學號：R08246009 系級：應數所碩二 姓名：許哲維

請實做以下兩種不同feature的模型，回答第 (1) ~ (2) 題：

(1) 抽全部9小時內的污染源feature當作一次項(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，其他的非數值(特殊字元)可以自己判斷

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

c. 第1-2題請都以題目給訂的兩種model來回答

d. 同學可以先把model訓練好，kaggle死線之後便可以無限上傳。

e. 根據助教時間的公式表示，(1) 代表 p = 9x18+1 而(2) 代表 p = 9*1+1

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

ANS：

model1 的 RMSE = (4.79129, 4.88643)；model2 的RMSE = (4.88643, 5.17965)，如果用所有的feature下去做迴歸，我們會有135個regressor加上一個bias，然而我們發現比model2只有用9個regressor加上一個bias的還要好，在直覺上是合理的，因為pm2.5應該會受其他環境變量影響，而非單只受到過去時間的pm2.5影響而已。

2. (1%)解釋什麼樣的data preprocessing 可以improve你的training/testing accuracy，ex. 你怎麼挑掉你覺得不適合的data points。請提供數據(RMSE)以佐證你的想法。

ANS：

針對數據我將”-”的值補上，此欄位的前一個值進行補值，而出現一些亂碼或是數字突然>700的資料筆數直接刪除。接著我在第一題的兩個model中，都刪除training data中pm2.5大於50的資料，主要原因是我發現testing set中的pm2.5大多都落在0~50的範圍內，只有少數幾筆大於50。若完全沒有做資料修剪得到的結果是model1的RMSE = (4.84498, 5.25181)； model2的RMSE = (4.91346, 5.06973)，可以發現都比有做修剪資料得來的差。

3.(3%) Refer to math problem

NO:

DATE: / /

1.(b).

We change the parameter of w.

Define $\begin{cases} W = \begin{pmatrix} b \\ w \end{pmatrix}_{k+1} \Rightarrow L_{ssq}(W) = \frac{1}{2N}\sum_{i=1}^{N}(y_i - X_i^T W)^2. \\ X_i = \begin{pmatrix} 1 \\ x_i \end{pmatrix}_{k+1}, \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \end{cases}$

$\mathbb{X} = (X_1, \cdots, X_N)_{(k+1)\times N}$ be a matrix. Then we can rewrite $L_{ssq}(W) = \frac{1}{2N}(Y - \mathbb{X}^T W)^T (Y - \mathbb{X}^T W)$.

$$\frac{\partial L_{ssq}}{\partial W} = \frac{1}{2N}\left[-2\mathbb{X}Y + 2(\mathbb{X}\mathbb{X}^T)W\right].$$

$$\left.\frac{\partial L_{ssq}}{\partial W}\right|_{\hat{W}} = 0 \Rightarrow 2(\mathbb{X}\mathbb{X}^T)\hat{W} = 2\mathbb{X}Y.$$

$$\Rightarrow \hat{W} = (\mathbb{X}\mathbb{X}^T)^{-1}\mathbb{X}Y = \begin{pmatrix} \hat{b} \\ \hat{w} \end{pmatrix} \#.$$

NO:

DATE: / /

1. (c).

$$\frac{\partial L_{reg}}{\partial W} = \frac{\partial}{\partial W}\frac{1}{2N}\sum_{i=1}^{N}(y_i - x_i^T W - b)^2 + \frac{\lambda}{2}\|W\|^2.$$

$$= \frac{-1}{2N} 2 \sum_{i=1}^{N} x_i(y_i - x_i^T W - b) + \lambda W = 0.$$

$$\Rightarrow \left[\sum_{i=1}^{N} x_i x_i^T + \lambda I_k\right] W + b\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i y_i \quad \text{——} \; ①$$

$$\frac{\partial L_{reg}}{\partial b} = \frac{-1}{2N} 2 \sum_{i=1}^{N} (y_i - x_i^T W - b) = 0.$$

$$\Rightarrow \sum_{i=1}^{N} x_i^T W + N b = \sum_{i=1}^{N} y_i \quad \text{——} \; ②$$

