HOMEWORK - 1

SSI PATTERN RECOGNITION

March 31, 2019

Submitted to

Desire SIDIBE

ASRAF ALI Abdul Salam Rasmi

Masters in Computer Vision

Centre Universitaire Condorcet

Universite de Bourgogne

1 MATRIX ALGEBRA

In this section, we need to prove some of the essential results of Matrix Algebra that can be used in Machine Learning technique.

1. $\frac{\partial(b^T a)}{\partial a} = b^T$

Solution:

Let a and b be the column vectors of size $(n \times 1)$, then

$$a = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_n \end{bmatrix} \qquad b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}$$

$$b^T a = [b_1 a_1 + b_2 a_2 + b_3 a_3 + \cdots + b_n a_n]$$

$$\therefore \quad \frac{\partial(b^T a)}{\partial a_i} = b^T \qquad \forall \ \ i = 1 \cdots n$$

To demonstrate this lets consider (2×1) column vectors,

$$\therefore \quad a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \qquad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

$$b^T a = [b_1 a_1 + b_2 a_2]$$

$$and$$

$$\frac{\partial(b^T a)}{\partial a_i} = \begin{bmatrix} \frac{\partial(b^T a)}{\partial a_1} & \frac{\partial(b^T a)}{\partial a_2} \end{bmatrix} = [b_1 \quad b_2] = b^T$$

Hence Proved.

2. $\frac{\partial(Aa)}{\partial a} = A$

Solution:

Let A be a matrix of size $(m \times n)$ and b be the column vector of size $(n \times 1)$, then

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1n} \\ A_{21} & A_{22} & A_{23} & \cdots & A_{2n} \\ A_{31} & A_{32} & A_{33} & \cdots & A_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & A_{m3} & \cdots & A_{mn} \end{bmatrix} \qquad a = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_n \end{bmatrix}$$

$$Aa = \begin{bmatrix} A_{11}a_1 + A_{12}a_2 + A_{13}a_3 + \cdots + A_{1n}a_n \\ A_{21}a_1 + A_{22}a_2 + A_{23}a_3 + \cdots + A_{2n}a_n \\ A_{31}a_1 + A_{32}a_2 + A_{33}a_3 + \cdots + A_{3n}a_n \\ \vdots \\ A_{m1}a_1 + A_{m2}a_2 + A_{m3}a_3 + \cdots + A_{mn}a_n \end{bmatrix}$$

$$\frac{\partial(Aa)}{\partial a_i} = \begin{bmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1n} \\ A_{21} & A_{22} & A_{23} & \cdots & A_{2n} \\ A_{31} & A_{32} & A_{33} & \cdots & A_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & A_{m3} & \cdots & A_{mn} \end{bmatrix} = A \qquad \forall \ i = 1 \cdots n$$

Hence Proved.

Note: This can be demonstrated in the same way as the previous example.

3. $\frac{\partial(a^T Aa)}{\partial a} = a^T(A + A^T)$

Solution:

Let A be a matrix of size $(n \times n)$ and b be the column vector of size $(n \times 1)$. We know that by **Product rule of Derivatives**,

$$\frac{d}{dx}(f(x).g(x)) = f(x).\frac{d}{dx}g(x) + g(x).\frac{d}{dx}f(x)$$

Here we can consider $f(a) = a^T$ and $g(a) = a$

$$\frac{\partial(a^T Aa)}{\partial a} = a^T A\frac{\partial}{\partial a}(a) + \frac{\partial}{\partial a}(a^T)Aa$$

Now from the first question we can say that,

$$\frac{\partial(b^T a)}{\partial a} = \frac{\partial(a^T b)}{\partial a} = b^T$$

From this we can conclude that,

$$a^T A\frac{\partial}{\partial a}(a) = a^T A$$

$$and$$

$$\frac{\partial}{\partial a}(a^T)Aa = (Aa)^T = a^T A^T$$

$$\therefore \quad \frac{\partial(a^T Aa)}{\partial a} = a^T A + a^T A^T = a^T(A + A^T)$$

To demonstrate this lets consider a (2×1) column vector and (2×2) matrix, therefore,

