學號：F05942036 系級： 電信博一 姓名：羅翊展

請實做以下兩種不同 feature 的模型，回答第 (1) ~ (3) 題：

(1) 抽全部 9 小時內的污染源 feature 的一次項(加 bias)

(2) 抽全部 9 小時內 pm2.5 的一次項當作 feature(加 bias)

備註：

a. NR 請皆設為 0，其他的數值不要做任何更動

b. 所有 advanced 的 gradient descent 技術(如: adam, adagrad 等) 都是可以用的

1. (2%)記錄誤差值 (RMSE)(根據 kaggle public+private 分數)，討論兩種 feature 的影響

9hr	kaggle public	Kaggle private	sum	Rmse
All feature	7.32418	5.27853	12.60271	5.34311792
pm2.5	7.79759	7.95881	15.7564	6.56451238

我的參數如下: weight 和 bias 一開始的設計是使用 0~1e-13 之間的 random, learning rate 為 0.01, iteration 次數為 100000 次。可以發現在一次時 all feature 在 kaggle public & private 的分數都比較好，我認為原因有可能是 pm2.5 的預測模型的確會受到其他的 feature 的影響，因此在之後的優化我是使用了幾三組 feature 來做(pm2.5 pm10 amp temp)，雖然結果是 public 不錯有過 baseline 但是 private 卻蠻慘的，有一部分當然是因為調 240 比少少的資料很容易造成 over fit 為了追求 pubic 低分而造成 private 誤差增加。

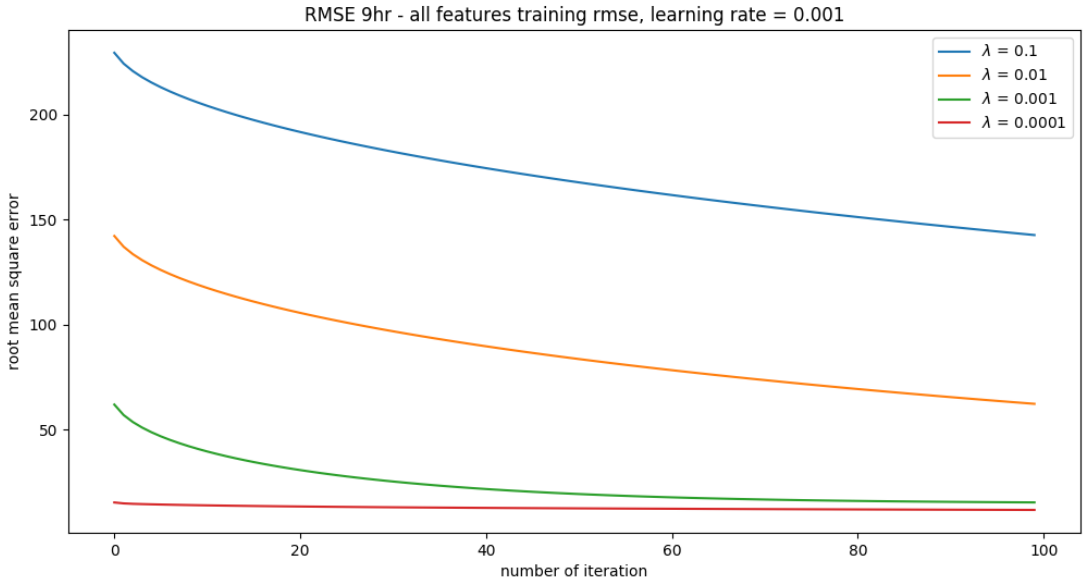
2. (1%)將 feature 從抽前 9 小時改成抽前 5 小時，討論其變化

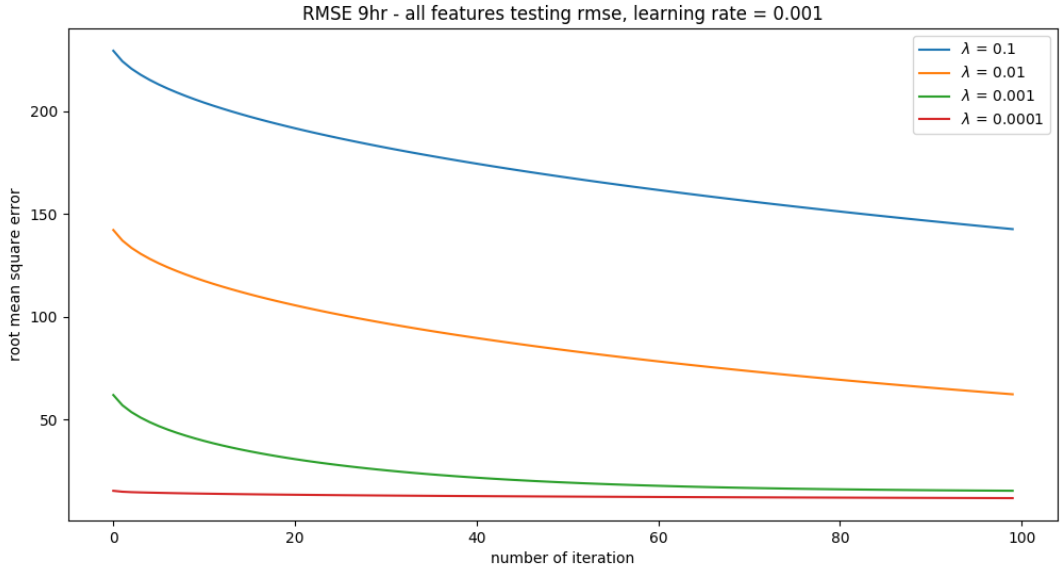
5hr	kaggle public	Kaggle private	sum	Rmse
All feature	7.57783	5.33990	12.91773	5.55835205
pm2.5	9.15773	7.27085	16.42858	6.87099375

