Introduction to Machine Learning

ELL784 | Assignment 3

Shivangi Bithel, Yukti Kaura, Vishnu Gorty

Index

- Polynomial Curve Fitting
- Genomic Sequence Analysis

Polynomial Curve Fitting

Problem Statement

1. To begin with, use only the first 20 data points in your file.
2. Solve the polynomial curve fitting regression problem using error function minimisation.
3. Define your own error function other than the sum-of-squares error.
4. Try different error formulations and report the results.
5. Use a goodness-of-fit measure for polynomials of different order.
6. Can you distinguish overfitting, underfitting, and the best fit?
7. Obtain an estimate for the noise variance.
8. Introduce regularisation and observe the changes. For quadratic regularisation, can you obtain an estimate of the optimal value for the regularisation parameter lamba? What is your corresponding best guess for the underlying polynomial? And the noise variance?
9. Now repeat all of the above using the full data set of 100 data points.
10. How are your results affected by adding more data? Comment on the differences.
11. What is your final estimate of the underlying polynomial? Why?

Introduction

Dataset - https://www.cse.iitd.ac.in/~sumantra/courses/ml/a3/group01.txt

Programming Language : Python
Libraries used :

1. sklearn.linear_model.LinearRegression : For Polynomial Regression
2. sklearn.linear_model.Ridge : For Regularized Polynomial Regression
3. sklearn.preprocessing.PolynomialFeatures : For polynomial features
4. sklearn.model_selection.KFold : For train-validation split of the data

Metrics(em):

1. Mean Squared Error(MSE)
2. Mean Absolute Error(MAE)
3. Goodness of Fit(R2)

Process Overview

Polynomial Degree Range : 0 to 10

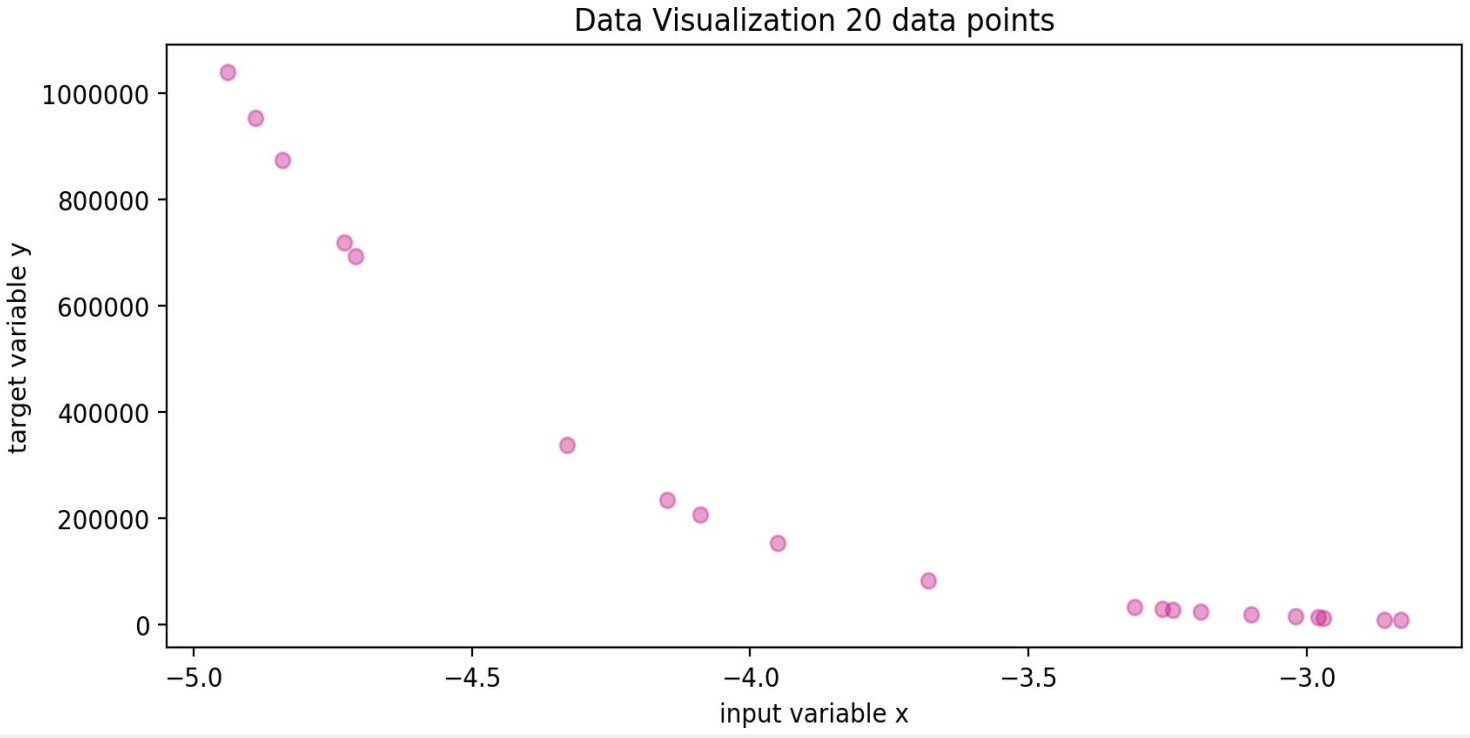
Alpha Range: 1e-2 to 1e2

Cross Validation constant (k) : 5

1. For each degree in degree range do:
    a. Fit model
    b. Calculate error metrics over k-folds (and alpha, if regularized)
2. Select the degree(and alpha, if regularized) which yields best goodness of fit

