University of Burgundy

Masters in Computer Vision and Robotics

SSI Pattern Recognition

Homework 1: Linear Regression

by

Vamshi Kodipaka

Supervisor: Dr. Desire Sidibe

1. Algebra : Matrix

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

let a, b be column vectors,

from definition,

Method-1:

Let u $= b^T a = \sum_{i=1}^{n} b_i a_i$, then

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

$$So, \quad \frac{\partial u}{\partial a} = b^T$$

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

Hence proved

Method-2:

Let u $= b^T a = b_1 a_1 + b_2 a_2 + \ldots\ldots + b_n a_n$. $Then \frac{\partial u}{\partial a_i} = b_i$

$So, \frac{\partial u}{\partial a} = b^T$ (As, per Numerator Layout Notation)

Check:

$$Let \quad a = \begin{bmatrix} x \\ \vdots \\ y \end{bmatrix} \qquad b = \begin{bmatrix} z \\ \vdots \\ w \end{bmatrix} \qquad b^T = [\, z \quad w \,] \qquad b^T a = [xz + yw]$$

$$\frac{\partial(b^T a)}{\partial a} = \begin{bmatrix} \frac{\partial\ (xz + yw)}{\partial x} \\ \vdots \\ \frac{\partial\ (xy + yw)}{\partial y} \end{bmatrix} = [z \quad w] = b^T \quad (Using\ Numerator\ Layout\ Notation)$$

Else we can also write,

$$\frac{\partial(b^T a)}{\partial a} = \begin{bmatrix} Z \\ \vdots \\ W \end{bmatrix} = b \qquad (Using\ Denominator\ Layout\ Notation)\ .$$

It is just matter of notation

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

Let A ∈ $R^{m\,x\,n}$ and $a \in R^n$.

Method-1:

$$Aa \begin{bmatrix} A_{11} & \dots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{n1} & \dots & A_{nn} \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} A_{11}a_1 + & \dots & + A_{1n}a_n \\ \vdots & \ddots & \vdots \\ A_{n1}a_1 + & \dots & + A_{nn}a_n \end{bmatrix}$$

$$\frac{\partial(Aa)}{\partial a} = \begin{bmatrix} \frac{\partial(A_{11}^T a_1)}{\partial a_1} & \dots\dots \\ \vdots & \\ \frac{\partial(A_{n1}^T a_1)}{\partial a_n} & \dots\dots \end{bmatrix}$$

from above proof, we can write the above as,

$$\frac{\partial(Aa)}{\partial a} = \begin{bmatrix} A_1^T \\ \vdots \\ A_n^T \end{bmatrix} = A \qquad (representing \text{ and } A_{1T}, \dots, A_{nT} \text{ be the rows of A.})$$

Hence proved.

Method-2:

Let u =$A^T a$ = $\sum_{j=1} A_{ij} a_j$, $\quad and\ \frac{\partial(A_i)}{\partial a_j} = a_{ij}.\quad So, \frac{\partial u}{\partial a_j} = A$

Hence proved

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

from product rule,

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

Since we know that:

$$\frac{\partial \mathbf{u}^\top \mathbf{v}}{\partial \mathbf{x}} = \mathbf{u}^\top \frac{\partial \mathbf{v}}{\partial \mathbf{x}} + \mathbf{v}^\top \frac{\partial \mathbf{u}}{\partial \mathbf{x}} \quad \text{(product rule)}$$

where $\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$ and $\frac{\partial \mathbf{v}}{\partial \mathbf{x}}$ are in numerator layout.

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

From **1.1 and 1.2** solved above; applying it first part of the sum,

We obtain $a^T A + (a^T A^T) = a^T A + (Aa)^T$

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

Hence proved.

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

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

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

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

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

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

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

Hence proved

Note : **tr(AB)** *is not equal to* ***tr(A).tr(B)***

1.5 Prove that $\boldsymbol{tr(ABC) = tr(CAB) = tr(BCA)}$

We know that trace is sum of diagonal elements of a matrix.

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

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

in the above equation the second and third terms are,

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

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

$$\text{so, } \boldsymbol{tr(ABC) = tr(CAB) = tr(BCA)}$$

Hence proved

2 Maximum Likelihood Estimate

2.1 We know that, the likelihood of the data given the parameters is given by

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

Let us assume, the parameters of a Gaussian distribution are the mean (μ) and variance (σ^2).

