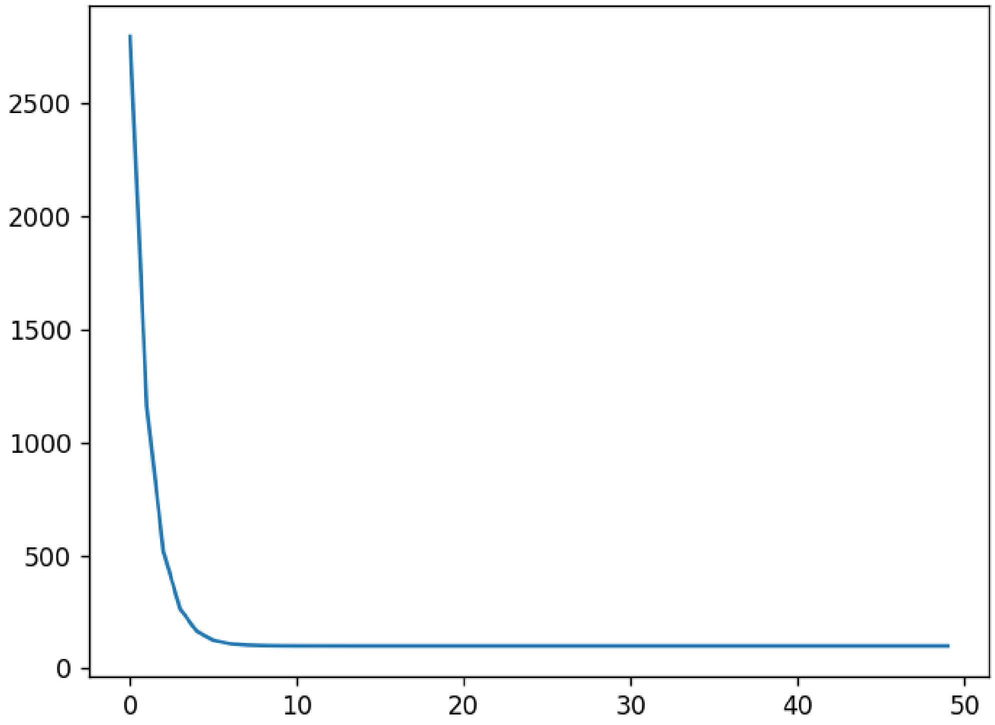
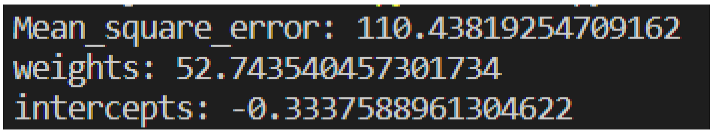
NYCU Introduction to Machine Learning, Homework 1

109550136 邱弘竣

Part. 1, Coding (60%):