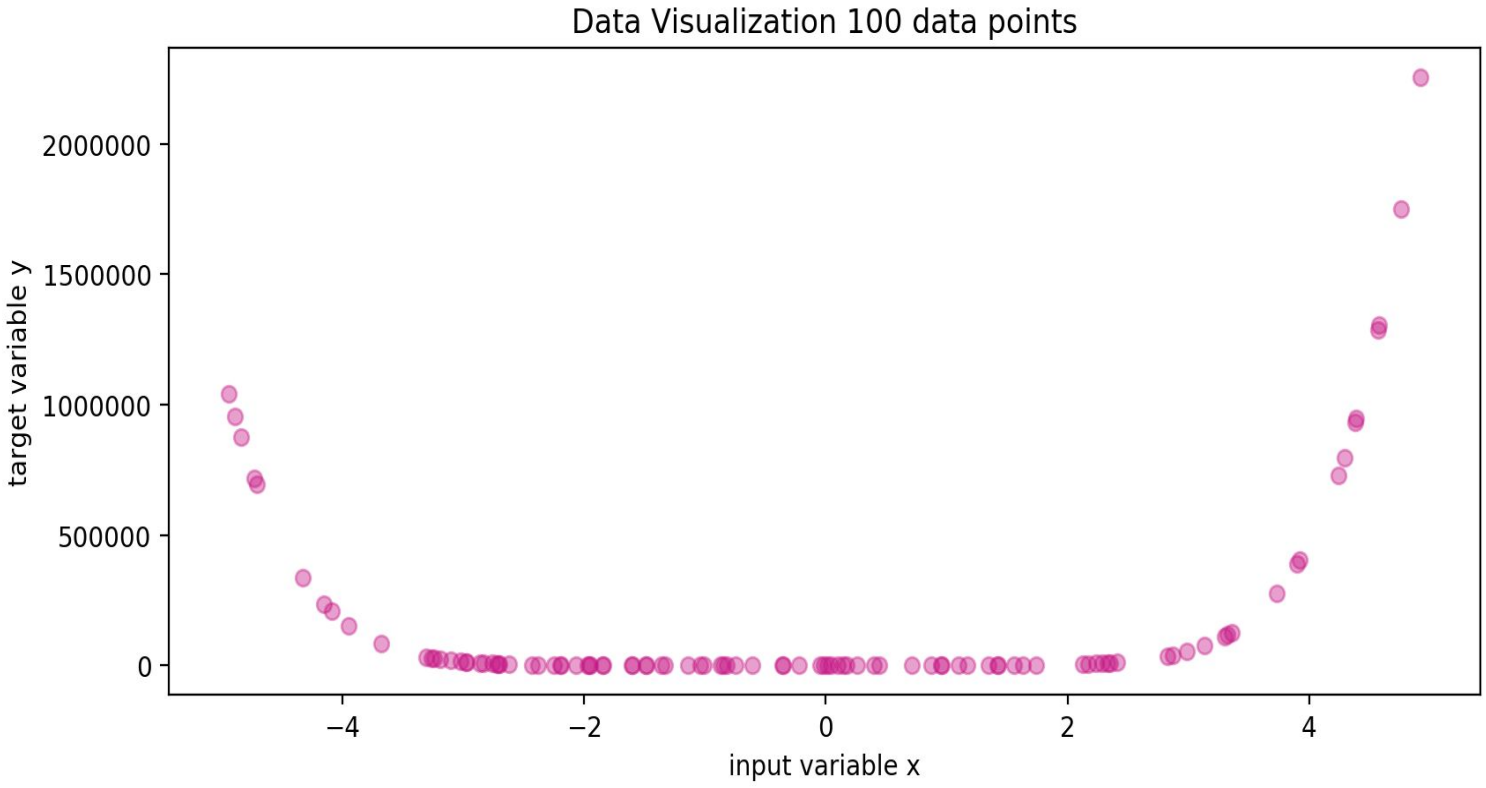
Data Visualization

Data Visualization with 20 data points



Y ranges from :
7990.54 to
1039755.18
X ranges from :
-4.94 to -2.83

Data Visualization with 100 data points



Y ranges from :
1.21 to 2254025.74
X ranges from :
-4.94 to 4.92

Polynomial Regression

20 Data Points

Data with top 20 data points

Errors and Goodness of Fit

MSE		
Degree	Train Error	TestError
0	1.18E+11	1.89E+11
1	1.51E+10	5.68E+10
2	8.04E+08	6.55E+09
3	2.84E+07	3.42E+08
4	2.90E+05	1.27E+07
5	1.26E+03	2.43E+05
6	1.49E+00	1.77E+03
7	3.19E-02	1.83E+02
8	2.86E-02	1.22E+02
9	2.03E-02	5.57E+03
10	2.06E-02	9.81E+03

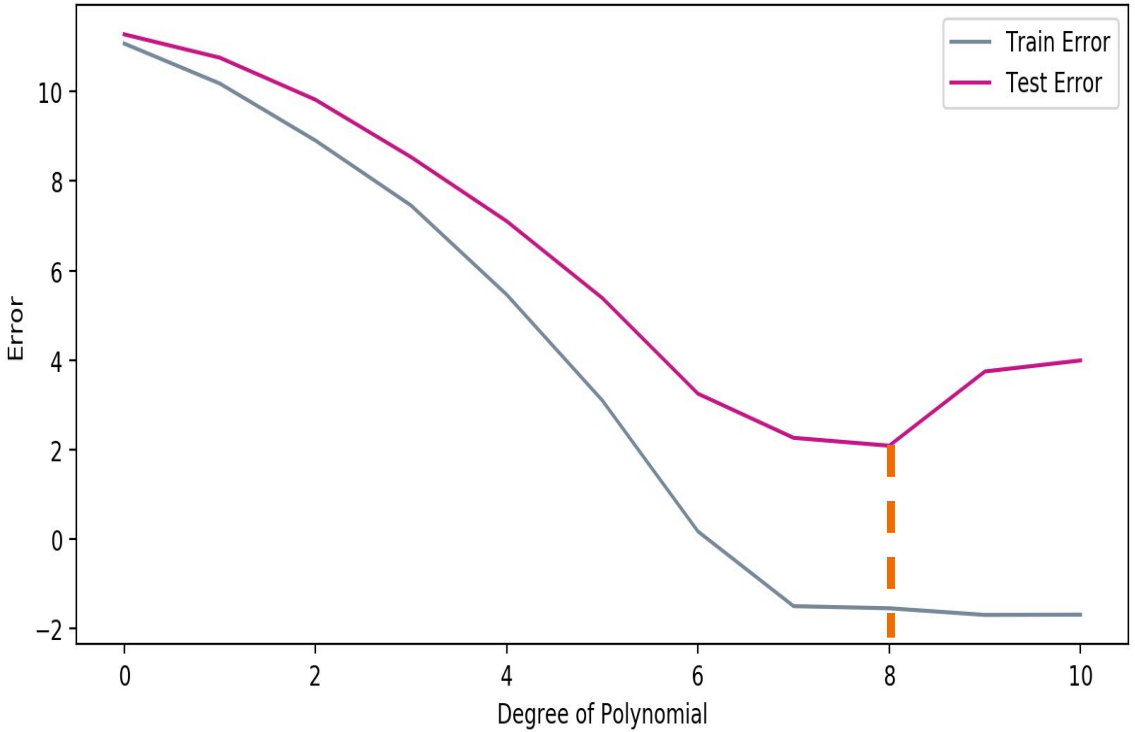
MAE		
Degree	Train Error	TestError
0	2.87E+05	3.64E+05
1	9.78E+04	1.95E+05
2	2.37E+04	7.02E+04
3	4.37E+03	1.36E+04
4	4.52E+02	2.55E+03
5	2.79E+01	2.95E+02
6	9.68E-01	2.18E+01
7	1.38E-01	5.77E+00
8	1.24E-01	4.58E+00
9	9.73E-02	2.99E+01
10	9.72E-02	3.79E+01

