Kernel Methods for Pattern Analysis
Assignment 1
Function Approximation

Jilt SebastianCS13D020
Jom KuriakoseCS13S028
Abil N GeorgeCS13M002

March 7, 2014

Contents

Part I
Linear Basis Function Models

Linear models for regression is a linear combination of functions of input variables which may or may not be linear.These are used for approximating functions by representing the approximation function y(x,w) as product of basis function and weight vector w.

0.1 Polynomial Curve Fitting

This is a simple regression problem in which the basis function to be approximated is of Polynomial form.We have a real valued input variable x and the function f(x) as the corresponding output variable.Our task is to predict the value of the target,t,which is equal to f(x) from the input variable x.We have a training data set consisting of N samples of input variable x and its corresponding target variables.For the functions of polynomial form, we can express the target variable as a linear combination of basis functions and weights where basis function,$Phi_j(x) = x^j, M = 0, 1..M, M$ being the model complexity.Here M is called as the order of the polynomial.Values of the weights ultimately determine the value of output variable.Hence output is a linear function of w even if it is a nonlinear function of input variable and thus it is called as a linear model.

The coefficients w is obtained by fitting the given polynomial to the training data by reducing the error function.Error function is a measure of dissimilarity or misfit between actual function and the approximated function.

The error function is given by,

$E(w) = \frac{1}{2}\sum\{y(x_n, w) - t_n\}^2$

Our aim is, hence to find out the values of w which will minimize the value of this error function.Selection of the appropriate model parameter M is empirical.Very small values of M gives poor fitting and very large values of M causes overfit, a case when even if the error is minimum, the function oscillates between successive data points.This is because as the M increases, the weight parameters becomes larger with positive and negative values in it.

0.1.1 Curve Fitting with Regularization

Regularization is used to control the effects of overfitting where the error function is modified by a term which prevents weight values from reaching larger values.The modified error function with regularization is given by,

The effect of regularization is controlled by the term Lamda.This regularization parameter is adjusted empirically to gain maximum approximation to the original function.

$E(w) = \frac{1}{2}\sum\{y(x_n, w) - t_n\}^2 + \frac{\text{Lamda}}{2}||w||^2$

0.2 Gaussian Basis function

The issue with polynomial basis function is it's target value is an entire function of input variable which causes it to have changes in one region causing a change in the other region.Gaussian basis function consists of mean values which determines the center of basis functions of the input variable x.The basis functions are multiplied by the weight vector to get the approximated output for the given input variables.Scale factors in gaussian functions are not important as the weight values takes care of these differences.

Gaussian Basis function is of the form,

$Phi_j(x) = exp\{\frac{-(x-u_j)^2}{s}\}$

0.3 Bias-Variance Trade Decomposition

For the function approximation task,we define loss function which is the expectation of the error for all samples of the training data.Average value of the loss function is the sum of noise factor, bias and variance.Bias is independent of the target and it depends on the difference between original output and mean of output obtained.Variance represents the differnence of obtained output and mean of it.These are terms in the average loss function.

Hence ,

Bias^2= $\frac{1}{N}\sum\{y(x_n) - h(x_n)\}^2$

variance $= \frac{1}{N}\sum\frac{1}{L}\sum\{y^l(x_n) - y(x_n)\}^2$

where,h(x_n) being the conditional expectation denoting the squared loss function.

For smaller values of model complexity,the term Bias^2 gives a high value and it is giving lower values as model complexity increases.But,behaviour of Variance vs Model complexity is exactly reverse i,e It has higher values at higher Model complexity.We need a model such that it is giving minimum for both bias^{2}and Variance.Thus there is a trade off between bias and variance values which will give us the optimum complexity.Similar to this,we have Bias^{2}and variance curve versus Lamda, the regularization parameter,which also has the same characteristic curves.

Part II
Experiments

1 Dataset 1(Univariate)

For the polynomial curve fitting with given data set, we were required to generate the function

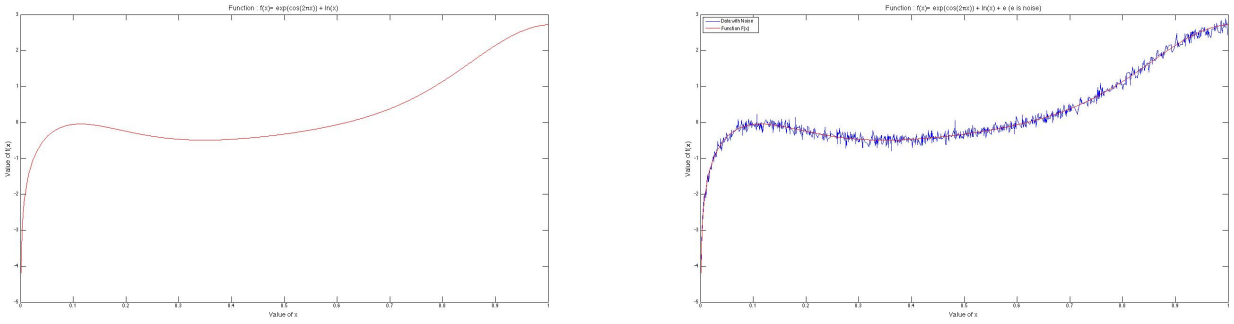
$f(x) = exp(cos(2 * pi * x)) + ln(x)$

1.1 Procedure

1. Input variable x is taken from the uniform distribution and noise is added from a gaussian distribution with mean 0 and variance 0.1.
2. Suitable number of points are taken which could yield good approximations.We have generated datasets of sizes 10,100 and 1000 samples and it is taken as 3 different training datasets.
3. Two datasets of size 200 each are generated for validation dataset and test dataset.
4. Weight values are computed from the design matrix and these values are used to obtain approximated function.The squared error is also found out for the particular model complexity.By varying value of M, we could get different approximations of the given function.

