CSCI3230 / ESTR3108: Fundamentals of Artificial Intelligence 2023-2024 Term 1
The Chinese University of Hong Kong

Assignment 1

Due date: 9 October 2023 (Mon) 23:59 — Full mark: 100
Expected time spent: 3-5 hours

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

Description:
In Assignment 1, 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. Next, you will practice using gradient descent to train a logistic regression model. Finally, you will explore how to tackle a multi-class classification problem.

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

Questions:

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

Training set

Index	x	y
1	5.86	0.74
2	1.34	1.18
3	3.65	0.51
4	4.69	-0.48
5	4.13	-0.07
6	5.87	0.37
7	7.91	1.35
8	5.57	0.30
9	7.30	1.64
10	7.89	1.75

Test set

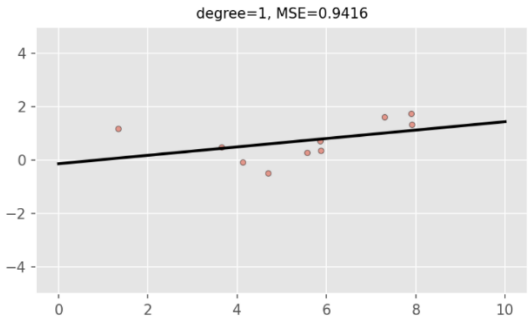
Index	x	y
1	5.80	0.93
2	0.57	1.87
3	4.30	-0.06
4	6.55	1.60
5	0.82	1.22
6	3.72	0.90
7	5.80	0.93
8	3.26	1.53
9	6.75	1.73
10	4.77	-0.51

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:

$$\boldsymbol{\Theta}^* = (\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\mathrm{Y}$$
$$= \begin{pmatrix} -0.124714 \\ 0.157483 \end{pmatrix}$$ ..... (3%)



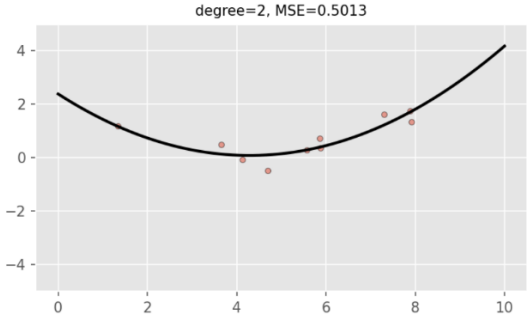
(2%)

(b) Suppose we want to increase the model complexity, by considering y as a linear function of both x and x^2: $\hat{f}_{\boldsymbol{\Theta}}(x) = \theta_0 + \theta_1 x + \theta_2 x^2$. 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, $\mathbf{X} = \begin{pmatrix} 1 & x^{(1)} & {x^{(1)}}^2 \\ \vdots & \vdots & \vdots \\ 1 & x^{(10)} & {x^{(10)}}^2 \end{pmatrix}$ and $\boldsymbol{\Theta} = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{pmatrix}$)

Solution:

$$\Theta^* = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathrm{Y} = \begin{pmatrix} 2.38902 \\ -1.07371 \\ 0.125279 \end{pmatrix}$$ (3%)



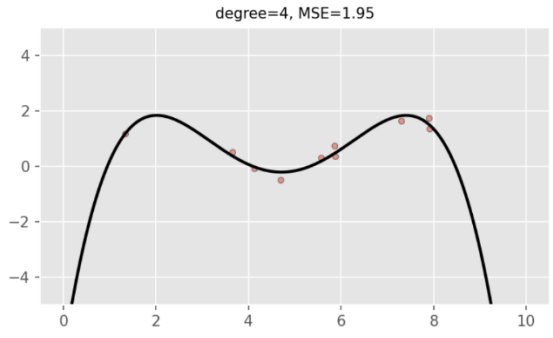
(2%)

