Assignment: 1
CMPT: 726 (Machine Learning) **Student Id: 301399532**

Q-1) Probabilistic Modeling

1) What is the parameter that explains the behaviour of the die in this case (in analogy to the μ for the coin)?

- μ is a vector of probabilities that a certain outcome will occur. In this case probability of occurrence of any one side of a dice is called as μ. Here μ_6 = (μ1, μ2, μ3, μ4, μ5, μ6).

2) What is the value of the parameter for a fair die (equal probability of rolling any number)?

- Value of parameter for a fair die is a vector of probabilities of rolling a particular digit. Thus value of μ vector for a fair die is given as μ_6 = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6).

3) What is the value of the parameter for a die that always rolls a 2?

- For a biased die that always rolls 2 the probability distribution over 2 will be one and over other outcomes will be zero. μ_6 = (0, 1, 0, 0, 0, 0).

4) Specify the domain of the parameter – which settings of the parameter are valid.

- The parameter μ_K can take any values between 0 and1 including both, hence creating a domain of [0, 1]. For k the domain is [1, 6]. Following settings of parameters are valid:
    i. Value of parameter cannot be less than 0.
    ii. Value of parameter cannot be greater than 1.
    iii. Sum of all components of μ should be equal to 1.

Q-2) Weighted Squared Error

TIRTH J. PATEL STUDENT NO: 301399532

CMPT: 726
(Assignment: 1)

Q.2 To derive optimal weights w given the weighted sum of squares error function.

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

Taking derivative w.r.t w

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

$$\nabla E(w) = -\sum_{n=1}^{N} (\phi(x_n))^T \alpha_n t_n + \sum_{n=1}^{N} w^T \phi(x_n)(\phi(x_n))^T \alpha_n$$

Setting derivative to zero

$\because \nabla E(w) = 0$

$$0 = -\sum_{n=1}^{N} (\phi(x_n))^T \alpha_n t_n + \sum_{n=1}^{N} w^T \phi(x_n)(\phi(x_n))^T \alpha_n$$

$$\therefore \sum_{n=1}^{N} (\phi(x_n))^T \alpha_n t_n = \sum_{n=1}^{N} w^T (\phi(x_n))^T \alpha_n \phi(x_n)$$

$$\therefore \Phi \alpha t^T = w^T \Phi \alpha \Phi^T$$

which can be written as:

$$\Phi^T \alpha t = w \Phi^T \alpha \Phi$$

$\Rightarrow$ Hence $\boxed{w = (\Phi^T \alpha \Phi)^{-1}(\Phi^T \alpha t)}$

where $\Phi = \begin{bmatrix} \phi_0(x_1) & \phi_1(x_1) & \dots & \phi_n(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \dots & \phi_n(x_2) \\ \vdots & \vdots & & \\ \phi_0(x_n) & \phi_1(x_n) & \dots & \phi_n(x_n) \end{bmatrix}$, $\begin{bmatrix} \alpha_1 & 0 & 0 & \dots \\ 0 & \alpha_2 & 0 & \\ 0 & 0 & \alpha_3 & \\ \vdots & & & \ddots \alpha_n \end{bmatrix}$

Q-3) Training vs. Test Error

1) **Suppose we perform unregularized regression on a dataset. Is the validation error always higher than the training error? Explain**

- Most of the time it is the case that validation error is higher than training error, taking into consideration data is evenly distributed and it is split properly. However, there might be exceptions where the chosen validation set lies in the path or very near to the fitted curve. In this case validation error might be lower than that of training error. If validation is performed using cross-validation then choosing fewer folds with easy prediction can also lead to the above exception.

2) **Suppose we perform unregularized regression on a dataset. Is the training error with a degree 10 polynomial always lower than or equal to that using a degree 9 polynomial? Explain**.

- Yes, the training error with a degree 10 polynomial will always be less or equal to that of degree 9 polynomial. The reason for this is as we increase the degree of polynomial the degrees of freedom of curve increase which gives a better fit for the training set. Also, the higher degree polynomial includes lower degree polynomial, as a result, it would be at least as good as lower degree polynomial curve.