Lets suppose say, given observations $X_1,\ldots,X_n$, the likelihood of those observations for a certain μ and σ^2 (assuming that the observations came from a Gaussian distribution) is:

$$\boldsymbol{P(X_1,\ldots,X_n|\mu,\sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-1/2\left(\frac{X_i-\mu}{\sigma}\right)^2}}$$

and, if the data is identically independent variables (i.i.d) of $X_1,\ldots,X_n$ then,

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

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

as data is i.i.d , we can write

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

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

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

So, likelihood function can be written as a product of Gaussians.

2.2.1 The log likelihood function

Using above problem's result:

now, taking log on both sides, then log likelihood estimate is as,

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

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

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

Effect of Regularization:

$$\tilde{l}(D;\mathrm{w}) \;=\; \overbrace{\sum_{i=1}^{n} \log P(y_i|\mathrm{x}_i,\mathrm{w})}^{\text{log-likelihood}} \overbrace{-\frac{1}{2\sigma^2}(w_1^2+w_2^2)}^{\text{log-prior}} + \text{const.}$$

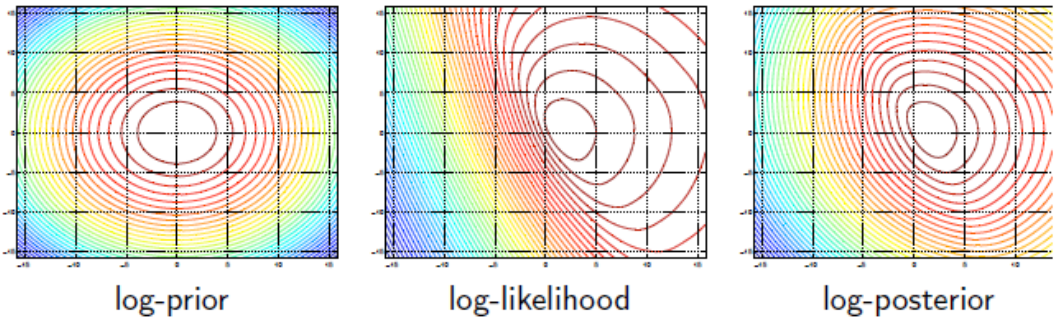


Figure a. Variations in log-prior, log-likelihood and log-posterior

$$\tilde{l}(D;\mathbf{w}) = \sum_{i=1}^{n} \log P(y_i|\mathbf{x}_i,\mathbf{w}) - \frac{1}{2\sigma^2} w_1^2 + \text{const.}$$

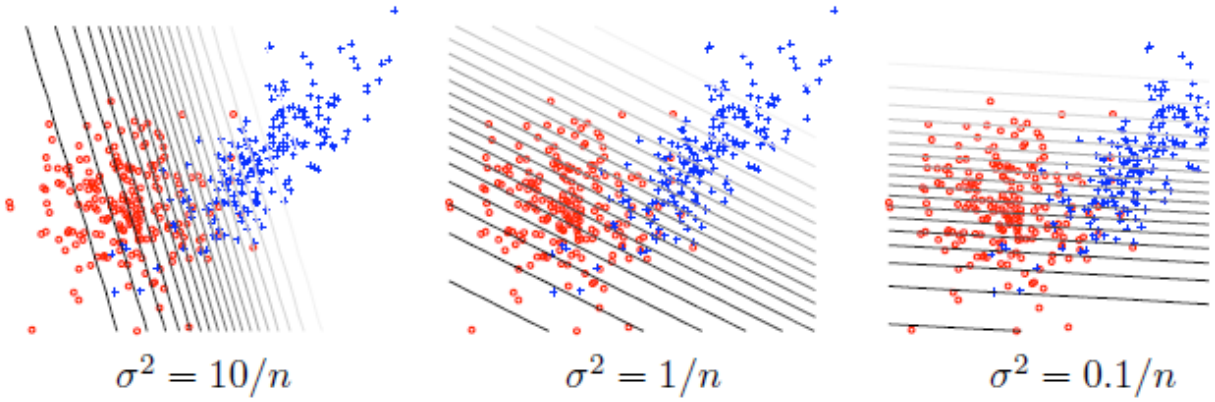


Figure b. Variations of log-prior with variations in variance

2.2.2 for mean and variance, so we take partial derivate the above equation with respect to μ *and* σ^2 and equate the results to zero.

So to get mean,

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

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

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

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

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

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

and now to get variance,

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

We take log likelihood estimate that we calculated above,

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

Now, by partial derivative with respect to σ^2

we get,

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

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

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

$$\boldsymbol{\sigma^2_{ML}} = \frac{\mathbf{1}}{\boldsymbol{N}}\sum(\boldsymbol{X_i} - \boldsymbol{\mu})^2$$

3 Linear Regression with Regularization

Linear Regression:

It is modelling the system with linear relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). If it is one explanatory variable, it is called simple linear regression.

