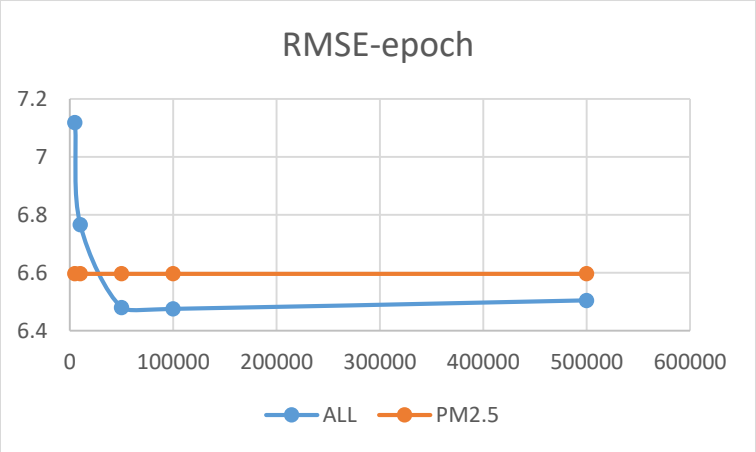
學號：B02123456 系級： 電機四 姓名：法拉利

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

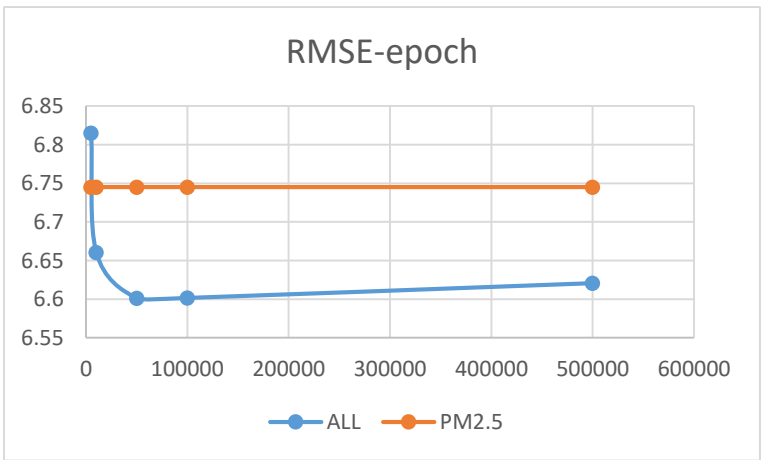
RMSE \ epoch	5000	10000	50000	100000	500000
feature (1)	7.118263438	6.7657952	6.4792726	6.47488070	6.50424812
feature (2)	6.596453407	6.59624142	6.59624142	6.59624142	6.59624142



由上述數據可知因為前 9 小時的 PM2.5 與第 10 小時的 PM2.5 有很大的關聯性，所以當訓練初期(epoch=5000、10000)，只採用 feature (1) 可以較準確的預測第 10 小時的 PM2.5。但同時因為參數量很少，所以 feature (1) 在 epoch 小於 10000 的時候就已收斂，RMSE 的值不再因 epoch 增加而改變。當 epoch 逐漸增加，feature (2) 的 RMSE 的值逐漸小於 feature (1)，可以得知 PM2.5 會受到其他 feature 影響，當訓練的誤差減小，也越來越能預測第 10 小時的 PM2.5。而根據設定的 learning rate，feature (2)在 epoch 等於 100000 附近可以得到最低的 RMSE。

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

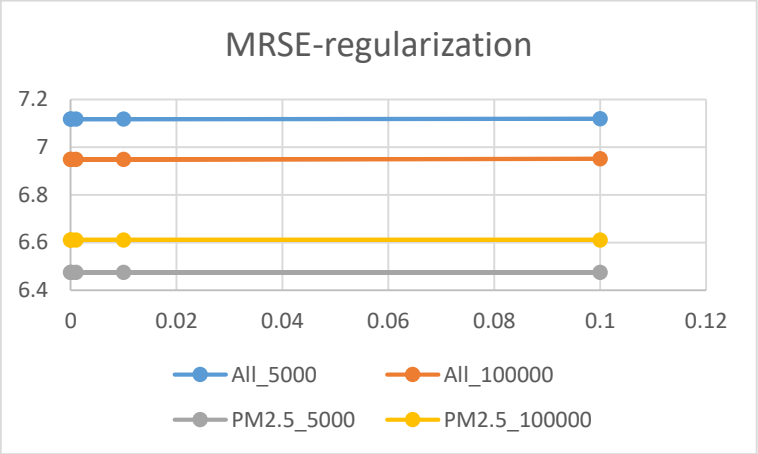
RMSE \ epoch	5000	10000	50000	100000	500000
feature (1)	6.814779223	6.66054603	6.60113733	6.6014402	6.62059975
feature (2)	6.744909392	6.74490939	6.74490939	6.74490939	6.74490939



由數據可知當 feature 變成抽取 5 個小時，參數量會變少但資料量會變多，所以 feature (1) 在 epoch 小於 5000 就收斂。但因為只能看到前 5 個小時的 feature，資訊量變少，所以無論是 feature (1) 或 feature (2) 的 RMSE 都比 問題 1. 大。

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

lambda	0	0.1	0.01	0.001
All_5000	7.118263	7.118263	7.118263	7.118263
All_100000	6.948976	6.948976	6.948976	6.948976
PM2.5_5000	6.474881	6.474881	6.474881	6.474881
PM2.5_100000	6.61103	6.611036	6.61103	6.61103



Regularization 這次的作業幾乎沒有影響，無論 model 是否收斂。另外印出 gradient 與 weight 出來發現，兩者約差了約三個數量級，所以 weight 的值乘上 learning rate 再除以每個 epoch 的 gradient 之和(實做 adagrad)後，得到的值約落在 10^{-4}~10^{-6} 次方，所以對參數的更新影響不大，上傳 kaggle 後再跟正確答案做 MRSE 影響就更微弱。故將取全部 feature 並跑 5000 epoch 的 model lambda 調成 100 後得到的 MRSE = 7.119134，即產生不同，但會造成 model 變差。

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 $\sum_{n=1}^{N}(y^n - x^n \cdot w)^2$ 。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)

$$\sum_{n=1}^{N}(y^n - x^n \cdot w)^2 = (y - Xw)^T(y - Xw)$$

$$\frac{\partial}{\partial w}(y - Xw)^T(y - Xw) = \frac{\partial}{\partial w}(y^Ty - y^TXw - w^TX^Ty + w^TX^TXw) = -2X^T(y - Xw) = 0$$

$$X^TXw = X^Ty, \quad \therefore w = (X^TX)^{-1}X^Ty \rightarrow \text{(C)}$$