

學號：R07942115 系級：電信碩一 姓名：謝硯澤

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**當作一次項(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

c. 第1-3題請都以題目給訂的兩種**model**來回答

d. 同學可以先把**model**訓練好，**kaggle**死線之後便可以無限上傳。

e. 根據助教時間的公式表示，(1) 代表 $p = 9 \times 18 + 1$ 而(2) 代表 $p = 9 \times 1 + 1$

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

| | Pubile Score | Private Score |
|----------------|--------------|---------------|
| 前9小時的所有feature | 8.02776 | 9.03806 |
| 前9小時只有PM2.5 | 6.09551 | 7.04791 |

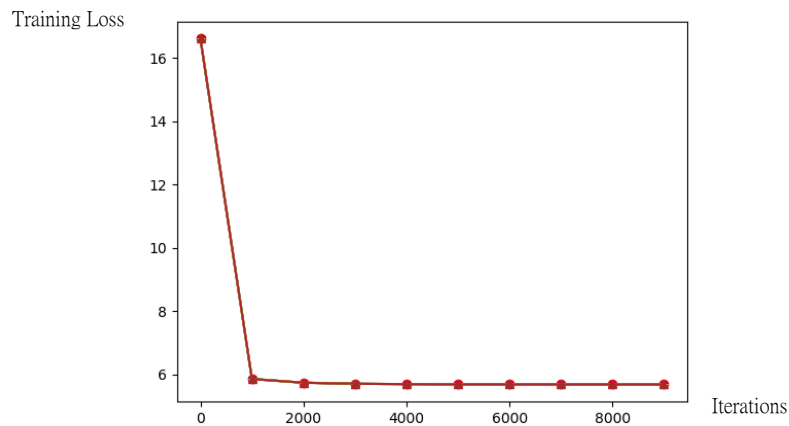
在我實作的程式中，只抽取**PM2.5**的模型，在**kaggle**上的分數是遠好於抽取全部**feature**的。但理論來說，如果**model**有**fit**的話，抽取全部的**feature**應該不會比只抽取**PM2.5**來得差。代表我的**model**可能還是**underfitting**的。

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

| | Pubile Score | Private Score |
|----------------|--------------|---------------|
| 前5小時的所有feature | 18.47992 | 18.90672 |
| 前5小時只有PM2.5 | 17.79629 | 19.89730 |

將**feature**從抽前9小時改成抽前5小時，只取**PM2.5**的模型，在**Kaggle**上的分數更低，可能的原因是**feature**數太少了($p=5+1$)，抽取全部**feature**的模型分數會稍微好一點，但兩種模型皆不理想。

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



| λ | AllFeature_public | AllFeature_private | PM2.5_public | PM2.5_private |
|-----------|-------------------|--------------------|--------------|---------------|
| 0.1 | 7.54748 | 8.68879 | 5.91274 | 7.2816 |
| 0.01 | 6.41488 | 10.8005 | 6.14957 | 7.41655 |
| 0.001 | 8.12723 | 8.36323 | 6.04844 | 7.44476 |
| 0.0001 | 7.23828 | 8.38745 | 5.73555 | 7.0238 |

加入Regularization後，表現並沒有提升，且各種 λ 畫出來的iterations-to-training loss的折線圖幾乎是重疊在一起的！推測可能是因為我的model並沒有overfitting，所以有沒有加入Regularization影響不大。

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 \mathbf{x}^n ，其標註(label)為一純量 y^n ，模型參數為一向量 \mathbf{w} (此處忽略偏權值 b)，則線性回歸的損失函數(loss function)為 $n = 1 \dots N$ $y^n - \mathbf{x}^n \cdot \mathbf{w}$ 。若將所有訓練資料的特徵值以矩陣 $\mathbf{X} = [\mathbf{x}^1 \mathbf{x}^2 \dots \mathbf{x}^N]^T$ 表示，所有訓練資料的標註以向量 $\mathbf{y} = [y^1 y^2 \dots y^N]^T$ 表示，請問如何以 \mathbf{X} 和 \mathbf{y} 表示可以最小化損失函數的向量 \mathbf{w} ？請選出正確答案。(其中 $\mathbf{X}^T \mathbf{X}$ 為invertible)

- (a) $(\mathbf{X}^T \mathbf{X}) \mathbf{X}^T \mathbf{y}$
- (b) $(\mathbf{X}^T \mathbf{X}) \mathbf{y} \mathbf{X}^T$
- (c) $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$
- (d) $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{y} \mathbf{X}^T$

Ans. (C)

- $loss = \sum (y - \mathbf{x} \mathbf{w})^2$
- $\frac{\partial loss}{\partial \mathbf{w}} = \sum 2(-\mathbf{x}^T)(y - \mathbf{x} \mathbf{w}) = 0$
- $\mathbf{x}^T \mathbf{x} \mathbf{w} = \mathbf{x}^T \mathbf{y}$
- $\mathbf{w} = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{y}$