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

	所有汙染物 features	pm2.5
RMSE	7.381871	6.245158

我覺得因為 pm2.5 對預測 pm2.5 的影響是比較直接的，但是當將所有的污染物 features 都納入時，因為有部分的汙染物 feature 是不會影響 pm2.5 的值，因此會導致 train 出來的 RMSE 會比較大。

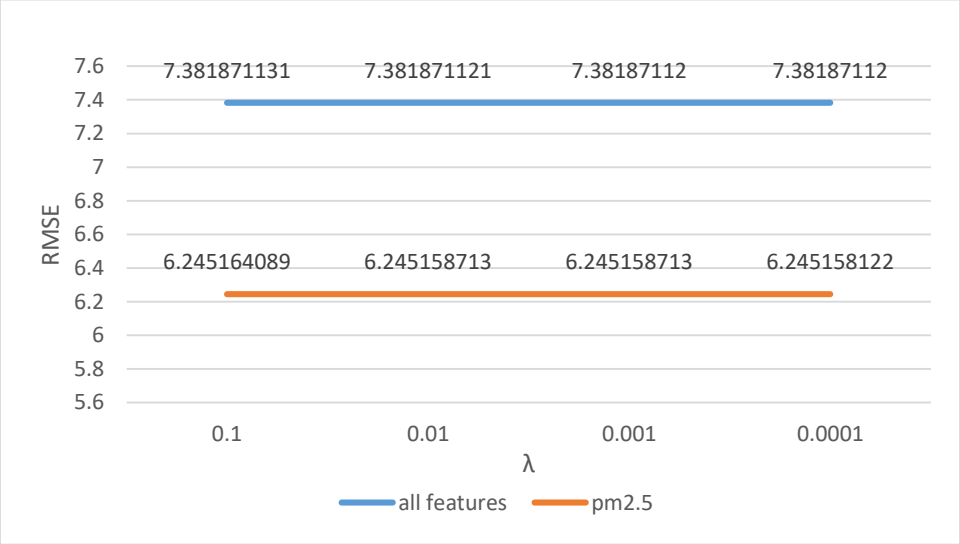
2. (1%)將 feature 從抽前 9 小時改成抽前 5 小時，討論其變化

	所有汙染物 features		pm2.5	
	5hr	9hr	5hr	9hr
RMSE	7.133596	7.381871	6.505429	6.245158

根據 RMSE，可以觀察到考慮所有汙染物 features 連續 5hr 會比連續 9hr 的 RMSE 還要小，但是相反的，在只考慮 pm2.5 的 case 中卻是連續 5hr 會比連續 9hr 的 RMSE 還要大。這可能的原因應該是 pm2.5 可能跟幾個小時內的 pm2.5 濃度比較相關，離越遠時間參考價值就越低。但是在考慮所有汙染物的時候可能裡面有些汙染物在長時間內都會影響 pm2.5 的濃度，所以時間減少預測出來的數值就比較差。

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

		0.1	0.01	0.001	0.0001
所有汙染物 features	RMSE	7.381871	7.381871	7.381871	7.381871
pm2.5		6.245164	6.245159	6.245159	6.245158



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

Sol:

$\text{Loss(w)} = (Xw - Y)^T(Xw - Y)$

$\nabla\text{Loss(w)} = \nabla((Xw - y)^T(Xw - y))$

$= \nabla((w^TX^T - y)^T(Xw - y))$

$= \nabla(w^TX^TXw - w^TX^Ty - y^TXw + y^Ty)$

$= \nabla(w^TX^TXw - 2w^TX^Ty + y^Ty)$

$= \nabla(w^TX^TXw) - \nabla(2w^TX^Ty) + \nabla(y^Ty)$

$= 2X^TXw - 2X^Ty$

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

故選 (c)