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \qquad a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

$$a^T Aa = [a_1^2 A_{11} + a_1 a_2 A_{12} + a_1 a_2 A_{21} + a_2^2 A_{22}]$$

$$\frac{\partial(a^T Aa)}{\partial a_i} = \begin{bmatrix} \frac{\partial(a^T Aa)}{\partial a_1} & \frac{\partial(a^T Aa)}{\partial a_2} \end{bmatrix}$$

$$\frac{\partial(a^T Aa)}{\partial a_i} = \begin{bmatrix} 2a_1A_{11} + a_2A_{12} + a_2A_{21} & a_1A_{21} + a_1A_{12} + 2a_2A_{22} \end{bmatrix} \quad (1)$$

$$A + A^T = \begin{bmatrix} 2A_{11} & A_{12} + A_{21} \\ A_{21} + A_{12} & 2A_{22} \end{bmatrix}$$

$$a^T(A + A^T) = \begin{bmatrix} 2a_1A_{11} + a_2A_{12} + a_2A_{21} & a_1A_{21} + a_1A_{12} + 2a_2A_{22} \end{bmatrix} \quad (2)$$

The R.H.S. of (1) and (2) are same, hence L.H.S. are equal. **Hence Proved.**

4. $\frac{\partial}{\partial A}(trace(BA)) = B$

Solution:

Let A_{ij} and B_{ij} be two matrices of size $(n \times n)$

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix} \qquad B = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1n} \\ B_{21} & B_{22} & \cdots & B_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1} & B_{n2} & \cdots & B_{nn} \end{bmatrix}$$

By the property of **Trace of a product**,

$$trace(BA) = \sum_{i=1}^{n} \sum_{j=1}^{n} B_{ij}A_{ji}$$

We can simplify the above equation as,

$$trace(BA) = \sum_{j=1}^{n} B_{1j}A_{j1} + \sum_{j=1}^{n} B_{2j}A_{j2} + \cdots + \sum_{j=1}^{n} B_{nj}A_{jn}$$

$$\therefore \frac{\partial}{\partial A_{ji}}(trace(BA)) = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1n} \\ B_{21} & B_{22} & \cdots & B_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1} & B_{n2} & \cdots & B_{nn} \end{bmatrix} = B$$

To demonstrate this lets consider two (2×2) matrices, therefore,

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \qquad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

$$BA = \begin{bmatrix} B_{11}A_{11} + B_{12}A_{21} & B_{11}A_{12} + B_{12}A_{22} \\ B_{21}A_{11} + B_{22}A_{21} & B_{21}A_{12} + B_{22}A_{22} \end{bmatrix}$$

$$trace(BA) = B_{11}A_{11} + B_{12}A_{21} + B_{21}A_{12} + B_{22}A_{22}$$

$$\therefore \quad \frac{\partial}{\partial A_{ji}}(trace(BA)) \;=\; \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \;=\; B$$

Hence Proved.

Note: For $\frac{\partial}{\partial A_{ij}}(trace(AB))$ we will get B^T since the index changes.

5. $trace(ABC) = trace(BCA) = trace(CAB)$

Solution:

By the property of **Trace of a product**,

$$trace(ABC) \;=\; \sum_i \sum_j \sum_k A_{ij}B_{jk}C_{ki}$$

We know that sum of product is commutative, therefore we can also write the above equation as,

$$\sum_i \sum_j \sum_k A_{ij}B_{jk}C_{ki} \;=\; \sum_j \sum_k \sum_i B_{jk}C_{ki}A_{ij} \;=\; trace(BCA)$$

$$\sum_i \sum_j \sum_k A_{ij}B_{jk}C_{ki} \;=\; \sum_k \sum_i \sum_j C_{ki}A_{ij}B_{jk} \;=\; trace(CAB)$$

$$\therefore trace(ABC) = trace(BCA) = trace(CAB)$$

Hence Trace is invariant under **Cyclic Permutation**. This is called the ***Cyclic Property of Trace***.

2 MAXIMUM LIKELIHOOD (ML) ESTIMATE

Suppose we are given a set of N observations of a scalar variable x, $\mathbf{x} = (x_1, x_2, \cdots, x_N)^T$. We assume that the observations are drawn independently from a Gaussian distribution whose mean μ and variance σ^2 are unknown. We would like to determine these parameters from the data set. The likelihood of the data given the parameters is given by

$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2)$$

1. Explain why the likelihood function can be written as a product of Gaussian?

Solution:

It is said that the observations are drawn independently from the same Gaussian distribution, which means all those observations are independent and identically distributed (i.i.d.; μ and σ^2 are same for all the observations).

Since the given data is i.i.d. we can write $p(\mathbf{x}|\mu,\sigma^2)$ as,

$$p(\mathbf{x}|\mu,\sigma^2) = p(x_1|\mu,\sigma^2) \cdot p(x_2|\mu,\sigma^2) \cdot p(x_3|\mu,\sigma^2) \cdots p(x_N|\mu,\sigma^2)$$

$$\therefore\ p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2)$$

Hence Likelihood Estimate can be written as **Product of Gaussian** if the observations are **i.i.d.**

2. ML aims to find the values of the parameters that maximize the likelihood function. Since $ln(x)$ is a monotonically increasing function of x, we can instead maximize the log of the likelihood.

