學號：R06522709 系級： 機械碩三 姓名：鄭呈毅

請實做以下兩種不同 feature 的模型，回答第 (1) ~ (2) 題：
抽全部 9 小時內的污染源 feature 當作一次項(加 bias)
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
a.抽全部 9 小時內的污染源 feature 當作一次項(加 bias):
Private score=5.51 Public score = 5.65
b.抽全部 9 小時內 pm2.5 的一次項當作 feature(加 bias)
privatecore=6.67 public score =6.66

如果以 9 小時內的所有汙染源資訊來 train model 的話 RMSE 會比只取 pm2.5 作為 feature 的結果還要小上很多，我想這是因為 pm2.5 本來就是我們想要預測的目標，而雖然前段時間的 pm2.5 必定與下一個時刻的 pm2.5 有一定的關係，但是顯然 pm2.5 也會受到其他 feature 的影響，因此在只考慮 pm2.5 本身的情況下，預測結果並非最好。此外我在 train 只取 pm2.5 作為 feature 的 model 時，有出現 model 發散的情形，iteration 為 1000 時則 kaggle 上顯示的 RMSE 為 55，甚至超過全部預測 0 所得到的 RMSE，但是如果 iteration 為 50 的話 RMSE 只有 6，因為我是採用 TA 手把手所提供的 model，理論上來說 gradient descent 不應該出現這種情況才對，後來發現似乎是對這個 feature 來說 learningg rate 0.001 太大，導致 model 發散，後來把 learning rate 調成 0.00001 之後就沒有這個問題了。

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。
a.不挑掉任何 data point:
private score=5.94 public score = 6.54
b.同 TA 手把手 code 挑掉 pm2.5<=2 or >=100:
private score=5.64 public score = 5.81

如同 TA 在手把手時段所傳授的，如果資料點本身就不合理，就不應該餵進 model 裡面，最明顯的就是 pm2.5 當中會出現負值，這是一定不合理的；pm2.5>65 就會對所有人的身體健康造成危害，而在台灣 pm2.5>80 已是非常嚴重的情況，所以若 pm2.5>100 已是 outlier，應該要予以剔除。另外在所有 feature 當中，rainfall 高頻率出現 NAN 的情形，一開始我認為這個 feature 紀錄狀況並不完整，應該不具參考價值，所以就將它剔除，但訓練結果不佳，我想這可能是因為降雨本身就會大幅影響 PM2.5 產生，像是雨季的 pm2.5 比起旱季的 pm2.5 少上很多，因此雖然該 feature 紀錄狀況不完整，但是一但有紀錄，則該資訊會大幅影響預測的正確性，因此仍然需要將此 feeature 餵進 model 中。

3.(3%) Refer to math problem

1-(a) $S=\{(1,1.2),(2,2.4),(3,3.5),(4,4.1)(5,5.6)\}$

$$L_{ssq}(w,b)=\frac{1}{2*5}\sum_{i=1}^{5}\left(y_i-(w^Tx_i+b)\right)^2$$

by normal equation $\theta=(X^TX)^{-1}X^Ty$

$$\begin{array}{c} y \\ \begin{bmatrix}1.2\\2.4\\3.5\\4.1\\5.6\end{bmatrix} \end{array} = \begin{array}{c} x_0\ x_1 \\ \begin{bmatrix}1&1\\1&2\\1&3\\1&4\\1&5\end{bmatrix} \end{array} \begin{array}{c} \theta \\ \begin{bmatrix}b\\w\end{bmatrix} \end{array}$$

$y=X\theta$

$$=\left(\begin{bmatrix}1&1&1&1&1\\1&2&3&4&5\end{bmatrix}\begin{bmatrix}1&1\\1&2\\1&3\\1&4\\1&5\end{bmatrix}\right)^{-1}\begin{bmatrix}1&1&1&1&1\\1&2&3&4&5\end{bmatrix}\begin{bmatrix}1.2\\2.4\\3.5\\4.1\\5.6\end{bmatrix}$$

$$=\frac{1}{50}\begin{bmatrix}55&-15\\-15&5\end{bmatrix}\begin{bmatrix}1&1&1&1&1\\1&2&3&4&5\end{bmatrix}\begin{bmatrix}1.2\\2.4\\3.5\\4.1\\5.6\end{bmatrix}$$

$$=\frac{1}{50}\begin{bmatrix}40&25&10&-5&-20\\-10&-5&0&5&10\end{bmatrix}\begin{bmatrix}1.2\\2.4\\3.5\\4.1\\5.6\end{bmatrix}=\frac{1}{50}\begin{bmatrix}10.5\\52.5\end{bmatrix}$$

$$=\begin{bmatrix}0.21\\1.05\end{bmatrix}=\begin{bmatrix}b\\w\end{bmatrix}$$

1-(b)

$$L_{ssq}(w,b) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - (w^T x_i + b)\right)^2$$

$$= \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \theta \cdot x_i\right)^2$$

$$= J(\theta)$$

$$\begin{array}{ccc} X & \theta & = y \\ \begin{bmatrix} - x(1) - \\ - x(2) - \\ \vdots \\ - x(n) - \end{bmatrix} & \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_d \end{bmatrix} & \approx \begin{bmatrix} y(1) \\ \vdots \\ y(n) \end{bmatrix} \end{array}$$

We want to minimize $J(\theta)$ $\Rightarrow$ let $\nabla_\theta(J(\theta)) = 0$.

