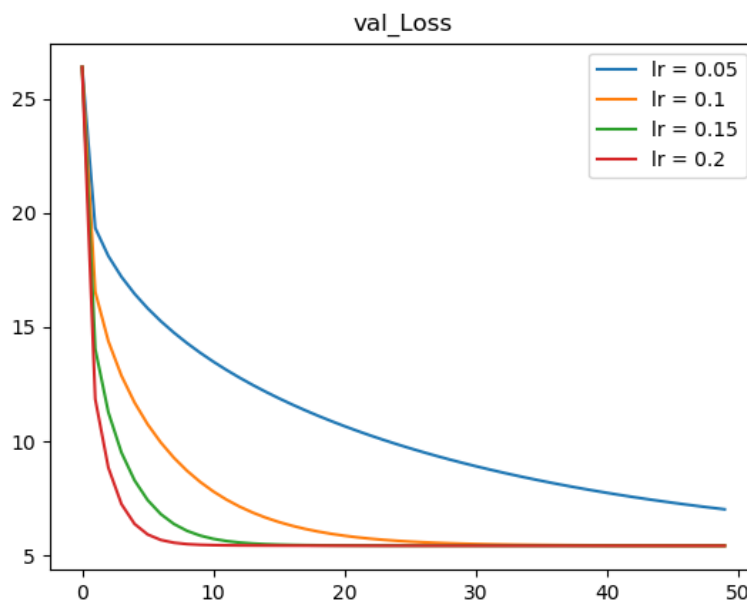


學號：B07902040 系級：資工二 姓名：吳承軒

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）。

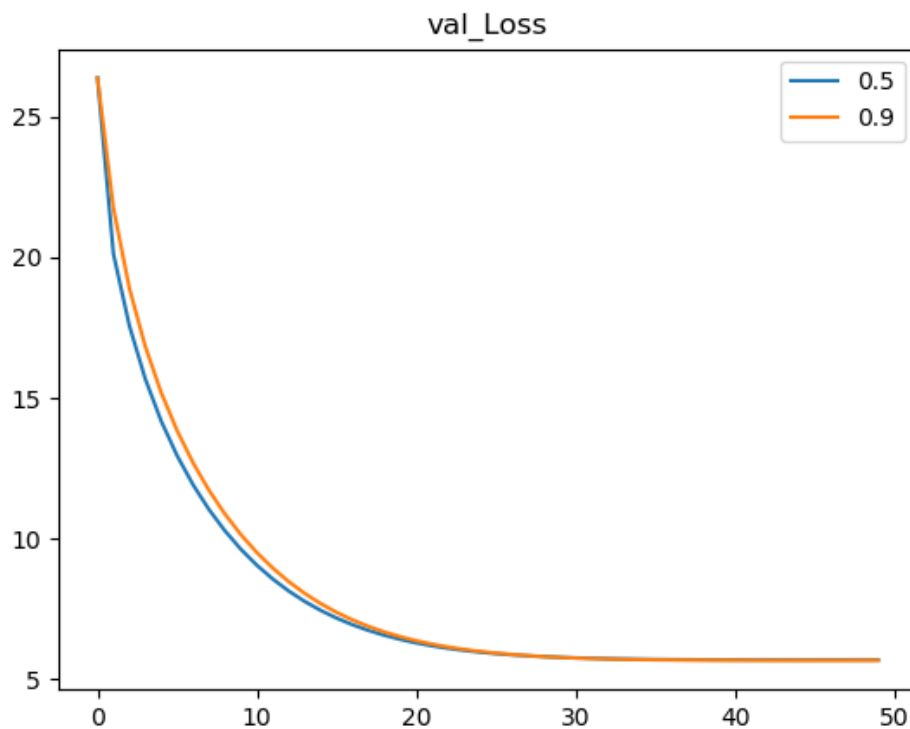


縱軸為 loss，橫軸為 itertime(\*1000)

可以發現隨著 learning rate 越大，loss 也下降的越快，且除了 0.05 的尚未到達底部，其他三個分別在不同次數時開始產生 overfitting。

