學號：R06946003 系級： 資料科學碩一 姓名：湯忠憲

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

Note: Model 均使用 SGD 和 adagrad，並且每次使用 98%的資料 training，2%資料 validation；由於 PM2.5 中有-1，因此取該點前一個時間點的 PM2.5 值 impute。

1. (2%)記錄誤差值 (RMSE)(根據 kaggle public+private 分數)，討論兩種 feature 的影響

只用 PM2.5 的一次項 feature 可以到比使用全部汙染源 features 還要好的成績。從結果可以得知 PM2.5 的一次像提供了大多數的資訊，並且全部汙染源特徵中可能有些助益不大，並且會影響回歸的擬合。

	Kaggle total score
All features used	14.42426
only PM2.5 used	13.26779

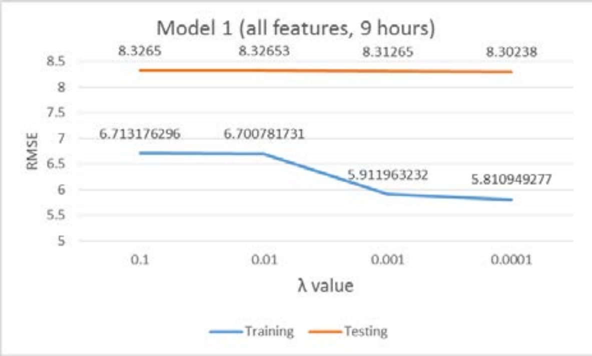
2. (1%)將 feature 從抽前 9 小時改成抽前 5 小時，討論其變化

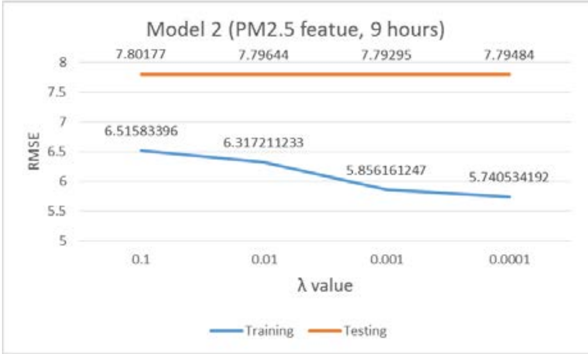
取前五小時的特徵後，在兩種不同 feature 的模型都得到了更低的 RMSE。因此篩選不必要的變數可以避免回歸過度擬合。而只用 PM2.5 的模型依然優於用上全部汙染源特徵的模型。

	Kaggle total score
All features used; 9 hours	14.42426
only PM2.5 used; 9 hours	13.26779
All feature; 5 hours	13.72514
PM2.5; 5 hours	13.10282

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

比較四組 regularization term 後發現各組的 testing RMSE 並沒有顯著的差異，不過 training RMSE 會隨著λ提高而上升，可發現 regularization term 造成的影響。兩種不同 feature 的模型中，只取 PM2.5 feature 者有依然有較低的 RMSE。





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 $X = [x^1\ x^2\ \ldots\ x^N]^T$ 表示，所有訓練資料的標註以向量 $y = [y^1\ y^2\ \ldots\ y^N]^T$ 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) $(X^TX)^{-2}X^Ty$

答案: (c)

Loss function = $\sum_{n=1}^{N}\left(y^n - x^n \cdot w\right)^2$

為求最小化損失函數，可讓損失函數對 w 微分，並令其等於零。

$\frac{\partial Loss\ function}{\partial w} = 2X^T\left(y - Xw\right) = 0$

$2X^TXw = 2X^Ty$

並且 X^TX 為 invertible

故 $w = (X^TX)^{-1}X^Ty$