CSCI3230 / ESTR3108: Fundamentals of Artificial Intelligence 2022-2023 Term 1
The Chinese University of Hong Kong

Assignment 1

Due date: 7 October 2022 (Fri) 23:59 Full mark: 100
Expected time spent: 3-5 hours

Aims: 1. Get familiar with AI concepts including accuracy, precision, recall, F1-score and AUC.
2. Understand the knowledge about linear regression and logistic regression.
3. Hands-on practice of the analytic solution of linear regression and ridge regression.
4. Hands-on practice of gradient descent in logistic regression.

Description:

In Assignment 1, you will first know some classic evaluation metrics used in machine learning. Next, you will calculate the analytic solution of a linear model on a training dataset and then try to increase the complexity of the model. You will also try to find the analytic solution for ridge regression. Finally, you will practice using gradient descent to train a logistic regression model.

For some calculations, you can use the toolbox in Python or MATLAB or any other programming languages you are familiar with.

Questions:

1. Assume that we use a logistic regression model to perform binary classification task. Given that the output of trained logistic regression model (i.e., $\sigma(f_\Theta(x)) = \frac{1}{1+e^{-f_\Theta(x)}}$) on six test samples is (0.56, 0.61, 0.43, 0.78, 0.12, 0.47), while the ground truth of the test data is (0, 1, 0, 1, 0, 1) respectively. In practice, we often set a threshold to perform binary classification (positive/negative) task using the output of logistic regression model. For example, if $\sigma(f_\Theta(x)) \geq threshold$, we consider the final prediction of sample x is positive and vice versa.

   (a) The F1-score combines the precision and recall of a classifier into a single metric by taking their harmonic mean (F1-score= $\frac{2(Precision*Recall)}{Precision+Recall}$)). Given that the threshold is set as 0.5 by default, use the above model predictions to calculate the accuracy and F1-score to evaluate this model. Then, show the confusion matrix. (6%)

   Solution:

   TP = 2, FP = 1, FN = 1, TN = 2.

   1) Accuracy = $\frac{TP+TN}{TP+FP+TN+FN} = 66.7\%$ ……(2%)
   2) Precision $\frac{TP}{TP+FP} = 66.7\%$
   3) Recall = $\frac{TP}{TP+FN} = 66.7\%$
   4) F1-score = $\frac{2(Precision*Recall)}{Precision+Recall}$ = 66.7% ……(2%)
   5) Confusion matrix:

		Predicted	
		Positive	Negative
Actual	Positive	2	1
	Negative	1	2

……(4%)

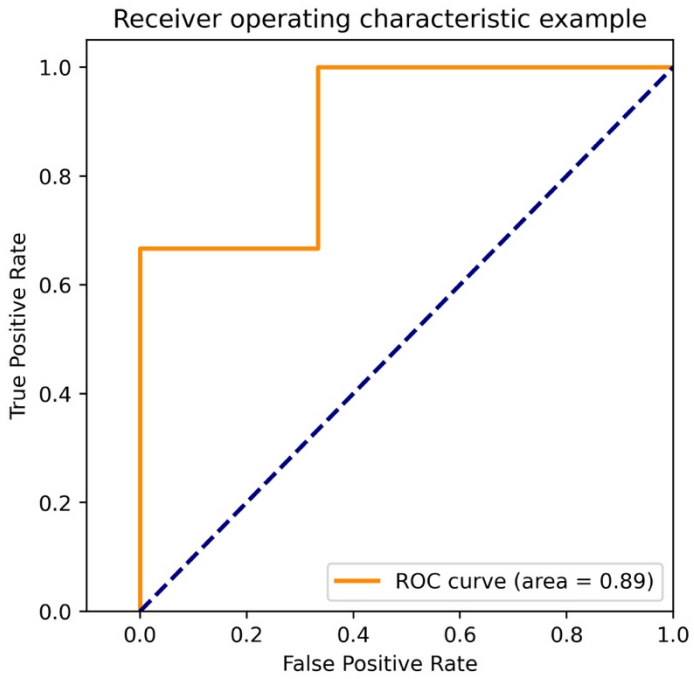
(b) Consider two scenarios: 1) If we want to use this model to decide whether to recommend a good to a customer (positive: recommend; negative: not recommend), we want to make the users be really interested in the good if we would like to recommend. How to change the threshold to achieve that? 2) If we want to use this model to develop an earthquake prediction system (positive: will be an earthquake; negative: will not be an earthquake), we don't want to miss any chance to save lives. How to change the threshold in this case? (4%)
(hint: from the perspectives of precision and recall)

