Homework 1 Report

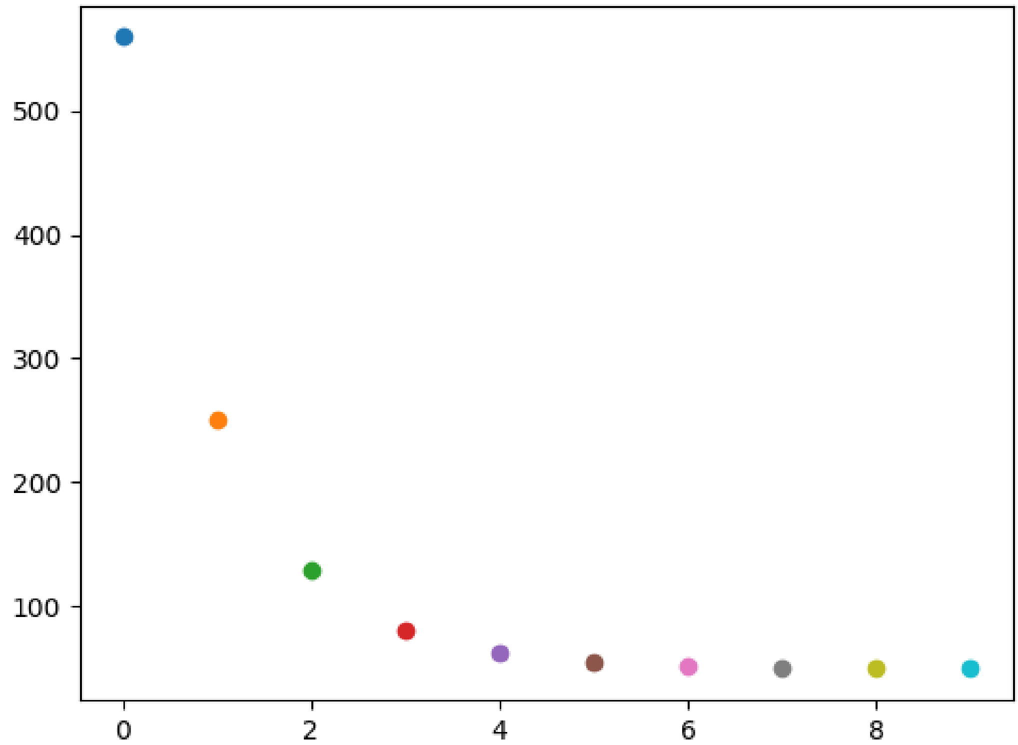
109550164 徐聖哲

Linear regression model

1.

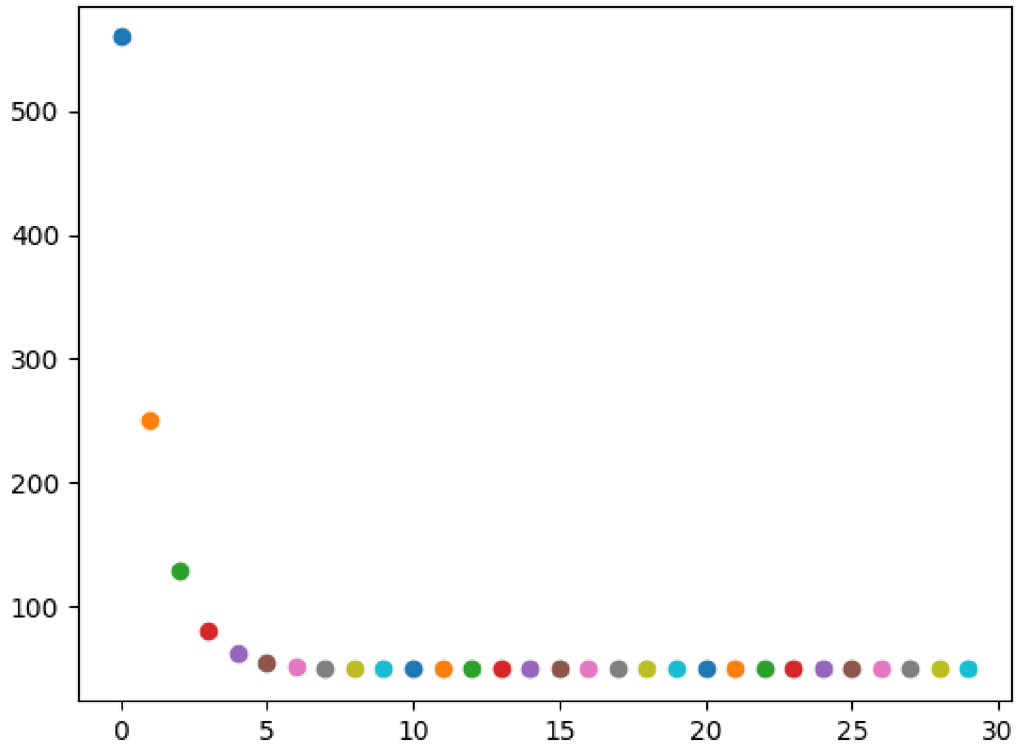
- **Iterations : 10**

```
PS C:\Users\danzel\Hsu\課程\大三上\機器學習>
機器學習/HW1/linear_regression.py
weights: 52.241568680275485
intercepts: -0.38341961689421783
Mean_square_error: 54.58001915343924
```