(a) Compute the log likelihood function: $ln\, p(\mathbf{x}|\mu,\sigma^2)$.

(b) By taking the partial derivatives of the log likelihood with respect to μ and to σ^2, find the ML solutions μ_{ML} and σ^2_{ML}.

Solution:

(a) The likelihood of the data given the parameters is given by,

$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2)$$

$$\therefore\ p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

Then the log-likelihood of the data is,

$$ln\ p(\mathbf{x}|\mu,\sigma^2) = \sum_{n=1}^{N} ln\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_n-\mu}{\sigma}\right)^2}\right]$$

$$= \sum_{n=1}^{N}\left[ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + ln\left(e^{-\frac{1}{2}\left(\frac{x_n-\mu}{\sigma}\right)^2}\right)\right]$$

$$= \sum_{n=1}^{N} ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2}\sum_{n=1}^{N}\left(\frac{x_n-\mu}{\sigma}\right)^2$$

$$\therefore\ ln\ p(\mathbf{x}|\mu,\sigma^2) = -\frac{N}{2}ln(2\pi) - \frac{N}{2}ln(\sigma^2) - \frac{\sum_{n=1}^{N}(x_n-\mu)^2}{2\sigma^2}$$

(b) By computing derivative of the log likelihood w.r.t. μ,

$$\frac{\partial}{\partial\mu} ln\ p(\mathbf{x}|\mu,\sigma^2) = 0$$

$$0 - \frac{1}{2\sigma^2}\left(-2\sum_{n=1}^{N}(x_n-\mu)\right) = 0$$

$$\sum_{n=1}^{N}(x_n-\mu) = 0$$

$$\sum_{n=1}^{N} x_n = \sum_{n=1}^{N}\mu = N\mu$$

$$\therefore\ \mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

By computing derivative of the log likelihood w.r.t. σ^2,

$$\frac{\partial}{\partial \sigma^2} ln\ p(\mathbf{x}|\mu,\sigma^2) = 0$$

$$0 - \frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\left(\sum_{n=1}^{N}(x_n-\mu)^2\right) = 0$$

$$\frac{1}{2\sigma^2}\left(-N + \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right) = 0$$

$$N = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2$$

$$\therefore\ \sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2$$

3 LINEAR REGRESSION WITH REGULARIZATION

3.1 Linear Regression for a Polynomial Function

The Regression parameters for different degree of polynomial order of it has been computed using MATLAB and the results are shown below.