1.1.1 Plot of underlying function and training data

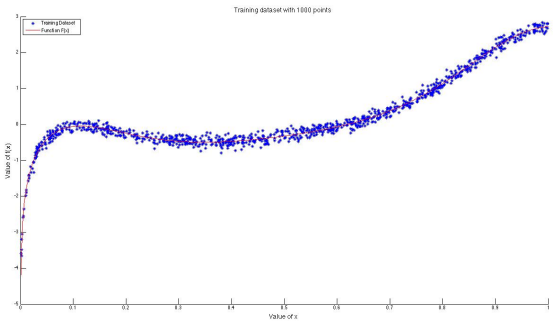
Below plots reprecents the actual underlying function (a) and function after adding the noise (b).This is the figure for a training samples of size 1000.Training samples of 10 points are randomly selected from this as shown in figure2.



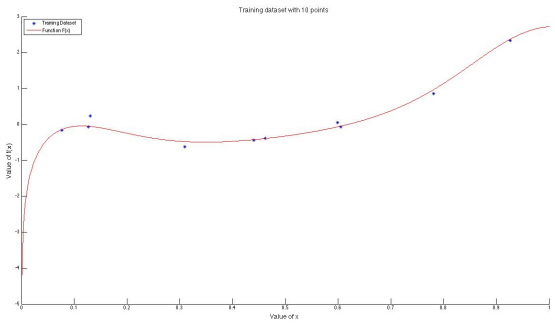
(a) Actual Function

(b) Training data(noisy realization)

Figure 1: Actual function and Noisy realization



(a) with 1000 points



(b) with 10 points

Figure 2: Training dataset of different sizes

For the approximated function,further smoothening of the weight values and thereby reduction of error is obtained by the use of regularization with regularization parameter lambda being varied between 0 and 1.It helps in reducing the overfitting by preventing weight values from reaching very large values and causing oscillations.For different values of this parameter,squared error between actual and approximated function is found out.The combination of parameters M,lambda are empirically chosen based on the experiment results.

2 Experiments and Observations on dataset 1

2.1 Experiment No.1

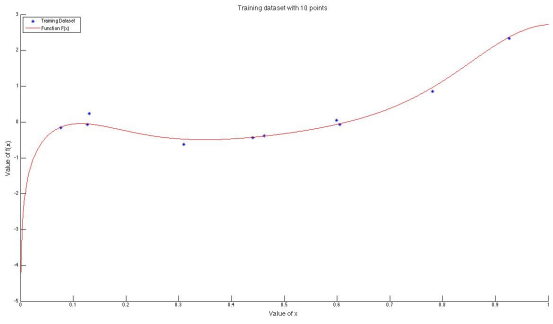
Experiment is conducted by changing the model complexity of polynomial basis function from 1 to 10 on the test dataset.Parameters of the model are found based on K-cross validation and this model is used to test the data.

2.1.1 Observations

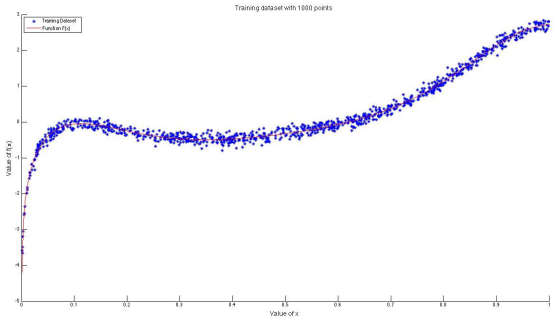
The best model complexity is selected by finding the complexity at which minimum error occurs for the test data.It is observed that as model complexity increases,the model is overfitted with the given training samples.Best Model complexity is found to be 7 for dataset of 10 points.

2.1.2 Plots comparing different Training data sets:

Below plots depicts the comparison of outputs with training sets of different sizes with a model complexity M=7.



(a) N=10

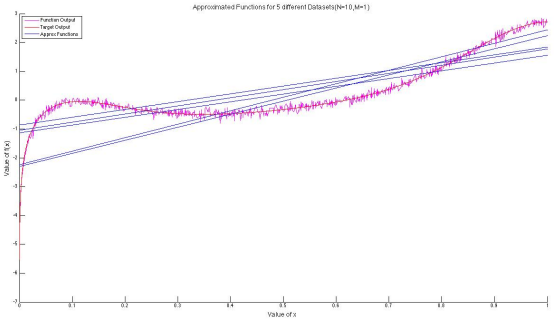


(b) N=1000

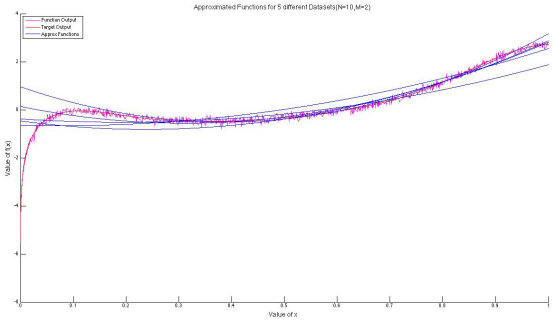
Figure 3: Plot comparing datasizes

2.1.3 Plots of Approximated function for different model complexities:

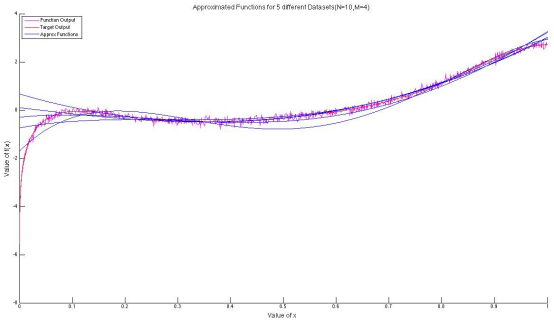
Following plot shows the approximated function for different model complexities.5 datasets of size 10 is used for plots.Changes with model complexity are as follows,