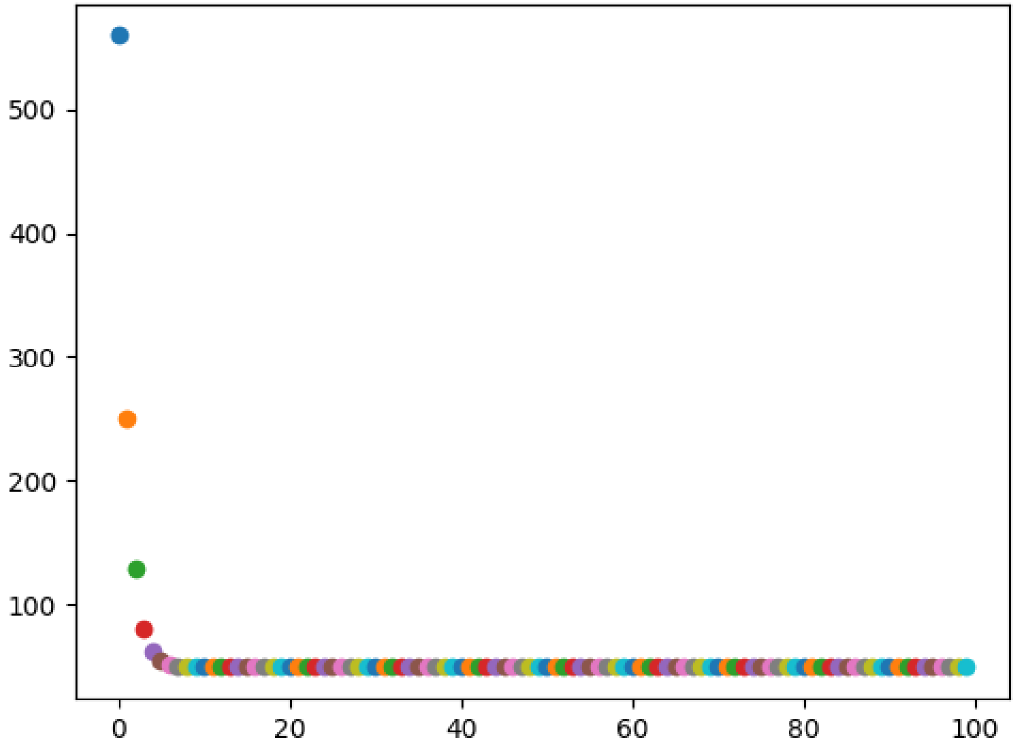
- **Iteratons : 30**

```
PS C:\Users\danzel\Hsu\課程\大三上\機器學習>
機器學習/HW1/linear_regression.py
weights: 52.74349369253078
intercepts: -0.3337684770833136
Mean_square_error: 55.21902353217067
```



- **Iterations : 100**

```
PS C:\Users\danzel\Hsu\課程\大三上\機器學習>
機器學習/HW1/linear_regression.py
weights:  52.74354046182485
intercepts:  -0.33375889502567535
Mean_square_error:  55.21909628062007
```

