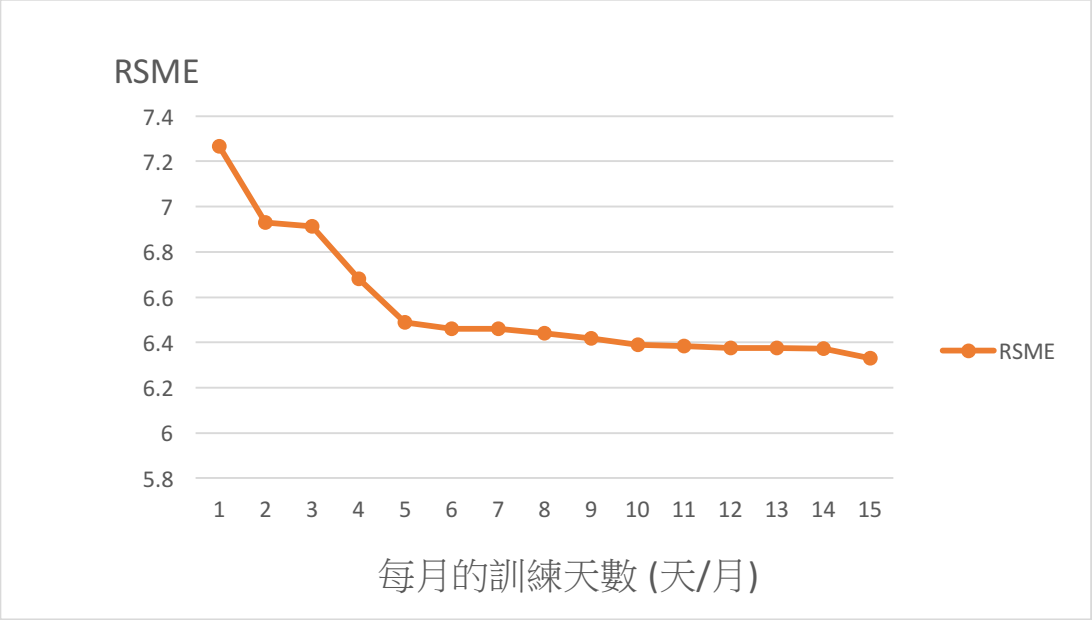
學號：R05942101 系級： 電信一 姓名：陳泓弦

1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)
答：
取前 9 個小時的 pm2.5 指標，第 9 個小時的 pm2.5 做一維和二維的 feature：
train_x = [x1,x2,x3,x4,x5,x6,x7,x8,x9,x9^2]

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：
取每一個月的前 1 天、前 2 天…到前 15 天的資料做為不同訓練資料量。
取每個月的第 16 天到第 20 天的資料做為 validation set 計算準確率。

下圖是訓練結果，橫軸為不同訓練資料量，縱軸為準確率，可以發現訓練資料量愈大，模型訓練的愈好，有較低的 RSME 值。




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：
用兩個模型，一個模型只取前九小時的 pm2.5 當作 feature(9 個)，另一個模型取前九小時全部 18 種空氣污染指標當作 feature(162 個)，結果發現取全部 18 種當 feature 的模型，其 Loss function 值一直在振盪無法收斂，iteration 一萬次在 kaggle 拿到了 37.92359，表現非常差。而只取 pm2.5 當 feature 的模型則很快就收斂，iteration 一千次在 kaggle 就拿到 7.30894，差 0.03 就能過 private 的 strong baseline。

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

| λ | Training | Validation |
|---|---|---|
| 0 | 6.585 | 6.754 |
| 1 | 6.395 | 6.566 |
| 10 | 6.378 | 6.621 |
| 100 | 7.863 | 8.406 |

regularization 能避免 overfitting，我做了 lamda=1,10,100 的模型，從上表可以看出在 trainging 和 validation 間的 RSME 沒有什麼影響，原因是我的模型並沒有 overfitting 的問題，而且當 lamda=100 時，表現會變差很多，故當模型沒有 overfitting 的問題時，是可以不用做 regularization。

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$\sum_{n=1}^{N} \quad (y^n - w \cdot x^n)^2$ 。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：

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