Machine Learning - Homework 1

資工四 B05902023 李澤諺

October 11, 2019

Part 1. Programming Problem

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 請皆設爲 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 $p = 9 \times 1 + 1$。

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

以下表格爲 data 未經過 preprocessing，並隨機選出 500 筆作爲 validation data，剩下的則作爲 training data，再使用助教所提供的 minibatch 函式 (除了 epoch 改爲 20 以外，其餘 hyper-parameter 皆不變)，於不同的 feature 下所得到的 RMSE：

Feature	Train	Validation	Test	
			Public	Private
$9 \times 18 + 1$	29.33821	26.03703	6.13045	5.67994
$9 \times 1 + 1$	29.53493	26.70052	6.44198	6.30330

兩種 feature 的 RMSE 皆偏高，除了 data 未經過 preprocessing 之外，推測可能的原因如下：$9 \times 1 + 1$ 個 feature 的 RMSE 偏高可能的原因爲 feature 的數目太少，model 不夠複雜，使得最終所得到的 linear function 和眞正的 target function 相去甚遠所致，而 $9 \times 18 + 1$ 個 feature 的 RMSE 偏高可能的原因爲 feature 的數目太多，使得 model 太過於複雜，導致了 overfit。由此可見，除了 data preprocessing 之外，feature 太多或太少，也皆有可能對 performance 造成不好的影響。

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。

以下表格中，我隨機選出 500 筆 data 作爲 validation data，剩餘的則作爲 training data，並使用助教所提供的 minibatch 函式 (除了 epoch 改爲 20 以外，其餘 hyper-parameter 皆不變)，以此進行 linear regression 並得出 RMSE。

首先，由於發現 feature 太多或太少皆會使得 performance 變差，因此計算 18 個汙染源和 PM2.5 之間的相關係數如下：

汙染源	汙染源和 PM2.5 之間的相關係數
AMB_TEMP	0.0448153729810179
CH4	0.05516971926973672
CO	0.09862870734664024
NMHC	0.026935172142996976
NO	0.01748232654065935
NO2	0.09256064805699277
NOx	0.06491192107671398
O3	0.0829880050037244
PM10	0.19017378762178272
PM2.5	1.0
RAINFALL	-0.020700122410982087
RH	-0.041557665426046025
SO2	0.1191309514421674
THC	0.04756004805889075
WD_HR	0.06979431138986333
WIND_DIREC	0.06346516798906802
WIND_SPEED	-0.01990944931047016
WS_HR	-0.03520712489750443

最後選取了相關係數較高的 6 個汙染源 (CO、NO2、O3、PM10、PM2.5、SO2)，用 $9 \times 6 + 1$ 個 feature 進行 linear regression，所得到的 RMSE 如下：

Train	Validation	Test	
		Public	Private
29.32333	24.94036	6.04361	5.62738

事實上 RMSE 並沒有多大的進步。接著，在觀察了 training data 後，發現 data 中含有許多異常值，推斷其爲造成 RMSE 偏高的主要原因，因此在經由人工判斷後，我將不滿足以下條件的 data point 視爲異常值：

汙染源	合理數值範圍
CO	$x \geq 0$
NO2	$x \geq 0$
O3	$x \geq 0$
PM10	$0 \leq x \leq 250$
PM2.5	$0 \leq x \leq 100$
SO2	$0 \leq x \leq 100$

在training data中，我將異常的data point直接丟棄，而在testing data中，則是在發現異常的汙染源數值後，便去尋找該汙染源前一個與下一個正常的數值，將兩者取平均後，取代該時間點異常的數值。在將data中的異常值去除或插值取代後，所得到的RMSE如下：

Train	Validation	Test	
		Public	Private
4.63518	4.62928	5.63850	5.46249

RMSE大幅改進，由此可以推斷異常值爲造成RMSE居高不下的最主要原因。以上即爲我所做的所有preprocessing。

Part 2. Math Problem

1. Closed-Form Linear Regression Solution

In the lecture, we've learnt how to solve linear regression problem via gradient descent. Here you will derive the closed-form solution for such kind of problems.

In the following questions, unless otherwise specified, we denote $S = \{(x_i, y_i)\}_{i=1}^N$ as a dataset of N input-output pairs, where $x_i \in \mathbb{R}^k$ denotes the vectorial input and $y_i \in \mathbb{R}$ denotes the corresponding scalar output.

1-(a)

Let's begin with a specific dataset

$$S = \{(x_i, y_i)\}_{i=1}^5 = \{(1, 1.2), (2, 2.4), (3, 3.5), (4, 4.1), (5, 5.6)\}$$

Please find the linear regression model $(w, b) \in \mathbb{R} \times \mathbb{R}$ that minimizes the sum of squares loss

$$L_{ssq}(w, b) = \frac{1}{2 \times 5} \sum_{i=1}^{5} (y_i - (w^T x_i + b))^2$$

1-(b)

Please find the linear regression model $(w, b) \in \mathbb{R}^k \times \mathbb{R}$ that minimizes the sum of squares loss

$$L_{ssq}(w, b) = \frac{1}{2 \times N} \sum_{i=1}^{N} (y_i - (w^T x_i + b))^2$$

1-(c)

A key motivation for regularization is to avoid overfitting. A common choice is to add a L2-regularization term into the original loss function

$$L_{reg}(w,b)=\frac{1}{2\times N}\sum_{i=1}^{N}(y_i-(w^Tx_i+b))^2+\frac{\lambda}{2}\|w\|^2$$

where $\lambda \geq 0$ and for $w=[w_1\ w_2\ \cdots\ w_k]^T$, one denotes $\|w\|^2=w_1^2+\cdots+w_k^2$.