R2		
Degree	Train Fit	Test Fit
0	0.00000000000	-5724.60742000000
1	0.86097601400	-1695.04210000000
2	0.99169480900	-358.51046500000
3	0.99975269900	-25.63434420000
4	0.99999752500	0.26824030100
5	0.99999999000	0.99033749900
6	1.00000000000	0.99994218600
7	1.00000000000	0.99999938400
8	1.00000000000	0.99999995100
9	1.00000000000	0.99997933500
10	1.00000000000	0.99998654400

As is seen for degree 8, R2 value is maximum we choose the degree 8 as the degree of the model

Optimal degree = 8

FIT OF THE MODEL



- **Degree = 0 :** Train & Test error both are very high : UNDERFIT
- **Degree > 8 :** Train error is low & Test error is high : OVERFIT
- **Degree = 8 :** both Train & Test are stable and minimum : BEST-FIT

Same has been Indicated by the goodness fit measure .

FINAL POLYNOMIAL REGRESSION MODEL

Polynomial Coefficients : 0.00000000e+00 -6.25701553e+05 -6.02581129e+05 -3.29180875e+05 -1.11568325e+05 -2.40209506e+04 -3.20479316e+03 -2.34696399e+02 -3.51654246e+00

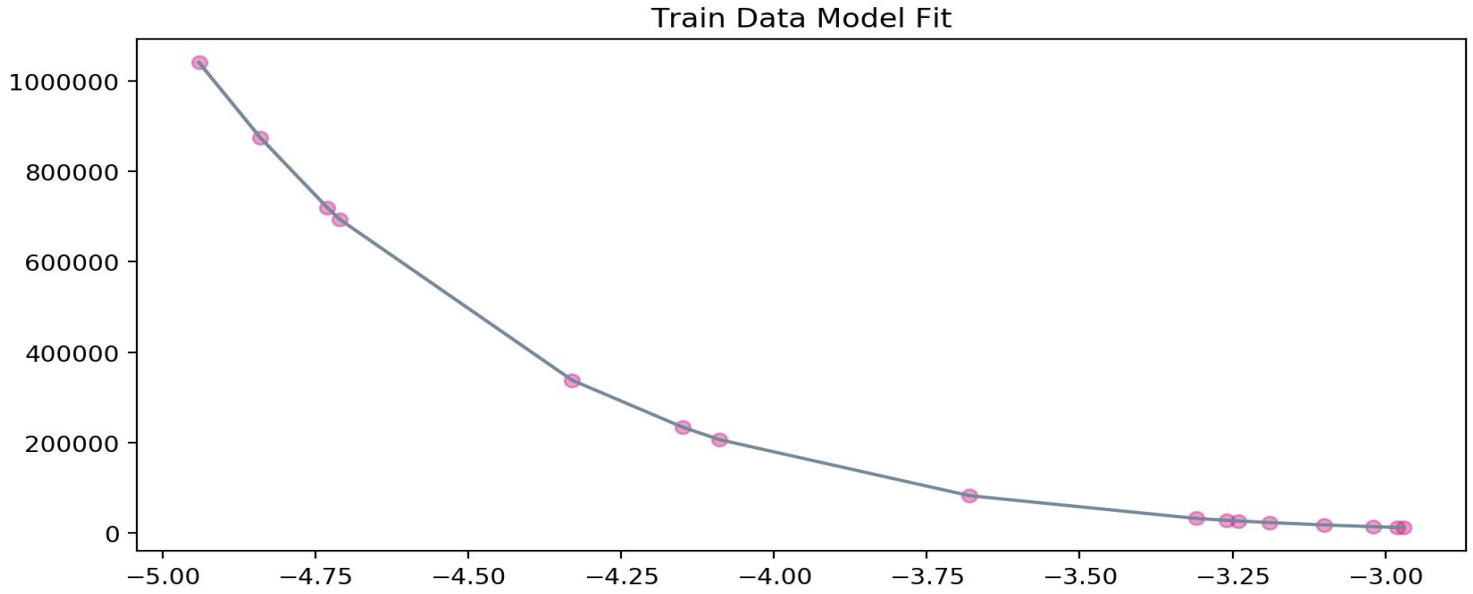
Intercept - -282167.77394048

Variance : .0442122972017O674

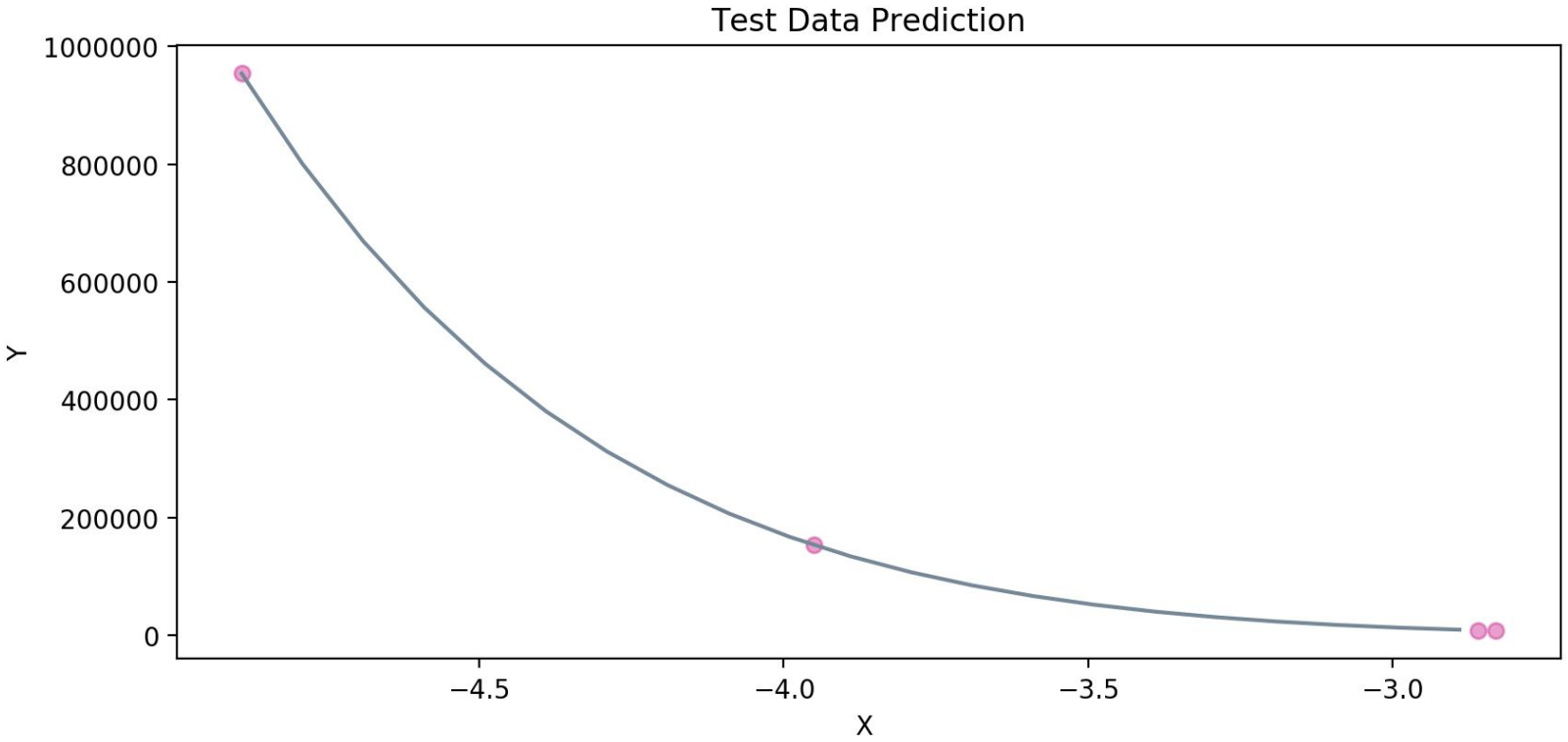
Optimal degree: 8

FITTING THE MODEL TO DATA

TRAIN TEST SPLIT = 80:20 (Non-Random, No Shuffle)



CONTD...


Test Data Prediction
1000000
800000
600000
400000
200000
0
Y
−4.5
−4.0
−3.5
−3.0
X

100 Data Points

Complete data

Errors and Goodness of Fit

MSE		
