學號：b04902053  系級： 資工二  姓名：鄭淵仁

**1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)**

**答：**

我總共寫了兩個版本：

hw1\_best.py取前9個小時的pm2.5指標做一維和二維的feature。

hw1.py則是取前9個小時的pm10、pm2.5、RAINFALL指標做一維和二維的feature。

**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：**

我把整個data set分一半，一半用來計算RMSE，另一半則是用來train。而train的時候，我只拿這一半資料中的共12種不同的資料量去train出結果。除此之外，為了怕結果因為資料量太小導致誤差太大，我又多取了不同的資料train出不同結果，再算出RMSE並平均起來。得到的結果如下圖：

從圖中可以發現：在資料量等距增加的時候，RMSE會下降得越來越慢。看起來很像是和資料量成反比的圖形。

而我在網路上查到以數學推導RMSE的教學，發現：數學推導出來的公式確實是和資料量成反比。（公式：，其中是資料量、是feature的數量）

**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：**

我先把資料shuffle一遍，再把資料切成兩塊，一塊拿來train，另一塊則是當作validation set。接著針對同一筆變數同時用1次、2次、3次的複雜度來train，算出RMSE，再分別對維度取平均。結果如下面兩個表格：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 表一 RMSE先標準化再平均 | |  | 表二 RMSE直接平均 | |
| dim | **RMSE先標準化再平均** |  | **dim** | **RMSE直接平均** |
| 1 | -0.815379555 |  | **1** | 11.17740882 |
| 2 | -0.07284098 |  | **2** | 11.2917164 |
| 3 | 0.888220534 |  | **3** | 16.28339887 |

在**表一**裡面，可以看出來1次的預測效果最好，2次的效果其次，3次的效果最差。

在**表二**裡面，可以看出來1次和2次的預測效果很接近，但是3次的效果明顯比較差，很像是overfitting的現象。

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

**答：**

我先把資料shuffle一遍，再把資料切成兩塊，一塊拿來train，另一塊當作validation set。而下表就是對於不同的維度和regular數值，算出的RMSE的結果。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 表三 對於不同的dim和regular數值，RMSE的結果 | | | | | | | |
| regular  dim | **0** | **10** | **100** | **1000** | **10000** | **100000** | **1000000** |
| 1 | 5.904 | 5.902 | 5.896 | 5.908 | 6.531 | 48.957 | 227.289 |
| 2 | 5.916 | 5.915 | 5.910 | 5.937 | 6.512 | 10.175 | 180.401 |
| 3 | 5.927 | 5.925 | 5.941 | 5.965 | 7.087 | 12.865 | 64.950 |

從表格中可以發現：在regular很小的時候，對資料不會有影響，但是當regular比10000大之後，不管維度是多少，RMSE都會變大。

**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xn，其標註(label)為一存量yn，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 。若將所有訓練資料的特徵值以矩陣X = [x1 x2 … xN]表示，所有訓練資料的標註以向量y = [y1 y2 … yN]T表示，請以X和y表示可以最小化損失函數的向量w。**

**答：**

設loss function：

則

故

設 ，則