard to find in the test data. Therefore, my

educated guess is that the linear regression from the training data may not generalize well on the test data.

(b). In regression.py

(c). In regression.py

(d). Implemented in regression.py

η	time	iterations	cost
0.0001	2.0966780185699463	10000	4.0863970
0.001	1.4808697700500488	7020	
0.01	0.20312237739562988	764	
0.0407	2.1020712852478027	10000	

η	time	#iterations	cost	coefficients ([w0, w1])
0.0001	2.09668	10000	4.08640	[2.2044798 -2.4064834]
0.001	1.48087	7020	3.91258	[2.4464058 -2.816353]
0.01	0.20312	764	3.91258	[2.44640703 -2.81635546]
0.0407	2.10207	10000	2.71092e+39	[-9.4047081e+18 -4.6822095e+18]

For $\eta = 10^{-4}$ and $\eta = 0.0407$, GD does not converge in 10000 steps. The reason is that 10^{-4} is too small while 0.0407 is too large, so for 10^{-4} it converging too slowly while for 0.0407 it is unstable. From $\eta = 10^{-4}$, $\eta = 10^{-3}$ to $\eta = 10^{-2}$, the number of iterations needed and the time used decrease. The coefficients of $\eta = 10^{-3}$ and $\eta = 10^{-2}$ are largely the same, since in these to cases the iterations has finished within 10000 steps. The coefficients of $\eta = 10^{-4}$ is close to them, and it is less accurate because the iterations was terminated before convergence. The coefficients of $\eta = 0.0407$ does not make sense.

(e). Implemented in regression.py.

The closed-form solution is: 3.91258

$[w_0, w_1] = [2.44640709, -2.81635359]$, cost = results

almost the same as the of GD when convergence is reached.

Both the $[w_0, w_1]$ and the cost are

Its time ≈ 0.0, so it is much faster than GD.

(f). Implemented in regression.py

time	cost	coefficients ([w0, w1])
0.26563	3.91258	[2.44640676, -2.81635297]

with # iterations = 1356

Both w and the cost are the same as the previous correct results, and its # iterations and time is between $\eta = 10^{-3}$ and $\eta = 10^{-2}$.

(g). In regression.py

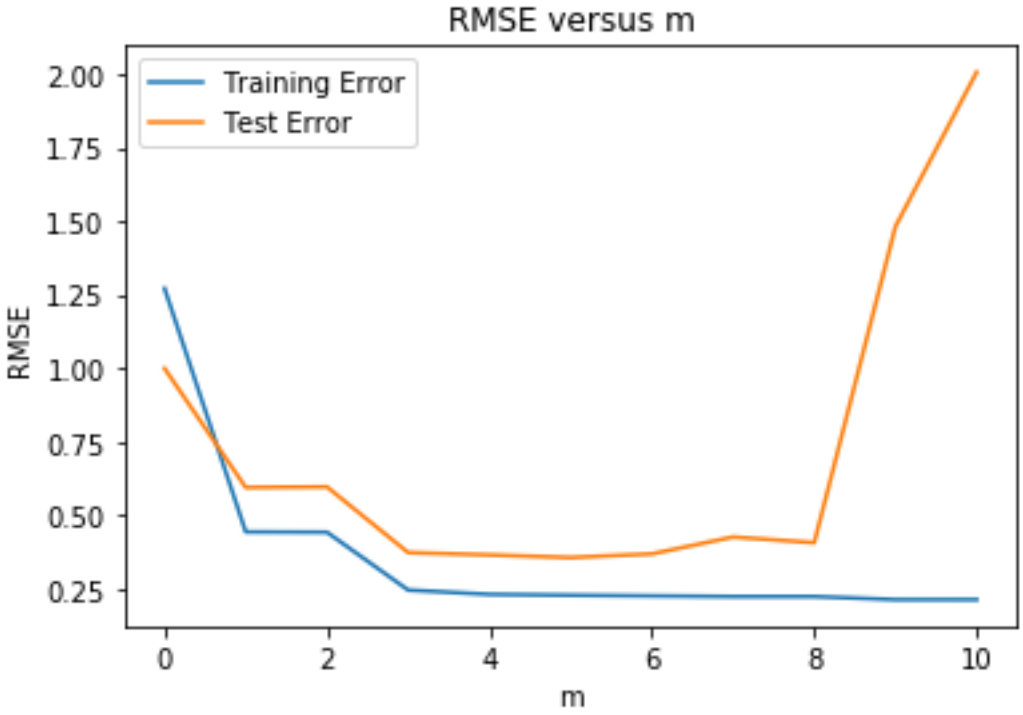
(h). Implemented in regression.py

Using RMSE can get rid of the influence of the size of data on the error, so errors of different models can be compared. It normalizes the error.

(i). As shown in the following plot, degrees of 4, 5 and 6 best fit the data. These values minimize the test error, and the training error is also very low.

When $m \leq 3$, both training error and test error are larger, indicating underfitting.

When $m \geq 8$, the training error continue to decrease while the test error increases dramatically, indicating overfitting.


RMSE versus m
Training Error
Test Error
RMSE
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0
2
4
6
8
10
m