Degree	Train Error	TestError
0	1.56E+11	1.89E+11
1	1.30E+11	3.65E+11
2	4.27E+10	7.55E+10
3	3.11E+10	9.59E+11
4	3.84E+09	2.63E+11
5	2.35E+09	2.11E+12
6	5.75E+07	5.43E+11
7	3.22E+07	9.83E+11
8	1.14E-01	5.15E+03
9	1.13E-01	8.02E+02
10	1.10E-01	1.22E+05

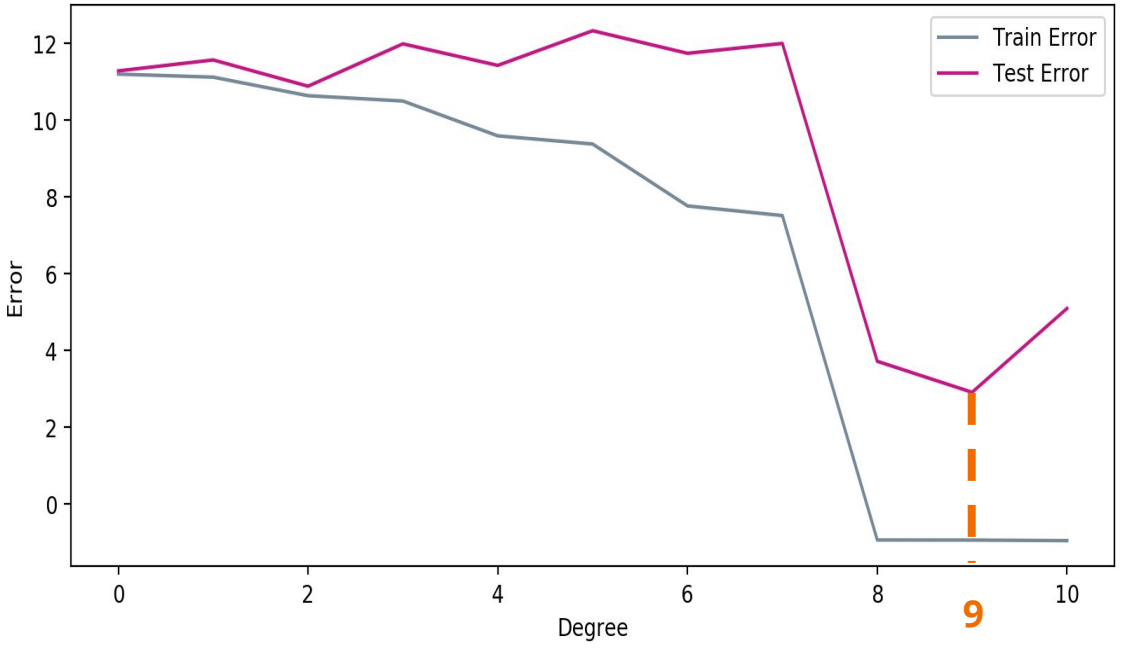
MAE		
Degree	Train Error	TestError
0	2.51E+05	2.87E+05
1	2.42E+05	4.14E+05
2	1.50E+05	1.99E+05
3	1.30E+05	5.48E+05
4	4.46E+04	2.90E+05
5	3.44E+04	6.74E+05
6	5.61E+03	3.32E+05
7	3.92E+03	4.01E+05
8	2.75E-01	2.56E+01
9	2.74E-01	8.95E+00
10	2.68E-01	1.07E+02

R2		
Degree	Train Fit	Test Fit
0	0.00000000000	-1891070290.00000000000
1	0.19165457600	-1687839700.00000000000
2	0.72201914100	-2144495170.00000000000
3	0.82033643800	-1854109420.00000000000
4	0.97608587700	-473225586.00000000000
5	0.98703890700	-287962289.00000000000
6	0.99966221000	-19492697.40000000000
7	0.99982871300	-9226391.68000000000
8	1.00000000000	0.99333263300
9	1.00000000000	0.99429207500
10	1.00000000000	0.99328728000

As is seen for degree 8, R2 value is maximum we choose the degree 8 as the degree of the model

Optimal degree = 9

FIT OF THE MODEL



- **Degree = 0 :** Train & Test error both are very high : UNDERFIT
- **Degree > 9 :** Train error is low & Test error is high : OVERFIT
- **Degree = 9 :** both Train & Test are stable and minimum : BEST-FIT

Same has been Indicated by the goodness fit measure .