(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:

$$\boldsymbol{\Theta}^* = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathrm{Y} = \begin{pmatrix} -6.80776 \\ 10.9168 \\ -4.60514 \\ 0.731709 \\ -0.0388404 \end{pmatrix}$$ (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%)

$$\text{Model1: } E_{mean} = \frac{1}{10}\sum_{i=1}^{10}\left\|\hat{f}_{\Theta}(X^{(i)}) - y^{(i)}\right\|_2^2 = 0.9416$$
$$\text{Model2: } E_{mean} = \frac{1}{10}\sum_{i=1}^{10}\left\|\hat{f}_{\Theta}(X^{(i)}) - y^{(i)}\right\|_2^2 = 0.5013$$
$$\text{Model3: } E_{mean} = \frac{1}{10}\sum_{i=1}^{10}\left\|\hat{f}_{\Theta}(X^{(i)}) - y^{(i)}\right\|_2^2 = 1.9504 \text{ ..... (5\%)}$$

2. Recall that we have learned ridge regression which is a shrinkage method to regularize the coefficients in linear models. Suppose we have M samples $\mathbf{X} = \begin{pmatrix} \mathbf{X^{(1)}} \\ \mathbf{X^{(2)}} \\ \vdots \\ \mathbf{X^{(m)}} \end{pmatrix}$. 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.

(a) First of all, in order to be more convenient for the derivation, we rewrite the error function as:

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

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:

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

Prove that, when $\theta_0 = \overline{Y} - \overline{\mathbf{X}}\boldsymbol{\Theta}$, we can get the minimal value for $J(\boldsymbol{\Theta}, \theta_0)$. (10%)

Solution:

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

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

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

$\cdots\cdots\cdots(2\%)$

(b) Next, let's define the centered input as $\mathbf{X}_c^{(i)}=\mathbf{X}^{(\mathrm{i})}-\overline{\mathbf{X}}$, and the corresponding centered label as $\mathrm{Y}_c^{(i)}=\mathrm{Y}^{(\mathrm{i})}-\overline{\mathrm{Y}}$. Plug the above θ_0 into the loss function, try to derive that:

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

(5%)

Solution:

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 $\overline{\mathrm{Y}}-\overline{\mathbf{X}}\boldsymbol{\Theta}$:

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

(c) Finally, 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:

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

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

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

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

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

Training set

Index	x_1	x_2	y
1	0.346	0.780	0
2	0.303	0.439	0
3	0.358	0.729	0
4	0.602	0.863	1
5	0.790	0.753	1
6	0.611	0.965	1

Test set

Index	x_1	x_2	y
1	0.959	0.382	0
2	0.750	0.306	0
3	0.395	0.760	0
4	0.823	0.764	1
5	0.761	0.874	1
6	0.844	0.435	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 $\sigma\big(f_\Theta(x_1, x_2)\big) = \frac{1}{1+e^{-f_\Theta(x_1,x_2)}}$. The error function 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 $\theta_0 = -1$, $\theta_1 = 1.5$, $\theta_2 = 0.5$. The learning rate is 0.1.

(a) Please write down the logistic model $P(\hat{y} = 1|x_1, x_2)$ and its cross-entropy error function. (5%)

Solution:

$$P(\hat{y} = 1|x_1, x_2) = \sigma\big(f_\Theta(x_1, x_2)\big) = \frac{1}{1+e^{-f_\Theta(x_1,x_2)}} = \frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}} \cdots\cdots\cdots\cdots (2\%)$$

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

(b) Use gradient descent to update θ_0, θ_1 and θ_2 for one iteration. Write down the updated logistic regression model. (10%)

Solution:

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

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

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

$$\theta_0: \theta_0 - lr \times \sum_{i=1}^{6} \frac{\partial E(\Theta)}{\partial \theta_0} = -1.01899756$$

$$\theta_1: \theta_1 - lr \times \sum_{i=1}^{6} \frac{\partial E(\Theta)}{\partial \theta_1} = 1.53210518$$

$$\theta_2: \theta_2 - lr \times \sum_{i=1}^{6} \frac{\partial E(\Theta)}{\partial \theta_2} = 0.51181202 \cdots\cdots\cdots\cdots\cdots\cdots (5\%)$$

(c) Use the above new model to make predictions for all the samples in the test dataset. Then, calculate the accuracy, precision and recall to evaluate this model. (10%)