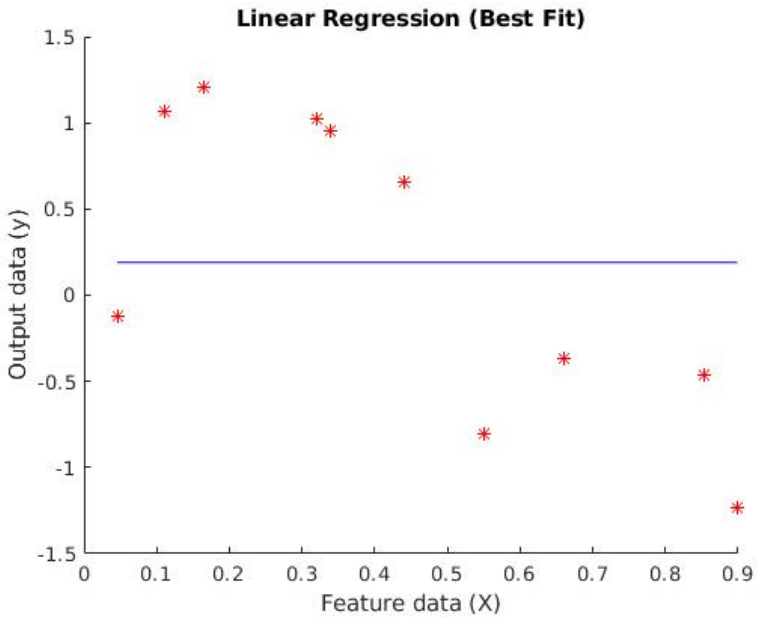

Figure 1: Output for M=0

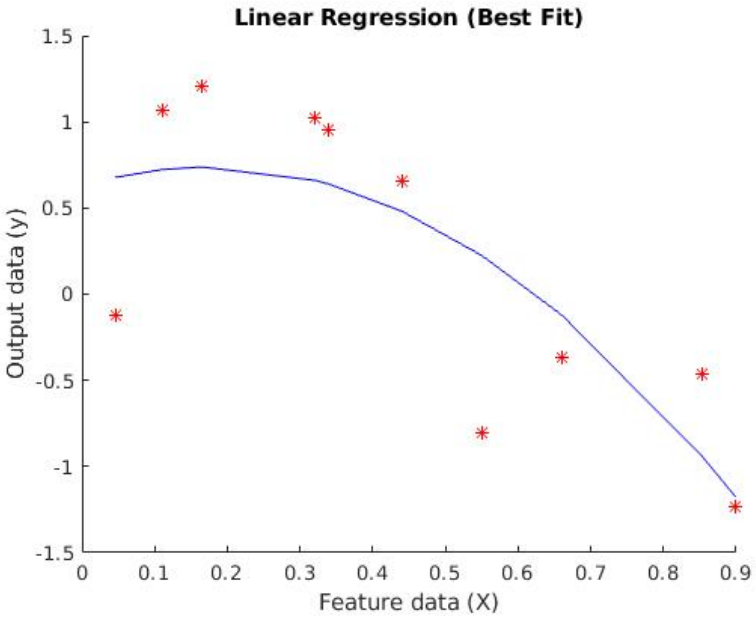

Figure 2: Output for M=2

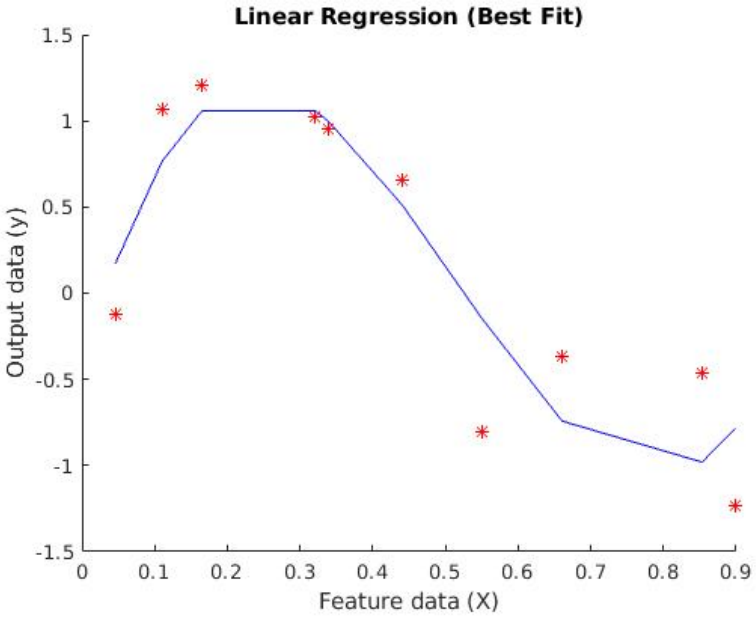

Figure 3: Output for M=3

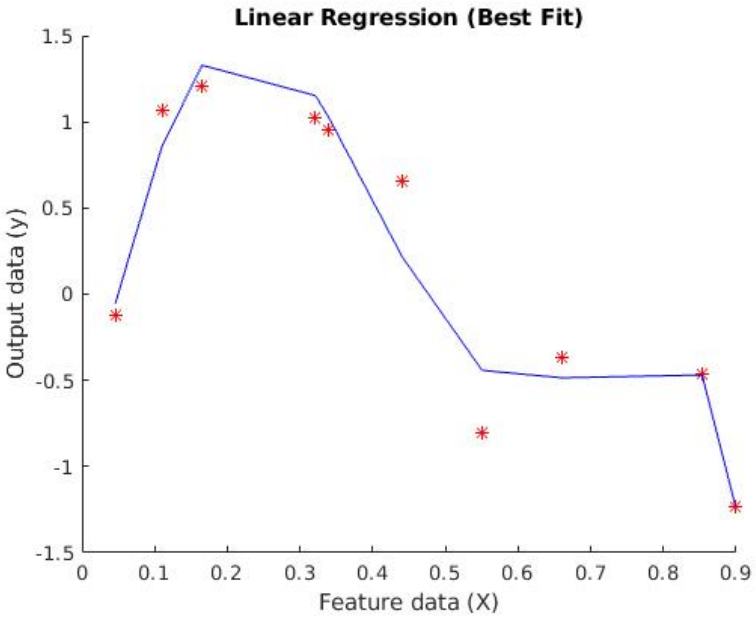

Figure 4: Output for M=5

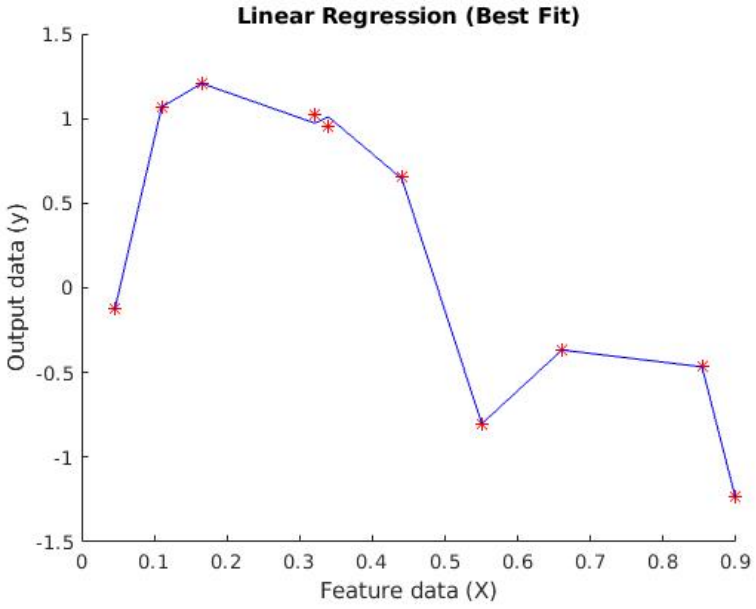

Figure 5: Output for M=8

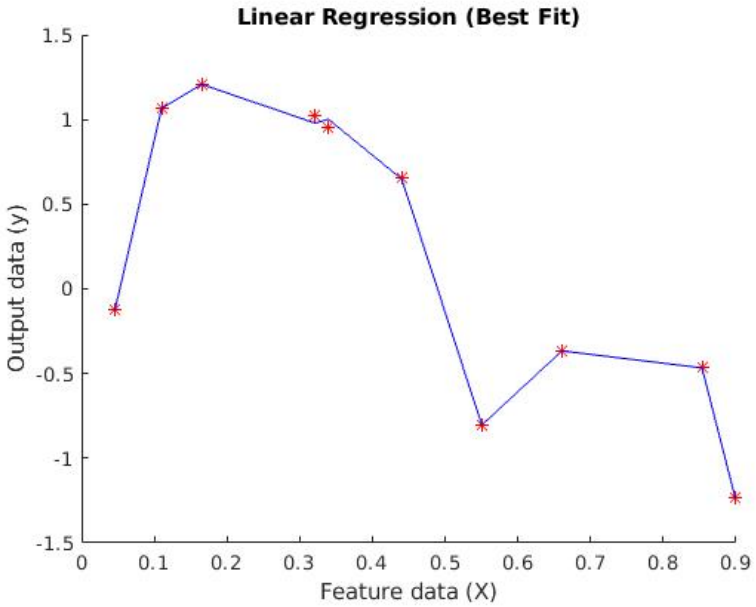

Figure 6: Output for M=10

