R05546030 **彭新翔**

實做以下兩種不同 Feature 的模型，並回答第 (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：

(1) NR 請皆設為 0，其他的數值不要做任何更動

(2) 所有 Advanced 的 Gradient Descent 技術（如：adam、adagrad…等）都是可以用的

1. 記錄誤差值（RMSE, 根據 Kaggle Public + Private 分數），討論兩種 feature 的影響

給定條件如下：

- Learning Rate: 10
- Iterations: 100000

分別以兩種模型和作業提供資料進行訓練上傳至 Kaggle 後跑分結果如下：

	Private Score	Public Score	RMSE
[Case 01] 所有汙染源	5.30105	7.46631	6.474832667
[Case 02] 僅考慮 PM2.5	5.62719	7.44013	6.596241419

觀察上述的結果可以知道，當考慮所有汙染源時在 Private Score 預測的準確度是較高的，而在僅考慮 PM2.5 時預測的準確度是較低的，這樣的結果其實並不令人意外，因為並非空氣中所有測項粒子的濃度都與實際 PM2.5 的濃度值具有正相關性。由於作業的起初其實不太清楚 Private 和 Public 排名的差異而都在嘗試著衝高排名，當時參考了《北京地區氣熔膠 PM2.5 粒子濃度的相關因子及其估算模型》中的內容，其中提及了在非常態性降雨地區，其濃度主要會與 NO_2、SO_2、O_3 和 PM2.5 的相關度較大。

R05546030 **彭新翔**

2. 將 Feature 從抽前 9 小時改成抽前 5 小時，討論其變化。

給定條件如下：

- Learning Rate: 10
- Iterations: 100000

分別以兩種模型和作業提供資料進行訓練上傳至 Kaggle 後跑分結果如下：

	Private Score	Public Score	RMSE
[Case 01] 抽 9 小時所有汙染源	5.30105	7.46631	6.474832667
[Case 02] 抽 9 小時僅考慮 PM2.5	5.62719	7.44013	6.596241419
[Case 03] 抽 5 小時所有汙染源	5.32290	7.66477	6.601383687
[Case 04] 抽 5 小時僅考慮 PM2.5	5.49187	7.57904	6.744909392

如上述結果所示，當所抽取的小時數越多會使得誤差越小。但這樣的結果似乎與其他文獻上所得到的結果有所不同，以上一題中提到的該篇報告中所提及的，當預測估計的時間越長會導致影響因子的增加（必須將當時的風速與風向納入考慮，甚至當中濕氣也會有變化導致預測難度增加），應是中國與台灣兩地的地理環境因素所影響，此外測站位於台中豐原地區，當地的風並不大與北京有所不同也是個可能因素。

R05546030 **彭新翔**

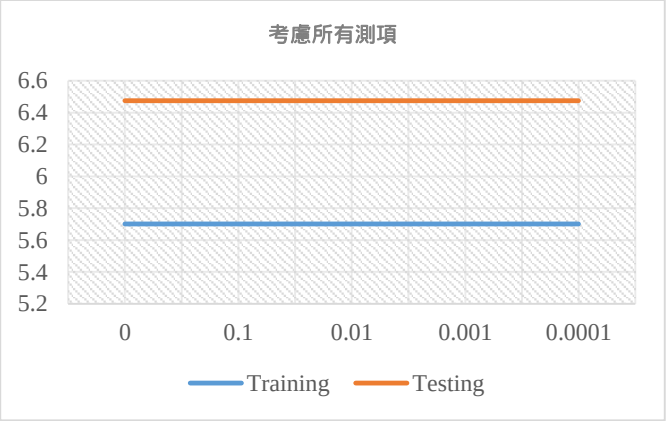
3. Regularization on all the weight with $\lambda = 0.1、0.01、0.001、0.0001$，並作圖。

給定條件如下：

- Learning Rate: 10
- Iterations: 100000

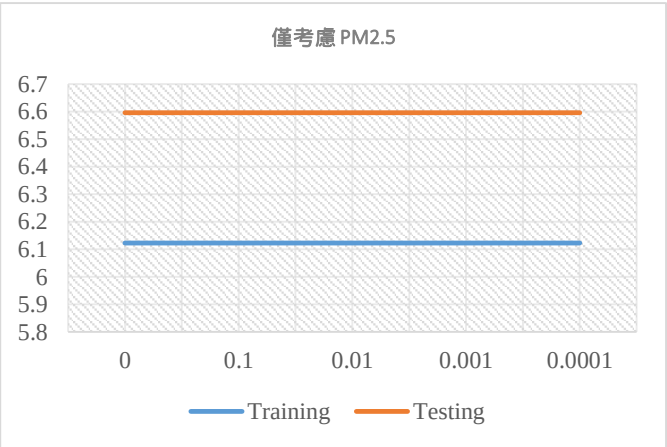
考慮所有測項，以前 9 小時進行訓練：

Lambda	Training Cost	Private	Public	RMSE
0	5.701669	5.301050	7.466310	6.474833
0.1	5.701678	5.301050	7.466310	6.474833
0.01	5.701670	5.301050	7.466310	6.474833
0.001	5.701669	5.301050	7.466310	6.474833
0.0001	5.701669	5.301050	7.466310	6.474833



考慮所有測項，以前 5 小時進行訓練：

Lambda	Training Cost	Private	Public	RMSE
0	6.123022	5.627190	7.440130	6.596241
0.1	6.123029	5.627190	7.440130	6.596241
0.01	6.123022	5.627190	7.440130	6.596241
0.001	6.123022	5.627190	7.440130	6.596241
0.0001	6.123022	5.627190	7.440130	6.596241



4. 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）為一純量 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 $X = [x^1\ x^2\ \cdots\ x^N]^T$ 表示，所有訓練資料的標註以向量 $Y = [y^1\ y^2\ \cdots\ y^N]^T$ 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) $(X^TX)^{-2}X^Ty$

已知損失函數為：

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

為求極值將方程式對 w 取一次偏微分，並令其偏導數為 0：

$$\frac{\partial L}{\partial w} = 2X^T(y - Xw) = 0$$

$$\implies 2X^TXw = 2X^Ty$$

又已知 X^TX 為可逆，故存在反矩陣，可得：

$$w = (X^TX)^{-1}X^Ty$$

答案選 (c)