.For more than one explanatory variable, the process is called 'multiple linear regression'. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

3.1 The code for linear regression with regularization is attached with the report and a README.txt file is included. A well commented code and documentation is done for future reference.

3.2 Different values of M to fit the given data. The results are as shown below:

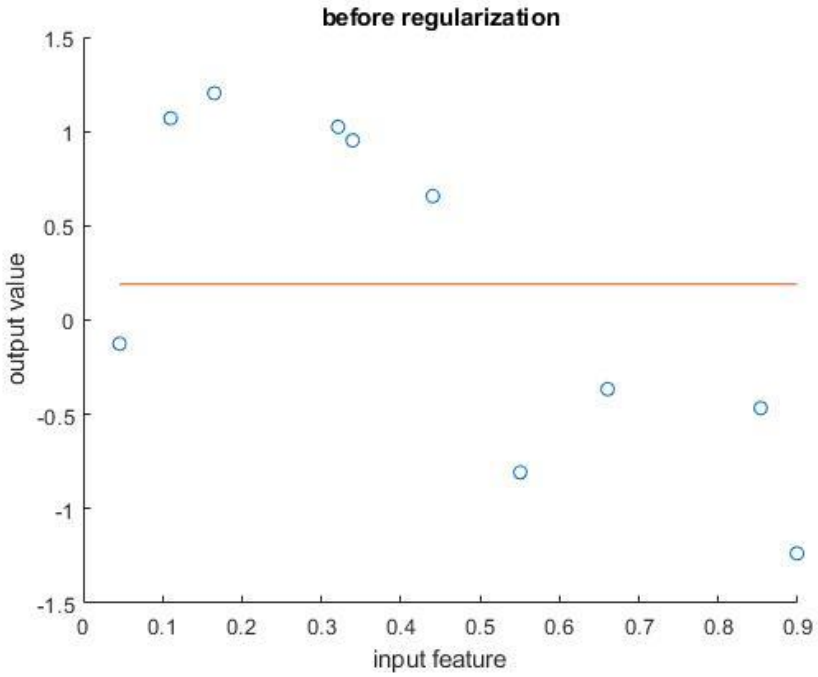


Figure 1 Curve fitting for M = 0

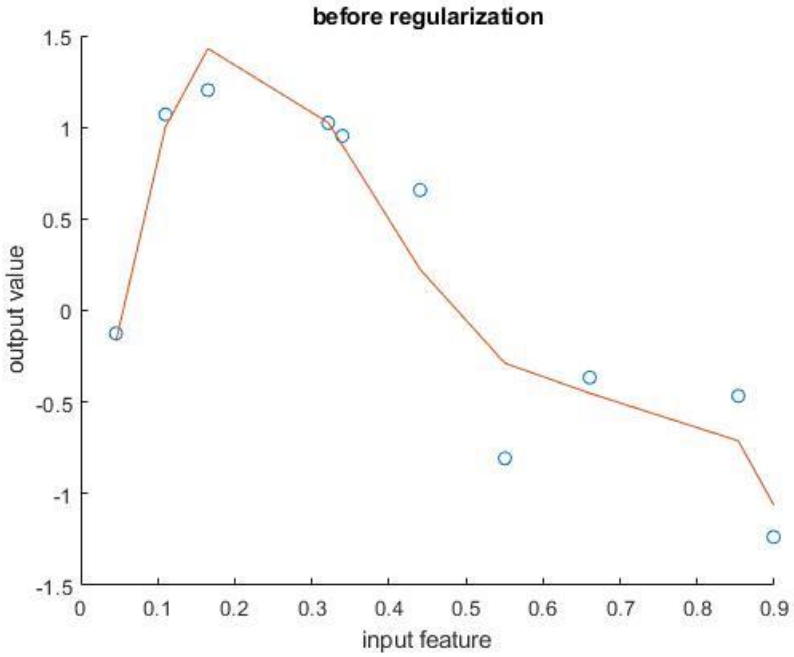


Figure 2 Curve fitting for M = 4

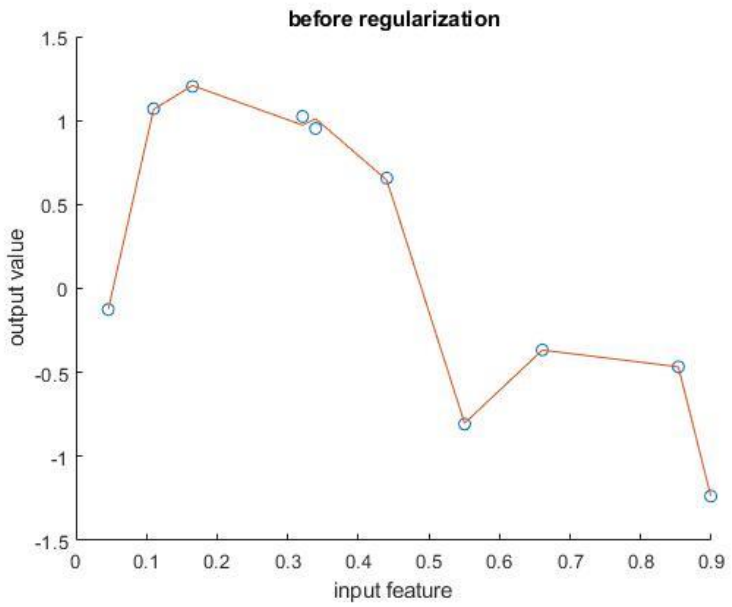