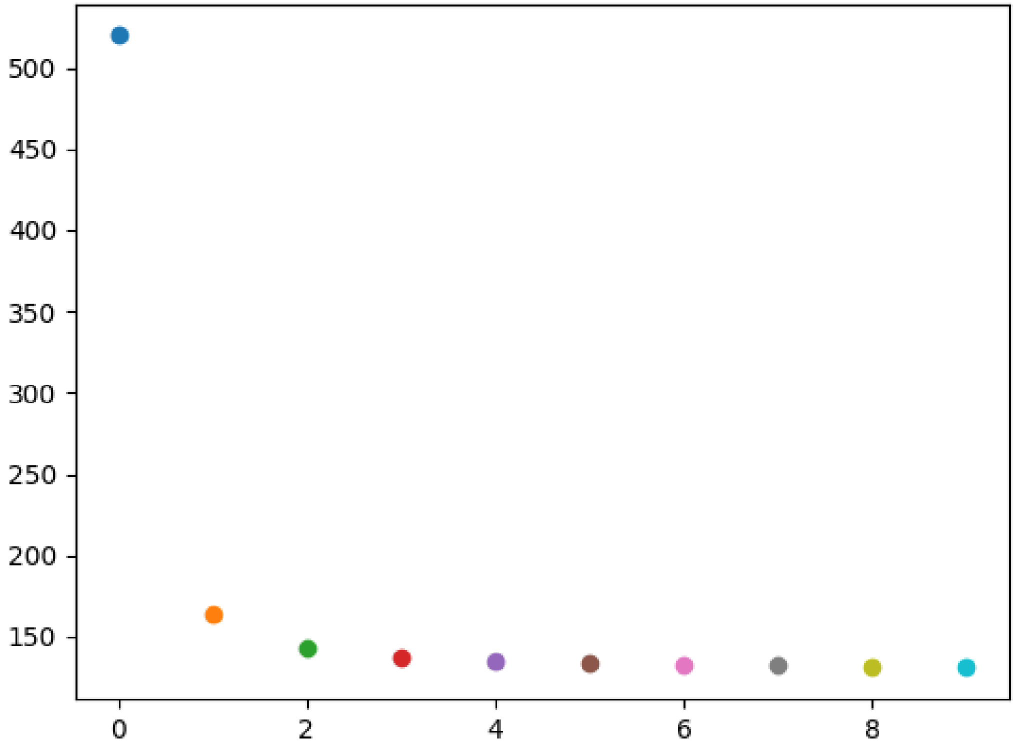


Logistic regression model

1.

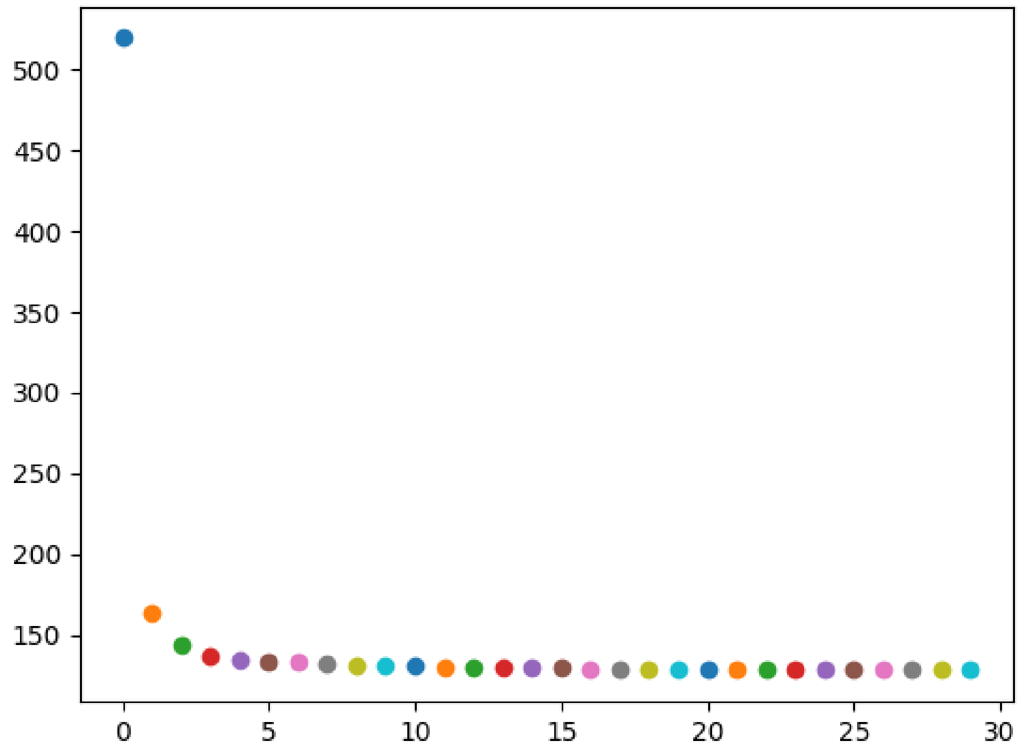
- **Iterations : 10**

```
PS C:\Users\danzel\Hsu\課程\大三上\機器學習>
機器學習/HW1/classification.py
weights: 5.525894194737832
intercepts: 2.108792369594169
Cross Entropy Error: 49.81141865717864
```



