CS5691: Pattern Recognition and Machine Learning
Programming Assignment 1

LINEAR REGRESSION WITH REGULARIZATION

September 23, 2022

Aanand Krishnan *(BE17B001)*
Manoranjan J *(NA17B112)*
Reneeth Krishna MG *(BS17B025)*
Team 17 - CCS Section Spring 2021

Contents

List of Tables

List of Figures

Linear Regression

A vital question in any field of study is the very necessity of the field itself. Along those lines, one could pose, why do we need machine learning rather than an algorithm that could carry out the task at hand? The necessity for machine learning arises out of the complexity of the problem and the requirement of adaptivity. Several tasks that are performed routinely by humans such as speech recognition become too complex when asked to program. On the other hand, there are tasks that become highly complex because of the large amount of data involved in them. Making sense of data is a fundamental requirement of science and machines with ever increasing processing power are the suitable tools for pattern recognition in large and complex data.

A machine learning approach involves taking in input data, referred to as the *training set* and using it as a learning medium for prediction using new, unseen data called the *test set*. The learning phase can be *supervised*, *unsupervised* or *reinforced*. Regression is a type of supervised learning, where the desired output consists of one or more continuous variables. This report presents the results of various linear models for univariate, bivariate and multivariate data implemented in python.

Underlying Mathematical Framework of the Models

The discussion that follows is based on our understanding of the chapters 1 and 3 from the book[?] and is added here to make the report as self contained as possible. The name 'linear regression' comes with the caveat of the assumption that the method is applicable only in fitting linear functions. But as will be seen in this report, it is an extremely powerful tool that can be used for fitting a wide range of functions, both linear and non linear.

A training data set of $N-\textit{samples}$ in supervised learning takes the general form,

$$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$$

Where, $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. The $d-\textit{dimensional}$ space is referred to as the *feature space* and y_i are the various labels associated with the data.

After learning from the data set, the goal is to predict y for an unknown $\mathbf{x}$. A **basis function** is one which does this job. The simplest one among such functions is,

$$f(\mathbf{x}, \mathbf{w}) = w^T X = \sum_{n=1}^{d} w_i x_i \qquad (w^d \in \mathbb{R}^d)$$

A more general form can be defined with the help of the function $\phi : \mathbb{R}^d \longrightarrow \mathbb{R}^M$,

$$f(\mathbf{x}, \mathbf{w}) = w^T \phi(\mathbf{x}) = \sum_{n=1}^{M} w_i \phi_i(\mathbf{x})$$

The characteristic to be noted here is that $\phi(\mathbf{x})$ can be a non-linear function but $w \in \mathbb{R}^M$ is always linear, hence the name linear regression. The basis function can be of various types such as identity function, polynomial function, radial basis function, wavelets and such. Solving such problems means finding the vector w.

Let $\mathbf{Z} = \phi(\mathbf{x})$. Then the optimum solution for w is given by,

$$min_w \left((y - \mathbf{Z}w)^T (y - \mathbf{Z}w)\right) \iff \mathbf{Z}^T \mathbf{Z} w = \mathbf{Z}^T y$$

In practical problems the chance of getting a square matrix (i.e) a matrix where features and the samples available are same is very rare and thus there won't be any exact solutions. Geometrically, the best solutions then available are the orthogonal projections and the above equation represents exactly that.

Regularization term is added to the objective function to avoid overfitting. A *ridge regressor* is one where the L2 norm of the parameter w is added to the objective function. Therefore to find the optimum w,

$$min_w \left[(y - \mathbf{Z}w)^T (y - \mathbf{Z}w) + \lambda w^T w\right] \qquad (\lambda \geq 0)$$

Where λ is called the hyperparameter and since it is an positive quantity,

$$\exists v : \lambda = v^2$$

The columns of $\mathbf{Z}$ are $d - vectors$ in the vector space $\mathbb{R}^N$. If we are to embed this space in to a larger space $\mathbb{R}^{N+d}$ by lengthening each of the column vectors. Such a lengthening of column vectors resolves any collinearity among vectors that was present in the $N - dimensional$ space.

$$\mathbf{Z}_* = \begin{pmatrix} \mathbf{Z} \\ vI \end{pmatrix}$$

Similarly,

$$y_* = \begin{pmatrix} y \\ 0_{p\times 1} \end{pmatrix}$$

The following statement can be clearly verified by matrix multiplication,

$$(y_* - \mathbf{Z}_*)^T(y_* - \mathbf{Z}_*) = (y - \mathbf{Z}w)^T(y - \mathbf{Z}w) + \lambda w^T w$$

Therefore,

$$min_w \left[(y - \mathbf{Z}w)^T(y - \mathbf{Z}w) + \lambda w^T w\right] \iff \mathbf{Z}_*^T\mathbf{Z}w = \mathbf{Z}^T y_*$$

The means by which the problem of solutions to the system of linear equations is converted in to a minimization problem as described above is generally referred to as the *Tikhonov regularization.*

I. Linear Regression for Univariate Data

The Basis Function takes the form,

$$\phi(\mathbf{x}) = \mathbf{x}^j \qquad (\ 0 \le j \le Degree)$$

I.1. Results of change in regularisation parameter and degree of fit for training data =10

Generally at lower degrees, the function does not well fit as can be observed from the plots [1]. For Degree 6 the function seems to fit well. On the contrary, for degree 9, it overfits in the case of no regularisation constant. This can be explained by the less number of samples used in the training set and the degree being high. Once the regularisation constant is increased, it fits well till a certain point.

