學號：R10922111 系級：資工所碩一 姓名：黃冠瑋

請實作以下兩種不同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。

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

(1) 5.51225 + 4.90217 = 10.41442

(2) 69.81766 + 69.522191 = 139.33985

第一項會考慮所有因素，而第二個只會考慮前九天的pm2.5，就連當天的氣體濃度或風速都沒加入參考，有可能前幾天都是晴天，而今天下大雨或是發生其他強烈變化，就會導致結果與前幾天有非常大的不同。

2. (1%)解釋什麼樣的data preprocessing可以improve你的training/testing accuracy，e.g.，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。

上網查相關資料，發現pm2.5與雨量、風向、風速等有較大的相關，因此可以將這幾個feature調高權重。

3.(4%) Refer to math problem

https://hackmd.io/@GfOkB4kgS66YhhM7j6TJew/BykqpjhEK

1-(a)

$$P(c_1|x) = \sigma\left(\sum_i w_i x_i + b\right) = \sigma(-7-3+50+1) = \frac{1}{1+e^{-41}}$$

1-(b)

$$-\ln L(w,b) = -\left(\ln f_{w,b}(x^1) + \ln f_{w,b}(x^2) + \ln(1 - f_{w,b}(x^3)) + \cdots + \ln f_{w,b}(x^N)\right)$$

$$= -\Big(\hat{y}^1 \ln f_{w,b}(x^1) + (1-\hat{y}^1)(1-f_{w,b}(x^1)) +$$

$$\hat{y}^2 \ln f_{w,b}(x^2) + (1-\hat{y}^2)(1-f_{w,b}(x^2)) +$$

$$\vdots$$

$$\hat{y}^N \ln f_{w,b}(x^N) + (1-\hat{y}^N)(1-f_{w,b}(x^N))\Big)$$

$$= -\sum_i \left[\hat{y}^i \ln f_{w,b}(x^i) + (1-\hat{y}^i) \ln f_{w,b}(x^i)\right]$$

1-(c) $$\frac{-\ln L(w,b)}{\partial w_j} = -\sum_i \left[\hat{y}^i \frac{\ln f(x^i)}{\partial w_j} + (1-\hat{y}^i)\frac{\ln(1-f(x^i))}{\partial w_j}\right]$$

$$\because \frac{\partial \ln f(x^i)}{\partial w_j} = \frac{\partial \ln f(x)}{\partial z}\frac{\partial z}{\partial w_j} = \frac{1}{f(x)} \cdot \frac{\partial f(x)}{\partial z} \cdot \frac{\partial(\sum w_j x_j^i + b)}{\partial w_j} = \frac{1}{\sigma(z)} \cdot \sigma(z)(1-\sigma(z)) \cdot x_j^i$$

$$= (1-\sigma(z)) \cdot x_j^i$$

$$\frac{\partial \ln(1-f(x^i))}{\partial w_j} = \frac{\partial \ln(1-f(x^i))}{\partial z}\frac{\partial z}{\partial w_j} = \frac{1}{\sigma(z)-1}\,\frac{\partial \sigma(z)}{\partial z} \cdot \frac{\partial(\sum w_j x_j^i + b)}{\partial w_j} = \frac{1}{\sigma(z)-1}\,\sigma(z)(1-\sigma(z)) \cdot x_j^i$$

$$= -\sigma(z) \cdot x_j^i$$

$$\therefore \frac{-\ln L(w,b)}{\partial w_j} = -\sum_i \left[\hat{y}^i (1-f(x^i)) \cdot x_j^i - (1-\hat{y}^i) \cdot f(x^i) \cdot x_j^i\right]$$

$$= -\sum_i \left[\hat{y}^i - \hat{y}^i \cancel{f(x^i)} - f(x^i) + \hat{y}^i \cancel{f(x^i)}\right] x_j^i = -\sum_i (\hat{y}^i - f_{w,b}(x^i)) \cdot x_j^i$$

