cccc

1.a *As it is given in that we have two vectors* $(y_1, y_2...., y_n)$ *and* $(x_1, x_2...., x_n)$ *of length n. Now considering that X is a matrix of length n*(d+1) we cannot make the assumption that the given data set has a non zero mean which means that* w_0 *cannot be assumed to be zero in this case.So the X matrix shapes up*

$$\begin{bmatrix} 1 & x_{11} & x_{12} & .... & x_{1d} \\ 1 & x_{21} & x_{22} & .... & x_{2d} \\ 1 & x_{31} & x_{32} & .... & x_{3d} \\ . & .... & .... & .... & .... \\ 1 & x_{n1} & x_{n2} & ....1 & x_{nd} \end{bmatrix}$$

*As it is clear that the matrix W(weight) should be dimension (d+1)*n. We can say that the weight matrix may look like:*

$$\begin{bmatrix} w_0 & w_0 & w_0 & .... & w_0 \\ w_{11} & x_{12} & x_{13} & .... & x_{1n} \\ w_{21} & x_{22} & x_{33} & .... & x_{2n} \\ . & .... & .... & .... & .... \\ w_{d1} & x_{d2} & w_{d3} & .... & x_{dn} \end{bmatrix}$$

Since the model that we are using to fit this data set is a multivariate linear regression model,so as to maintain the property of matrix multiplication of matrix X and W, so the matrix w had to be of the shape mentioned above.Now as the model used is a linear regression we can say that the predicted value of the model can be of the type:

$$\widetilde{y} = Xw$$

It is evident that y is a square matrix of dimension (d+1)(d+1).As we know that the mean square error is the square of the difference between the expected and the predicted value over all the data points.This can be written as:*

$$MSE(w) = \frac{\sum_{i=1}^{n}(y - Xw)^2}{n}$$

Now this looks similar to the the L2 norm of the term y-Xw ,if we take the square of the L2 norm of $y - \widetilde{y}$ *we get:*

$$\sum_{i=1}^{n} ||y - Xw|\,|_2^2$$

$$= \sum_{i=1}^{n} \sqrt{(y - Xw)^2})^2 = \sum_{i=1}^{n}(y - Xw)^2$$

This result is equivalent to MSE of the function $y - \widetilde{y}$.

1.b *For getting the optimal values of MSE(w), we can calculate the gradient or the partial derivative w.r.t w . As derived from the above problem we have the MSE(w) as*

$$MSE(w) = \frac{1}{n}(y - Xw)^T(y - Xw)$$
$$= \frac{1}{n}(y^T - X^T w^T)((y - Xw))$$
$$= \frac{1}{n}(y^T y - X^T w^T y - y^T Xw - X^T X w^T w)$$

Taking gradient of MSE we get:

$$\nabla(MSE) = \frac{\partial MSE}{\partial w}$$
$$= 2X^T Xw - 2X^T y$$

For the optimum solution $\nabla(MSE) = 0$. *Since it is assumed that* $rank(x) = k(fullrank)$, *then* $X^T X$ *is a positive definite and unique solution of the normal equation is*

$$X^T X\widehat{w} = X^T y$$

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

Few other assumptions that we need to consider is as follows :
(i) X is a non-stochastic matrix
(ii) X has to be a singular matrix to get its inverse
(iii)$\lim_{x \to +\infty}(\frac{X^T X}{n}) = \Delta$ *exists and is a non-stochastic and non singular matrix (with finite elements).*