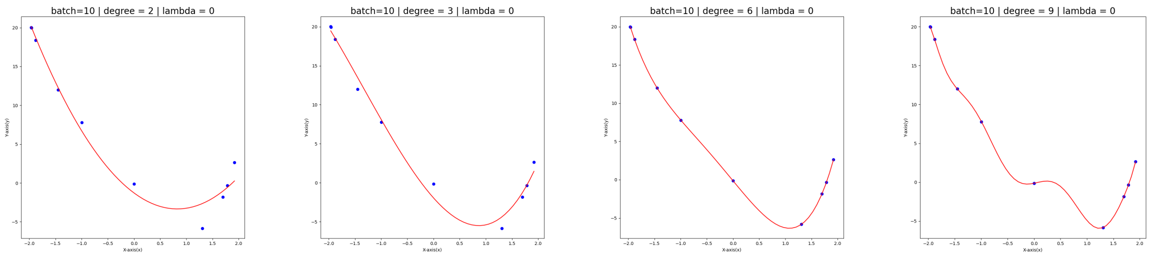


Figure 1: Plots of Polynomials having various orders of degree for a fixed regularisation parameter $\lambda = 0$ using batch size=10

$\mathbf{E}_{rms}$ values for different degrees			
Degree	$\mathbf{E}_{rms-train}$	$\mathbf{E}_{rms-test}$	$\mathbf{E}_{rms-valid}$
2	1.38	1.83	2.00
3	0.99	1.74	1.39
6	0.01	0.24	0.10
9	5.4e-4	0.19	0.42

Table 1: Error comparisons for varying degrees of $\phi(\mathbf{x})$ for Dataset 1 using batch size=10

Since degree 9 is overfit, we experiment with different regularisation parameter lambda the of quadratic regularisation term to see how it affects the fit. For lambda=0.001 and 0.01, the validation Erms decreases than what was before and from 0.1 it increases again. For higher values of lambda the fit is not proper as can be seen from the figures below[3].

