學號：B03901103 系級： 電機三 姓名：陳學平

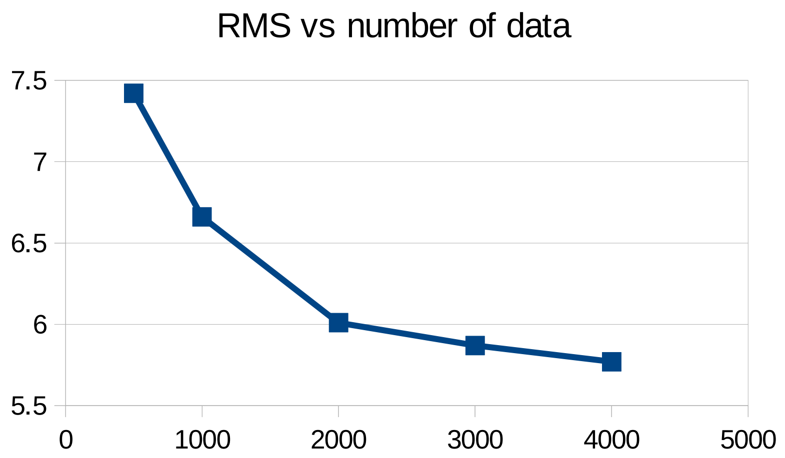
1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

答：將前九個小時的 18 個 feature，以及這些 feature 的平方值、兩兩相乘值經過 normalize 後，共 1701 個值作為輸入特徵，經過線性迴歸可以得到每個特徵值的權重，將這些權重的絕對值排序，取最大的 100 個，這 100 個權重對應的 features，即為最後所要抽取的 features。

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：本題固定使用的 features 為前九個小時的全部 feature，表格中為 RMS error

Training size	Training set	Validation set(size=1000)
100	368.78	1386
500	4.55	7.42
1000	5.29	6.66
2000	5.50	6.01
3000	5.59	5.84
4000	5.65	5.77



顯然的，隨著資料量的提升，validation set 上的 error 也就愈小，這是由於大的資料量可以降低 training data 的 noise 對模型造成之干擾。

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：表格中為 RMS error

	Training set(size = 4642)	Validation set(size = 1000)
只用 PM2.5 作為 feature	6.19	5.83
18 個 features 皆用	5.69	5.75
18 個 features、features 平方	5.51	5.78
18 個 features、平方、兩兩相乘	4.25	20.4

隨著模型的複雜度增加，在 validation set 上的 root mean square error 是先下降，而又上升。一開始 rms error 會下降的原因是由於複雜的模型之 bias 變小，但隨著模型愈來愈複雜，overfitting 的現象會愈來愈嚴重，導致雖然在 training set 上之 error 很小，但 validation set 上之 error 很大。

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

答：本題使用的 features 為 18 個 feature 加上平方項、兩兩相乘，使用 L2 – regularization

表格中為 RMS error

Lamda	Training set(size = 4642)	Validation set(size = 1000)
0	4.25	20.4
0.1	4.30	17.7
1	4.36	13.9
10	4.43	9.41
100	4.54	7.66
1000	4.68	7.12
100000	5.05	6.31
1000000	12.12	12.07

regularization 可以用來抑制複雜模型之 overfitting 的問題，由上表可知，隨著調增 regularization 的比重，validation set 上的 error 是可以達到顯著的下降的，不過如表格最後一列顯示，若是 regularization 的比重過重，則又可能會對模型造成 bias，因而又使 error 提高。

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 n=1Nyn-wxn2 。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：

Loss Function = $(X^Tw-y)^T(X^Tw-y)$ ，由 Loss Function 對 w 微分等於零可得 w 如下：

$w = (XX^T)^{-1}Xy$