Report of

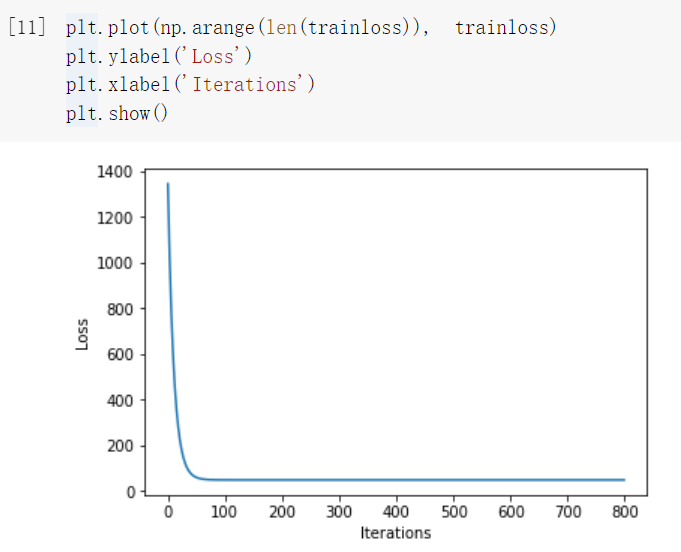
NYCU Introduction to Machine Learning, Homework 1

**109550027 紀竺均**

**Part. 1, Coding (60%)**:

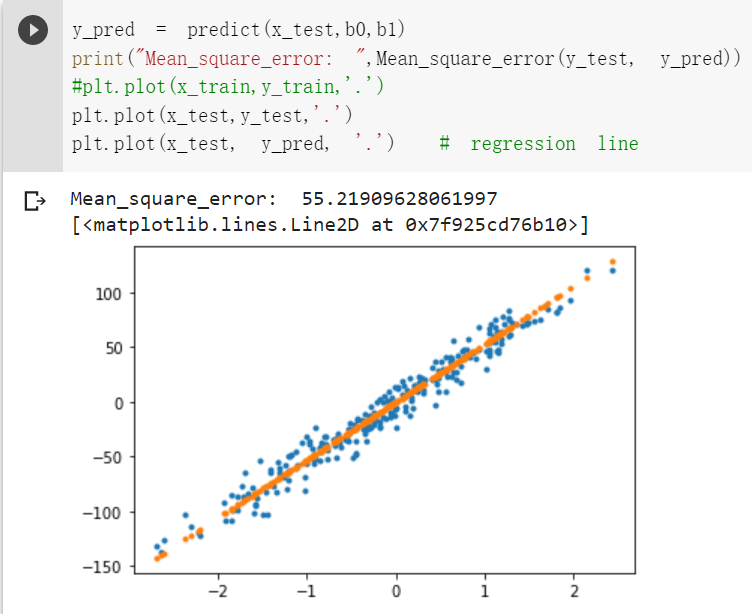
**Linear regression model**

1. (10%) Plot the [learning curve](https://miro.medium.com/max/1718/1*69r3IZNMnWPngwzbeSTz_Q.png) of the training, you should find that loss decreases after a few iterations and finally converge to zero  (x-axis=iteration, y-axis=loss, Matplotlib or other plot tools is available to use)



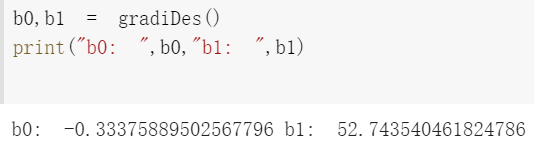
1. (10%) What’s the [Mean Square Error](https://zh.wikipedia.org/zh-tw/%E5%9D%87%E6%96%B9%E8%AF%AF%E5%B7%AE) of your prediction and ground truth?

Ans: Mean\_square\_error: 55.21909628061997



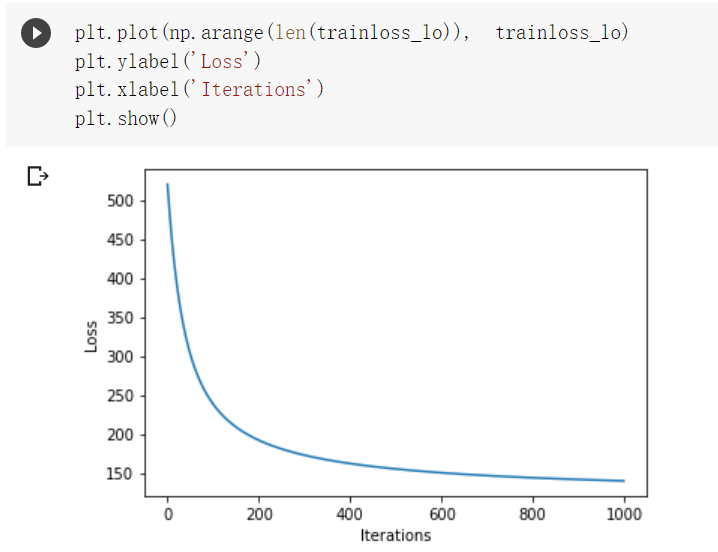
1. (10%) What’re the weights and intercepts of your linear model?

