學號：B03902093　系級：　資工四　姓名：張庭維

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	Private	Train Data Loss
All feature	8.98307	5.89224	6.35396309386
PM 2.5	7.68443	5.50123	6.32698089358

我先只取 PM2.5 的 data，用 learning rate = 0.0000003，Train 了 1000 iteration，得到對 Training data 的 RMSE 約為 6.32，再來我取全部 feature 並 train 至 Training data 的 loss 跟 6.32 差不多。

並比較兩者上傳 kaggle 後的表現如上表，可發現兩者 public 的誤差差了約 1.2，推測是全部的 feature 干擾過多，導致結果表現較差

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

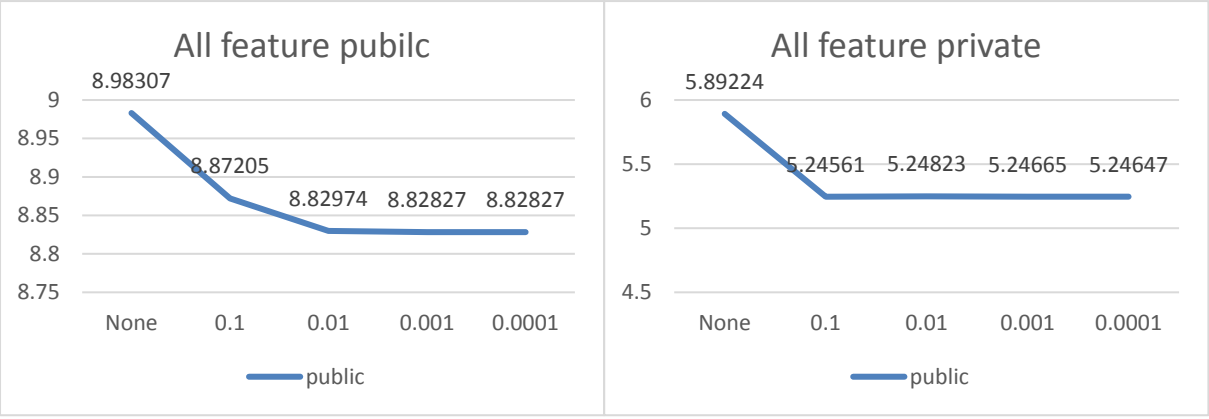
	Public	Private	Train Data Loss
All feature (9hr)	8.98307	5.89224	6.35396309386
All feature (5hr)	8.90333	5.79146	6.32698089358
PM 2.5 (9hr)	7.68443	5.50123	6.62705009183
PM 2.5 (5hr)	7.83516	5.66189	6.2489979221

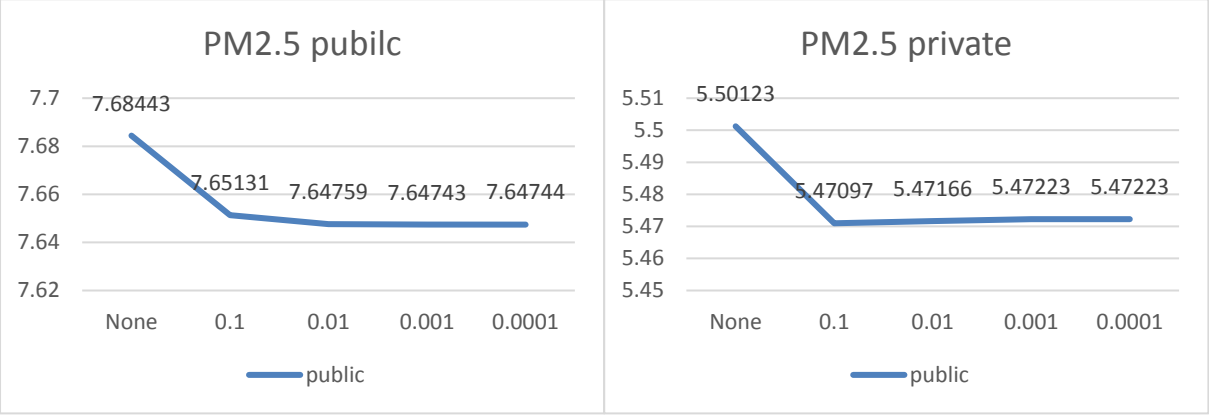
可看出只取 PM2.5 時，若只取 5 個小時的 data，一樣用 learning rate = 0.0000003，Train 了 1000 iteration，會出現了 Train data 的 loss 雖然較 9 個小時的低，但上傳 kaggle 後，表現的結果卻更差的狀況。我判斷是當指取 5 個 feature 時，會出現 underfitting 的現象，導致雖然 training 的 RMSE 略低，但 testing data 的 RMSE 卻更差的現象。

但當事與全部的 feature 時，只取五個小時的 data 卻會比取九個小時有更低的 public 跟 private 分數，可能是因為取全部 feature 時，五個小時 feature 有 90 項，九個小時則有 162 項，因此沒有出現 underfitting 的問題，但因為干擾減少，因此比全部 feature 取九個小時的表現好。

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

可發現取全部 feature 時，加入 regularization 會起到比較明顯的效果，而只取 PM2.5 時，變化不明顯甚至有時略差。





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

Ans. C

Loss function: $\sum_{n=1}^{N}(y^n - x^n \cdot w)^2 = (Y - X*W)^T(Y - X*W)$

$= (Y^T - W^TX^T)(Y - X*W)$

$= Y^TY - Y^TXW - W^TX^TY + W^TX^TX*W$

optimal: $\frac{d}{dw}\sum_{n=1}^{N}(y^n - x^n \cdot w)^2 = 0$

$$\frac{d}{dw}Y^TY - Y^TXW - W^TX^TY + W^TX^TX*W = -X^TY + X^TX*W = 0$$

$\underline{W = (X^TX)^{-1}X^TY}$