學號：R06922002 系級： 資工碩一 姓名：姚嘉昇

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

	Private Score	Public Score	Testing Set Score	Training Set Score
18 features, 9 hours	5.63635	8.01539	6.928741523	6.251824
1 feature, 9 hours	5.83648	7.39590	6.661975518	6.568829

取 18 個 features 雖然 private score 較好，但其 public score 與 testing set score 都比只取 1 個 feature 還要差。取 18 個 features 可能有 overfitting 的現象發生。

如果比對最後再把 model 丟回去 training set 去算它的 RMSE，會發現 18 個 features 的 training set score 只有 6.25，而 testing set score 卻是 6.82，很明顯是 overfitting。而只取 1 個 feature 的話雖然 training set 的 RMSE 較高一點，但其實這個 model 在 testing set 比較能夠 fitting。

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

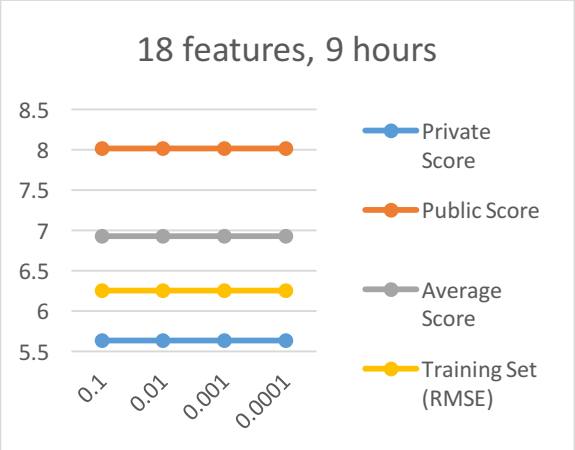
	Private Score	Public Score	Testing Set Score	Training Set Score
18 features, 5 hours	5.40360	7.71723	6.661626372	6.007679
1 feature, 5 hours	5.86272	7.52172	6.743432270	6.369811

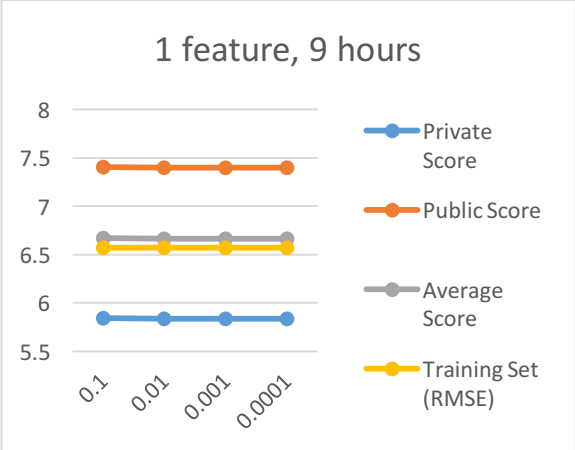
先看 18 features 從 9 小時改為 5 小時，他的 private、public、testing set、training set 分數都有進步，甚至比只取 1 個 feature 還要棒。再看 1 feature 從 9 小時改為 5 小時，他只有自己在 training set 上有 fitting，在 testing set 上是退步的。所以只取 1 個 feature 又只取 5 小時很明顯是維度不夠高，是 underfitting。

綜合以上，可發現 18 features, 5 hours (90 維)是最棒的；1 feature, 9 hours (9 維)次之；1 feature, 5 hours (5 維)則是第三 (underfitting)；18 features, 9 hours (162 維)最糟糕 (overfitting)。

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

	λ	Private Score	Public Score	Testing Set Score	Training Set Score
18 features 9 hours	0.1	5.63627	8.01304	6.927349766	6.255044
	0.01	5.63637	8.01492	6.928477805	6.252148
	0.001	5.63635	8.01532	6.928701034	6.251857
	0.0001	5.63635	8.01536	6.928724170	6.251828
1 features 9 hours	0.1	5.84309	7.40644	6.670721633	6.568908
	0.01	5.83714	7.39696	6.662853016	6.568829
	0.001	5.83655	7.39601	6.662067240	6.568829
	0.0001	5.83649	7.39591	6.661985449	6.568829

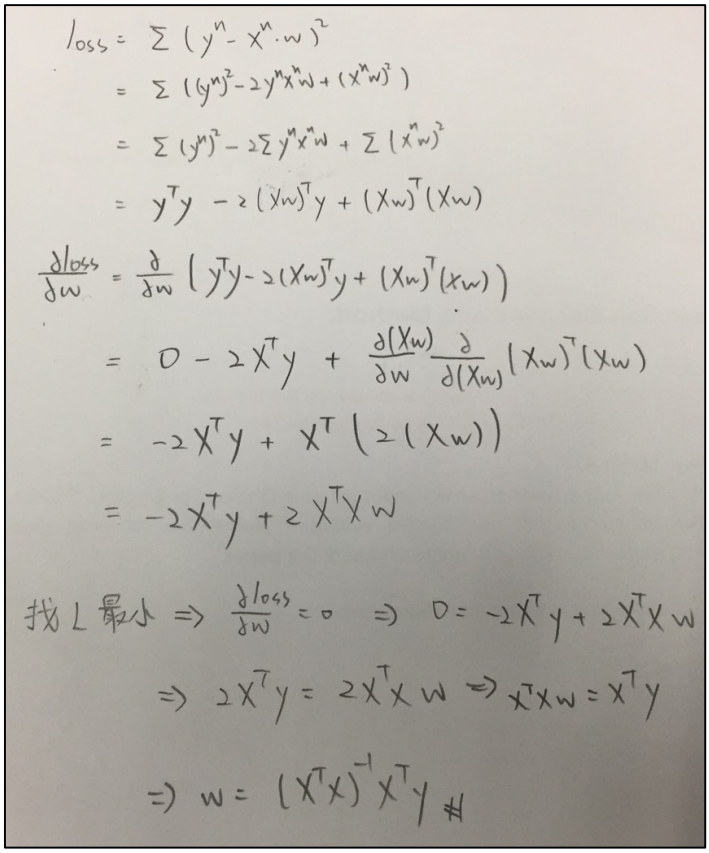




4. (1%)在線性回歸問題中，假設有 N 筆訓練資料，每筆訓練資料的特徵 (feature) 為一向量 x^n，其標註(label)為一存量 y^n，模型參數為一向量 w (此處忽略偏權值 b)，則線性回歸的損失函數(loss function)為$\sum_{n=1}^{N} (y^n - x^n \cdot w)^2$ 。若將所有訓練資料的特徵值以矩陣 $X = [x^1\ x^2 \dots x^N]^T$ 表示，所有訓練資料的標註以向量 $y = [y^1\ y^2 \dots y^N]^T$ 表示，請問如何以 X 和 y 表示可以最小化損失函數的向量 w ？請寫下算式並選出正確答案。(其中 X^TX 為 invertible)

答案：(c) $(X^TX)^{-1}X^Ty$

證明 1：



證明 2：

為了使損失函數最小化，必須讓 $y - Xw = 0$，則 $y = Xw$

兩邊同乘 X^T 後，則 $X^Ty = X^TXw$

兩邊同乘$(X^TX)^{-1}$ 後，則 $(X^TX)^{-1}X^Ty = (X^TX)^{-1} X^TXw$

右式整理後，可得 $w = (X^TX)^{-1}X^Ty$