University of Burgundy

SSI Pattern Recognition

Homework 1 Linear Regression

by

Majeed

Supervisor: Dr.Desire Sidibe

I. Matrix Algebra

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

Given a, b are vectors and A is matrix. Let a, b are column vectors then from the definition we can write as follows

$$\frac{\partial \sum_{i=1}^{n} b_i a_i}{\partial a_i} = b_i$$

Replace b with b^T we get

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

1.2 Prove that $\frac{\partial(Aa)}{\partial a} = A$

Here Given A is matrix and let A $\in R^{m\,x\,n}$ and $a \in R^n$ and $A_1^T, \ldots, A_n^T$ be the rows of A.

$$Aa = \begin{bmatrix} A_1^T \\ \vdots \\ A_n^T \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} A_1^T a \\ \vdots \\ A_n^T a \end{bmatrix}$$

$$\frac{\partial(Aa)}{\partial a} = \begin{bmatrix} \frac{\partial(A_1^T a)}{\partial a} \\ \vdots \\ \frac{\partial(A_n^T a)}{\partial a} \end{bmatrix}$$

Therefore we can write the above as,

$$\frac{\partial(Aa)}{\partial a} = \begin{bmatrix} A_1^T \\ \vdots \\ A_n^T \end{bmatrix} = A$$

1.3 Prove that $\frac{\partial(a^T A a)}{\partial a} = a^T(A + A^T)$

Using product rule, we can write given equation as follows

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

where $\bar{a}$ is constant.

From using above equations, we can prove the given statements.

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

From the above equation we can use it to prove as follows

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

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

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

1.4 Prove that $\frac{\partial trac(BA)}{\partial A} = B$

Here in Matrix A and B, let $b_1^T, \ldots, b_n^T$ be the rows of B and $a_1, \ldots, a_n$ be the columns of A

$$tr(BA) = tr \begin{bmatrix} b_1^T \\ \vdots \\ b_n^T \end{bmatrix} [a_1 \quad \ldots\ldots \quad a_n]$$

$$tr\,(BA) = tr \begin{bmatrix} b_1^T a_1 & \ldots & b_1^T a_n \\ \vdots & \ddots & \vdots \\ b_n^T a_1 & \ldots & b_n^T a_n \end{bmatrix}$$

$$tr\,(BA) = b_1^T a_1 + b_2^T a_2 + \ldots + b_n^T a_n$$

$$tr\,(BA) = \sum_{i=1}^{m} b_{1i} a_{i1} + \sum_{i=1}^{m} b_{2i} a_{i2} + \ldots + \sum_{i=1}^{m} b_{ni} a_{in}$$

$$\boldsymbol{\frac{\partial tr(BA)}{\partial a} = b_{ji}^T = b_{ij} = B}$$

1.5 Prove that $tr(ABC) = tr(CAB) = tr(BCA)$

Trace of the matrix is equals to some of the diagonal elements.

$$tr(ABC) = \sum_{i,j,k} A_{ij}\, B_{jk} C_{ki}$$

Trace of A,B,C is equal to trace of C,A,B and B,CA

$$\sum_{i,j,k} A_{ij}\, B_{jk} C_{ki} \;=\; \sum_{i,j,k} C_{ki}\, A_{ij}\, B_{jk} \;=\; \sum_{i,j,k} B_{jk} C_{ki}\, A_{ij}$$

Where $\sum_{i,j,k} C_{ki}\, A_{ij}\, B_{jk} \;=\; tr(CAB)$

$\sum_{i,j,k} B_{jk} C_{ki}\, A_{ij} \;=\; tr(BCA)$

Therefore, we can state as follows.