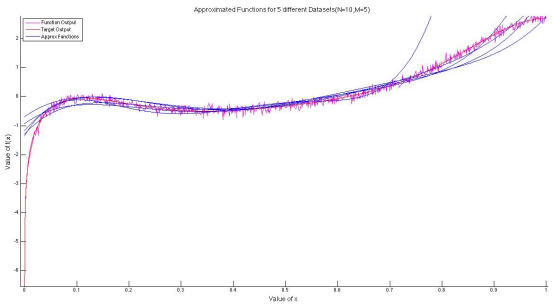
(a) For M=1



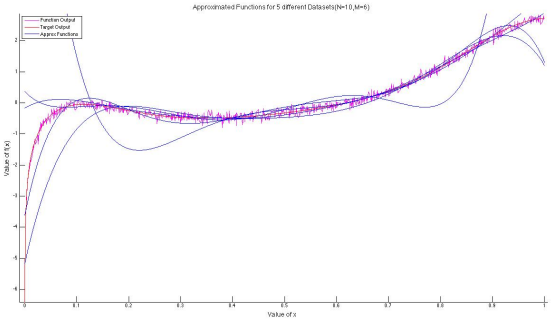
(b) For M=2



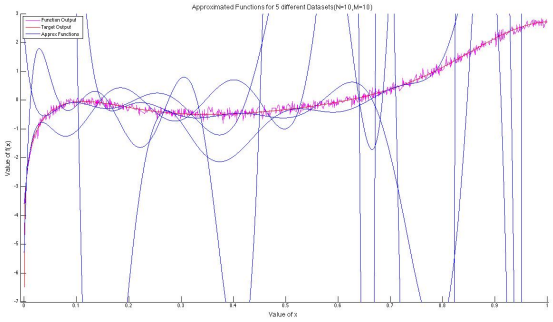
(c) For M=4



(d) For M=5



(e) For M=6



(f) For M=10

2.1.4 Observations

- For smaller values of M ,the approximation is poor.As the model complexity increases,error reduces and fit becomes more and more proper.
- When model complexity is very large,even if error reduces,oscillations are produced as shown .It is called overfitting
- Effect of overfitting increases if we increase M further.

2.1.5 Table showing weight parameters:

Below Table shows the values of weights of the model obtained using training samples of size 10.

W vs M	M = 0	M = 2	M = 4	M = 7	M = 10
w0*	0.26	-0.37	-1.67	-3.09	-3.47
w1*		-2.62	19.80	78.98	128.94
w2*		5.758	-79.52	-705.37	-1991.41
w3*			114.54	2904.80	16446.39
w4*			-50.45	-6338.38	-81409.90
w5*				7558.41	253442.74
w6*				-4632.38	-506367.11
w7*				1139.84	647105.51
w8*					-510786.29
w9*					226696.22
w10*					-43259.12

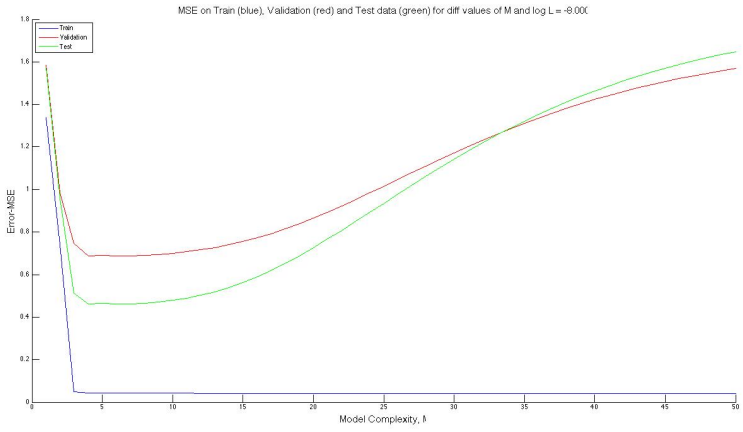
2.1.6 Inferences

From the table we can see that when the model complexity increases,the training data is overfitted by the approximated function. For a model complexity of 10, the error is negligible but error on test and validation dataset sharply rises because of overfitting. This is prominent if the size of dataset is smaller.

2.1.7 Plot of Mean Squared Error :

Following plot reprecents the change of mean squared error with respect to Model complexity for a dataset of size 10.

Figure 4: Plot of MSE vs Model complexity ,N=10



2.2 Experiment No.2

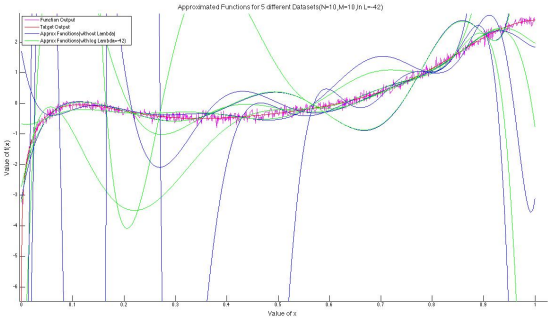
The problem of overfitting is takencare by introducing regularization parameter which changes the weight values and prevents it from going to higher amplitudes of opposite sign.The plots are done with a model complexity of 10.

2.2.1 Observations

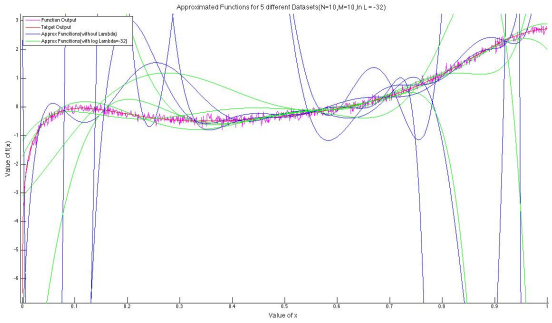
Best regularization parameter is found to be -21.This is obtaned by fining the minimum average error for the validation dataset for best values of M and lamda.Even for less number of sampleswe are getting a good approximation with the help of regularization parameter.

2.2.2 Plots of Approximated function for different Regularization Parameters

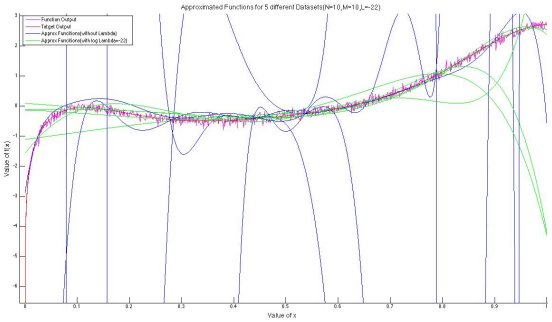
Below are the plots of approximated function for different regularization parameters.lamda is varied from zero to one to obtaon the below plots.