3.2 Select a "good" Model

In order to select good model we tried to compute Regression parameter for another set of data and the RMS error for both data-set is shown below.

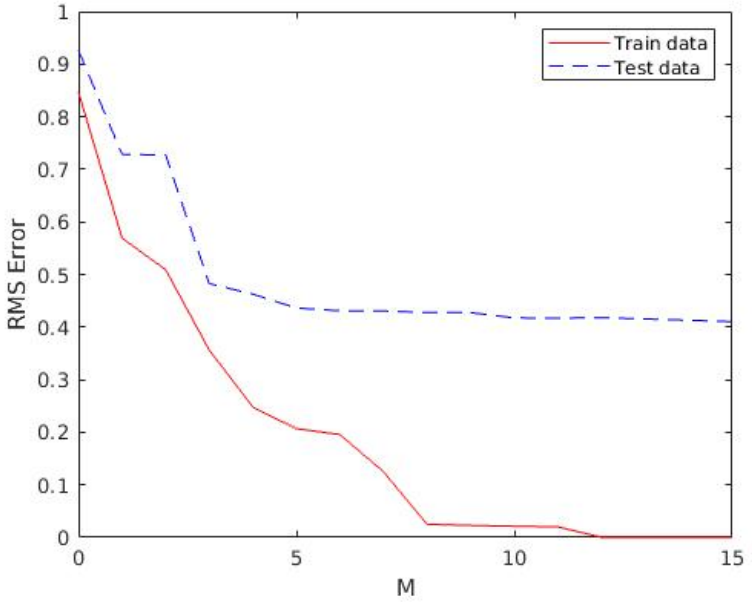

Figure 7: RMS Error for different M value

From the above figure we can say that both the data has less error for the value $M = 12$. Hence the model that best fit both test and train data is $M = 12$.