$$\boldsymbol{tr(ABC) \;=\; tr(CAB) \;=\; tr(BCA)}$$

II. Maximum Likelihood Estimate

2.1 The likelihood of the data given the parameters is given by

$$P(X|\mu, \sigma^2) \;=\; P(X_1, \ldots, X_n|\mu, \sigma^2)$$

It is assumed to be identically independent variables of $X_1, \ldots, X_n$ then,

$$Z_n \;=\; \sum_{i\,=\,1}^{n} X_i$$

$$\lim_{n \to \infty} Z_n \;\to\; \mathcal{N}(\mu, \sigma)$$

Therefore we can rewrite above equation as follows:

$$P(X|\mu, \sigma^2) \;=\; P(X_1|\mu, \sigma^2)\; P(X_2|\mu, \sigma^2) \ldots\; P(X_n|\mu, \sigma^2)$$

$$P(X|\mu, \sigma^2) \;=\; \prod P(X_i|\mu, \sigma^2)$$

$$\boldsymbol{P(X|\mu, \sigma^2) \;=\; \prod \mathcal{N}(X_i|\mu, \sigma^2)}$$

From the above equation we can say that likelihood function can be written as a product of Gaussians.

2.2. The log likelihood function

From above equation, we have

$$P(X|\mu, \sigma^2) = \prod \mathcal{N}(X_i|\mu, \sigma^2)$$

$$P(X|\mu,\sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-1/2\left(\frac{X_i-\mu}{\sigma}\right)^2}$$

Apply log on both sides

$$\ln P(X|\mu,\sigma^2) = \sum_{i=1}^{N} ln \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-1/2\left(\frac{X_i-\mu}{\sigma}\right)^2}\right]$$

$$\ln P(X|\mu,\sigma^2) = \sum_{i=1}^{N} \left[ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + ln\left(e^{-1/2\left(\frac{X_i-\mu}{\sigma}\right)^2}\right)\right]$$

$$\boldsymbol{\ln P(X|\mu,\sigma^2) = \sum_{i=1}^{N} ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{X_i-\mu}{\sigma}\right)^2}$$

2.2.b

Here in order to get mean and variance we partial derivate the above equation w.r.t to μ *and* σ^2 and equate to zero.

To obtain mean first we partial derivate w.r.t to μ

$$\frac{\partial \ln P(X|\mu,\sigma^2)}{\partial\mu} = 0$$

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

$$\sum_{i=1}^{N}(X_i-\mu) = 0$$

$$\sum_{i=1}^{N} X_i = \sum_{i=1}^{N}\mu$$

$$\sum_{i=1}^{N} X_i = N\mu$$

$$\boldsymbol{\mu_{ml} = \frac{1}{N}\sum_{i=1}^{N} X_i}$$

To obtain variance first we partial derivate w.r.t to σ^2

$$\frac{\partial \ln P(X|\mu,\sigma^2)}{\partial \sigma^2} = 0$$

As we know from log likelihood estimate,

$$\ln P(X|\mu,\sigma^2) = \frac{-N}{2} ln(2\pi) - \frac{-N}{2} ln(\sigma^2) - \frac{\sum(X_i-\mu)^2}{2\sigma^2}$$

