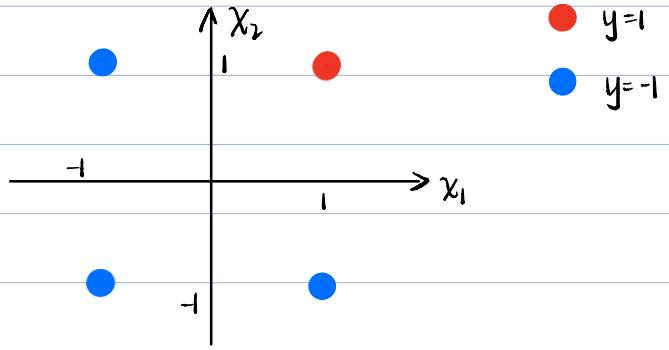
005230642 Rui Xu

1.(a). AND

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



A valid linear model exists.

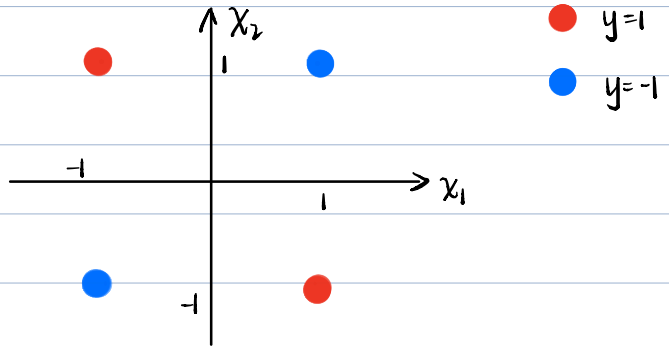
$x_1 + x_2 = \frac{3}{2} \Rightarrow \vec{w} = [1, 1]^T \; b = -\frac{3}{2} \Rightarrow y = \text{sign}(x_1 + x_2 - \frac{3}{2})$

The valid linear model is not unique. Another valid linear model:

$x_1 + x_2 = \frac{5}{4} \Rightarrow \vec{w} = [1, 1]^T \; b = -\frac{5}{4} \Rightarrow y = \text{sign}(x_1 + x_2 - \frac{5}{4})$

(b). XOR

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



No such linear model exists because XOR is not linearly seperable.

Assume a valid linear model exists:

$\vec{w} = [w_1, w_2]^T, b \;.\; y_i = \text{sign}(\vec{w}^T \vec{x_i} + b) \quad i = 1, 2, 3, 4 \Rightarrow y_i(\vec{w}^T \vec{x_i} + b) \geq 0$

$$\Rightarrow \begin{cases} (-1)(w_1 + w_2 + b) \geq 0 \\ (1)(w_1 - w_2 + b) \geq 0 \\ (1)(-w_1 + w_2 + b) \geq 0 \\ (-1)(-w_1 - w_2 + b) \geq 0 \end{cases} \Rightarrow \begin{cases} w_1 + w_2 + b \leq 0 & ① \\ w_1 - w_2 + b \geq 0 & ② \\ -w_1 + w_2 + b \geq 0 & ③ \\ -w_1 - w_2 + b \leq 0 & ④ \end{cases}$$

$① + ④ \Rightarrow \begin{cases} b \leq 0 \\ b \geq 0 \end{cases}$ $\quad ② + ③$

$\Rightarrow b = 0$

$$\Rightarrow \begin{cases} w_1 + w_2 \le 0 \\ w_1 - w_2 \ge 0 \\ -w_1 + w_2 \ge 0 \\ -w_1 - w_2 \le 0 \end{cases} \Rightarrow \begin{cases} w_1 + w_2 \le 0 \\ w_1 - w_2 \ge 0 \\ w_1 - w_2 \le 0 \\ w_1 + w_2 \ge 0 \end{cases} \Rightarrow \begin{cases} w_1 + w_2 = 0 \\ w_1 - w_2 = 0 \end{cases} \Rightarrow \begin{cases} w_1 = 0 \\ w_2 = 0 \end{cases}$$

$\Rightarrow \vec{w} = [0 \ 0]^T \quad b = 0 \Rightarrow y_i = \text{sign}(\vec{w}^T \vec{x_i} + b) = \text{sign}(0) = 1 \quad , i = 1, 2, 3, 4$

$\Rightarrow$ This is not a valid linear model

$\Rightarrow$ No such linear model exists.