Please find the linear regression model (w,b) that minimizes the aforementioned regularized sum of squares loss.

solution

1-(a)

因爲

$$\begin{aligned}L_{ssq}(w,x)&=\frac{1}{10}\sum_{i=1}^{5}(y_i-(wx_i+b))^2\\&=\frac{1}{10}((w+b-1.2)^2+(2w+b-2.4)^2+(3w+b-3.5)^2+\\&\qquad(4w+b-4.1)^2+(5w+b-5.6)^2)\end{aligned}$$

所以

$$\begin{aligned}\frac{\partial}{\partial w}L_{ssq}&=\frac{1}{10}(2(w+b-1.2)\cdot 1+2(2w+b-2.4)\cdot 2+2(3w+b-3.5)\cdot 3+\\&\qquad 2(4w+b-4.1)\cdot 4+2(5w+b-5.6)\cdot 5)\\&=11w+3b-12.18\\\frac{\partial}{\partial b}L_{ssq}&=\frac{1}{10}(2(w+b-1.2)\cdot 1+2(2w+b-2.4)\cdot 1+2(3w+b-3.5)\cdot 1+\\&\qquad 2(4w+b-4.1)\cdot 1+2(5w+b-5.6)\cdot 1)\\&=3w+b-3.36\end{aligned}$$

令

$$\frac{\partial}{\partial w}L_{ssq}=11w+3b-12.18=0 \tag{1}$$

$$\frac{\partial}{\partial b}L_{ssq}=3w+b-3.36=0 \tag{2}$$

則有

$$\begin{aligned}(1)-3\cdot(2):2w-2.1=0,\ w=1.05\\11\cdot(2)-3\cdot(1):2b-0.42=0,\ b=0.21\end{aligned}$$

故 $(w, b) = (1.05, 0.21)$ 爲 L_{ssq} 的一個 critical point。 □

1-(b)

令

$$\mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_K \end{pmatrix} \in \mathbb{R}^K,\ \mathbf{w}' = \begin{pmatrix} w'_0 \\ w'_1 \\ w'_2 \\ \vdots \\ w'_K \end{pmatrix} = \begin{pmatrix} b \\ w_1 \\ w_2 \\ \vdots \\ w_K \end{pmatrix} \in \mathbb{R}^{K+1}$$

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ \vdots \\ x_{iK} \end{pmatrix} \in \mathbb{R}^K,\ \mathbf{x}'_i = \begin{pmatrix} x'_{i0} \\ x'_{i1} \\ x'_{i2} \\ \vdots \\ x'_{iK} \end{pmatrix} = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{iK} \end{pmatrix} \in \mathbb{R}^{K+1}$$

$$X = \begin{pmatrix} \mathbf{x}_1'^T \\ \mathbf{x}_2'^T \\ \mathbf{x}_3'^T \\ \vdots \\ \mathbf{x}_N'^T \end{pmatrix} = \begin{pmatrix} x'_{10} & x'_{11} & x'_{12} & \cdots & x'_{1K} \\ x'_{20} & x'_{21} & x'_{22} & \cdots & x'_{2K} \\ x'_{30} & x'_{31} & x'_{32} & \cdots & x'_{3K} \\ \vdots & \vdots & \vdots & & \vdots \\ x'_{N0} & x'_{N1} & x'_{N2} & \cdots & x'_{NK} \end{pmatrix} \in \mathbb{R}^{N\times(K+1)}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N$$

因爲

$$\mathbf{w}^T\mathbf{x}_i + b = \sum_{j=1}^{K} w_j x_{ij} + b \cdot 1 = \sum_{j=1}^{K} w'_j x'_{ij} + w'_0 x'_{i0} = \sum_{j=0}^{K} w'_j x'_{ij}$$

所以

$$\begin{aligned}
L_{ssq} &= \frac{1}{2N}\sum_{i=1}^{N}(y_i - (\mathbf{w}^T\mathbf{x}_i + b))^2 \\
&= \frac{1}{2N}\sum_{i=1}^{N}(y_i - \sum_{j=0}^{K} w'_j x'_{ij})^2 \\
&= \frac{1}{2N}\sum_{i=1}^{N}(\sum_{j=0}^{K} w'_j x'_{ij} - y_i)^2
\end{aligned}$$

所以

$$\begin{aligned}
\frac{\partial}{\partial w'_n} L_{ssq} &= \frac{\partial}{\partial w'_n}(\frac{1}{2N}\sum_{i=1}^{N}(\sum_{j=0}^{K} w'_j x'_{ij} - y_i)^2) \\
&= \frac{1}{2N}\sum_{i=1}^{N}(2(\sum_{j=0}^{K} w'_j x'_{ij} - y_i)\cdot x'_{in}) \\
&= \frac{1}{N}\sum_{i=1}^{N}(x'_{in}(\sum_{j=0}^{K} w'_j x'_{ij} - y_i))
\end{aligned}$$

