

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

(1)抽全部 9 小時內的污染源 **feature** 的一次項(加 **bias**)：

public RMSE: 7.46631

private RMSE: 5.30105

$$\rightarrow \sqrt{\frac{7.46631^2 + 5.30105^2}{2}} \approx 6.47483$$

(2)抽全部 9 小時內 **pm2.5** 的一次項當作 **feature**(加 **bias**)：

public RMSE: 7.44013

private RMSE: 5.62719

$$\rightarrow \sqrt{\frac{7.44013^2 + 5.62719^2}{2}} \approx 6.59624$$

可以看出這兩個 case 的結果極為相近。以 public RMSE 而言，case(2)會比較低可能是因為 public test data 本身比較偏向由 **PM2.5** 組成，或也可能有些污染源 **feature** 會使結果走偏，亦即有些污染源 **feature** 並不應該被考慮。以 private RMSE 而言，case(1)比較低可能是因為 private test data 考慮了包含 **PM2.5** 以外的的污染源 **feature**。總體而言，case(1)比 case(2)稍好

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

(1)抽全部 5 小時內的污染源 **feature** 的一次項(加 **bias**)：

public RMSE: 21.75661

private RMSE: 16.23906

$$\rightarrow \sqrt{\frac{21.75661^2 + 16.23906^2}{2}} \approx 19.19709$$

(2)抽全部 5 小時內 **pm2.5** 的一次項當作 **feature**(加 **bias**)：

public RMSE: 22.56670

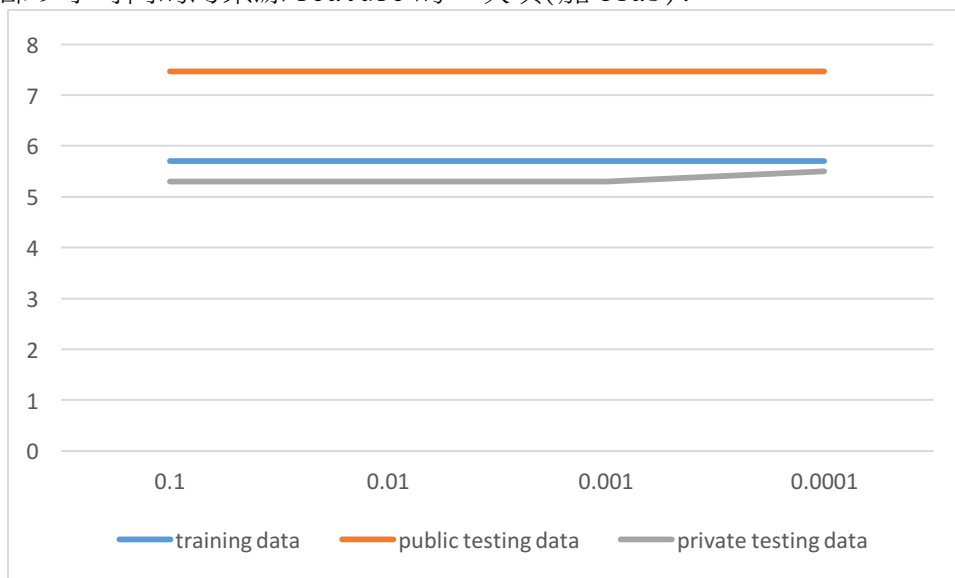
private RMSE: 16.73367

$$\rightarrow \sqrt{\frac{22.56670^2 + 16.73367^2}{2}} \approx 19.86544$$

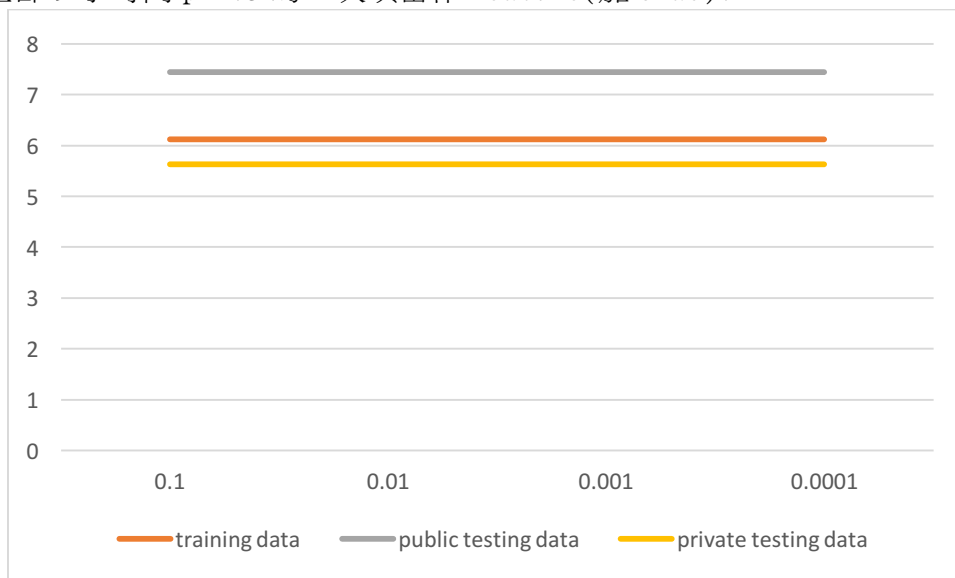
總體而言，抽 5 小時內的 case 比抽 9 小時內的 case 的 RMSE 高上許多，亦即維度增加有助於降低 cost。case(1)以 public RMSE 而言，case(2)反而比較高;以 private RMSE 而言，case(1)仍然比較低。總體而言，case(1)也比 case(2)稍好

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

(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):



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)為一存量 \mathbf{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 $\sum_{n=1}^N (\mathbf{y}^n - \mathbf{x}^n \cdot \mathbf{w})^2$ 。若將所有訓練資料的特徵值以矩陣 $\mathbf{X} = [\mathbf{x}^1 \mathbf{x}^2 \dots \mathbf{x}^N]$ 表示，所有訓練資料的標註以向量 $\mathbf{y} = [\mathbf{y}^1 \mathbf{y}^2 \dots \mathbf{y}^N]^T$ 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} ？請寫下算式並選出正確答案。

- (a) $(\mathbf{X}^T \mathbf{X}) \mathbf{X}^T \mathbf{y}$
- (b) $(\mathbf{X}^T \mathbf{X})^{-0} \mathbf{X}^T \mathbf{y}$
- (c) $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$
- (d) $(\mathbf{X}^T \mathbf{X})^{-2} \mathbf{X}^T \mathbf{y}$

$$L = \sum_{n=1}^N (\mathbf{y}^n - \mathbf{x}^n \cdot \mathbf{w})^2$$

$$L = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$$

$$\nabla L(\mathbf{w}) = 2\mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{X}^T \mathbf{y}$$

$$2\mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{X}^T \mathbf{y} = 0$$

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Ans:(c)