

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:

抽全部 feature 的誤差值(public+private)為 13.2843，僅用 pm 2.5 的誤差值為 13.23238。僅使用 pm2.5 的成績較使用全部 feature 要來的好，可能因為抽全部 feature 是一個很複雜的 model，很可能會 overfitting，導致在 testing 的結果變差。然而單純使用 pm2.5 可能又是太簡單的模型，但仍然比使用全部 feature 來的好一點。

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

ANS:

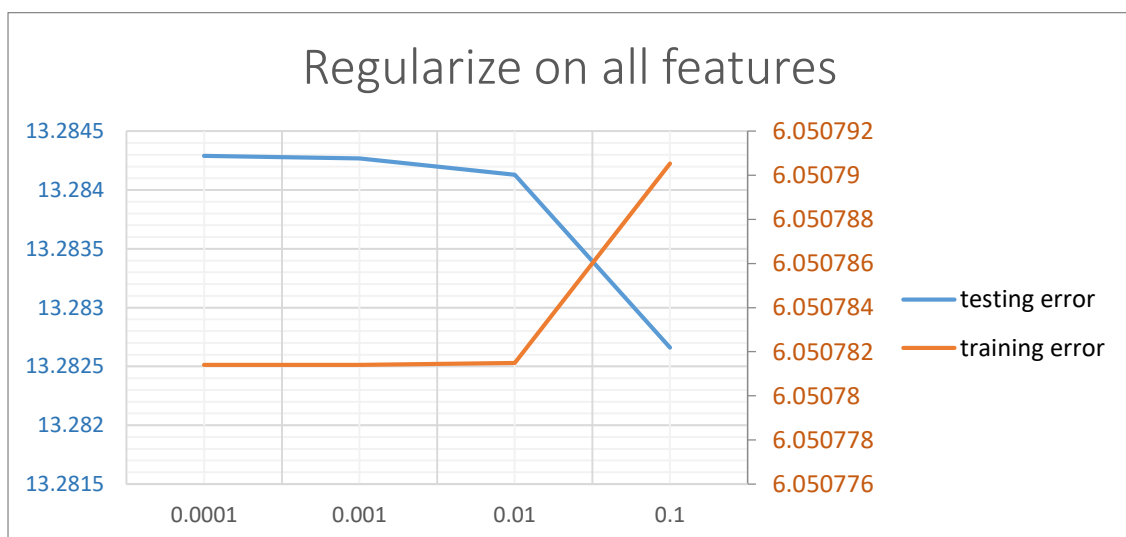
全部 feature: 從 9 小時改成 5 小時後，總分從 13.2843 變成 13.08004。總分有滿大的進步，很可能因為 9 小時加上所有 feature 的 model 太過複雜，改成 5 小時之後 model 的複雜度下降，減輕 training data 的 overfitting，得到比較好的結果。

僅用 pm2.5: 從 9 小時改成 5 小時後，public 的分數從 13.23238 變成 13.38444，這邊則變差了一點，不同於全部 feature 的情況。可以想到的原因是僅使用 pm2.5 已經是一個很簡單的模型了，再從 9 小時改成 5 小時讓 model 太簡單了，function set 太小，結果因此變差。

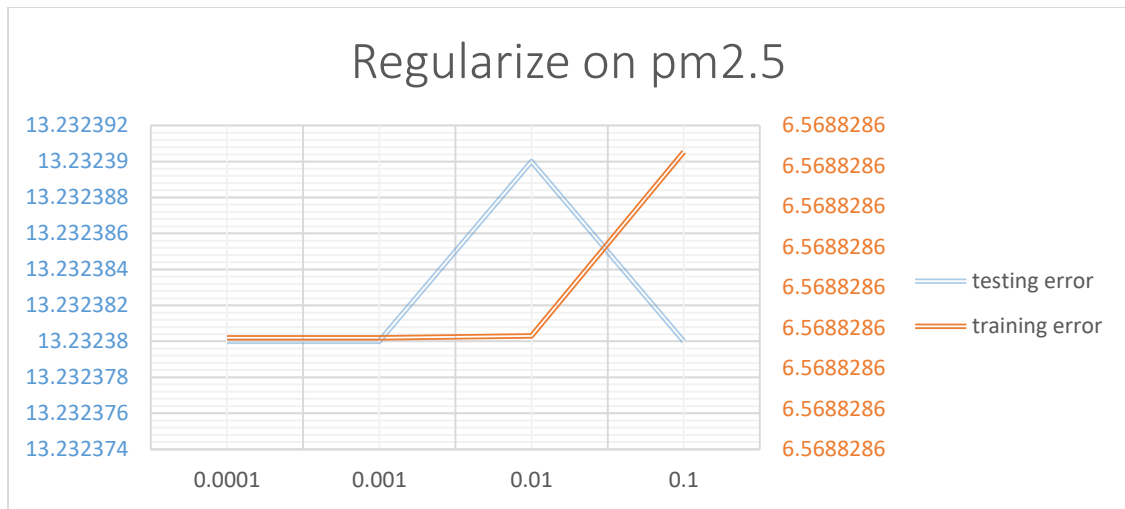
3. (1%)Regularization on all the weight with $\lambda=0.1$ 、 0.01 、 0.001 、 0.0001 ，並作圖

ANS:

All features:



Pm 2.5 only:



上兩圖中，testing 跟前面一樣使用 public+private 的分數，training 則是用自己切的 validation 的 error。使用全部 feature 時，當 λ 愈大，testing error 逐漸減小同時 training error 逐漸增加。Training error 增加如同預期，因為當 regularize 佔的比例愈大，loss 就愈被忽略，所以 error 就逐漸增加。而 testing error 在此則是逐漸下降，可能因為 model 過於複雜，讓 function 平滑反而比較好。理論上當 λ 過大時，又會讓 Loss function 太過傾向 regularize 那一項而讓整體 error 上升，但在此處沒有觀察到這種結果。至於僅用 pm2.5 的情況，則是幾乎沒有變化，error 在非常小的區間變動，結果 testing error 沒有如預期般變動，如果 λ 調大一點應該才能比較看得出趨勢。

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 $\sum_{n=1}^N (y^n - x^n * w)^2$ 。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:

將 Loss function 寫成矩陣表達，即可寫成 $L(w) = \|y - Xw\|_2^2$

將其對 w 做微分，設結果為 0，由於 function 是 convex，微分 0 的地方即是最低點:

$$\frac{\partial L(w)}{\partial w} = \frac{\partial \|y - Xw\|_2^2}{\partial w} = 0 \rightarrow -2X^T(y - Xw) = 0 \rightarrow X^T y - X^T X w = 0 \rightarrow w = (X^T X)^{-1} X^T y$$

因此答案為(c)