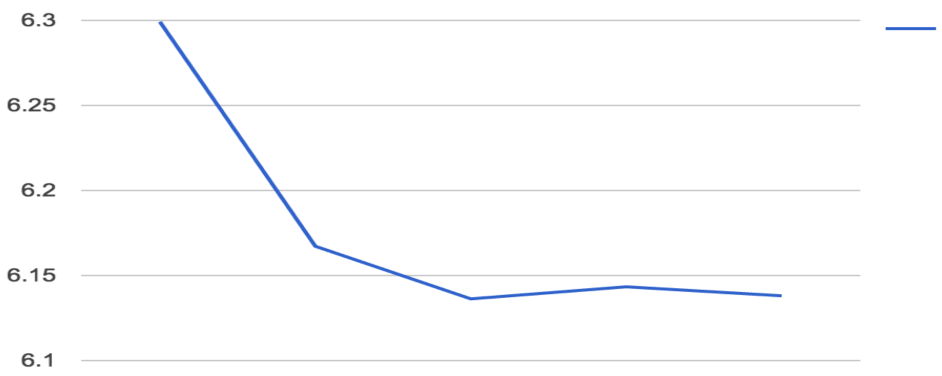
學號：B03902059 系級： 資工三 姓名：紀典佑

1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)
答：

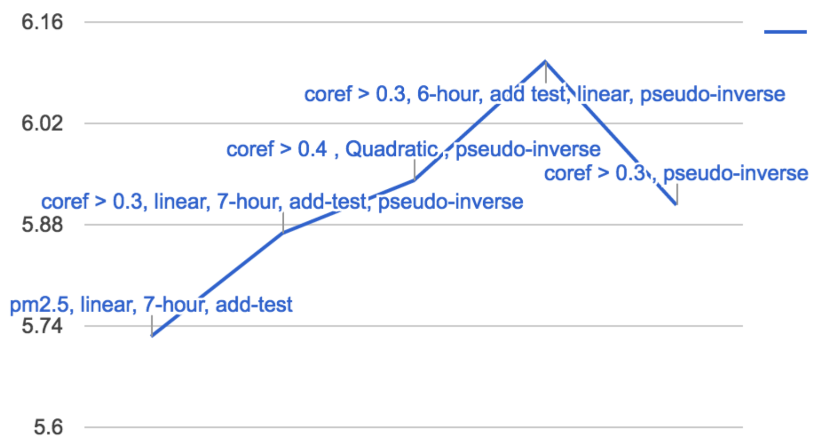
一開始先抽取 9*18 個完整的 feature 來 train，之後用相關係數取出所有 feature 中大於 0.5 的幾個來 train，也做過各個指標與 PM2.5 的相關係數取出大於 0.4 的來 train,而最好的結果(public)是拿七小時的 PM2.5 train 出來的。

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：



各個折點分別為 1000,2000,3000,4000 及全部資料所測出來的 validation error,使用的 model 為與 PM2.5 的相關係數大於 0.4 的幾個指標當做 feature 的結果，正常的情況下，資料量愈大，train 出來的結果 error 會比較小，但是若是在多加入的 data 有 noise 的狀況時，可能會使得 error 上升，圖表中 1000 與 2000 比之間的差距可能為前者，而後折線的浮動原因可能為後者。

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：



資料複雜度由左而右分別為：

a. 只取 PM2.5 七個小時的 feature，並且將 test data 切割七小時取出併入 training data 裡頭，做一次方的 linear regression。
b. 取指標與 PM2.5 的相關係數大於 0.3 的七個小時 feature，一樣加入七小時的 test data，做一次方的 linear regression。
c. 取指標與 PM2.5 的相關係數大於 0.4 的九個小時 feature，做二次方的 linear regression。
d. 取指標與 PM2.5 的相關係數大於 0.3 的六個小時 feature，加入六小時的 test data，做一次方的 linear regression。
e. 取指標與 PM2.5 的相關係數大於 0.3 的九個小時 feature，直接做一次方的 linear regression。

以上除了第一個沒用最佳解外其他皆有使用，可以直接判斷該模型能到達 Gradient descent 的最低點，取相關係數高的原因是為了讓一些與結果較無關的資料不要加入 training 過程，但取相關係數高的 feature 做出來的結果並未比只取 PM2.5 的結果好，另外取相近的幾個小時的結果也有些浮動，我推測的原因是因為 PM2.5 在空氣中存在的時間有關，另外使用二次的結果比一次的差一點，可能發生了 overfitting，因為在整體的 training data error 的結果，二次是比一次好的。

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

答：

正規化主要是為了避免 training 的結果 overfitting，而發生 overfitting 的主要原因是選取 model 的 degree 過大，使得結果接近於 training data，若 training data 的 noise 過多，就會讓結果跟 test 相去甚遠。

這次的作業中，我實作了二次的 linear regression，加入 regularization 並且調整 lambda，lambda 變大，在整體 training data 的 error 會稍微變大，但是將得到的結果放上 Kaggle 做測試，並未得到比一次 linear regression 小的 error，我推測可能的原因是這次資料的 feature 與結果的關係，比較適合使用一次 model 來 training。

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 $\mathbf{x}^n$，其標註(label)為一存量 $\mathbf{y}^n$，模型參數為一向量 $\mathbf{w}$ (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 $\mathbf{X} = [\mathbf{x}^1\ x^2 \ldots x^N]$ 表示，所有訓練資料的標註以向量 $\mathbf{y} = [\mathbf{y}^1\ y^2 \ldots y^N]^T$ 表示，請以 $\mathbf{X}$ 和 $\mathbf{y}$ 表示可以最小化損失函數的向量 $\mathbf{w}$ 。

答：

$$\mathrm{L} = (\mathrm{y} - \mathrm{w} \cdot \mathrm{X})^2 = (\mathrm{y} - \mathrm{w} \cdot \mathrm{X})(\mathrm{y} - \mathrm{w} \cdot \mathrm{X})^T \Rightarrow \frac{\partial \mathrm{L}}{\partial \mathrm{w}} = -2\mathrm{X}^T \cdot (\mathrm{y} - \mathrm{w} \cdot \mathrm{X})$$

$$\min \rightarrow \frac{\partial \mathrm{L}}{\partial \mathrm{w}} = 0 \quad \Rightarrow -2\mathrm{X}^T \cdot (\mathrm{y} - \mathrm{w} \cdot \mathrm{X}) = 0$$

$$\mathrm{X}^T y = w \cdot X\mathrm{X}^T$$

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