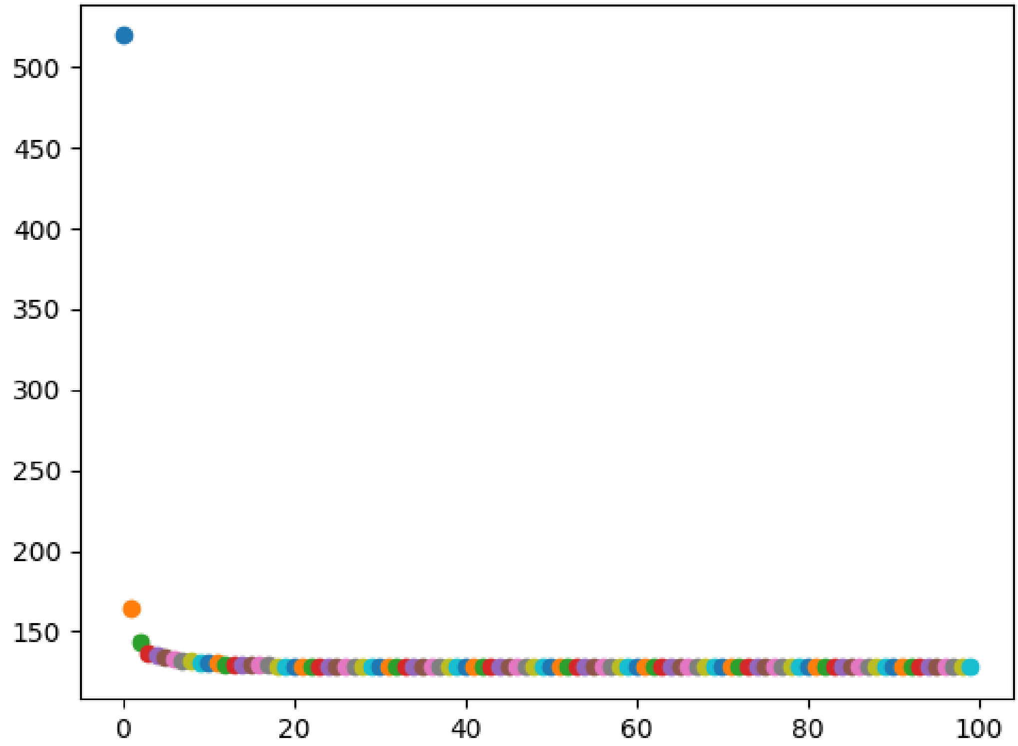
- **Iterations : 30**

```
PS C:\Users\danzel\Hsu\課程\大三上\機器學習>
機器學習/HW1/classification.py
weights: 5.002169900946224
intercepts: 1.788060301989326
Cross Entropy Error: 47.69360685469468
```



- **Iterations : 100**

```
PS C:\Users\danzel\Hsu\課程\大三上\機器學習>
機器學習/HW1/classification.py
weights: 4.877128620284377
intercepts: 1.711767920212915
Cross Entropy Error: 47.248435018985404
```



Part. 2, Questions (40%):

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

Ans: difference between Gradient Descent, Mini-Batch Gradient Descent, and Stochastic Gradient Descent is the amount of data been chosen. Stochastic Gradient Descent do training every piece of data and Mini-Batch Gradient Descentdo training every n pieces of data, while Gradient Descent do training for all data in once.

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

Ans: Different values of learning rate do affect the convergence of optimization. Learning rate controls how we adjust weights with respect to loss error. The smaller the value, the slower we travel through the descent slope. It may be good if we slowly walk through the gradient but we might stuck if rate is too small. So it is difficult to choose the right learning rate at first.

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

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

3. ① $\sigma(a) = \frac{1}{1+e^{-a}}$

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

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

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

② let $y = \frac{1}{1+e^{-x}}$

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

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

$-x = \ln\left(\frac{1-y}{y}\right)$

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

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

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

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

Hints:

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

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

4. $E(w_1, \dots, w_k) = -\ln p(T | w_1, \dots, w_k) = -\sum_{n=1}^{N} \sum_{k=1}^{k} t_{nk} \ln y_{nk}$

$\nabla_{w_j} E(w_1, \dots, w_k) = \sum_{n=1}^{N} (y_{nj} - t_{nj}) \phi_n$

$\Rightarrow \quad \frac{\partial E}{\partial y_{nk}} = -\frac{t_{nk}}{y_{nk}}$

Since We have $a_k = w_k^T \phi$

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

$\Rightarrow \quad \frac{\partial E}{\partial a_{nj}} = \sum_{k=1}^{k} \frac{\partial E}{\partial y_{nk}} \frac{\partial y_{nk}}{\partial a_{nj}}$

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

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

$\Rightarrow$ show $\nabla_{w_j} E(w_1, \dots, w_k) = \sum_{n=1}^{N} (y_{nj} - t_{nj}) \phi_n$