3) **Suppose we perform both regularized and unregularized regression on a dataset. Is the testing error with a degree 20 polynomial always lower using regularized regression compared to unregularized regression? Explain.**

- No, it is not always the case. It depends on how the testing dataset is distributed. Performing regularization removes the overfitting of the curve. However, if the testing data lies along the overfitted curve then performing regularization will increase the testing error. If the testing data is skewed then performing regularization will decrease.

Q-4) Basis Function Dependent Regularization

TIRTH J. PATEL STUDENT NO: 301399532

CMPT: 726
(Assigment: 1)

Q.4 Formula of the gradient for the regularized squared error loss function in this scenario.

→ Let J_1 be the set of indices having L_1 regularization and J_2 is set of indices having L_2 regularization.
λ_n is the tradeoff parameter for each weight w_n

→ Error equation can be written as:

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}(t_n - w^T\phi(x_n))^2 + \frac{1}{2}\sum_{i\in J_1}\lambda_i|w_i| + \frac{1}{2}\sum_{j\in J_2}\lambda_j|w_j|^2$$

Taking derivative:

$$\nabla E(w) = -\frac{2}{2}\sum_{n=1}^{N}(t_n - w^T\phi(x_n))\phi^T(x_n) + \frac{1}{2}\sum_{i\in J_1}\lambda_i + \frac{2}{2}\sum_{j\in J_2}\lambda_j w_j$$

$$\Longrightarrow \boxed{\nabla E(w) = -\sum_{n=1}^{N}(t_n - w^T\phi(x_n))\phi^T(x_n) + \frac{1}{2}\sum_{i\in J_1}\lambda_i + \sum_{j\in J_2}\lambda_j w_j}$$

Q-5) Regression

5.1) Getting Started

1) Which country had the highest child mortality rate in 1990? What was the rate?

- Niger had the highest child mortality rate in 1990 which was 313.7.

2) Which country had the highest child mortality rate in 2011? What was the rate?

- Sierra Leone had the highest child mortality rate in 2011 which was 185.3.

3) Some countries are missing some features (see original .xlsx/.csv spreadsheet). How is this handled in the function assignment1.load_unicef_data()?

- In assigment1.load_unicef_data() missing values are replaced by taking mean of the respective columns.

5.2) Polynomial Regression

1) Figure 1 shows the graph of training error vs testing error for polynomial of degree 1 to 6 without normalization, while figure 2 shows plot of errors with normalization.

