Machine Learning 2017 HW1 Report

Predict PM 2.5

B03901156 Yu Xuan Huang

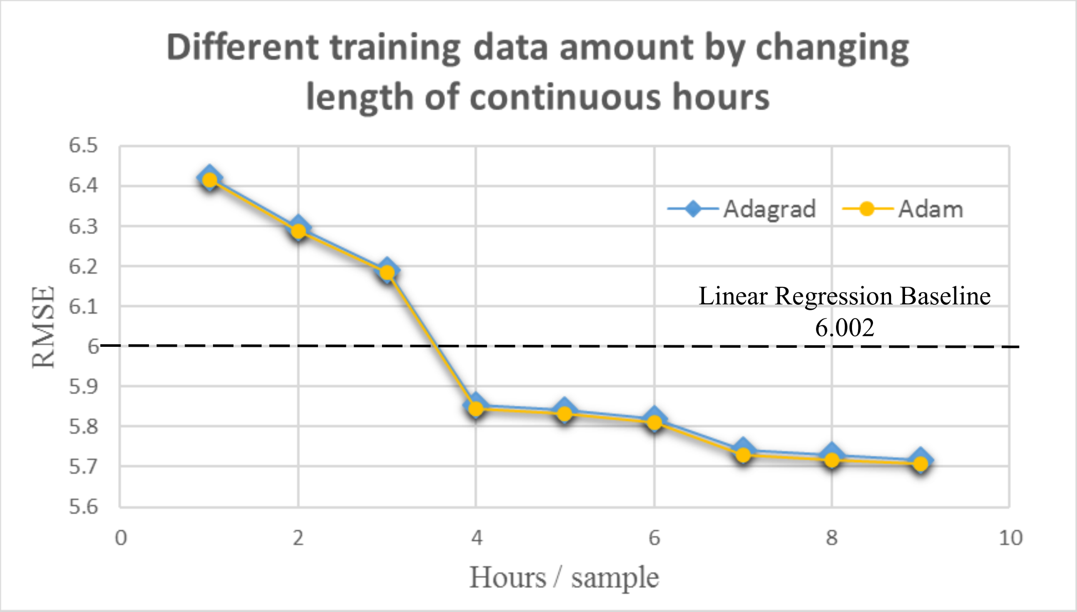
1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

ANS. The raw train.csv file consists of 18*24 features in each day. Since my goal is to predict the 10th hour's PM 2.5 value in train_X.csv file according to the previous 9 hours' features, I extract the training pairs (x, y) by concatenating the first 9 hours' features (dimension = 9*18 = 162) and utilizing the 10th hour's PM 2.5 to be its label. In addition, I replace non-numeric feature value "NR" in train.csv and test_X.csv with 0.

In addition, I also declare feature_list in my code to choose the adequate features from the 18 air pollution indexes for training in order to discuss about the answer of problem 2.

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

Ans. In this problem, I change the length of continuous hours to change the amount of training data.



▲ Experiment result for different training data amount by changing length of continuous hours (#iteration = 200000)

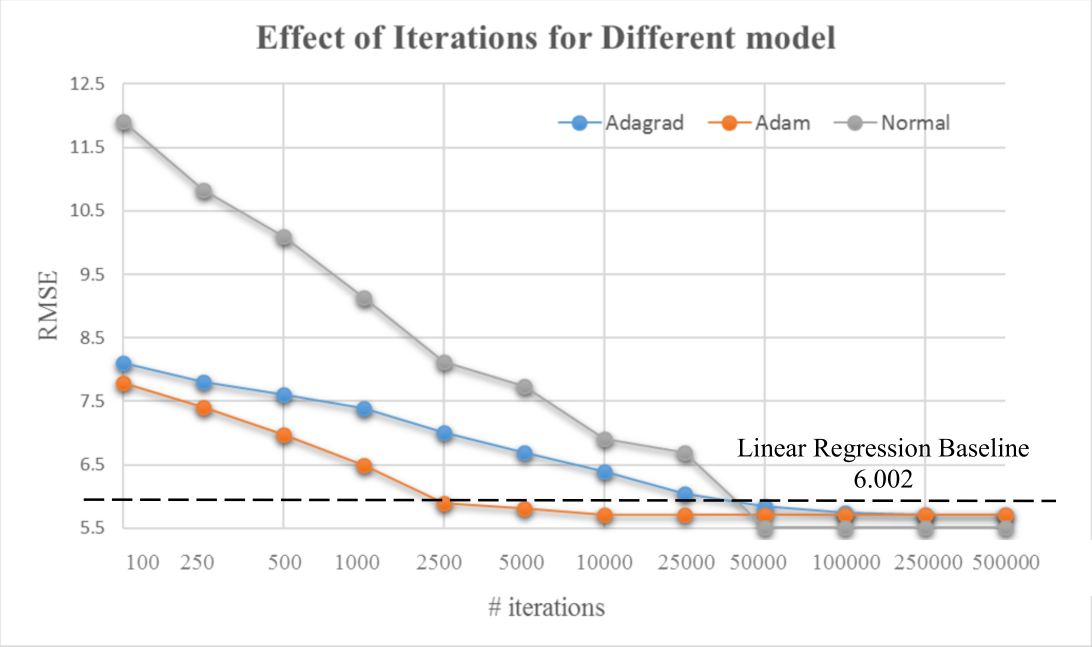
3. 請比較不同複雜度的模型對於 PM2.5 預測準確率的影響

Ans. After doing the experiments with normal, adagrad and adam model, there are some observations obtained from the result:

a. The normal model may generate the best result if given enough iteration for running the training process.

b. The adagrad and adam model can obtain better results if the number of iteration is limited.

c. From the observation of the result of the experiments, regularization doesn't improve the result much better, and sometimes even generates worse result.



▲ Experiment result for effect of iterations for different model (hours / sample = 9)

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

Ans. After adding lambda to regularize the computing formula, the experiment result seems that it doesn't perform better than the original one (from the view of RMSE on the below) and even gets worse (from the error rate on the kaggle). In my viewpoints, I think the reasons why regularization doesn't help training process to get better result may be:

a. The increasing of error rate of the public set when regularization is utilized on the training process seems to hint that my model doesn't overfit the training data, which indicates the model may be more adequate when applied on other data.

b. Since my model presents functions using multivariate linear equations which is less possible to overfit the training data than quadratic functions.

I have done regularization experiments with different value of lambda, and the following is the result. In the theory, the larger the value of lambda, the harder to gain lower RMSE (root mean square error), and the result on the below fits this hypothesis.

Lambda with learning rate = 1e-10	**RMSE**
0	5.8470885
0.01	5.8470885
0.1	5.8470890
1	5.8470938
10	5.8471414

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 $X = [x^1\ x^2 \ldots x^N]$ 表示，所有訓練資料的標註以向量 $y = [y^1\ y^2 \ldots y^N]^T$表示，請以 X 和 y 表示可以最小化損失函數的向量 w 。

Ans. $w = yX^{-1}$