① + ②

$$\begin{pmatrix} \sum_{i=1}^{N} x_i x_i^T + \lambda I_k & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i^T & N \end{pmatrix} \begin{pmatrix} \hat{W} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} x_i y_i \\ \sum_{i=1}^{N} y_i \end{pmatrix}.$$

$$\Rightarrow \begin{pmatrix} \hat{W} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} \sum x_i x_i^T + \lambda I_k & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i^T & N \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{N} x_i y_i \\ \sum_{i=1}^{N} y_i \end{pmatrix}. \#$$

(a).

This is the simple linear regression.

So, $\hat{w} = \frac{Sxy}{Sxx}$ $\hat{b} = \bar{y} - \hat{w}\bar{x}$

$$\hat{w} = \frac{\sum_{i=1}^{5}(X_i - \bar{X})(y_i - \bar{y})}{\sum_{i=1}^{5}(X_i - \bar{X})^2} = \frac{10.5}{10} = 1.05 \#$$

$$\hat{b} = 3.36 - 1.05 \cdot 3 = 0.21 \#$$

No.
Date

2.

$$E\left(\frac{1}{2N}\sum_{i=1}^{N}\left(f_{w,b}(x_i+y_i)-y_i\right)^2\right)$$

$$=\frac{1}{2N}\sum_{i=1}^{N}E\left(w^T(x_i+y_i)+b-y_i\right)^2=\frac{1}{2N}\sum_{i=1}^{N}E\left[(w^Tx_i+b-y_i)+w^Ty_i\right]^2$$

$$=\frac{1}{2N}\sum_{i=1}^{N}\left[E\left(f_{w,b}(x_i)-y_i\right)^2+E\left(w^Ty_i\right)^2+2\left(f_{w,b}(x_i)-y_i\right)\underbrace{E(w^Ty_i)}_{0}\right]$$

$$=\frac{1}{2N}\sum_{i=1}^{N}\left(f_{w,b}(x_i)-y_i\right)^2+\frac{1}{2N}\sum_{i=1}^{N}\underbrace{Var(w^Ty_i)}_{w^T Var(y_i)w=\|w\|^2\sigma^2}$$

$$=\frac{1}{2N}\sum_{i=1}^{N}\left(f_{w,b}(x_i)-y_i\right)^2+\frac{\sigma}{2}\|w\|^2 \quad \#$$

3.

(a)

$$e_0=\frac{1}{N}\sum_{i=1}^{N}g_k(x_i)^2$$

$$e_k=\frac{1}{N}\sum_{i=1}^{N}g_k(x_i)^2+\frac{1}{N}\sum_{i=1}^{N}y_i^2-2\frac{1}{N}\sum_{i=1}^{N}g_k(x_i)y_i$$

$$=s_k+e_0-\frac{2}{N}\sum_{i=1}^{N}g_k(x_i)y_i$$

$$\Rightarrow \sum_{i=1}^{N}g_k(x_i)y_i=\frac{N}{2}\left(s_k+e_0-e_k\right) \quad \forall k=1,\dots,K+1 \quad \#$$

(b).

$$\frac{\partial}{\partial\alpha_k}\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(\alpha_k g_k(x_i)-y_i\right)^2\Big|_{\alpha_k=\hat{\alpha}_k}=0.$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\alpha}_k g_k(x_i)-y_i\right)g_k(x_i)=0.$$

$$\hat{\alpha}_k\sum_{i=1}^{N}g_k(x_i)=\sum_{i=1}^{N}y_i g_k(x_i) \Rightarrow \hat{\alpha}_k=\frac{\sum_{i=1}^{N}y_i g_k(x_i)}{\sum_{i=1}^{N}g_k(x_i)} \quad \forall k=1,\dots,K \quad \#$$