Niraj Mahajan.
180050069

CS 335/337
Assignment. 1

1.1 $$mse(w,b) = \sum_{i} \frac{1}{2N} \left((wx_i+b) - y_i \right)^2.$$

$$\nabla mse(w,b) = \begin{pmatrix} \frac{\partial mse}{\partial w} \\ \frac{\partial mse}{\partial b} \end{pmatrix}$$

$$\therefore \frac{\partial mse}{\partial w} = \sum \frac{1}{2N} \cdot 2(wx_i+b-y_i) \cdot (x_i).$$

$$= \sum_{i} \frac{x_i(wx_i+b-y_i)}{N}$$

$$\therefore \frac{\partial mse}{\partial b} = \sum \frac{1}{2N} \cdot 2(wx_i+b-y_i).\ (1)$$

$$= \sum_{i} \frac{(wx_i+b-y_i)}{N}.$$

$$\nabla mse(w,b) = \sum_{i} \begin{bmatrix} \frac{x_i(wx_i+b-y_i)}{N} \\ \frac{(wx_i+b-y_i)}{N} \end{bmatrix}$$

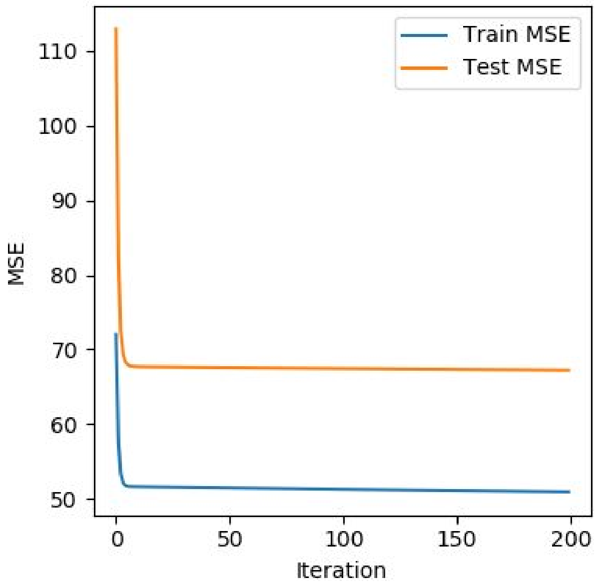
$$= \sum_{i} \left(\frac{wx_i+b-y_i}{N}\right) \begin{bmatrix} x_i \\ 1 \end{bmatrix}$$

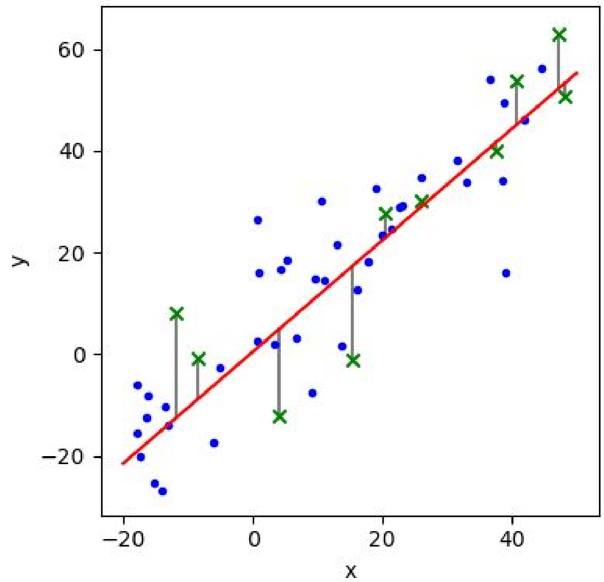
Vectorised :-

(Note:- All summations are $\sum_{i=0}^{n\_samples}$).

Assignment 1 Graphs: Niraj Mahajan: 180050069

Q1.2)d)





Q 1.2) (d) The desired image is on the previous page.
The plot on the right has x, y coordinates of the data scattered in the 2d plane. The red line is the least square fit (line that minimises sum of squared distance from the line) of the data. (train).

The green points are the test datapoints and the subsequent green lines are the deviation of the test data from theire prediction.

So, assuming the train data follows a linear relation, the red line is the best possible function that estimates / predicts the train data most appropriately by minimising the total MSE error.

2.1] Assuming :- Bias term is included in W while preprocessing X.

For a given sample $X_i = [x_{i_1} \; x_{i_2} \; --- \; x_{id}]^T$
and for given $W = [w_1 \; w_2 \; --- \; w_d]^T$

$$\text{prediction} = (w_1 x_{i_1} + w_2 x_{i_2} + \cdots + w_d x_{id}).$$

Hence, for $X_{N \times D}$, $W_{D \times 1}$

$$\boxed{\hat{Y}_{N \times 1} = X_{N \times D} \cdot W_{D \times 1}}$$

$\{$ For each individual $\hat{y}_i = x_{i_{1 \times D}}^T \cdot W_{D \times 1}$ $\}$

b) $$mse = \frac{1}{2N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

$$= \frac{1}{2N} \sum_{i=1}^{N} (w_1 x_{i_1} + w_2 x_{i_2} + \cdots w_d x_{id} - y)^2$$

$$\frac{\partial mse}{\partial \vec{w}} = \left[\frac{\partial mse}{\partial w_1} \quad \frac{\partial mse}{\partial w_2} \quad \cdots \quad \frac{\partial mse}{\partial w_d} \right]^T$$

For a general j, st $d \geq j \geq 1$,

$$\frac{\partial mse}{\partial w_j} = \frac{1}{\cancel{2}N} \sum_{i=1}^{N} \cancel{2} \cdot (w_1 x_{i_1} + w_2 x_{i_2} + \cdots w_d \cdot x_{id} - y) \times x_{ij}$$

$$\frac{\partial mse}{\partial w_j} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y} - y) \, x_{ij}$$

Vectorized Formula.

$$\frac{\partial mse}{\partial w_j} = \frac{1}{N} \, sum\left((X_{N \times D} \cdot W_{D \times 1} - Y_{N \times 1}) .* X[:, j] \right).$$

$$\frac{\partial mse}{\partial \vec{w}_{D \times 1}} = \frac{1}{N} \left[(X_{N \times D} \cdot W_{D \times 1} - Y_{N \times 1})^T \cdot X_{N \times D} \right]^T$$

Assuming typo in problem statement here

c). In case of Ridge Regression:-

$$mse = \underbrace{\frac{1}{2N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2}_{mse_1} + \underbrace{\lambda \|W\|^2}_{mse_2}.$$

Let

Define mse_1 and mse_2 as above.

