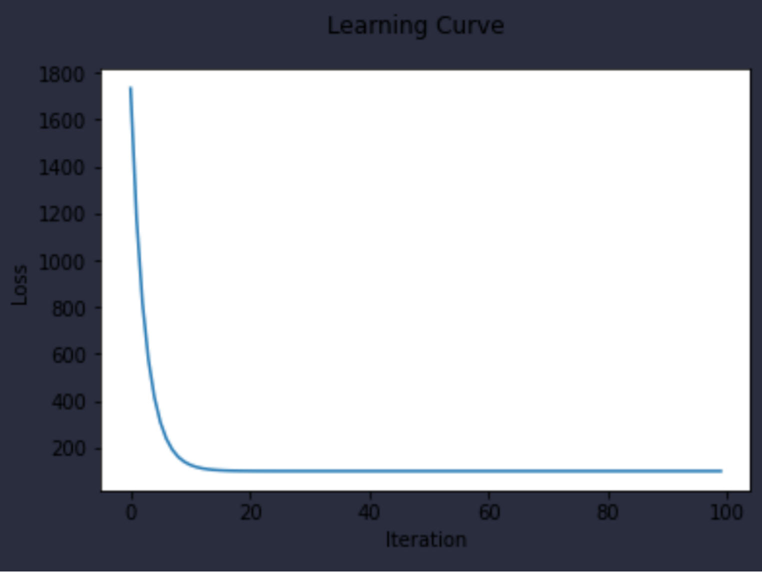
NYCU Introduction to Machine Learning, Homework 1

Part. 1, Coding (60%):

Linear regression model

1. (10%) Plot the learning curve of the training, you should find that loss decreases after a few iterations and finally converge to zero (x-axis=iteration, y-axis=loss, Matplotlib or other plot tools is available to use)
2. (10%) What's the Mean Square Error of your prediction and ground truth?
3. (10%) What're the weights and intercepts of your linear model?

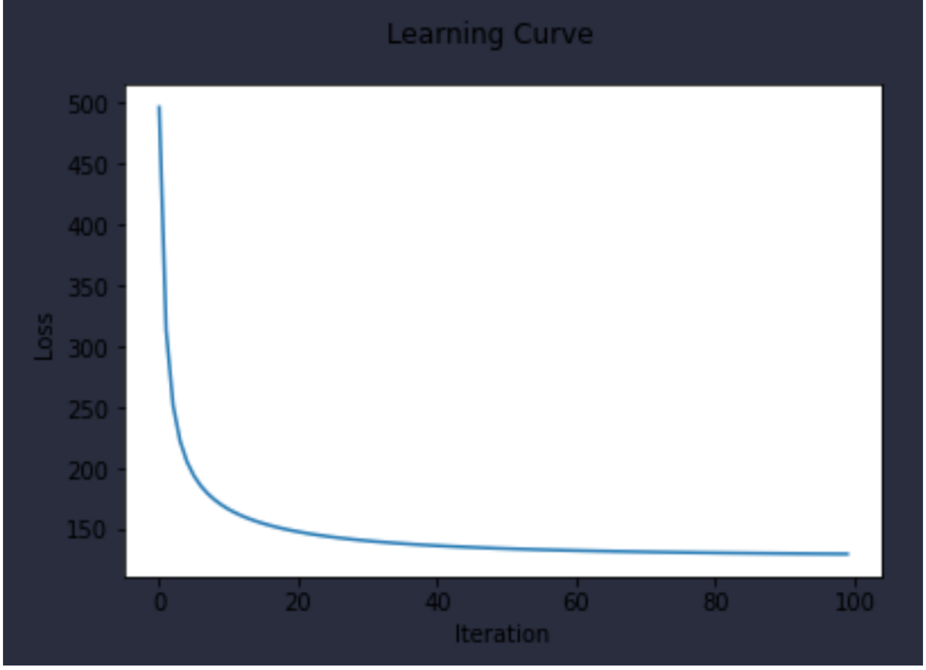
```
Weight: 52.74354039983602
Intercept: -0.33375890919599194
Mean Square Error: 110.43819236775595
```



Logistic regression model

1. (10%) Plot the learning curve of the training, you should find that loss decreases after a few iterations and finally converge to zero (x-axis=iteration, y-axis=loss, Matplotlib or other plot tools is available to use)
2. (10%) What's the Cross Entropy Error of your prediction and ground truth?
3. (10%) What're the weights and intercepts of your linear model?

```
Weight: 4.2002744308136055
Intercept: 1.30169881702314
Cross Entropy Error: 45.38122864007148
```



Part. 2, Questions (40%):

1. What’s the difference between Gradient Descent, Mini-Batch Gradient Descent, and Stochastic Gradient Descent?

The main difference is the number of data used for updating the parameters in a single iteration. Gradient Descent uses all the data in the training set, Mini-Batch Gradient Descent uses a small subset of the training set and Stochastic Gradient Descent usually uses only one sample of the training set.