Solution:

1) Increase the threshold to achieve high precision. ……(2%)
2) Decrease the threshold to achieve high recall. ……(2%)

(c) A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as **its discrimination threshold is varied**. The ROC curve is created by plotting the true positive rate (TPR=$\frac{TP}{TP+FN}$) against the false positive rate (FPR=$\frac{FP}{FP+TN}$) at various threshold settings. The area under the curve (AUC) is the area between the curve and the coordinate axis, which can be used as another evaluation metrics that ignore the influence of threshold. Please draw the ROC curve and calculate the AUC of the trained logistic regression model.
(hint: https://en.wikipedia.org/wiki/Receiver_operating_characteristic. You can progressively change the threshold to find the critical points to draw the plot.) (10%)

Solution:



……(5%)
AUC = 0.89……(5%)

2. Assume we have a training set and a test set as follows:

Training set

Index	x	y
1	5.51	0.81
2	1.25	1.22
3	3.60	0.43
4	4.72	-0.51
5	3.91	-0.13
6	6.13	0.44
7	8.05	1.49
8	5.55	0.31
9	7.33	1.59
10	7.59	1.61

Test set

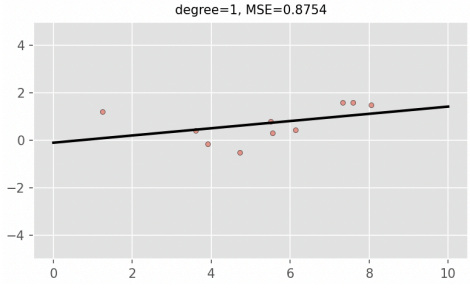
Index	x	y
1	5.84	0.89
2	0.61	1.79
3	4.23	-0.15
4	6.50	1.63
5	0.89	1.27
6	3.75	0.91
7	5.73	0.88
8	3.10	1.41
9	6.47	1.69
10	4.59	-0.46

Let's try to find some linear models to fit the training data. We use the RSS objective function $J(\boldsymbol{\Theta}) = \left\|\hat{f}_{\boldsymbol{\Theta}}(\mathbf{X}) - \boldsymbol{Y}\right\|_2^2$.

(a) Calculate the analytic solution of the linear model $\hat{f}_{\boldsymbol{\Theta}}(\mathrm{x}) = \theta_0 + \theta_1 \mathrm{x}$. Then, plot the line of your obtained linear model together with the data points in training set. (5%)

Solution:

$\boldsymbol{\Theta}^* = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathrm{Y} = \begin{pmatrix} -0.0893031 \\ 0.151995 \end{pmatrix}$ ..... (3%)



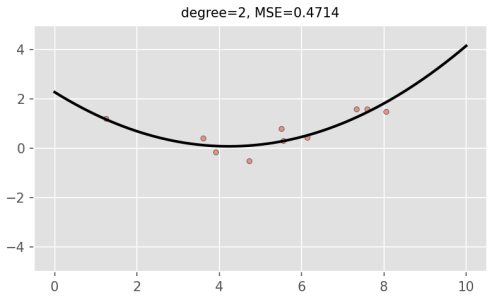
..... (2%)

(b) Suppose we want to increase the model complexity, by considering y as a linear function of both x and x^2: $\hat{f}_{\boldsymbol{\Theta}}(\mathrm{x}) = \theta_0 + \theta_1 \mathrm{x} + \theta_2 \mathrm{x}^2$. In this case, calculate the analytic solution of the model and plot the curve of the model, together with the data points in training set. (5%)