→ To find $\frac{\partial mse_1}{\partial \vec{W}}$ refer to part (b).

~~mse~~ $\cancel{\frac{\partial mse_2}{\partial \vec{w}}}$

→ To find $\frac{\partial mse_2}{\partial \vec{w}}$.

$$mse_2 = \lambda . W^T . W.$$
$$= \lambda (w_1^2 + w_2^2 + \cdots w_D^2).$$

$$\therefore \frac{\partial mse_2}{\partial \vec{w}} = \left[\frac{\partial mse_2}{\partial w_1} \quad \frac{\partial mse_2}{\partial w_2} \quad \cdots \quad \frac{\partial mse_2}{\partial w_D} \right]^T$$

For a general $j \in [0, D]$

$$\frac{\partial mse_2}{\partial w_j} = 2\lambda w_j$$

$$\therefore \boxed{\frac{\partial mse_2}{\partial \vec{w}} = 2\lambda \vec{w}} \quad \text{—— (1)}$$

classmate
Date
Page

Hence, $\frac{\partial mse}{\vec{w}} = \frac{\partial mse_1}{\vec{w}} + \frac{\partial mse_2}{\vec{w}}$

From part (b), eqn ①

$$\boxed{\therefore \frac{\partial mse}{\vec{w}_{D\times 1}} = \frac{1}{N}\left[\left(X_{N\times D}\cdot W_{D\times 1} - Y_{N\times 1}\right)^T \cdot X_{N\times D}\right]^T + 2\lambda \vec{W}_{D\times 1}}$$

Q-3.1

$$E(w) = \frac{1}{2n}\sum_{i=1}^{n} r_i^2 \left(y_i - w^T x_i\right)^2.$$

$$= \frac{1}{2n}\sum_{i=1}^{n} \left(y_i r_i - w^T x_i r_i\right)^2.$$

each x_i, y_i are being transformed by r_i

Consider a diagonal matrix $R = diag(r_1, r_2, \dots r_n)$

Hence, in our linear regression, we have the following substitution,

$X_{N\times D} \rightarrow R_{N\times N}\cdot X_{N\times D}$

$Y_{N\times 1} \rightarrow R_{N\times N}\cdot Y_{N\times 1}$

→ For a linear regression, $X_{N\times D} W_{D\times 1} = Y_{N\times 1}$, the closed form can be obtained as

$X^T X \cdot W = X^T Y$

$W = (X^T X)^{-1} X^T Y$

$$\boxed{W = (X^T X)^{-1} X^T Y}$$

Hence, in our case, with the transformation of R, closed form solution:-

$$W = \left((RX)^T (RX)\right)^{-1} (RX)^T (RY)$$

$$\boxed{W = \left(X^T R^2 X\right)^{-1} X^T R^2 Y}$$

Q. 4.1 The program throughs error when X^TX is singular, as it will be non invertible.
If we look at the dataset we observe that X_0 and X_2 are related by the following relation:-

$$X_2 = 3X_0.$$

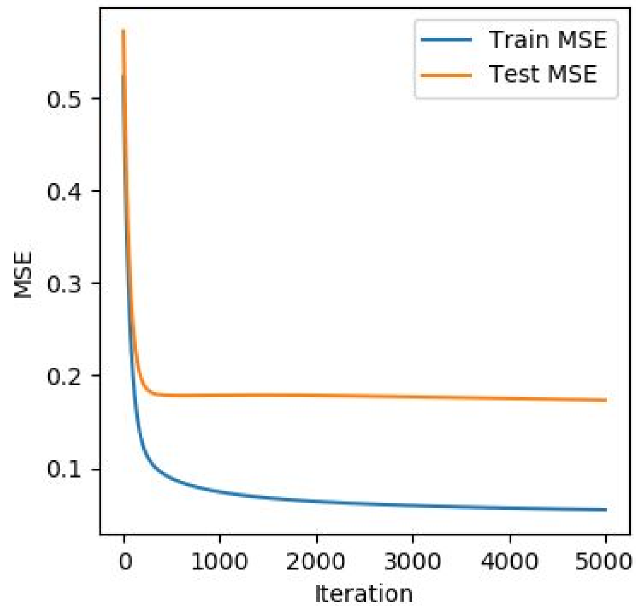
This is the reason for the singularity of X^TX. Hence, we can simply remove either of X_0, X_2 as one of them is redundant and not adding anything to our training process. Also, doing so will make X^TX non singular.

Q. 4.2 There exists no solution for the closed form of OLS when the matrix X is rank deficient, ie, one or more columns of X ~~are~~ correlated ~~by~~ by some linear relation. In this case X^TX will be singular and no closed form solution will exist for the linear regression.

→ On the other hand, Gradient Descent will still converge to some solution as the loss function (MSE) is convex and the algorithm will try to reach the global optima.

Assignment 1 Graphs: Niraj Mahajan: 180050069

Q2) OLS



Q2) Ridge