2. Will different values of learning rate affect the convergence of optimization? Please explain in detail.

Definitely. A large learning rate will make the model converge more quickly, however, the model may be overfitted when the learning rate is too large because it may converge to a suboptimal solution. A small learning rate requires more training iterations to converge since it makes a small change to the parameter in each iteration, but a learning rate that is too small might make the process stuck.

3. Show that the logistic sigmoid function (eq. 1) satisfies the property $\sigma(-a) = 1 - \sigma(a)$ and that its inverse is given by $\sigma^{-1}(y) = \ln \{y/(1-y)\}$.

$$\sigma(a) = \frac{1}{1+\exp(-a)} \tag{4.59}$$

(eq. 1)

1. $\sigma(-a) = 1 - \sigma(a)$:

$$\sigma(-a) = \frac{1}{1+e^{a}} = \frac{1}{1+\frac{1}{e^{-a}}} = \frac{e^{-a}}{e^{-a}+1}$$
$$= \frac{(1+e^{-a})-1}{1+e^{-a}} = 1 - \frac{1}{1+e^{-a}} = 1 - \sigma(a)$$

2. $\sigma^{-1}(y) = \ln \{y/(1-y)\}$:

$$\sigma(a) = y = \frac{1}{1+e^{-x}}$$
$$\Longrightarrow 1+e^{-x} = \frac{1}{y} \qquad \Longrightarrow e^{-x} = \frac{1}{y} - 1$$
$$\Longrightarrow e^{-x} = \frac{1-y}{y} \qquad \Longrightarrow ln(e^{-x}) = ln(\frac{1-y}{y})$$
$$\Longrightarrow -x = ln(\frac{1-y}{y}) \qquad \Longrightarrow x = -ln(\frac{1-y}{y})$$
$$\Longrightarrow x = ln(\frac{y}{1-y}) \qquad \Longrightarrow \sigma^{-1}(y) = x = ln(\frac{y}{1-y})$$

4. Show that the gradients of the cross-entropy error (eq. 2) are given by (eq. 3).

$$E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\ln p(\mathbf{T}|\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk} \qquad (4.108)$$

(eq. 2)

$$\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = \sum_{n=1}^{N} (y_{nj} - t_{nj})\, \boldsymbol{\phi}_n \qquad (4.109)$$

(eq. 3)

Hints:

$$a_k = \mathbf{w}_k^{\mathrm{T}} \boldsymbol{\phi}. \qquad (4.105)$$

(eq. 4)

$$\frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j) \qquad (4.106)$$

(eq. 5)

$$\because\ a_{nj} = w_j^T \phi_n \implies \nabla_{w_j} a_{nj} = \phi_n\ ,$$

$$\frac{\partial E}{\partial y_{nk}} = -\frac{t_{nk}}{y_{nk}}\ ,\ \frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_i)\ ,$$

$$\frac{\partial E}{\partial a_{nj}} = \sum_{k=1}^{K} \frac{\partial E}{\partial y_{nk}} \frac{\partial y_{nk}}{\partial a_{nj}} = -\sum_{k=1}^{k} \frac{t_{nk}}{y_{nk}} y_{nk}(I_{kj} - y_{nj})$$

$$= -\sum_{k=1}^{k} t_{nk}(I_{kj} - y_{nj}) = -t_{nj} + \sum_{k=1}^{k} t_{nk} y_{nj} = y_{nj} - t_{nj}$$

$$\therefore\ \nabla_{w_j} E(w_1, \ldots, w_k) = \sum_{n=1}^{N} \frac{\partial E}{\partial a_{nj}} \nabla_{w_j} a_{nj} = \sum_{n=1}^{N} (y_{nj} - t_{nj}) \phi_n$$