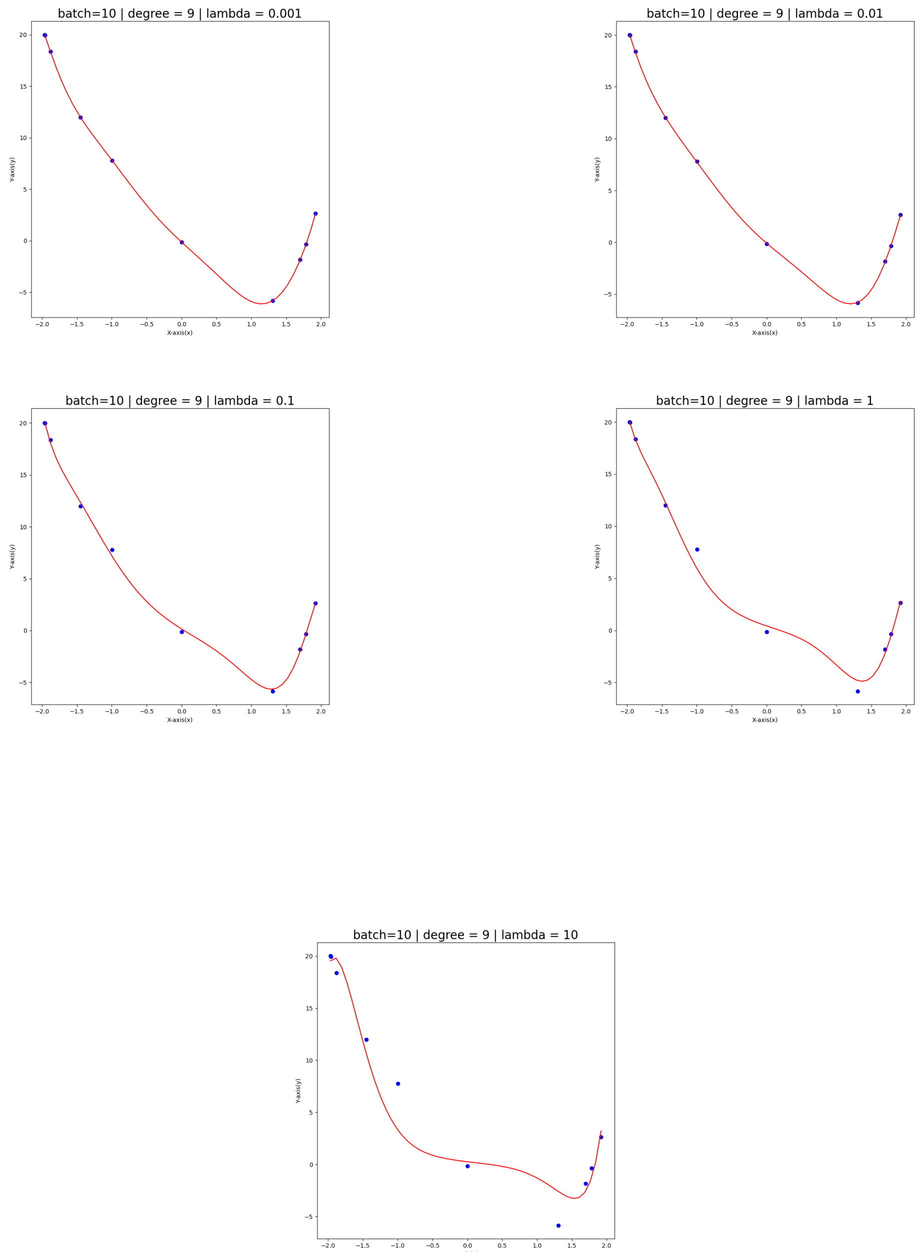


Figure 3: Plots of degree 9 Polynomials having various orders of regularisation parameter using batch size=10

$\mathbf{E}_{rms}$ values for different regularisation parameters			
lambda	$\mathbf{E}_{rms-train}$	$\mathbf{E}_{rms-test}$	$\mathbf{E}_{rms-valid}$
0.001	0.01	0.16	0.14
0.01	0.05	0.09	0.21
0.1	0.28	0.20	0.45
1	0.71	0.41	1.15
10	1.91	0.64	2.57

Table 2: Error comparisons for various regularisation parameters with degree=9 for Dataset 1 using batch size=10

I.2. Results for change in regularisation parameter and degree of fit for training data = 200

Generally at lower degrees, the function does not well fit as can be observed from the plots [4]. The fit for degree 6 9 seems to be good. Since the validation accuracy is good and the model doesn't seem to overfit, regularisation experiment is not done here.

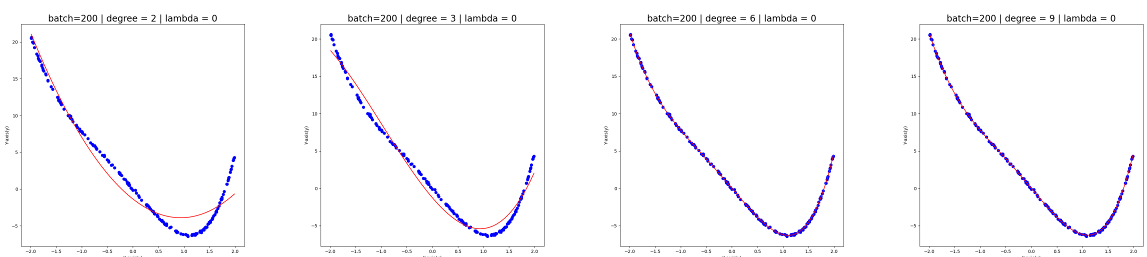


Figure 4: Plots of Polynomials having various orders of degree for a fixed regularisation parameter $\lambda = 0$ using batch size200

$\mathbf{E}_{rms}$ values for different degrees			
Degree	$\mathbf{E}_{rms-train}$	$\mathbf{E}_{rms-test}$	$\mathbf{E}_{rms-valid}$
2	1.64	1.73	1.70
3	1.05	0.85	1.07
6	0.09	0.09	0.11
9	0.09	0.09	0.11

Table 3: Error comparisons for varying degrees of $\phi(\mathbf{x})$ for Dataset 1 using batch size=200

II. Linear Regression for Bivariate Data

The Polynomial Basis Function used takes the form,

$$\phi(\mathbf{x}) = \{\mathbf{x}_1^i \mathbf{x}_2^j\}$$

Where, $0 \leq i, j \leq Degree$ and $i + j \leq Degree$ In case of data points with two features, the polynomial basis functions includes,

$$\phi(\mathbf{x}) = \{1, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_1^2, \mathbf{x}_2^2, \mathbf{x}_1\mathbf{x}_2, \mathbf{x}_1^2\mathbf{x}_1^2\}$$

II.1. Experiments & Observations on Dataset 2:

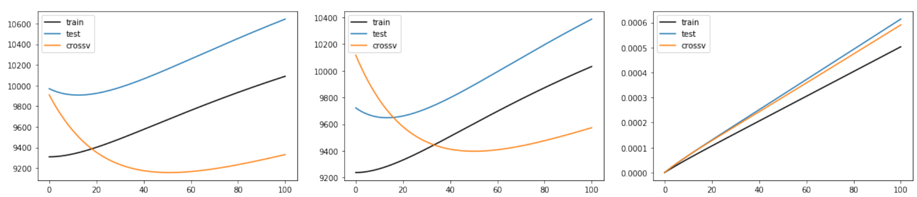
II.1.1. RMS comparison for varying regularisation parameter λ (quadratic regularisation):

Different λ values were tested across different models with the goal of finding any λ value that might be able to ameliorate any overfitting. However across all models, the increase of λ value was accompanied with the strict increase of RMS value. Therefore, subsequent analysis has been carried out without regularization. The best model which was found to be of $batch = 500$ and $degree = 6$ with no regularisation performed with significant accuracy on the training, test and cross validation without any instances of overfitting. The effect of λ is illustrated in the figure[5]. The table[3] illustrates the analysis of different λ values across the $degree = 6$ model and it can be inferred without doubt from the table that the best fitting model has $\lambda = 0$.

$\mathbf{E}_{rms}$ values for different data			
λ	$\mathbf{E}_{rms-train}$	$\mathbf{E}_{rms-test}$	$\mathbf{E}_{rms-valid}$
1	7.58e-06	1.06e-05	7.51e-06
0	1.21e-07	1.15e-07	1.17e-07
0.1	7.75e-06	1.06e-06	7.77e-07
0.01	1.04e-07	1.31e-07	1.01e-07
0.001	1.18e-07	1.08-07	1.17e-07
0.0001	3.30e-08	3.2e-08	3.08e-08
1e-05	1.77e-07	1.65e-07	1.69e-07

Table 4: Error comparisons for varying values of λ for Dataset 2 and a model of $degree = 6$

.



(a) Degree, $M = 2$ (b) Degree, $M = 3$ (c) Degree, $M = 6$

Figure 5: Effect of λ

II.1.2. RMS comparison across different models

Tables [5], [6], [7], and [??] provides an error based analysis of the model implemented.

$\mathbf{E}_{rms}$ values for different data			
Degree	$\mathbf{E}_{rms-train}$	$\mathbf{E}_{rms-test}$	$\mathbf{E}_{rms-valid}$
2	8472.09	7916.24	11603.31
3	7746.87	10209.22	9055.03
6	1.11e-06	7.12e-06	1.032e-06

Table 5: Error comparisons for varying degrees $\phi(\mathbf{x})$ for Dataset 2 using 50 samples

.

$\mathbf{E}_{rms}$ values for different data			
Degree	$\mathbf{E}_{rms-train}$	$\mathbf{E}_{rms-test}$	$\mathbf{E}_{rms-valid}$
2	7836.93	9102.67	10203.59
3	7410.93	8955.31	9615.30
6	9.07e-08	1.14e-07	9.06e-08

Table 6: Error comparisons for varying degrees $\phi(\mathbf{x})$ for Dataset 2 using 100 samples
.

$\mathbf{E}_{rms}$ values for different data			
Degree	$\mathbf{E}_{rms-train}$	$\mathbf{E}_{rms-test}$	$\mathbf{E}_{rms-valid}$
2	9007.85	9941.50	7712.90
3	8891.04	10448.97	7669.04
6	2.77e-08	3.15e-08	3.17e-08

Table 7: Error comparisons for varying degrees $\phi(\mathbf{x})$ for Dataset 2 using 200 samples
.

$\mathbf{E}_{rms}$ values for different data			
Degree	$\mathbf{E}_{rms-train}$	$\mathbf{E}_{rms-test}$	$\mathbf{E}_{rms-valid}$
2	9308.74	9968.55	9908.81
3	9237.93	9721.98	10117.44
6	2.54e-08	2.92e-08	2.51e-08

Table 8: Error comparisons for varying degrees $\phi(\mathbf{x})$ for Dataset 2 using 500 samples
.

II.1.3. Surface Plots with Training set superimposed

Figures [6], [7], [8] and [9] are the various surface plots.

