學號：B04902045 系級： 資工二 姓名：孫凡耘

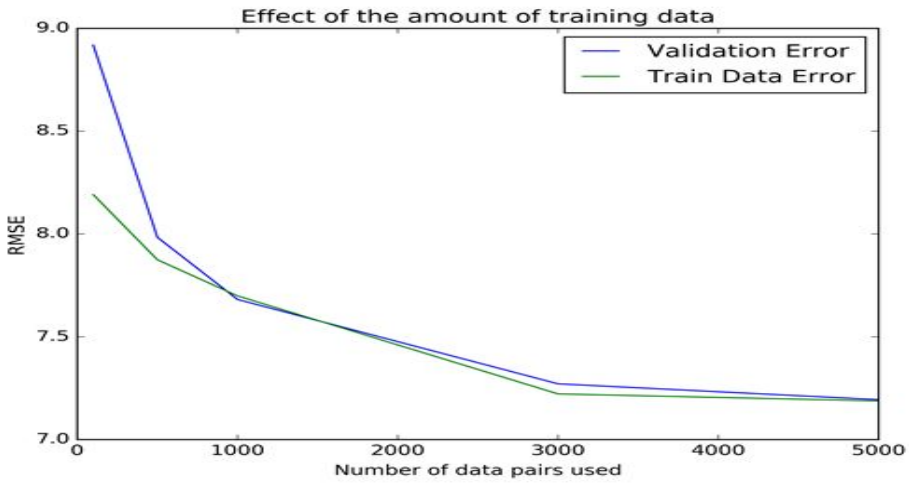
1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

I calculated the Pearson Correlation coefficient between all features and PM2.5. Aside from this, I read science articles about what affects PM2.5. I manually selected PM2.5, PM10, NO2 as effective factors. But, how many previous days of each factor should I choose ? I applied cross validation using K-fold method(K = 12), that is, every time I train my model with 11 months of data and let the month left out be the validation set. I wrote a shell script that tries different amount of features and compare their results by the error calculated by cross validation. At last, I decided on the following parameters.

PM2.5 : 9
PM2.5**2 : 9
PM10 : 8
PM10**2 : 6
O3 : 3
RAINFALL : 2
NO2 : 1
All the comparisons made below are based on this set of parameters.

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

K-fold cross validation with K = 12
RMSE of Validation Error is obtained by calculating the mean of the K RMSEs
Parameters: 1000 iterations, initial_learning_rate = 1e-2, lambda = 100



As we can see, generally the accuracy increases if I use more data pairs under a single iteration. But this doesn't mean that more iterations gives better results, since too many iterations may result in overfitting(not discussed in this problem).

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：

K-fold cross validation with K = 12

Using only PM2.5*9 :

RMSE on training data: 6.196922

Mean of RMSE with K-fold cross validation: 6.123827

Using the parameter I chose at last (Please refer to Q.1):

RMSE on training data: 5.825633

Mean of RMSE with K-fold cross validation: 5.920034

Using all parameters possible(including quadratic term of every feature):

RMSE on training data: 5.509979

Mean of RMSE with K-fold cross validation: 7.215646

Using more complex models can fit better to the training data, but not necessary to the for testing data set. As you can see, the rmse on training data decreases as the complexity of the model increases, but obviously it didn't fit better to the testing data since too complex model might result in overfitting.

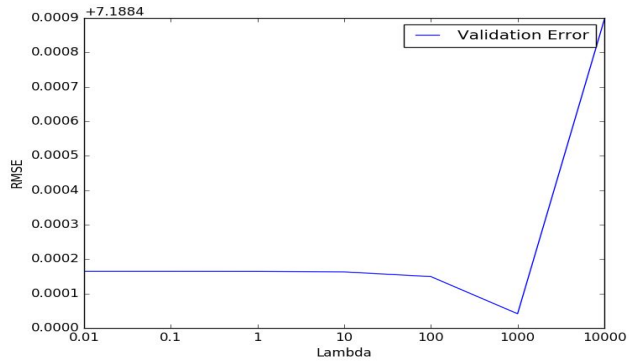
4. 請討論正規化(regularization)對於PM2.5預測準確率的影響

K-fold cross validation with K = 12

RMSE of Validation Error is obtained by calculating the mean of the K RMSEs

Parameters: 500 iterations, initial_learning_rate = 1e-2

Lambda value: 1e-1, 1, 10, 1e2, 1e3, 1e4



We can see that when lambda = 1000, the performance of this model will be the best. However, if the value of lambda is not chosen correctly, the result will be even worse.

5. $w = (X^T X)^{-1} X^T y$