2.(a). $J(w) = -\sum_{n=1}^{N} [y_n \log h_w(x_n) + (1-y_n) \log (1-h_w(x_n))]$

$$\frac{\partial J}{\partial w_j} = \frac{\partial \{-\sum_{n=1}^{N} [y_n \log h_w(x_n) + (1-y_n) \log (1-h_w(x_n))]\}}{\partial w_j}$$

$$= -\sum_{n=1}^{N} \frac{\partial [y_n \log h_w(x_n) + (1-y_n) \log (1-h_w(x_n))]}{\partial w_j}$$

$$= -\sum_{n=1}^{N} \left(y_n \frac{\partial \log h_w(x_n)}{\partial w_j} + (1-y_n) \frac{\partial \log (1-h_w(x_n))}{\partial w_j} \right)$$

$$= -\sum_{n=1}^{N} \left(y_n \frac{1}{h_w(x_n)} \frac{\partial h_w(x_n)}{\partial w_j} + (1-y_n) \frac{-1}{1-h_w(x_n)} \frac{\partial h_w(x_n)}{\partial w_j} \right)$$

$$\frac{\partial h_w(x_n)}{\partial w_j} = \frac{\partial \alpha(w^T x_n)}{\partial w_j} = \frac{\partial \alpha(w^T x_n)}{\partial (w^T x_n)} \cdot \frac{\partial (w^T x_n)}{\partial w_j} = \frac{\partial \alpha(w^T x_n)}{\partial (w^T x_n)} \cdot x_{n,j}$$

$$\frac{\partial \alpha(w^T x_n)}{\partial (w^T x_n)} = \alpha(w^T x_n)(1 - \alpha(w^T x_n)) = h_w(x_n)(1 - h_w(x_n))$$

$$\Rightarrow \frac{\partial h_w(x_n)}{\partial w_j} = \frac{\partial \alpha(w^T x_n)}{\partial (w^T x_n)} \cdot x_{n,j} = h_w(x_n)(1-h_w(x_n)) x_{n,j}$$

$$\Rightarrow \frac{\partial J}{\partial w_j} = -\sum_{n=1}^{N} \Big(y_n \frac{1}{h_w(x_n)} \cdot h_w(x_n)(1-h_w(x_n)) x_{n,j} + (1-y_n) \frac{-1}{1-h_w(x_n)} h_w(x_n)(1-h_w(x_n)) x_{n,j} \Big)$$

$$= -\sum_{n=1}^{N} \left(y_n (1-h_w(x_n)) x_{n,j} - (1-y_n) h_w(x_n) x_{n,j} \right)$$

$$= -\sum_{n=1}^{N} [(y_n - h_w(x_n)) x_{n,j}]$$

$$= \sum_{n=1}^{N} [(h_w(x_n) - y_n) x_{n,j}]$$

$$\Rightarrow \frac{\partial J}{\partial w_j} = \sum_{n=1}^{N} [(h_w(x_n) - y_n) x_{n,j}]$$

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

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

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

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

(b). $\frac{\partial J}{\partial w_0} = 2 \sum_{n=1}^{N} a_n (w_0 + w_1 x_{n,1} - y_n) = 0 \Rightarrow \sum_{n=1}^{N} a_n (w_0 + w_1 x_{n,1} - y_n) = 0$

$\Rightarrow \sum_{n=1}^{N} a_n w_0 + \sum_{n=1}^{N} a_n w_1 x_{n,1} - \sum_{n=1}^{N} a_n y_n = 0$

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

$\frac{\partial J}{\partial w_1} = 2 \sum_{n=1}^{N} a_n x_{n,1} (w_0 + w_1 x_{n,1} - y_n) = 0 \Rightarrow \sum_{n=1}^{N} a_n x_{n,1} (w_0 + w_1 x_{n,1} - y_n) = 0$