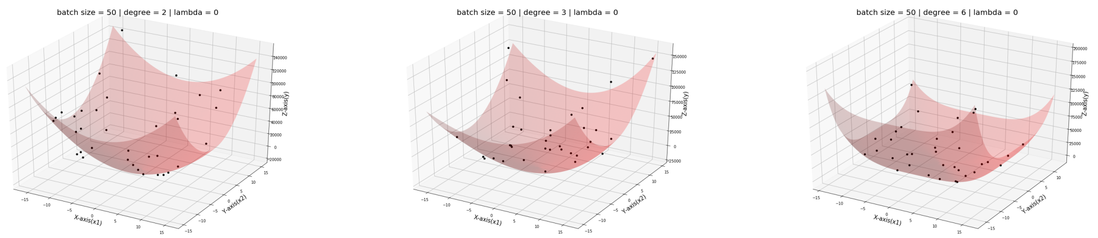


Figure 6: Surface Plots using various degree for a fixed best regularisation parameter $\lambda = 0$ using 50 samples

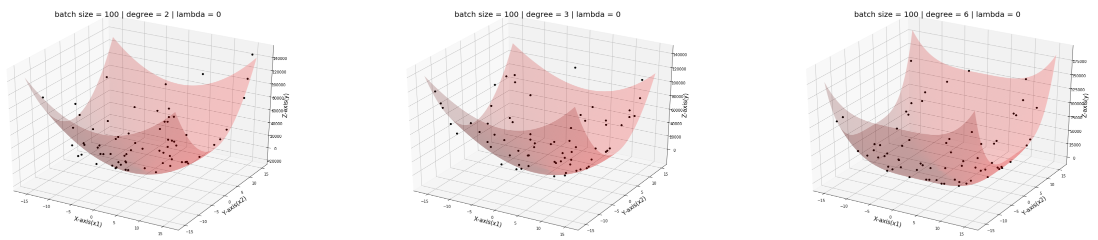


Figure 7: Surface Plots using various degree for a fixed regularisation parameter $\lambda = 0$ using 100 samples

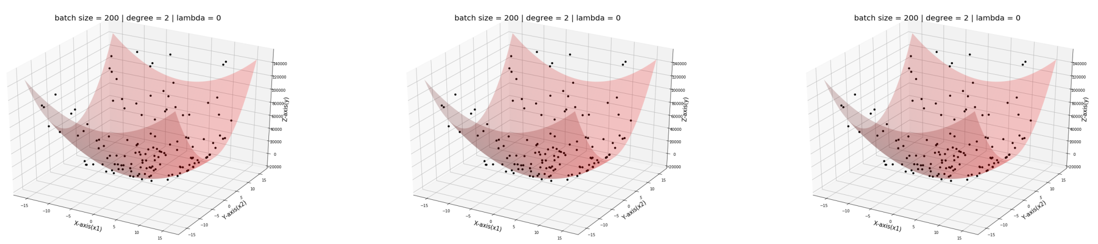


Figure 8: Surface Plots using various degree for a fixed regularisation parameter $\lambda = 0$ using 200 samples

II.1.4. Scatter Plots of the Best Model

The best model was found to be of batch size 500 and degree 6 without any regularization and fared significantly better than the other degree 2,3 models as it's E_{rms} values were found to be smaller in orders of magnitude 10^9. Scatter plots of the best model are represented in figure [10]

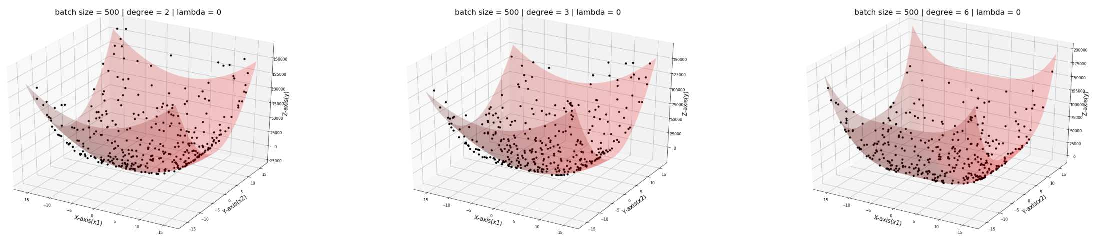
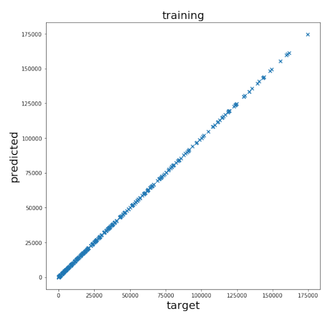
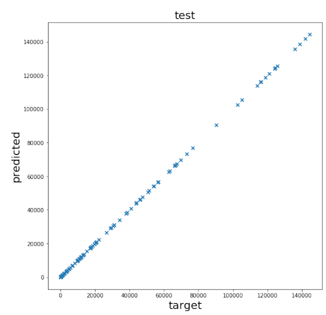


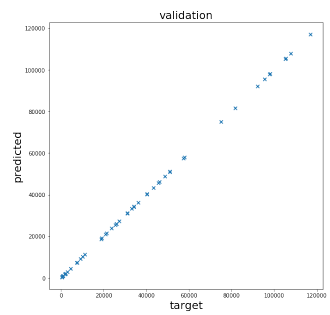
Figure 9: Surface Plots using various degree for a fixed regularisation parameter $\lambda = 0$ using 500 samples



