學號：B03502125 系級： 機械四 姓名：倪嘉宏

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

(1)9小時，使用18種feature : RMSE = 8.71906
(2)9小時，使用1種feature : RMSE = 7.04051

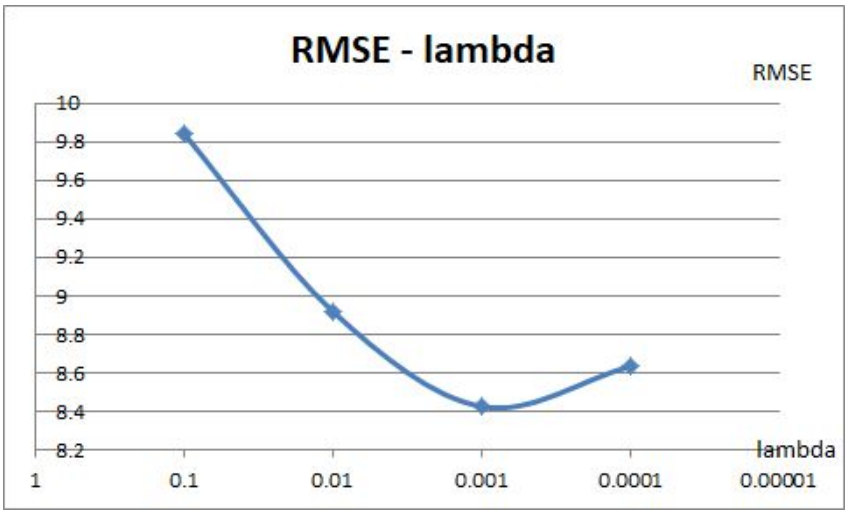
使用18種feature時所得到的結果與只使用PM2.5所得到的結果有非常顯著的差異。推測使用全部18種feature時的結果會遠遠遜色於只使用PM2.5的關係是因為18種Feature當中有太多與預測結果無關的資訊，使得linear regression的過程當中受到那些無關資訊的影響。而這些影響導致即使linear regression抵達一個穩定的狀態後，該組參數拿去預測其他筆資料時的結果不如只使用PM2.5所得出的參數。

2. (1%)將feature從抽前9小時改成抽前5小時，討論其變化

(1)5小時，使用18種feature : RMSE = 8.51013
(2)5小時，使用1種feature : RMSE = 7.24286
我們依然可以從這個結果看出採用18種feature時因為有太多的參數，而使得結果遠遜色於只使用1種feature時的結果。
不過只採用5小時的資料則分別對兩種情況產生剛好相反的結果，於採用18種feature時的結果當中是能幫助改善結果，而於只使用PM2.5的結果則是稍微降低了準確度，相當的耐人尋味。
為何會有這樣的結果?我推測是因為參數數量及品質的關係而造成的。有18種feature時又使用9小時的資料使得參數過多，尤其當中又有很多與預測似乎難以有關聯的參數，比如說現在的濕度對於9小時後的PM2.5濃度有多少的關聯?顯然相較於8小時後的PM2.5的濃度對於9九時後PM2.5的關聯性，是微乎其微的。降為只使用5小時的資料可以改善這個問題，使得表現更好。相對的，只使用PM2.5作為預測資訊時9個參數已經是很理想的參數數量，與要預測的數據也有很明顯的相關，再砍掉部分的參數則會使得預測品質下降，似乎也頗為合理。

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

lambda	RMSE
0.0001	8.63514
0.001	8.42617
0.01	8.92081
0.1	9.84461



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)

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

(c)可產出最小化損失函數

$$X = \begin{bmatrix} x_1^1 & \cdots & x_n^1 \\ \vdots & \ddots & \vdots \\ x_1^m & \cdots & x_n^m \end{bmatrix}, Y = \begin{bmatrix} y^1 \\ \vdots \\ y^m \end{bmatrix}, W = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$$

$$H(X) = XW = \begin{bmatrix} w_1x_1^1 & \cdots & w_nx_n^1 \\ \vdots & \ddots & \vdots \\ w_1x_1^m & \cdots & w_nx_n^m \end{bmatrix} = \begin{bmatrix} h_w(x^1) \\ \vdots \\ h_w(x^m) \end{bmatrix}$$

$$H_w(X) - Y = \begin{bmatrix} h_w(x^1) - y^1 \\ \vdots \\ h_w(x^m) - y^m \end{bmatrix}$$

$$\text{cost function} = J(W) = \sum_{j=1}^{m}\left[h_w(x^j) - y^j\right]^2 = (XW - Y)^T(XW - Y)$$

$$\frac{\partial}{\partial W}J(W) = \frac{\partial}{\partial W}(XW - Y)^T(XW - Y)$$

$$= \frac{\partial}{\partial W}(W^TX^TXW - W^TX^TY - Y^TXW + Y^TY)$$

$$= 2(X^TXW - X^TY)$$

$$\text{To Find minimal of } J(W), \text{Let} \frac{\partial}{\partial W}J(W) = 0$$

$$2(X^TXW - X^TY) = 0$$

$$W = (X^TX)^{-1}X^TY$$