Figure 3 Curve fitting for M = 8

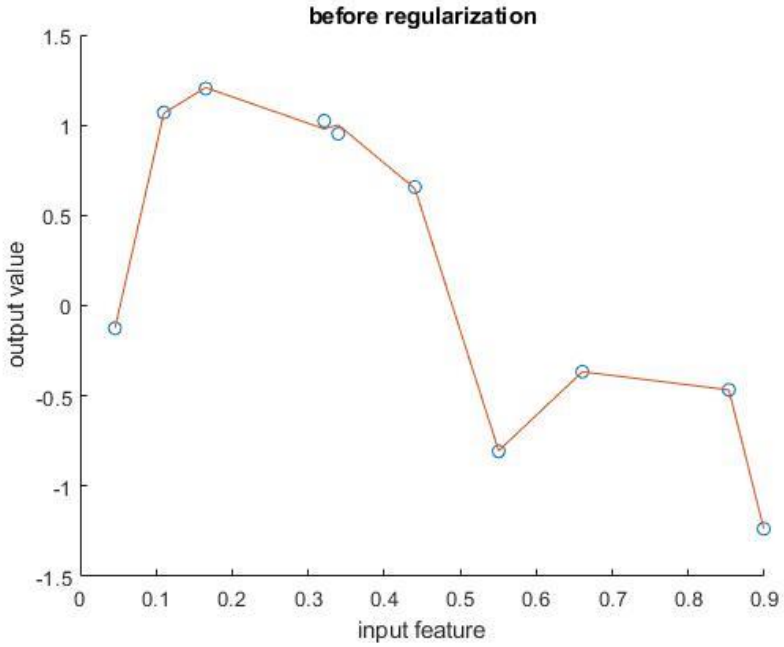


Figure 4 The result for M = 10

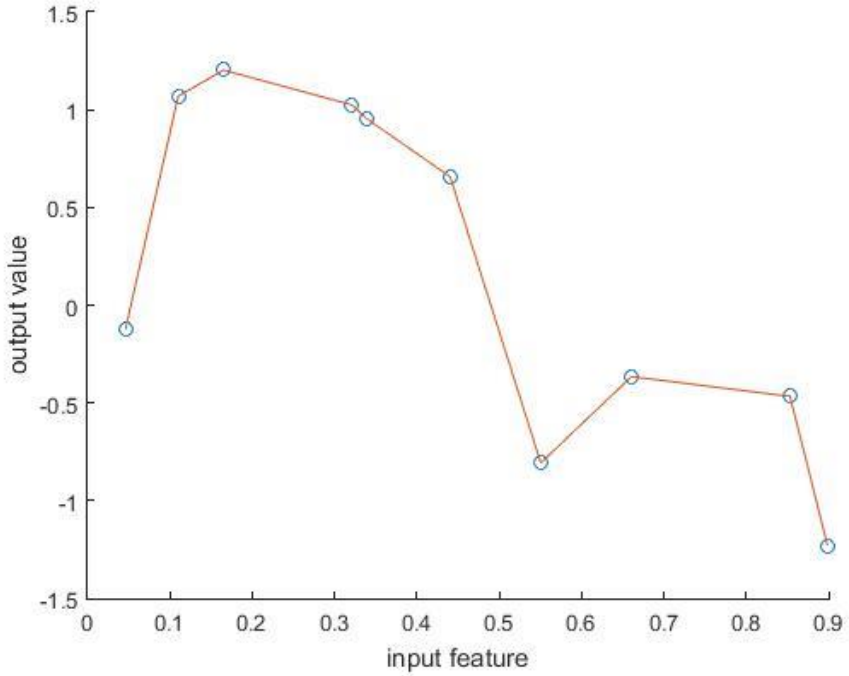


Figure 5 Curve fitting for M = 10

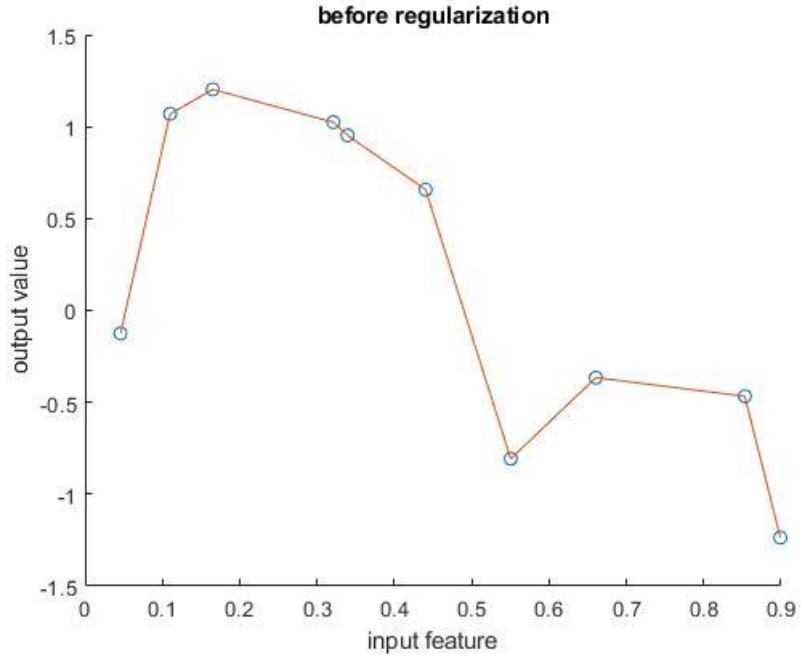


Figure 6 Curve fitting for M = 12

Observation: The more is the value of M the better we observe the curve fitting the data

3.3 To choose M:

We consider, the training and test error for different M values as shown in figure below.

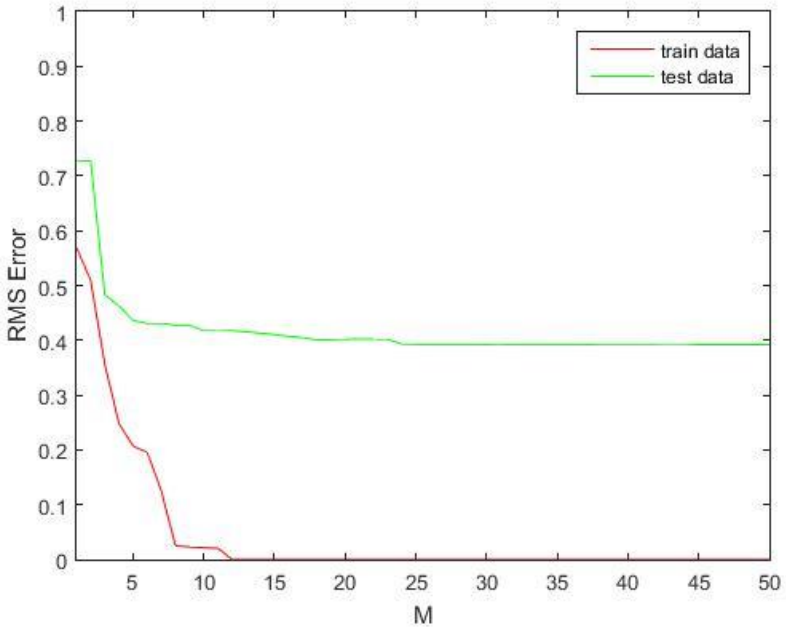


Figure 5 RMS Error for training and test data for different values of M

From the above Fig: For M = 12 has less training error and test error.

Reason for Regularization: *If M is high, gives high normalization values which increases the model capacity(but also has high complexity to build such model).*

3.4 To fit a model of order M = 10 to the data.

Here, we observe the value of M is high but we have few data points, so if we want to reduce the capacity of model we need to regularize the model.

Regularization: Process of adding a bias to cost function which regulates the complex models.

Regression with regularization term is referred as **Ridge Regression**.

The regularization bias is said to be **weighted decay**.