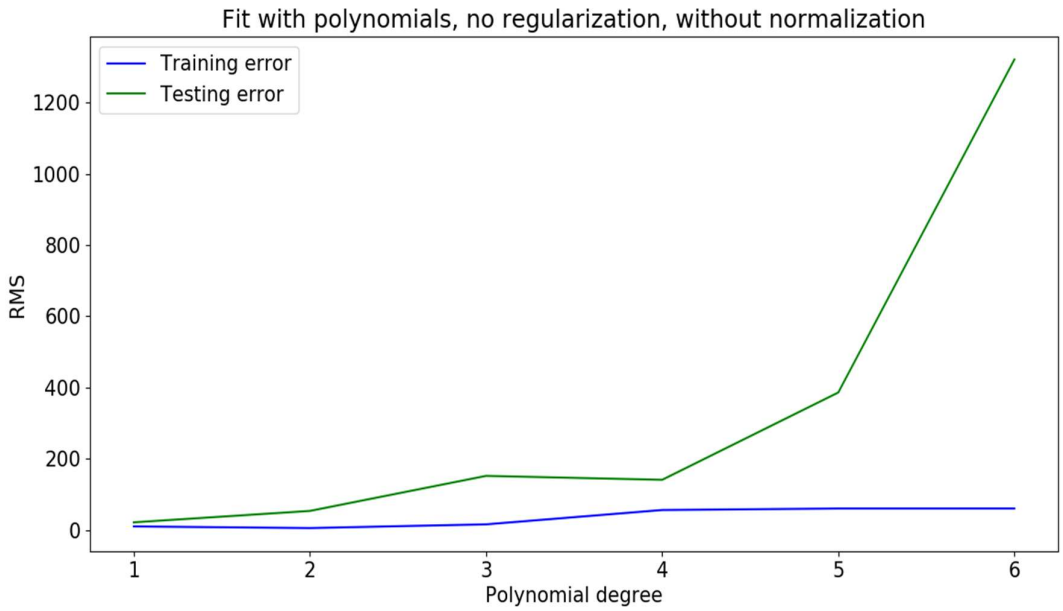

Figure: 1 Training vs testing error without normalization

- In the above graph the testing error increasing considerably after degree 4, this shows the polynomial curve over fits the data. Also after a certain point, the training error increases this is because the data is not normalized.

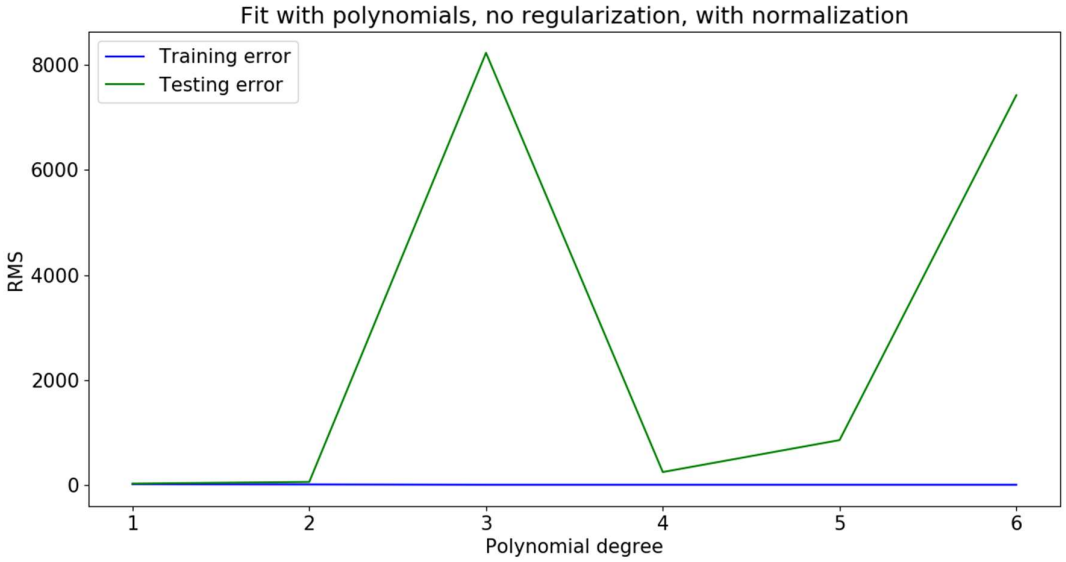

Figure: 2 Training vs testing error with normalization

2) Figure 3 shows the plot of training error and test error (in RMS error) for each of the 8 features (8-15) using a degree 3 polynomial with bias term and figure 4 shows plot of errors for same features but without bias term.

