學號：B02705027 系級： 資管四 姓名：陳信豪

1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

答：

我資料抽取的方式是先將前九小時的所有 feature （一維）取出來

做成長度為162 （9 hr \* 18 feature）的一維陣列 stat\_list

二維或三維數據則是藉由直接將 stat\_list 裡頭的數據做平方或立方

再跟原本的 stat\_list 合併

the\_stat\_list = stat\_list

stat\_list = stat\_list + [x\*\*2 for x in the\_stat\_list]

另外我有設三個 Boolean 陣列

feature\_Take 對應 18 個天氣指標

time\_Take 對應前九小時到前一小時

param\_Take 對應全部 162 個 feature

如果只取前 7 小時的 PM2.5：

feature\_Take = [

False, False, False, False, False, False, False, False, False,

True, False, False, False, False, False, False, False, False]

time\_Take = [False, False, True, True, True, True, True, True, True]

param\_Take = np.repeat(feature\_Take, 9) \* np.tile(time\_Take, 18)

x\_Take = train\_X[:, param\_Take]

# train\_X 是 stat\_list 的總集合，其 train\_X.shape[1] 是 162

# x\_Take 是 feature 塞選後的要訓練的資料集合，其 x\_Take.shape[1] 是 7

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：

(作圖：hw1\_q2.py)

取前9 小時 PM2.5 一維作為 features

切 50% 總資料作為 Validation Set

再從剩下的 50% 總資料中取 10%、20%、30% ... 100% 的資料作為訓練用

訓練出模型，觀察並紀錄 Validation Set 的 RMSE

**結論是資料量與 Valid RMSE 呈負相關**

**也就是資料量越多，模型預測 PM2.5 的準確率愈高。**

|  |  |  |
| --- | --- | --- |
| 訓練資料比例 | 訓練資料量 | Valid RMSE |
| 10% | 274 | 5.80577499154 |
| 20% | 549 | 5.75516028391 |
| 30% | 823 | 5.74218464446 |
| 40% | 1098 | 5.71816274647 |
| 50% | 1373 | 5.70927341202 |
| 60% | 1647 | 5.73849114808 |
| 70% | 1922 | 5.70737649023 |
| 80% | 2196 | 5.69982396921 |
| 90% | 2471 | 5.70800780233 |
| 100% | 2746 | 5.70426224388 |



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：

取前9 小時 PM2.5 的資料

分別配對一維、二維和三維作為 features

並切 50% 總資料作為 Validation Set

以下為各組合（複雜度）訓練出來後的 RMSE

**結論是模型複雜度並不是越複雜越好**

**我們可以看出在此情況下，一維就已有不錯的效果**

|  |  |  |
| --- | --- | --- |
|  | Valid RMSE | |
| 複雜度 | max\_iteration = 1000 | max\_iteration = 3000 |
| 一維 | 5.75648 | 5.70759 |
| 二維 | 11.44504 | 10.46105 |
| 三維 | 17.61086 | 17.54742 |
| 一維 ＋ 二維 | 6.90109 | 6.53077 |
| 二維 ＋ 三維 | 51.00183 | 10.76510 |
| 一維 ＋ 三維 | 10.04785 | 9.44541 |
| 一維 ＋ 二維 ＋ 三維 | 9.05310 | 8.32828 |

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

答：

(作圖：hw1\_q4.py)

取前9 小時 PM2.5 一維作為 features

切 50% 總資料作為 Validation Set

另外 50% 通通用來做訓練

分別以lambda = [0, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10] 來正規化

將原本的 loss function 加上 lambda \*

訓練出模型並觀察並紀錄 Validation Set 的 RMSE

**結論是 lambda 有助於降低 Valid RMSE，其值通常不必太高**

**在這邊的情況是 lambda 約等於 0.05 時最合適**

|  |  |
| --- | --- |
| lambda | Valid RMSE |
| 0 | 5.75661445687 |
| 0.001 | 5.75658710753 |
| 0.005 | 5.75648372625 |
| 0.01 | 5.75636803907 |
| 0.05 | 5.75598471407 |
| 0.1 | 5.75686352377 |
| 0.5 | 5.81837150599 |
| 1 | 6.00503199987 |
| 5 | 8.75156959083 |
| 10 | 14.9610170793 |



(左圖為全覽，右圖為放大 x 軸 0~0.1)

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：

假設有 P 個 feature

w 維度：1 \* P

xn 維度：1 \* P

y 維度：N \* 1

X 維度：N \* P

w = yT X(XT X) -1