Using above equation, we can partial derivate with respect to σ^2 , we get,

$$0 - \frac{N}{2}\frac{1}{\sigma^2} + \frac{\sum(X_i-\mu)^2}{2\sigma^4} = 0$$

$$\frac{1}{2\sigma^2}\left[-N + \frac{\sum(X_i-\mu)^2}{\sigma^2}\right] = 0$$

$$\frac{\sum(X_i-\mu)^2}{\sigma^2} = N$$

$$\boldsymbol{\sigma_{ml}^2 = \frac{1}{N}\sum(X_i-\mu)^2}$$

III. Linear Regression with Regularization

3.1 The code for this part is attached with this report in Homework folder.

3.2 Output results using different values of M are as follows

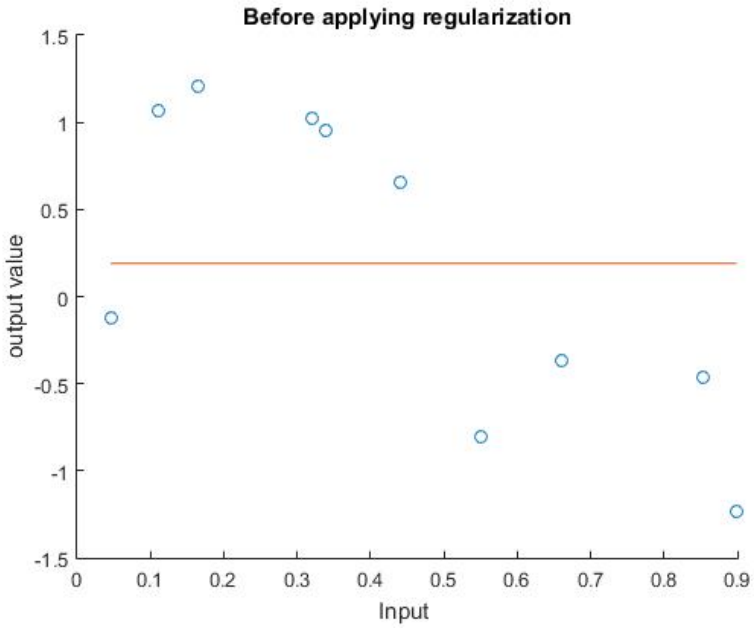

Figure 1 When M = 0

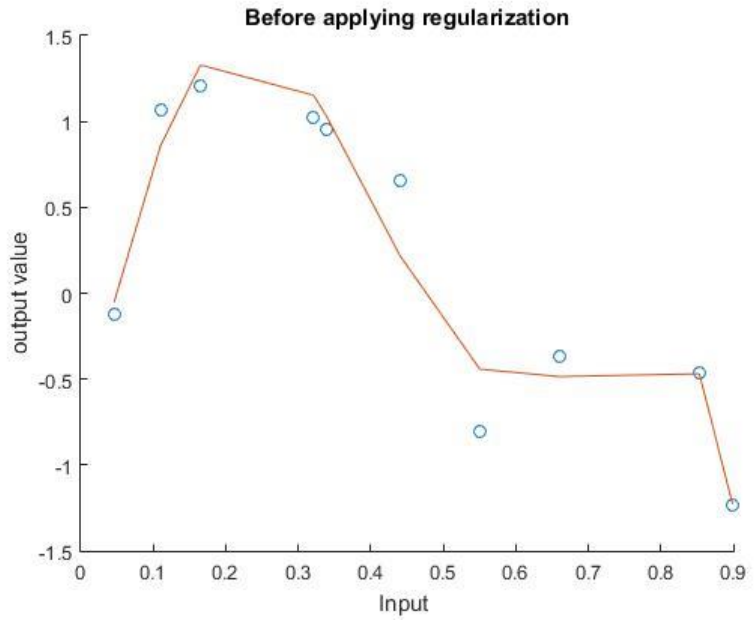

Figure 1 When M = 5

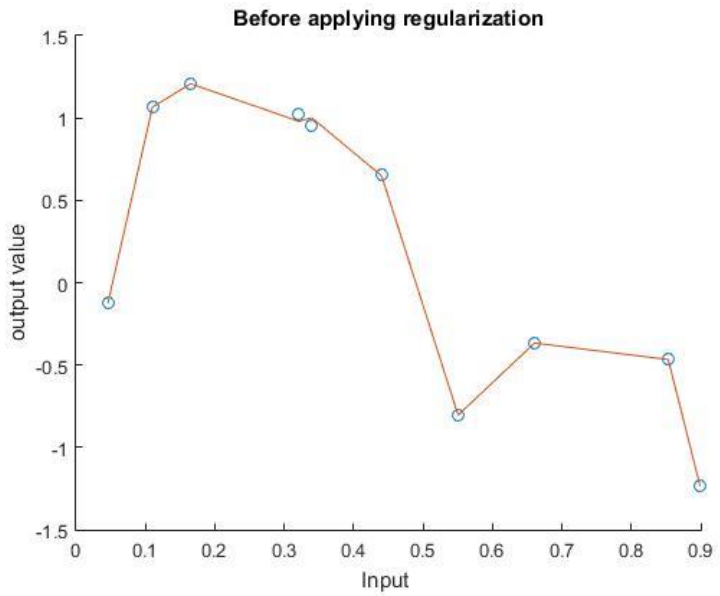

Figure 1 When M = 10

3.3 For different values of M, best M can be chosen by observing the below figure.

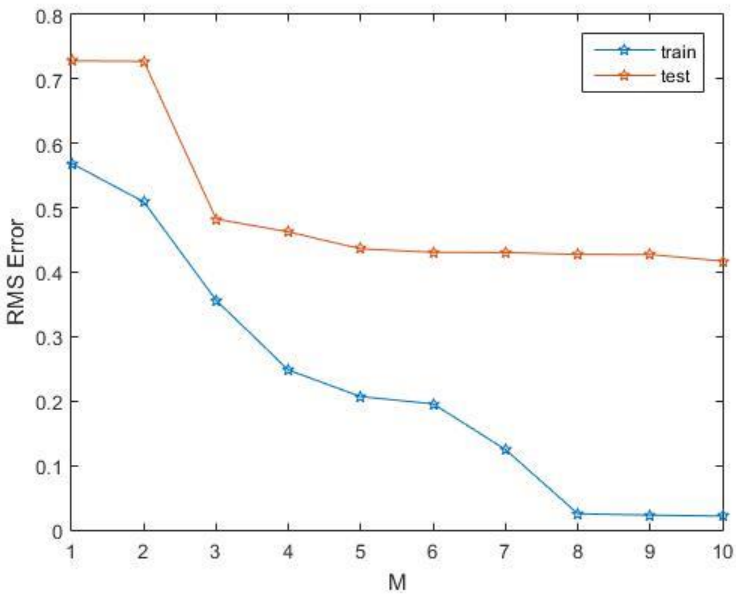

Figure 5 RMS Error vs M

From the figure we can see that after M = 10 there is less training happening. We can apply regularization to deal with overfitting issues.

3.4 We now want to fit a model of order M = 10 to the data.

Now we want to reduce the capacity of model we need to regularize the model by adding a term to cost function that penalizes the complex models.

So, we updated the cost function as,

$$\tilde{E}(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2 + \frac{\lambda}{2}\|w\|^2$$

where λ is the hyper parameter to control the amount of regularization.

And, finally the parameters can be estimated as,

$$w = (\lambda I + \phi^T\phi)^{-1}\phi^T t$$

3.5 To Find good regularization parameter.

In order to find best lambda value we can just plot the error vs lambda value and use those plot to figure out the best one.

