學號：b07902034　系級：資工二　　姓名：王昱凱

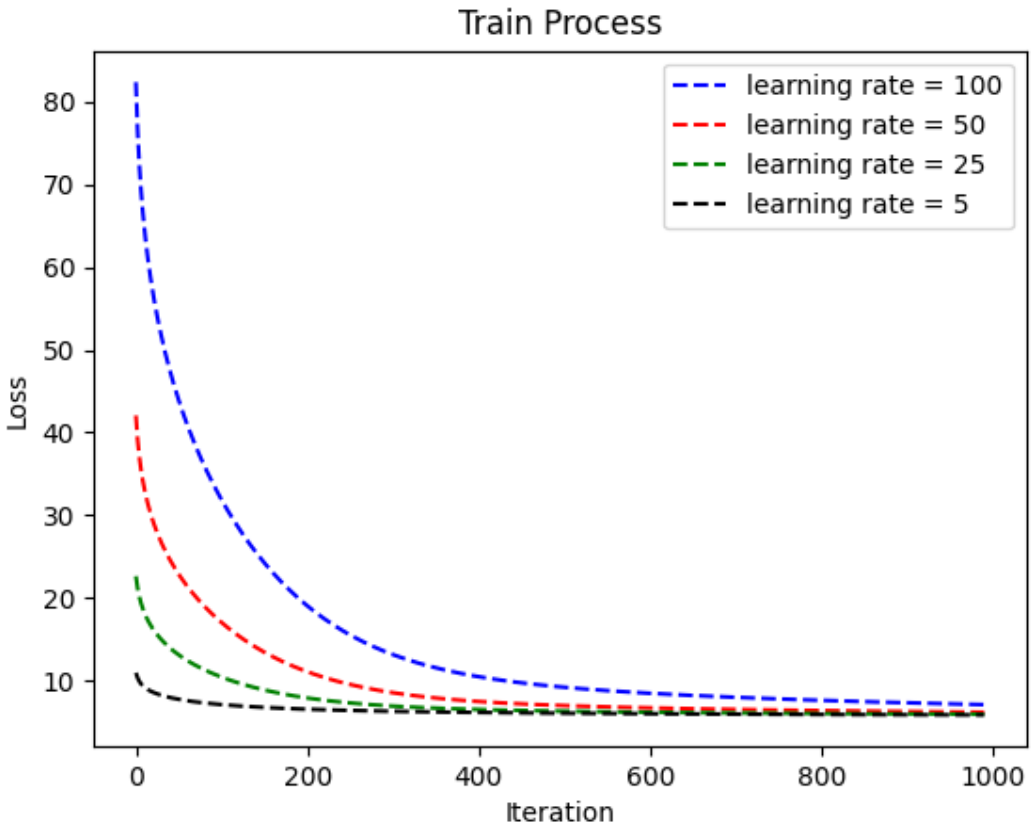
備註：

a. 1~3 題的回答中，NR 請皆設為 0，其他的數值不要做任何更動。

b. 可以使用所有 advanced 的 gradient descent 技術（如 Adam、Adagrad）。

c. 1~3 題請用 linear regression 的方法進行討論作答。

1. (2%) 使用四種不同的 learning rate 進行 training (其他參數需一致)，作圖並討論其收斂過程（橫軸為 iteration 次數，縱軸為 loss 的大小，四種 learning rate 的收斂線請以不同顏色呈現在一張圖裡做比較）。



我挑選的四種 learning rate 為 100, 50, 25, 5，透過作圖顯示我發現 learning rate 越大收斂過程的幅度就越大，而由於使用的是 adagrad，所以隨著 iteration 的次數越多，收斂的幅度會越小，但或許是因為 iteration 的次數還不夠多，因此四種不同 learning rate 沒有收斂到一樣的大小，仍然有一點點的差距，而當 learning rate 為 5 的時候，有著最小的 loss，因此為四種情況中較佳的 learning rate 大小。

2. (1%) 比較取前 5 hrs 和前 9 hrs 的資料（5*18 + 1 v.s 9*18 + 1）在 validation set 上預測的結果，並說明造成的可能原因（1. 因為 testing set 預測結果要上傳 Kaggle 後才能得知，所以在報告中並不要求同學們呈現 testing set 的結果，至於什麼是 validation set 請參考：https://youtu.be/D_S6y0Jm6dQ?t=1949 2. 9hr:取前 9 小時預測第 10 小時的 PM2.5；5hr:在前面的那些 features 中，以 5~9hr 預測第 10 小時的 PM2.5。這樣兩者在相同的 validation set 比例下，會有一樣筆數的資料）。

以助教在 colab 提供的 training model 進行測試下，取 5 小時在 validation set 上所得到的 loss 為 5.753714491511453，而取 9 小時的 loss 則為 5.912205466286512，由於在取 5 小時的情況下我們是取 5~9 小時的資料，因此由 loss 的結果我們可以推測前 4 小時的資料對於預測第 10 小時的 PM2.5 是較為沒有幫助的，而為了驗證這點，我也測試了取 1~4 小時的資料來預測，結果 loss 的大小為 13 左右，因此造成取 9 小時的資料做預測有較差結果的原因就在於前 4 小時的資料對預測第 10 小時沒有太大的幫助。

3. (1%) 比較只取前 9 hrs 的 PM2.5 和取所有前 9 hrs 的 features（9*1 + 1 vs. 9*18 + 1）在 validation set 上預測的結果，並說明造成的可能原因。

以助教在 colab 提供的 training model 進行測試下，我只取 PM2.5 在 validation set 上所得到的 loss 為 5.861175212293，而取所有 features 的 loss 則為 5.912205466286512，取所有 features 的結果較差一些，我認為原因是有許多 features 的數值與 PM2.5 是無關的，因此如果把這些 feature 的數值都考慮進去並進行 training，那麼可能會造成預測結果的偏差。

4. (2%) 請說明你超越 baseline 的 model(最後選擇在 Kaggle 上提交的) 是如何實作的（例如：怎麼進行 feature selection, 有沒有做 pre-processing、learning rate 的調整、advanced gradient descent 技術、不同的 model 等等）。

我最後選擇的 model 是使用 adagrad，我也嘗試過 adam 和 momentum 等 gradient descent 的技術，但是我發現與 adagrad 的差異並不大甚至差一些，而 feature 我只挑選了 18 項數據中的 6 項，我挑選的 feature 有 AMB_TEMP, NO, NO2, NOx, PM2.5, SO2，原因是我透過網路上的查詢和我自己的測試發現 PM2.5 和氮氧化物、硫氧化物、溫度等因素較為相關，而 learning rate 我使用的是 1，原因是讓一開始的 learning

rate 是 100，如果 iteration 次數夠多的話結果是不會差到太多，但我選擇的 iteration 次數是 100000，在這情況下如果 learning rate 太小就會收斂得太慢，而太大的話次數不夠多，因此我最後認為 learning rate 設成 1 較為合適。