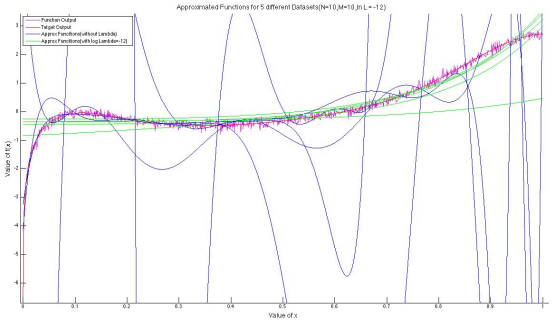
(a) For ln lamda=-42



(b) For ln lamda=-32



(c) For ln lamda=-22

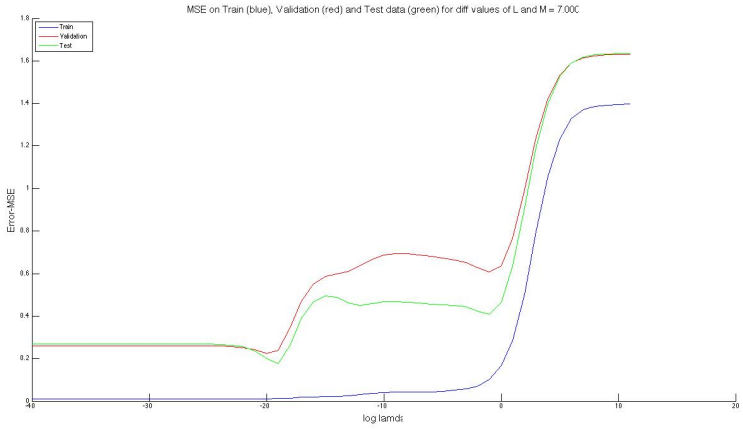


(d) For ln lamda=-12

2.2.3 Plot of Mean Squared Error vs Regularization Parameter:

Following is the MSE vs lamda plot.This is done for N=10 and M=7.

Figure 5: Plot of MSE vs Regularization Parameter ,N=10



2.2.4 Table showing weight with Regularization Parameter.

Below table shows the change in weight values for best M with changes in regularization parameter.

W vs ln lamda	ln lamda = $-\infty$	ln lamda = -15	ln lamda = -5	ln lamda= 0
w0*	-3.47	-2.61	-0.899	-0.6069
w1*	128.94	53.05	4.237	0.3163
w2*	-1991.41	-355.17	-8.812	0.4341
w3*	16446.39	958.82	1.764	0.6338
w4*	-81409.90	-980.09	4.736	0.6698
w5*	253442.74	-95.37	4.030	0.6042
w6*	-506367.11	534.15	2.370	0.4921
w7*	647105.51	267.06	0.713	0.3640
w8*	-510786.29	-252.93	-0.702	0.2358
w9*	226696.22	-369.21	-1.871	0.1149
w10*	-43259.12	245.01	-2.839	0.0046

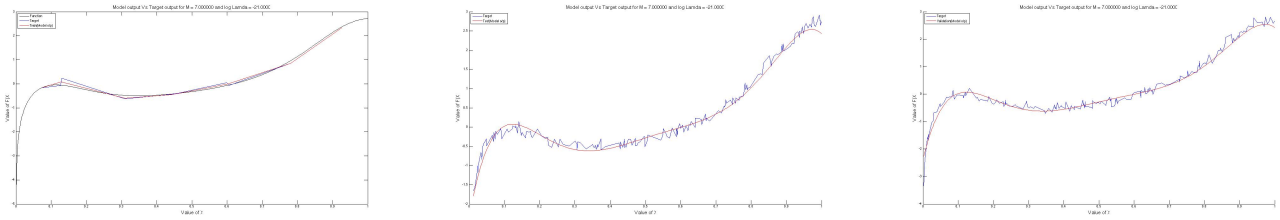
- It is observed from the table that, when regularization is not used, weight values are very high.When lamda increases, weight values decreases.

2.3 Experiment No.3

The experiment is done to generate the plots of model complexity with target output.For this we used datasets of different samples and for a model complexity of 10 and regularization parameter of -21.

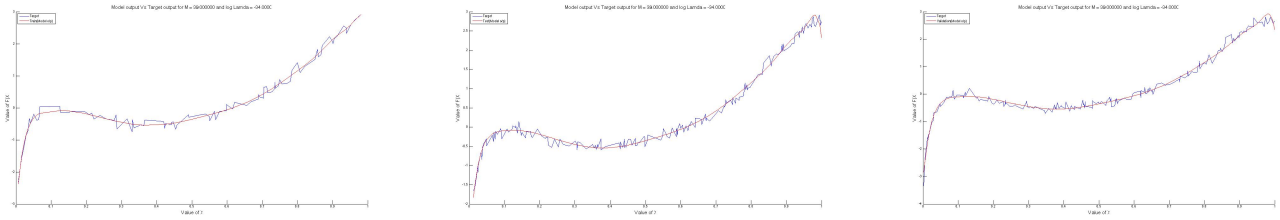
2.3.1 Plots of Target versus Model output

Below plots shows the comparison between the target and model output on how well they are approximated.



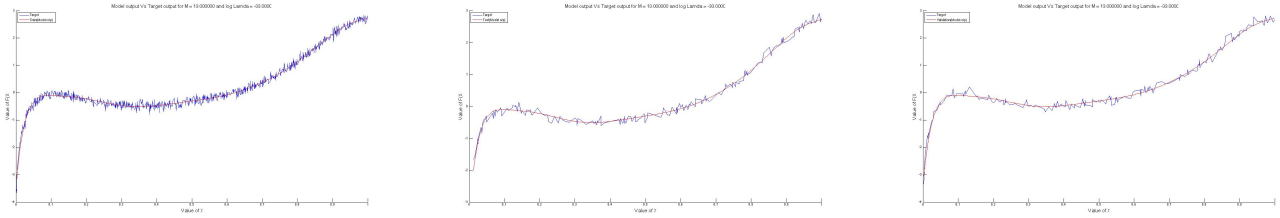
(a) Train data (b) Test data (d) Validation data