$$\begin{bmatrix} \frac{\partial}{\partial\theta_1} J(\theta) \\ \frac{\partial}{\partial\theta_2} J(\theta) \\ \vdots \\ \frac{\partial}{\partial\theta_d} J(\theta) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$J(\theta) = \frac{1}{2n}\sum_{t=1}^{n}\left[y^{(t)} - (x_1^{(t)}\theta_1 + x_2^{(t)}\theta_2 + \dots x_d^{(t)}\theta_d)\right]^2$$

$\Downarrow$ by chain rule.

$$\frac{\partial}{\partial\theta_1}(J(\theta)) = \frac{1}{n}\sum_{i=1}^{n}\left[y_1^{(i)} - (x_1^{(i)}\theta_1 + x_2^{(i)}\theta_2 + \dots x_d^{(i)}\theta_d)\right](-x_1^{(t)})$$

$$= \frac{1}{n}\sum_{t=1}^{n}\left[y^{(t)} - (x^{(t)})^T\theta\right](-x_1^{(t)}) = 0$$

$$\Rightarrow \cancel{\frac{1}{n}}\sum_{t=1}^{n} x_1^{(t)}(x^{(t)})^T\theta = \cancel{\frac{1}{n}}\sum_{t=1}^{n} x_1^{(t)} y^{(t)}$$

$$\Rightarrow \overset{X^T}{\begin{bmatrix} | & | & & | \\ x^{(1)} & x^{(2)} & \dots & x^{(n)} \\ | & | & & | \end{bmatrix}} \overset{X}{\begin{bmatrix} - x^{(1)} - \\ - x^{(2)} - \\ \vdots \\ - x^{(n)} - \end{bmatrix}} \theta = \overset{X^T}{\begin{bmatrix} | & | & & | \\ x^{(1)} & x^{(2)} & \dots & x^{(n)} \\ | & | & & | \end{bmatrix}} y$$

$$\Rightarrow \theta = (X^T X)^{-1} X^T y.$$

$\theta = w \rightarrow w = (X^T X)^{-1} X^T y$

1-(c) $J(\theta) = \frac{1}{2n} \sum_{t=1}^{n} [y - w x^{(t)}]^2 + \frac{\lambda}{2} \|w\|^2$

Since $\frac{\lambda}{2}\|\theta\|^2 = \frac{\lambda}{2}(w_1^2 + w_2^2 + w_3^2 + w_4^2 \dots w_d^2)$

$\Rightarrow \frac{\partial}{\partial \theta_i}(J(\theta)) = \frac{1}{n} \sum_{t=1}^{n} [y^{(t)} - (x^{(t)})^T w](-x_i^{(t)}) + \lambda w_i$

setting partial derivative to 0.

$\Rightarrow \frac{1}{n} \sum_{t=1}^{n} x^{(t)} (x^{(t)})^T W + \lambda W = \frac{1}{n} \sum_{t=1}^{n} x^{(t)} y^{(t)}$.

$\Rightarrow \frac{1}{n} X^T X W + \lambda W = \frac{1}{n} X^T y$.

$\Rightarrow [\frac{1}{n} X^T X + \lambda I] w = \frac{1}{n} X^T y$

$\Rightarrow w = [\frac{1}{n} X^T X + \lambda I]^{-1} \frac{1}{n} X^T y$.

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

Show that

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

$f_{w,b}(x) = w^T x + b$.

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

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

$E[\eta_{ij}] = 0$ $\quad E[w^2 \eta_i^2]$

$= \frac{1}{2N} \sum_{i=1}^{N} (w^T(x_i) + b - y_i) + \frac{1}{2N} \cdot N \cdot \sigma^2 \cdot w^2$

only N term with value

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

3-(a) $e_k = \frac{1}{N}\sum_{i=1}^{N}(g_k(x_i) - y_i)^2$, $k = 0, 1, \dots, K$.

$$s_k = \frac{1}{N}\sum_{i=1}^{N}(g_k(x_i))^2, \qquad e_0 = \frac{1}{N}\sum_{i=1}^{N} y_i^2 \quad (g_0 \text{ involved})$$

express $\sum_{i=1}^{N} g_k(x_i) y_i$ in terms N, e_k, s_k.

$$e_k = \frac{1}{N}\sum_{i=1}^{N}\left[g_k(x_i)^2 - 2g_k(x_i)y_i + y_i^2\right] \qquad k = 0, 1, 2, \dots, K.$$

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

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

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

3-(b)

$$\min_{\alpha_1 \dots \alpha_K} L_{test}\left(\sum_{k=1}^{K} \alpha_k g_k\right) = \min\left[\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K}\alpha_k g_k(x_i) - y_i\right)^2\right].$$

$\Rightarrow$ find optimal $\alpha_1 \dots \alpha_K$.

$$\frac{\partial}{\partial \alpha} L_{test}\left(\sum_{k=1}^{K}\alpha_k g_k\right) = \frac{2}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(\alpha_k g_k(x_i) - y_i)\, g_k(x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{N}\sum_{k=1}^{K}\left\{\alpha_k [g_k(x_i)]^2 - y_i g_k(x_i)\right\} = 0$$

$$\Rightarrow \sum_{k=1}^{K}\sum_{i=1}^{N} \alpha_k [g_k(x_i)]^2 = \sum_{k=1}^{K}\sum_{i=1}^{N} g_k(x_i) y_i$$

$$\Rightarrow \sum_{k=1}^{K} \alpha_k s_k = \sum_{k=1}^{K}(s_k + e_0 - e_k) \times \frac{N}{2}$$

$$\Rightarrow \alpha_k = \frac{1}{s_k}(s_k + e_0 - e_k) \times \frac{N}{2}$$