5hr 我是由第十個小時往前回推 5 個小時去做，rmse 的結果是比 9 小時的略差，我認為可能的原因是 9 小時的包含了 5 小時的參數，而多出來的 4 小時可以對最後的預測結果多一些準確度貢獻。

3. (1%)Regularization on all the weight with λ=0.1、0.01、0.001、0.0001，並作圖

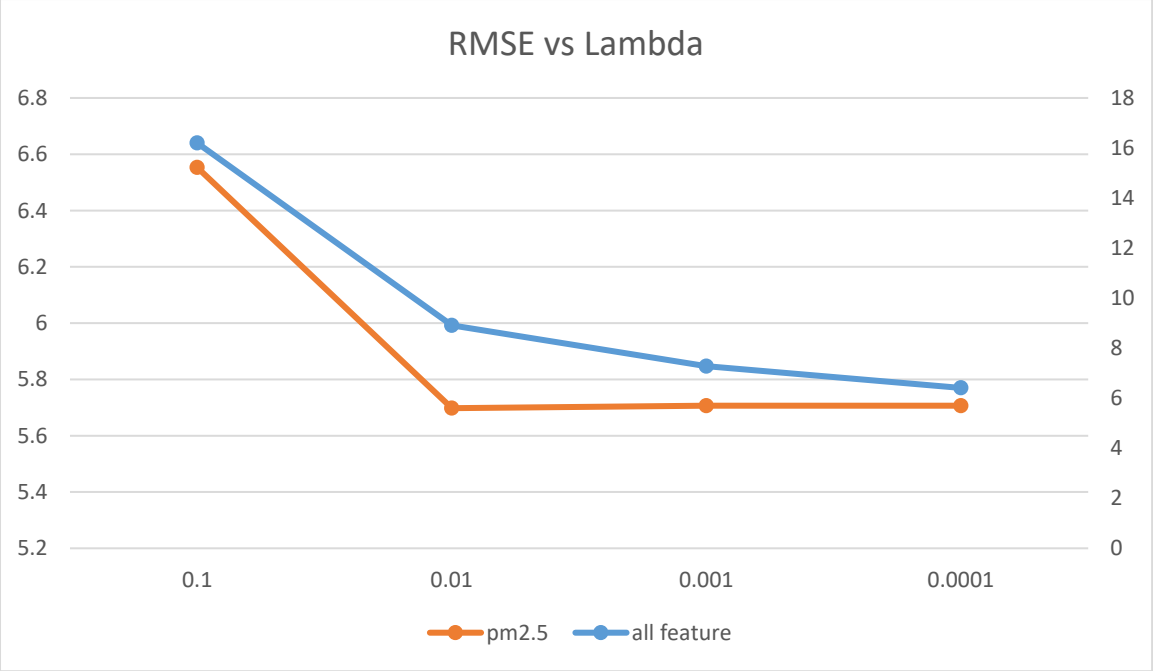
一開始的 weight 是用 0 到 0.1 之間的 random，然後考慮不同 regularization 的結果(相同的 initial weight)做圖如下:





可以發現 rmse 會隨著 lambda 的下降而減小，在最後的，上圖為 iteration 100 次的變化，在 iterate 100000 次之後，用 kaggle 的分數做比較結果如下:

RMSE\lambda	0.1	0.01	0.001	0.0001
All feature	16.199	8.90845	7.27858	6.41056
PM2.5	6.5528	5.6978	5.06633	5.70660



我是直接用在社團中助教所上傳的 test.csv 標準答案下去算 RMSE 結果顯示當 lambda 越小，RMSE 也越小，推論是 weight 的值可能比較大(smooth term 為 sum of [lambda 乘 weigh 的平方])，所以當我們把 lambda 條大的時候反而會造成 RMSE 的上升。

4. (1%)在線性回歸問題中，假設有 N 筆訓練資料，每筆訓練資料的特徵 (feature) 為一向量 x^n，其標註(label)為一存量 y^n，模型參數為一向量 w (此處忽略偏權值 b)，則線性回歸的損失函數(loss function)為$\sum_{n=1}^{N} (y^n - w \cdot x^n)^2$。若將所有訓練資料的特徵值以矩陣 $X = [x^1\ x^2 \dots x^N]^T$ 表示，所有訓練資料的標註以向量 $y = [y^1\ y^2 \dots y^N]^T$ 表示，請問如何以 X 和 y 表示可以最小化損失函數的向量 w ？請寫下算式並選出正確答案。(其中 X^TX 為 invertible)

(a) $(X^TX)X^Ty$
(b) $(X^TX)^{-0}X^Ty$
(c) $(X^TX)^{-1}X^Ty$
(d) $(X^TX)^{-2}X^Ty$

C

推導如下:

$(y-x^Tw)^2=(y-x^Tw)\ (y-x^Tw)^T$

$=(y-x^Tw)\ (y^T-xw^T)= yy^T- y\ xw^T- x^Tw\ y^T+ x^Tw\ xw^T$

對 w 微分-> $y\ x- x^Ty^T+ x^T\ xw^T+ x^T\ wx =0$

$2X^TX\ w= 2X^Ty$

$X^TX\ w= X^Ty$

$w=(X^TX)^{-1}X^Ty$

Q.E.D