

請實做以下兩種不同feature的模型，回答第(1)~(2)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，其他的非數值(特殊字元)可以自己判斷
- b. 所有 advanced 的 gradient descent 技術(如: adam, adagrad 等)都是可以用的
- c. 第1-2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  $p = 9 * 1 + 1$

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

Adam, learning\_rate = 0.003, early stop step = 30 (20% training data for validation)

RMSE result:

Features	Public	Private
抽全部9小時內的污染源feature當作一次項(加bias)	5.49596	5.48863
抽全部9小時內PM2.5的一次項當作feature(加bias)	5.99751	5.89860

可以看到使用全部features跟只使用PM2.5比較在成績上有較好的表現。不過單是PM2.5就有一定程度的預測能力，可見PM2.5是相當關鍵的一個feature，但仍有其他的features也同樣有一定的影響力不容忽視。

2. (1%)解釋什麼樣的data preprocessing可以improve你的training/testing accuracy，ex. 你怎麼挑掉你覺得不適合的data points。請提供數據(RMSE)以佐證你的想法。

在training data的處理部分，假如有包含特殊符號或是NaN的資料部分，都直接過濾掉不使用，主要是因為含有特殊符號的部分有些數值遠遠高於平均值很多，影響訓練收斂，光是在validation set上的結果就有很顯著的差異。

RMSE result:

Data process	Public	Private
僅濾掉NaN的資料	8.78112	6.47980
濾掉所有含特殊符號或是NaN的資料	5.49596	5.48863

3.(3%) Refer to math problem

<https://hackmd.io/RFiu1FsYR5uQTrrpdxUvlw?view>

Math problem:

1.

discuss with r08521604 王鈞平

ref: <http://web.mit.edu/zoya/www/linearRegression.pdf>

$$\begin{aligned}
 & \text{1-(a)} \quad \sum_{i=1}^5 x_i = 1+2+3+4+5 = 15, \quad \sum_{i=1}^5 y_i = 1.2+2.4+3.5+4.1+5.6 = 16.8 \\
 & \frac{\partial L}{\partial b} = \frac{1}{5} \sum_{i=1}^5 (y_i - (wx_i + b))(-1) \quad \frac{\partial L}{\partial w} = \frac{1}{5} \sum_{i=1}^5 (y_i - wx_i - b)(-x_i) \\
 & \text{set } \frac{\partial L}{\partial w} = 0 \\
 & \text{set } \frac{\partial L}{\partial b} = 0 \\
 & \Rightarrow \sum_{i=1}^5 (y_i - wx_i - b) = 0 \quad \Rightarrow \sum_{i=1}^5 (x_i y_i - wx_i^2 - bx_i) = 0 \\
 & \Rightarrow b = \frac{1}{5} \sum_{i=1}^5 y_i - w \frac{1}{5} \sum_{i=1}^5 x_i \\
 & b = \frac{1}{5} \times 16.8 - w \cdot \frac{1}{5} \times 15 \quad \left( \begin{array}{l} \sum_{i=1}^5 x_i y_i = 1 \times 1.2 + 2 \times 2.4 + 3 \times 3.5 + 4 \times 4.1 + 5 \times 5.6 = 60.9 \\ \sum_{i=1}^5 x_i^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 = 55 \end{array} \right) \\
 & b = 3.36 - 3w \\
 & \Rightarrow 60.9 - 55w - (3.36 - 3w) \times 15 = 0 \\
 & \Rightarrow 60.9 - 55w - 50.4 + 45w = 0 \\
 & \Rightarrow 10w = 10.5 \Rightarrow w = 1.05 \\
 & \therefore b = 0.21 \\
 & \therefore (w, b) = (1.05, 0.21)
 \end{aligned}$$

$$1-(b) \quad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \quad \tilde{w} = \begin{bmatrix} b \\ w \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1}$$

$$\begin{aligned} L_{\text{sq}}(w, b) &= \frac{1}{2N} (y - X\tilde{w})(y - X\tilde{w})^T \\ &= \frac{1}{2N} (y^T - \tilde{w}^T X^T)(y - X\tilde{w}) \\ &= \frac{1}{2N} (y^T y - y^T X \tilde{w} - \tilde{w}^T X^T y + \tilde{w}^T X^T X \tilde{w}) \\ &= \frac{1}{2N} (y^T y - 2\tilde{w}^T X^T y + \tilde{w}^T X^T X \tilde{w}) \end{aligned}$$

$$\frac{\partial L}{\partial \tilde{w}} = \frac{1}{2N} (0 - 2X^T y + 2X^T X \tilde{w}) \stackrel{\text{set } 0}{=} 0.$$

$$\Rightarrow X^T y = X^T X \tilde{w}$$

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

$$\begin{bmatrix} b \\ w \end{bmatrix}$$

1-(c) from 1(b).