(a) Training Data



(b) Test Data



(c) Validation Data

Figure 10: Best Scatter Plot

III. Linear Regression for Multivariate Real World Data & Bivariate Data

The model implemented here makes use of a Gaussian Basis Function,

$$\phi_j(\mathbf{x}) = \exp\left(-\frac{(\mathbf{x} - \mu_j)^2}{\sigma^2}\right)$$

III.1. Experiments & Observations on Dataset 2:

III.1.1. RMS error comparison for varying dimensions and width of $\phi(\mathbf{x})$:

A total of 500 samples from the given bivariate dataset was chosen at random and divided in to training, validation and test data(70%, 20% and 10%). Using the training data thus generated an attempt is made to find the best model. At first the case without regularisation is considered. For each dimension of the gaussian basis function, different values of width has been experimented here. As can be concluded from the tables, a dimension of 75 and width of 10 seems to be optimal model for the regression task at hand. Also, one can observe over-fitting when the dimensions are increased to 340 and for width of 1. The value of training data, E_{rms} is drastically lower than that of validation data.

$\mathbf{E}_{rms}$ values for dimension 25			
Width	$\mathbf{E}_{rms-train}$	$\mathbf{E}_{rms-val}$	$\mathbf{E}_{rms-test}$
1	3.40×10^4	6.42×10^4	4.16×10^4
10	3.07×10^3	4.12×10^3	3.17×10^3
50	1.96×10^5	2.01×10^5	2.08×10^5

$\mathbf{E}_{rms}$ values for dimension 50			
Width	$\mathbf{E}_{rms-train}$	$\mathbf{E}_{rms-val}$	$\mathbf{E}_{rms-test}$
1	2.70×10^4	4.39×10^4	5.50×10^4
10	3.33×10^2	4.87×10^2	4.46×10^2
50	7.01×10^5	6.86×10^5	6.72×10^5

$\mathbf{E}_{rms}$ values for dimension 75			
Width	$\mathbf{E}_{rms-train}$	$\mathbf{E}_{rms-val}$	$\mathbf{E}_{rms-test}$
1	2.28×10^4	3.56×10^4	4.79×10^4
10	1.18×10^2	1.47×10^2	1.34×10^2
50	3.84×10^5	3.94×10^5	4.05×10^5

$\mathbf{E}_{rms}$ values for dimension 340			
Width	$\mathbf{E}_{rms-train}$	$\mathbf{E}_{rms-val}$	$\mathbf{E}_{rms-test}$
1	1.01×10^3	2.51×10^4	1.97×10^4
10	2.74×10^6	3.30×10^6	3.41×10^6
50	5.22×10^5	5.29×10^5	5.32×10^5

Table 9: Error comparisons for varying values of $\phi(\mathbf{x})$ and width(σ) for Dataset 2

III.1.2. Effect of regularisation in case of over-fitting:

As inferred from the previous section, the over-fit case of dimension 340 and width 1 is experimented for different values of λ, thus a regularisation term is added resulting in various models for fixed dimension and width. Both quadratic and tikhonov regulariser are experimented with here. This helps us to see the impact of regularisation in case of over-fitting. As can be seen from the error values, though adding a regulariser doesn't better fit the model but it helps in ruling out models which have essentially memorized the training data-set but perform poorly on test data set.

$\mathbf{E}_{rms}$ values for different data			
λ	$\mathbf{E}_{rms-train}$	$\mathbf{E}_{rms-val}$	$\mathbf{E}_{rms-test}$
10e-5	1.07×10^3	2.517×10^4	1.979×10^4
10e-3	1.08×10^3	2.520×10^4	1.974×10^4
10e-1	1.17×10^3	2.529×10^4	1.983×10^4
0	1.01×10^3	2.516×10^4	1.979×10^4
10	2.84×10^4	3.852×10^4	3.782×10^4
10e+03	5.11×10^4	5.919×10^4	6.081×10^4
10e+05	5.20×10^4	6.001×10^4	6.170×10^4

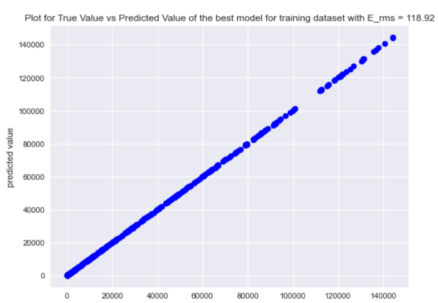
Table 10: Error comparisons for quadratic regularisation by varying λ and fixed dimension =340 and width = 1 for Dataset 2

$\mathbf{E}_{rms}$ values for different data			
λ	$\mathbf{E}_{rms-train}$	$\mathbf{E}_{rms-val}$	$\mathbf{E}_{rms-test}$
10e-3	1.007×10^3	2.543×10^4	1.970×10^4
10e-1	1.046×10^3	3.088×10^4	2.763×10^4
0	1.007×10^3	2.516×10^4	1.979×10^4
10	3.161×10^4	4.01×10^4	3.930×10^4
10e+03	5.119×10^4	5.920×10^4	6.080×10^4
10e+05	5.204×10^4	6.005×10^4	6.170×10^4

