

Machine Learning Homework Report

PM2.5 Prediction

學號：r05922145 系級：資工碩二 姓名：郁錦濤

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)

備註：

a. NR 請皆設為 0，其他的數值不要做任何更動

b. 所有 advanced 的 gradient descent 技術(如: adam, adagrad 等) 都是可以用的

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

model	training data RMSE	kaggle public RMSE	kaggle private RMSE
(1)	5.698564	7.53453	5.39402
(2)	6.226667	7.85510	5.87430

解答：

模型 (1) 抽取全部 9 小時內的污染源 feature 的一次項 (加 bias)，共有 $12 * (480 - 9) = 5652$

筆測資，每筆測資則有 $18 * 9 = 162$ 個 feature。

模型 (2) 抽取全部 9 小時內的 pm2.5 的一次項 (加 bias)，共有 $12 * (480 - 9) = 5652$ 筆測資，

每筆測資則有 9 個 feature。

模型（1）抽取的 feature 比較多，所以在 RMSE 表現上好於模型（2），但是計算速度明顯慢於模型（2）。相反，模型（2）抽取的 feature 比較少，雖然計算速度明顯快於模型（1），但是造成的結果是 RMSE 的表現上比模型（1）差很多。所以，選取合適的 feature，保證 feature 的數目，這樣的 performance 會比較好一點，計算速度也比較快。

根據兩種模型的對比，發現選取特徵較多的模型一般比選取模型較少的 performance 更好。這兩種模型都只考慮了 feature 的一次方，所以在 public set 和 private set 上表現都不太好。我的模型考慮了相關 feature 的一次方和二次方，效果比以上兩種模型都要好很多。

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

model	(1)Training RMSE	(1) Public RMSE	(1) Private RMSE	(2)Training RMSE	(2)Public RMSE	(2)Private RMSE
前 9 小時	5.698564	7.53453	5.39402	6.226667	7.85510	5.87430
前 5 小時	5.823100	7.71000	5.39192	6.340770	8.08842	6.03934

解答：

在模型（1）中，將 feature 從抽前 9 小時改成抽前 5 小時，這樣每個測資的 feature 數會由原來的 162 個減少到 90 個。這種變化會使 public set RMSE 會有變化，但是幅度不是很大。在抽取前 9 小時和前 5 小時的全部 feature 的兩種模型中，抽取前 5 小時全部 feature 模型的 public set 的 RMSE 雖然沒有前 9 個小時的模型好，但是在 private set 的 RMSE 則好於抽取前 9 個小時的模型。

在模型（2）中，將 feature 從抽前 9 小時改成抽前 5 小時，這樣每個測資的 feature 數會由原來的 9 個減少到 5 個，這樣造成的結果是 public set 和 private set 的 RMSE 會變大，也就是 performance 不好。

模型（1）在抽前 9 個小時和抽前 5 個小時的 feature 最後得到 performance 差別不是很大，因為還是有大部分的 feature 進行了保留。而模型（2）在抽前 9 個小時和抽前 5 個小時的 feature 最後得到的 performance 差別挺大，因為選取的 feature 太少。

通過比較說明，在某些情況下，選取特定的 feature 會有更好的 performance，每個 feature 的權重是不一樣的。feature 的數量大小也會直接影響到 performance。

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

解答：

My Model:

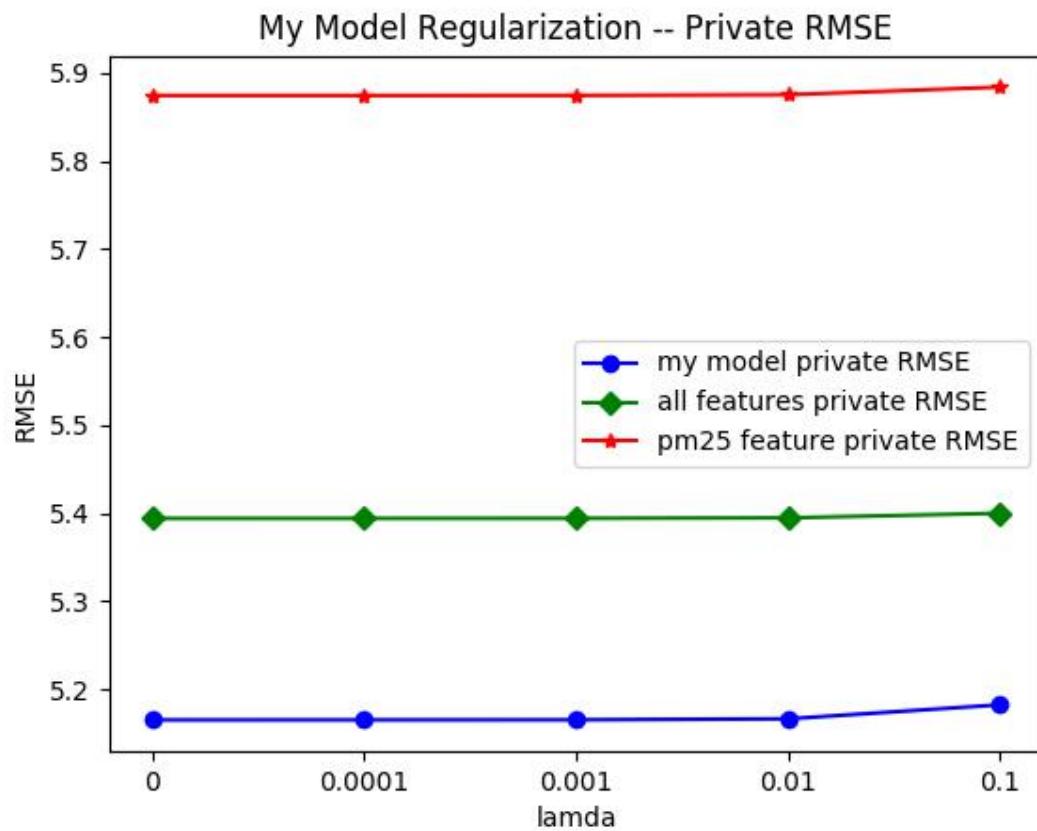
$\lambda =$	0.1	0.01	0.001	0.0001	0
Train RMSE:	5.687335	5.682595	5.682388	5.682372	5.682370
Public RMSE:	6.50708	6.49006	6.48836	6.48820	6.48818
Private RMSE:	5.18226	5.16634	5.16522	5.16512	5.16510

Model(1):

$\lambda =$	0.1	0.01	0.001	0.0001	0
Train RMSE:	5.701145	5.698791	5.698584	5.698564	5.698562
Public RMSE:	7.55890	7.53694	7.53475	7.53453	7.53450
Private RMSE:	5.39984	5.39458	5.39407	5.39402	5.39401

Model(2):

$\lambda =$	0.1	0.01	0.001	0.0001	0
Train RMSE:	6.230814	6.227074	6.226704	6.226667	6.226663
Public RMSE:	7.87279	7.85685	7.85526	7.85510	7.85508
Private RMSE:	5.88383	5.87516	5.87437	5.87430	5.87429



從測資中可以看出， λ 對於模型(1)和(2)的兩種 feature 選取的一次方的模型最後 public RMSE 和 private RMSE 結果影響不大。隨著 λ 變小，RMSE 也隨之變小，但是幅度很小，差距不是很大。

對於我的二次模型，當 lamda 取 0.1 和 0.01 時，private RMSE 會從 5.18226 降至 5.16634。

分析如下。由於 regularization 所添加的是 $\text{lamda}^* \mathbf{w}^2$ ，這個對於選取 feature 的一次方的模型 performance 是不明顯的。但是對於有選取 feature 的二次方的模型是有幫助的，能夠降低 RMSE，當 lamda 越小時，RMSE 表現更好。

我認為 regularization 主要是增加在取過多的參數和過高的 weight 時的 penalty，當我取過多的參數時添加 regularization 就會有效果，當取過少的參數 regularization 效果會不明顯。

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 \mathbf{x}^n ，其標註 (label) 為一存量 y^n ，模型參數為一向量 \mathbf{w} (此處忽略偏權值 b)，則線性回歸的損失函數 (loss function) 為 $\frac{1}{2} \sum_{n=1}^N (\mathbf{y}^n - \mathbf{X}\mathbf{w})^2$ 。若將所有訓練資料的特徵值以矩陣 $\mathbf{X} = [\mathbf{x}^1 \ \mathbf{x}^2 \ \dots \ \mathbf{x}^N]^T$ 表示，所有訓練資料的標註以向量 $\mathbf{y} = [y^1 \ y^2 \ \dots \ y^N]^T$ 表示，請問如何以 \mathbf{X} 和 \mathbf{y} 表示可以最小化損失函數的向量 \mathbf{w} ？請寫下算式並選出正確答案。(其中 $\mathbf{X}^T\mathbf{X}$ 為 invertible)

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

解答：選擇(c)。

理由：

在取出特徵後，我們的預測結果為 $\mathbf{X}\mathbf{w}$ ，因此我們想找到 \mathbf{w} 使 $(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})$ 最小(也就是讓我們定義的 loss function 最小)，將 loss function 對 \mathbf{w} 做偏微分後得到 $-2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w})$ 。令其等於 0，得到 $\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ 。