

## 機器學習 HW1

學號：R04522631 系級：機械碩二 姓名：盧玄真

1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

答：

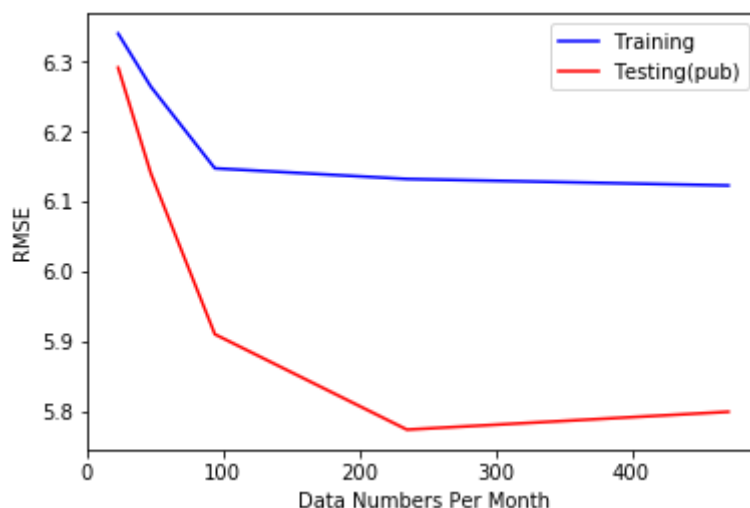
**概念：**先將特徵照月份分成 12 份，然後以每 9 小時為單位跨日將每個月分成 471 筆，因此最後輸入特徵會變成一個 [12(月份), 471(筆數), 9(輸入特徵  $x_1 \sim x_9$ )] 的陣列。

**實作：**

```
data = pd.read_csv(Train_D, encoding = "big5") #讀入原始資料
def Tfprocess(data, S):                        #data 為原始資料, S 為所選特徵
    V = data[data["測項"]==S]                  #讀取所選特徵
    V = V.drop(['日期', '測站', '測項'], axis = 1)
    V = np.array(V, float)
    v = np.reshape(V, (12, 480))              #將資料照月份分好
    datarow = len(v)                           #月份
    datacol = len(v[0])                         #總小時數
    DSV = np.zeros((datarow, datacol-9, 9), float) #宣告儲存陣列
    for i in range(0, datarow):                 #將輸入特徵每九小時一筆分好
        for j in range(0, datacol-9):
            for k in range(0, 9):
                DSV[i, j, k] = v[i, j+k]
    return DSV                                #回傳 feature 陣列
```

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：如上題所示，我最多一個月有 471 筆 training data，所以本題減少訓練資料量是選取一個月的前幾筆 (Batch size) 作為控制資料量的方法。如下圖我設定每月資料量由左到右分別為 [23, 47, 94, 235, 471] 筆，從圖中可以看出 train error 的確隨著訓練資料量越大則越小，但是資料量大到一定程度時 training error 就不太會下降了，但是 testing error 卻在 235 筆的時候出現一個極值，因為隨著 batch size 增加，達到一定精準度的 epoch 變少，但是計算成本增加因此在 235 筆時達到最佳的 batch size。



### 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：在本次作業中我都是選取 pm2.5 作為 feature 因此假設 pm2.5 的特徵向量為  $x$ ， $b$  為偏移量， $w_i$  為參數向量，下表為不同複雜度 training error 以及 testing error 比較。

model	Training RMSE	Testing RMSE(public)	Testing RMSE(private)
$b+w_1x$	6.12302	5.79970	7.30647
$b+w_1x+w_2x^2$	6.26164	5.92952	6.30983
$b+w_1x+w_2x^2+w_3x^3$	6.26164	5.90239	7.25834

其中為了讓訓練結果在同樣的訓練次數得到相近的結果，我對不同的特徵( $x, x^2, x^3$ )作 feature scaling。在表中可以看到 training error 跟 public testing error 的趨勢是相近的，但是 private testing error 在 model 最高階為二次式的時候有最好的結果，代表 pm2.5 這個特徵最高階為二次的 model 是最適合 pm2.5 預測的 model，到了最高階為三次的时候就發生了過度訓練(overfitting)。

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

答：理論上正規化是簡化過度訓練的模型，但是因為我所取的特徵為前九個小時的 pm2.5 之一次式，所以本身就已經 underfitting 因此正規化只會讓誤差更加增大。在此沒有正規化才是最佳的模型。

lamda	Training RMSE	Testing RMSE(pub)	Testing RMSE(private)
1	6.132267081	5.88590	7.40784
0.1	6.123131404	5.80813	7.31656
0.001	6.123021534	5.79977	7.30657
0.0001	6.123021522	5.79970	7.30648
0	6.123021522	5.79969	7.30647

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  $n=1N y_n - w x_n^2$ 。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：最小損失函數向量  $w$  即為小二乘方解  $w = (X^T X)^{-1} X^T y$