Figure 6: For N=10



(a) Train data (b) Test data (d) Validation data

Figure 7: For N=100



(a) Train data (b) Test data (d) Validation data

Figure 8: For N=1000

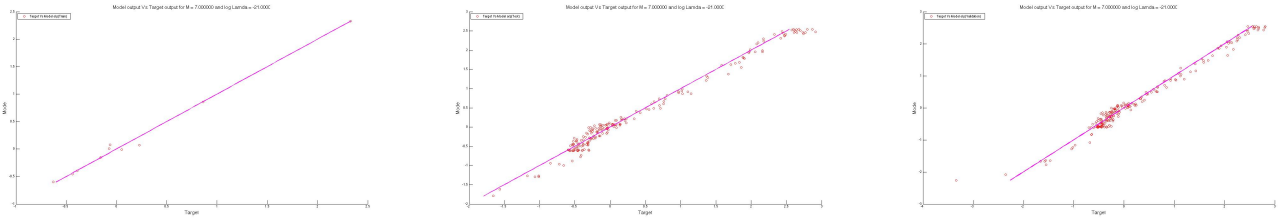
- If the number of points are less,then for higher value of model complexity, overfitting happens.
- For high value of M and with less data-points,Number of inflection points is M-2.Hence inflection points are seen to be more.This makes the curve "lumpy".

2.4 Experiment No.4:

In this experiment we are generating the scatter plot showing target vs model outputs.For this we used a dataset of N=10 and lamda=-21 for a model complexity of 10.

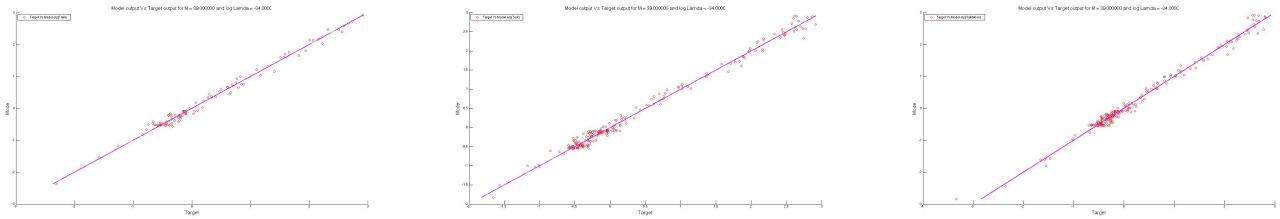
2.4.1 Scatter Plot of Target versus Model Output

Below scatter plots are done for different datasets to find how well the model is built.



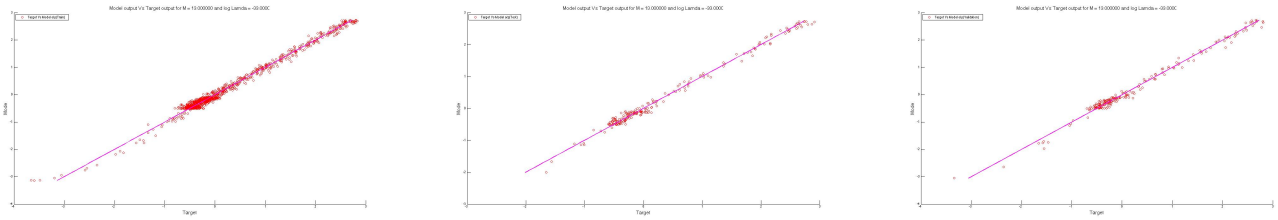
(a) Train data (b) Test data (d) Validation data

Figure 9: For N=10



(a) Train data (b) Test data (d) Validation data

Figure 10: For N=100



(a) Train data (b) Test data (d) Validation data

Figure 11: For N=1000

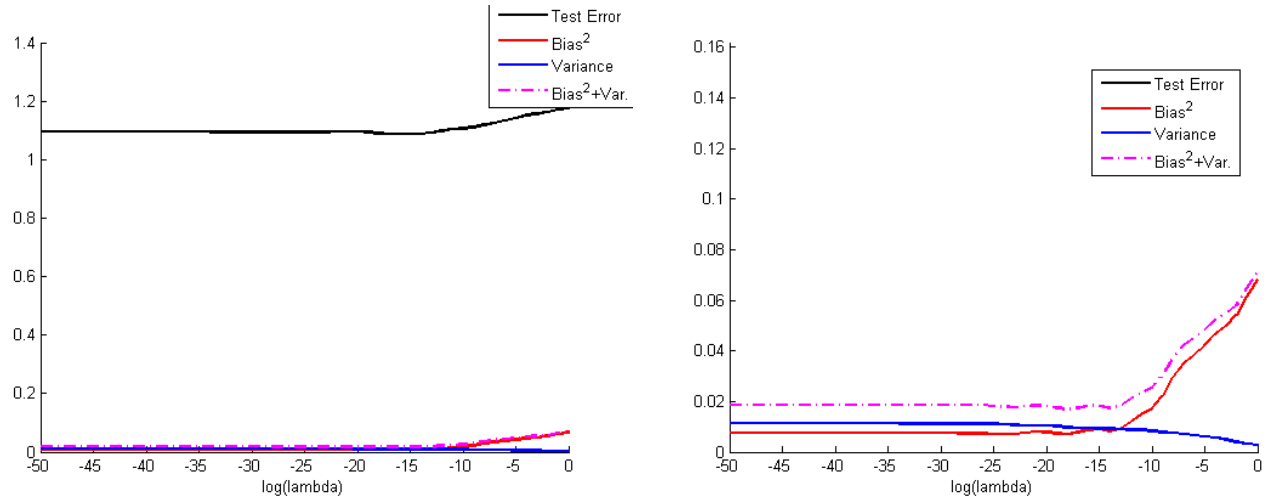
- Target vs output gives approximately same function as the actual function because the curve goes through y=x point with slope 45 degree w.r.to x axis.

2.5 Experiment No.5

In this experiment we are plotting the Bias variance trade off curve for different values of regularization parameter and different values of model complexity.