Solution:

The result: $\hat{y} = [1,1,0,1,1,1]$ ····· (4%)

1) TP = 3, FP = 2, FN = 0, TN = 1.
2) Accuracy = $\frac{TP+TN}{TP+FP+TN+FN} = 66.6\%$ ……(2%)
3) Precision = $\frac{TP}{TP+FP} = 60\%$ ……(2%)
4) Recall = $\frac{TP}{TP+FN} = 100\%$ ……(2%)
5) Confusion matrix:

		Predicted	
		Positive	Negative
Actual	Positive	3	0
	Negative	2	1

4. In multi-class tasks we discriminate between three or more classes. As usual, we imagine that we have one sample $\mathbf{X} = \begin{pmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(n)} \end{pmatrix}$ with n-dimensional features. Suppose there are K classes in this task. Instead of representing the output labels as 1, 2, …, K, we present the vector to be 0 in each component except for the one corresponding to the correct label. For example, if there are 5 classes and a particular sample belongs to class 3, then its label vector is [0, 0, 1, 0, 0]. In other words, the label $y_3 = 1$ and $y_1 = y_2 = y_4 = y_5 = 0$.

   In many cases, we perform multi-class tasks by *softmax* function, with the probability written as:

$$P(\hat{y}_i = 1|\mathbf{X}) = \frac{e^{\mathbf{X}^\mathbf{T}\boldsymbol{\Theta}_i}}{\sum_{k=1}^{K} e^{\mathbf{X}^\mathbf{T}\boldsymbol{\Theta}_k}}$$

   Here $\mathbf{X}^\mathbf{T}\boldsymbol{\Theta}_i$ is the linear model. As there are K classes in this task, we have a set of K weight vectors for multi-class classification: $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \ldots, \boldsymbol{\Theta}_K$. The $P(\hat{y}_i = 1|\mathbf{X})$ indicates the probability that the prediction for the sample is class i.

   (a) Calculate that the derivative of $P(\hat{y}_i = 1|\mathbf{X})$ w.r.t. $\mathbf{X}^\mathbf{T}\boldsymbol{\Theta}_j$ is:

$$\frac{\partial P(\hat{y}_i = 1|\mathbf{X})}{\partial \mathbf{X}^\mathbf{T}\boldsymbol{\Theta}_j} = \begin{cases} P(\hat{y}_i = 1|\mathbf{X})\left(1 - P\left(\hat{y}_j = 1\middle|\mathbf{X}\right)\right), j = i \\ -P(\hat{y}_i = 1|\mathbf{X})P\left(\hat{y}_j = 1\middle|\mathbf{X}\right), \quad j \neq i \end{cases} \quad (10\%)$$

   (hint: you can separately discuss the two situations for $j = i$ and $j \neq i$)

   Solution:

$$\frac{\partial P(\hat{y}_i = 1|\mathbf{X})}{\partial \mathbf{X}^\mathbf{T}\boldsymbol{\Theta}_j} = \frac{\partial \frac{e^{\mathbf{X}^\mathbf{T}\boldsymbol{\Theta}_i}}{\sum_{k=1}^{K} e^{\mathbf{X}^\mathbf{T}\boldsymbol{\Theta}_k}}}{\partial \mathbf{X}^\mathbf{T}\boldsymbol{\Theta}_j}$$

   We discuss two situations. In the case of $j = i$, we have:

$$\frac{\partial P(\hat{y}_i = 1|\mathbf{X})}{\partial \mathbf{X}^\mathbf{T}\boldsymbol{\Theta}_j} = \frac{\partial \frac{e^{\mathbf{X}^\mathbf{T}\boldsymbol{\Theta}_i}}{\sum_{k=1}^{K} e^{\mathbf{X}^\mathbf{T}\boldsymbol{\Theta}_k}}}{\partial \mathbf{X}^\mathbf{T}\boldsymbol{\Theta}_j} = \frac{e^{\mathbf{X}^\mathbf{T}\boldsymbol{\Theta}_i}\sum_{k=1}^{K} e^{\mathbf{X}^\mathbf{T}\boldsymbol{\Theta}_k} - e^{\mathbf{X}^\mathbf{T}\boldsymbol{\Theta}_j} e^{\mathbf{X}^\mathbf{T}\boldsymbol{\Theta}_i}}{\left(\sum_{k=1}^{K} e^{\mathbf{X}^\mathbf{T}\boldsymbol{\Theta}_k}\right)^2}$$

