

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

Feature	Private Score	Public Score	RMSE
All feature	5.40079	7.66556	6.63058605018
PM2.5	5.66065	7.45878	6.62104060971

由上面實驗結果得知，使用 PM2.5 跟使用全部的 feature 差別並不是很明顯。

備註: 我的程式是每 100 次 iteration 做一次 cost 紀錄，假如兩次 cost < 0.001 就停止訓練。

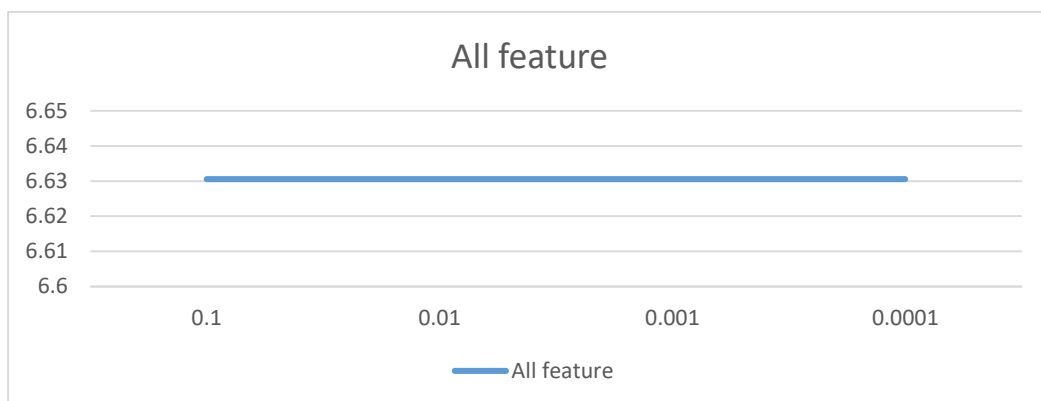
2. (1%)將 feature 從抽前 9 小時改成抽前 5 小時，討論其變化

Feature	Private Score	Public Score	RMSE
All feature	5.39173	7.75789	6.68040454033
PM2.5	5.80890	7.59657	6.76207050299

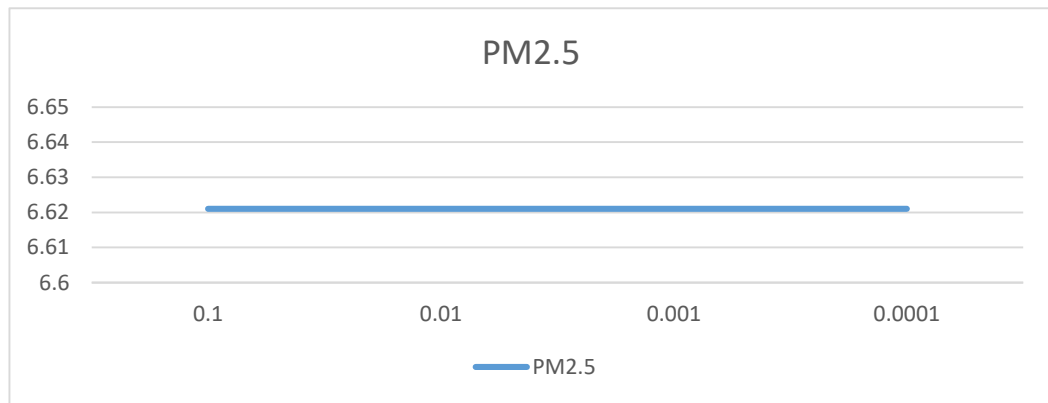
由上面的實驗結果得知，指抽取前 5 小時，表現比抽取前 9 小時還差。

2. (1%)Regularization on all the weight with $\lambda=0.1$ 、 0.01 、 0.001 、 0.0001 ，並作圖

9 小時 All feature，不同 λ 的 RMSE 變化:



9 小時 PM2.5，不同 λ 的 RMSE 變化:



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 $n=1N y^n - x^n w^2$ 。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 $X^T X$ 為 invertible)

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

Ans:

$$XW=Y$$

$$\Rightarrow X^T XW = X^T Y$$

$$\Rightarrow (X^T X)^{-1} (X^T X)W = (X^T X)^{-1} X^T Y$$

$$\Rightarrow W = (X^T X)^{-1} X^T Y$$

故選 C