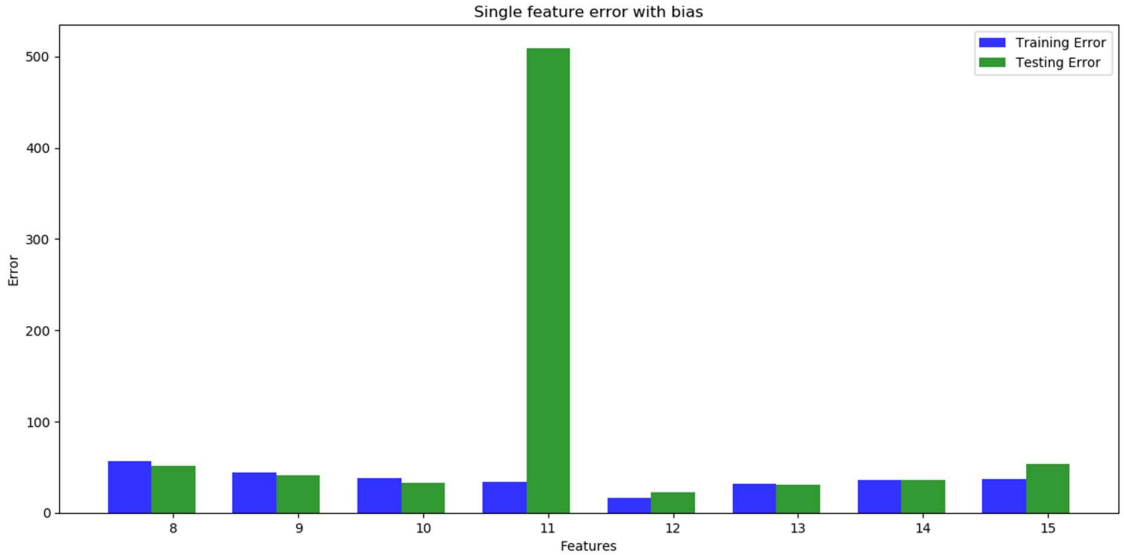

Figure: 3 Training vs testing error for feature 8-15 with bias

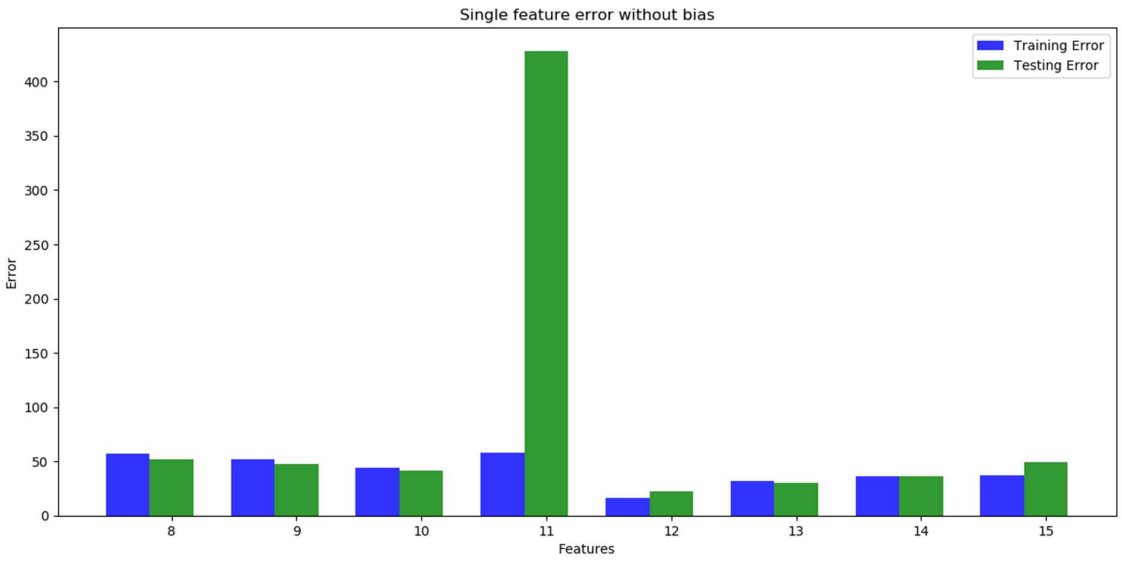

Figure: 4 Training vs testing error for feature 8-15 without bias

Figure 5, 6 and 7 shows plots of the fits for degree 3 polynomials for features 11 (GNI), 12 (Life expectancy), 13 (literacy) with bias.