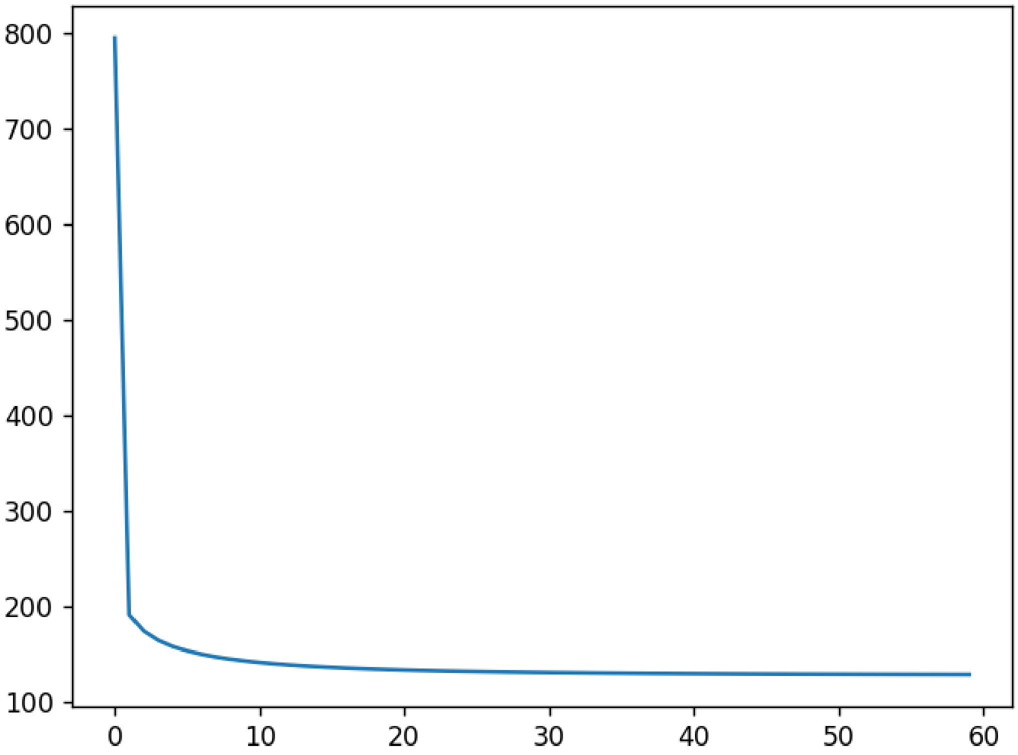
Linear regression model

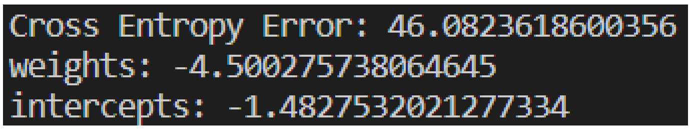




```
Mean_square_error: 110.43819254709162
weights: 52.743540457301734
intercepts: -0.3337588961304622
```

Logistic regression model





```
Cross Entropy Error: 46.0823618600356
weights: -4.500275738064645
intercepts: -1.4827532021277334
```

Part. 2, Questions (40%):

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

(1) Gradient Descent

$$w = w - \alpha \nabla_w J(w)$$

a. It takes the sum of all training data to run a single iteration.

b. It takes too much time.

c. It produces no noise and gives a lower standard error

(2) Mini-Batch Gradient Descent

$$w = w - \alpha \nabla_w J(x^{\{i:i+b\}}, y^{\{i:i+b\}}; w)$$

a. It takes 50-256 training set to run a single iteration.

b. The pros and cons are same as Gradient Descent while the computations are faster and the results are more accurate and precise than that.

(3) Stochastic Gradient Descent

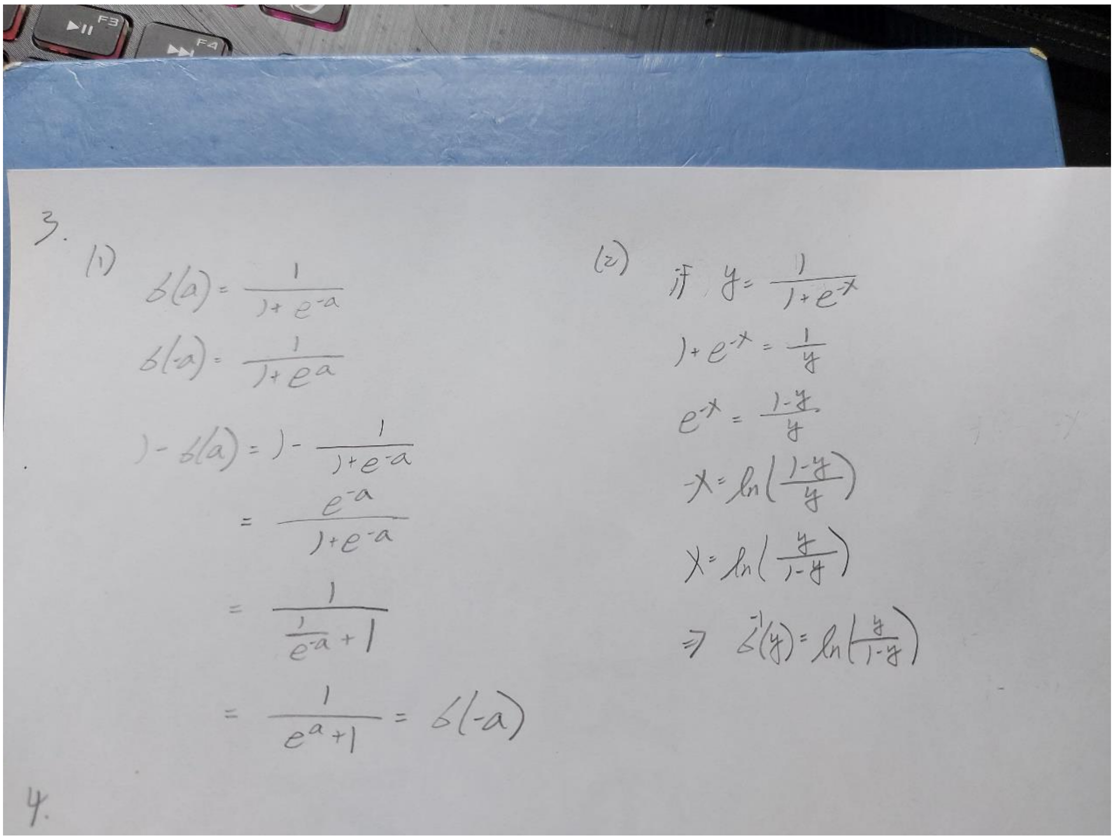
$$w = w - \alpha \nabla_w J(x^i, y^i; w)$$

