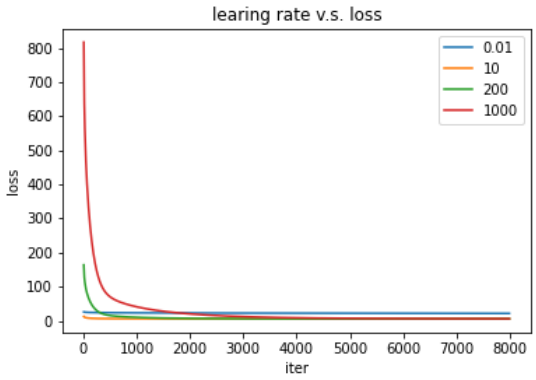
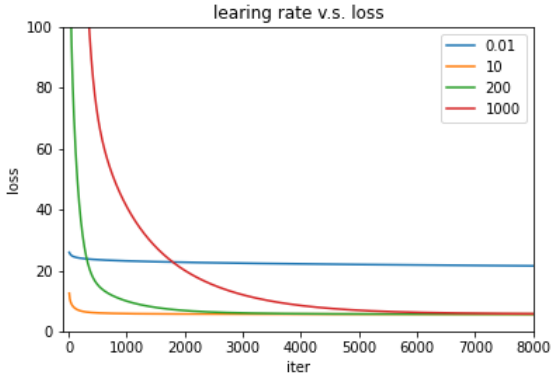
HW1 Report

學號：b07901039　　系級：電機二　　姓名：劉知穎

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）。





使用 Adagrad gradient descent。

Learning Rate	收斂時的 iteration(大約)
0.01	>8000
10	300
200	3000
1000	6000

當 learning rate 太大或太小時，需要更多的 iteration 才會收斂。（當 learning rate 為 0.01 時，至 iteration 為 8000 時都還未收斂，依然有緩慢地降低 loss。）經過實驗，learning rate 大約在 5 到 10 之間能達到最好的效果，大約在 iteration 200 到 300 之間收斂。

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。

	9 hrs		5 hrs	
	train loss	val loss	train loss	val loss
0	5.689748	5.918724	5.946058	5.294696
1	5.588991	6.300199	5.828269	5.816074
2	5.706921	5.828209	5.891053	5.536816
3	5.585783	6.261362	5.779897	6.009823
4	5.665061	5.980282	5.782203	6.005159
5	5.651441	6.062477	5.849459	5.73085
6	5.767891	5.538575	5.771555	6.04321
7	5.760855	5.594234	5.812187	5.89095
8	5.603745	6.214653	5.775929	6.057391
9	5.73018	5.700647	5.758881	6.101499
avg	5.675062	5.939936	5.819549	5.848647

經過十次實驗，每次將前 5hrs 和前 9hrs 的資料進行相同的 shuffle。只取前 5hrs 的資料在 validation set 上的預測結果較好。推測原因為取前 9hrs 的資料時造成”overfitting”的現象。

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。

	all features		only pm2.5	
	train loss	val loss	train loss	val loss
0	5.656782	5.985104	6.112105	6.176684
1	5.739233	5.660917	6.156142	5.996668
2	5.531615	6.457651	5.997874	6.61532
3	5.78032	5.537788	6.204739	5.799829
4	5.672091	5.926444	6.134959	6.08681
5	5.716021	5.801193	6.184203	5.884498
6	5.61382	6.192817	6.007995	6.575708
7	5.768247	5.565631	6.198125	5.838065
8	5.616255	6.152458	6.084799	6.290843
9	5.638325	6.081619	6.087447	6.268224
avg	5.673271	5.936162	6.116839	6.153265

經過十次實驗，每次將取全部 features 和只取 PM2.5 的資料進行相同的 shuffle。取全部 features 在 validation set 上的預測結果較好。推測原因是決定 PM2.5 濃度的因素，除了前幾小時的 PM2.5 之外，還有其他重要的因素。因此只取前 9 hrs 的 PM2.5 太過於簡化模型，預測的準確值降低。

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）。

(1) pre-processing：

因為測資中的 18 個 features 都應該是正值，在做資料處理時卻發現有些值是負值，推測這些負值的資料是有偏差的，因此將資料中的負值補成零。（有嘗試過以平均值取代負值的資料，但效果較差。）

(2) feature selection：

以 Standardize Beta Coefficient 評估 162 筆(18*9) features 的重要程度。(有嘗試過以原本的 18 個 features 評估，但效果較差，一次評估全部 162 個 features 能將時間的因素也考慮進去。) 經過實驗，留下前 50 個 features 的預測較準確。

對於前幾個較重要的 features，加上次方項，經過實驗決定加上四項。

(3) advanced gradient descent：

使用 Adagrad 的 gradient descent。

(4) model：

用手刻的 linear regression。

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。