

學號：B03901161 系級：電機四 姓名：楊耀程

請實做以下兩種不同 feature 的模型，回答第 (1) ~ (3) 題：

1. 抽全部 9 小時內的污染源 feature 的一次項(加 bias)
2. 抽全部 9 小時內 pm2.5 的一次項當作 feature(加 bias)

1. (2%)記錄誤差值 (RMSE)(根據 kaggle public+private 分數)，討論兩種 feature 的影響

Ans:

取全部污染項與只取 pm2.5 的結果，皆 9hr

| | 全部污染項 | 只取 pm2.5 |
|------|---------|----------|
| RMSE | 7.74344 | 7.49832 |

討論: 只取 pm2.5 比取全部污染項的 RMSE 更低，可見並不是 feature 取越多越好，很可能全部污染項中有的資訊不是太影響 pm2.5，全部採用反而會造成 overfitting。反之，只取 pm2.5 雖然資訊略少，但是也不易發生 overfitting 的現象，因此只取 pm2.5 在這邊結果還更好，我認為最理想的情況是 pm2.5 搭配數個較有用的污染項，可能多幾遍嘗試就能降低 testing data 的 RMSE(這也是我找 hw1_best model 的過程)。

2. (1%)將 feature 從抽前 9 小時改成抽前 5 小時，討論其變化

Ans:

取連續 9hr 與 5hr 的結果，皆 18 個 features(全部污染項)

| | 連續 9hr | 連續 5hr |
|-----|---------|---------|
| MSE | 7.74344 | 7.84615 |

討論: 取連續九小時比取連續五小時來的 RMSE 更低，這樣的結果蠻符合預期，因為 testing data 也都是連續九小時，如果都只用到前五小時，會少很多資訊，造成精準度的下降。也有可能是因為 pm2.5 model 的複雜度就是比較適合連續九小時的資料，用連續五小時的資料可能會資訊不足(沒用到前九到六小時)或是 overfitting(資料筆數更多)。

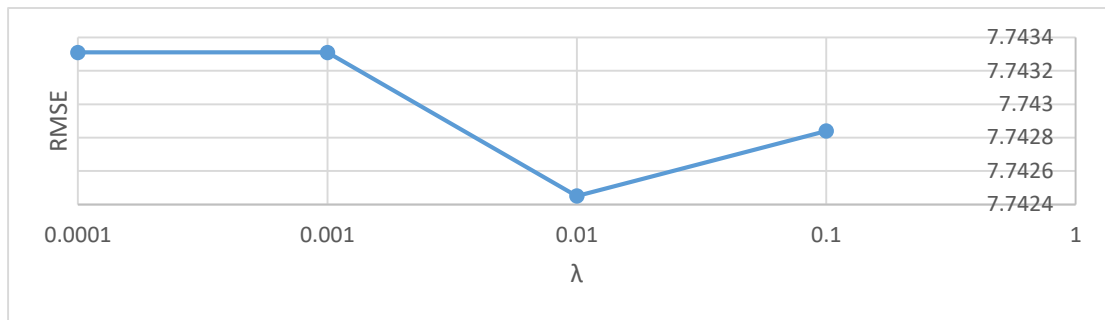
3. (1%)Regularization on all the weight with $\lambda=0.1$ 、 0.01 、 0.001 、 0.0001 ，並作圖

Ans:

皆採用 18 個 features(全部污染項)，9hr

| λ | 0 | 0.1 | 0.01 | 0.001 | 0.0001 |
|-----------|---------|---------|---------|---------|---------|
| RMSE | 7.74344 | 7.74284 | 7.74245 | 7.74331 | 7.74331 |

作圖:



討論:

採用較多的 feature，比較容易有 overfitting 的現象，為了看清楚 regularization 的效果，因此我都採用全部污染項去 train。在這樣的條件下，有 regularization 確實降低了 MSE，在 $\lambda = 0.01$ 時得到了最小的 MSE，不過改善幅度非常小，差距在小數點三位之後，可見此處 regularization 影響沒非常大。

4. (1%)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 $\sum_{n=1}^N (y^n - x^n \cdot w)^2$ 。若將所有訓練資料的特徵值以矩陣 $X = [x^1 x^2 \dots x^N]^T$ 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，請問如何以 X 和 y 表示可以最小化損失函數的向量 w ？請寫下算式並選出正確答案。

- (a) $(X^T X) X^T y$
- (b) $(X^T X)^{-0} X^T y$
- (c) $(X^T X)^{-1} X^T y$
- (d) $(X^T X)^{-2} X^T y$

Ans: (c)

算式:

$$L = \sum_{n=1}^N (y^n - w \cdot x^n)^2 = (y - Xw)^T (y - Xw)$$

所求為 Minimize L 時的 w 。令 L 對 w 做偏微分為 0 得

$$\frac{\partial L}{\partial w} = 0$$

$$\begin{aligned} & \frac{\partial ((y - Xw)^T (y - Xw))}{\partial w} \\ &= 2(X^T X w - X^T y) = 0 \\ & w = (X^T X)^{-1} X^T y \end{aligned}$$

故選(c)