學號：r06921011 系級： 電機碩一 姓名：劉逸霖

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

	public	private	Public+private
(1) all feature	7.48161	5.46557	6.551601
(2) pm2.5	7.36405	5.61963	6.550171

用全部的 feature 預測出來的分數跟 pm2.5 差不多，可以算是在誤差範圍內，觀察不出甚麼現象，不予討論。我覺得只用 pm2.52 的結果應該必較差，但是因為只有一階，所以效果不明顯。

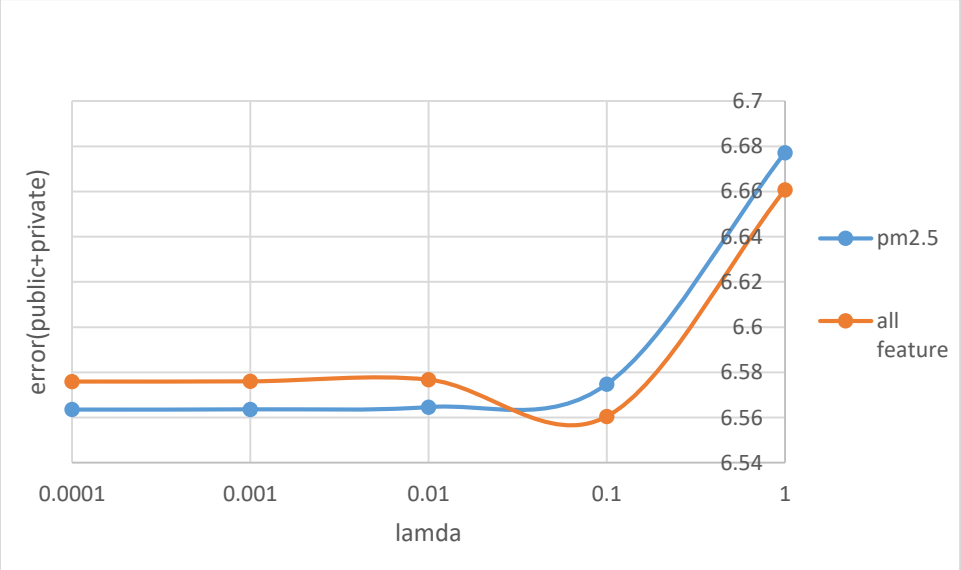
2. (1%)將 feature 從抽前 9 小時改成抽前 5 小時，討論其變化

	public	private	Public+private
(1) all feature_5hours	7.60120	5.46750	6.62087
(2) pm2.5_5hours	7.59423	5.86378	6.78440

改成前 5 小時，用全部 feature 的效果比較好。跟第一小題比起來，誤差變大了，因為可以用的 feature 變少了，會降低預測的準確性。

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

	λ=1	λ=0.1	λ=0.01	λ=0.001	λ=0.0001
(1) all feature	6.660709	6.56044	6.576726	6.575962	6.575887
(2) pm2.5	6.677138	6.574684	6.564588	6.563581	6.563483



Lamda 越大，會使預測出來的 model 變越平滑，能夠降低 overfitting 的情況，但依照我的程式而言，還沒有到達 overfitting，所以 lamda 會使誤差增加，但影響不明顯。從圖出會發現 pm2.5 的誤差不會一直大於全部 feature，但所有資料的差距都非常小，視為隨機誤差。

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 $X = [x^1\ x^2 \dots x^N]^T$ 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)

$$y = Xw$$

因為 X 不是方陣，無法直接算反矩陣，因此:

$$w = pinv(X)y \ ,\ pinv(X) = (X^TX)^{-1}X^T$$

Proof:

$$\text{Loss function: } L = \sum [y - Xw]^2 = (y - Xw)^T(y - Xw)$$

$$\frac{\partial}{\partial w} L = \frac{\partial}{\partial w}(y^Ty - y^TXw - w^TX^Ty + w^TX^TXw) = 2(X^TXw - X^Ty)$$

$$\text{Let } \frac{\partial}{\partial w} L = 0$$

$$w = (X^TX)^{-1}X^Ty$$

(a) $(X^TX)X^Ty$
(b) $(X^TX)^{-0}X^Ty$
(c) $(X^TX)^{-1}X^Ty$
(d) $(X^TX)^{-2}X^Ty$