

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)all feature

9hr cost 5.79 public 8.44 priv 5.63884 >>>RMSE=7.18

(2)PM2.5

9hr cost 6.12 public 7.44 priv 5.6271 >>>RMSE= 6.596

>>All feature 可能會參雜太多的雜訊,導致 model 失準, 只用 PM2.5 預測的話,會比較準, 可能 all feature 裡有些與 PM2.5 無關的資訊影響結果

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

(1)all feature

5hr cost 5.83 public 8.41 priv 5.5603 >>>> RMSE =7.129

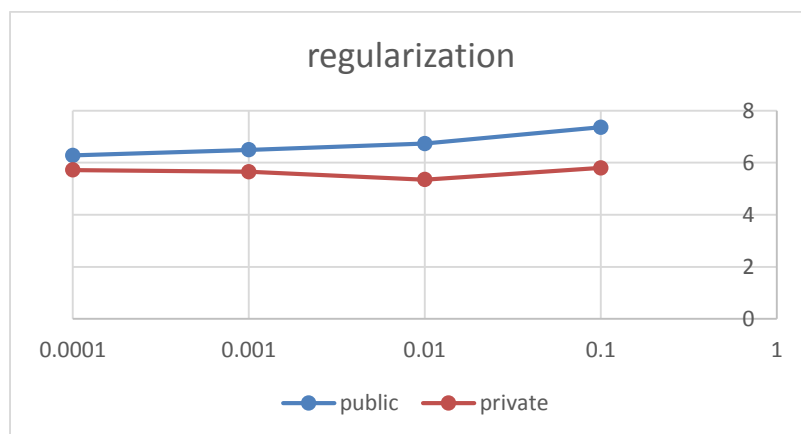
>>All feature 可能因為取的資料數變少,使 model 受到更少的不相干因素影響 導致 RMSE 下降

(2)PM2.5

5hr cost 6.207 public 7.57 priv 5.7921>>>RMSE =6.74

>>PM2.5 則因為是預測 model 裡最重要的因子,因取較少小時,接收到較少的資訊量 導致 model 誤差上升

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



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 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 $\sum_{n=1}^N (\hat{y}^n - \mathbf{x}^n \cdot \mathbf{w})^2$ 。若將所有訓練資料的特徵值以矩陣 $\mathbf{X} = [\mathbf{x}^1 \mathbf{x}^2 \dots \mathbf{x}^N]^T$ 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} ？請寫下算式並選出正確答案。(其中 $\mathbf{X}^T \mathbf{X}$ 為 invertible)

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

記分欄 從此處開始寫起。試卷用紙務須節用。非經主試認可不得續用其他紙張作答。
Please write from here.

$$\text{Loss function } L = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 = (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) = \mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T \mathbf{X}\mathbf{w} + \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w}$$

$$\frac{\partial L}{\partial \mathbf{w}} = -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \mathbf{w} = 0$$

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