令

$$
\begin{aligned}
&\begin{pmatrix} \frac{\partial}{\partial w'_0} L_{ssq} \\ \frac{\partial}{\partial w'_1} L_{ssq} \\ \frac{\partial}{\partial w'_2} L_{ssq} \\ \vdots \\ \frac{\partial}{\partial w'_K} L_{ssq} \end{pmatrix} = 0 \\
\Leftrightarrow &\begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N}(x'_{i0}(\sum_{j=0}^{K} w'_j x'_{ij} - y_i)) \\ \frac{1}{N}\sum_{i=1}^{N}(x'_{i1}(\sum_{j=0}^{K} w'_j x'_{ij} - y_i)) \\ \frac{1}{N}\sum_{i=1}^{N}(x'_{i2}(\sum_{j=0}^{K} w'_j x'_{ij} - y_i)) \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{N}(x'_{iK}(\sum_{j=0}^{K} w'_j x'_{ij} - y_i)) \end{pmatrix} = 0 \\
\Leftrightarrow &\frac{1}{N}\begin{pmatrix} x'_{10} & x'_{20} & x'_{30} & \cdots & x'_{N0} \\ x'_{11} & x'_{21} & x'_{31} & \cdots & x'_{N1} \\ x'_{12} & x'_{22} & x'_{32} & \cdots & x'_{N2} \\ \vdots & \vdots & \vdots & & \vdots \\ x'_{1K} & x'_{2K} & x'_{3K} & \cdots & x'_{NK} \end{pmatrix} \cdot \\
&\left(\begin{pmatrix} x'_{10} & x'_{11} & x'_{12} & \cdots & x'_{1K} \\ x'_{20} & x'_{21} & x'_{22} & \cdots & x'_{2K} \\ x'_{30} & x'_{31} & x'_{32} & \cdots & x'_{3K} \\ \vdots & \vdots & \vdots & & \vdots \\ x'_{N0} & x'_{N1} & x'_{N2} & \cdots & x'_{NK} \end{pmatrix} \begin{pmatrix} w'_0 \\ w'_1 \\ w'_2 \\ \vdots \\ w'_K \end{pmatrix} - \begin{pmatrix} y'_1 \\ y'_2 \\ y'_3 \\ \vdots \\ y'_N \end{pmatrix}\right) = 0 \\
\Leftrightarrow &X^T(X\mathbf{w}' - \mathbf{y}) = 0 \\
\Leftrightarrow &X^T X\mathbf{w}' = X^T\mathbf{y}
\end{aligned}
$$

若 $X^T X$ 爲 invertible，則有 $\mathbf{w}' = (X^T X)^{-1} X^T\mathbf{y} = X^\dagger\mathbf{y}$。
若 $X^T X$ 不爲 invertible，則可取 $\mathbf{w}' = X^\dagger\mathbf{y}$，其爲 normal equation $X^T X\mathbf{w}' = X^T\mathbf{y}$ 的最佳解之一。
綜合以上所述，可得

$$
\begin{pmatrix} b \\ w_1 \\ w_2 \\ \vdots \\ w_K \end{pmatrix} = \mathbf{w}' = X^\dagger\mathbf{y} = \begin{pmatrix} 1 & \mathbf{x}_1^T \\ 1 & \mathbf{x}_2^T \\ 1 & \mathbf{x}_3^T \\ \vdots & \vdots \\ 1 & \mathbf{x}_N^T \end{pmatrix}^\dagger \mathbf{y}
$$

爲 L_{ssq} 的一個 critical point。 □

1-(c)

符號沿用自 **1-(b)**。
由 **1-(b)** 可知

$$\mathbf{w}^T\mathbf{x}_i + b = \sum_{j=0}^{K} w'_j x'_{ij}$$

而

$$\|\mathbf{w}\|^2 = \sum_{i=1}^{K} w_i^2 = \sum_{i=1}^{K} w_i'^2 = \sum_{i=0}^{K} w_i'^2 - w_0'^2 = \sum_{i=0}^{K} w_i'^2 - b^2$$

所以

$$\begin{aligned} L_{reg} &= \frac{1}{2N}\sum_{i=1}^{N}(y_i - (\mathbf{w}^T\mathbf{x}_i + b))^2 + \frac{\lambda}{2}\|w\|^2 \\ &= \frac{1}{2N}\sum_{i=1}^{N}(y_i - \sum_{j=0}^{K} w'_j x'_{ij})^2 + \frac{\lambda}{2}(\sum_{i=0}^{K} w_i'^2 - b^2) \\ &= \frac{1}{2N}\sum_{i=1}^{N}(\sum_{j=0}^{K} w'_j x'_{ij} - y_i)^2 + \frac{\lambda}{2}(\sum_{i=0}^{K} w_i'^2 - b^2) \end{aligned}$$