$$= \frac{e^{\mathbf{X}^T\boldsymbol{\Theta}_i}}{\sum_{k=1}^{K} e^{\mathbf{X}^T\boldsymbol{\Theta}_k}} \frac{\left(\sum_{k=1}^{K} e^{\mathbf{X}^T\boldsymbol{\Theta}_k} - e^{\mathbf{X}^T\boldsymbol{\Theta}_j}\right)}{\sum_{k}^{K} e^{\mathbf{X}^T\boldsymbol{\Theta}_k}}$$
$$= P(\hat{y}_i = 1|\mathbf{X})\left(1 - P\left(\hat{y}_j = 1\middle|\mathbf{X}\right)\right) \cdots\cdots\cdots\cdots (5\%)$$

In the case of $j \neq i$, we have:

$$\frac{\partial P(\hat{y}_i = 1|\mathbf{X})}{\partial \mathbf{X}^T\boldsymbol{\Theta}_j} = \frac{\partial \frac{e^{\mathbf{X}^T\boldsymbol{\Theta}_i}}{\sum_{k=1}^{K} e^{\mathbf{X}^T\boldsymbol{\Theta}_k}}}{\partial \mathbf{X}^T\boldsymbol{\Theta}_j}$$
$$= \frac{-e^{\mathbf{X}^T\boldsymbol{\Theta}_j} e^{\mathbf{X}^T\boldsymbol{\Theta}_i}}{\left(\sum_{k=1}^{K} e^{\mathbf{X}^T\boldsymbol{\Theta}_k}\right)^2}$$
$$= -\frac{e^{\mathbf{X}^T\boldsymbol{\Theta}_i}}{\sum_{k=1}^{K} e^{\mathbf{X}^T\boldsymbol{\Theta}_k}} \frac{e^{\mathbf{X}^T\boldsymbol{\Theta}_j}}{\sum_{k=1}^{K} e^{\mathbf{X}^T\boldsymbol{\Theta}_k}}$$
$$= -P(\hat{y}_i = 1|\mathbf{X})P\left(\hat{y}_j = 1\middle|\mathbf{X}\right) \cdots\cdots\cdots\cdots (5\%)$$

(b) Calculate the derivative of $P(\hat{y}_i = 1|\mathbf{X})$ w.r.t. $\boldsymbol{\Theta}_j$. (5%)
(hint: $\frac{\partial P(\hat{y}_i = 1|\mathbf{X})}{\partial \boldsymbol{\Theta}_j} = \frac{\partial P(\hat{y}_i = 1|\mathbf{X})}{\partial \mathbf{X}^T\boldsymbol{\Theta}_j} \frac{\partial \mathbf{X}^T\boldsymbol{\Theta}_j}{\partial \boldsymbol{\Theta}_j}$)

Solution:
In the case of $j = i$, we have:

$$\frac{\partial P(\hat{y}_i = 1|\mathbf{X})}{\partial \boldsymbol{\Theta}_j} = \frac{\partial P(\hat{y}_i = 1|\mathbf{X})}{\partial \mathbf{X}^T\boldsymbol{\Theta}_j} \frac{\partial \mathbf{X}^T\boldsymbol{\Theta}_j}{\partial \boldsymbol{\Theta}_j}$$
$$= \mathbf{X}P(\hat{y}_i = 1|\mathbf{X})\left(1 - P\left(\hat{y}_j = 1\middle|\mathbf{X}\right)\right)$$

