

1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

答：

我實作兩類模型：線性函數 ($y = b + w \cdot x$) 及二次函數 ($y = b + w_1 \cdot x + w_2 \cdot x^2$)

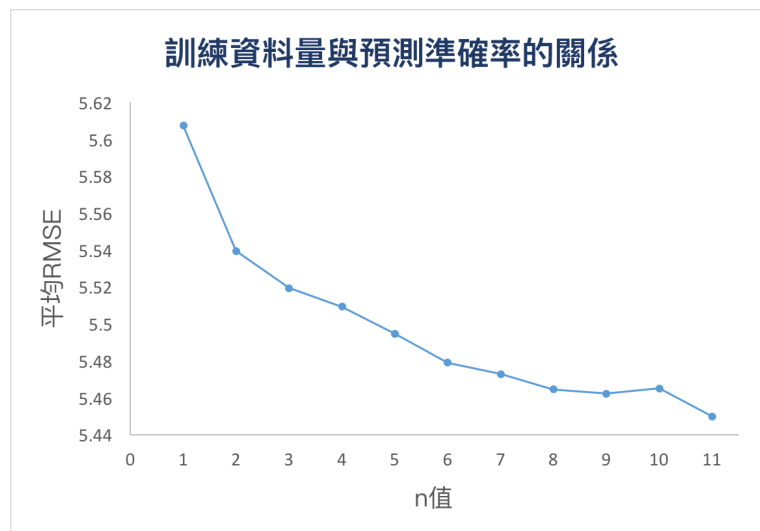
- 線性函數只取前九小時的 pm2.5 指標做一維 feature
 $\text{train_x} = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9]$
- 二次函數使用前九小時的 pm2.5 指標做一維及二維 feature
 $\text{train_x} = [x_1, x_2, \dots, x_8, x_9, x_1^2, x_2^2, \dots, x_8^2, x_9^2]$

另外，在抽取 feature 之前，我先將 train.csv 裡的 -1 以插值處理。

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：

以 12 月的 pm2.5 指標作為 validation set，前 11 個月的資料則以「連續 n 個月之 pm2.5 指標」來訓練一維模型，n 為 1 至 11，若 n = 3，則取 1~3、2~4...8~10、9~11 月各自訓練一維模型並預測。不同 n 值的平均 RMSE 如下圖：



由上圖可知，訓練資料量越大，準確率越高，而 n = 1 及 n = 2 的差距最明顯，可見訓練資料量過低時對於模型有非常負面的影響。

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：

承第 1 題，我實作了兩種模型，分別為線性函數 ($y = b + w \cdot x$) 及二次函數 ($y = b + w_1 \cdot x + w_2 \cdot x^2$)

- 以 12 個月的所有 pm2.5 指標訓練一維及二維模型，且皆不做正規化，上傳至 Kaggle 後一維模型的 RMSE 為 5.70451，二維則為 5.82360。
→ 使用 Gradient Descent 時，一維模型在 testing set 裡的資料表現較好。

- 以 12 月作為 validation set、前 11 月的 pm2.5 指標訓練一維及二維模型，皆不做正規化，一維模型的 RMSE 為 5.44989047462，二維則為 5.82333093214。

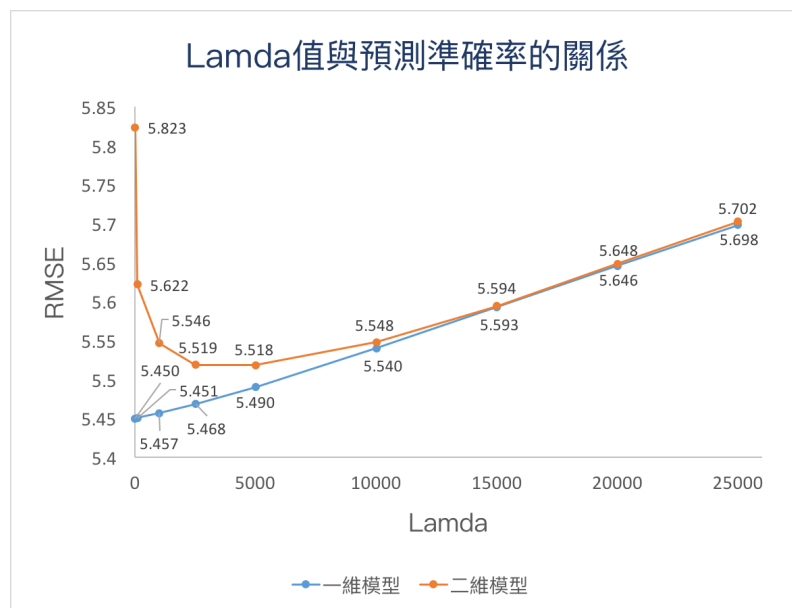
→ 一維模型在 validation set 的表現也較二維模型好。

由以上兩點，我推論：使用 Gradient Descent 預測 2014 年的 pm2.5 指標時，一維模型的表現較二維模型好。

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

答：

我以 12 月作為 validation set、前 11 月的 pm2.5 指標訓練一維及二維模型，比較 Lamda = 0, 100, 1000, 2500, 5000, 10000, 15000, 20000, 25000 時於一維模型及二維模型上的表現，結果如下圖：



→ Lamda 越大，一維模型在 validation set 上的表現越差，二維模型則在 Lamda = 5000 時表現最好。另外，Lamda ≥ 10000 時，兩種模型之準確率與 Lamda 值的關係很相似，此時 Lamda 值變大對準確率有負面影響。

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 $X = [x^1 x^2 \dots x^N]$ 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： $X^{-1}y$