所以

$$\begin{aligned} \frac{\partial}{\partial w'_n} L_{reg} &= \frac{\partial}{\partial w'_n}(\frac{1}{2N}\sum_{i=1}^{N}(\sum_{j=0}^{K} w'_j x'_{ij} - y_i)^2 + \frac{\lambda}{2}(\sum_{i=0}^{K} w_i'^2 - b^2)) \\ &= \frac{1}{2N}\sum_{i=1}^{N}(2(\sum_{j=0}^{K} w'_j x'_{ij} - y_i)\cdot x'_{in}) + \frac{\lambda}{2}\cdot 2w'_n \\ &= \frac{1}{N}\sum_{i=1}^{N}(x'_{in}(\sum_{j=0}^{K} w'_j x'_{ij} - y_i)) + \lambda w'_n \end{aligned}$$

令

$$\begin{pmatrix} \frac{\partial}{\partial w'_0} L_{reg} \\ \frac{\partial}{\partial w'_1} L_{reg} \\ \frac{\partial}{\partial w'_2} L_{reg} \\ \vdots \\ \frac{\partial}{\partial w'_K} L_{reg} \end{pmatrix} = 0$$

$$\Leftrightarrow \begin{pmatrix} \frac{1}{N} \sum_{i=1}^{N} (x'_{i0}(\sum_{j=0}^{K} w'_j x'_{ij} - y_i)) + \lambda w'_0 \\ \frac{1}{N} \sum_{i=1}^{N} (x'_{i1}(\sum_{j=0}^{K} w'_j x'_{ij} - y_i)) + \lambda w'_1 \\ \frac{1}{N} \sum_{i=1}^{N} (x'_{i2}(\sum_{j=0}^{K} w'_j x'_{ij} - y_i)) + \lambda w'_2 \\ \vdots \\ \frac{1}{N} \sum_{i=1}^{N} (x'_{iK}(\sum_{j=0}^{K} w'_j x'_{ij} - y_i)) + \lambda w'_K \end{pmatrix} = 0$$

$$\Leftrightarrow \frac{1}{N} \begin{pmatrix} x'_{10} & x'_{20} & x'_{30} & \cdots & x'_{N0} \\ x'_{11} & x'_{21} & x'_{31} & \cdots & x'_{N1} \\ x'_{12} & x'_{22} & x'_{32} & \cdots & x'_{N2} \\ \vdots & \vdots & \vdots & & \vdots \\ x'_{1K} & x'_{2K} & x'_{3K} & \cdots & x'_{NK} \end{pmatrix} .$$

$$\left(\begin{pmatrix} x'_{10} & x'_{11} & x'_{12} & \cdots & x'_{1K} \\ x'_{20} & x'_{21} & x'_{22} & \cdots & x'_{2K} \\ x'_{30} & x'_{31} & x'_{32} & \cdots & x'_{3K} \\ \vdots & \vdots & \vdots & & \vdots \\ x'_{N0} & x'_{N1} & x'_{N2} & \cdots & x'_{NK} \end{pmatrix} \begin{pmatrix} w'_0 \\ w'_1 \\ w'_2 \\ \vdots \\ w'_K \end{pmatrix} - \begin{pmatrix} y'_1 \\ y'_2 \\ y'_3 \\ \vdots \\ y'_N \end{pmatrix} \right) + \lambda \begin{pmatrix} w'_0 \\ w'_1 \\ w'_2 \\ \vdots \\ w'_K \end{pmatrix} = 0$$

$$\Leftrightarrow \frac{1}{N}(X^T(X\mathbf{w}' - \mathbf{y})) + \lambda \mathbf{w}' = 0$$

$$\Leftrightarrow (X^T X + N\lambda I_{K+1})\mathbf{w}' = X^T \mathbf{y}$$

其中，因爲 $\forall\ \mathbf{u} \in \mathbb{R}^{K+1}$ 且 $\mathbf{u} \neq 0$，皆有

$$\begin{aligned} & \mathbf{u}^T(X^T X + N\lambda I_{K+1})\mathbf{u} \\ = & \mathbf{u}^T X^T X \mathbf{u} + N\lambda \mathbf{u}^T \mathbf{u} \\ = & (X\mathbf{u})^T (X\mathbf{u}) + N\lambda \mathbf{u}^T \mathbf{u} \\ = & \|X\mathbf{u}\|^2 + N\lambda \|\mathbf{u}\|^2 > 0 \end{aligned}$$

所以 $X^T X + N\lambda I_{K+1}$ 爲 positive definite。
所以 $X^T X + N\lambda I_{K+1}$ 爲 invertible。

因此可得

$$\begin{pmatrix} b \\ w_1 \\ w_2 \\ \vdots \\ w_K \end{pmatrix} = \mathbf{w}' = (X^T X + N\lambda I_{K+1})^{-1} X^T \mathbf{y}$$

$$= \left(\begin{pmatrix} 1 & \mathbf{x}_1^T \\ 1 & \mathbf{x}_2^T \\ 1 & \mathbf{x}_3^T \\ \vdots & \vdots \\ 1 & \mathbf{x}_N^T \end{pmatrix}^T \begin{pmatrix} 1 & \mathbf{x}_1^T \\ 1 & \mathbf{x}_2^T \\ 1 & \mathbf{x}_3^T \\ \vdots & \vdots \\ 1 & \mathbf{x}_N^T \end{pmatrix} + N\lambda I_{K+1} \right)^{-1} \begin{pmatrix} 1 & \mathbf{x}_1^T \\ 1 & \mathbf{x}_2^T \\ 1 & \mathbf{x}_3^T \\ \vdots & \vdots \\ 1 & \mathbf{x}_N^T \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{pmatrix}$$