3.3 Linear Regression with Regularization

We want fit the data for the model $M = 10$. We can perform this by adding a term λ to penalize model's complexity [1]. This concept is called Regularization and the Regression with a regularization term λ is also called **Ridge Regression**.The regularization term is often known as **weight decay**[1]. We need to find λ such that the RMS Error and the Norm of the Regression Parameter should minimize. The output is shown below for different λ values.

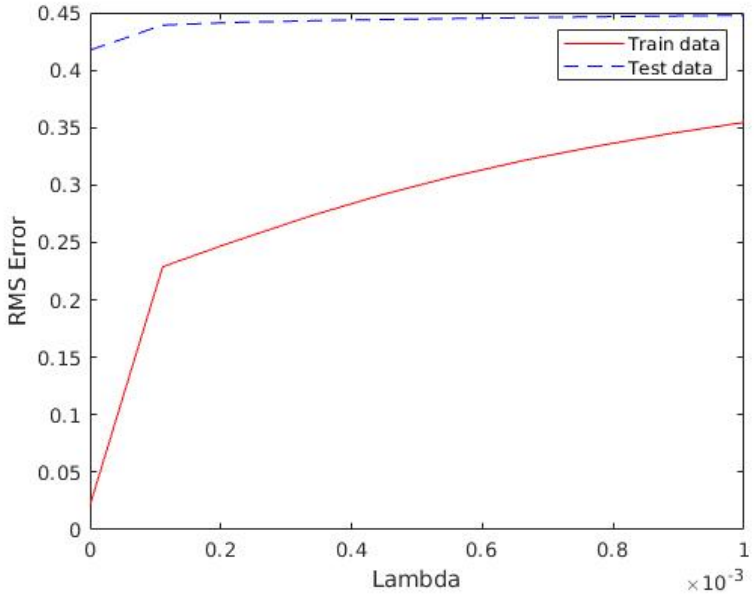

Figure 8: RMS Error for different Lambda value

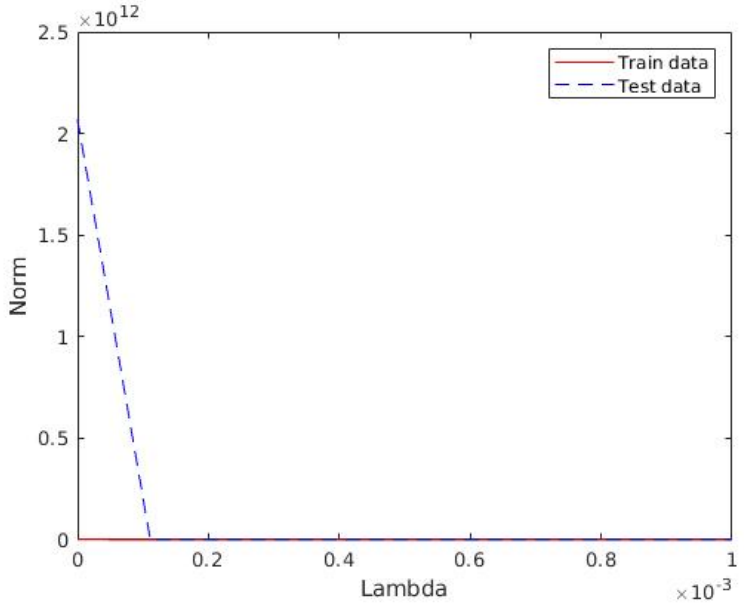

Figure 9: Norm for different Lambda value

We found that the best value for λ is 0.001 and for $M = 10$ the best fit for Training and test data is given below.