In the case of $j \neq i$, we have:

$$\frac{\partial P(\hat{y}_i = 1|\mathbf{X})}{\partial \boldsymbol{\Theta}_j} = \frac{\partial P(\hat{y}_i = 1|\mathbf{X})}{\partial \mathbf{X}^T\boldsymbol{\Theta}_j} \frac{\partial \mathbf{X}^T\boldsymbol{\Theta}_j}{\partial \boldsymbol{\Theta}_j}$$
$$= -\mathbf{X}P(\hat{y}_i = 1|\mathbf{X})P\left(\hat{y}_j = 1\middle|\mathbf{X}\right) \cdots\cdots\cdots\cdots (5\%)$$

(c) We use the following negative log-likelihood loss function for this multi-class task:

$$E(\boldsymbol{\Theta}) = -\sum_{i=1}^{K} y_i \ln P(\hat{y}_i = 1|\mathbf{X})$$

where y_i indicates the label for class *i*. Prove that the derivative of $E(\boldsymbol{\Theta})$ w.r.t. $\boldsymbol{\Theta}_j$ is

$$\frac{\partial E(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}_j} = \mathbf{X}P\left(\hat{y}_j = 1\middle|\mathbf{X}\right) - \mathbf{X}y_j \quad (10\%)$$

(hint: $\frac{\partial E(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}_j} = \frac{\partial E(\boldsymbol{\Theta})}{\partial P(\hat{y}_i = 1|\mathbf{X})} \frac{\partial P(\hat{y}_i = 1|\mathbf{X})}{\partial \mathbf{X}^T\boldsymbol{\Theta}_j} \frac{\partial \mathbf{X}^T\boldsymbol{\Theta}_j}{\partial \boldsymbol{\Theta}_j}$)

Solution:

$$\frac{\partial E(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}_j} = \frac{\partial E(\boldsymbol{\Theta})}{\partial P(\hat{y}_i = 1|\mathbf{X})} \frac{\partial P(\hat{y}_i = 1|\mathbf{X})}{\partial \mathbf{X}^T\boldsymbol{\Theta}_j} \frac{\partial \mathbf{X}^T\boldsymbol{\Theta}_j}{\partial \boldsymbol{\Theta}_j}$$
$$= -\sum_{i=1}^{K} y_i \frac{1}{P(\hat{y}_i = 1|\mathbf{X})} \frac{\partial P(\hat{y}_i = 1|\mathbf{X})}{\partial \boldsymbol{\Theta}_j} \cdots\cdots\cdots\cdots (3\%)$$

$$= -\frac{y_j}{P(\hat{y}_j = 1|\mathbf{X})}\frac{\partial P(\hat{y}_j = 1|\mathbf{X})}{\partial \boldsymbol{\Theta}_j} - \sum_{i\neq j}^{K} y_i \frac{1}{P(\hat{y}_i = 1|\mathbf{X})}\frac{\partial P(\hat{y}_i = 1|\mathbf{X})}{\partial \boldsymbol{\Theta}_j} \quad \cdots\cdots(2\%)$$

$$= -\frac{y_j}{P(\hat{y}_j = 1|\mathbf{X})}\mathbf{X}P(\hat{y}_j = 1|\mathbf{X})\left(1 - P(\hat{y}_j = 1|\mathbf{X})\right) + \sum_{i\neq j}^{K} y_i \frac{1}{P(\hat{y}_i = 1|\mathbf{X})}\mathbf{X}P(\hat{y}_i = 1|\mathbf{X})P(\hat{y}_j = 1|\mathbf{X})$$

$$= -\mathbf{X}y_j + \mathbf{X}y_j P(\hat{y}_j = 1|\mathbf{X}) + \sum_{i\neq j}^{K} y_i \mathbf{X}P(\hat{y}_j = 1|\mathbf{X}) \quad \cdots\cdots(2\%)$$

$$= -\mathbf{X}y_j + \sum_{i=1}^{K} y_i \mathbf{X}P(\hat{y}_j = 1|\mathbf{X})$$

$$= -\mathbf{X}y_j + \mathbf{X}P(\hat{y}_j = 1|\mathbf{X})\sum_{i=1}^{K} y_i \quad \cdots\cdots(3\%)$$

$$= \mathbf{X}P(\hat{y}_j = 1|\mathbf{X}) - \mathbf{X}y_j$$

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 2023-24 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.