

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 當作一次項(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

c. 第 1-3 題請都以題目給訂的兩種 model 來回答

d. 同學可以先把 model 訓練好, kaggle 死線之後便可以無限上傳。

e. 根據助教時間的公式表示, (1) 代表  $p = 9 \times 18 + 1$  而 (2) 代表  $p = 9 \times 1 + 1$

所有模型 learning rate = 300, 跑 20000 個 iterations, 初始的 weight 全部為零

1. (2%) 記錄誤差值 (RMSE) (根據 kaggle public+private 分數), 討論兩種 feature 的影響

(1) 抽全部污染源 feature

RMSE: public = 5.68190, private = 7.26508

(2) 只抽取全部 pm2.5 的 feature

RMSE: public = 5.90263, private = 7.22356

由結果顯示, 只取 pm2.5 訓練的 model 在 public data 上表現不如 private data, 但在 private data 卻反之。可以推測其他污染源的 features 對於 pm2.5 的濃度預測確實存在影響, 但影響的關係可能並非簡單的線性迴歸可以模擬。

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

(1) 抽全部污染源 feature

Training final loss = 5.8055

Kaggle RMSE: public = 5.97570, private = 7.22232

(2) 只抽取全部 pm2.5 的 feature

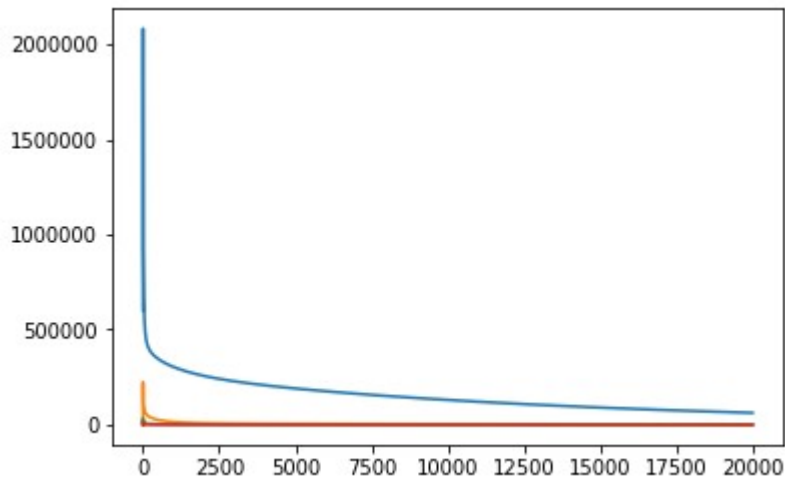
Training final loss = 6.2070

Kaggle RMSE: public = 6.22732, private = 7.22552

若單純就結果而論, 表現都明顯變得比較差, 甚至連 training 的 loss 也高出許多 (至少高出 0.5 以上)。但是在 training 的過程最顯而易見的變化是, 大概才到不到四分之一的 iteration 過程 loss 幾乎就收斂很難再降低了。從跟前面比較起來, 原因應該是因為參數本身就比較少, 使得收斂的速度加快, 但是也比較難得到較低的 loss (結果比較不準確)。

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 當作一次項(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) 為一純量  $y^n$ ，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  $\sum_{n=1}^N (y^n - \mathbf{x}^n \cdot \mathbf{w})^2$ 。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)  $(\mathbf{X}^T \mathbf{X}) \mathbf{y} \mathbf{X}^T$
- (c)  $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$
- (d)  $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{y} \mathbf{X}^T$

答案：(c)

理由