FINAL POLYNOMIAL REGRESSION MODEL

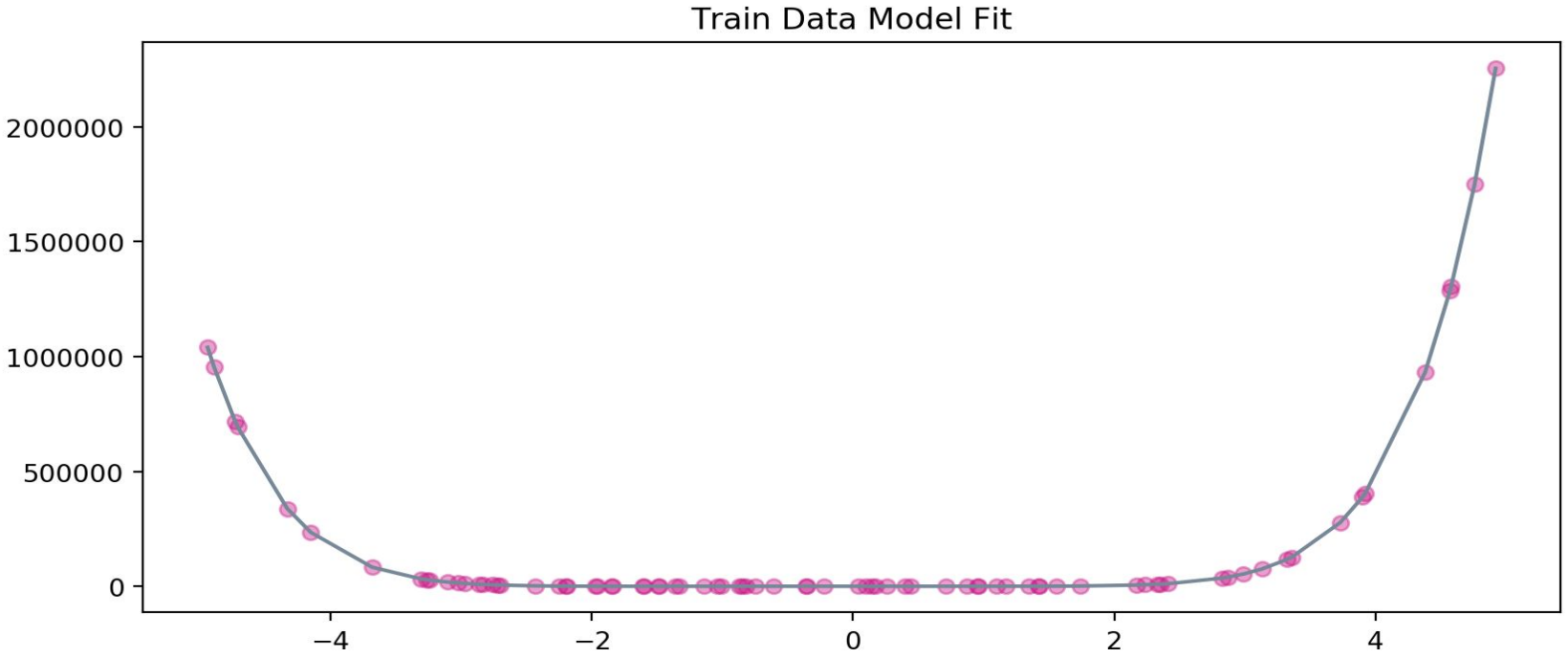
Polynomial Coefficients : 0.6.93942057 8.63781465 7.07980821 5.50764703 5.80178037 5.69965776 8.69995532 4.50000523