爲 L_{reg} 的一個 critical point。 □

2. Noise and regulation

Consider the linear model $f_{w,b} : \mathbb{R}^k \to \mathbb{R}$, where $w \in \mathbb{R}^k$ and $b \in \mathbb{R}$, defined as

$$f_{w,b}(x) = w^T x + b$$

Given dataset $S = \{(x_i, y_i)\}_{i=1}^N$, if the inputs $x_i \in \mathbb{R}^k$ are contaminated with input noise $\eta_i \in \mathbb{R}^k$, we may consider the expected sum-of-squares loss in the presence of input noise as

$$\tilde{L}_{ssq}(w, b) = E[\frac{1}{2N} \sum_{i=1}^N (f_{w,b}(x_i + \eta_i) - y_i)^2]$$

where the expectation is taken over the randomness of input noises $\eta_1, \cdots, \eta_N$.

Now assume the input noises $\eta_i = [\eta_{i,1}\ \eta_{i,2}\ \cdots\ \eta_{i,k}]$ are random vectors with zero mean $E[\eta_{i,j}] = 0$, and the covariance between components is given by

$$E[\eta_{i,j}\eta_{i',j'}] = \delta_{i,i'}\delta_{j,j'}\sigma^2$$

where $\delta_{i,i'} = \begin{cases} 1, & if\ i = i' \\ 0, & otherwise. \end{cases}$ denotes the Kronecker delta.

Please show that

$$\tilde{L}_{ssq}(w, b) = \frac{1}{2N} \sum_{i=1}^N (f_{w,b}(x_i) - y_i)^2 + \frac{\sigma^2}{2} \|w\|^2$$

That is, minimizing the expected sum-of-squares loss in the presence of input noise is equivalent to minimizing noise-free sum-of-squares loss with the addition of a L2-regularization term on the weights.

- Hint: $\|x\|^2 = x^T x = Trace(xx^T)$.

solution

因爲

$$\begin{aligned} f_{\mathbf{w},b}(\mathbf{x}_i + \eta_i) &= \mathbf{w}^T(\mathbf{x}_i + \eta_i) + b \\ &= (\mathbf{w}^T\mathbf{x}_i + \mathbf{w}^T\eta_i) + b \\ &= (\mathbf{w}^T\mathbf{x}_i + b) + \mathbf{w}^T\eta_i \\ &= f_{\mathbf{w},b}(\mathbf{x}_i) + \mathbf{w}^T\eta_i \end{aligned}$$

所以

$$\begin{aligned} \tilde{L}_{ssq}(\mathbf{w}, b) &= E[\frac{1}{2N}\sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x}_i + \eta_i) - y_i)^2] \\ &= E[\frac{1}{2N}\sum_{i=1}^{N}((f_{\mathbf{w},b}(\mathbf{x}_i) + \mathbf{w}^T\eta_i) - y_i)^2] \\ &= E[\frac{1}{2N}\sum_{i=1}^{N}((f_{\mathbf{w},b}(\mathbf{x}_i) - y_i) + \mathbf{w}^T\eta_i)^2] \\ &= E[\frac{1}{2N}(\sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)^2 + \sum_{i=1}^{N}2(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)(\mathbf{w}^T\eta_i) + \\ &\qquad \sum_{i=1}^{N}(\mathbf{w}^T\eta_i)^2] \\ &= \frac{1}{2N}(E[\sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)^2] + E[\sum_{i=1}^{N}2(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)(\mathbf{w}^T\eta_i)] + \\ &\qquad E[\sum_{i=1}^{N}(\mathbf{w}^T\eta_i)^2]) \end{aligned}$$

其中

$$E[\sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)^2] = \sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)^2$$

而

$$
\begin{aligned}
& E[\sum_{i=1}^{N} 2(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)(\mathbf{w}^T \eta_i)] = \sum_{i=1}^{N} E[2(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)(\mathbf{w}^T \eta_i)] \\
= & \sum_{i=1}^{N}(2(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i) E[\mathbf{w}^T \eta_i]) = \sum_{i=1}^{N}(2(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i) E[\sum_{j=1}^{K} w_j \eta_{ij}]) \\
= & \sum_{i=1}^{N}(2(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i) \sum_{j=1}^{K}(w_j E[\eta_{ij}])) = \sum_{i=1}^{N}(2(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i) \sum_{j=1}^{K}(w_j \cdot 0)) = 0
\end{aligned}
$$