(hint: in this case, $\mathbf{X} = \begin{pmatrix} 1 & \mathrm{x}^{(1)} & {\mathrm{x}^{(1)}}^2 \\ \vdots & \vdots & \vdots \\ 1 & \mathrm{x}^{(10)} & {\mathrm{x}^{(10)}}^2 \end{pmatrix}$ and $\boldsymbol{\Theta} = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{pmatrix}$)

Solution:

$\Theta^* = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathrm{Y} = \begin{pmatrix} 2.28533 \\ -1.03753 \\ 0.122519 \end{pmatrix}$ ..... (3%)



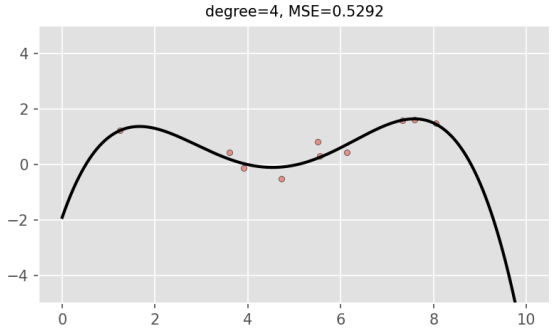
····· (2%)

(c) Let's further increase the model complexity, by assuming y is related to higher-order forms of x, i.e., $\hat{f}_{\boldsymbol{\Theta}}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$. Again, calculate the analytic solution of the model, plot the curve of the function, together with the data points in training set. (5%)

(hint: in this case, $\mathbf{X} = \begin{pmatrix} 1 & x^{(1)} & x^{(1)^2} & x^{(1)^3} & x^{(1)^4} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x^{(10)} & x^{(10)^2} & x^{(10)^4} & x^{(10)^4} \end{pmatrix}$ and $\boldsymbol{\Theta} = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix}$)

Solution:

$$\boldsymbol{\Theta}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y = \begin{pmatrix} -1.90766 \\ 4.81139 \\ -2.29185 \\ 0.385993 \\ -0.0210055 \end{pmatrix}$$ ····· (3%)



····· (2%)

(d) Observe the above three functions, please point out which could be faced with underfitting, which could be faced with overfitting, and which one is relatively a good one? Then, you can calculate the values of prediction error on the test data to verify your thoughts. (10%)

Solution:

The first model is faced with underfitting, the third one is faced with overfitting and the second one is the best. ····· (5%)

Model1: $E_{mean} = \frac{1}{10}\sum_{i=1}^{10}\left\|\hat{f}_{\boldsymbol{\Theta}}(x^{(i)}) - y^{(i)}\right\|_2^2 = 0.8754$

Model2: $E_{mean} = \frac{1}{10}\sum_{i=1}^{10}\left\|\hat{f}_{\boldsymbol{\Theta}}(x^{(i)}) - y^{(i)}\right\|_2^2 = 0.4714$

Model3: $E_{mean} = \frac{1}{10}\sum_{i=1}^{10}\left\|\hat{f}_{\boldsymbol{\Theta}}(x^{(i)}) - y^{(i)}\right\|_2^2 = 0.5292$ ····· (5%)

3. Recall that we have learned ridge regression which is a shrinkage method to regularize the coefficients in linear models. Suppose we have M samples $\mathbf{X} = \left(\mathbf{X^{(1)}}, \mathbf{X^{(2)}}, \dots, \mathbf{X^{(m)}}\right)^{\boldsymbol{T}}$. The ridge regression penalizes L2 norm of the model parameters: $\widehat{\boldsymbol{\Theta}}^{\text{ridge}} = \arg\min_{\boldsymbol{\Theta}} \|\mathbf{X\Theta} + \theta_0 - \mathbf{Y}\|_2^2 + \lambda\|\boldsymbol{\Theta}\|_2^2$, with λ as a positive scalar ($\lambda > 0$). Let's find the analytic solution for the ridge regression. First of all, in order to be more convenient for the derivation, we rewrite the error function as:

$$J(\boldsymbol{\Theta}, \theta_0) = \sum_{i=1}^{m}\left[\mathbf{X^{(i)}\Theta} + \theta_0 - \mathrm{Y^{(i)}}\right]^2 + \lambda\|\boldsymbol{\Theta}\|_2^2.$$