Intercept - 7.23321344

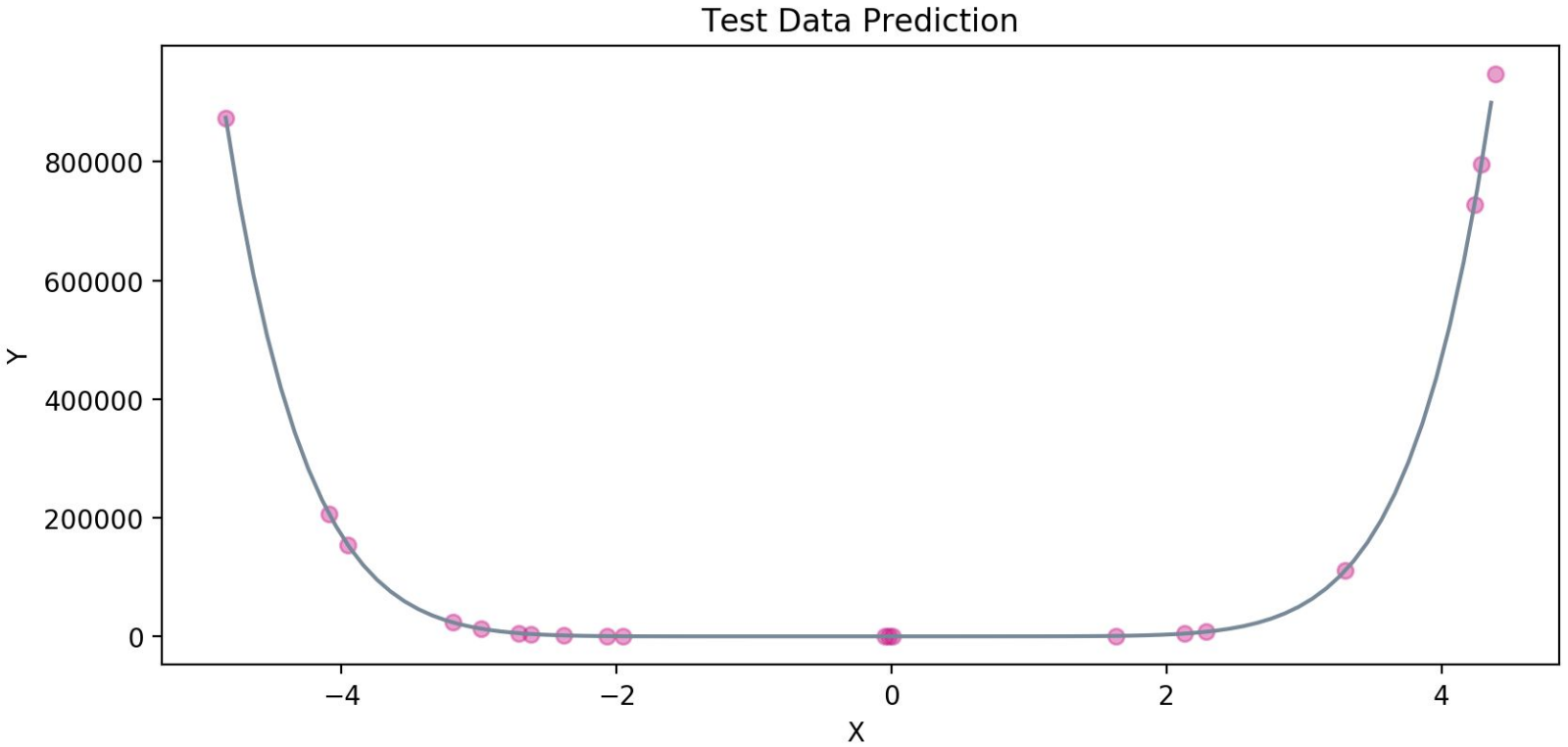
Variance : .115205897524308 6

Optimal degree: 9

FITTING THE MODEL TO DATA -TRAINING SET



FITTING MODEL TO DATA - TEST SET



Observation

For 20 data points the degree was 8 and now for 100 data points we get a optimal degree of 9 and variance increases.

Therefore, as we increase the data points the model learns better.

Regularized Regression

(With a regularization parameter lambda)

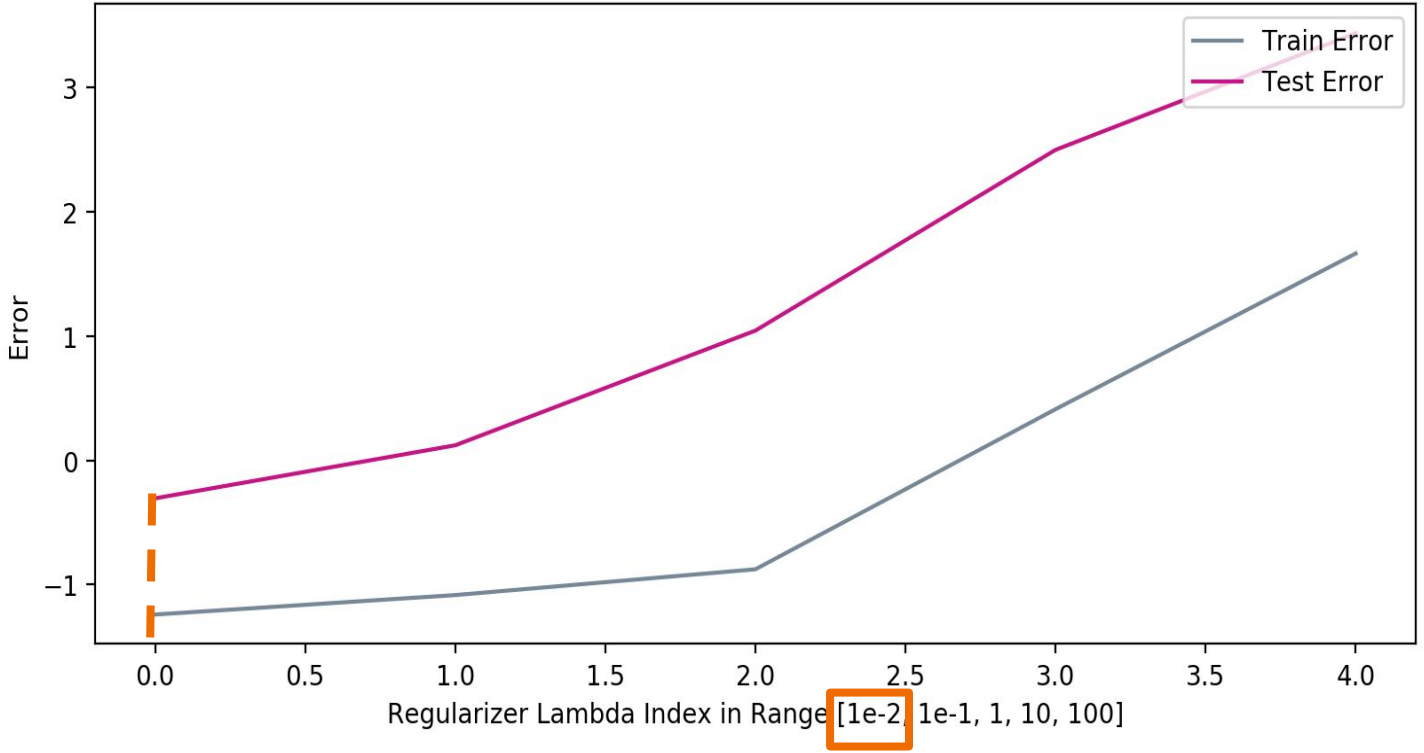
20 Data Points

Data with top 20 data points

GOODNESS OF FIT

		Regularizer Lambda				
		0.01	0.10	1.00	10.00	100.00
Polynomial Degree	0	-5724.607420000	-5724.607420000	-5724.607420000	-5724.607420000	-5724.607420000
	1	-1683.188160000	-1579.786540000	-814.146627000	-693.218113000	-4541.870560000
	2	-26.816775800	-554.819952000	-1062.003530000	-1017.106060000	-287.752809000
	3	-142.726112000	-102.674688000	-15.314375000	-481.823343000	-692.565182000
	4	-8.822655790	-46.378206400	-49.728552600	-12.682515800	-118.132866000
	5	-0.737101495	0.959729915	-8.546933230	-15.507427600	-9.296941640
	6	0.813430417	0.667409511	0.898374191	0.121916267	-2.579981500
	7	0.999930743	0.994728304	0.977901344	0.984677344	0.976626545
	8	0.999999965	0.999999983	0.999999823	0.999998096	0.999999193
	9	0.999999994	0.999999986	0.999999988	0.999998119	0.999968696
	10	0.999999989	0.999999931	0.999999247	0.999999095	0.999997149

FIT OF THE MODEL



ƛ = 100 : Train and Test error both are very high : UNDERFIT

ƛ = 0.01 : both are stable and minimum : BEST-FIT.