2.5.1 Bias^2 - Variance Plot : With Regularization Parameter

Below plot is bias variance curve for different values of lamda

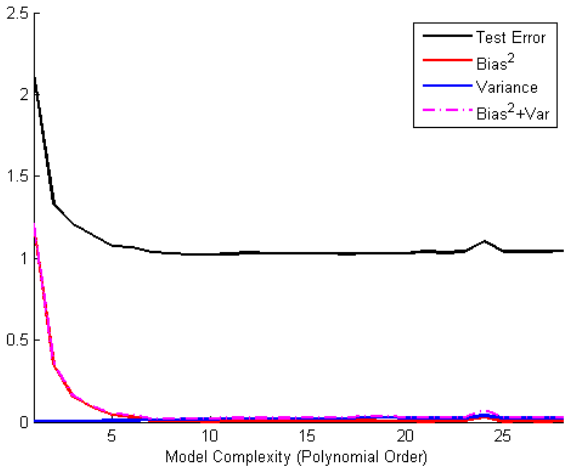


(c) Bias Variance plot with Different L

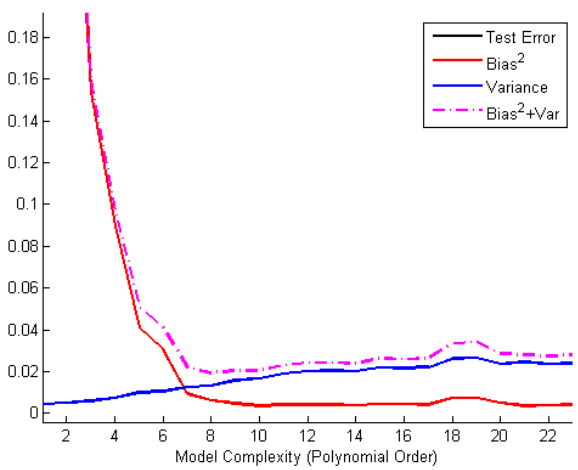
Figure 12: Bias Variance plot with Different L

2.5.2 Bias^2 - Variance Plot : With Model Complexity

Below plot is bias variance curve for different values ofModel complexity.



(b) With Model complexity



(e) With Model complexity

(e) Bias Variance plot with Different M

Figure 13: Bias Variance plot with Different M

- Minimum error point is the intersecting point between Bias and variance, at this value of lambda and model complexity, Bias and Variances are minimum.
- For the given problem,M~7 and Lambda=-13 provides the intersection point.

2.6 Inferences

- Overfitting occurs when M is large, it is because the model is trying to fit a simpler function with more complex functions and trying to reduce the error. Even if the error decreases,the weight values becomes higher positive and negative, causing oscillations between successive data points.
- As Model complexity increases,weights increases.With larger M error reduces in training data. However it decreases for test and validation datasets as it overfits and gives poor generalization.
- With regularization parameter, large weight values are penalized and hence the error reduces. For lower lambda, it will behave same as without regularization.With larger lambda values, error again increases because it penalizes on weight too much.

3 Dataset 2 (Bivariate Data)

Approximating the underlying function of the given training dataset for which the input variable is bivariate in nature using Gaussian Basis Function is done here.The true function is approximated by a linear combination of M gaussian functions of the training input variables where M is the model complexity.These M gaussian functions are centered around its cluster mean.Given dataset is a two dimensional input data.Train, Validation and Test datasets are provided with us for experiments.

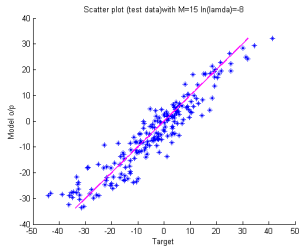
3.1 Procedure

1. Represent the data as M clusters using K means clustering.This is a discriminative approach for clustering the data-points.
2. The combination of M and lamda is taken for getting the best possible approximation.
3. Model parameters are computed on the validation dataset and Squared error is computed on the test dataset.

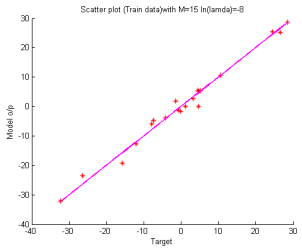
4 Experiments and Observations with Bivariate Data

4.0.1 Scatter Plot of Target versus Model Output

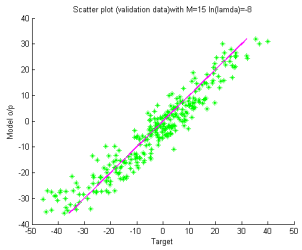
a).When Number of training data,N=20



(b) with N=20 on test data



(d) with N=20,on training data



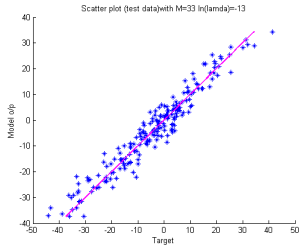
(f) with N=100 on validation data

Figure 14: Target versus Model Output

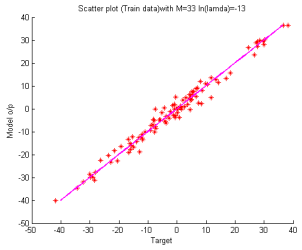
Results.

- Least Error on Validation Data = 7.654501e+03
- Least Erms value on Validation Data= 5.051238e+00
- Best M = 18
- log(lambda) = -9, lambda = 1.670170e-05
- Error on Test Data (for best model) = 5.370663e+03
- Erms on Test Data (for best model)= 5.182018e+00

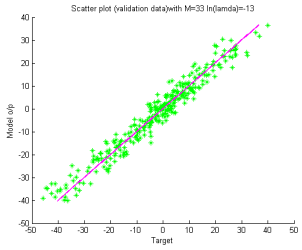
b).When N=100



(b) with N=100 on test data



(d) with N=100,on training data



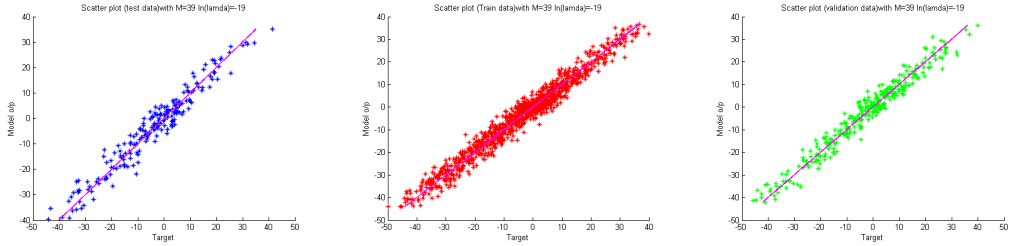
(f) with N=100 on validation data

Figure 15: Target versus Model Output

Results.