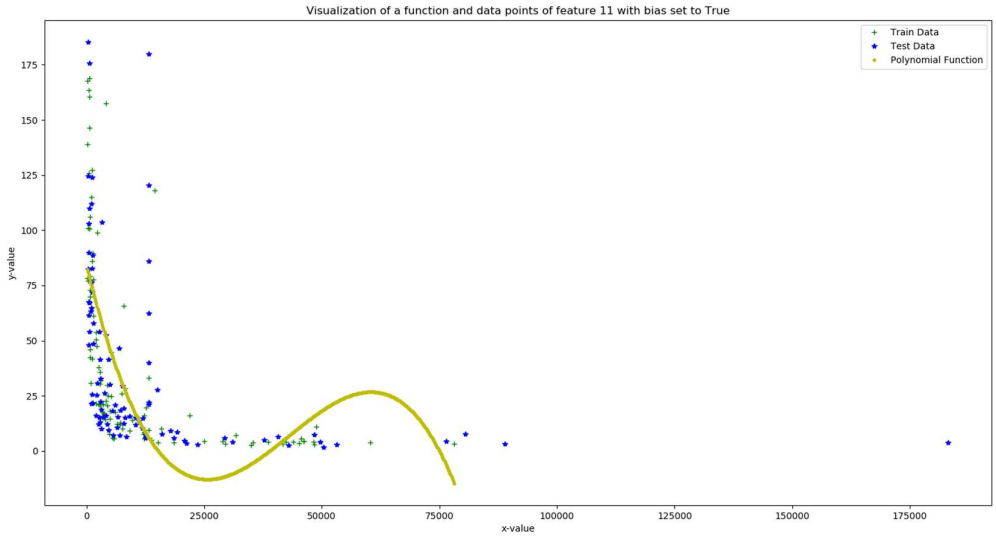

Figure 5: Training, testing points and polynomial regression curve for feature 11

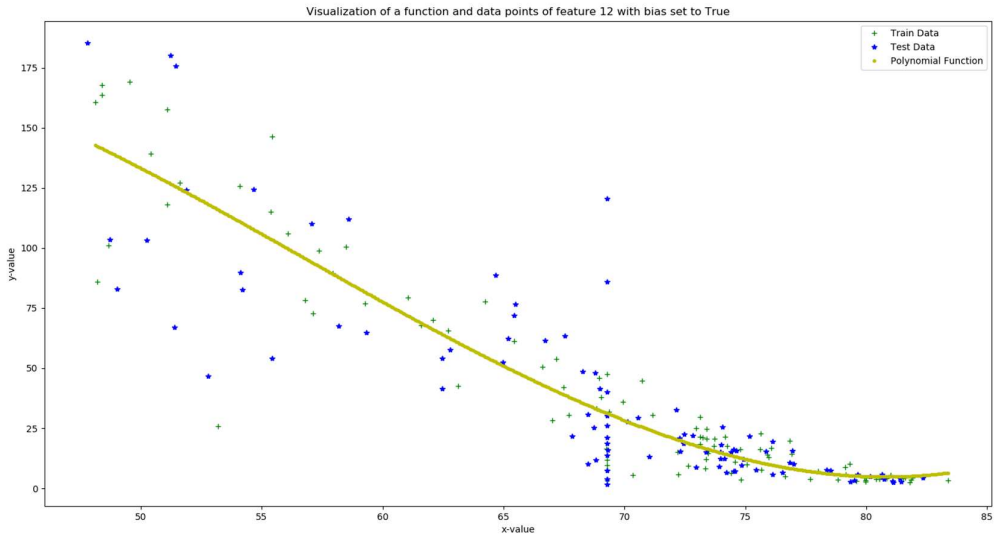

Figure 6: Training, testing points and polynomial regression curve for feature 12