2. (1%) 比較取前 5 hrs 和前 9 hrs 的資料 ( $5 \times 18 + 1$  v.s  $9 \times 18 + 1$ ) 在 validation set 上預測的結果，並說明造成的可能原因（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：

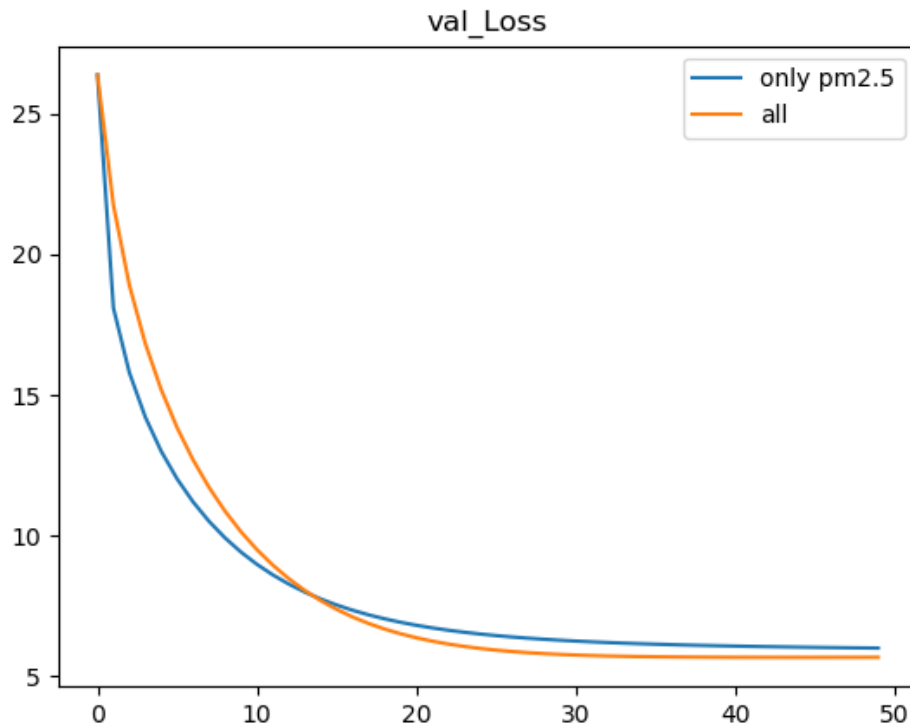
[https://youtu.be/D\\_S6y0Jm6dQ?t=1949](https://youtu.be/D_S6y0Jm6dQ?t=1949)）。



縱軸為 **loss**，橫軸為 **itertime(\*1000)**

可以發現只取前五小時資料時，loss 的下降速度較取 9 小時來的快，可能是因為前五小時的資料距離所要預測的時間較近，相關性更高，但在次數足夠多之後，取五小時和取九小時的 loss 就趨近相同了。

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



縱軸為 **loss**，橫軸為 **itertime(\*1000)**

可以發現一開始時，只取用 pm2.5 作為 train data 的 loss 是下降比較快的，可能是因為我們要預測的對象本身就是 pm2.5，只取用 pm2.5 的 data 的相關性比較高。但在 itertime 到 13000 左右時，取用全部資料的 loss 開始小於 pm2.5 的，可能是因為其他資料雖然相關性較低，但也非完全無關，所以雖然一開始因整體相關性不如 pm2.5 來得高，而使 loss 下降較慢，但其所能達到 loss 的下限是相對較低的。

**4. (2%) 請說明你超越 baseline 的 model 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）。**

設定 itertime 為 10000，learning rate 為 1，對 train data 和 test data 做 pre-processing: 將所有非溫度的項，小於 0 的改為 0(因為估計非溫度的項小於 0 應是觀測錯誤或是記錄錯誤)。扣除了跟溫度及和風速有關的 WD\_HR、WIND\_DIREC、WIND\_SPEED、WS\_HR 五項 feature，並使用 **Adagrad** 改進。