- Least Error on Validation Data= 4.188529e+03
- Least Erms value on Validation Data= 3.736544e+00
- Best M = 31
- log(lambda) = -10 lambda = 6.144212e-06
- Error on Test Data (for best model) = 3.361006e+03
- Erms on Test Data (for best model)= 4.099394e+00

c).When N=1000



(b) with N=1000 on test data (d) with N=1000,on training data (f) with N=1000 on validation data

Figure 16: Target versus Model Output

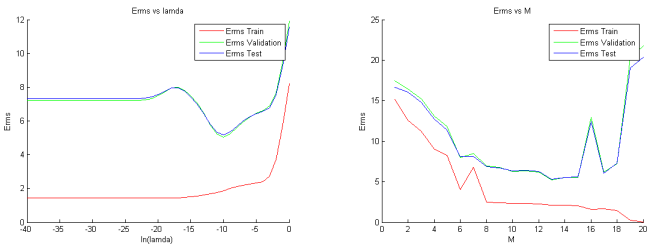
Results.

- Least Error on Validation Data = 3.652858e+03
- Least Erms on Validation Data= 3.489440e+00
- M = 34
- log(lambda) = -14 lambda = 1.125352e-07
- Error on Test Data (for best model) = 2.834502e+03
- Erms on Test Data (for best model)= 3.764640e+00

4.0.2 Observations

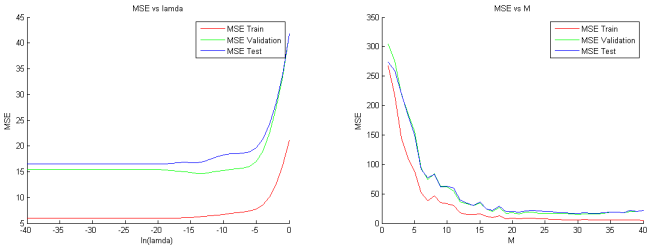
- When the number of data-points on training data set increases,the approximation is observed to be better.
- Bivariate data performs reasonably good approximation even with less number of data-points on validation data set.
- Best value of S is found to be ˜20.

4.0.3 Plot of RMS Error vs Regularization Parameter and Plot of RMS Error vs Model Complexity



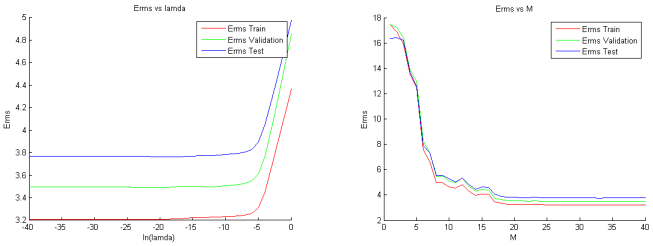
(b) ERMS vs lambda (d) ERMS vs M

Figure 17: When N=20



(b) ERMS vs lambda (d) ERMS vs M

Figure 18: When N=100

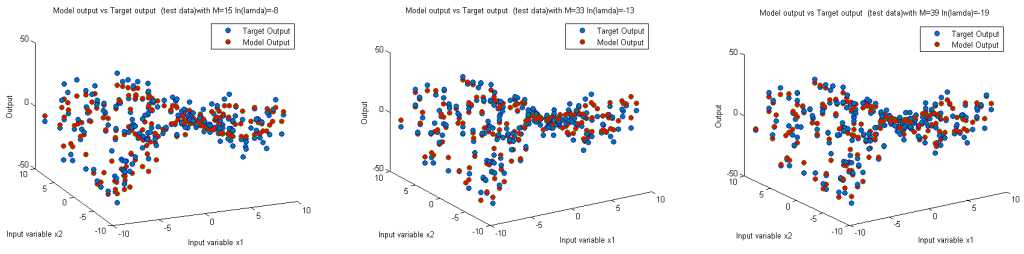


(b) ERMS vs lambda (d) ERMS vs M

Figure 19: When N=1000

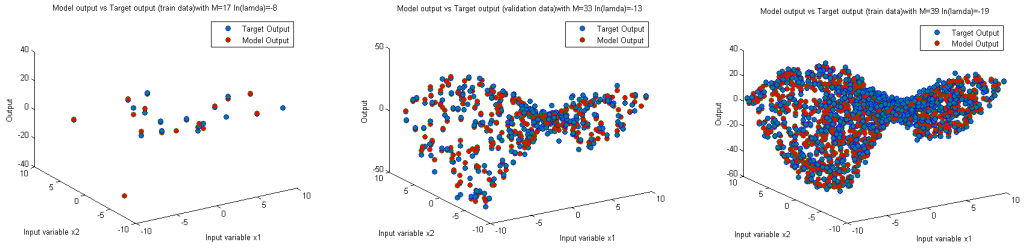
- RMS Error value decreases when larger number of data points are taken for training
- When training data is large ,error sharply increases for larger values of lambda.For N=100, error increases slowly as compared to N=1000.This is because, even if larger number of N gives better approximation, the effect of changes in weights are more for it.
- Validation data set performance increases with increase in training data points.
- When model complexity increases, training error decreases but validation and test data error increases leading to poor generalization at higher M values.This is more evident with less number of data points.
- The difference in errors for train and test data sets are lesser when we go for higher number of samples, as seen in ERMS vs M graphs.This is because with higher number of samples, approximation becomes better.

4.0.4 Realization of Approximation function



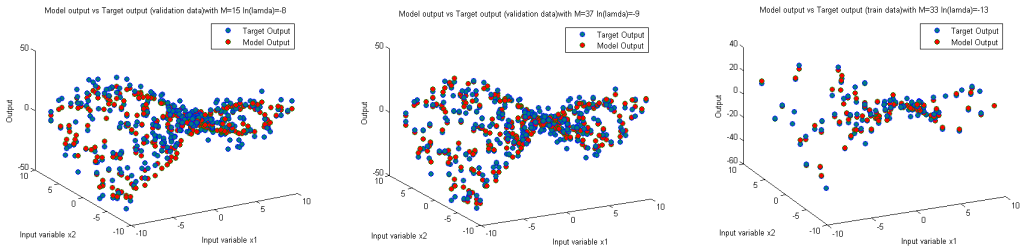
(b) with N=20 on test data (d) with N=100,on test data (f) with N=1000 on test data

Figure 20: Realization on test data



(b) with N=20 on training data (d) with N=100,on training data (f) with N=1000 on training data

Figure 21: Realization on training data



(b) with N=20 on validation data (d) with N=100,on validation data (f) with N=1000 on validation data

Figure 22: Realization on validation data

- Radial basis functions such as Gaussian function works based on the distance from a particular point to the mean of the function.The centers of the gaussians after K means clustering are these mean points.
- In Gaussian function, every point can be represented by a set of K gaussian function values where K is the number of clusters.Here it refers to model complexity.
- As the value of S changes,the cluster shapes changes.The model is trying to fit to the actual function in terms of it's cluster gaussians.