Ans: weights: 52.743540461824786, intercepts: -0.33375889502567796



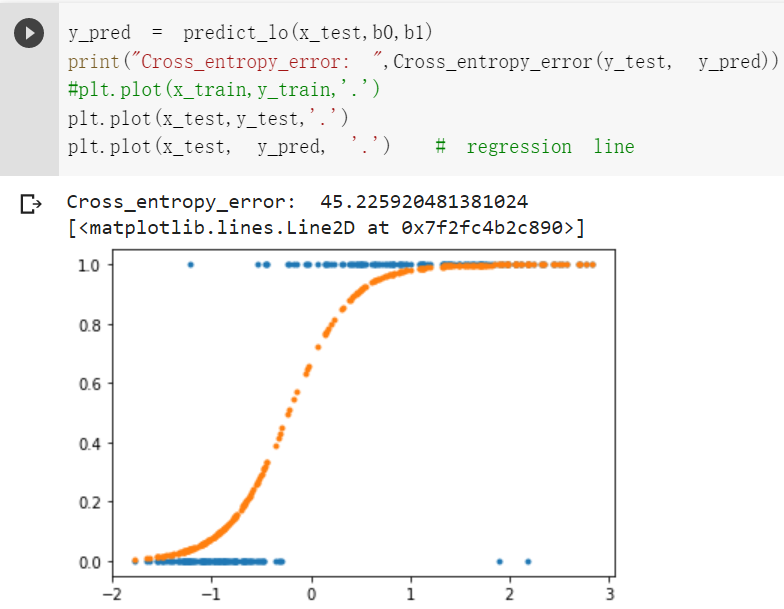
**Logistic regression model**

1. (10%) Plot the [learning curve](https://miro.medium.com/max/1718/1*69r3IZNMnWPngwzbeSTz_Q.png) of the training, you should find that loss decreases after a few iterations and finally converge to zero  (x-axis=iteration, y-axis=loss, Matplotlib or other plot tools is available to use)



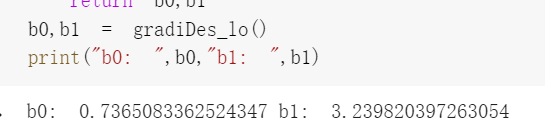
1. (10%) What’s the [Cross Entropy Error](https://en.wikipedia.org/wiki/Cross_entropy) of your prediction and ground truth?

Ans: Cross\_entropy\_error: 45.225920481381024



1. (10%) What’re the weights and intercepts of your linear model?

Ans: weights: 3.239820397263054, itercepts: 0.7365083362524347



**Part. 2, Questions (40%):**

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

Ans: The main difference is the number of training data taken into consideration within a single step. For Gradient Descent, it takes **all the training data** into consideration; For Stochastic Gradient Descent, it only consider **one example** at a time; Finally, Mini-Batch Gradient Descent is a mixture of Gradient Descent and SGD, it takes **a batch of fixed number** of training data into consideration.

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

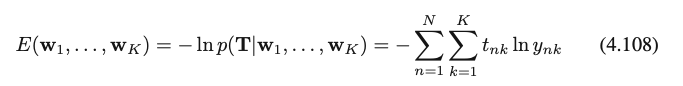
Ans: Yes, if the learning rate is too large, gradient descent might overshoot the minimum. It may fail to converge. On the other hand, if the learning is too small, it would take more steps and more time to reach local minimum.

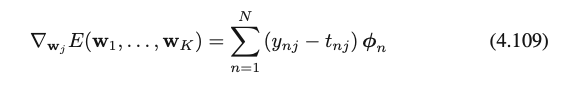
3. Show that the logistic sigmoid function (eq. 1) satisfies the property σ(−a) = 1 − σ(a) and that its inverse is given by σ−1(y) = ln {y/(1 − y)}.



Ans: Next Page

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

(eq.2)

(eq.3 )

https://lh5.googleusercontent.com/iCWqnmzycccmBbmJwLP9X8I93NUhp233pKAxMgd5n_4FpT7c_MzDwXNSzdm1FvfwB_nHxQMujc1xVHvZGHn7F2SXcgEmX5ajSGQdNI6nDC-mJIdaaZEfkXgttpR0JtqNPVVSjr0DpDXUJZ9vNCZYdbWAIGslU8irTvjYvvit5fcc5kXUZq0aDS3F(eq. 4)

(eq. 5)

Ans: Next Page

