

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
(public set err/private set err)

	(1)	(2)
error	7.49725/5.43869	7.47101/5.64620

Discussion: 在public set 中，只抽9小時內PM2.5的一次項feature效果比較好；不過在private set中卻是反過來。兩模型的誤差各自做完平均後，(1) 6.46797 / (2) 6.558605，以抽取全部feature的training 所做出的模型error比較低。不過，因為我是以固定跑100萬次iteration作為termination的依據，所以時間上來說只抽取PM2.5會快很多。而且另外在自己實做的模型中，減少部相關的feature確實會使誤差減少。我自己試出最好的模型其實只有用到7個feature(其中五個是有到二次)。另外，檢測這結果發生的原因，可能單純是因為參數變多而使誤差下降，抽取全feature使誤差減小。

2. (1%)將feature從抽前9小時改成抽前5小時，討論其變化

(public set err/private set err)

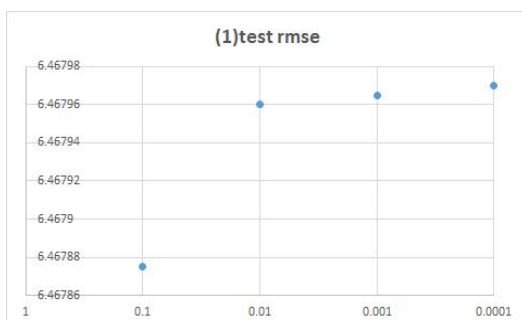
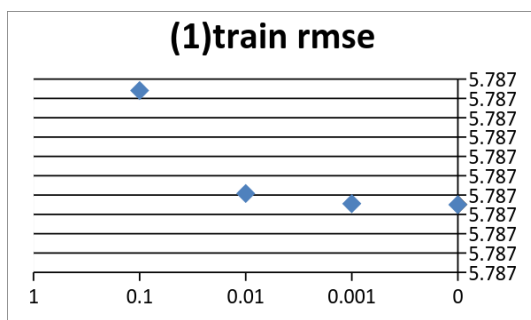
	(1)	(2)
9hr	7.49725/5.43869	7.47101/5.64620
5hr	7.67945/5.41822	7.59849/5.80119

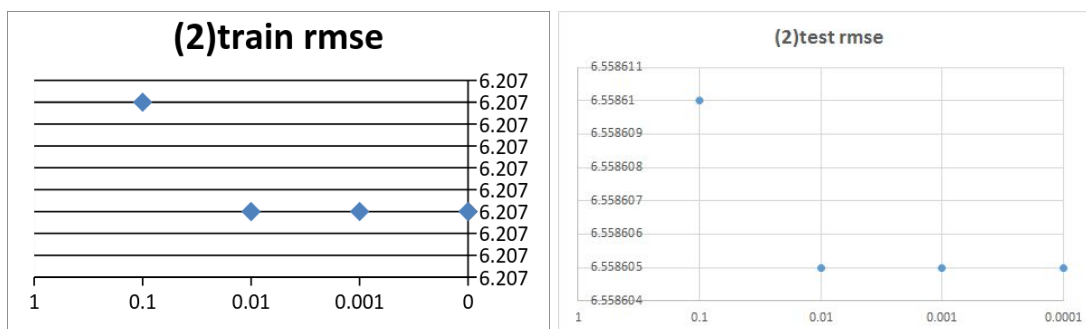
Discussion: 相比於9小時的兩個模型，5小時的兩個模型其實都相對誤差比較大，不論是在public 或是private set。推測其可能原因，有可能是因為參數變少使誤差上升，也有可能是因為6~9小時前的各項feature對於預測下一小時的PM2.5都有一定的相關性。整體而言，取較多小時的資料會使預測變精確，但同時也會使運算時間增長。此外，兩者的平均誤差: (1) 5hr: 6.54884 (2) 5hr: 6.69984，(1) 的誤差上升比率1.25%，(2) 的誤差上升比率2.15%，以(1)上升較少。

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

(training set err/ testing set err(avg))

	(1)	(2)
0.1	5.78719881699/6.467875	6.20711447202/6.55861
0.01	5.78718816526/6.46796	6.20711447201/6.558605
0.001	5.78718710541/6.467965	6.20711447201/6.558605
0.0001	5.78718699948/6.46797	6.20711447201/6.558605





Discussion: 隨著 λ 越來越高，代表輸出的回歸線要越平滑，因此在 training set 的 error 會越來越大；然而，在 (2) 中 test set 上竟然也越來越大，這個與理論所學不甚相符，究所學應該會有最佳的 λ test set error 最小，但是在這次作業中似乎無法觀察到此現象。而在 (1)，就比較符合預期的結果，有成功使 test set 的 error 隨著 λ 上升而減少。

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 $X^T X$ 為 invertible) **Ans: (c)**

- (a) $(X^T X) X^T y$
- (b) $(X^T X)^{-1} X^T y$
- (c) $(X^T X)^{-1} X^T y$
- (d) $(X^T X)^{-2} X^T y$

$$\begin{aligned}
 \text{let } \hat{y} &= X \cdot w \\
 \sum_j (\hat{y}(x_j) - y_j)^2 &= (\hat{y}(x_1) - y_1, \dots, \hat{y}(x_n) - y_n) \begin{pmatrix} \hat{y}(x_1) - y_1 \\ \vdots \\ \hat{y}(x_n) - y_n \end{pmatrix} = (\hat{y} - y)^T (\hat{y} - y) \\
 \text{又 } \hat{y} &= \begin{pmatrix} \hat{y}(x_1) \\ \vdots \\ \hat{y}(x_n) \end{pmatrix} = X \cdot w \quad \Rightarrow \quad \sum_j (\hat{y}(x_j) - y_j)^2 = (Xw - y)^T (Xw - y) \\
 \frac{\partial}{\partial w} (Xw - y)^T (Xw - y) &= \frac{\partial}{\partial w} (w^T X^T X w - w^T X^T y - y^T X w + y^T y) \\
 &= 2X^T X w - 2X^T y = 0 \quad \Rightarrow \quad X^T X w - X^T y = 0 \\
 \Rightarrow w &= (X^T X)^{-1} X^T y \quad (\text{if } X^T X \text{ is invertible})
 \end{aligned}$$