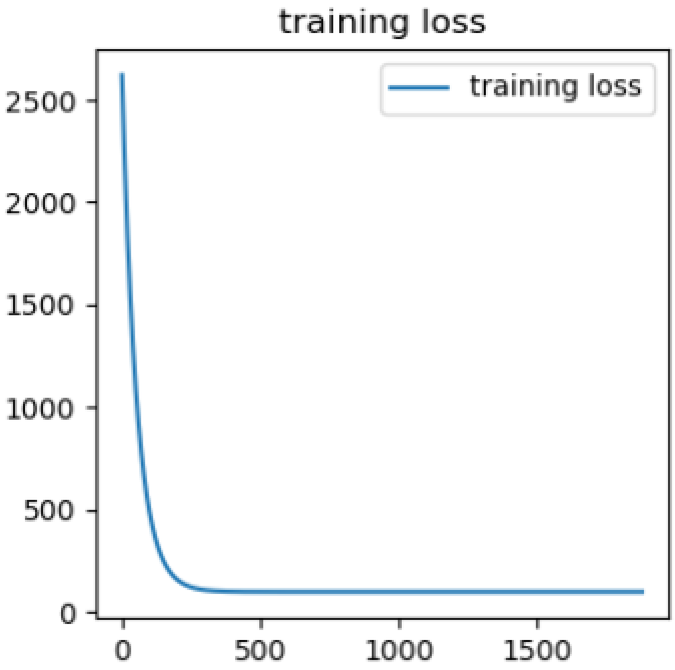
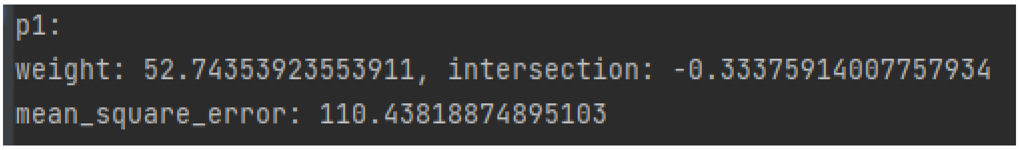
Part.1 Coding

Linear regression model

1. Learning Curve:



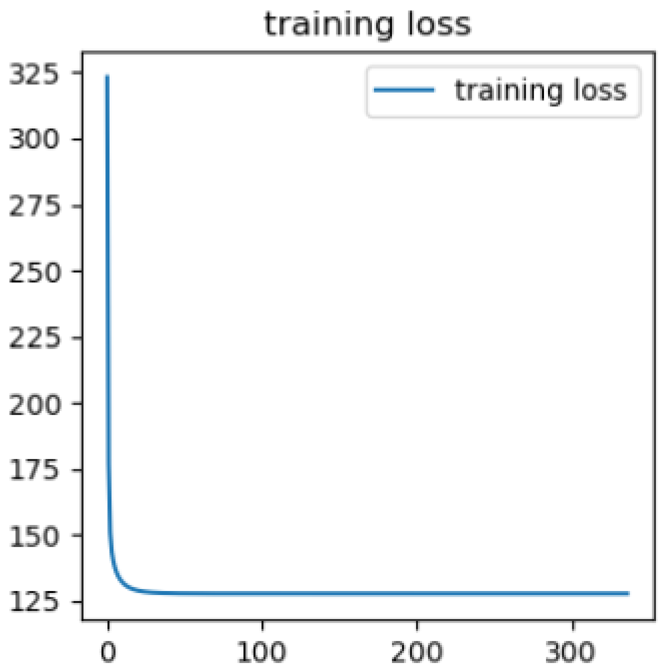
2. Mean square error: 110.43
3. Weight: 52.74, intersection: -0.33



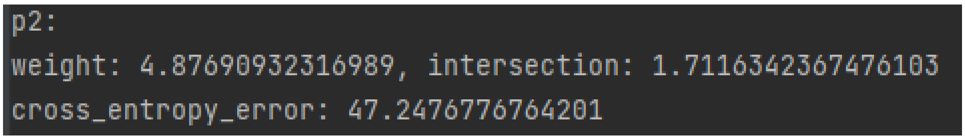
```
p1:
weight: 52.74353923553911, intersection: -0.33375914007757934
mean_square_error: 110.43818874895103
```