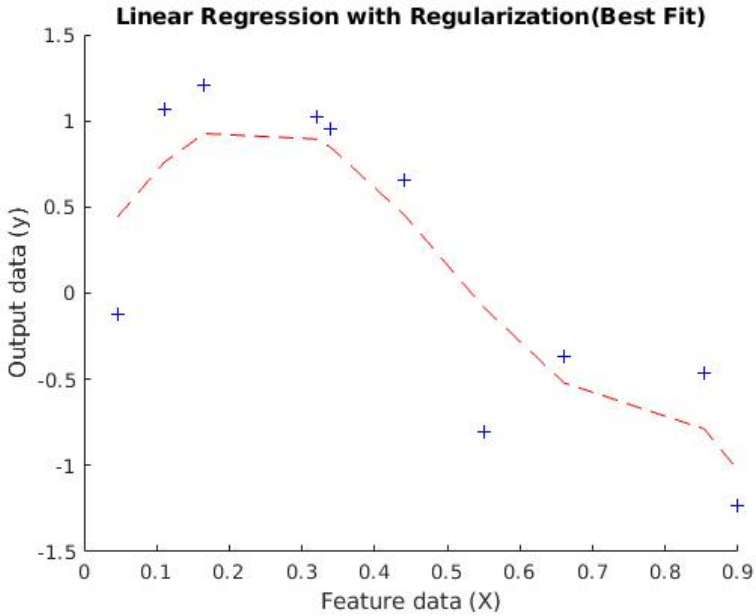

Figure 10: Result for Train data with M = 10 and lambda = 0.001

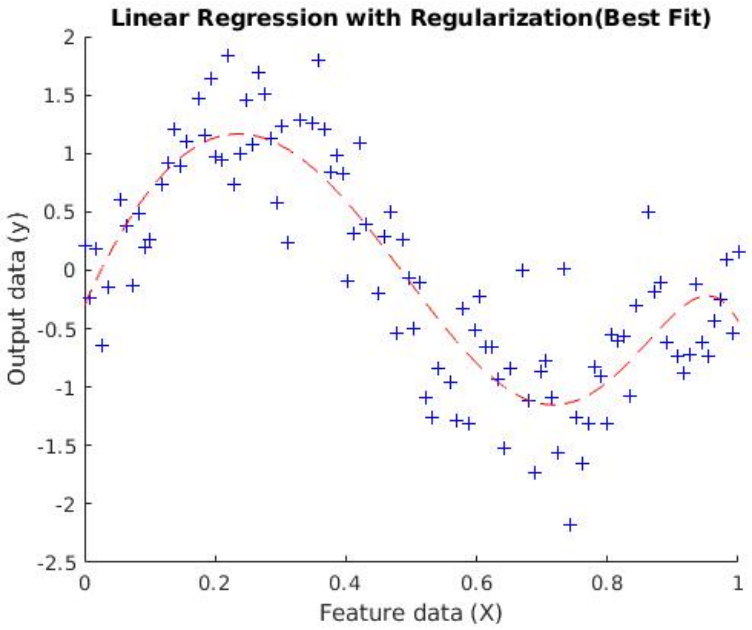

Figure 11: Result for Test data with M = 10 and lambda = 0.001

4 MLE & MAP

We consider a regression problem given a training data set $\mathbf{x} = (x_1, x_2, \cdots, x_N)^T$ with corresponding target values$\mathbf{t} = (t_1, t_2, \cdots, t_N)^T$. Let assume that given the value of x, the corresponding target value of t has a Gaussian distribution with mean equal to $y(x, w)$. Thus,

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})$$

where β^{-1} is the precision parameter equal to the inverse variance of the distribution. As in the case above, we want to use the training data $\{x, t\}$ to determine the values of the unknown parameters $\mathbf{w}$ and β. If the data is i.i.d., then the likelihood function is given by,

$$p(t|x, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})$$

1. Show that maximizing the likelihood function, to find $\boldsymbol{w}$, is equivalent to minimizing the sum of squares error function.

Solution:

From section 2, we can compute $ln\ p(t|x, \mathbf{w}, \beta)$ as,

$$ln\ p(t|x, \mathbf{w}, \beta) = -\frac{N}{2} ln(2\pi) + \frac{N}{2} ln(\beta) - \frac{\sum_{n=1}^{N} (t_n - y(x, \mathbf{w}))^2}{2\beta^{-1}}$$

Note: Here, I considered β^{-1} as **Variance (not as Inverse Variance)** of the distribution. Hence,

$$ln\ p(t|x, \mathbf{w}, \beta) = -\frac{N}{2} ln(2\pi) + \frac{N}{2} ln(\beta) - \frac{\beta \sum_{n=1}^{N} (t_n - y(x, \mathbf{w}))^2}{2}$$

To find the maxima of the likelihood function, we need to compute derivative w.r.t. $\mathbf{w}$, therefore,

$$\frac{\partial}{\partial \mathbf{w}} ln\ p(t|x, \mathbf{w}, \beta) = -\frac{\partial}{\partial \mathbf{w}} \left(\frac{\beta \sum_{n=1}^{N} (t_n - y(x, \mathbf{w}))^2}{2} \right)$$

The negative sign clearly justifies that maximizing the L.H.S. is equivalent to minimizing the R.H.S.

$$\therefore \quad \arg\max_{w} ln\ p(t|x, \mathbf{w}, \beta) = \beta \arg\min_{w} E_D(\mathbf{w})$$

where, $E_D(\mathbf{w})$ is the sum-of-square errors[1].

