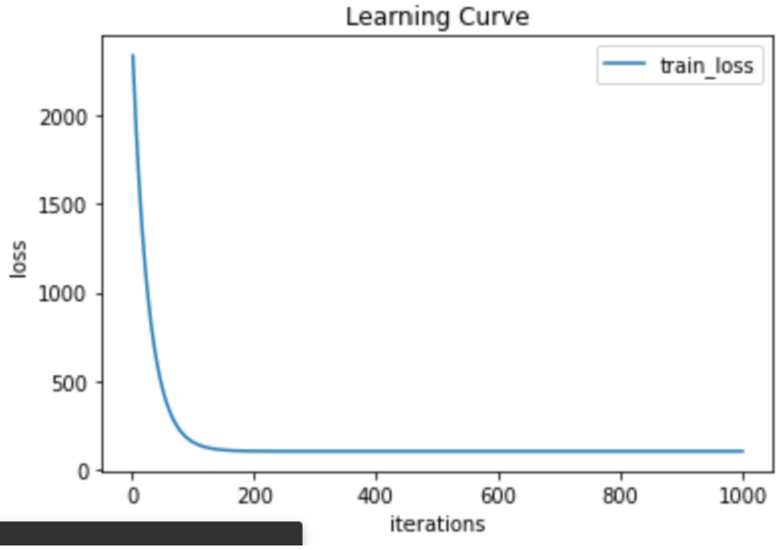
Part 1

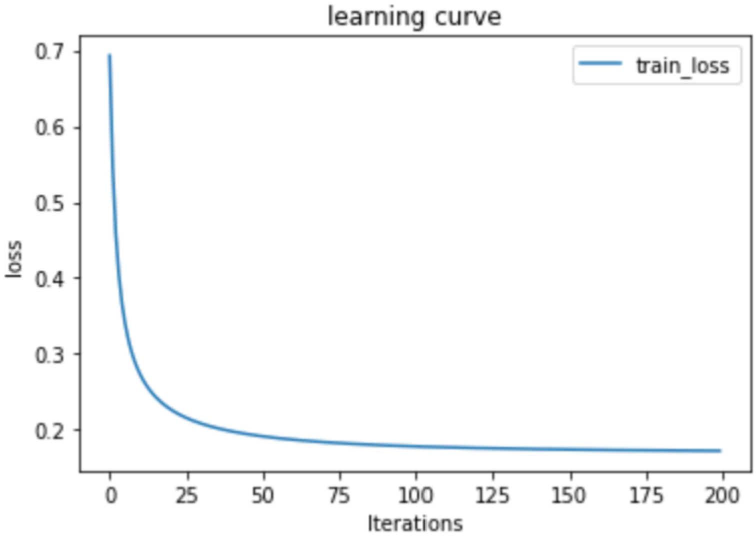
1. Linear Regression

```
Mean Square Error= [103.64391225]
Weight= [52.69971675]
Intercept=  [-2.40820525]
```



2. Logistic regression

```
Weight: [1.13868119]
Intercept: [4.95507395]
Cross Entropy Error: [1.92967234]
```



Part. 2, Questions

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

   這三種資料處理的方式主要差別在一次處理的資料量多少。

   Gradient Descent每次運算都會把整筆資料train過一遍，所以所花費的時間和cost也會較大。但也因為資料多，可以找到較為精確的gradient。

   Stochastic Gradient Descent跟Gradient Descent相反，一次運算只train一筆資料。雖然速度較快，但所得到的gradient不一定是朝我們預期的方向發展。所以即使單筆運算速度快，也還是可能會花費不少時間才能找到較為精確的gradient。

   Mini-Batch Gradient Descent則是介於Gradient Descent和Stochastic Gradient Descent中間。每次運算會train固定的m筆資料，如此既能快速確定gradient的運算方向，也能大幅縮短時間。另外因為每次train的資料數量固定，也有利於電腦進行運算，能提高efficiency。

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

   過小的learning rate會導致運算次數大幅增加，進而需要花更長的時間才能取得答案甚至可能導致program stuck。而過大的learning rate則可能會使regression model 過快收斂，導致找到的值是相對最小值而不是絕對最小值。

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

$$\sigma(a) = \frac{1}{1+\exp(-a)} \qquad \text{(eq. 1)}$$

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

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

$$y = \frac{1}{1+e^{-x}}$$

inverse of

$$\Rightarrow y(1+e^{-x}) = 1 \qquad \Rightarrow \qquad \sigma^{-1}(y) = \ln\{y/(1-y)\}$$

$$\Rightarrow y \cdot e^{-x} = 1-y$$

$$\Rightarrow e^{-x} = \frac{1-y}{y}$$

$$\Rightarrow -x = \ln\{(1-y)/y\}$$

$$\Rightarrow x = \ln\{y/(1-y)\}$$

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

$$E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\ln p(\mathbf{T}|\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk} \qquad \text{(eq. 2)}$$

$$\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = \sum_{n=1}^{N} (y_{nj} - t_{nj})\, \boldsymbol{\phi}_n \qquad \text{(eq. 3)}$$

Hints:

$$a_k = \mathbf{w}_k^{\mathrm{T}} \boldsymbol{\phi}. \qquad \text{(eq. 4)}$$

$$\frac{\partial y_k}{\partial a_j} = y_k (I_{kj} - y_j) \qquad \text{(eq. 5)}$$

consider $t_{nk} \ln y_{nk}$ from

$$E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\ln p(\mathbf{T}|\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk} \qquad \text{(eq. 2)}$$

derive it with regard to a_j

$$\Rightarrow \frac{\partial t_{nk} \ln y_{nk}}{\partial w_j} = \frac{\partial t_{nk} \ln y_{nk}}{\partial y_{nk}} \cdot \frac{\partial y_{nk}}{\partial a_j} \cdot \frac{\partial a_j}{\partial w_j}$$

$$= t_{nk} \cdot \frac{1}{\cancel{y_{nk}}} \cdot \cancel{y_{nk}} (I_{kj} - y_{nj}) \cdot \phi_n, \text{ since}$$

$$a_k = \mathbf{w}_k^{\mathrm{T}} \phi. \qquad \text{(eq. 4)}$$

$$\frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j) \qquad \text{(eq. 5)}$$

$$= t_{nk} \cdot (I_{kj} - y_{nj}) \cdot \phi_n$$

the summation over n, k :

$$\nabla_{w_j} E = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} (I_{kj} - y_{nj}) \cdot \phi_n$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} (y_{nj} - I_{kj}) \cdot \phi_n$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} y_{nj} \cdot \phi_n - \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} I_{kj} \phi_n$$

since $\sum_{k=1}^{K} t_{nk} = 1 \ \forall n.$

$$= \sum_{n=1}^{N}\left(\left(\sum_{k=1}^{K} t_{nk}\right) y_{nj} \phi_n \right) - \sum_{n=1}^{N} t_{nj} \phi_n$$

$$= \sum_{n=1}^{N} y_{nj} \phi_n - \sum_{n=1}^{N} t_{nj} \phi_n$$

$$= \sum_{n=1}^{N} (y_{nj} - t_{nj}) \phi_n .$$