而

$$
\begin{aligned}
& E[\sum_{i=1}^{N}(\mathbf{w}^T \eta_i)^2] = \sum_{i=1}^{N} E[(\mathbf{w}^T \eta_i)^2] \\
= & \sum_{i=1}^{N} E[(\sum_{j=1}^{K} w_j \eta_{ij})^2] = \sum_{i=1}^{N} E[\sum_{1 \le j, j' \le K} w_j w_{j'} \eta_{ij} \eta_{ij'}] \\
= & \sum_{i=1}^{N}(\sum_{1 \le j, j' \le K} w_j w_{j'} E[\eta_{ij} \eta_{ij'}]) = \sum_{i=1}^{N}(\sum_{1 \le j, j' \le K} w_j w_{j'} \cdot \delta_{ii} \delta_{jj'} \sigma^2) \\
= & \sum_{i=1}^{N}(\sum_{1 \le j = j' \le K} w_j w_{j'} \cdot \delta_{ii} \delta_{jj'} \sigma^2) + \sum_{i=1}^{N}(\sum_{1 \le j \ne j' \le K} w_j w_{j'} \cdot \delta_{ii} \delta_{jj'} \sigma^2) \\
= & \sum_{i=1}^{N}(\sum_{1 \le j = j' \le K} w_j w_{j'} \cdot 1 \cdot 1 \cdot \sigma^2) + \sum_{i=1}^{N}(\sum_{1 \le j \ne j' \le K} w_j w_{j'} \cdot 1 \cdot 0 \cdot \sigma^2) \\
= & \sum_{i=1}^{N}(\sum_{j=1}^{K} w_j^2 \sigma^2) = \sum_{i=1}^{N}(\sigma^2 \sum_{j=1}^{K} w_j^2) = \sum_{i=1}^{N}(\sigma^2 \|\mathbf{w}\|^2) = N\sigma^2 \|\mathbf{w}\|^2
\end{aligned}
$$

因此可得

$$
\begin{aligned}
\tilde{L}_{ssq}(\mathbf{w}, b) &= \frac{1}{2N}(\sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)^2 + 0 + N\sigma^2 \|\mathbf{w}\|^2) \\
&= \frac{1}{2N} \sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)^2 + \frac{\sigma^2}{2} \|\mathbf{w}\|^2 \ \Box
\end{aligned}
$$

3. Kaggle Hacker

In the lecture, we've learnt the importance of validation. It is said that fine tuning your model based on Kaggle public leaderboard always causes ”disaster” on private test dataset.

Let's not talk about whether it'll lead to disastrous results or not. The fact is that most students even don't know how to ”overfit” public leaderboard except for submitting many and many times.

In this problem, you'll see how to take advantages of public leaderboard in hw1 kaggle competition. (In theory XD)

Suppose you have trained $K+1$ models $g_0,\ g_1,\ \cdots,\ g_K$, and in particular $g_0(x)=0$ is the zero function.

Assume the testing dataset is $\{(x_i, y_i)\}_{i=1}^N$, where you only know x_i while y_i is hidden. Nevertheless, you are allowed to observe the sum of squares testing error

$$e_k = \frac{1}{N}\sum_{i=1}^{N}(g_k(x_i) - y_i)^2,\ k = 0,\ 1\ ,\cdots,\ K$$

Of course, you know $s_k = \frac{1}{N}\sum_{i=1}^{N}(g_k(x_i))^2$.

3-(a)

Please express $\sum_{i=1}^N g_k(x_i)y_i$ in terms of $N,\ e_0,\ e_1,\ \cdots,\ e_K,\ s_1,\ \cdots,\ s_K$. Prove your answer.

- Hint: $e_0 = \frac{1}{N}\sum_{i=1}^N y_i^2$

3-(b)

For the given $K+1$ models in the previous problem, explain how to solve

$$min_{\alpha_1,\cdots,\alpha_K} L_{test}(\sum_{k=1}^{K}\alpha_k g_k) = min[\frac{1}{N}\sum_{i=1}^{N}(\sum_{k=1}^{K}\alpha_k g_k(x_i) - y_i)^2]$$

and obtain the optimal weights $\alpha_1,\ \cdots,\ \alpha_K$.

solution

3-(a)

因爲

$$\begin{aligned}
e_k &= \frac{1}{N}\sum_{i=1}^{N}(g_k(\mathbf{x}_i) - y_i)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}((g_k(\mathbf{x}_i))^2 - 2g_k(\mathbf{x}_i)y_i + y_i^2) \\
&= \frac{1}{N}\sum_{i=1}^{N}(g_k(\mathbf{x}_i))^2 - \frac{2}{N}\sum_{i=1}^{N}g_k(\mathbf{x}_i)y_i + \frac{1}{N}\sum_{i=1}^{N}y_i^2 \\
&= s_k - \frac{2}{N}\sum_{i=1}^{N}g_k(\mathbf{x}_i)y_i + e_0
\end{aligned}$$

所以

$$\sum_{i=1}^{N}g_k(\mathbf{x}_i)y_i = \frac{N}{2}(s_k - e_k + e_0) \ \square$$

3-(b)

令

$$\mathbf{Z} = \begin{pmatrix} g_1(\mathbf{x}_1) & g_2(\mathbf{x}_1) & g_3(\mathbf{x}_1) & \cdots & g_K(\mathbf{x}_1) \\ g_1(\mathbf{x}_2) & g_2(\mathbf{x}_2) & g_3(\mathbf{x}_2) & \cdots & g_K(\mathbf{x}_2) \\ g_1(\mathbf{x}_3) & g_2(\mathbf{x}_3) & g_3(\mathbf{x}_3) & \cdots & g_K(\mathbf{x}_3) \\ \vdots & \vdots & \vdots & & \vdots \\ g_1(\mathbf{x}_N) & g_2(\mathbf{x}_N) & g_3(\mathbf{x}_N) & \cdots & g_K(\mathbf{x}_N) \end{pmatrix} \in \mathbb{R}^{N\times K}$$

$$\mathbf{a} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_K \end{pmatrix} \in \mathbb{R}^K,\ \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N,\ \mathbf{s} = \begin{pmatrix} \sum_{i=1}^{N} g_1(\mathbf{x}_i)y_i \\ \sum_{i=1}^{N} g_2(\mathbf{x}_i)y_i \\ \sum_{i=1}^{N} g_3(\mathbf{x}_i)y_i \\ \vdots \\ \sum_{i=1}^{N} g_K(\mathbf{x}_i)y_i \end{pmatrix} \in \mathbb{R}^K$$

