學號：B04902099 系級： 資工三 姓名：黃嵩仁

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

	Public score	Private score	Loss
(1)所有汙染源	7.59856	5.40484	6.59357
(2)pm2.5	7.68747	5.77560	6.79907

註：(1) $\text{loss} = \sqrt{\frac{public^2+private^2}{2}}$, 以下同

(2)訓練停止條件：training loss 兩次之間誤差小於 1e-6，以下同

由上表可知，只取 pm2.5 來 training 的 loss 會比取所有汙染物來 train 的高，***表示在 18 種汙染物中，有其他的汙染物也會影響第 10 小時的 pm2.5 數值(結果)***，須列入考慮。但是，有些修課同學所測出得 Loss 與我所得到的結果相反，我覺得有可能在 18 種汙染物中，***有些汙染物會對第 10 小時 pm2.5 的預測結果產生嚴重誤差***，而他們得訓練方法(or 停止條件)與我不同，導致所訓練出的 model 對於那些"不好的"汙染物有明顯的反應，故得到該結果。

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

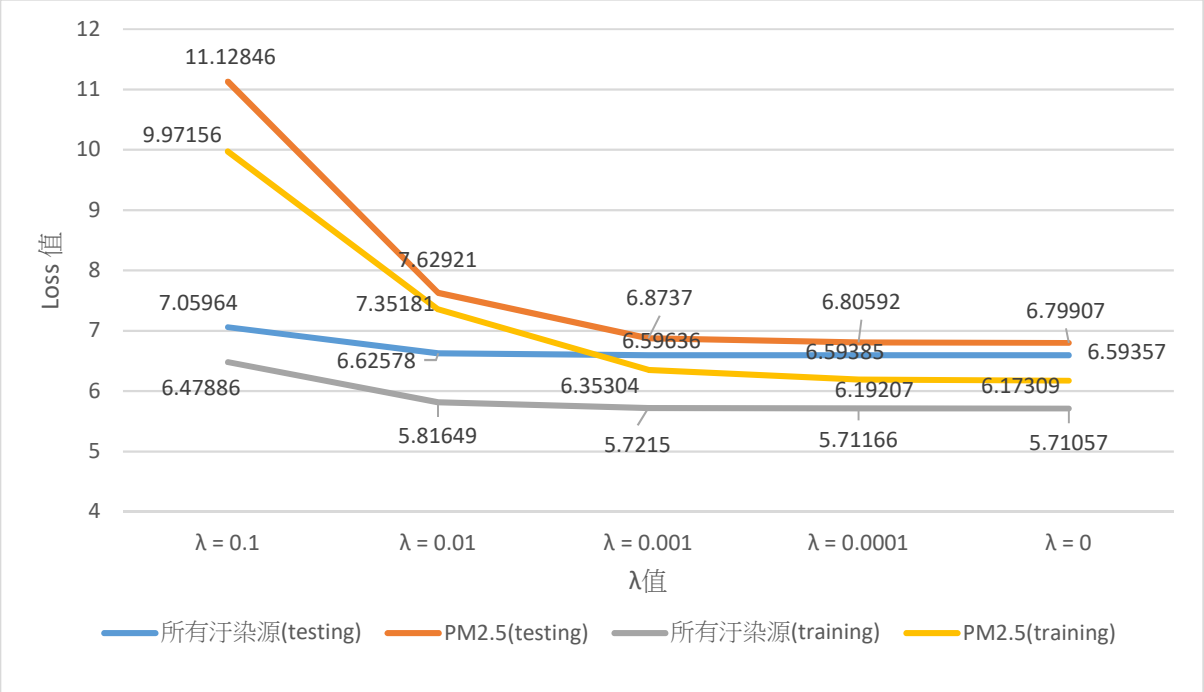
	Public score	Private score	Loss
所有汙染源前 9 小時	7.59856	5.40484	6.59357
所有汙染源前 5 小時	7.81637	5.39328	6.71502
Pm2.5 前 9 小時	7.68747	5.77560	6.79907
Pm2.5 前 5 小時	7.88416	5.92986	6.97578

由上表可知，不論是取所有汙染物或是只取 pm2.5 來做 training，抽取前 9 小時所得的成績都比僅抽取前 5 小時來得好，可以推論***頭 4 個小時得汙染物對於預測結果得正確率有一定程度的影響***。

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

(1)	Public score	Private score	Loss(testing)	Loss(training)
所有汙染源, λ = 0.1	8.24109	5.63575	7.05964	6.47886
所有汙染源, λ = 0.01	7.64711	5.41515	6.62578	5.81649
所有汙染源, λ = 0.001	7.60284	5.40563	6.59636	5.72150
所有汙染源, λ = 0.0001	7.59899	5.40492	6.59385	5.71166
所有汙染源, λ = 0	7.59856	5.40484	6.59357	5.71057
(2)	Public score	Private score	Loss	Loss(training)

pm2.5, λ = 0.1	13.09160	8.73473	11.12846	9.97156
pm2.5, λ = 0.01	8.71389	6.36224	7.62921	7.35181
pm2.5, λ = 0.001	7.77711	5.83199	6.87370	6.35304
pm2.5, λ = 0.0001	7.69568	5.78081	6.80592	6.19207
pm2.5, λ = 0	7.68747	5.77560	6.79907	6.17309



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)為一純量 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) $(X^TX)^{-2}X^Ty$

Ans：

$$\sum_{n=1}^{N}(y^n - x^n \cdot w)^2 = \sum_{n=1}^{N}(x^n \cdot w - y^n)^2 = (w \cdot x^1 - y^1 \dots \dots w \cdot x^N - y^N)\begin{pmatrix} w \cdot x^1 - y^1 \\ \dots \\ w \cdot x^N - y^N \end{pmatrix}$$

$$= (w \cdot X - Y)^T(w \cdot x - Y)$$

$$\frac{\partial}{\partial w}(w \cdot X - Y)^T(w \cdot x - Y) = \frac{\partial}{\partial w}(w^TX^TXw - w^TX^TY - Y^TXw + Y^TY) = X^TXw - 2X^TY$$

(上式結果由$\frac{\partial}{\partial M}M^TM = 2M$ 及 $\frac{\partial}{\partial M}AM = A^T$)

令$X^TXw - 2X^TY = 0$，得到$w = (X^TX)^{-1}X^TY$，故答案為(c)