學號：B04901146　系級：　電機三　姓名：黃禹傑

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

	Private RMSE	Public RMSE	Test RMSE
全部污染源	5.53562	7.46237	6.570009696
PM2.5	5.62719	7.44013	6.596241419

兩者在 public 以及 private 各有長處，但總體而言，以全部 9 小時的污染源作為考慮的結果稍稍優於只考慮 pm2.5 的結果，可能是因為 18 個中只取 pm2.5，並且又只有一次項的模型無法足夠精準描述其分佈，而 18 個污染源雖然可能其中有些較不需要考慮，但在這邊仍然可以比較精準的預測出結果，包括在 training error 的部分其實取全部參數的表現也最是比較佳的。

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

	Private RMSE	Public RMSE	Test RMSE
全部污染源	5.44092	7.65925	6.643332033
PM2.5	7.57904	5.79187	6.744909392

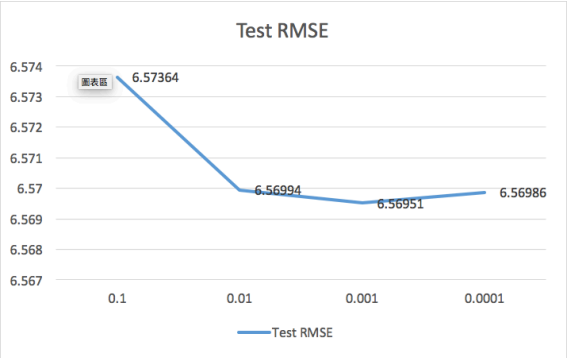
相較於取 9 小時，error 都上升了，而考慮全部污染源的結果仍然是優於只考慮 pm2.5 的結果。

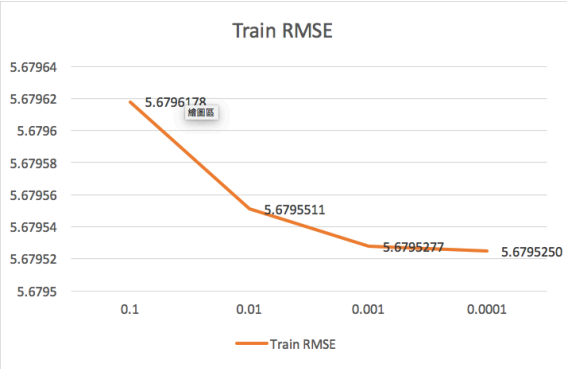
3. (1%)Regularization on all the weight with λ=0.1、0.01、0.001、0.0001，並作圖

(橫軸為 lamda，縱軸為 RMSE)

(1) 抽全部 9 小時內的污染源 feature 的一次項(加 bias)

lamda	private	public	Test RMSE	Train RMSE
0.1	5.53614	7.46837	6.57364	5.6796178
0.01	5.53534	7.46246	6.56994	5.6795511
0.001	5.53455	7.46229	6.56951	5.6795277
0.0001	5.53476	7.46275	6.56986	5.6795250

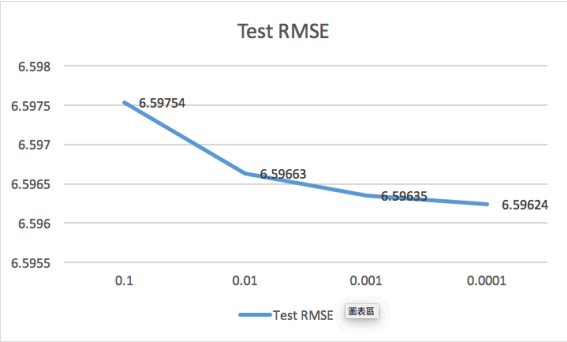


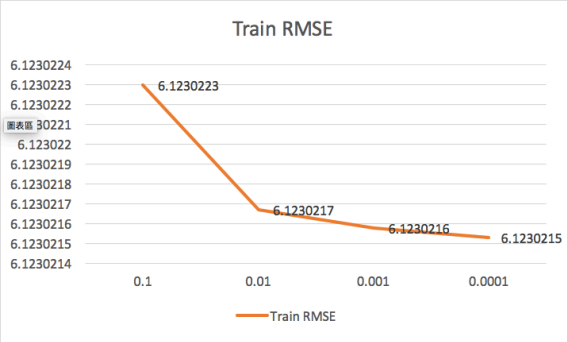


在 train 時 lamda 越小表現似乎越好，但 test 的結果顯示卻為 lamda=0.001 時的表現較佳。

(2) 抽全部 9 小時內 pm2.5 的一次項當作 feature(加 bias)

lamda	private	public	Test RMSE	Train RMSE
0.1	5.62768	7.44206	6.59754	6.1230223
0.01	5.62753	7.44056	6.59663	6.1230217
0.001	5.62739	7.44017	6.59635	6.1230216
0.0001	5.62725	7.44009	6.59624	6.1230215





在只考慮 pm2.5 的情況下，在 train 以及 test 的結果中，都是 lamda 越小表現越好。

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$X^TX$為 invertible)

(a) $(X^TX)X^Ty$

(b) $(X^TX)^{-0}X^Ty$

(c) $(X^TX)^{-1}X^Ty$

(d) $(X^TX)^{-2}X^Ty$

Answer : (c)

Derivation :

$Let\ L(w) be\ the\ loss\ function, that\ is,$

$$L(w) = \sum_{n=1}^{N}(y^n - \sum_{m=1}^{M} x_m^n \cdot w_m)^2$$

$to\ obtain\ \hat{w} = argmin(L(w))$

$$\frac{\partial L}{\partial w_k} = \sum_{n=1}^{N}(y^n - \sum_{m=1}^{M} x_m^n \cdot w_m)(-2)(x_k^n) = 0, \forall k = 1, 2, \dots, M$$

$rearrange\ the\ summations$

$$\sum_{n=1}^{N}\sum_{m=1}^{M} x_k^n \cdot x_m^n \cdot \hat{w}_m = \sum_{n=1}^{N} x_k^n \cdot y^n, \forall k = 1, 2, \dots, M$$

$leading\ to$

$$(X^TX)\hat{w} = X^Ty$$

$equivalently,$

$$\hat{w} = ((X^TX)^{-1}X^Ty, \quad Q.E.D.$$