a. We just run one example of the training data set in each iteration.
b. Compute faster since it run just one example in each iteration.
c. Result in larger variance.
d. There tends to be more noise which allows for the improved generalization error.

2. **Will different values of learning rate affect the convergence of optimization? Please explain in detail.**
   Learning rate determines how fast or slow we move towards the optimal weights. If the learning rate is large, we might skip the solution. If it is small, we might need too many iterations to converge to the solution.
3. **Show that the logistic sigmoid function (eq. 1) satisfies the property σ(−a) = 1 − σ(a) and that its inverse is given by σ −1 (y) = ln {y/(1 − y)}.**

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

(eq. 1)



3.

(1)

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

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

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

(2) 汙 $y = \frac{1}{1+e^{-x}}$

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

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

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

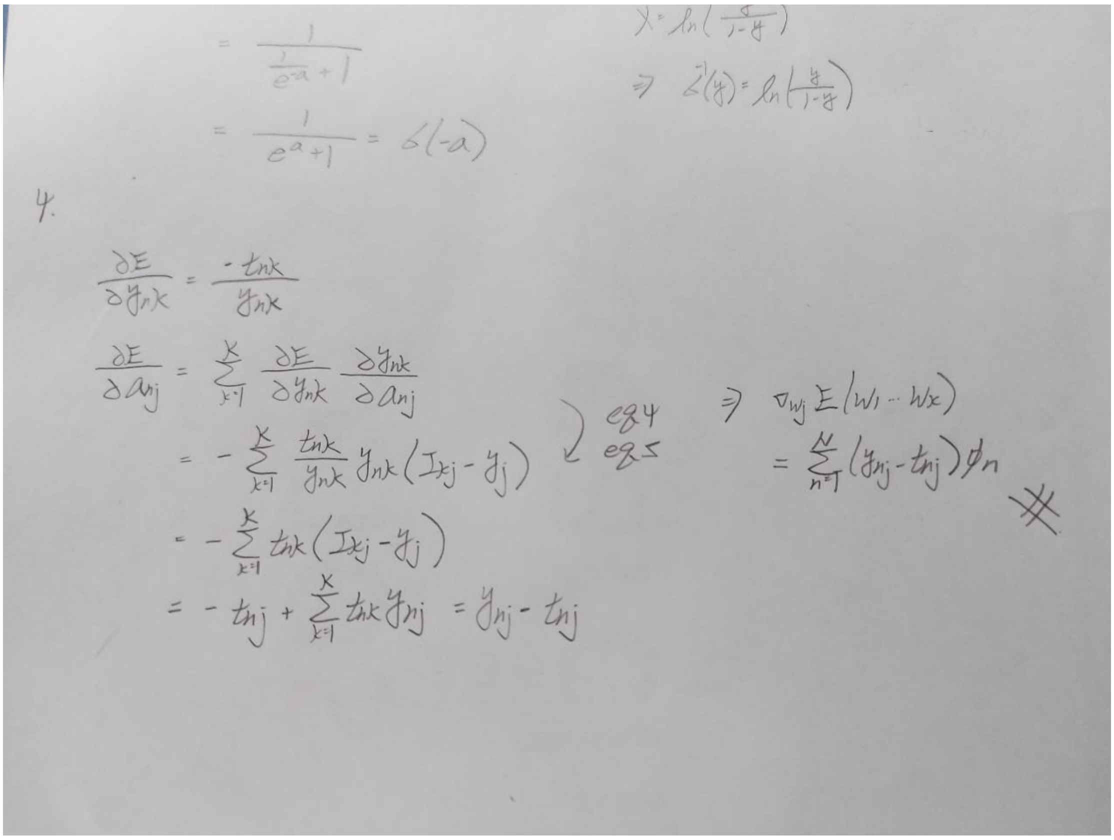
$$x = \ln\left(\frac{y}{1-y}\right)$$

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

4.

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

3).



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

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

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

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

4.

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

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

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

eq4 eq5

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

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

$$\Rightarrow \nabla_{w_j} E(w_1 \cdots w_K)$$

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

#