學號：R07946013 系級： 資料科學碩二 姓名：吳泓毅

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1-3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

因為沒做normalize，所以對每個Feature 加上normal(mean=0, var = self.var / 10)的noise；下面顯示的error是cross validation 10次取平均，每次切1/10當 cross validation的testing set再算 root mean square error得到的。

Case 1:

沒加noise:15.644075914132111

有加noise:14.927462885316675

Case 2:

沒加noise:54.28306031220069

有加noise:55.66705445214008

由(3)可以得知在資料裡面加上noise 會有做regularization 的效果，而Case 加了Noise 得到更好的表現，代表regularization 產生了效果，避免了些overfitting。而Case2，由於本身的錯誤率太高，所以做Regularization的效果不明顯，甚至導致錯誤率提高。

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。

在做完預處理後，資料的大小變成(7100,162)，對於簡單的模型(ex: 線性回歸)來說來說這個資料量算非常足夠了，所以決定做一些資料的 selection。

Row selection:

再拿到初始資料時，PM2.5 的variancezo 非常大(>7000)，想說有些資料可能可以當異常值直接放棄掉，在原本助教的手把手code上有把 PM2.5 < 5 或 PM2.5>100 的值去掉，得到大小(7100,162)的資料，所以我就把這個值在限縮，得到大小(5726,162)的資料，且在testing也取得更好的準確率(5.65126 --> 5.37674)。

Feature selection:

最初做看到資料以為應該只有PM2.5這個測項 是最重要的，但看了各個測項的相關係數發現，每一項都跟PM2.5有些線性關係，而確實如果單純把某個feature 直接看調不用的話，Training跟Testing的準確率都降低了。(包括降雨、風向、風速)

不過有可以選擇比較不激進的方式拿掉feature，就是每個測項都取前幾個小時，不用把全部9個小時來用，原本想用Cross validation來找最適合每個測項的小時數，不過不確定有沒有比暴力法(18*9 種可能)更快的方法，就先沒做了。

3.(3%) Refer to math problem

https://hackmd.io/RFiu1FsYR5uQTrrpdxUvlw?view

1-(a) $S = \{(1, 1.2), (2, 2.4), (3, 3.5), (4, 1), (5, 5.6)\}$. $\Rightarrow A = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \\ 5 & 1 \end{bmatrix}$, $y = \begin{bmatrix} 1.2 \\ 2.4 \\ 3.5 \\ 4.1 \\ 5.6 \end{bmatrix}$

By 1-(b), we have $\begin{bmatrix} \tilde{w} \\ \tilde{b} \end{bmatrix} = (A^T A)^{-1} A^T y = \begin{bmatrix} 1.05 \\ 0.21 \end{bmatrix}$.

1-(b) For each x_i, append 1 to the end of the vector, we have $\tilde{x}_i = \begin{bmatrix} x_i \\ 1 \end{bmatrix}$.

$\therefore L_{ssq}(w,b) = \frac{1}{2N} \sum_{i=1}^{N} [y_i - (w^T x_i + b)]^2 = \frac{1}{2N} \sum [y_i - [w^T \; b] \begin{bmatrix} x_i \\ 1 \end{bmatrix}]^2 = \frac{1}{2N} \| y - \begin{bmatrix} -x_1- & 1 \\ \vdots & \vdots \\ -x_N- & 1 \end{bmatrix} \begin{bmatrix} w \\ b \end{bmatrix} \|^2$.

By linear algbra, min occurs when $\begin{bmatrix} \hat{w} \\ \hat{b} \end{bmatrix} = \left(\begin{bmatrix} -x_1- & 1 \\ -x_N- & 1 \end{bmatrix}^T \begin{bmatrix} -x_1- & 1 \\ \vdots & \vdots \\ -x_N- & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} -x_1- & 1 \\ -x_N- & 1 \end{bmatrix}^T y$.