Set $\overline{\mathbf{X}}$ be the mean vector of all the raw vectors of $\mathbf{X}$, then, if we change the form of $J(\boldsymbol{\Theta}, \theta_0)$ into the following rewritten function:

$$J(\boldsymbol{\Theta}, \theta_0) = \sum_{i=1}^{m}\left[\left(\mathbf{X^{(i)}} - \overline{\mathbf{X}}\right)\boldsymbol{\Theta} + \theta_0 + \overline{\mathbf{X}}\boldsymbol{\Theta} - \mathrm{Y^{(i)}}\right]^2 + \lambda\|\boldsymbol{\Theta}\|_2^2$$

(a) Prove that, when $\theta_0 = \overline{\mathrm{Y}} - \overline{\mathbf{X}}\boldsymbol{\Theta}$, we can get the minimal value for $J(\boldsymbol{\Theta}, \theta_0)$. Then, let's define the centered input as $\mathbf{X}_c^{(i)} = \mathbf{X^{(i)}} - \overline{\mathbf{X}}$, and the corresponding centered label as $\mathrm{Y}_c^{(i)} = \mathrm{Y^{(i)}} - \overline{\mathrm{Y}}$. Plug the above θ_0 into the loss function and derive that:

$$J(\boldsymbol{\Theta}, \theta_0) = J_c(\boldsymbol{\Theta}) = \sum_{i=1}^{m}\left[\mathbf{X}_c^{(i)}\boldsymbol{\Theta} - \mathrm{Y}_c^{(i)}\right]^2 + \lambda\|\boldsymbol{\Theta}\|_2^2$$

(15%)

Solution:

$$\frac{\partial J(\boldsymbol{\Theta}, \theta_0)}{\partial \theta_0} = 2\sum_{i=1}^{m}\left[\left(\mathbf{X^{(i)}} - \overline{\mathbf{X}}\right)\boldsymbol{\Theta} + \theta_0 + \overline{\mathbf{X}}\boldsymbol{\Theta} - \mathrm{Y^{(i)}}\right] = 0 \cdots\cdots\cdots (5\%)$$

$$\Rightarrow \sum_{i=1}^{m}\left(\mathbf{X^{(i)}} - \overline{\mathbf{X}}\right)\boldsymbol{\Theta} + m\theta_0 + \sum_{i=1}^{m}\overline{\mathbf{X}}\boldsymbol{\Theta} - \sum_{i=1}^{m}\mathrm{Y^{(i)}} = 0 \cdots\cdots\cdots (3\%)$$

As $\overline{\mathbf{X}}$ is the mean of all the input, $\sum_{i=1}^{m}\left(\mathbf{X^{(i)}} - \overline{\mathbf{X}}\right)\boldsymbol{\Theta} = 0$, $\theta_0 = \frac{1}{m}\sum_{i=1}^{m}\mathrm{Y^{(i)}} - \overline{\mathbf{X}}\boldsymbol{\Theta} = \overline{\mathrm{Y}} - \overline{\mathbf{X}}\boldsymbol{\Theta}$ ····· (2%)

Note that when $\theta_0 = \overline{\mathrm{Y}} - \overline{\mathbf{X}}\boldsymbol{\Theta}$, we get the minimized value of $J(\boldsymbol{\Theta}, \theta_0)$. So we replace θ_0 by $\overline{\mathrm{Y}} - \overline{\mathbf{X}}\boldsymbol{\Theta}$:

$$\begin{aligned}
&\sum_{i=1}^{m}\left[\mathbf{X^{(i)}\Theta} + \theta_0 - \mathrm{Y^{(i)}}\right]^2 + \lambda\|\boldsymbol{\Theta}\|_2^2 \\
&= \sum_{i=1}^{m}\left[\mathbf{X^{(i)}\Theta} + \overline{\mathrm{Y}} - \overline{\mathbf{X}}\boldsymbol{\Theta} - \mathrm{Y^{(i)}}\right]^2 + \lambda\|\boldsymbol{\Theta}\|_2^2 \\
&= \sum_{i=1}^{m}\left[\left(\mathbf{X^{(i)}} - \overline{\mathbf{X}}\right)\boldsymbol{\Theta} - \left(\mathrm{Y^{(i)}} - \overline{\mathrm{Y}}\right)\right]^2 + \lambda\|\boldsymbol{\Theta}\|_2^2 \\
&= \sum_{i=1}^{m}\left[\mathbf{X}_c^{(i)}\boldsymbol{\Theta} - \mathrm{Y}_c^{(i)}\right]^2 + \lambda\|\boldsymbol{\Theta}\|_2^2 \cdots\cdots\cdots (5\%)
\end{aligned}$$

(b) With calculating the first-order derivatives of $J_c(\boldsymbol{\Theta})$, try to find the analytic solution $\widehat{\boldsymbol{\Theta}}$ for the ridge regression. (10%)

Solution:

$$\begin{aligned}
J_c(\boldsymbol{\Theta}) &= \|\mathbf{X_c\Theta} - \mathbf{Y_c}\|_2^2 + \lambda\|\boldsymbol{\Theta}\|_2^2 \\
&= (\mathbf{X_c\Theta} - \mathbf{Y_c})^{\mathrm{T}}(\mathbf{X_c\Theta} - \mathbf{Y_c}) + \lambda\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta} \\
&= \boldsymbol{\Theta}^{T}\mathbf{X_c}^{T}\mathbf{X_c\Theta} - \mathbf{Y_c}^{T}\mathbf{X_c\Theta} - \boldsymbol{\Theta}^{T}\mathbf{X_c}^{T}\mathbf{Y_c} - \mathbf{Y_c}^{T}\mathbf{Y_c} \\
&\quad + \lambda\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta}
\end{aligned}$$

$$\frac{\partial J_c(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} = 2\mathbf{X_c}^{T}\mathbf{X_c\Theta} - \mathbf{X_c}^{T}\mathbf{Y_c} - \mathbf{X_c}^{T}\mathbf{Y_c} + 2\lambda\boldsymbol{\Theta} = 0 \cdots\cdots\cdots (3\%)$$

$$\frac{\partial^2 J_c(\boldsymbol{\Theta})}{\partial^2 \boldsymbol{\Theta}} = 2\mathbf{X_c}^T\mathbf{X_c} + 2\lambda > 0 \cdots\cdots (2\%)$$

$$\widehat{\boldsymbol{\Theta}} = \left(\mathbf{X_c}^T\mathbf{X_c} + \lambda\mathbf{I}\right)^{-1}\mathbf{X_c}^T\mathbf{Y_c} \cdots\cdots (5\%)$$

(c) Use ridge regression with $\lambda = 0.1$ to train the linear model of $\hat{f}_{\boldsymbol{\Theta}}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$ in **Question 2**. Then, compare the values of prediction error of this model with/without ridge regression, and give a brief description about the influence of ridge regression. (10%)

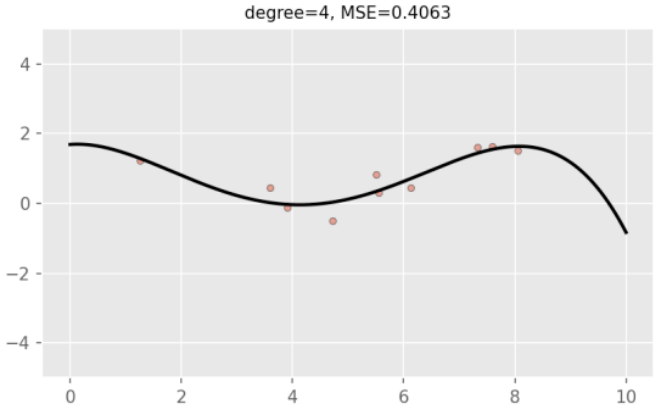
Solution:

$$\widehat{\boldsymbol{\Theta}} = \left(\mathbf{X_c}^T\mathbf{X_c} + \lambda\mathbf{I}\right)^{-1}\mathbf{X_c}^T\mathbf{Y_c} = \begin{pmatrix} 0.1198 \\ -0.4762 \\ 0.1120 \\ -0.0068 \end{pmatrix} \cdots\cdots (3\%)$$

$$\theta_0 = \bar{Y} - \bar{\mathbf{X}}\widehat{\boldsymbol{\Theta}} = 1.6715 \cdots\cdots (2\%)$$

$$E_{mean} = \frac{1}{10}\sum_{i=1}^{10}\left\|\hat{f}_{\widehat{\boldsymbol{\Theta}}}(X^{(i)}) - y^{(i)}\right\|_2^2 = 0.4063 \cdots\cdots (3\%)$$

