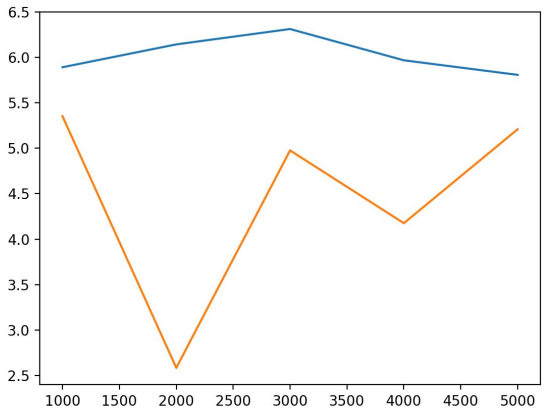
學號：B03502040 系級： 資工三 姓名：劉君猷

1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)
答：使用部分測項(CO, O3, PM10, PM2.5, RAINFALL, WD_HR, WIND_DIREC, WIND_SPEED, WS_HR)的連續9個小時的數值為一次方，加上PM10和PM2.5的兩次方與三次方作為features。
train_x = [x2, x7, x8, x9, x10, x14, x15, x16, x17, x8^2, x9^2, x8^3, x9^3]

2.請作圖比較不同訓練資料量對於PM2.5預測準確率的影響
答：藍色線是training loss rate (in RMS)；橘色線是valid loss rate (in RMS)。因為切valid的時候都是切固定的，所以可能會有bias。



3. 請比較不同複雜度的模型對於PM2.5預測準確率的影響
答：只使用一維參數當features時，loss rate大約在5.7~5.9之間；加上features的平方項後，loss rate降到5.6~5.7之間；調成三次方項後，loss rate大概在5.6左右。

一次項	二次項	三次項	Training loss	Public score
2, 7, 8, 9, 10, 14, 17			5.85	5.76
2, 7, 8, 9, 10, 14, 15, 16, 17			5.84	5.75
0 ~ 17			5.68	5.94
2, 7, 8, 9, 10, 14, 15, 16, 17	8, 9		5.81	5.65
2, 7, 8, 9, 10, 14, 15, 16, 17	8, 9	8, 9	5.79	5.61

(0 ~ 17 對應到 18 個測項)

4. 請討論正規化(regularization)對於PM2.5預測準確率的影響
答：在同樣的模型與參數輸入下，加上regularization效果沒有進步多少。我覺得是因為我太早做regularization的測試，以至於在當時那個model與參數下還underfitting，所以應該要再去調整model與參數，而不是就先加上regularization測試。

Lambda value	Training loss rate	Valid loss rate
0	5.791388	5.891997
1e-7	5.791392	5.892001
1e-5	5.791765	5.892368
1e-3	5.829029	5.929006

5. 在線性回歸問題中，假設有 N 筆訓練ji資料，每筆訓練資料的特徵 (feature) 為一向量 x^n，其標註(label)為一存量 y^n，模型參數為一向量w (此處忽略偏權值 b)，則線性回歸的損失函數(loss function)為 $\sum_{n=1}^{N} (y^n - w \cdot x^n)^2$ 。若將所有訓練資料的特徵值以矩陣 $X = [x^1\ x^2 \dots x^N]$ 表示，所有訓練資料的標註以向量 $y = [y^1\ y^2 \dots y^N]^T$表示，請以 X 和 y 表示可以最小化損失函數的向量 w 。
答： $w = (X^TX)^{-1}X^Ty$