

學號: R05944043 系級: 網媒碩一 姓名: 宋焱檳

1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

答:

Linear-regression 嘗試過的提取方式。(一) 大多數情況下優於 (二)

(一):

取連續每 9 個小時的 pm2.5 指標做一維和二維的 feature:

$\text{train\_x} = [[x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9], [x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}], \dots]$

並且忽略掉十二月二十日的 15-23 時這種情況(因為, 沒有下一小時的 PM2.5 指標) 因為一個月只有 20 天的 data, 故每 480 (20\*24) 組 data 的後 9 組要 delete 掉。最終得到需要的 feature

(二):

取前 9 個小時的所有空氣污染指標(17 種, 除了 RAINFALL, 因為該指標多數為空, 不具有參考價值):  $\text{train\_x} = [x_1, x_2, \dots, x_{153}]$

$\text{train\_x} = [[x_1, x_2, \dots, x_{153}], [x_2, x_3, \dots, x_{154}], \dots]$

並且忽略掉十二月二十日的 15-23 時這種情況(因為, 沒有下一小時的 PM2.5 指標) 因為一個月只有 20 天的 data, 故每 480 (20\*24) 組 data 的後 9 組要 delete 掉。最終得到需要的 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:

Model:  $y = b + w_1 * x$

Loss function =  $(y_{\text{head}} - y)^2$

Iteration : 1,000,000 Adagrad

Feature	LR	LOSS	Public Score
PM2.5 (5652 * 1)	1.0	37.235	5.799
11 個月 PM2.5(5082 * 1)	1.0	38.12	5.818
18 種氣體(5652 * 18)	1.0	32.522	5.954

因為 feature 的種類多了, 無關 data 的影響也大了, 因為不確定哪些 data 是和 PM2.5 直接相關。所以, 在 Public 上的結果變差了。但是由於取的參數多, 所以 LOSS 變低了。用全部 PM2.5 的 data train 的時候, 在 Public 上的分數比較好, 過了 strong line, 用 11 個月的 PM2.5 的 data train 時, LOSS 變高了, Public 的表現變差了, 但是在 Private 上很可能比全部 train 出來的效果好, 因為這是根據對於自己設定的 Test data 的 LOSS 選出來的, 可以將 PublicScore 視作 Private 的得分, 而且, 選擇 model 時不會因為 Public Score 而改變, 不會出現 Overfitting 的情況。

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：

$$\text{Loss function} = (\text{yhead} - y)^2 + 0.1 * \sum (w^2)$$

Iteration :1,000,000 LR = 1.0 Adagrad

Model	Feature	LOSS	Public Score
$y = b + w1 * x$	PM2.5 (5652 * 1)	37.235	5.799
$y = b + w1 * x + w2 * x$	PM2.5 (5652 * 1)	37.374	5.757

在相同情況下，複雜度高的模型，得到的 LOSS 相對比較高，但是由於考慮的情況包含低複雜度的模型的情況，所以在 PublicScore 上的表達較好。

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

答：

$$\text{Model: } y = b + w1 * x$$

Iteration :1,000,000 LR = 1.0 Adagrad

Loss function	Feature	LOSS	Public Score
$\text{Loss} = (\text{yhead} - y)^2$	PM2.5 (5652 * 1)	39.01	5.845
$\text{Loss} = (\text{yhead} - y)^2 + 0.1 * w1^2$	PM2.5 (5652 * 1)	37.235	5.799

在複雜度相同的情況下，Regularization 使得 variance 變小了，也減小了 LOSS，所以在 Public Testdata 上的表現變好了。

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  $\sum_{n=1}^N (y^n - w * x^n)^2$ 。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：運用最小平方法，求最小化 Loss 在  $w$  的微分為 0 時出現，

先對 Lossfunction 求微分 求 MIN(Lossfunction)，

$$X^T X w = X^T y, \text{ 解方程}$$

$$\text{Thus, } w = (X^T X)^{-1} X^T y$$

心得與總結：

這次的實驗，需要注意對數據的觀測，儘早發現日期不是連續的，每個月只抽取了 20 天作為 train 的 data，處理上要注意。一般情況下，需要先在自已設定的 validation 上得到不錯的 Loss，選擇好 model，在經過所有 data train，即使在 PublicScore 上可能比不那麼高，但是綜合 private 後，應該會有很大的提升。畢竟不會因為一半的 Score 影響整體 model。