學號：T08902019　系級：　電機三　　姓名：賈成銪

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

所有污染源 feature 當作一次項(加 bias)：kaggle 分數：5.67943
pm2.5 的一次項當作 feature(加 bias): kaggle 分數：6.58617
可以看出，當選取所有污染源 feature 當做一次項時，預測結果會更好
主要原因可能是當相關變量變多時，預測的穩定性會提高，受個別特殊數據的影響會減小。

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。

我做的 data preprocessing 主要包括：
（1）去除無記錄數據，如：NR，nan，空
（2）去除不合理數據，如：PM2.5 為負值或大於一百
當完全未處理數據時：RMSE：12.52191
將無記錄或不合理數據設為零時：RMSE：6.58617
將不合理數據剔除時：RMSE：5.71415
可見，將不合理數據剔除會有比較好的結果。

3.(3%) Refer to math problem
https://hackmd.io/RFiu1FsYR5uQTrrpdxUvlw?view

1. 1-(a)

$$L(w,b) = \frac{1}{2\times 5}\sum_{i=1}^{5}(y_i - (w^T x_i + b))^2$$

$$\frac{d(L(w,b))}{dw} = \frac{1}{2\times 5}\sum_{i=1}^{5} -2(y_i - wx_i - b)x_i = 0 \qquad ①$$

$$\frac{d(L(w,b))}{db} = \frac{1}{2\times 5}\sum_{i=1}^{5} -2(y_i - wx_i - b) = 0 \qquad ②$$

combinate ①, ②:

$w = 1.05$ $\qquad\qquad b = 0.21$

1-(b) 令 $w^* = (w, b)$ $\qquad X^* = (X, 1)$

$$L(w^*) = \frac{1}{2N}\sum_{i=1}^{n}(y_i - x_i^{*T}w^*)^2 = \frac{1}{2N}(y - X^*w)^T(y - X^*w)$$

$$\frac{d(L(w^*))}{dw^*} = \frac{1}{2N}\frac{d(y^Ty - y^TX^*w^* - w^{*T}X^{*T}y + w^{*T}X^{*T}X^*w^*)}{dw^*}$$

$$= \frac{1}{2N}(-X^{*T}y - X^{*T}y + 2X^{*T}X^*w^*)$$

$$= 0$$

$$\longrightarrow \quad X^{*T}X^*w^* = X^{*T}y$$

$$w^* = (X^{*T}X)^{-1}X^{*T}y$$

$$\underline{(w, b) = ((X,1)^T(X,1))^{-1}(X,1)^Ty}$$

1-(c).

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

$$\theta = (w, b)$$

$$L_{reg}(\theta) = \frac{1}{2N}\left[\sum_{i=1}^{N} (\theta^T x_{(i)}) - y^{(i)})^2 + N\lambda \|\theta\|^2\right.$$

gradient descent:

$$\theta_0 := \theta_0 - \alpha \frac{1}{N} \sum_{i=1}^{N} (\theta^T x^{(i)} - y^{(i)}) x_0^{(i)}$$

$$\theta_j := \theta_j - \alpha \left[\frac{1}{N} \sum_{i=1}^{N} (\theta^T x^{(i)} - y^{(i)}) x_j^{(i)} + \frac{\lambda N}{m}\theta_j\right] \quad j = 1, 2, 3, \cdots$$

merge:

$$\theta_j := \theta_j \underbrace{(1 - \alpha \frac{N\lambda}{m})}_{<1} - \alpha \frac{1}{m} \sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)}) x_j^{(i)}$$

also can be solved by normal equation:

$$X = \begin{bmatrix} (x^{(1)})^T \\ (x^{(2)})^T \\ \cdots \\ (x^{(m)})^T \end{bmatrix} \qquad Y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \cdots \\ y^{(m)} \end{bmatrix}$$

☆ $\theta = (w, b) = \left[X_*^T X_* + N\lambda \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & 1 & \\ & & & \ddots, 1 \end{bmatrix} \right]^{-1} X_*^T Y$

$X_* = (w, 1)$

2. sorry, no idea!

3. (a)

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

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

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

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

(b)

$$d(L_{test}(\alpha_k))$$

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

$$= 0$$

$$\therefore \sum_{i=1}^{N} \sum_{k=1}^{K} \alpha_k g_k(x_i) \sum_{k=1}^{K} g_k(x_i)$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} g_k(x_i)\, y_i$$

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

so:

solve:

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

the result are the optimal weights $\alpha_1 \cdots \alpha_k$.