4.1 Inferences

- When the value of S varies,accuracy also varies.This is because the gaussian basis function varies with S,which in-turn changes the points in the same cluster.Hence it tries to fit differently to the approximated function.
- Training error decreases with complexity but validation and test data error increases because of overfitting leading to poor generalization ability.
- For the artificial data,the accuracy and fit is better as compared to real data.Given data gives model with very less error, because it is an artificially generated dataset.

5 Gaussian Curve fitting for Multivariate Data

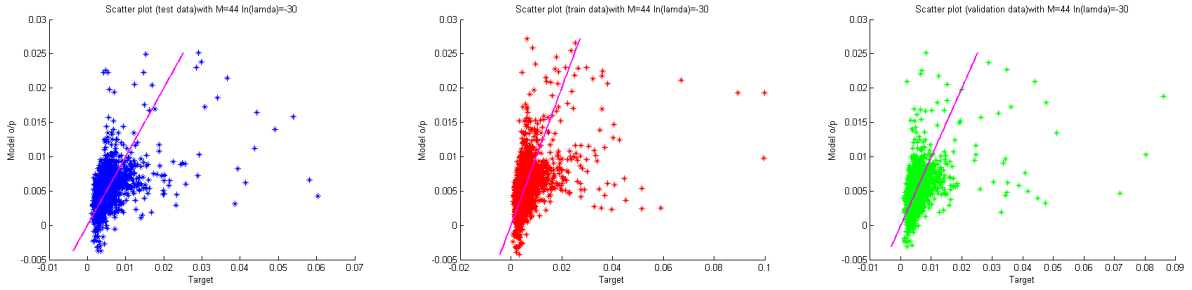
For the real world data consisting of multiple variables as the input.Hence visualization of data is not directly possible.The experiments are done in the same way as in Bivariate case

5.1 Results.

- Least Error on Validation Data = 4.015349e-02
- Least Erms on Validation Data= 5.173876e-03
- Best M = 40

- log(lambda) = -31 lambda = 4.658886e-15
- Error on Test Data (for best model) = 2.850782e-02
- Erms on Test Data (for best model)= 4.553346e-03

5.1.1 Scatter Plot of Target versus Model Output



(b) Target vs Model o/p for test data (d) Target vs Model o/p for train data (f) Target vs Model o/p for validation data

Figure 23: Target vs Model output

- In all cases it approximates target function well and the plot is nearly linear in nature.
- Value of S which gives best result is ~120

5.1.2 Plots of RMS Error

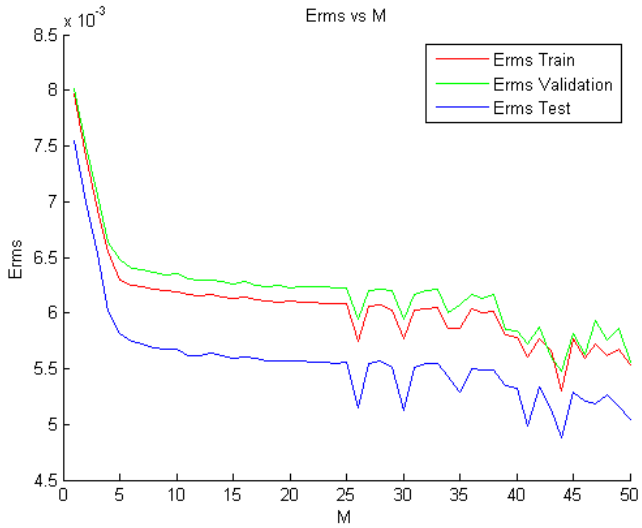


Figure 24: ERMS vs Model complexity

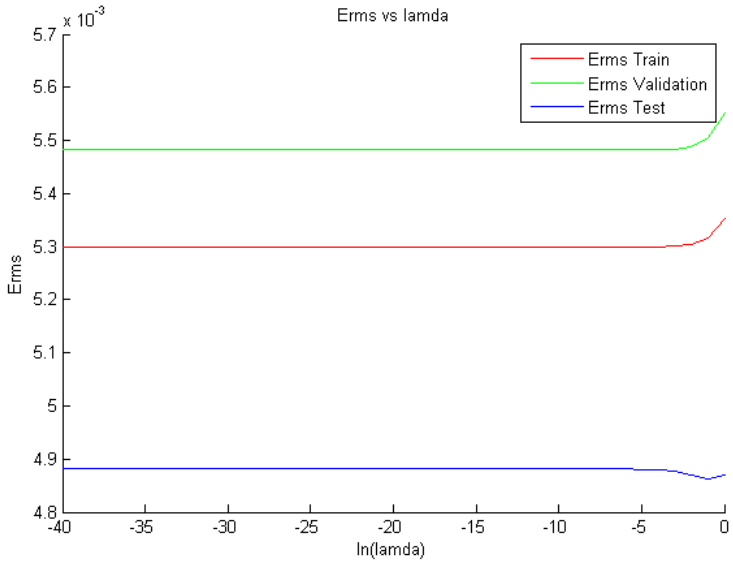


Figure 25: ERMS vs ln(Lamda)

- ERMS decreases with model complexity and the error is found to be minimum for test data set, maximum for validation dataset.
- For larger values of lambda, error increases on validation data set.

5.2 Inferences

- For real data, the approximation is bad compared to bivariate data. But error is comparatively low because output values are below 0.1.
- For very small values of S (s<80), error is very large.It is because the data given has high variance and it gives poor fit for low variance data.
- For same value of M and lambda, error value changes slightly due to different initialization of K-means clustering.