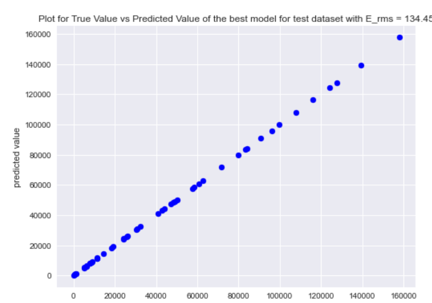
Table 11: Error comparisons for tikhonov regularisation by varying λ and fixed dimension =340 and width = 1 for Dataset 2

III.1.3. Scatter Plots of Best Model

Figure [11] shows the the scatter plot of predicted and actual target variable for the training and the test data.



(a) Training Data



(b) Test Data

Figure 11: Scatter Plot of the Gaussian Basis Function of the best performing model for Dataset 2

III.2. Experiments & Observations on Dataset 3:

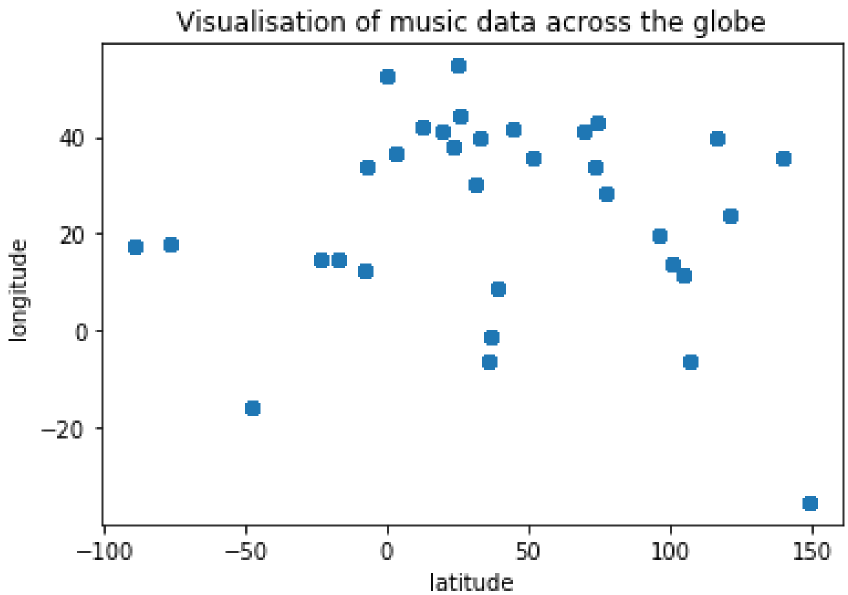


Figure 12: Training Data

III.2.1. RMS error comparison for varying dimensions and width of $\phi(\mathbf{x})$:

A total of 500 samples from the given bivariate dataset was chosen at random anddivided in to training, validation and test data(70%, 20% and 10%). Using the training data thus generated an attempt is made to find the best model. At first the case without regularisation is considered. For each dimension of the gaussian basis function, different values of width has been experimented here. As can be concluded from the tables, a dimension of 100 with width of 10 seems to be the optimal model for the regression task at hand. Also one can observe overfitting when the dimensions are increased to 340. Here x,y in each column of the table represents the Erms in both the output dimensions

$\mathbf{E}_{rms}$ values for dimension 25			
Width	$\mathbf{E}_{rms-train}$	$\mathbf{E}_{rms-val}$	$\mathbf{E}_{rms-test}$
1	17.19,48.11	19.49,52.58	16.25,47.46
10	15.23,43.45	19.17,45.01	15.50,42.72
50	15.01,42.80	19.52,45.40	15.16,46.30

$\mathbf{E}_{rms}$ values for dimension 50			
Width	$\mathbf{E}_{rms-train}$	$\mathbf{E}_{rms-val}$	$\mathbf{E}_{rms-test}$
1	15.91,44.73	19.44,52.32	16.29,46.97
10	13.82,40.69	19.32,45.60	13.56,42.47
50	14.24,39.84	19.19,45.61	14.53,43.38

$\mathbf{E}_{rms}$ values for dimension 100			
Width	$\mathbf{E}_{rms-train}$	$\mathbf{E}_{rms-val}$	$\mathbf{E}_{rms-test}$
1	13.21,42.71	19.21,46.37	19.88,48.96
10	11.99,33.57	18.61,42.23	13.18,39.20
50	12.12,33.54	22.19,46.99	14.34,45.70

$\mathbf{E}_{rms}$ values for dimension 340			
Width	$\mathbf{E}_{rms-train}$	$\mathbf{E}_{rms-val}$	$\mathbf{E}_{rms-test}$
1	1.73,6.06	23.07,68.55	22.38,62.18
10	1.90,5.90	26.20,50.01	14.12,46.23
50	2.23,5.98	26.04,57.54	23.33,66.33

III.2.2. RMS error comparison for varying regularisation parameters of quadratic and Tikhonov regularisation

Only the regularisation was changed here and dimensions were set to 100 with a variance of 3. The Erms increases as the regularisation parameter increases which tells that the model is now trying to reduce the overfit. As can be observed, the difference between E_{rms} of validation and test set is less with the E_{rms} on the training set is now less but it comes as a result of increase in overall E_{rms}. Table [13] demonstrates this.