Logistic regression model

1. Learning Curve:



2. Cross entropy error: 47.24
3. Weight: 4.87, intersection: 1.71



```
p2:
weight: 4.87690932316989, intersection: 1.7116342367476103
cross_entropy_error: 47.2476776764201
```

Part 2. Questions

1. What’s the difference between Gradient Descent, Mini-Batch Gradient Descent, and Stochastic Gradient Descent?
   - Gradient Descent: Feed in all of the training data for every iteration. May be slow for each iteration when dataset is too huge.
   - Mini-Batch Gradient Descent: Feed in a subset of training data for every iteration. Capable to deal with large dataset, but needs more iteration to cover the whole dataset.
   - Stochastic Gradient Descent: Feed in a single data to train for every iteration. Learns very fast, but can’t ensure it’s converging to correct direction.
2. Will different values of learning rate affect the convergence of optimization? Please explain in detail.
   - High learning rate would cause the model to converge faster, but setting to high might cause the model keep swinging around the minimum and not converging.

- Low learning rate gives the model more precision, but slower to converge. Setting too low might cause the model stuck in a local minimum, instead of global minimum.

3.

$$\sigma(\mathrm{a}) = \frac{1}{1+\exp(-\mathrm{a})} = \mathrm{y} = \sigma\left(\sigma^{-1}(\mathrm{y})\right)$$

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

$$\exp(-\mathrm{a}) = \frac{1-\mathrm{y}}{\mathrm{y}}$$

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

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

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

4.

$$\nabla_{\mathrm{w_j}}\mathrm{E} = -\sum_{\mathrm{n}=1}^{\mathrm{N}}\sum_{\mathrm{k}=1}^{\mathrm{K}}\frac{\partial}{\partial \mathrm{a_j}}\frac{\partial \mathrm{a_j}}{\partial \mathrm{w_j}}\mathrm{t_{nk}}\ln \mathrm{y_{nk}}$$

$$\frac{\partial \mathrm{a_j}}{\partial \mathrm{w_j}} = \phi_{\mathrm{n}}$$

$$\Rightarrow -\sum_{\mathrm{n}=1}^{\mathrm{N}}\sum_{\mathrm{k}=1}^{\mathrm{K}}\frac{\partial}{\partial \mathrm{a_j}}\frac{\partial \mathrm{a_j}}{\partial \mathrm{w_j}}\mathrm{t_{nk}}\ln \mathrm{y_{nk}} = -\sum_{\mathrm{n}=1}^{\mathrm{N}}\sum_{\mathrm{k}=1}^{\mathrm{K}}\mathrm{t_{nk}}\,\phi_{\mathrm{n}}\frac{\partial}{\partial \mathrm{a_j}}\ln \mathrm{y_{nk}}$$

$$= -\sum_{\mathrm{n}=1}^{\mathrm{N}}\sum_{\mathrm{k}=1}^{\mathrm{K}}\phi_{\mathrm{n}}\frac{\mathrm{t_{nk}}}{\mathrm{y_{nk}}}\left(\mathrm{I_{kj}} - \mathrm{y_{nj}}\right)\mathrm{y_{nk}}$$

$$= -\sum_{\mathrm{n}=1}^{\mathrm{N}}\sum_{\mathrm{k}=1}^{\mathrm{K}}\phi_{\mathrm{n}}\mathrm{t_{nk}}\left(\mathrm{I_{kj}} - \mathrm{y_{nj}}\right)$$

$$= -\sum_{\mathrm{n}=1}^{\mathrm{N}}\phi_{\mathrm{n}}\left(\sum_{\mathrm{k}=1}^{\mathrm{K}}\mathrm{t_{nk}}\mathrm{I_{kj}} - \sum_{\mathrm{k}=1}^{\mathrm{K}}\mathrm{t_{nk}}\mathrm{y_{nj}}\right)$$

Note that I_{kj} = {1 when k=j, 0 otherwise} and sum of t_{nk} over k is 1.

So, we can get $\sum_{k=1}^{K} t_{nk}I_{kj} = t_{nj}$ and $\sum_{k=1}^{K} t_{nk}y_{nj} = y_{nj}$.

$$\boxed{\Rightarrow \nabla_{w_j} E = -\sum_{n=1}^{N} \phi_n \left(t_{nj} - y_{nj} \right) = \sum_{n=1}^{N} \phi_n \left(y_{nj} - t_{nj} \right) \quad \blacksquare}$$