So, we modify 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 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 a good regularization parameter.

We check for different values of lambda which would lead to *less norm of parameter to estimate value by compensating with the fitting error.*

The below plots show the RMS error and the value of norm of parameter estimate, of both training and test data for different values of lambda.

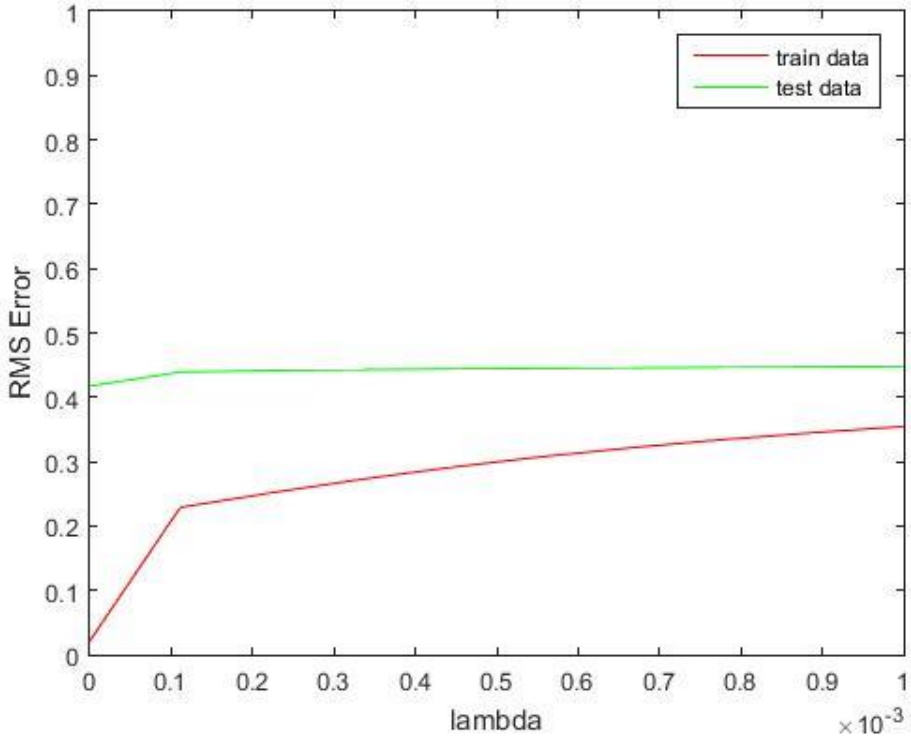


Figure 6 RMS Error of training and test data for different values of lambda.

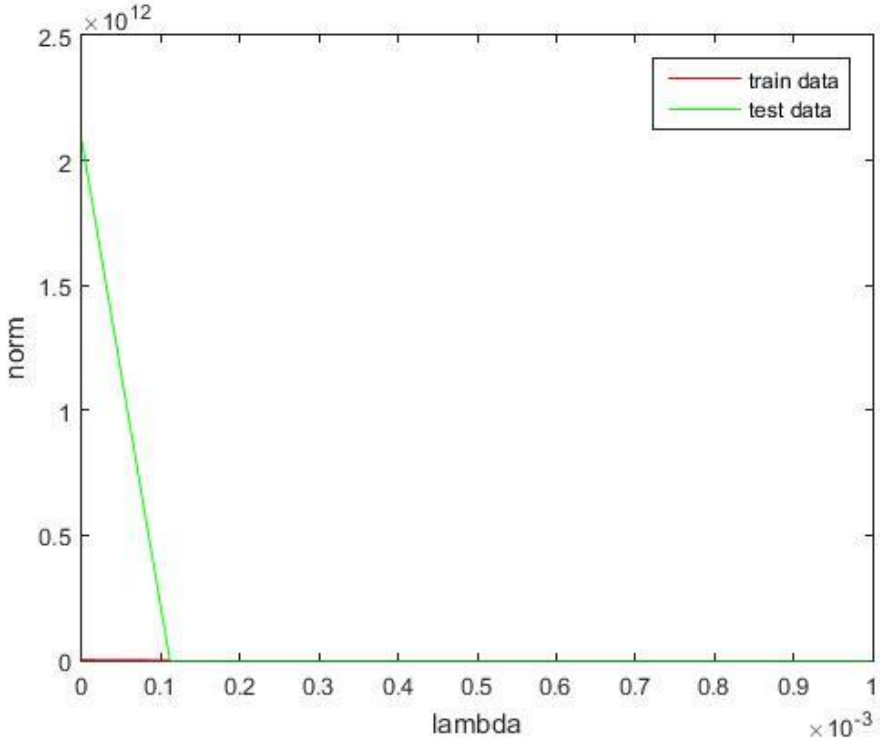
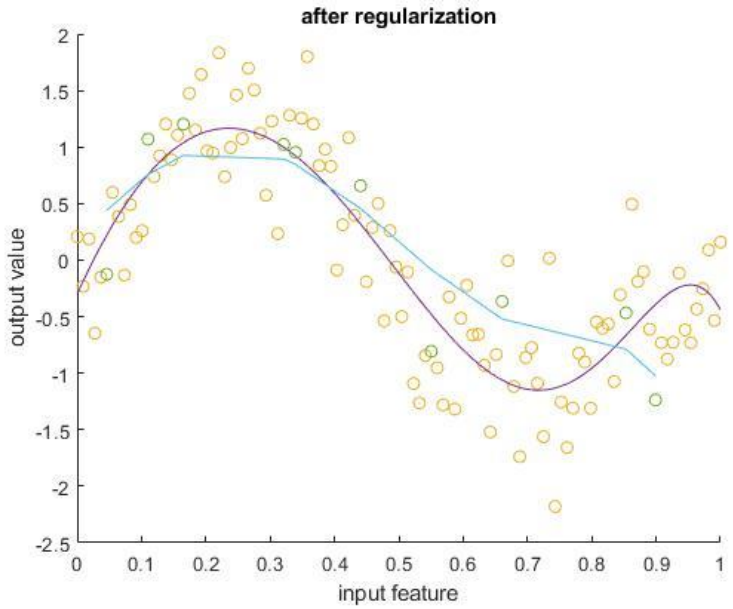


