

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

feature	<i>Private</i>	<i>Public</i>	<i>Sum</i>
pm2.5	5.83648	7.39590	13.23238
all	5.59229	7.62214	13.21443

從結果可以看到,當只考慮 pm2.5 時, RMSE 較考慮所有 feature 來的高, 意思是當提供所有資訊的預測結果較好. 這是很合邏輯的, 因為我們知道事實上 pm2.5 的確會受到比方說降雨等因素影響 (比方說下雨時 pm2.5 會下降, 一定程度上相關), 這意味着當提供這些 extra 的資訊可能會幫助預測.

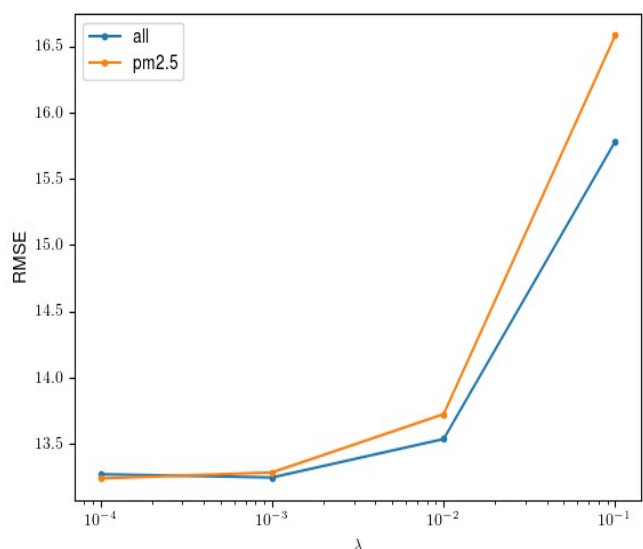
2. (1%)將 feature 從抽前 9 小時改成抽前 5 小時，討論其變化

feature	<i>Private</i>	<i>Public</i>	<i>Sum</i>
pm2.5	5.86272	7.52172	13.38444
all	5.44282	7.63383	13.07665

將 feature 改成 5hr 後, 發現在考慮所有的 feature 情況, RMSE 是下降的 (但其實變化不大) 但在只考慮 pm2.5 feature 時卻上升, 有可能是因為在只考慮 pm2.5 時, 從 9hr 變成 5hr, 實際上資訊量是變少的, 故導致預測較差, 但在考慮所有 feature 時, 可能某些短期變化對 pm2.5 的影響高, 故在考慮只有 5hr 時, 較長時間的 data 不須 fitting 情形下, fitting 能夠得到較好的預測結果.

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

λ	Pm2.5 RMSE (public+private)	All RMSE (public+private)
0.0001	13.23742	13.26916
0.001	13.28273	13.24471
0.01	13.72251	13.53497
0.1	16.58079	15.78113



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

Ans: (c)

推導：

我們想求未知 w ，使得其儘可能滿足式 (1)

$$y = Xw \quad (1)$$

$$X^T y = X^T X w \quad (2)$$

$$(X^T X)^{-1} X^T y = (X^T X)^{-1} (X^T X) w \quad (3)$$

$$(X^T X)^{-1} X^T y = w \quad (4)$$

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