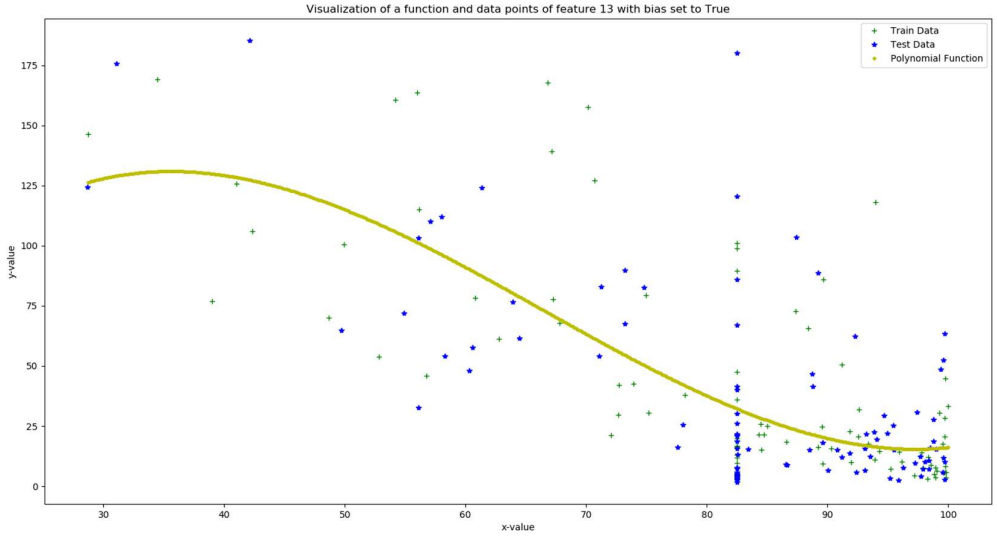

Figure 7: Training, testing points and polynomial regression curve for feature 13

Figure 8, 9 and 10 shows plots of the fits for degree 3 polynomials for features 11 (GNI), 12 (Life expectancy), 13 (literacy) without bias.

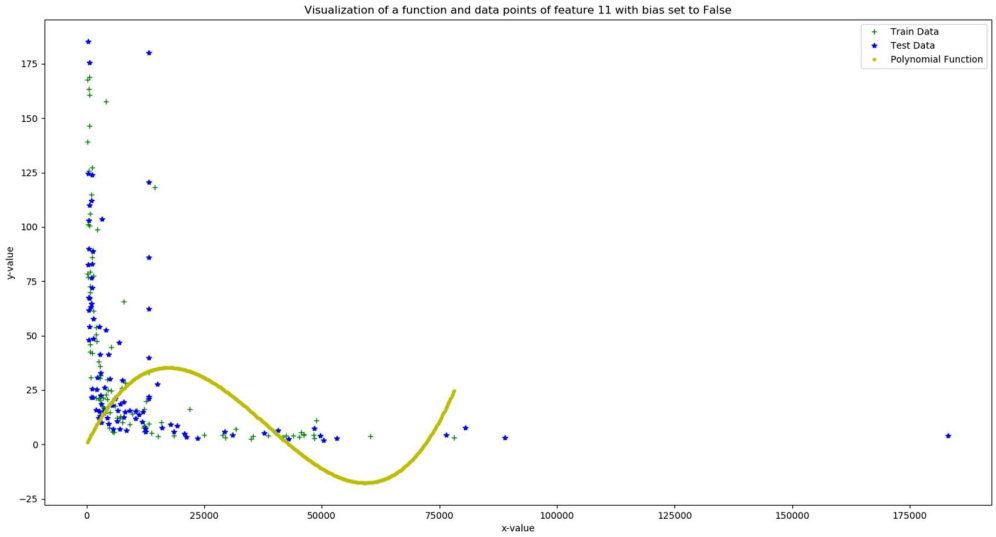

Figure 8: Training, testing points and polynomial regression curve for feature 11

