學號:R07922163 系級: 資工所二 姓名:徐浩翔

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

	Private Score	Public Score
只用 PM2.5	5.98013	6.06713
用所有 feature	5.91050	5.76324

單單 PM2.5 這個 feature RMSE 就可以壓在 6 左右表示 PM2.5 這個 feature 是 很重要的。

但保留所有 feature 的模型整體而言 RMSE 會比只考慮 PM2.5 的模型還要小,表示除了 PM2.5 這個 feature 還有其他因素也需要考慮。

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

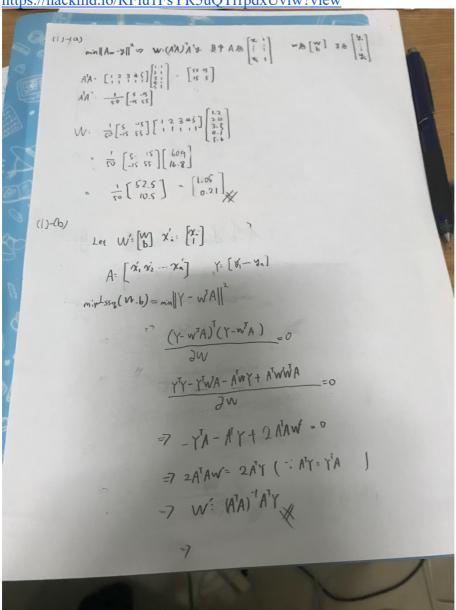
我去查了和 pm2.5 相關的空污指標,得 CO、NO、NO2、O3、PM10、PM2.5、SO2 與其正相關,我使用七項 feature 做 Gradient descent,效果比放所有的 18 項 feature 來得好

選這七種 feature 都可以過 public 和 private 的 strong base line, 另外有一些極端值像是超過 100 的我會把他們設成 80 小於零的把他設為 10 因為我算出了全年 pm2.5 平均大概是 16 所以就設了一個比較有可能的數字,這樣可以讓 traing data 變得比較多如果直接把他去掉會少好幾千筆 data

項目	Private Score	Public Score
取7個feature	5.56758	5.72966
只用 PM2.5	5.98013	6.06713
用所有 feature	5.91050	5.76324

## 3.(3%) Refer to math problem

https://hackmd.io/RFiu1FsYR5uQTrrpdxUvlw?view



(1)-(c) Let W= [ ] A= [ x, x, - x, ] Y=[y, - y,]  $\min L_{ag}(w,b) = \prod_{min} || Y - w'^T A ||^2 + \frac{2}{2} || w ||^2 - \frac{b^2}{2} ( : w' \cdot ["])$  $= 7 \qquad (Y - W'A)^{T} (Y - W'A) + \frac{\alpha}{2} W^{T} W - b^{2}$ => YTY-YWA-ATMY + ATWWA + ZWW = -2 YA +2AAW + 2W = 0 ( )[ ]] (ATA+IA) NF YA TIN (INSTAN) -W F (2) fwb(x) = Wixtb L = E [ ] 2 (5mb(x+n) - yi)]  $= \frac{1}{2N} \tilde{E} \left[ \sum_{i=1}^{N} \left( 5(x_i + \eta_i) - f(x_i) + 5(x_i) - y_i \right)^2 \right] = \frac{1}{2N} \tilde{E} \left[ \sum_{i=1}^{N} \left( f(x_i) - y_i \right)^2 + 2 \left( f(x_i + \eta_i) - f(x) \right) \left( f(x_i) - y_i \right) + f(x_i) + f(x_i) \right]$ = 1 2 f(wb(M2) - 82) + 2N = [ 2 (wing)] J(E[WW]=2, 1:4,2:1,)  $= \frac{1}{2N} \sum_{i=1}^{N} (5nb(\pi_i) - y_i)^2 + \frac{1}{26} \sum_{i=1}^{N} |w|^2$ 

retter in in in in in the con-