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

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

$\Rightarrow \begin{cases} (\sum_{n=1}^{N} a_n) w_0 + (\sum_{n=1}^{N} a_n x_{n,1}) w_1 = \sum_{n=1}^{N} a_n y_n & ① \\ (\sum_{n=1}^{N} a_n x_{n,1}) w_0 + (\sum_{n=1}^{N} a_n x_{n,1}^2) w_1 = \sum_{n=1}^{N} a_n x_{n,1} y_n & ② \end{cases}$

For simplicity, we define $\bar{x} = \frac{\sum_{n=1}^{N} a_n x_{n,1}}{\sum_{n=1}^{N} a_n}$

$\bar{y} = \frac{\sum_{n=1}^{N} a_n y_n}{\sum_{n=1}^{N} a_n}$

$\overline{x^2} = \frac{\sum_{n=1}^{N} a_n x_{n,1}^2}{\sum_{n=1}^{N} a_n}$

$\overline{xy} = \frac{\sum_{n=1}^{N} a_n x_{n,1} y_n}{\sum_{n=1}^{N} a_n}$

$\Rightarrow \frac{①}{\sum_{n=1}^{N} a_n} \Rightarrow w_0 + \bar{x} w_1 = \bar{y}$

$\frac{②}{\sum_{n=1}^{N} a_n} \Rightarrow \bar{x} w_0 + \overline{x^2} w_1 = \overline{xy}$

$$\Rightarrow \begin{vmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{vmatrix} = \overline{x^2} - (\bar{x})^2 = \frac{\sum_{n=1}^{N} \partial_n x_{n,1}^2}{\sum_{n=1}^{N} \partial_n} - \left(\frac{\sum_{n=1}^{N} \partial_n x_{n,1}}{\sum_{n=1}^{N} \partial_n}\right)^2 \neq 0$$

$$\Rightarrow w_0 = \frac{\begin{vmatrix} \bar{y} & \bar{x} \\ \overline{xy} & \overline{x^2} \end{vmatrix}}{\overline{x^2} - (\bar{x})^2} = \frac{\overline{x^2}\,\bar{y} - \bar{x}\,\overline{xy}}{\overline{x^2} - (\bar{x})^2}$$

$$w_1 = \frac{\begin{vmatrix} 1 & \bar{y} \\ \bar{x} & \overline{xy} \end{vmatrix}}{\overline{x^2} - (\bar{x})^2} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}$$

$$\Rightarrow \begin{cases} w_0 = \dfrac{\overline{x^2}\,\bar{y} - \bar{x}\,\overline{xy}}{\overline{x^2} - (\bar{x})^2} \\ w_1 = \dfrac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - (\bar{x})^2} \end{cases}$$

$$\bar{x} = \frac{\sum_{n=1}^{N} \partial_n x_{n,1}}{\sum_{n=1}^{N} \partial_n}$$

$$\bar{y} = \frac{\sum_{n=1}^{N} \partial_n y_n}{\sum_{n=1}^{N} \partial_n}$$

$$\overline{x^2} = \frac{\sum_{n=1}^{N} \partial_n x_{n,1}^2}{\sum_{n=1}^{N} \partial_n}$$

$$\overline{xy} = \frac{\sum_{n=1}^{N} \partial_n x_{n,1} y_n}{\sum_{n=1}^{N} \partial_n}$$

4.(a). D is linearly separable $\Rightarrow$ There exists $\vec{w}, b$ such that

$$y_i = \begin{cases} 1 & \text{if } \vec{w}^T\vec{x_i} + b \geq 0 \\ -1 & \text{if } \vec{w}^T\vec{x_i} + b < 0 \end{cases} \qquad \forall (\vec{x_i}, y_i) \in D$$

Define two subsets, S_1, S_2, such that $S_1 \cup S_2 = D$, $S_1 \cap S_2 = \phi$ and

$$\begin{cases} \vec{w}^T\vec{x_i} + b \geq 0 & y_i = 1 \quad \forall (\vec{x_i}, y_i) \in S_1 \\ \vec{w}^T\vec{x_i} + b < 0 & y_i = -1 \quad \forall (\vec{x_i}, y_i) \in S_2 \end{cases}$$