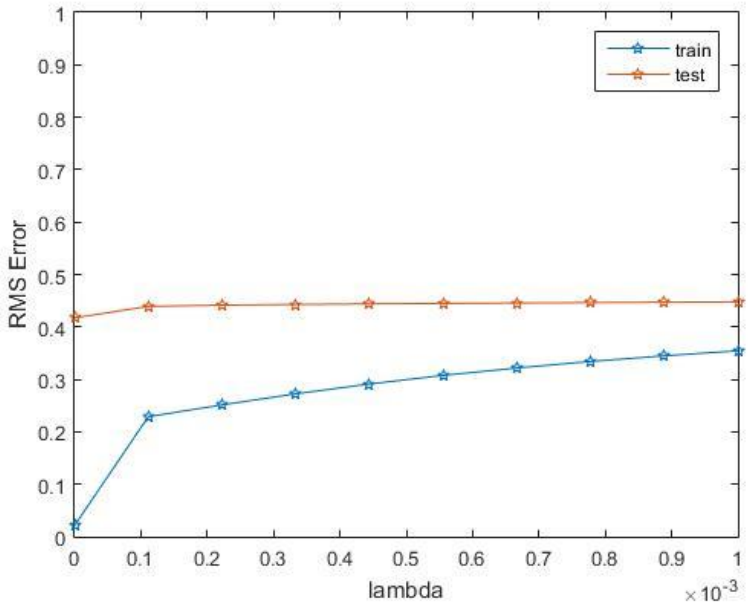

Figure 6 RMS Error vs lambda

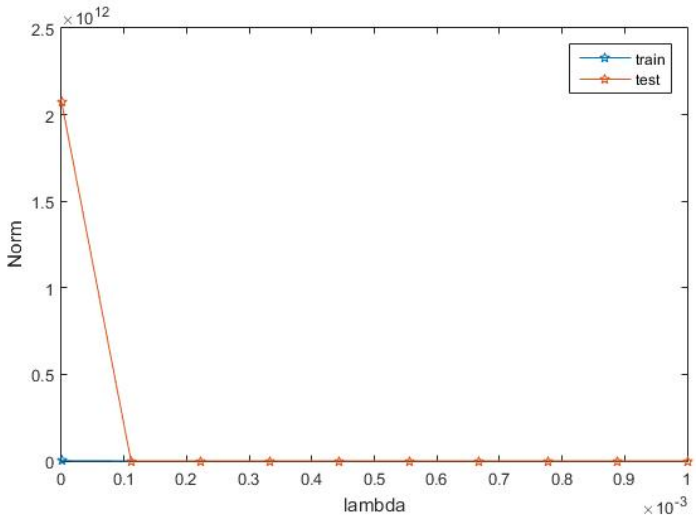

Figure 7 Norm of parameter vs lambda

There is always a trade-off between regularization value and fitting error to reduce the complexity of the model. The best way is to choose small values here we used $\lambda = 0.0001$ using this we plot before regularization and after regularization figures as follows

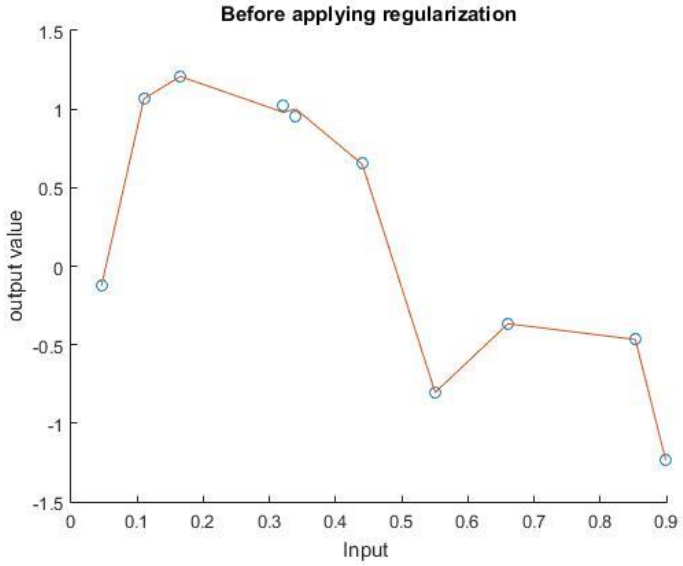

Figure 8 M = 10 and before regularization

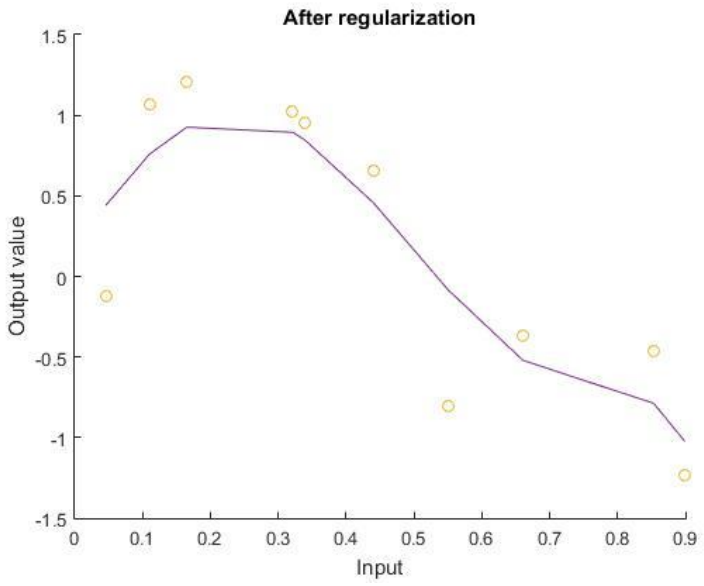

Figure 9 M = 10 and after regularization with lambda = 0.0001

After regularization the models looks better compared to before regularization.