

ML2017FALL HW1 Report

學號：R05945012 系級：生醫電資碩二 姓名：張凱歲

- Method: Adam
- Parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\eta = 1e-4$, $\max_iteration = 1e6$

1. (2%)記錄誤差值 (RMSE)，討論兩種 feature 的影響

Model	All pollutant, 9hr, 1 st order	Only pm2.5, 9hr, 1 st order,
Training score	5.679520	6.123022
Testing score	6.569768	6.596244

運用較多污染物去訓練的模型必定會有較好的 training score，但 testing score 則不一定有這樣的關係，因為做 feature selection 有機會得到更好的結果。而根據以上表格我們可以看到，包含所有 18 種污染物的模型仍然有較好的 testing score，應該是因為若只用 pm2.5 做預測，feature 會太單一的關係。實際上我也測試了其他污染物的組合，發現使用 pm2.5, pm10, O₃, CO 的組合可以得到最好的結果。

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

Model	All pollutant, 5hr, 1 st order	Only pm2.5, 5hr, 1 st order
Training score	5.805064	6.207004
Testing score	6.643333	6.744917

如同問題一，在相同小時數下，所有污染物的模型得到的分數(包含 training score 和 testing score)比只用 pm2.5 的模型還要好。另外，比較 9hr 與 5hr，我們也可以發現 9hr 的各項分數都比 5hr 還要好，代表 9hr 的模型更能夠準確預測 pm2.5 的數值。

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

All pollutant, 9hr, 1st order:

Model	$\lambda=0$	$\lambda=0.0001$	$\lambda=0.001$	$\lambda=0.01$	$\lambda=0.1$
Training score	5.679520	5.679521	5.679522	5.679540	5.679714
Testing score	6.569768	6.570187	6.569766	6.569715	6.569237

因 λ 小，而導致分數的變化不大。因此我額外再多做了幾種 λ 數值。

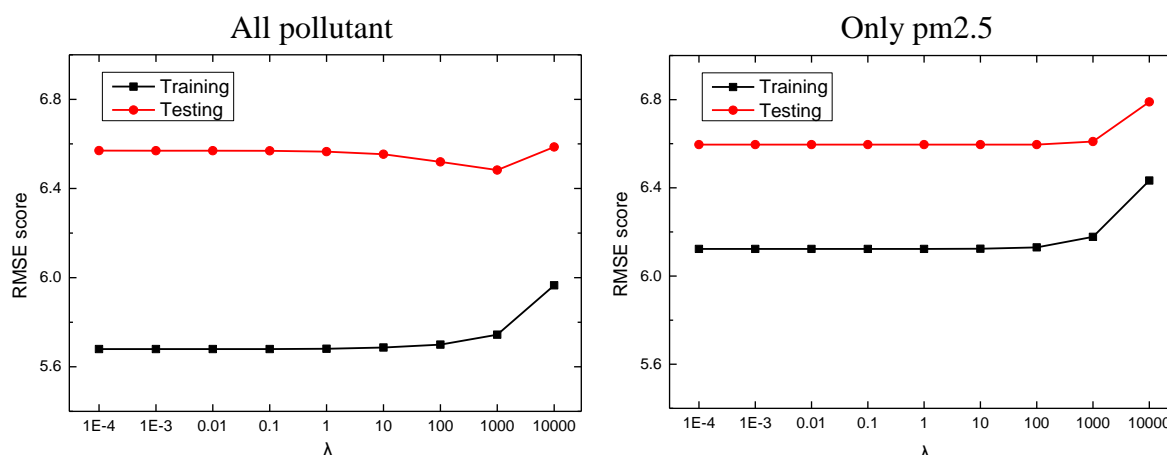
Model	$\lambda=1$	$\lambda=10$	$\lambda=100$	$\lambda=1000$	$\lambda=10000$
Training score	5.681114	5.686518	5.699580	5.744231	5.965416
Testing score	6.565624	6.553855	6.519410	6.482602	6.586142

Only pm2.5, 9hr, 1st order:

Model	$\lambda=0$	$\lambda=0.0001$	$\lambda=0.001$	$\lambda=0.01$	$\lambda=0.1$
Training score	6.123022	6.123022	6.123022	6.123022	6.123029
Testing score	6.596244	6.596244	6.596244	6.596244	6.596244

同樣因變化不大，加上更多種 λ 後作圖。

Model	$\lambda=1$	$\lambda=10$	$\lambda=100$	$\lambda=1000$	$\lambda=10000$
Training score	6.123093	6.123730	6.129891	6.177838	6.432964
Testing score	6.596239	6.596198	6.596096	6.609867	6.789559



對於訓練所有污染物的模型，隨著 λ 提高，我們可以看到 training score 逐漸升高，但 testing score 卻有逐漸下降的趨勢，直到 $\lambda=1000$ 時到達最小值。只用 pm2.5 的模型也有相似的趨勢，雖然 testing score 的變化非常小，但也會逐漸下降，並在 $\lambda=100$ 時達到最小值。

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。(其中 $X^T X$ 為 invertible)

- (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 為正確答案。推導如下：

$$\begin{aligned}
 L &= \sum_{n=1}^N (y^n - w \cdot x^n)^2 = \sum_{n=1}^N (w \cdot x^n - y^n)^2 = (Xw - Y)^T (Xw - Y) \\
 L(w) &= \min(L), \text{ when } \frac{\partial L}{\partial w} = \frac{\partial}{\partial w} ((Xw - Y)^T (Xw - Y)) = 0 \\
 &\Rightarrow \frac{\partial}{\partial w} (w^T X^T X w - w^T X^T Y - Y^T X w + Y^T Y) = 0 \\
 &\Rightarrow X^T X w - X^T Y = 0 \\
 &\Rightarrow w = (X^T X)^{-1} X^T Y
 \end{aligned}$$