學號：R06943153 系級： 電子碩一 姓名：蘇旻彥

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

	kaggle public score	kaggle private score
(1) All feature	9.16839	8.80755
(2) Only pm2.5	6.32403	6.27491

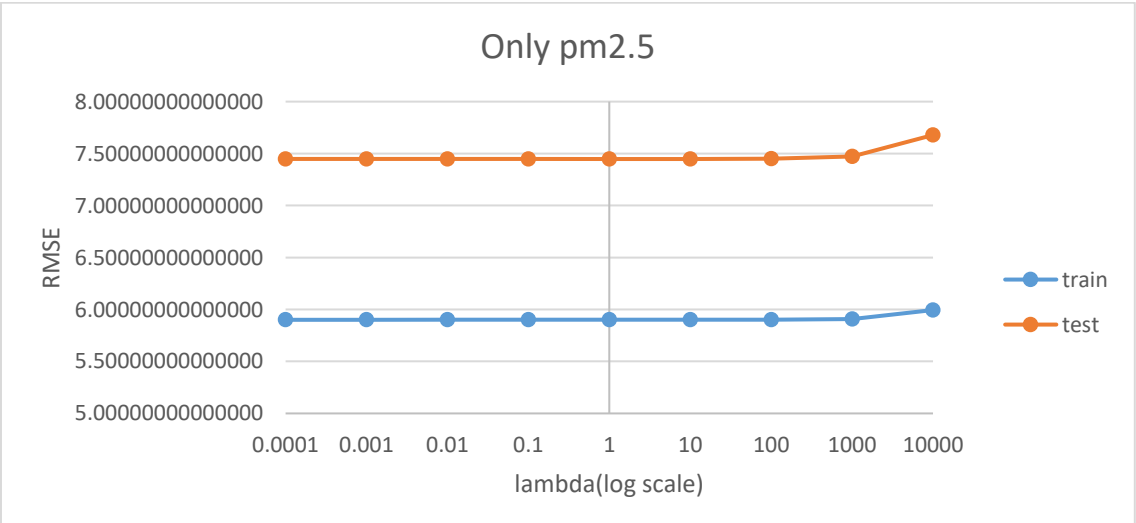
由以上表格可以看到只有抽取 pm2.5 的資料，訓練結果會比抽取所有汙染物還要好。我猜測是因為所有 18 種汙染物中有很是跟 pm2.5 沒有直接關係，造成訓練的結果 overfitting training data。若能找到哪些汙染物和 pm2.5 有直接的影響，只使用這些 feature 做訓練，有可能可以使訓練結果更好。

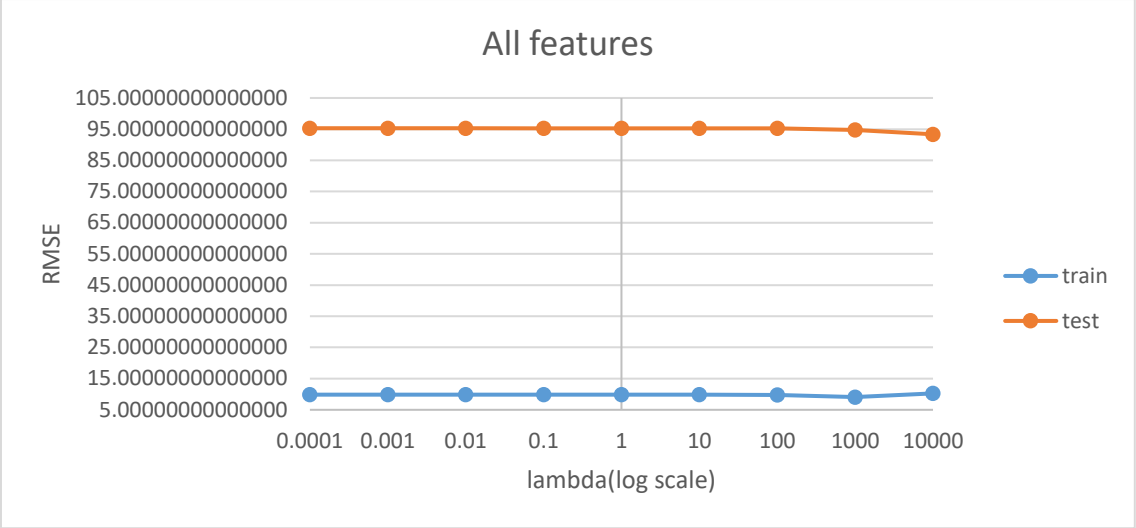
2. (1%)將 feature 從抽前 9 小時改成抽前 5 小時，討論其變化

	kaggle public score	
	9 hrs	5 hrs
(1) All feature	7.96306	8.23130
(2) Only pm2.5	7.29604	129.37119

由於用的 model 不同，所以重新做了一次抽取前小時的結果。如表所示，和第一題一樣，在抽取前九小時的狀況下，只抽取 pm2.5 的成績比抽取所有汙染物還要好；但是在抽取前五小時的狀況下，反而是只有抽取 pm2.5 的成績變差很多。我猜測是因為抽取的 feature 太少導致學習的狀況變差。

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





(以上圖表中的 test RMSE 是取自 Kaggle 上的 Public 分數)

由圖表可以看出在這裡用了 regularization，不論 lambda 多少都沒有影響，個人猜測是因為使用的數據當中 scale 相差甚多，沒有先做 normalization 的關係。

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

解: (C)

目標是要達到平方誤差的最小值：

$$\min \sum_{n=1}^{N}(y^n - wx^n)^2$$

我們先把以上式子換成矩陣運算之後定義 cost function 如下：

$$\frac{1}{2}(\mathrm{wX} - \mathrm{y})^T(wX - y)$$

要找到一個 w 可以使 cost function 達到最小值，我們用 cost function 對 w 進行偏微分：

$$\frac{\partial}{\partial w}\left(\frac{1}{2}(\mathrm{wX}-\mathrm{y})^T(\mathrm{wX}-\mathrm{y})\right)$$

$$=\frac{1}{2}\times\frac{\partial}{\partial w}(w^TX^TXw-w^TX^Ty-y^TXw+y^Ty)$$

$$=X^TXw-X^Ty$$

為了找到達到最小值時的 w，我們讓偏微分的結果等於 0:

$$X^TXw-X^Ty=0$$

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

得證，故解答為(C)