$|S_2| \leq m$ is finite $\Rightarrow$ There exists a small enough $\varepsilon > 0$ such that

$$\begin{cases} \vec{w}^T\vec{x_i} + b \geq 0 & y_i = 1 \quad \forall (\vec{x_i}, y_i) \in S_1 \\ \vec{w}^T\vec{x_i} + b \leq -\varepsilon & y_i = -1 \quad \forall (\vec{x_i}, y_i) \in S_2 \end{cases}$$

$$\Rightarrow \begin{cases} \vec{w}^T\vec{x_i} + b + \frac{\varepsilon}{2} \geq \frac{\varepsilon}{2} & y_i = 1 \quad \forall (\vec{x_i}, y_i) \in S_1 \\ \vec{w}^T\vec{x_i} + b + \frac{\varepsilon}{2} \leq -\frac{\varepsilon}{2} & y_i = -1 \quad \forall (\vec{x_i}, y_i) \in S_2 \end{cases}$$

$$\Rightarrow \begin{cases} \frac{2}{\varepsilon}\vec{w}^T\vec{x_i} + \frac{2}{\varepsilon}b + 1 \geq 1 & y_i = 1 \quad \forall (\vec{x_i}, y_i) \in S_1 \\ \frac{2}{\varepsilon}\vec{w}^T\vec{x_i} + \frac{2}{\varepsilon}b + 1 \leq -1 & y_i = -1 \quad \forall (\vec{x_i}, y_i) \in S_2 \end{cases}$$

$$\Rightarrow y_i\left(\frac{2}{\varepsilon}\vec{w}^T\vec{x_i} + \frac{2}{\varepsilon}b + 1\right) \geq 1 \quad \forall (\vec{x_i}, y_i) \in D$$

Let $\vec{w}' = \frac{2}{\varepsilon}\vec{w}$, $b' = \frac{2}{\varepsilon}b + 1 \Rightarrow y_i(\vec{w}'^T\vec{x_i} + b') \geq 1$

$\Rightarrow y_i(\vec{w}'^T\vec{x_i} + b) \geq 1 - \delta$, δ can be 0

$\Rightarrow$ There is an optimal solution $\vec{w}', b'$ to the linear program (2) with $\delta = 0$

(b). There is an optimal solution with $\delta = 0 \Rightarrow$ There exists $\vec{w}, b$ such that

$$y_i(\vec{w}^T\vec{x_i} + b) \geq 1 \quad \forall (\vec{x_i}, y_i) \in D$$

$$\Rightarrow \begin{cases} \text{if } y_i = 1 \Rightarrow \vec{w}^T\vec{x_i} + b \geq 1 \geq 0 \\ \text{if } y_i = -1 \Rightarrow \vec{w}^T\vec{x_i} + b \leq -1 < 0 \end{cases} \Rightarrow y_i = \begin{cases} 1 & \text{if } \vec{w}^T\vec{x_i} + b \geq 0, \quad \forall (\vec{x_i}, y_i) \in D \\ -1 & \text{if } \vec{w}^T\vec{x_i} + b < 0 \end{cases}$$

$\Rightarrow D$ satifies condition (1) $\Rightarrow D$ is linearly separable

(d). ① If $1-\delta > 0 \Rightarrow 0 < \delta < 1$

$\Rightarrow y_i(\vec{w}^T\vec{x_i} + b) \geq 1-\delta > 0, \forall (\vec{x_i}, y_i) \in D \Rightarrow y_i(\vec{w}^T\vec{x_i} + b) > 0$

$\Rightarrow$ if $y_i = 1 \Rightarrow \vec{w}^T\vec{x_i} + b > 0 \geq 0 \Rightarrow y_i = 1$ if $\vec{w}^T\vec{x_i} + b \geq 0$, $\forall (\vec{x_i}, y_i) \in D$

if $y_i = -1 \Rightarrow \vec{w}^T\vec{x_i} + b < 0 \Rightarrow y_i = -1$ if $\vec{w}^T\vec{x_i} + b < 0$

