學號：B03705006 系級： 資管三 姓名：侯舜元

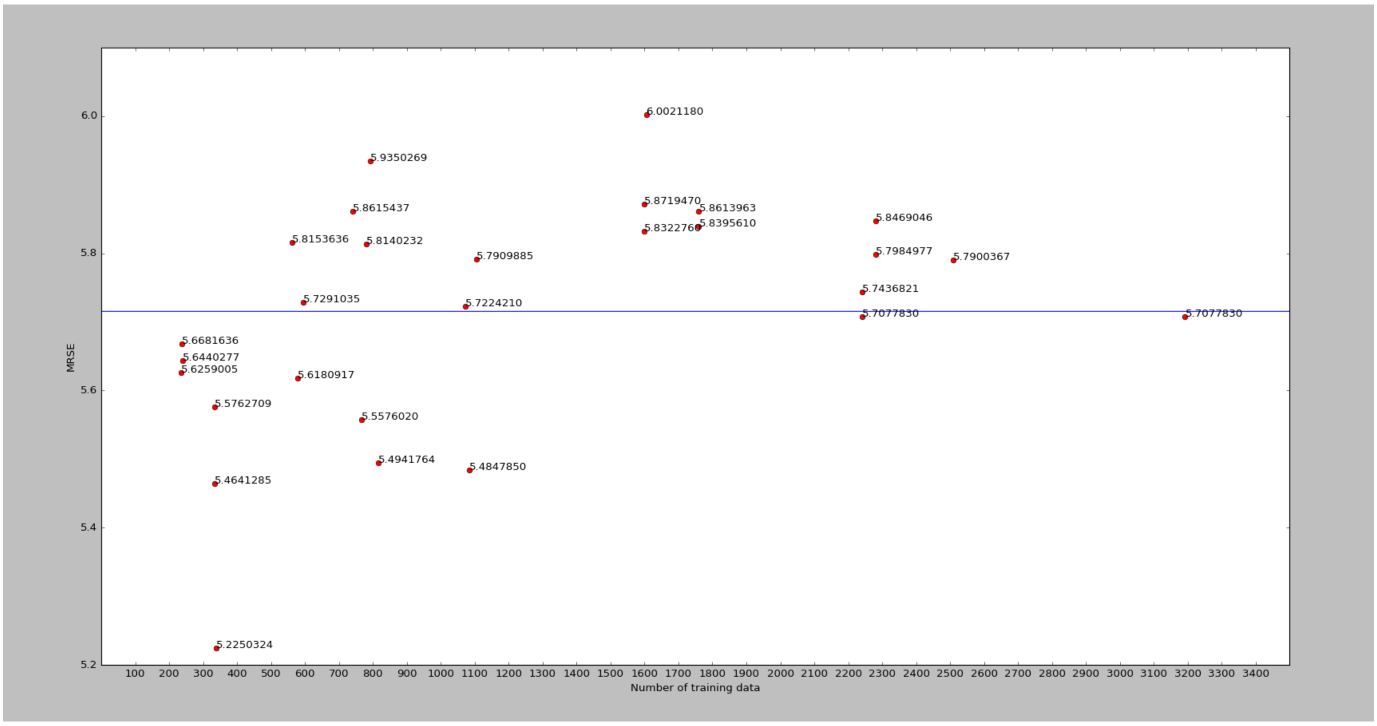
1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

答：

在一筆(train_x,train_y)的 pair 之中，train_x 裡面包含每天隨機的連續九小時的 PM2.5 資料。train_y 第十小時資料為 answer。(每天隨機取 n 筆共取 240 天)

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：



藍色水平線為平均誤差的值：5.715308372528903

原本是預想越大的訓練資料量，PM2.5 的預測會越準，但是圖上所顯示的不盡然如此。圖中可以看出，當訓練資料量大時，其誤差皆蠻接近平均的，而較小的資料量其誤差的分布的 variance 比較大。而在小的訓練量時，我推估因為 bias 的關係剛好蠻多組合都有較小的誤差。

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：

一次式的模型 $y = b + \sum_{i=1}^{9} w_i x_i$ （變數：$b, w_1, \ldots, w_9$）

特徵值：$x_1, x_2 \ldots . x_9$，為一連續九小時的 PM2.5 資料。

以此模型在 2500 筆訓練測資訓練出的最佳模型，其平均誤差為 5.798497729184221

二次式的模型 $y = b + \sum_{i=1}^{9} w_i x_i + \sum_{i=10}^{18} w_i {x_{i-9}}^2$（變數：$b, w_1, \ldots, w_{18}$）

特徵值：$x_1, x_2 \ldots . x_9$，為一連續九小時的 PM2.5 資料。

以此模型在 2500 筆訓練測資訓練出的最佳模型，其平均誤差為 7.217467116729584

一次式的模型搭配取 PM2.5 和 PM10 為 feature

$$y = b + \sum_{i=1}^{9} w_i x_i + \sum_{i=10}^{18} w_i x_{i-9}$$

（變數：$b, w_{1,} \dots, w_{18}$）

特徵值：$x_1, x_2 \dots . x_9$，為一連續九小時的 PM2.5 資料，$x_{10}, x_{11} \dots . x_{18}$為與 PM2.5 同時段的連續 PM10 資料。

以此模型在 2500 筆訓練測資訓練出的最佳模型，其平均誤差為 5.729103470087808

可以看出，越複雜的模型不見得有比較好的結果。最低的誤差為使用一次模型搭配 PM10，但 Kaggle 上最佳的結果 **5.60939** 為只使用 PM2.5 一次式模型。故推測有可能有 Overfitting 的情況發生。

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

答：

$y = b + \sum_{i=1}^{9} w_i x_i + \lambda \sum_{i=1}^{9} {w_i}^2$ (2500 筆資料，iteration 10000，learning rate 0.1)

$\lambda = 0, \mathrm{MRSE} = 5.82365$	$\lambda = 0.5, \mathrm{MRSE} = 5.8347853$	$\lambda = 1, \mathrm{MRSE} = 5.850734$
$\lambda = 5, \mathrm{MRSE} = 6.022398$	$\lambda = 10, \mathrm{MRSE} = 6.2203094$	$\lambda = 50, \mathrm{MRSE} = 7.135542$

λ 值越高，MRSE 越高，正規化並無法提升準確率。

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

$$\widehat{w} = argminS(w)\ \mathrm{X}$$

$$\mathrm{S(w)} = |\mathrm{y} - \mathrm{Xw}|^2$$

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

最小化損失函數的向量 $\widehat{w} = (\mathrm{X}^{\mathrm{T}} X)^{-1} X^T y$