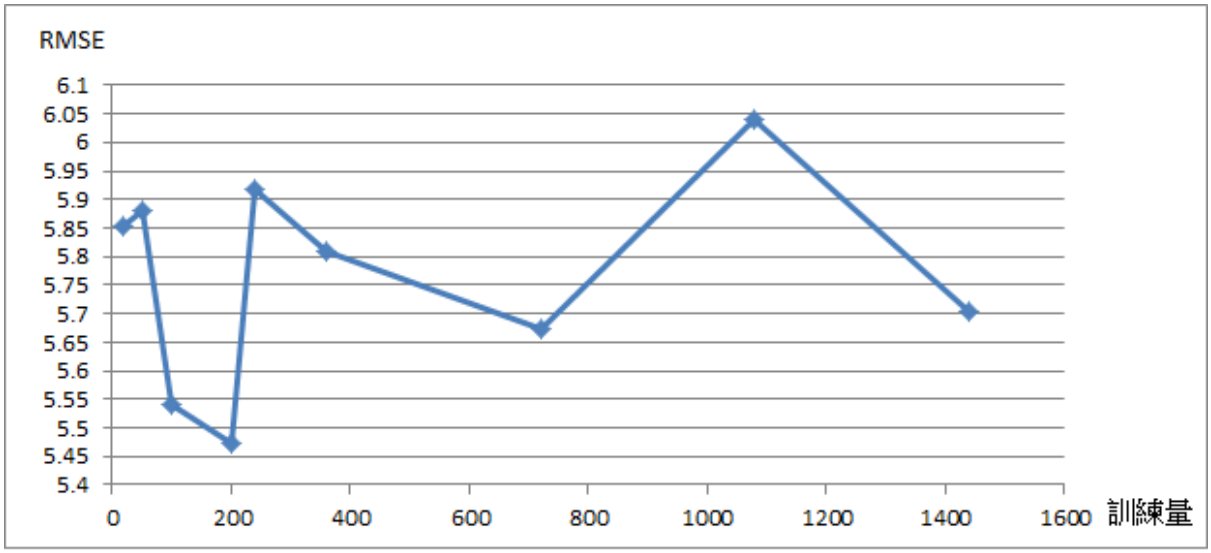
學號：B03901076 系級： 電機三 姓名：林尚謙

1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)
答：一開始先只考慮 PM2.5 這項輸入特徵，找出應該取幾小時內的資料做預測較準確（前 9 個小時～前 1 個小時）。找到一個最佳的時間範圍後，再陸續加上相同時間內的其他特徵（温度、PM10 等）做比較，一次加上一個特徵，去看有沒有提升預測準確度，若有則將此項特徵留下，若無則刪去，最後即為我模型的輸入特徵。
P.S.因為在這次實作中我沒有另外從 training set 切出 validation set，只能由 Kaggle 上的分數判斷模型的好壞，能做的嘗試有限（Kaggle 每日僅能上傳 5 筆資料），所以這部份忽略了一些情況，例如對 PM10 這項特徵的最佳時間範圍可能與 PM2.5 不同、温度與濕度須一起加入考慮等等。

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：在不同訓練資料量對於預測準確率影響的比較中，我模型使用前 7 個小時的 PM2.5 作為特徵，固定初始參數值、初始 learning rate、iteration 次數等，僅改變用來訓練的（train_x,train_y）數量，總共做了 9 筆資料，實驗結果如下圖：




可以發現在訓練量達一定數量後（約 100～200 筆），訓練出來的模型已經可以有良好的表現，然而當訓練量繼續往上增加，結果並沒有繼續變好，且彼此之間看不出明顯的關係。在做一些其他的實驗後，我發現即使在相同模型與相同訓練量之下，當選取的資料不同（例如平移後再選取），所產生的結果也會有很大的不同，所以我覺得在這部份決定結果好壞的是選了哪些資料來做訓練，和訓練量沒有直接的關連。

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：如第 1 小題中所述，我在不同複雜度的實作上可分成兩部份，分別是找出最佳的時間範圍與最佳的輸入特徵。在時間複雜度的實作上，我以 PM2.5 指標做實驗，把不同小時的 PM2.5 視為不同的複雜度，最後找出拿前 7 小時的數據做訓練後，能得到預測最準確的模型。而在不同輸入特徵的考量，我以前 7 小時的 PM2.5 為基準，一次加上另一項特徵，結果發現預測準確率皆呈現下降，也就是說只考慮 PM2.5 反而會有最佳的結果。有可能是在加上其他特徵後，會使模型 overfitting，也可能是沒考量到特徵之間的交互影響，還有各項特徵的最佳時間範圍不同所致。

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

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

$$\mathrm{X} = \begin{pmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nd} \end{pmatrix},\ \mathrm{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix},\ \mathrm{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix},\ \mathrm{L} = \sum_{n=1}^{N}(y_n - w\cdot x_n)^2$$

$$\text{假設 } \mathrm{y} + e = Xw\text{，則 } \mathrm{L} = |\mathrm{e}|^2 = e^Te = (Xw-y)^T(Xw-y)$$

$$\text{要使 L 最小，將 L 對 w 做偏微分：}$$

$$\begin{aligned}\nabla \mathrm{L} &= \nabla[(Xw-y)^T(Xw-y)] \\ &= \nabla[(w^TX^T - y^T)(Xw-y)] \\ &= \nabla(w^TX^TXw - w^TX^Ty - y^TXw + y^Ty) \\ &= 2X^TXw - X^Ty - (y^TX)^T + 0 \\ &= 2X^TXw - 2X^Ty\end{aligned}$$

$$\text{令}\nabla \mathrm{L} = 0\text{，可得使 L 最小的 w 為：}$$

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