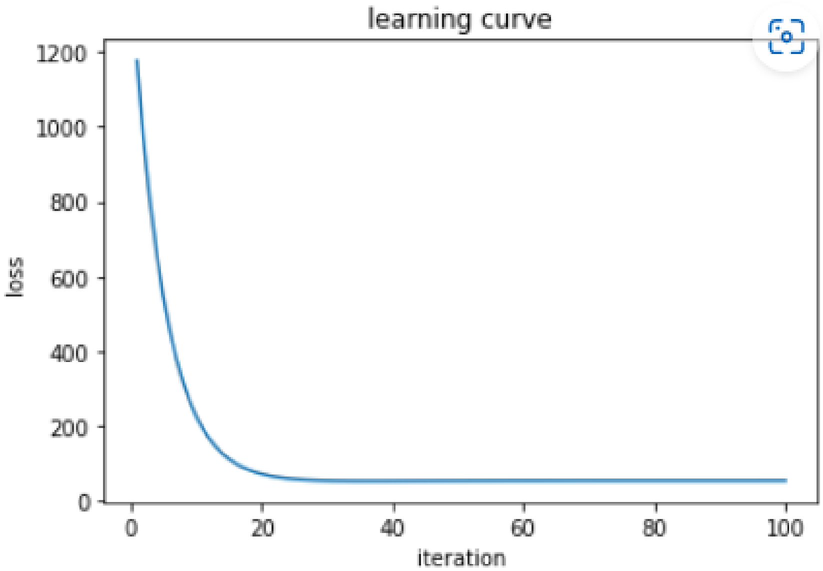
Part. 1

Linear regression model

1.

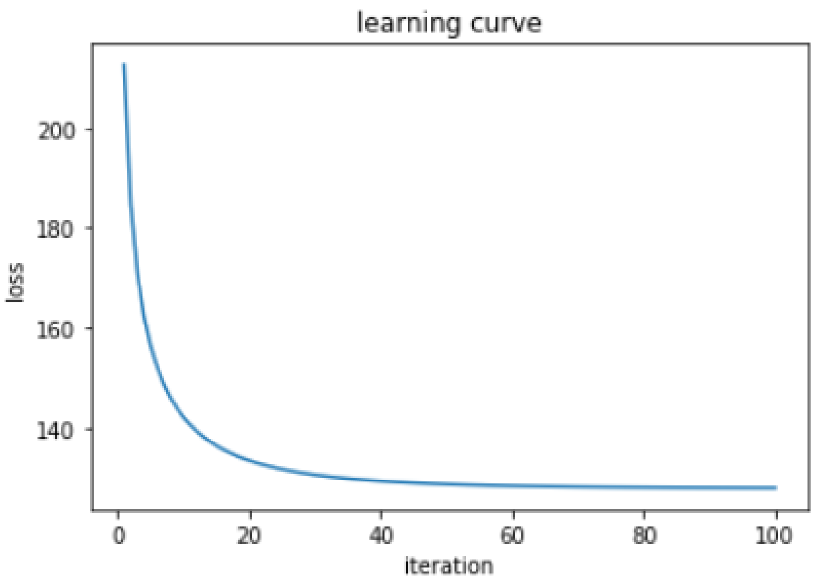


2+3.

```
Mean Square Error:  55.21442307485028
weights:  52.74050648059757
intercept:  -0.33418947537034915
```

Logistic regression model

1.



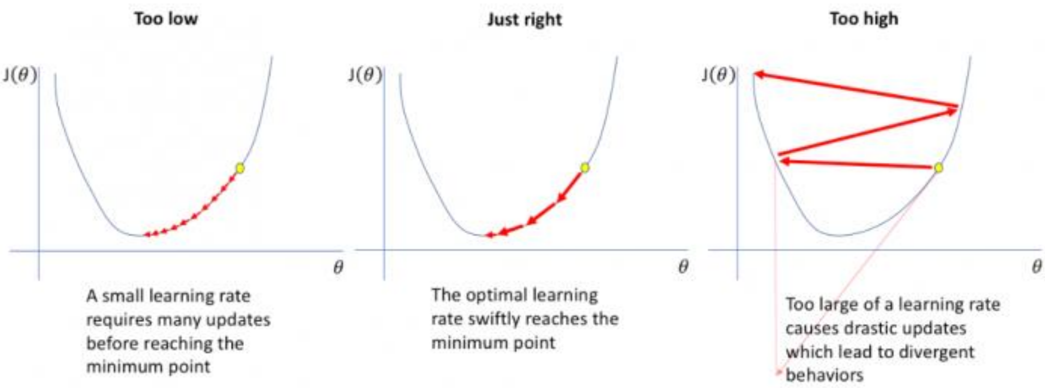
2+3.

```
Cross Entropy Error:  46.75820789486666
weights:  4.729475975042 1145
intercept:  1.621861070934789
```

Part. 2

1. Gradient Descent 是看過整筆資料過後再 update 一次 weight 和 bias； Mini-Batch Gradient Descent 是會先決定一個 batch size，接著每看完一個 batch 就 update 一次；Stochastic Gradient Descent 則是每看過一個點都會直接 update 一次參數。也就是說，這三者的差異主要在 update 參數的頻率。

2. 會，learning rate 要設剛好才會 converge。首先如果太大的話，update 參數的級距會太大，造成做 training 的時候一直來回震盪(一直 update 過頭)；如果太小，則可能讓 training 卡住(因為每次只 update 一點點)，或是訓練太慢。



3.

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

$$= \left(\frac{1+e^{a}}{1+e^{a}}\right)\cdot\frac{1}{1+e^{-a}} + \left(\frac{1+e^{-a}}{1+e^{-a}}\right)\cdot\frac{1}{1+e^{a}}$$

$$= \frac{2+e^{a}+e^{-a}}{(1+e^{-a})(1+e^{a})}$$

$$= \frac{2+e^{a}+e^{-a}}{1+e^{-a}+e^{a}+e^{0}} = \frac{2+e^{a}+e^{-a}}{2+e^{a}+e^{-a}} = 1$$

$\Rightarrow \sigma(-a) = 1 - \sigma(a)$

② let $y = \frac{1}{1+e^{-x}}$ $(y = \sigma(x))$

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

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

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

$-x = \ln \frac{1-y}{y}$

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

4.

4. $a_{nj} = w^{T}\phi_n$

$\nabla_{w_j} a_{nj} = \phi_n$

$\frac{\partial E}{\partial y_{nk}} = -\frac{t_{nk}}{y_{nk}}$ $\frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j)$

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

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

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