$\Rightarrow$ D satifies condition (1) $\Rightarrow$ D is linearly separable

② If $1-\delta \leq 0 \Rightarrow \delta \geq 1$

(1). If this hyperplane is the optimal solution corresponds to the minimum δ

$\Rightarrow$ There exists $(\vec{x_i}, y_i)$ such that $y_i(\vec{w}^T\vec{x_i} + b) = 1-\delta$

$\Rightarrow$ for this $(\vec{x_i}, y_i)$, $y_i(\vec{w}^T\vec{x_i} + b) \leq 0$

$\Rightarrow$ if $y_i = 1 \Rightarrow \vec{w}^T\vec{x_i} + b \leq 0$ for this $(\vec{x_i}, y_i)$

if $y_i = -1 \Rightarrow \vec{w}^T\vec{x_i} + b \geq 0$

$\Rightarrow$ D does not satify condition (1) $\Rightarrow$ D is not linearly separable

(2). If this hyperplane is not the optimal solution

$\Rightarrow$ There may exist another hyperplane such that $0 \leq \delta < 1$

Then D is linearly separable.

Or there does not exist any hyperplane such that $0 \leq \delta < 1$

Then D is not linearly separable

$\Rightarrow$ If $0 < \delta < 1$, D is linearly separable

If $\delta \geq 1$ & δ is the minimal δ, D is not linearly separable

If $\delta \geq 1$ & δ is not the minimal δ, D is linearly separable or D is not linearly separable

(e). optimal solution gives minimal δ

one optimal solution could be $\vec{w} = \vec{0}$, $b = 0$ regardless of D

such that $y_i(\vec{w}^T\vec{x_i} + b) = y_i(\vec{0}^T\vec{x_i} + 0) = 0 \geq 0 \geq -\delta, (\delta \geq 0 \Rightarrow -\delta \leq 0), \forall (\vec{x_i}, y_i) \in D$

then δ could be 0, which is the minimal value of δ we can get

$\Rightarrow$ Then when this formulation tries to give us a optimal solution,

it may give $\vec{w} = \vec{0}$, $b = 0$, which is not a hyperplane.

$\Rightarrow$ This is the issue with such a formulation.

(f). only two examples in D $\Rightarrow$ D is separable

$\Rightarrow$ according to (a), there exists an optimal solution $\vec{w}^T, b$ to the linear program (2) with $\delta = 0 \Rightarrow y_i(\vec{w}^T\vec{x_i} + b) \geq 1 \quad \forall (\vec{x_i}, y_i) \in D$

Let $\vec{w} = [w_1, w_2, w_3]^T$

plug D into formulation (2).

$$\Rightarrow \begin{cases} w_1 + w_2 + w_3 + b \geq 1 \\ (+)(-w_1 - w_2 - w_3 + b) \geq 1 \Rightarrow w_1 + w_2 + w_3 - b \geq 1 \end{cases}$$

$$\Rightarrow \begin{cases} w_1 + w_2 + w_3 \geq 1 - b \\ w_1 + w_2 + w_3 \geq 1 + b \end{cases}$$

$\Rightarrow w_1 + w_2 + w_3 \geq \max\{1-b, 1+b\}$

$\max\{1-b, 1+b\} = 1+b$ if $b \geq 0$

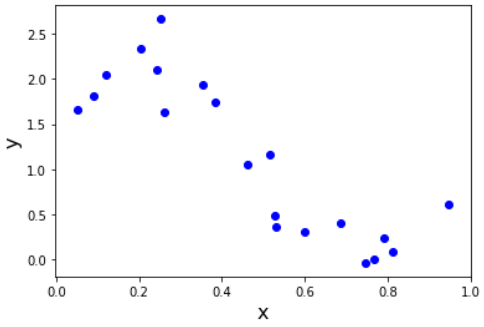
$= 1-b$ if $b < 0$

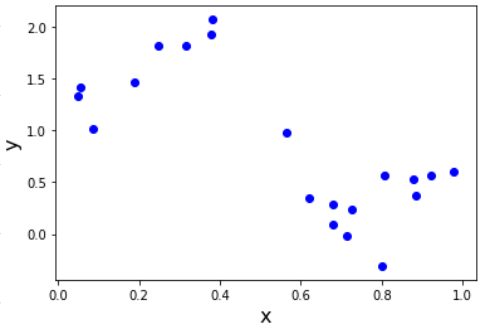
$\Rightarrow w_1 + w_2 + w_3 \geq 1 + |b|$

