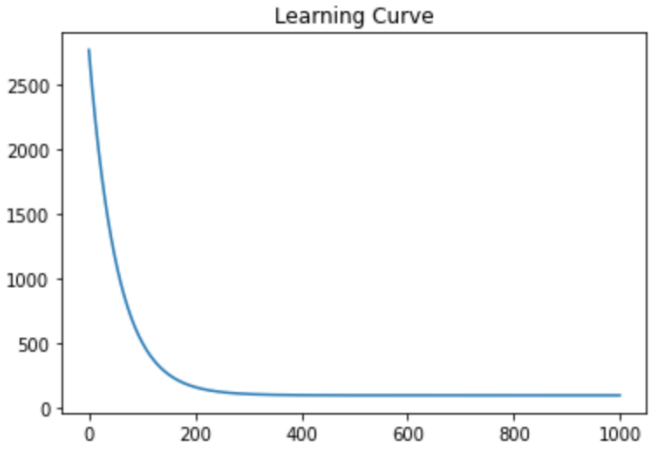
ML HW1 Report

109550171 陳存佩

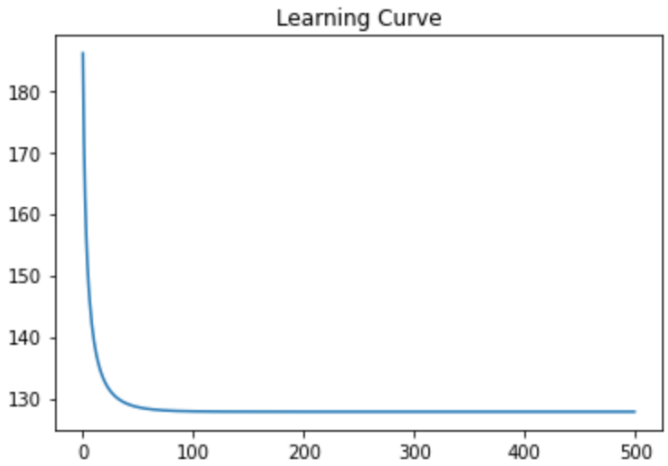
Part. 1, Coding (60%):

Linear regression model




```
Mean Square Error: 110.42352089914914
weights: 52.73882513770308
intercepts: -0.3344699947594075
```

Logistic regression model




```
cross_entropy: 47.24761018449381
weights: -4.876889778933664
intercepts: -1.7116223226069782
```

Part. 2, Questions (40%):

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

	Gradient Descent	Mini-Batch GD	Stochastic GD
Formula	$w_{i+1} = w_i - a \cdot \nabla_{w_i} J(w_i)$	$w_{i+1} = w_i - a \cdot \nabla_{w_i} J(x^{i:i+b}, y^{i:i+b}; w_i)$	$w_{i+1} = w_i - a \cdot \nabla_{w_i} J(x^i, y^i; w_i)$
Desc.	An **iterative** optimization algorithm: find the minimum of any **differentiable** function	Considered to be the cross-over between GD and SGD. Split the dataset into small subsets (**batches**) and compute the gradients for each subset.	A simplification of GD. **Randomly choose** partition of the dataset instead of using whole dataset to compute iterations.
Pros. & Cons.	1. Time-consuming and computationally expensive 2. May converge to a local minimum	1. Can use vectorized implementation for faster computations	1. Converge faster for larger datasets 2. May lead to noisier results than the GD

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

   Learning rate affect the speed of convergence.
   Larger learning rate leads to faster convergence, but it might skip the solution.
   Smaller learning rate causes slower convergence, so it needs more iterations to converge.

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)

3. $1-\sigma(a)=1-\frac{1}{1+\exp(-a)}=\frac{1+\exp(-a)-1}{1+\exp(-a)}=\frac{\exp(-a)}{1+\exp(-a)}$

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

et $f(x)=\frac{1}{1+\exp(-x)}$

$\frac{1}{f(x)}=1+\exp(-x)$

$\exp(-x)=\frac{1}{f(x)}-1$

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

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

$y=f(x),\ f^{-1}(y)=x$

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

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

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)

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

$\frac{\partial E}{\partial a_{nj}}=\sum_{k=1}^{K}\frac{\partial E}{\partial y_{nk}}\frac{\partial y_{nk}}{\partial a_{nj}}$ → eq5

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

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

$\Rightarrow \nabla_{w_j}E(w_1,\ldots,w_K)=\sum_{n=1}^{N}(y_{nj}-t_{nj})\phi_n$ #