學號：b04902053　系級：資工二　姓名：鄭淵仁

1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

答：

我總共寫了兩個版本：

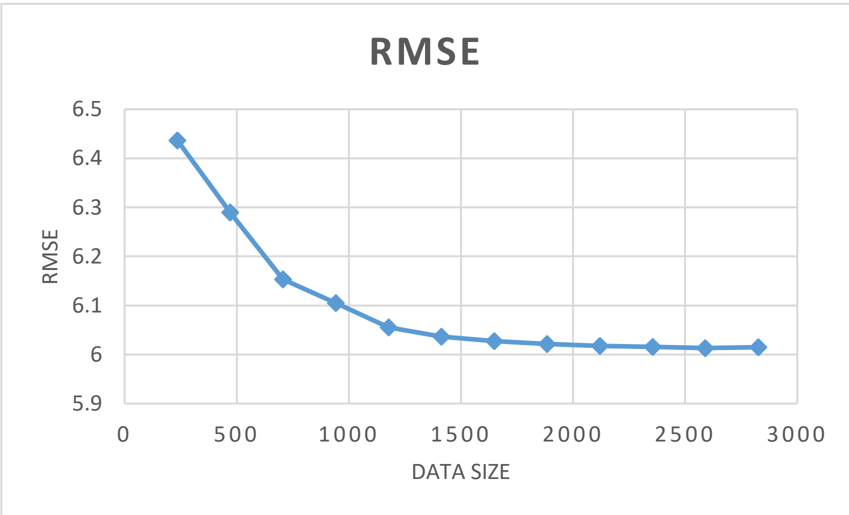
hw1_best.py 取前 9 個小時的 pm2.5 指標做一維和二維的 feature。

hw1.py 則是取前 9 個小時的 pm10、pm2.5、RAINFALL 指標做一維和二維的 feature。

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：

我把整個 data set 分一半，一半用來計算 RMSE，另一半則是用來 train。而 train 的時候，我只拿這一半資料中的$\frac{1}{12}\sim\frac{12}{12}$共 12 種不同的資料量去 train 出結果。除此之外，為了怕結果因為資料量太小導致誤差太大，我又多取了不同的資料 train 出不同結果，再算出 RMSE 並平均起來。得到的結果如下圖：



從圖中可以發現：在資料量等距增加的時候，RMSE 會下降得越來越慢。看起來很像是和資料量成反比的圖形。

而我在網路上查到以數學推導 RMSE 的教學，發現：數學推導出來的公式確實是和資料量成反比。（公式：$\mathrm{E_{out}} = noise\_level \cdot \left(1-\frac{d+1}{N}\right)$，其中$N$是資料量、$d$是 feature 的數量）

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：

我先把資料 shuffle 一遍，再把資料切成兩塊，一塊拿來 train，另一塊則是當作 validation set。接著針對同一筆變數同時用 1 次、2 次、3 次的複雜度來 train，算出 RMSE，再分別對維度取平均。結果如下面兩個表格：

表一 RMSE 先標準化再平均

dim	RMSE 先標準化再平均
1	-0.815379555
2	-0.07284098
3	0.888220534

表二 RMSE 直接平均

dim	RMSE 直接平均
1	11.17740882
2	11.2917164
3	16.28339887

在**表一**裡面，可以看出來 1 次的預測效果最好，2 次的效果其次，3 次的效果最差。

在**表二**裡面，可以看出來 1 次和 2 次的預測效果很接近，但是 3 次的效果明顯比較差，很像是 overfitting 的現象。

4. 請討論正規化(regularization)對於 PM2.5 預測準確率的影響

答：

我先把資料 shuffle 一遍，再把資料切成兩塊，一塊拿來 train，另一塊當作 validation set。而下表就是對於不同的維度和 regular 數值，算出的 RMSE 的結果。

表三 對於不同的 dim 和 regular 數值，RMSE 的結果

dim \ regular	0	10	100	1000	10000	100000	1000000
1	5.904	5.902	5.896	5.908	6.531	48.957	227.289
2	5.916	5.915	5.910	5.937	6.512	10.175	180.401
3	5.927	5.925	5.941	5.965	7.087	12.865	64.950

從表格中可以發現：在 regular 很小的時候，對資料不會有影響，但是當 regular 比 1000 大之後，不管維度是多少，RMSE 都會變大。

5. 在線性回歸問題中，假設有 N 筆訓練資料，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 $X = [x^1\ x^2\ \ldots\ x^N]$ 表示，所有訓練資料的標註以向量 $y = [y^1\ y^2\ \ldots\ y^N]^T$ 表示，請以 X 和 y 表示可以最小化損失函數的向量 w。

答：

設 loss function：$E_{in} = \sum_{n=1}^{N}(y^n - w \cdot x^n)^2$

則 $E_{in} = \|X \cdot w^T - y\|^2 = wX^TXw^T - 2wX^Ty + y^Ty$

故 $\nabla E_{in} = 2(X^TXw^T - X^Ty)$

設 $0 = \nabla E_{in} = 2(X^TXw^T - X^Ty)$ ，則

$X^TXw^T = X^Ty$

$\Rightarrow w^T = (X^TX)^{-1}X^Ty$

$\Rightarrow w = [(X^TX)^{-1}X^Ty]^T$