$\Rightarrow \vec{w}, b, \delta = 0$ is the optimal solution

if $w_1 + w_2 + w_3 \geq 1 + |b|$, $\delta = 0$ $\quad \vec{w} = [w_1 \; w_2 \; w_3]^T$

5. (a). Training data: Test data:





In this instance, linear regression is the linear combination of X and a bias, which is a line. It is obvious that both the training data and test data shown above cannot be predicted well by a line.
The linear regression cannot predict the data well but polynomial regression may work.

(d)

learning rate	coefficients	number of iterations	final value of the objective function	time (s)
10^{-4}	[2.270448, -2.460648]	10000	4.086397	0.1459422
10^{-3}	[2.446407, -2.816353]	7020	3.912576	0.1044039
10^{-2}	[2.446407, -2.816353]	764	3.912576	0.0118754
0.0407	$[-9.405\times10^{18}, -4.652\times10^{18}]$	10000	2.711×10^{39}	0.1422105

The coefficients from 10^{-4}, 10^{-3} and 10^{-2} are similar.
Theoretically, the smaller learning rate is, the more accurate coefficients are, the slower the convergence is.
The coefficients from 10^{-3} and 10^{-2} are the same with same lowest final objective function value. The coefficients from 10^{-4} give higher final objective function value. The convergence of 10^{-4} learning rate may be too slow, so that after 10000 iterations, the model still does not converge yet.
The 0.0407 learning rate is too large so that the model cannot converge.

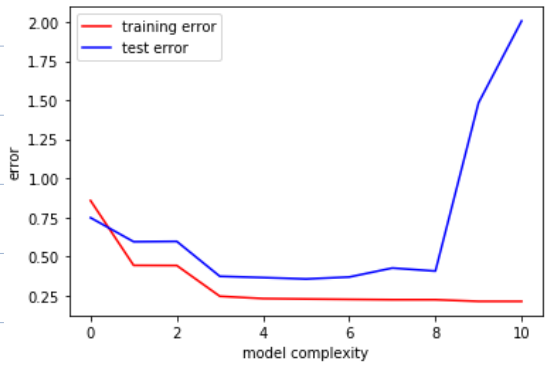
⇒ The coefficients of the linear regression model can be accurate enough with some relatively larger learning rate.
The learning rate cannot be too large, otherwise the model will not converge.
The learning rate maynot be too small, otherwise the convergence is too slow.

(e). The closed form solution is: coefficients: [2.446407, -2.816354]
final value of the objective function: 3.912576
The coefficients and the cost of the closed-form solution is almost the same with those obtained by GD with 10^{-3} or 10^{-2} learning rate.
The closed-form solution runs much faster than GD. (time = 9.502×10^{-4} s)

(f). Set the learning rate η for GD to be $\eta_k = \frac{1}{k+1} = \frac{1}{t+2}$ for each iteration.
k is the kth iteration. t is the index in the GD function.
It takes 615 iterations to converge to the same solution yielded by the closed-form optimization. (time = 0.0100265 s)

(h)① Because RMSE uses $\frac{1}{N}$, which eliminate the effect of the size of the data set. Then even if the sizes of training data set and test data set are different, we can still compare their errors with respect to different model complexity. Thus, RMSE is more useful.
② RMSE uses a square root.
It represents the standard deviation of the predictions from real values.
⇒ The scale of RMSE is the same as the predictions while the scale of J(w) is the square of the predictions'. Thus, RMSE is more reasonable.

(i).



① Degree 3 best fits the data.

Degree 3 is the lowest degree with relatively small training error and test error. In addition, the errors of degree 3 are both on the inflection point.

⇒ Degree 3 balances the errors and the degree well.

② When the degree increases, the training error decreases.

the test error first decreases, then increases.

The increasing of the test error shows the overfitting.

The overfitting is obvious when degree is larger than 8

The high training and test error for degree 0,1,2 shows the underfitting.

The underfitting is obvious when degree is smaller than 3.