$\mathbf{E}_{rms}$ values for different data			
λ	$\mathbf{E}_{rms-train}$	$\mathbf{E}_{rms-val}$	$\mathbf{E}_{rms-test}$
10e-5	1.97,5.96	21.77 ,46.14	11.34, 41.73
10e-3	5.79, 13.55	18.91,43.57	12.08, 37.75
10e-1	12.27, 35.94	18.39, 42.86	13.56,38.78
0	1.90, 5.90	26.20 50.01	14.12,46.23
10	14.98, 43.41	18.58,45.88	15.10,43.18
10e+03	21.92, 55.43	22.41,59.02	21.42,53.03
10e+05	32.75, 63.30	31.42,67.81	31.97,61.44

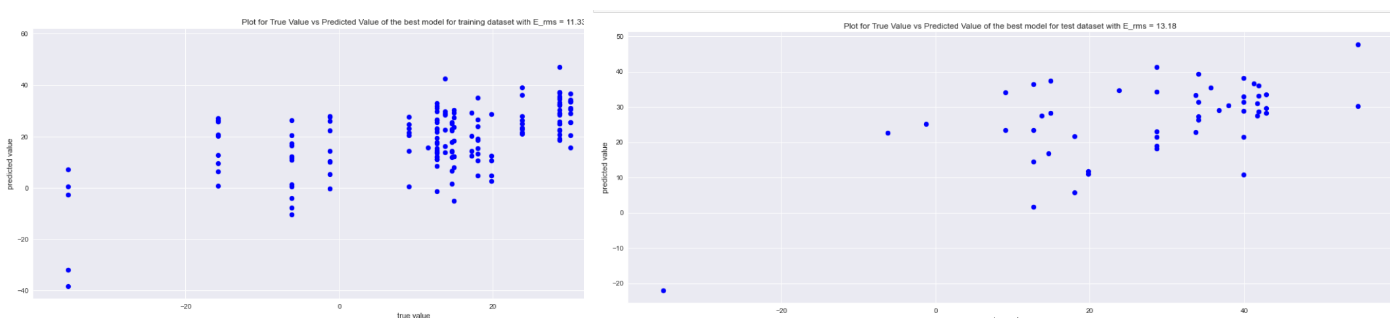
Table 12: Error comparisons for quadratic regularisation by varying λ and fixed dimension =340 and width = 10 for Dataset 3

$\mathbf{E}_{rms}$ values for different data			
Variance	$\mathbf{E}_{rms-train}$	$\mathbf{E}_{rms-val}$	$\mathbf{E}_{rms-test}$
10e-3	2.37, 6.52	19.40,44.16	12.32, 38.92
10e-1	11.23, 32.31	18.21, 42.89	12.95,38.52
0	1.90, 5.90	26.20 50.01	14.12,46.23
10	15.43, 44.86	18.59,48.12	14.89,44.27
10e+03	32.06, 62.60	30.76,67.07	31.26,60.72
10e+05	33.10, 63.58	31.73,68.11	32.33,61.74

Table 13: Error comparisons for tikhonov regularisation by varying λ and fixed dimension =340 and width = 10 for Dataset 3

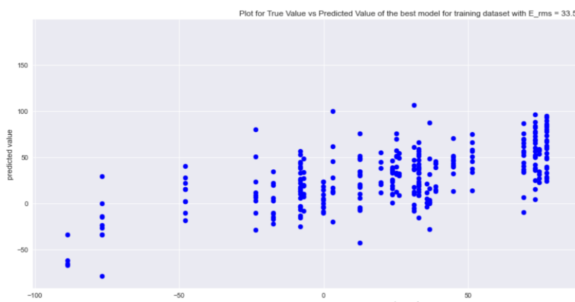
III.2.3. Scatter Plots of Best Model:

Figure [14] shows the the scatter plot of predicted and actual target variable for the training and the test data.



(a) Training Data (b) Test Data

Figure 13: Scatter Plot of the Gaussian Basis Function of the best performing model for Dataset 3 for first column



(a) Training Data

(b) Test Data

Figure 14: Scatter Plot of the Gaussian Basis Function of the best performing model for Dataset 3 for second column