Same has been Indicated by the goodness fit measure

FINAL POLYNOMIAL REGRESSION MODEL

Polynomial Coefficients : 0.00000000e+00 1.49778155e-01 -4.74272849e-01 8.62626551e-01 -6.74076261e-01 -1.03036688e+00 1.91346890e+00 7.65801496e+00 4.35992945e+00 -7.38110718e-03

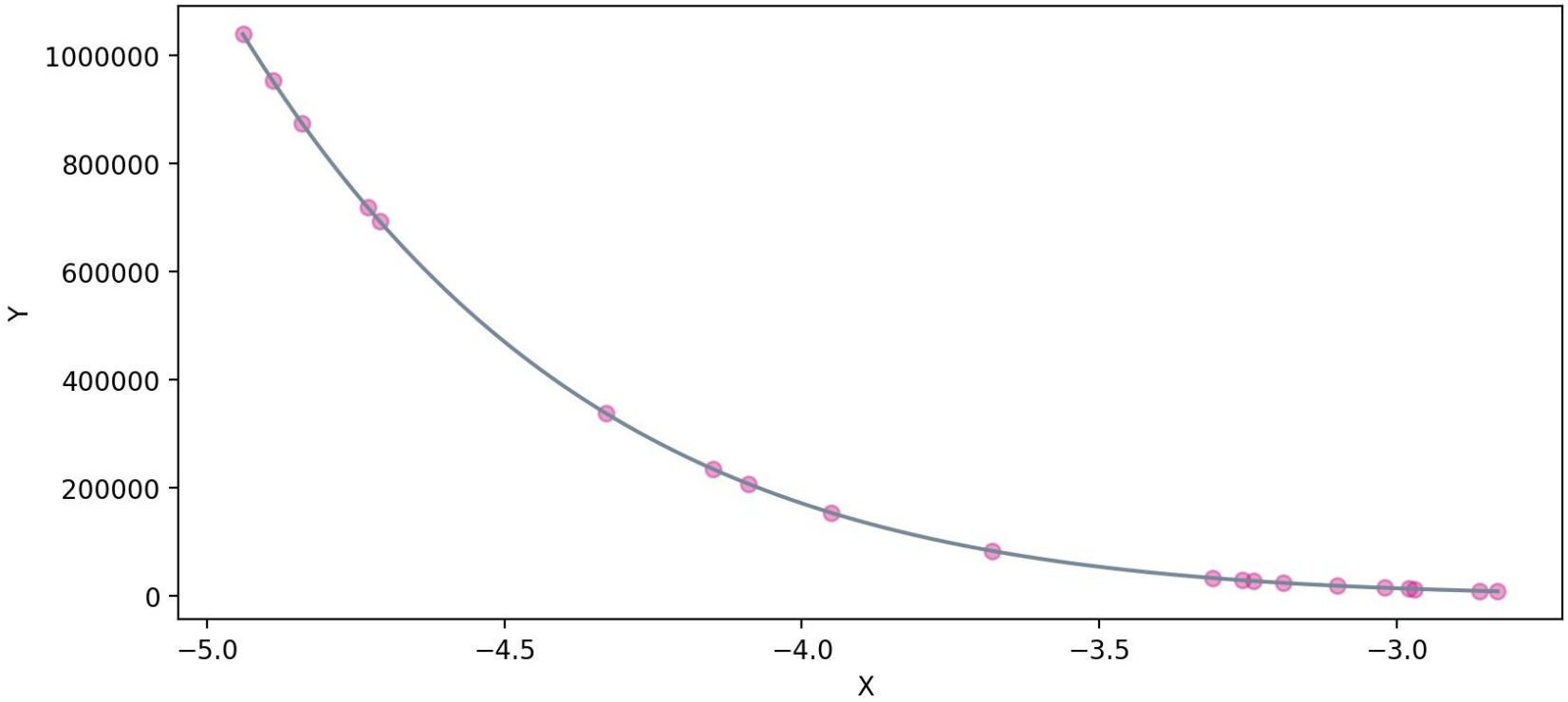
Intercept : -3.38420919

Variance : 0.06180557404041734 5

Optimal degree: 9

Optimal lambda: 0.01

FITTING DATA TO MODEL



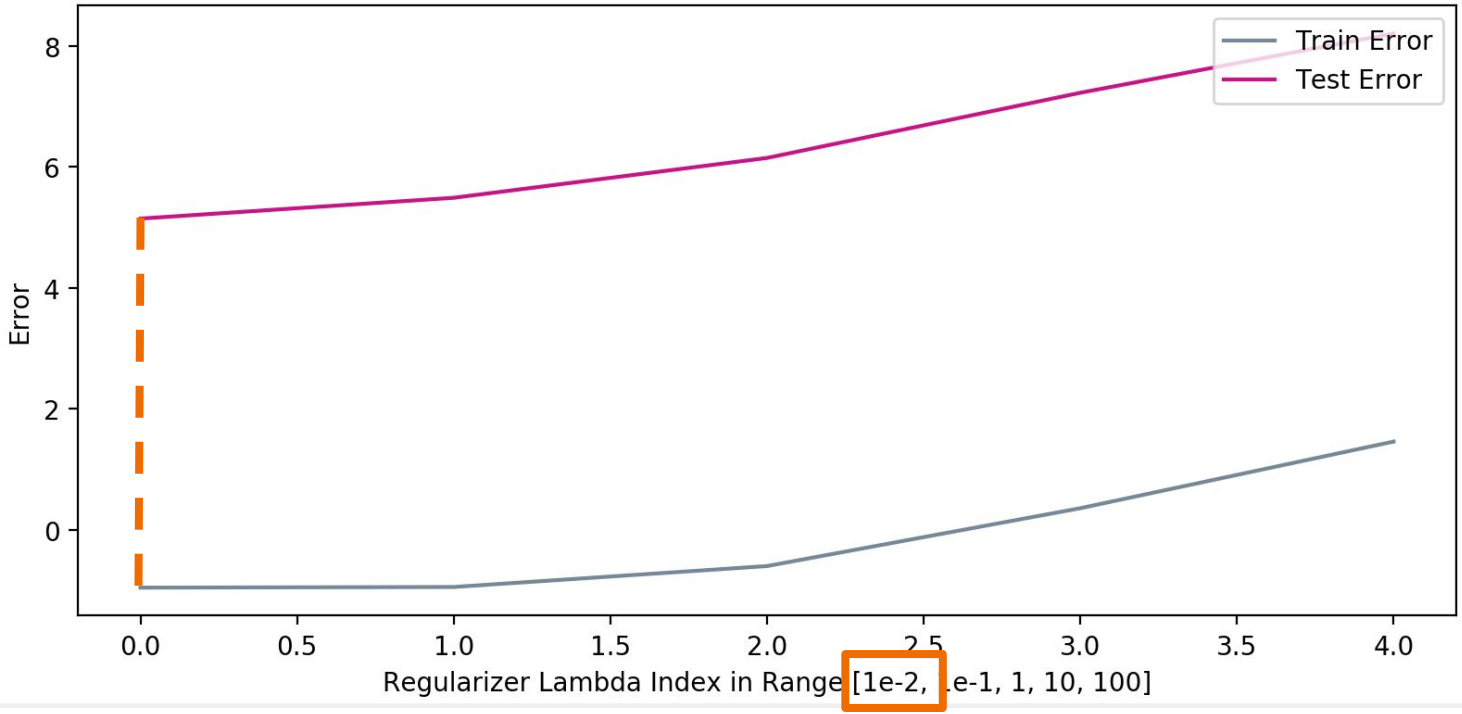
100 Data Points

Complete data

GOODNESS OF FIT

		Regularizer Lambda				
		0.01	0.10	1.00	10.00	100.00
Polynomial Degree	0	-1891070290.000000000	-1891070290.000000000	-1891070290.000000000	-1891070290.000000000	-1891070290.000000000
	1	-1687842180.000000000	-1687864500.000000000	-1688087400.000000000	-1690290350.000000000	-1709984420.000000000
	2	-2144470870.000000000	-2144252190.000000000	-2142067020.000000000	-2120372590.000000000	-1917906430.000000000
	3	-1854092660.000000000	-1853941770.000000000	-1852432100.000000000	-1837269650.000000000	-1685220520.000000000
	4	-473138261.000000000	-472353210.000000000	-464587683.000000000	-394683357.000000000	-85832520.200000000
