學號：B05705011 系級：資管四 姓名：楊子霖

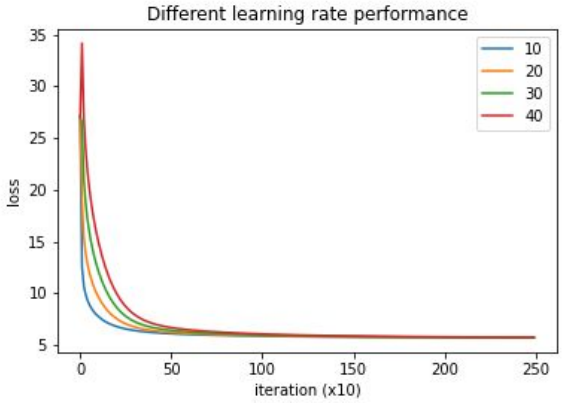
備註：

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）。

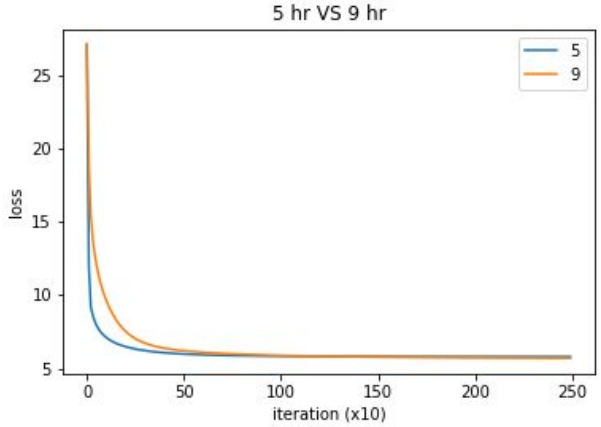
c. 1~3題請用linear regression的方法進行討論作答。

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）。



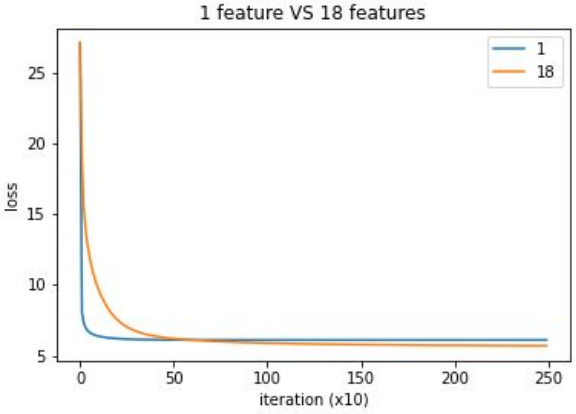
嘗試了 learning rate 分別為 10, 20, 30, 40，並且 iter_time = 2500，並且每 10 次 iteration 紀錄一次 loss，可發現反而是數字愈小的 learning rate 愈快趨於平穩，推測可能是 learning rate 設得不夠小的關係，才會在每次調 model 的時候調過頭直到 learning rate 影響降低為止，以至於 learning rate 小的反而最快趨於平穩。

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）。



嘗試了 learning rate = 10，並且 iter_time = 2500，可發現用每 5 ~ 9 個小時預測的 loss 比較快達到平緩，然而最後的 loss 卻是略高於每 9 個小時預測的 loss，前者約為 5.82，後者則約為 5.72，每 9 小時的結果較好的可能原因是因為 model 比 5 ~ 9 小時的複雜，並且它的 model 可能性可以涵蓋 5 ~ 9 小時的 model 的所有可能性，因此隨著 iteration 的增加，結果肯定只會更有可能比對方好而已。

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。



嘗試了 learning rate = 10，並且 iter_time = 2500，只使用 PM25 預測的結果 loss 大約為 6.12，使用所有 features 的結果則約為 5.72，後者結果較好的原因和第二題的狀況類似，後者的 model 結果所形成的集合涵蓋前者的 model 集合，而且只取 PM2.5 的 model 太過簡單，因此預測結果比使用全部 features 不好的可能性很大。

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）。

我一開始是隨機刪除幾個 feature 並丟到 kaggle 試試，分別是AMB_TEMP, CH4, RAINFALL, RH, SO2, THC, WD_HR, WIND_DIREC, WIND_SPEED, WS_HR，learning rate 設為 10、iteration 次數為 5000，並且使用 Adagrad，就過 strong baseline 了，所以也沒有嘗試其他 feature selection 的方式。