學號：B05901189 系級： 電機一 姓名：吳祥叡

1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

答:

(1) for hw1_best.sh

取連續 9 小時的 pm2.5 數據[x0,x1,…,x8]

以及每兩個數據點的相乘:[x1*x1,x1*x2…].總共 9+9*10/2=54 項.

(2)for hw1.sh

取連續 9 小時的 pm2.5 和 pm10 數據[x0,x1,…,x17]

以及每兩個數據點的相乘:[x1*x1,x1*x2…].總共 18+18*19/2=189 項.

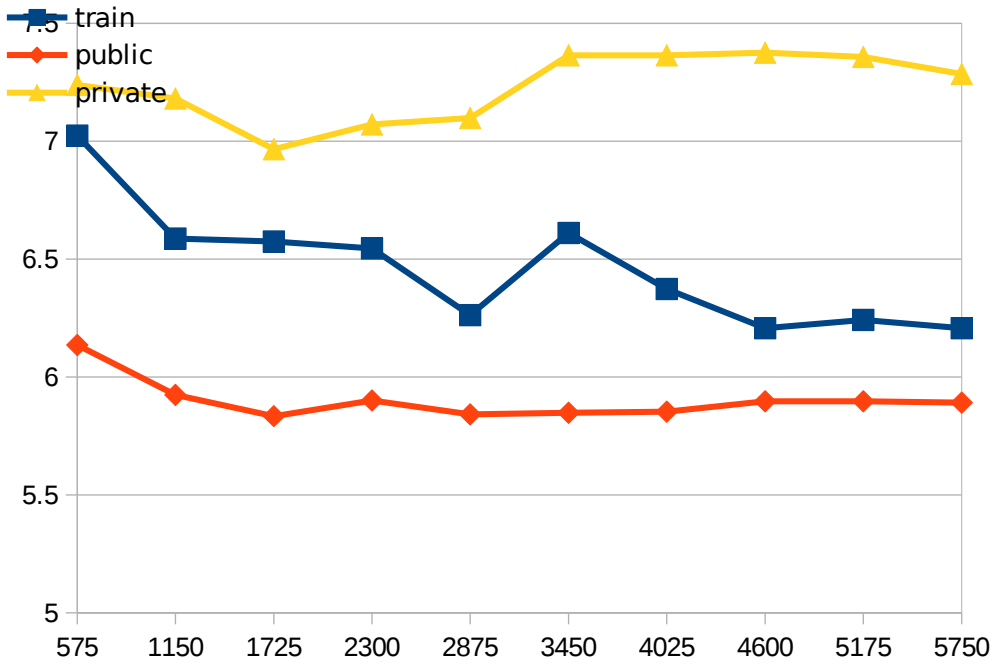
2.請作圖比較不同訓練資料量對於 PM2.5 預測準確率的影響

答:

以下均用簡化模型探討,只輸入 9 小時的 pm2.5 數據當作 feature.

最大訓練資料量定為 24*240-10 組,由 train.csv 給定.

下圖 y 軸為 RMS Error,x 軸為訓練資料量.



一階模型在 training set 上看起來訓練資料量越大會越準確,但在 test set 上其實沒有明顯進步.

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:

模型	RMS Error after 10 min training
All feature, 1st order	5.90903
Feature pm2.5, 1st order	5.82595
Feature pm2.5, 2nd order	5.73658
Feature pm2.5, 3rd order	5.75398
Feature pm2.5 and pm10, 2nd order	5.89915

K Nearest Neighbor, K=1	10.2160
KNN, K=all, weight=exp(-distance**3/10)	7.58452
KNN, K=all weight=distance**-1	10.69424
KNN, K=all weight=distance**-3	7.30354
KNN, K=all weight=distance**-5	8.11685

以上除了 K Nearest Neighbor 模型以外均為 non-stochastic adagrad gradient descent linear regression without regularization 的結果,且 1 階 2 階 3 階的意思在第一題有說明.KNN method 的 distance 定義為 9 個 pm2.5 數據標準化後的方差合.

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

答:

Method	Resulting RMS Error
2nd order lambda=0	5.73658
2nd order lambda=1000*r_learn/sigma	5.80373
2nd order lambda=10*r_learn/sigma	5.74601
3rd order lambda=1000*r_learn/sigma	5.80230
3rd order lambda=0.01r_learn	5.75398

以上均為 non-stochastic adagrad gradient descent linear regression, 取 pm2.5 的數據當作 feature.

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

答:

$w=(X^TX)^{-1}X^Ty$

Reference: http://cs229.stanford.edu/notes/cs229-notes1.pdf page11

##To 助教：我的程式會跑 10 分鐘 而且會生出暫存的 txt 檔 希望不會有意外