$$\frac{\partial L}{\partial \tilde{w}} = \frac{1}{N} (X^T X \tilde{w} - X^T y) + \boxed{\lambda \tilde{w}} \stackrel{\text{set } 0}{=} 0$$

$$\Rightarrow \frac{1}{N} X^T X \tilde{w} + \lambda \tilde{w} = \frac{1}{N} X^T y$$

$$\Rightarrow \left[ \frac{1}{N} X^T X + \lambda I \right] \tilde{w} = \frac{1}{N} X^T y$$

$$\Rightarrow \tilde{w} = \left[ \frac{1}{N} X^T X + \lambda I \right]^{-1} \frac{1}{N} X^T y$$

$$\begin{bmatrix} b \\ w \end{bmatrix}$$

2.

ref: <https://medium.com/autonomous-agents/mathematical-foundation-for-noise-bias-and-variance-in-neuralnetworks-4f79ee801850>

$$\begin{aligned}
2. \quad \tilde{L}_{\text{sig}}(w, b) &= E \left[ \frac{1}{2N} \sum_{i=1}^N \left( \underbrace{f_{w,b}(x_i + \eta_i)}_{f_{w,b}(x_i) + w^T \eta_i} - y_i \right)^2 \right] \\
&= E \left[ \frac{1}{2N} \sum_{i=1}^N \left( (f_{w,b}(x_i) - y_i) + w^T \eta_i \right)^2 \right] \\
&= E \left[ \frac{1}{2N} \sum_{i=1}^N (f(x_i) - y_i)^2 + \frac{1}{2N} \sum_{i=1}^N 2(f(x_i) - y_i) w^T \eta_i + \frac{1}{2N} \sum_{i=1}^N (w^T \eta_i)^2 \right] \\
&\quad E(\eta_i) = 0 \quad E[\eta_i \eta_j] = \sigma^2 \\
&= E \left[ \frac{1}{2N} \sum_{i=1}^N (f(x_i) - y_i)^2 \right] + 0 + \frac{1}{2N} \|w\|^2 \underbrace{E \left( \sum_{i=1}^N \eta_i^T \eta_i \right)}_{N \cdot \sigma^2} \\
&= \frac{1}{2N} \sum_{i=1}^N (f(x_i) - y_i)^2 + \frac{\sigma^2}{2} \|w\|^2
\end{aligned}$$

3.

ref: <https://arxiv.org/pdf/1802.04947.pdf>

3-(a)

$$\begin{aligned}
e_k &= \frac{1}{N} \sum_{i=1}^N (g_k(x_i) - y_i)^2 \\
&= \frac{1}{N} \sum_{i=1}^N \left( (g_k(x_i))^2 - 2 g_k(x_i) y_i + y_i^2 \right) \\
&= S_k - 2 \frac{1}{N} \sum_{i=1}^N g_k(x_i) y_i + e_0 \\
\Rightarrow \sum_{i=1}^N g_k(x_i) y_i &= \frac{N}{2} (S_k + e_0 - e_k) \quad k=1, 2, \dots, K
\end{aligned}$$

$$\begin{aligned}
 & \text{3-(b)} \quad \frac{1}{N} \sum_{i=1}^N \left( \sum_{k=1}^K \alpha_k g_k(x_i) - y_i \right)^2 \\
 & = \frac{1}{N} \| G \alpha - y \|^2 \\
 & \text{由 1(b). } \underbrace{\alpha = (G^T G)^{-1} G^T y}_{*} \quad \left( G^T y = \left[ \frac{1}{2} (\epsilon_k + \epsilon_0 - \epsilon_k) \right]_{k=1,2,\dots,K} \right)
 \end{aligned}$$