Figure 7 Norm of parameter estimate of training and test data for different values of lambda.



From the above results we can infer that,

Table 1. Error and norm values for different data types and regularization parameters

Lambda value	RMS Error	Norm	Data Type
$\lambda = 0$	0.0214	1.6e9	Training

	0.4174	2e12	Test
$\lambda = 0.0001$	0.3545	486	Training
	0.4478	1304	Test

So, a small regularization value will drastically reduce the norm value and at the cost of fitting error. So there is always a *trade-off between regularization value and fitting error* to reduce the complexity of the model.

So, with the value of $lambda = 0.0001$, the data is fitted and is as shown below, before (without regularization) and after regularization.

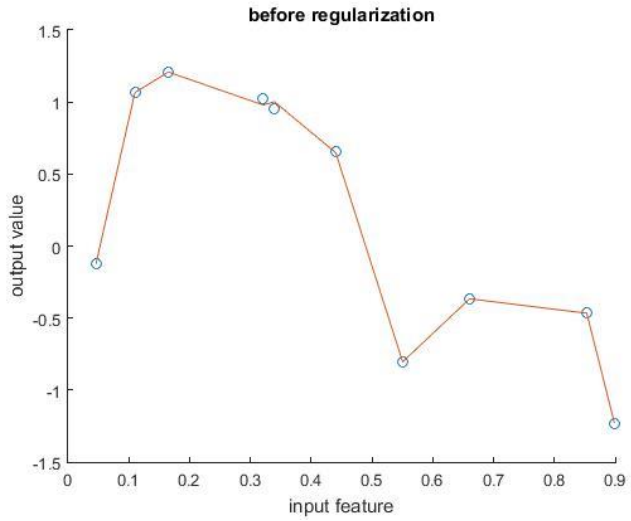


Figure 8 Result for M = 10 and before regularization

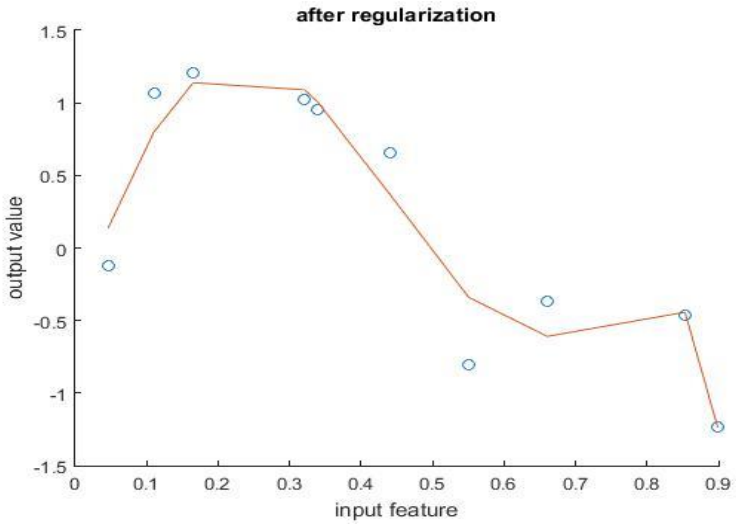


Figure 9 Result for M = 10 and after regularization with lambda = 0.0001

From above details, we observe that *regularization plays very significant role on reducing the model complexity at a cost of increase of fitting error.*