Hence from this expression we can conclude that **maximizing likelihood** is equivalent to **minimizing sum-of-squares** error function. In a linear model, if the errors belong to a normal distribution the least squares estimators are also the maximum likelihood estimators.

2. We can also introduce a prior information over the polynomial coefficients $\boldsymbol{w}$ *as follows.*

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I})$$

where α^{-1} *is the precision of the distribution. Using Bayes theorem, the posterior distribution for w is proportional to the product of the prior distribution and the likelihood function:*

$$p(\mathbf{w}|\mathbf{x}, t, \alpha, \beta) \propto p(t|\mathbf{x}, \mathbf{w}, \beta)\, p(\mathbf{w}|\alpha)$$

(a) Show that maximizing this posterior distribution, to find $\boldsymbol{w}$*, is equivalent to minimizing the regularized sum-of-squares error function.*

Solution:

The above expression can be expressed in log,

$$ln\ p(\mathbf{w}|\mathbf{x}, t, \alpha, \beta) = ln\ p(t|\mathbf{x}, \mathbf{w}, \beta) + ln\ p(\mathbf{w}|\alpha)$$

And the log-likelihood of the prior function can be expressed as,

$$ln\ p(\mathbf{w}|\alpha) = -\frac{N}{2} ln(2\pi) + \frac{N}{2} ln(\alpha) - \frac{\alpha \sum_{n=1}^{N} \mathbf{w}^2}{2}$$

To compute the maxima,

$$\frac{\partial}{\partial \mathbf{w}}\ ln\ p(t|\mathbf{x}, \mathbf{w}, \beta) + \frac{\partial}{\partial \mathbf{w}}\ ln\ p(\mathbf{w}|\alpha) = -\frac{\partial}{\partial \mathbf{w}}\left(\frac{\beta \sum_{n=1}^{N} (t_n - y(x, \mathbf{w}))^2}{2}\right) - \frac{\partial}{\partial \mathbf{w}}\left(\frac{\alpha \sum_{n=1}^{N} \mathbf{w}^2}{2}\right)$$

$$= -\frac{1}{2}\left(\beta \frac{\partial}{\partial \mathbf{w}} \sum_{n=1}^{N} (t_n - y(x, \mathbf{w}))^2 + \alpha \frac{\partial}{\partial \mathbf{w}} \sum_{n=1}^{N} \mathbf{w}^2\right) \tag{3}$$

From the above expression we can conclude that **maximizing posterior distribution**, to find **w**, is equivalent to **minimizing regularized sum-of-squares error** function (The negative sign clearly justifies that maximizing the L.H.S. is equivalent to minimizing the R.H.S.).

(b) Give the value of the regularized parameter λ in this case.

Solution:

From the technique of Maximum A Posteriori (MAP),

$$\mathbf{w}_{MAP} = \arg\max_{w} \; p(\mathbf{w}|\mathbf{x}, t, \alpha, \beta) = \arg\max_{w} \; p(t|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\alpha)$$

Dividing (3) by β, we get,

$$\frac{\partial}{\partial \mathbf{w}} \; ln \; p(t|\mathbf{x}, \mathbf{w}, \beta) + \frac{\partial}{\partial \mathbf{w}} \; ln \; p(\mathbf{w}|\alpha) = -\frac{1}{2}\left(\frac{\partial}{\partial \mathbf{w}} \sum_{n=1}^{N} (t_n - y(x, \mathbf{w}))^2 + \frac{\alpha}{\beta} \; \frac{\partial}{\partial \mathbf{w}} \; \sum_{n=1}^{N} \mathbf{w}^2 \right)$$

$$= -\frac{1}{2}\left(\frac{\partial}{\partial \mathbf{w}} \sum_{n=1}^{N} (t_n - y(x, \mathbf{w}))^2 + \lambda \; \frac{\partial}{\partial \mathbf{w}} \; \sum_{n=1}^{N} \mathbf{w}^2 \right)$$

where $\lambda = \frac{\alpha}{\beta}$ is the regularization parameter.

Note: If β^{-1} is considered as **Inverse Variance** of the distribution, then $\lambda = \alpha\beta$.

Bibliography

[1] D. Sidibe, “Pattern recognition - lecture notes,” 2019.