學號：B04902004 系級： 資工二 姓名：王佑安

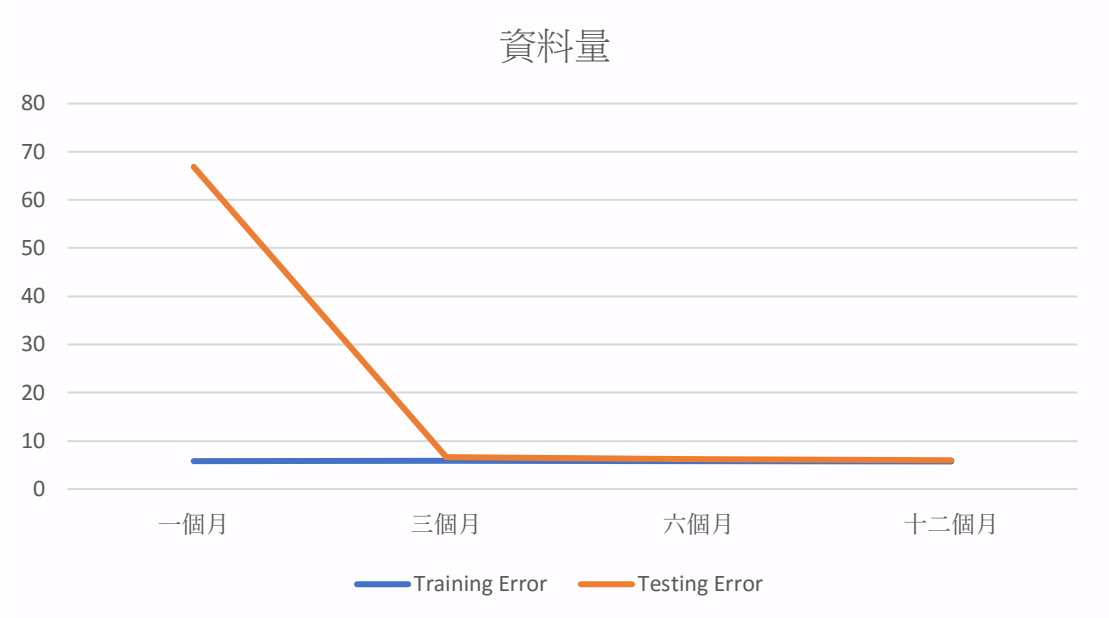
1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

答：

計算每個 feature 與 PM2.5 的相關係數，選擇相關係數較高的

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：



資料量	前一個月	前三個月	前六個月	全部十二個月
Training error	5.839	5.862	5.737	5.719
Testing error	66.857	6.679	6.136	5.963
* 以上數據為每次隨機取全部資料中 20%作 validation set 後 test 五次下的平均				

由表格可以看出，越多資料量，預測準確度越高。

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：

n	1	2	3	4
Training error	5.746	5.719	5.695	5.684
Testing error	5.857	5.963	6.895	9.657
* 以上數據為每次隨機取全部資料中 20%作 validation set 後 test 五次下的平均 * n 表示將選取的 feature 取 n 次方後做 normalization 成為新的 feature				

可以看出，複雜度越高的 model 在 training error 上越低，但在 testing data 上會出現 overfitting 的情形。雖然在 validation test 上 n=1 的表現比 n=2 好，但在 kaggle 上的 public data n=1 的 model 出現較高的 error，因此最後選擇 n=2 的 model。

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

答：

λ	0	1e-3	1	1000
Training error	5.719	5.719	5.725	7.473
Testing error	5.963	5.963	5.984	9.701
* 以上數據為每次隨機取全部資料中 20%作 validation set 後 test 五次下的平均				

在這個 Linear regression 的例子中，正規化似乎對準確率沒有明顯的幫助，只在 λ 過大時讓準確率下降。

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 … x^N] 表示，所有訓練資料的標註以向量 y = [y^1 y^2 … 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：

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