學號：B03705006 系級： 資管三 姓名：侯舜元

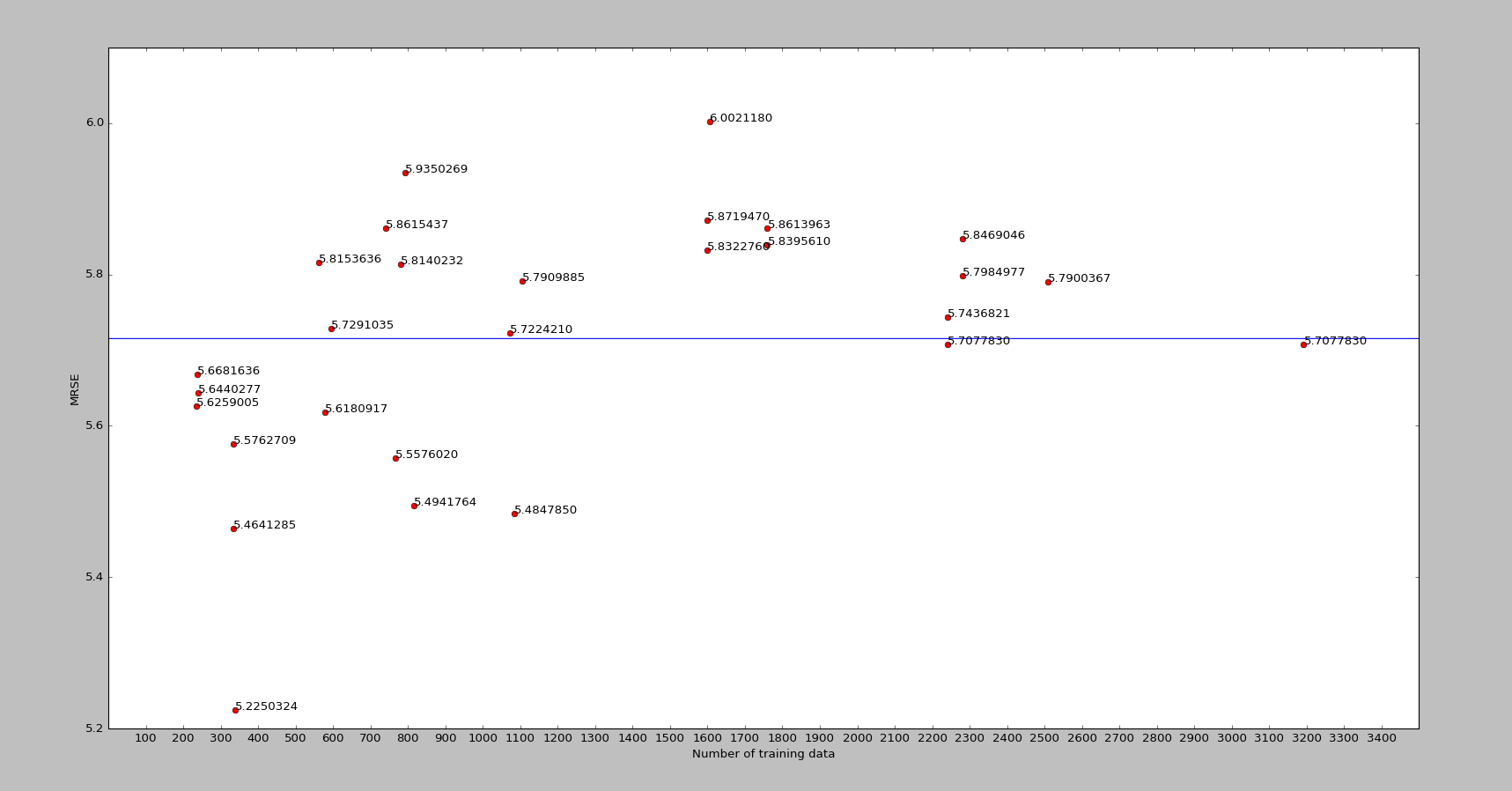
1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

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 （變數： …,）

特徵值：，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5.823414191898094

二次式的模型（變數： …,）

特徵值：，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

（變數： …,）

特徵值：，為一連續九小時的PM2.5資料，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：

(2500筆資料，iteration 10000，learning rate 0.1)

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

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 xn，其標註(label)為一存量 yn，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 。若將所有訓練資料的特徵值以矩陣 X = [x1 x2 … xN] 表示，所有訓練資料的標註以向量 y = [y1 y2 … yN]T表示，請以 X 和 y 表示可以最小化損失函數的向量 w 。

答：