1-(c) Similar to above, denote $\tilde{w} = \begin{bmatrix} w \\ 1 \end{bmatrix}$, $\tilde{X} = \begin{bmatrix} -x_1- & 1 \\ \vdots & \vdots \\ -x_N- & 1 \end{bmatrix}$, $y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$

$$L_{reg}(w,b) = \frac{1}{2N} \sum_{i=1}^{N} [y_i - (w^T x_i + b)]^2 + \frac{\lambda}{2} \|w\|^2 = \frac{1}{2N} \| y - \tilde{X}\tilde{w} \|^2 + \frac{\lambda}{2} \|w\|^2$$

$$= \frac{1}{2N} \| y - \tilde{X} \begin{bmatrix} w \\ 1 \end{bmatrix} \|^2 + \frac{\lambda}{2} \| 0 - [I_k \; \vdots] \begin{bmatrix} w \\ 1 \end{bmatrix} \|^2$$

$$= \frac{1}{2N} \| \begin{bmatrix} y \\ 0 \\ \vdots \\ 0 \end{bmatrix} - \begin{bmatrix} -x_1- & 1 \\ -x_N- & 1 \\ 2N I_k & 0 \end{bmatrix} \tilde{w} \|^2$$

Denote $y' = \begin{bmatrix} y_1 \\ \vdots \\ y_N \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^{N+k}$, $X' = \begin{bmatrix} -x_1- & 1 \\ -x_N- & 1 \\ 2N I_k & 0 \end{bmatrix} \in \mathbb{R}^{(N+k) \times (k+1)}$

$\therefore$ similarly, $\hat{X'} = (X'^T X')^{-1} X'^T y'$. #

2.

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

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

$$= \frac{1}{2N}\sum\left\{Var[w^T(x_i+\eta_i)+b-y_i] + (E[w^T(x_i+\eta_i)+b-y_i])^2\right\}$$

$$= \frac{1}{2N}\sum\left\{Var(w^T\eta_i) + (w^Tx_i+b-y_i)^2\right\}$$

$$= \frac{1}{2N}\sum(w^Tx_i+b-y_i)^2 + \frac{1}{2N}\sum cov(w^T\eta_i,\ w^T\eta_i)$$

$$= \frac{1}{2N}\sum_{i=1}^{N}(f_{w,b}(x_i)-y_i)^2 + \frac{1}{2N}\sum w^T(Var\,\eta_i)\,w,\quad Var\,\eta_i = \sigma^2 I,\ \forall i.$$

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

-(a)

3-(a)

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

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

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

3-(b) $\min_{\alpha} L_{test}\left(\sum_{k=1}^{K} \alpha_k g_k\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K} \alpha_k g_k(x_i) - y_i\right)^2$

Let $\frac{\partial L}{\partial \alpha_t} = 0 \Rightarrow \frac{2}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} \alpha_k g_k(x_i) g_t(x_i) - \frac{2}{N}\sum_{i}\sum_{k} y_i g_t(x_i) = 0$

$\Rightarrow \frac{2}{N}\sum_{k} \alpha_k \sum_{i} g_k(x_i) g_t(x_i) - \frac{2k}{N}\sum_{i} y_i g_t(x_i) = 0$

$\Rightarrow \sum_{k} \alpha_k \sum_{i} g_k(x_i) g_t(x_i) = k\cdot(s_t - e_t + e_0)$, $t = 1, \cdots, K$

Denote $\sum_{i=1}^{N} g_k(x_i)\cdot g_t(x_i) = a_{kt}$, $t = 1, \cdots, K$.

$\therefore \begin{bmatrix} a_{11} & \cdots & a_{K1} \\ \vdots & \ddots & \vdots \\ a_{1K} & \cdots & a_{KK} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_K \end{bmatrix} = K \begin{bmatrix} s_1 - e_1 + e_0 \\ \vdots \\ s_K - e_K + e_0 \end{bmatrix}$.

Since a_{kt} is known, $A = [a_{kt}]_{k,t}$. is also known.

$\therefore \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_K \end{bmatrix} = A^{-1}\cdot K \begin{bmatrix} s_1 - e_1 + e_0 \\ \vdots \\ s_K - e_K + e_0 \end{bmatrix}$, $A = \begin{bmatrix} a_{11} & \cdots & a_{K1} \\ \vdots & \ddots & \vdots \\ a_{1K} & \cdots & a_{KK} \end{bmatrix}$ and $a_{kt} = \sum_{i=1}^{N} g_k(x_i)\cdot g_t(x_i)$ #.