The regularization term imposes a penalty to the linear model and allows some of coefficient values to go to the small value. In other words, it can reduce model complexity so that can alleviate overfitting····· (2%)



4. Assume that we have a training set as follows:

Training set

Index	x_1	x_2	y
1	0.315	0.761	0
2	0.129	0.413	0
3	0.358	0.752	0
4	0.611	0.855	1
5	0.765	0.782	1
6	0.613	0.967	1

Use these data to implement a logistic regression classifier. We use the linear model $f_\Theta(x_1, x_2) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$ and the logistic regression function as $\sigma\left(f_\Theta(x_1, x_2)\right) = \frac{1}{1+e^{-f_\Theta(x_1,x_2)}}$. The error function $E(\boldsymbol{\Theta})$ uses cross-entropy error function.

Now, let's use the gradient descent to update the model in the training set. The initial weights are set as $\theta_0 = -2.5$, $\theta_1 = 6.0$, $\theta_2 = 0.5$.

(a) Calculate the derivative of $E(\boldsymbol{\Theta})$ w.r.t. θ_0, θ_1 and θ_2 (don't need the specifc values). (10%) (hint: first write down the logistic model $P(\hat{y}=1|x_1,x_2)$ and its cross-entropy error function)

Solution:

$$P(\hat{y}=1|x_1,x_2)=\sigma\big(f_\Theta(x_1,x_2)\big)=\frac{1}{1+e^{-f_\Theta(x_1,x_2)}}=\frac{1}{1+e^{-(\theta_0+\theta_1x_1+\theta_2x_2)}}\cdots\cdots\cdots\cdots(2\%)$$

$$\text{Cross-entropy error}=-y\ln P(\hat{y}=1|x_1,x_2)-(1-y)\ln\left(1-P(\hat{y}=1|x_1,x_2)\right)\cdots\cdots(3\%)$$

$$\frac{\partial E(\Theta)}{\partial\theta_0}=\left(P\left(\hat{y}^{(i)}=1\middle|{x_1}^{(i)},{x_2}^{(i)}\right)-y^{(i)}\right)\cdot 1$$

$$\frac{\partial E(\Theta)}{\partial\theta_1}=\left(P\left(\hat{y}^{(i)}=1\middle|{x_1}^{(i)},{x_2}^{(i)}\right)-y^{(i)}\right)\cdot x_1$$

$$\frac{\partial E(\Theta)}{\partial\theta_2}=\left(P\left(\hat{y}^{(i)}=1\middle|{x_1}^{(i)},{x_2}^{(i)}\right)-y^{(i)}\right)\cdot x_2\cdots\cdots\cdots\cdots\cdots\cdots(5\%)$$

(b) Given that $\sum_{i=1}^{6}\frac{\partial E(\Theta)}{\partial\theta_0}$=0.723, $\sum_{i=1}^{6}\frac{\partial E(\Theta)}{\partial\theta_1}$=0.084 and $\sum_{i=1}^{6}\frac{\partial E(\Theta)}{\partial\theta_2}$=0.433, use gradient descent to update θ_0, θ_1 and θ_2 for **one iteration**. Then, compare the errors (i.e., $\sum_{i=1}^{6}E(\boldsymbol{\Theta})$) of the learning rate lr=1, lr=0.1 and lr=0.01 in the **second** iteration. Which learning rate is the best considering the error? Why? (10%)

Solution:

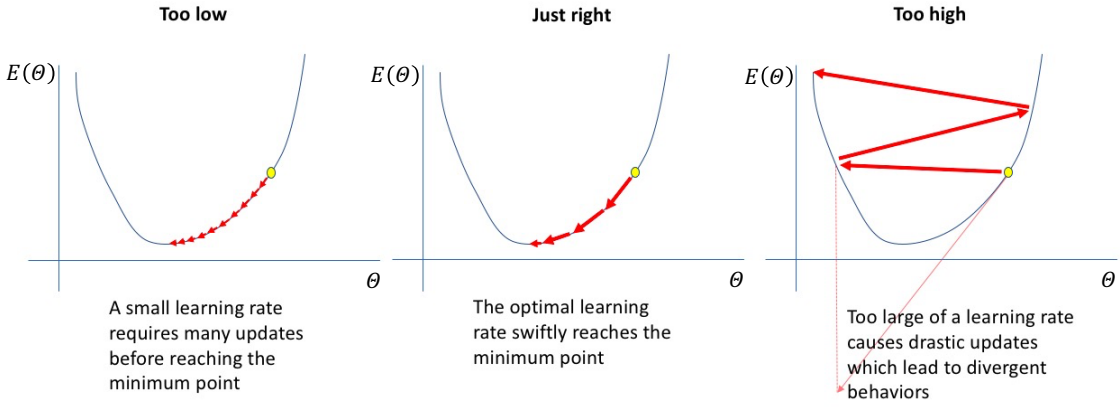
$$\theta_0:\theta_0-lr\times\sum_{i=1}^{6}\frac{\partial E(\Theta)}{\partial\theta_0}$$

$$\theta_1:\theta_1-lr\times\sum_{i=1}^{6}\frac{\partial E(\Theta)}{\partial\theta_1}$$

$$\theta_2:\theta_2-lr\times\sum_{i=1}^{6}\frac{\partial E(\Theta)}{\partial\theta_2}\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots(2\%)$$

$Lr=1$:
$\theta_0=-3.223,\ \theta_1=5.916,\theta_2=0.067,\ Loss=1.8315$
$Lr=0.1$:
$\theta_0=-2.572,\ \theta_1=5.992,\theta_2=0.457,\ Loss=1.8614$
$Lr=0.01$:
$\theta_0=-2.507,\ \theta_1=5.999,\theta_2=0.496,\ Loss=1.9201$ ……(3%)

$Lr=1$ is the best. The learning rates of 0.1 and 0.01 are too slow for one iteration in this scenario (Open-ended question. In addition, the scores will not be deducted if you update the model with two iterations. Another analysis can be seen in the figure below.) ……(5%)



Submission:

Submit a single file named <ID>_asmt1.pdf, where <ID> is your student ID.
Your file should contain the following header. Contact Professor Dou before submitting the assignment if you have anything unclear about the guidelines on academic honesty.

```
CSCI3230 / ESTR3108 2022-23 First Term Assignment 1

I declare that the assignment here submitted is original except
for source material explicitly acknowledged, and that the same or
closely related material has not been previously submitted for
another course. I also acknowledge that I am aware of University
policy and regulations on honesty in academic work, and of the
disciplinary guidelines and procedures applicable to breaches of
such policy and regulations, as contained in the following
websites.

University Guideline on Academic Honesty:
http://www.cuhk.edu.hk/policy/academichonesty/
Faculty of Engineering Guidelines to Academic Honesty:
http://www.erg.cuhk.edu.hk/erg-
intra/upload/documents/ENGG_Discipline.pdf

Student Name: <fill in your name>
Student ID  : <fill in your ID>
```

Submit your files using the Blackboard online system.

Notes:

1. Remember to submit your assignment by 23:59pm of the due date. We may not accept late submissions.
2. If you submit multiple times, **ONLY** the content and time-stamp of the **latest** one would be considered.

University Guideline for Plagiarism

Please pay attention to the university policy and regulations on honesty in academic work, and the disciplinary guidelines and procedures applicable to breaches of such policy and regulations. Details can be found at http://www.cuhk.edu.hk/policy/academichonesty/. With each assignment, students will be required to submit a statement that they are aware of these policies, regulations, guidelines and procedures.