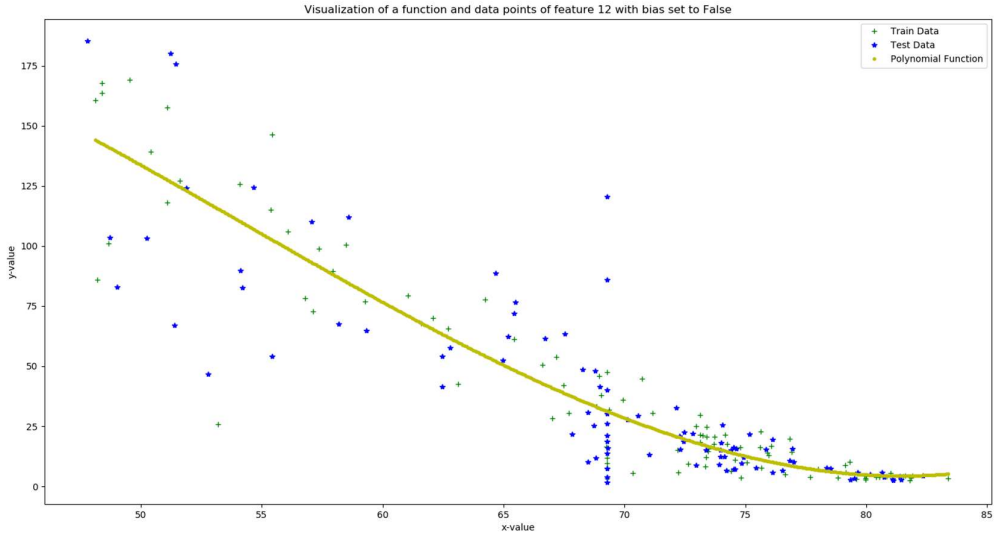

Figure 9: Training, testing points and polynomial regression curve for feature 12

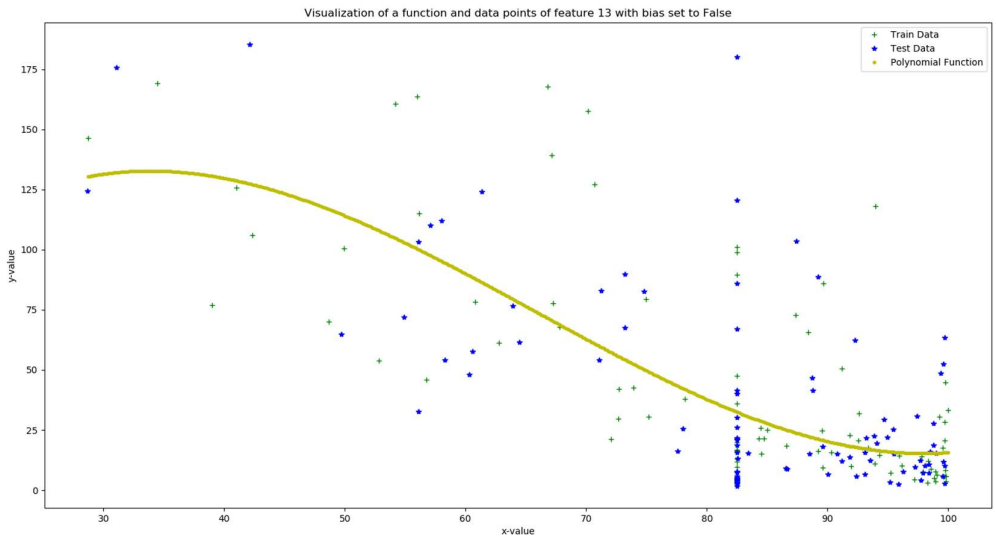

Figure 10: Training, testing points and polynomial regression curve for feature 13

5.3) Sigmoid Basis Function

1) Figure 11 shows regression using sigmoid basis functions for a single input feature (11^{th} feature). Two sigmoid basis functions, with μ = 100, 10000 and s = 2000.0 are taken. Features are not normalized and bias term is included.