So, $w_i \leftarrow w_i - \eta \sum_i -(\hat{y}^i - f(x^i))\, x_j^i$

2-(a) $$\frac{\partial L(w,b)}{\partial b} = \frac{2}{10}\sum (y_i - wx_i - b)(-1) = 0 \Rightarrow 1.5 - w - b + 2.4 - 2w - b$$

$$+ 3.5 - 3w - b + 4.1 - 4w - b + 5.3 - 5w - b$$

$$= -15w - 5b + 16.8 = 0$$

$$\frac{\partial L(w,b)}{\partial w} = \frac{2}{10}\sum (y_i - wx_i - b)(-x_i) = 0 \Rightarrow (1.5 - w - b) + (2.4 - 2w - b) \cdot 2 + (3.5 - 3w - b) \cdot 3$$

$$+ (4.1 - 4w - b) \cdot 4 + (5.3 - 5w - b) \cdot 5 = 59.7 - 55w - 15b = 0$$

$$\begin{cases} 15w + 5b = 16.8 \\ 55w + 15b = 59.7 \end{cases} \Rightarrow w = 0.93,\ b = 0.57 \Rightarrow (w, b) = (0.93, 0.57)$$

2-1b)

$x_i \in R^k, \forall i \in [1, N]$

let $X = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{pmatrix}$, $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$, then we can rewrite $L(w, b)$ as $\qquad$ let $\begin{pmatrix} w \\ b \end{pmatrix} = \beta$

$$L(w, b) = \frac{1}{2N} \left\| \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} - \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} & 1 \\ x_{21} & x_{22} & \cdots & x_{2k} & 1 \\ \vdots & & & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nk} & 1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \\ b \end{pmatrix} \right\|^2 = \frac{1}{2N} \left\| Y - X \begin{pmatrix} w \\ b \end{pmatrix} \right\|^2 = \frac{1}{2N} (Y - X\beta)^T (Y - X\beta)$$

To minimize L, $\frac{\partial L}{\partial \beta} = 0 = \frac{\partial}{\partial \beta} (Y - X\beta)^T (Y - X\beta) = \frac{\partial}{\partial \beta} (Y^T - \beta^T X^T)(Y - X\beta)$

$$= \frac{\partial}{\partial \beta} (Y^T Y - Y^T X \beta - \beta^T X^T Y + \beta^T X^T X \beta)$$

$\because Y^T X \beta \in R^1$

$\therefore (Y^T X \beta)^T = \beta^T X^T Y = Y^T X \beta$ $\quad \overset{\sim}{=} \frac{\partial}{\partial \beta} (Y^T Y - 2 \beta^T X^T Y + \beta^T X^T X \beta)$

Now, we prove $\frac{\partial}{\partial \beta} (\beta^T X^T X \beta) = 2 X^T X \beta$

$\because (X^T X)^T = X^T X$ $\quad \therefore$ replace $X^T X$ as A, and its entry as a_{ij}

$$\beta^T A \beta = \sum_i^{k+1} \sum_j^{k+1} a_{ij} \beta_i \beta_j = \sum_{i=1}^{k+1} a_{i1} \beta_i \beta_1 + \sum_{i=1}^{k+1} a_{1j} \beta_1 \beta_j + \sum_{i=2}^{k+1} \sum_{j=2}^{k+1} a_{ij} \beta_i \beta_j$$

$$\frac{\partial \beta^T A \beta}{\partial \beta_1} = \sum_{i=1}^{k+1} a_{i1} \beta_i + \sum_{i=1}^{k+1} a_{1j} \beta_j \underset{\text{since } a_{ij} = a_{ji}}{=} \sum a_{i1} \beta_i + \sum a_{i1} \beta_i = 2 \sum a_{i1} \beta_i$$

$$\therefore \frac{\partial \beta^T A \beta}{\partial \beta} = \begin{pmatrix} 2 \sum a_{1i} \beta_i \\ 2 \sum a_{2i} \beta_i \\ \vdots \\ 2 \sum a_{k+1\, i} \beta_i \end{pmatrix} = 2 \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1(k+1)} \\ & & \vdots & \\ & & & a_{(k+1)(k+1)} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{k+1} \end{pmatrix} = 2 A \beta$$

So, $\frac{\partial L}{\partial \beta} = 0 = -2 X^T Y + 2 X^T X \beta \quad \Rightarrow \quad \beta = (X^T X)^{-1} (X^T Y) = \begin{pmatrix} w \\ b \end{pmatrix}$