	5	-287909373.000000000	-287433794.000000000	-282742238.000000000	-241220636.000000000	-53371130.800000000
	6	-19473336.800000000	-19300064.700000000	-17659495.600000000	-7517913.020000000	-1677431.630000000
	7	-9214986.260000000	-9113241.580000000	-8177785.390000000	-3199821.660000000	-2274609.540000000
	8	0.993314505	0.992681348	0.948701564	-0.245267120	-3.804268960
	9	0.993302486	0.992233554	0.908559214	-0.376771174	-4.054807530
	10	0.993464808	0.992987838	0.875984954	-0.308628031	-4.274547820

FIT OF THE MODEL



ƛ = 100 : Train and Test error both are very high : UNDERFIT

ƛ = 0.01 : both are stable and minimum : BEST-FIT after this test and train error starts increasing.

Same has been Indicated by the goodness fit measure

FINAL POLYNOMIAL REGRESSION MODEL

Polynomial Coefficients : 0.00000000e+00 6.93550528e+00 8.78616548e+00 7.08446796e+00 5.45249467e+00 5.80063154e+00 5.70665822e+00 8.70003833e+004.49965071e+00 -1.78521148e-06 6.17856419e-06

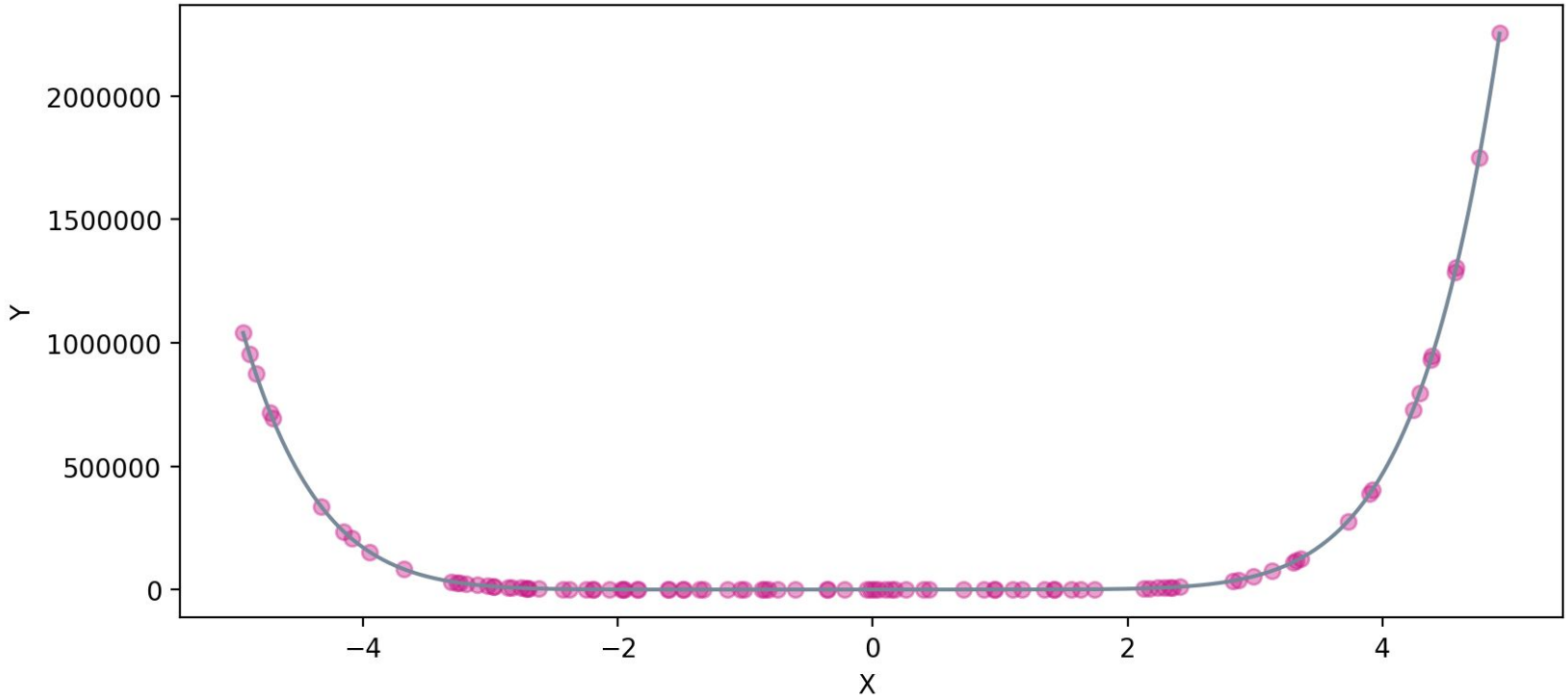
Intercept : 7.17396127

Variance : 0.1108132442535524

Optimal degree: 10

Optimal lambda: 0.01

FITTING DATA TO MODEL



Genomic Sequence Analysis

Thank You