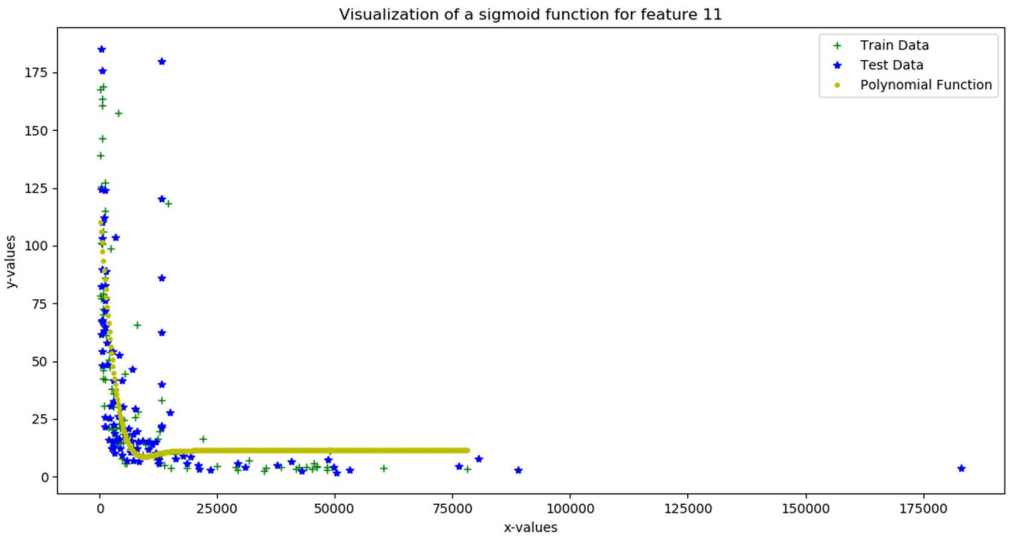

Figure 11: Plot of data points and regression model using sigmoid basis function for feature 11

✓ **Training error for the above sigmoid basis function is 28.457 and testing error is 33.806.**

5.4) Regularized Polynomial Regression

1) Figure 12 shows a plot of average validation set error versus λ for L2-regularized regression using λ = {0, .01, .1, 1, 10, 10^2, 10^3, 10^4}. Inputs are normalized and bias term is included. Also 10-fold cross-validation to decide on the best value for λ.

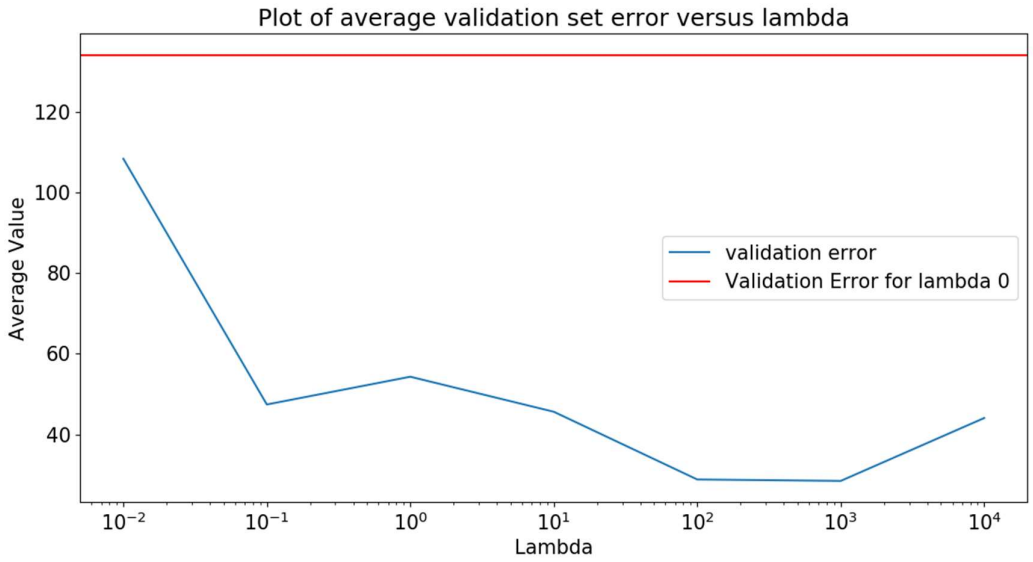

Figure 12: Plot of average validation set error versus λ

✓ **From the given plot it can be seen that when value of λ is 10^3, validation error is minimum. Hence the best value of λ is 10^3.**