所以

$$\begin{aligned}
\frac{\partial}{\partial\alpha_n}L_{test} &= \frac{\partial}{\partial\alpha_n}(\frac{1}{N}\sum_{i=1}^{N}(\sum_{k=1}^{K}\alpha_k g_k(\mathbf{x}_i) - y_i)^2) \\
&= \frac{1}{N}\sum_{i=1}^{N}(2(\sum_{k=1}^{K}\alpha_k g_k(\mathbf{x}_i) - y_i)\cdot g_n(\mathbf{x}_i)) \\
&= \frac{2}{N}\sum_{i=1}^{N}(g_n(\mathbf{x}_i)(\sum_{k=1}^{K}\alpha_k g_k(\mathbf{x}_i) - y_i))
\end{aligned}$$

令

$$
\begin{aligned}
&\begin{pmatrix} \frac{\partial}{\partial\alpha_1}L_{test} \\ \frac{\partial}{\partial\alpha_2}L_{test} \\ \frac{\partial}{\partial\alpha_3}L_{test} \\ \vdots \\ \frac{\partial}{\partial\alpha_K}L_{test} \end{pmatrix} = 0 \\
\Leftrightarrow &\begin{pmatrix} \frac{2}{N}\sum_{i=1}^{N}(g_1(\mathbf{x}_i)(\sum_{k=1}^{K}\alpha_k g_k(\mathbf{x}_i) - y_i)) \\ \frac{2}{N}\sum_{i=1}^{N}(g_2(\mathbf{x}_i)(\sum_{k=1}^{K}\alpha_k g_k(\mathbf{x}_i) - y_i)) \\ \frac{2}{N}\sum_{i=1}^{N}(g_3(\mathbf{x}_i)(\sum_{k=1}^{K}\alpha_k g_k(\mathbf{x}_i) - y_i)) \\ \vdots \\ \frac{2}{N}\sum_{i=1}^{N}(g_K(\mathbf{x}_i)(\sum_{k=1}^{K}\alpha_k g_k(\mathbf{x}_i) - y_i)) \end{pmatrix} = 0 \\
\Leftrightarrow\ &\frac{2}{N}\begin{pmatrix} g_1(\mathbf{x}_1) & g_1(\mathbf{x}_2) & g_1(\mathbf{x}_3) & \cdots & g_1(\mathbf{x}_N) \\ g_2(\mathbf{x}_1) & g_2(\mathbf{x}_2) & g_2(\mathbf{x}_3) & \cdots & g_2(\mathbf{x}_N) \\ g_3(\mathbf{x}_1) & g_3(\mathbf{x}_2) & g_3(\mathbf{x}_3) & \cdots & g_3(\mathbf{x}_N) \\ \vdots & \vdots & \vdots & & \vdots \\ g_K(\mathbf{x}_1) & g_K(\mathbf{x}_2) & g_K(\mathbf{x}_3) & \cdots & g_K(\mathbf{x}_N) \end{pmatrix} \cdot \\
&\left(\begin{pmatrix} g_1(\mathbf{x}_1) & g_2(\mathbf{x}_1) & g_3(\mathbf{x}_1) & \cdots & g_K(\mathbf{x}_1) \\ g_1(\mathbf{x}_2) & g_2(\mathbf{x}_2) & g_3(\mathbf{x}_2) & \cdots & g_K(\mathbf{x}_2) \\ g_1(\mathbf{x}_3) & g_2(\mathbf{x}_3) & g_3(\mathbf{x}_3) & \cdots & g_K(\mathbf{x}_3) \\ \vdots & \vdots & \vdots & & \vdots \\ g_1(\mathbf{x}_N) & g_2(\mathbf{x}_N) & g_3(\mathbf{x}_N) & \cdots & g_K(\mathbf{x}_N) \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_K \end{pmatrix} - \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{pmatrix}\right) = 0 \\
\Leftrightarrow &Z^T(Z\mathbf{a} - \mathbf{y}) = 0 \\
\Leftrightarrow &Z^T Z\mathbf{a} = Z^T\mathbf{y}
\end{aligned}
$$

