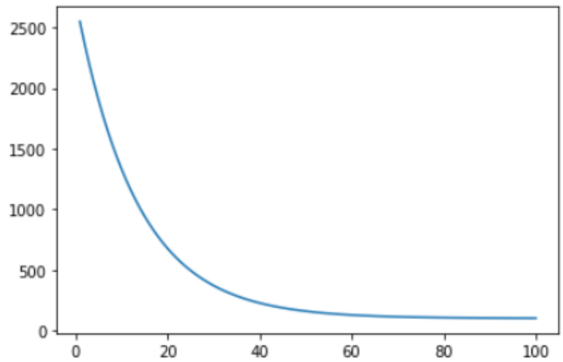
0816170 郭建良

Part 1

Linear regression model

1. Learning Curve



2. Mean Square Error: 108.282857706965

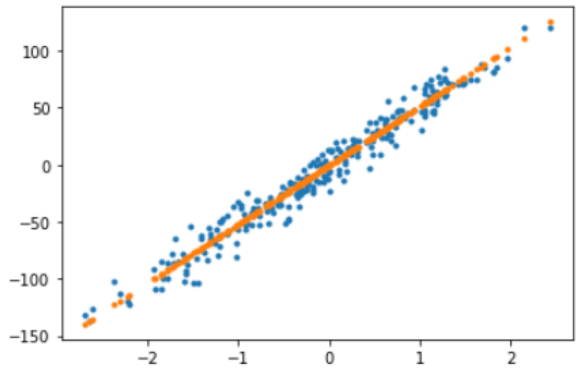
```
Mean Square Error: 108.282857706965
```

3.

Weights: 51.576268072290326

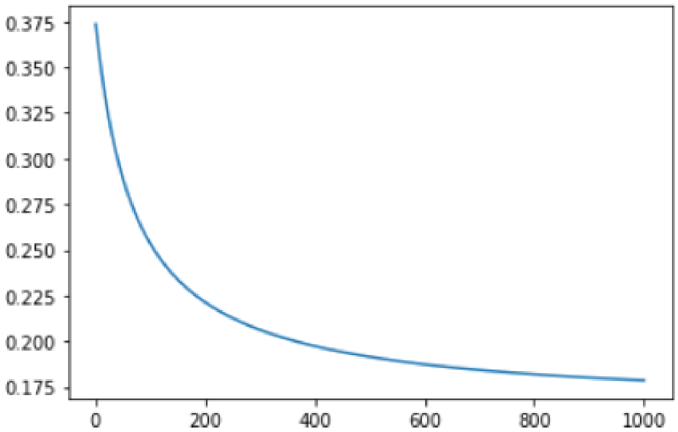
Intercepts: -0.4202922059016378

```
Mean Square Error: 108.282857706965

Weights: 51.576268072290326
intercepts: -0.4202922059016378

[<matplotlib.lines.Line2D at 0x7fd77a486590>]
```



Logistic regression model

1. Learning Curve



2. Cross Entropy Error: 0.17922123068337245

```
Cross entropy 0.17922123068337245
```

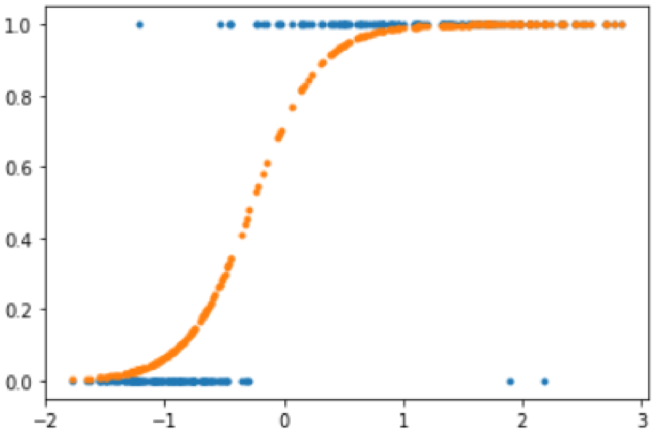
3.

Weights: 3.636819873424943

Intercepts: 0.9657610009242119

```
Weights: 3.636819873424943
intercepts: 0.9657610009242119
```

```
[<matplotlib.lines.Line2D at 0x7fd778fcb9d0>]
```



Part 2

1.

Gradient Descent: 使用整個 dataset(或 training set)來計算梯度以找到最佳解，會一直往最佳解的方向前進，最後可能停在 local 或 global 最佳解。因為每次都要使用整個 dataset 來進行訓練，所以若資料量大，每更新一次都要花很多時間。

Mini-Batch Gradient Descent: 將 dataset 分割成很多小的區塊(batch)，每次更新參數的時候隨機挑選其中一個區塊出來訓練，既不用遍歷整個 dataset，也不用一次只採用一個 sample，兼顧穩定性跟計算效率。

Stochastic Gradient Descent: 每次訓練只隨機挑選其中一筆 data 進行計算，每次訓練時間都非常短，但因為資料量龐大，所以單一 data 的可靠程度並不高，有可能會往錯誤的方向前進，因此找到最佳解的速度也不一定比較快。

2.

Learning rate 代表每次訓練時，更新 parameters 的幅度，越高的學習率，代表每次往最佳解前進的步伐越大。學習率的控管非常重要，過猶不及，太大的話，更新幅度太高，可能因此錯過最佳解，收斂到次佳解；若學習率太小，則有可能因為更新幅度太小，導致設定的 epoch 不足以達到收斂，且會耗費相當長的訓練時間。適當的學習率，可以兼顧準確性與計算效率，因此學習率的調整相當重要。

Part 2 3.

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

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

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

$$= \frac{\frac{1}{e^{a}}}{1+\frac{1}{e^{a}}}$$

$$= \frac{\frac{1}{e^{a}}}{\frac{e^{a}+1}{e^{a}}}$$

$$= \frac{1}{e^{a}+1}$$

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

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

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

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

$$\Rightarrow \ln e^{-a} = \ln\left(\frac{1}{y} - 1\right) = \ln \frac{1-y}{y}$$

$$= -a$$

$$\Rightarrow a = \ln\left(\frac{1-y}{y}\right)^{-1} = \ln \frac{y}{1-y}$$

$$\Rightarrow \sigma^{-1}(y) = \ln \frac{y}{1-y}$$

Part2 4.

According to the question,

we have $\frac{\partial E}{\partial y_{nk}} = -\frac{t_{nk}}{y_{nk}}$

According to eq.4

we have $\nabla_{w_j} a_{nj} = \phi_n$ — ①

Given $\frac{\partial E}{\partial y_{nk}} = -\frac{t_{nk}}{y_{nk}}$ and eq.5

we can compute that

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

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

$$= -t_{nj} + \sum_{k=1}^{K} t_{nk} y_{nj}$$

because $\sum_{k=1}^{K} t_{nk} = 1$, $= y_{nj} - t_{nj}$ — ②

① + ② $\Rightarrow \nabla_{w_j} E(w_1, \dots w_K) = \sum_{n=1}^{N} \frac{\partial E}{\partial a_{nj}} \nabla_{w_j} a_{nj} = \sum_{n=1}^{N} (y_{nj} - t_{nj}) \phi_n$