2-1c)

$$L_{reg}(w, b) = \frac{1}{2N} \left\| Y - X \begin{pmatrix} w \\ b \end{pmatrix} \right\|^2 + \frac{\lambda}{2} \left\| \begin{pmatrix} w \\ b \end{pmatrix} \right\|^2 - \frac{\lambda b}{2}$$

$$= \frac{1}{2N} (Y - X\beta)^T (Y - X\beta) + \frac{\lambda}{2} \beta^T \beta - \frac{\lambda b}{2}$$

$\frac{\lambda}{2} \beta^T \beta = \frac{\lambda}{2} \beta^T I \beta$

$\therefore \frac{\partial}{\beta} \frac{\lambda}{2} \beta^T I \beta = 2 \cdot \frac{\lambda}{2} I \beta$

$$\frac{\partial L}{\partial \beta} = (-2 X^T Y + 2 X^T X \beta) \frac{1}{2N} + \lambda I \cdot \beta = 0 \quad \Rightarrow \quad \frac{-X^T Y + X^T X \beta}{N} + \lambda \beta = 0$$

$$\Rightarrow \left(\frac{X^TX}{N} + \lambda I\right)\beta = \frac{X^TY}{N} \Rightarrow \beta = \left(\frac{X^TX}{N} + \lambda I\right)^{-1}\frac{X^TY}{N} = \begin{pmatrix} w \\ b\end{pmatrix}$$

3.

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

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

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

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

$$E(w^T\eta_i) = E(w_1\eta_{i1}+w_2\eta_{i2}+\dots+w_k\eta_{ik})$$
$$= w_1E(\eta_{i1})+w_2E(\eta_{i2})+\dots+w_kE(\eta_{ik})$$
$$= 0 \quad (\because E(\eta_{ij})=0)$$

$$E\left[\sum_{i=1}^{N}(w^T\eta_i)^2\right]$$

$$= \sum_{i=1}^{N} E\left[(w^T\eta_i)^2\right] = \sum_{i=1}^{N} E\left[(w_1\eta_{i1}+w_2\eta_{i2}+\dots+w_k\eta_{ik})^2\right]$$

$$= \sum_{i=1}^{N} E\left[\sum_{j=1}^{k} w_j^2\eta_{ij}^2 + 2\sum_{j=1}^{k}\sum_{\substack{k=1\\ j\neq k}}^{k} w_jw_k\eta_{ij}\eta_{ik}\right]$$

since $E[\eta_{ij}\eta_{i'j'}] = \delta_{ii'}\delta_{jj'}\sigma^2$

$$= \sum_{i=1}^{N}\left[\sum_{j=1}^{k} E\left[w_j^2\eta_{ij}^2\right] + 2E\left[\sum_{j}\sum_{\substack{k\\ j\neq k}} w_jw_k\eta_{ij}\eta_{ik}\right]\right]$$

$$= \sum_{i=1}^{N}\left[\sum_{j=1}^{k} w_j^2\sigma^2\right] + 2\sum_{j}\sum_{\substack{k\\ j\neq k}} w_jw_k E\left[\eta_{ij}\eta_{ik}\right]\Big]$$

$$= \sigma^2 N\cdot\left[\sum_{j=1}^{k} w_j^2\right] + 0$$

So, $$\tilde{L}_{ssq}(w,b) = \frac{1}{2N}\sum\left(f_{w,b}(x_i)-y_i\right)^2 + \frac{\sigma^2 N}{2N}\sum_{j=1}^{k} w_j^2$$

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

$$\|w\|^2 = w^Tw = (w_1\, w_2 \cdots w_k)\begin{pmatrix}w_1\\ \vdots \\ w_k\end{pmatrix} = w_1^2+\dots+w_k^2 = \sum_{j=1}^{k} w_j^2$$

-(a)

$$e_k = \frac{1}{N}\left[g_k(x_1)^2 + y_1^2 - 2g_k(x_1)y_1 + g_k(x_2)^2 + y_2^2 - 2g_k(x_2)y_2 + \cdots \right]$$

$$= \frac{1}{N}\left[\sum_i g_k(x_i)^2 + \sum_i y_i^2 - 2\sum_i g_k(x_i)y_i \right]$$

since $\frac{1}{N}\sum_i g_k(x_i)^2 = s_k$, $\frac{1}{N}\sum_i y_i^2 = e_0$, $\therefore \sum_i g_k(x_i)y_i = \frac{N}{-2}(e_k - s_k - e_0)$

-(b)