其中

$$
\begin{aligned}
Z^T\mathbf{y} &= \begin{pmatrix} g_1(\mathbf{x}_1) & g_1(\mathbf{x}_2) & g_1(\mathbf{x}_3) & \cdots & g_1(\mathbf{x}_N) \\ g_2(\mathbf{x}_1) & g_2(\mathbf{x}_2) & g_2(\mathbf{x}_3) & \cdots & g_2(\mathbf{x}_N) \\ g_3(\mathbf{x}_1) & g_3(\mathbf{x}_2) & g_3(\mathbf{x}_3) & \cdots & g_3(\mathbf{x}_N) \\ \vdots & \vdots & \vdots & & \vdots \\ g_K(\mathbf{x}_1) & g_K(\mathbf{x}_2) & g_K(\mathbf{x}_3) & \cdots & g_K(\mathbf{x}_N) \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{pmatrix} \\
&= \begin{pmatrix} \sum_{i=1}^{N} g_1(\mathbf{x}_i)y_i \\ \sum_{i=1}^{N} g_2(\mathbf{x}_i)y_i \\ \sum_{i=1}^{N} g_3(\mathbf{x}_i)y_i \\ \vdots \\ \sum_{i=1}^{N} g_K(\mathbf{x}_i)y_i \end{pmatrix} = \mathbf{s}
\end{aligned}
$$

因此可得

$$Z^T Z\mathbf{a} = Z^T\mathbf{y} \Leftrightarrow Z^T Z\mathbf{a} = \mathbf{s}$$

設 Z^TZ 爲 invertible，則

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_K \end{pmatrix} = \mathbf{a} = (Z^TZ)^{-1}\mathbf{s}$$

$$= \left(\begin{pmatrix} g_1(\mathbf{x}_1) & g_1(\mathbf{x}_2) & g_1(\mathbf{x}_3) & \cdots & g_1(\mathbf{x}_N) \\ g_2(\mathbf{x}_1) & g_2(\mathbf{x}_2) & g_2(\mathbf{x}_3) & \cdots & g_2(\mathbf{x}_N) \\ g_3(\mathbf{x}_1) & g_3(\mathbf{x}_2) & g_3(\mathbf{x}_3) & \cdots & g_3(\mathbf{x}_N) \\ \vdots & \vdots & \vdots & & \vdots \\ g_K(\mathbf{x}_1) & g_K(\mathbf{x}_2) & g_K(\mathbf{x}_3) & \cdots & g_K(\mathbf{x}_N) \end{pmatrix} \cdot \right.$$

$$\left. \begin{pmatrix} g_1(\mathbf{x}_1) & g_2(\mathbf{x}_1) & g_3(\mathbf{x}_1) & \cdots & g_K(\mathbf{x}_1) \\ g_1(\mathbf{x}_2) & g_2(\mathbf{x}_2) & g_3(\mathbf{x}_2) & \cdots & g_K(\mathbf{x}_2) \\ g_1(\mathbf{x}_3) & g_2(\mathbf{x}_3) & g_3(\mathbf{x}_3) & \cdots & g_K(\mathbf{x}_3) \\ \vdots & \vdots & \vdots & & \vdots \\ g_1(\mathbf{x}_N) & g_2(\mathbf{x}_N) & g_3(\mathbf{x}_N) & \cdots & g_K(\mathbf{x}_N) \end{pmatrix} \right)^{-1} \cdot$$

$$\begin{pmatrix} \sum_{i=1}^N g_1(\mathbf{x}_i)y_i \\ \sum_{i=1}^N g_2(\mathbf{x}_i)y_i \\ \sum_{i=1}^N g_3(\mathbf{x}_i)y_i \\ \vdots \\ \sum_{i=1}^N g_K(\mathbf{x}_i)y_i \end{pmatrix}$$

$$= \left(\begin{pmatrix} g_1(\mathbf{x}_1) & g_1(\mathbf{x}_2) & g_1(\mathbf{x}_3) & \cdots & g_1(\mathbf{x}_N) \\ g_2(\mathbf{x}_1) & g_2(\mathbf{x}_2) & g_2(\mathbf{x}_3) & \cdots & g_2(\mathbf{x}_N) \\ g_3(\mathbf{x}_1) & g_3(\mathbf{x}_2) & g_3(\mathbf{x}_3) & \cdots & g_3(\mathbf{x}_N) \\ \vdots & \vdots & \vdots & & \vdots \\ g_K(\mathbf{x}_1) & g_K(\mathbf{x}_2) & g_K(\mathbf{x}_3) & \cdots & g_K(\mathbf{x}_N) \end{pmatrix} \cdot \right.$$

$$\left. \begin{pmatrix} g_1(\mathbf{x}_1) & g_2(\mathbf{x}_1) & g_3(\mathbf{x}_1) & \cdots & g_K(\mathbf{x}_1) \\ g_1(\mathbf{x}_2) & g_2(\mathbf{x}_2) & g_3(\mathbf{x}_2) & \cdots & g_K(\mathbf{x}_2) \\ g_1(\mathbf{x}_3) & g_2(\mathbf{x}_3) & g_3(\mathbf{x}_3) & \cdots & g_K(\mathbf{x}_3) \\ \vdots & \vdots & \vdots & & \vdots \\ g_1(\mathbf{x}_N) & g_2(\mathbf{x}_N) & g_3(\mathbf{x}_N) & \cdots & g_K(\mathbf{x}_N) \end{pmatrix} \right)^{-1} \cdot$$

$$\begin{pmatrix} \frac{N}{2}(s_1 - e_1 + e_0) \\ \frac{N}{2}(s_2 - e_2 + e_0) \\ \frac{N}{2}(s_3 - e_3 + e_0) \\ \vdots \\ \frac{N}{2}(s_K - e_K + e_0) \end{pmatrix}$$

(其中 α_0 可以爲任意實數而不會影響 L_{test}) □