Assignment 1 | Machine Learning

Tarush Goyal

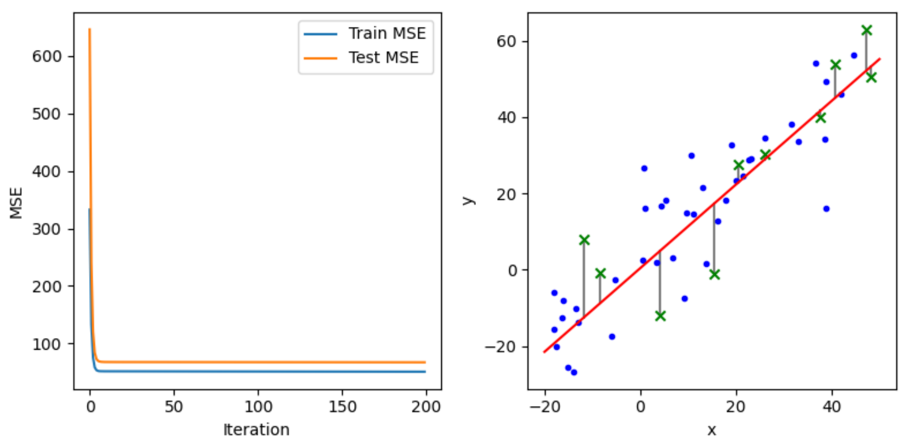
September 5, 2020

1 OLS

1.1

$$\begin{aligned}
&\nabla mse(w, b) \\
&= \begin{pmatrix} \frac{\partial mse(w,b)}{\partial w} \\ \frac{\partial mse(w,b)}{\partial b} \end{pmatrix} \\
&= \frac{1}{2N} \begin{pmatrix} \frac{\partial \sum_{i=1}^{m} (y_i - wx_i - b)^2}{\partial w} \\ \frac{\partial \sum_{i=1}^{m} (y_i - wx_i - b)^2}{\partial b} \end{pmatrix} \\
&= \frac{1}{N} \begin{pmatrix} \sum_{i=1}^{m} -x_i(y_i - wx_i - b) \\ \sum_{i=1}^{m} -(y_i - wx_i - b) \end{pmatrix}
\end{aligned}$$

1.2



The blue dots represent the training data and red line represents the best fit line for training data(that minimises the mse). The green points represent the test data. It shows that the line that minimises mse is actually a good fit for train data. However it does not seem to fit that well for test data because there was no regularisation.

2 OLS and Ridge

(a)

$$\hat{Y} = X.W$$

(b)

$$\begin{aligned} mse &= \frac{1}{2N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \\ &= \frac{1}{2N}\sum_{i=1}^{N}(\sum_{k=1}^{P} x_{i,k}w_k - y_i)^2 \end{aligned}$$

$$\begin{aligned} \frac{\partial mse}{\partial w_j} &= \frac{1}{2N}\frac{\partial \sum_{i=1}^{N}(\sum_{k=1}^{P} x_{i,k}w_k - y_i)^2}{\partial w_j} \\ &= \frac{1}{2N}\sum_{i=1}^{N}\frac{\partial(\sum_{k=1}^{P} x_{i,k}w_k - y_i)^2}{\partial w_j} \\ &= \frac{1}{2N}\sum_{i=1}^{N} 2x_{i,j}(\sum_{k=1}^{P} x_{i,k}w_k - y_i) \\ &= \frac{1}{N}X_j^T(XW - Y) \end{aligned}$$

where X_j^T is jth row of X^T

$$\begin{aligned} \frac{\partial mse}{\partial W} &= \begin{pmatrix} \frac{\partial mse}{\partial w_1} & \frac{\partial mse}{\partial w_2} & \dots & \frac{\partial mse}{\partial w_p} \end{pmatrix}^T \\ &= \frac{1}{N}\begin{pmatrix} X_1^T & X_2^T & \dots & X_P^T \end{pmatrix}^T (XW - Y) \\ &= \frac{1}{N}X^T(XW - Y) \end{aligned}$$

(c)

$$mse' = \frac{1}{2N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 + \lambda\|W\|^2$$

$$\frac{\partial mse'}{\partial W} = \frac{\partial mse}{\partial W} + \lambda\frac{\partial W^T W}{\partial W}$$

$$\frac{\partial mse'}{\partial W} = \frac{\partial mse}{\partial W} + \lambda\begin{pmatrix}\frac{\partial W^T W}{\partial w_1} & \frac{\partial W^T W}{\partial w_2} & ... & \frac{\partial W^T W}{\partial w_p}\end{pmatrix}^T$$

$$\frac{\partial mse'}{\partial W} = \frac{1}{N}X^T(XW - Y) + 2\lambda W$$

3 Weighted Linear Regression

$$\begin{aligned}
E(w) &= \frac{1}{2n}\sum_{i=1}^{n} r_i^2(y_i - w^T x_i)^2 \\
&= \frac{1}{2N}[R(XW - Y)]^T[R(XW - Y)] \\
&\left(\text{where R} = \begin{bmatrix} r_1 & 0 & 0 & .. & 0 \\ 0 & r_2 & 0 & .. & 0 \\ 0 & 0 & r_3 & .. & 0 \\ .. & .. & .. & .. & .. \\ 0 & 0 & 0 & .. & r_n \end{bmatrix}\right) \\
&= \frac{1}{2N}(X'W - Y')^T(X'W - Y') \\
&\big(\text{where X' = RX and Y' = RY}\big)
\end{aligned}$$

Thus,

$$\begin{aligned}
w^* &= (X'^T X')^{-1} X'^T Y' \text{ (from result derived in class)} \\
&= ((RX)^T(RX))^{-1}(RX)^T RY \\
&= (X^T R^2 X)^{-1} X^T R^2 Y
\end{aligned}$$

4 Failure cases of Linear Regression

4.1

$X^T X$ is singular implies X does not have full column rank. So one of the columns can be expressed as linear combination of others. We can remove that column as it does not contain any additional information and then check the same in the remaining 3 columns. Upon inspecting it can be observed that X2 is 3 times the X0. Also a 3 x 3 determinant of any 3 samples with X0, X1, X3 as columns is non zero implying that they are linearly independent. So removing X2 suffices.

4.2

Closed form is dependent on the non-singularity of $X^T X$. So if X does not have full column rank, closed form offers no solution.

Gradient descent on the other hand, is not dependent on it. Gradient descent will always converge to a solution (if done correctly) as it seeks to reduce the target loss function without caring about the redundancy of the data. However in the case of multiple solutions, it may reach a different solution every time we run it.