

ML2017FALL-HW1-report

學號：B03801039 系級：電機四 姓名：楊福恩

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=7.83378 ; private: RMSE=5.50413

public+private : RMSE=6.76991 ;

- (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

public+private : RMSE=6.59624 ;

由以上誤差值可知，模型(2)的誤差值小於模型(1)，造成此結果的原因應為模型 (1) 所抽的 featur 太多，使得在 training data 上，通過 learning rate=10，且重複進行 10000 次的參數迭代後，RMSE 僅 5.87 左右，但在 testing data 上卻出現較大誤差，可推測因為選擇了太多 feature，使得模型(1)的 variance 較大而出現 overfitting。而由模型(2)的 RMSE 值可知誤差值較小，此模型的 feature 數目應較適當，讓預測結果較好。

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

- (1)抽前 5 小時內的污染源 feature 的一次項(加 bias):

public: RMSE= 7.73754; private: RMSE= 5.37815 ;

public+private : RMSE=6.66311 ; training: RMSE= 5.851922

和上題(1)比較可發現 testing data 上的誤差減小，原因應該為將模型(1)的 feature 數目減少，可讓模型的 variance 下降，使誤差減小。

- (2)抽前 5 小時內 pm2.5 的一次項當作 feature(加 bias):

public: RMSE= 7.57904 ; private: RMSE= 5.79187 ;

public+private : RMSE=6.74491 ; training: RMSE= 6.207004

和上題(2)比較可發現 testing data 上的誤差變大，原因應該為將模型(2)的 feature 數目減少，使得 feature 數目太少而 underfitting。

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

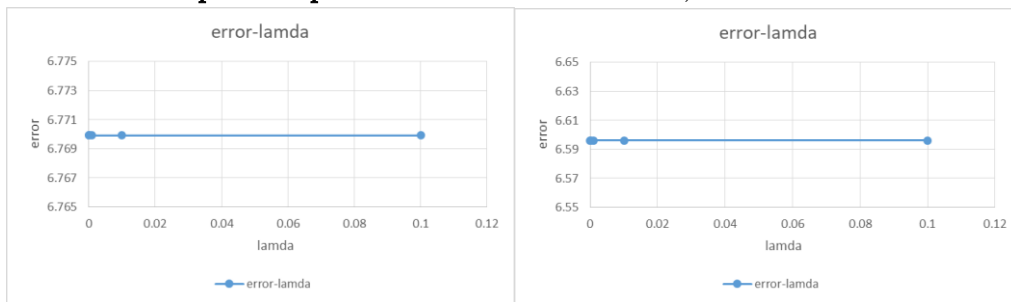
- (1) 抽全部 9 小時內的污染源 feature 的一次項(加 bias)

$\lambda=0.1$: public: RMSE= 7.83378 ; private: RMSE= 5.50414

public+private : RMSE=6.76992 ;

$\lambda=0.01$: public: RMSE= 7.83378 ; private: RMSE= 5.50413

public+private : RMSE=6.76992;
 $\lambda=0.001$: public: RMSE= 7.83378 ; private: RMSE= 5.50413
 public+private : RMSE=6.76992 ;
 $\lambda=0.0001$: public: RMSE= 7.83378 ; private: RMSE= 5.50413
 public+private : RMSE=6.76992 ;
 (2)抽全部 9 小時內 pm2.5 的一次項當作 feature(加 bias):
 $\lambda=0.1$: public: RMSE= 7.44012 ; private: RMSE= 5.62720
 public+private : RMSE=6.59624 ;
 $\lambda=0.01$: public: RMSE= 7.44013 ; private: RMSE= 5.62719
 public+private : RMSE=6.59624 ;
 $\lambda=0.001$: public: RMSE= 7.44013 ; private: RMSE= 5.62719
 public+private : RMSE=6.59624 ;
 $\lambda=0.0001$: public: RMSE= 7.44013 ; private: RMSE= 5.62719
 public+private : RMSE=6.59624 ;



以上結果可發現兩個模型在原模型其他參數皆無調整的情況下，加上 λ 作 Regularization，幾乎對原本結果沒有影響，原因應該為所用的 λ 皆很小，所以幾乎沒有造成太大的影響。

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 $\sum_{n=1}^N (y^n - x^n \cdot w)^2$ 。若將所有訓練資料的特徵值以矩陣 $X = [x^1 \ x^2 \ \dots \ x^N]^T$ 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。

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

答: $J(w) = \sum_{n=1}^N (y^n - x^n \cdot w)^2 = (w^T X - y)^T (w^T X - y)$
 $= (y^T - X^T w) (w^T X - y)$
 $= y^T (w^T X) - y^T y - X^T w w^T X + y (X^T w)$

(因為 $y^T (w^T X) = y (X^T w)$) $= 2(X^T w) y - y^T y - X^T w w^T X$

故欲求最小化損失函數的向量 w ，則讓 $J(w)$ 對 w 偏微 $\rightarrow \frac{\partial J(w)}{\partial w} = 0$
 $\rightarrow 2X^T y - 2X^T X w = 0 \rightarrow X^T X w = X^T y \rightarrow w = (X^T X)^{-1} X^T y$

答案選(c)