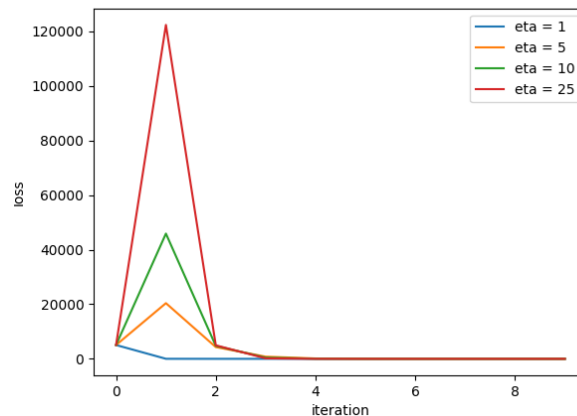
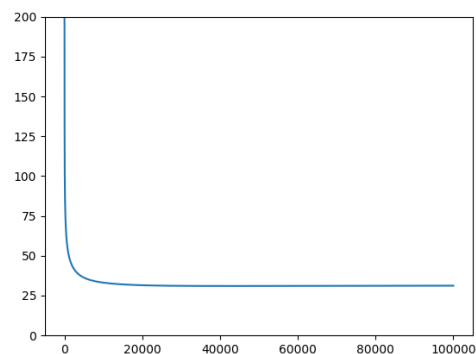
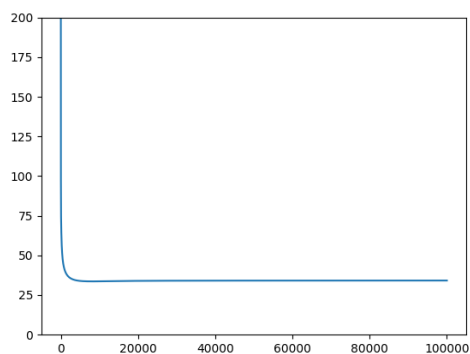


1. (2%) 使用四種不同的 learning rate 進行 training (其他參數需一致)，作圖並討論其收斂過程 (橫軸為 iteration 次數，縱軸為 loss 的大小，四種 learning rate 的收斂線請以不同顏色呈現在一張圖裡做比較)。



Learning rate 較大時，有可能一更新就跳到 loss 更大的地方，因此在 eta=25 時，有一個顯著的峰。然而 learning rate 越大時，在沒有跳到錯誤區段的前提下，更新的速度也確實有比較快。

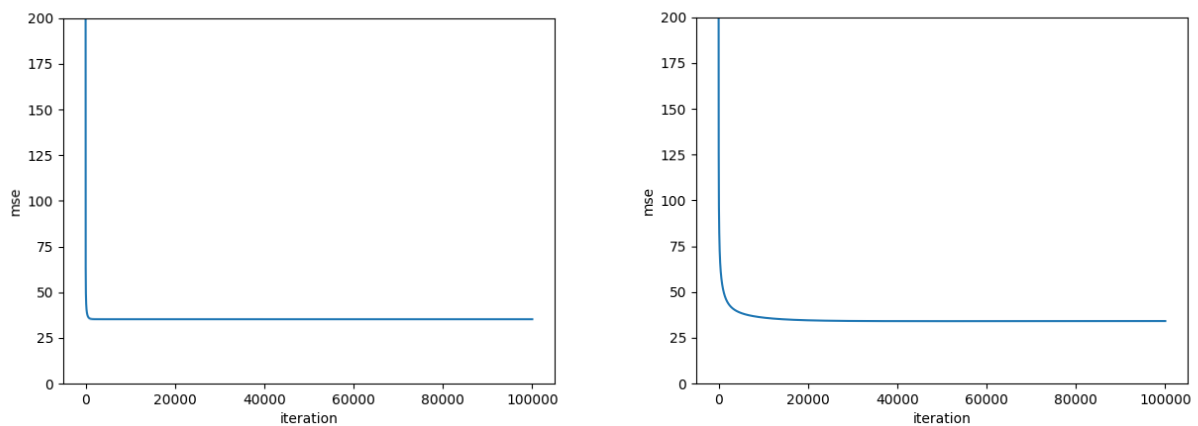
2. (1%) 比較取前 5 hrs 和前 9 hrs 的資料 ($5 \times 18 + 1$ v.s $9 \times 18 + 1$) 在 validation set 上預測的結果，並說明造成的可能原因 (1. 因為 testing set 預測結果要上傳 Kaggle 後才能得知，所以在報告中並不要求同學們呈現 testing set 的結果，至於什麼是 validation set 請參考：https://youtu.be/D_S6y0Jm6dQ?t=1949 2. 9hr:取前 9 小時預測第 10 小時的 PM2.5；5hr:在前面的那些 features 中，以 5~9hr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



- 第二、三題的 validation set 都是取 2:8

在只取前五小時（左圖）的訓練上，相較於取九小時（右圖），訓練的速度較快、較早達到 overfit（約 8000 次 iteration 相對於 44000 次）但 MSE 也比較大（33.6 相對於 31.0）。這樣的結果可能是因為取五小時的資料明顯不足於取九個小時，因此在 validation set 上理所當然的會有較大的錯誤。

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。



在只取前 PM2.5（左圖）的訓練上，相較於取全部 features（右圖），訓練的速度較快、較早達到 overfit（約 29000 次 iteration 相對於 75000 次）但 MSE 也比較大（35 相對於 34）。這樣的結果是因為只取一項 feature，可能會少考慮到很多影響結果的因素，就像不判斷寶可夢的物種便難以預測 cp 值一樣，因此也在 validation set 上有較大的 error。

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）。

訓練過程為：a. 將九小時的全部 feature 當作 X，再加上一個 bias=1，共 163 維；並把接下來的 PM2.5 值當作 y。b. 處理負數：因值不該有複數，所以對所有負值做線性插值來去掉 noise。c. 按照投影片規定的方式做 gradient descent，iteration 採用差不多會 overfit 的值，learning rate 由第一題，取個中規中矩的值，即可過 baseline。另外我不知道有提供 code 所以可能跟別人的不太一樣，還請包涵。