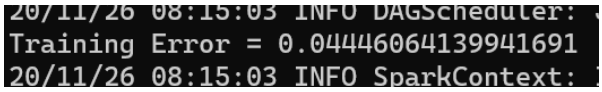
Cloud Computing and Security HW4

電機四 B06901087 翁瑋襄

Github URL: https://github.com/jill880829/CloudComputingHW4.git

PART 1: Reproduce the result on Spark (in run-spark-ex.py):

I created three docker containers for Spark service with the .yml file offered by the professor. Then I installed numpy package while entering the master container with root. I also modified the sample code to run this Logistic Regression example. One is wrapping data features and labels with LabeledPoint(), and the other is change the notation of tuple in lambda function. Finally, I ran the sample code with spark-submit and get the following result.



PART 2: Implement Logistic Regression with numpy module (in LR_numpy.py):

I implemented Logistic Regression by gradient descent with the following parameters: w_init = [0, 0, 0, 0], b_init = 0, 10000 epoches, learing_rate=5/sigma**0.5 where sigma equals to the summation of gradient**2 in each iteration. The result is as following.

Comparison:

I found that training error of the Spark version is higher than the other one. I guess the reason is the initial setting of the two platforms are different. Besides, due to the small size of dataset, the running time on Spark is longer than pure python. If the dataset is large enough, Spark will show the power of computation.