4.1 (Additional)

Show that maximizing the likelihood function, to find w, is equivalent to minimizing the sum-of squares error function as defined in class.

from 2.2.2 above we get,

$$ln\, p(t|w,\beta) = \frac{N}{2} ln(\beta) - \frac{N}{2} ln(2\pi) - \beta E_D(w) \text{ where}$$

$$E_D(w) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - w^T\phi(x_n)\}^2$$

Similarly,

4.2 For Posterior distribution::

The MLE with Regularization is given as,

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

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

Using Bayes Rule,

$$p(w|t) = p(t|X,w,\beta).p(w|\alpha) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n,w),\beta^{-1}).\mathcal{N}(w|0,\alpha^{-1}I)$$

$$ln\, p(w|t) = -\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - w^T\phi(x_n)\}^2 - \frac{\alpha}{2}w^Tw + const$$

From above observations, thus Maximization of Posterior is equivalent to minimization of the sum-of-squares

And we estimate *w* by:: $E(w) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - w^T\phi(x_n)\}^2 - \frac{\lambda}{2}w^Tw$

4.2 Value of λ

with addition of quadratic regularization term w^Tw :: with $\lambda = \alpha/\beta$

REFERENCES:

Questions on Matrix Algebra: 1.1 to 1.3

1. Matrix Differentiation
   https://www.comp.nus.edu.sg/~cs5240/lecture/matrix-differentiation.pdf

2. J. E. Gentle, Matrix Algebra: Theory, Computations, and Applications in Statistics, Springer, 2007.

   http://pws.npru.ac.th/sartthong/data/files/Matrix%20Algebra%20theory%20computations%20and%20applications%20in%20statistics.pdf

3. K. B. Petersen and M. S. Pedersen, The Matrix Cookbook, 2012
   http://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

4. Matrix Calculus https://en.wikipedia.org/wiki/Matrix_calculus

5. Review of Matrix Algebra for Regression A.Colin Cameron
   http://cameron.econ.ucdavis.edu/e240a/matrixalgebra.pdf

6. H. Lutkepohl, Handbook of Matrices, John Wiley & Sons, 1996

Questions on Matrix Algebra: 1.4 and 1.5

7. Properties of the Trace and Matrix Derivatives ::
   https://web.stanford.edu/~jduchi/projects/matrix_prop.pdf
8. Matrix Differentiation – NUS Computing
   https://www.comp.nus.edu.sg/~cs5240/lecture/matrix-differentiation-c.pdf
9. Matrix Calculus – Notes on the Derivatives of the Trace∷
   http://cal.cs.illinois.edu/~johannes/research/matrix%20calculus.pdf

Questions on Probability (MLE): 2.1 and 2.2

10. Estimation and Multivariate Gaussians by Shubhendu Trivedi::
    https://ttic.uchicago.edu/~shubhendu/Slides/Estimation.pdf
11. Maximum Likelihood Estimation::
    http://www.cs.princeton.edu/courses/archive/spr08/cos424/scribe_notes/0214.pdf

Questions on Linear Regression with Regularization: Q3

12. Dr. Desire Sidibe's Class Notes
13. Wikipedia :: https://en.wikipedia.org/wiki/Linear_regression
14. MIT Machine learning: lecture 6 Tommi S. Jaakkola::
    http://www.ai.mit.edu/courses/6.867-f04/lectures/lecture-6-ho.pdf

Questions on Maximum a posteriori : Q4

15. Bayesian Interpretations of Regularization Charlie Frogner
http://www.mit.edu/~9.520/spring09/Classes/class15-bayes.pdf
16. Bayesian Linear Regression by Sargur Srihari
https://cedar.buffalo.edu/~srihari/CSE574/Chap3/3.4-BayesianRegression.pdf
17. Parameter Estimation in Probabilistic Models, Linear Regression and Logistic Regression Piyush Rai :: https://www.cs.utah.edu/~piyush/teaching/20-9-print.pdf
18. Probability Theory and Parameter Estimation II ::
https://dcc.ufrj.br/~sadoc/machinelearning/mlslides4-sadoc.pdf