Navishaa Agarwaal
2023101136

Assignment -2

Bias-Variance Trade off

Task 1 : Gradient Descent

a) Gradient descent is an iterative optimization algorithm used to minimise a function by adjusting its parameters.

b) For one independent and one dependant variable i.e. simple linear regression

$$\hat{y} = w_1 x + w_0$$

(w_1: parameter, w_0: bias)

x: independent variable
y: dependent variable

So, gradient descent tries to find w_1 & w_0 such that, the ~~minimum error~~ mean square error is least.

Let $\hat{y}$ be predicted value & y the actual value

$\hat{y} - y \Rightarrow$ error

w_0 & $w_1 \propto$ fitness of model

thus we define loss function to measure fitness of model with least square error.

For linear regression, cost function is MSE

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

Gradient of MSE : wrt slope (w_1) & intercept (w_0) is calculated (using partial differentiation)

$$\frac{\partial MSE}{\partial w_1} = \frac{2}{N} \sum_{i=1}^{N} x_i (\hat{y}_i - y_i)$$

$$\frac{\partial MSE}{\partial w_0} = \frac{2}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)$$

Now initially, we assign random values to w_1 & w_0, then update coefficients using gradients calculated as above. Update is done as follows:

$w_{1,new} = w_{1,old} - \alpha \frac{\partial MSE}{\partial w_1}$ α : learning rate

$w_{0,new} = w_{0,old} - \alpha \frac{\partial MSE}{\partial w_0}$

We go in opposite direction of gradient to reduce the error → ∴ -ve α.

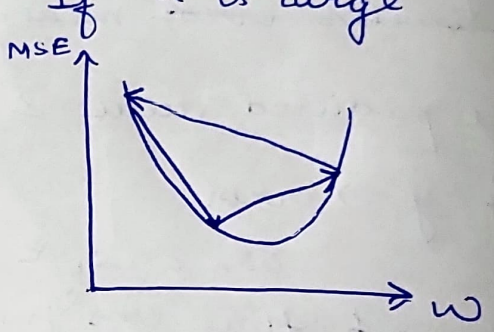
The above steps are repeated until $w_{1,new} \approx w_{1,old}$ and $w_{0,new} \approx w_{0,old}$, i.e. convergence of algorithm.

The final values represent $w_{1,final}$ & $w_{0,final}$

Note: learning rate greatly affects the speed of algorithm

If α is small

If α is large



b) Multiple independent variables and single dependant variable, multi-variable linear regression model

Here: $\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$

$\hat{y}$ ↙ dependent variable; w_0 ↓ intercept

x_i : independent variable

w_i : coefficient

We initialise w_i to some random values

Error ⇒ $\hat{y} - y$ ↳ actual value

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

~~$MSE = \frac{1}{N} \sum_{i=1}^{N} (y^{(n)} - f(x^{(n)}))^2$~~

Gradient is calculated with all w_i

$$\frac{\partial MSE}{\partial w_0} = \frac{2}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)$$

$$\frac{\partial MSE}{\partial w_j} = \frac{2}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i) x_j$$

Now update the coefficients based on α.

$$w_0^{new} = w_0^{old} - \alpha \frac{\partial MSE}{\partial w_0}$$

$$w_j^{new} = w_j^{old} - \alpha \frac{\partial MSE}{\partial w_j}$$

Repeat steps till $w_{new} \approx w_{old}$

The coefficients will be final coefficients

This is how gradient descent works for multiple independent variables and a dependent variable.

Task-2

To find bias^2, variance, MSE & verify MSE = Bias^2 + Variance (given $\sigma^2 = 0$)

x: [-3, -1, 0, 2, 3, 4]

y: [10, 2, 3, 8, 18, 30]

Models:

$f_1(x) = x^2 + x + 1$

$f_2(x) = 2x^2 + 2x + 2$

$f_3(x) = x^2 + 2x + 2$

$f_1(x)$:

$f_1(-3) = (-3)^2 + (-3) + 1 = 7$

$f_1(-1) = (-1)^2 + (-1) + 1 = 1$

$f_1(0) = 0^2 + 0 + 1 = 1$

$f_1(2) = 2^2 + 2 + 1 = 7$

$f_1(3) = 3^2 + 3 + 1 = 13$

$f_1(4) = 4^2 + 4 + 1 = 21$

$f_2(x)$: $f_2(-3) = 2(-3)^2 + 2(-3) + 2 = 14$

$f_2(-1) = 2(-1)^2 + 2(-1) + 2 = 2$

$f_2(0) = 2(0)^2 + 2(0) + 2 = 2$

$f_2(2) = 2(2)^2 + 2(2) + 2 = 14$

$f_2(3) = 2(3)^2 + 2(3) + 2 = 26$

$f_2(4) = 2(4)^2 + 2(4) + 2 = 42$

$f_3(x)$:

$f_3(-3) = (-3)^2 + 2(-3) + 2 = 5$

$f_3(-1) = (-1)^2 + 2(-1) + 2 = 1$

$f_3(0) = 0^2 + 2(0) + 2 = 2$

$f_3(2) = 2^2 + 2(2) + 2 = 10$

$f_3(3) = 3^2 + 2(3) + 2 = 17$

$f_3(4) = 4^2 + 2(4) + 2 = 26$

Let's define 3 tuples of 6 elements each representing values of $f_1(x), f_2(x), f_3(x)$ for our data set.

$f_1(x) = (7, 1, 1, 7, 13, 21)$

$f_2(x) = (14, 2, 2, 14, 26, 42)$

$f_3(x) = (5, 1, 2, 10, 17, 26)$

For finding bias^2:

Bias $= E_i[f_i(x)] - f(x)$ (as $\sigma^2 = 0$)

Bias $= E_i[f_i(x)] - y$ here $y = f(x) + \phi$

$\therefore, y = f(x)$

(this is tuple of 6 elements) (formula is for individual data point x) (i: across all models)

$\therefore$ Bias $(x) = E_i[f_i(x)] - y$

$= \frac{f_1(x) + f_2(x) + f_3(x)}{3} - y$ $[y = f(x)]$

for $x = [-3, -1, 0, 2, 3, 4]$

we will get 6 values of bias

Bias$(-3) = \frac{7+14+5}{3} - 10 = -\frac{4}{3}$

Bias$(-1) = \frac{1+2+1}{3} - 2 = -\frac{2}{3}$

Bias$(0) = \frac{1+2+2}{3} - 3 = -\frac{4}{3}$

Bias$(2) = \frac{7+14+10}{3} - 8 = \frac{1}{3}$

Bias$(3) = \frac{13+26+17}{3} - 18 = \frac{2}{3}$

Bias$(4) = \frac{21+42+26}{3} - 30 = -\frac{1}{3}$

Bias $= (-1.33, -0.67, -1.33, 2.33, 0.67, -0.33)$

Bias$^2 = (1.77, 0.44, 1.77, 5.44, 0.44, 0.11)$

Bias $= (-\frac{4}{3}, -\frac{2}{3}, -\frac{4}{3}, \frac{7}{3}, \frac{2}{3}, -\frac{1}{3})$

Bias$^2 = (\frac{16}{9}, \frac{4}{9}, \frac{16}{9}, \frac{49}{9}, \frac{4}{9}, \frac{1}{9})$

Avg. bias$^2 = \frac{1}{6}\left(\frac{16}{9} + \frac{4}{9} + \frac{16}{9} + \frac{49}{9} + \frac{4}{9} + \frac{1}{9}\right)$

$= \frac{90}{54} = \frac{5}{3}$

For finding variance:

$$\text{Variance} = E_i\left[\left(\hat{f}_i(x) - E_i[\hat{f}_i(x)]\right)^2\right]$$

Let us first find $E_i[\hat{f}_i(x)] = \frac{f_1(x) + f_2(x) + f_3(x)}{3}$

$$\text{Mean} = E_i[\hat{f}_i(x)] = \left(8, \frac{24}{3}, \frac{4}{3}, \frac{5}{3}, \frac{31}{3}, \frac{54}{3}, \frac{89}{3}\right)$$

(Already calculated for bias)

Variance: $E\left[\hat{f}_i(x) - \text{mean}(x)^2\right] = E[\hat{f}_i(x)^2] - \text{mean}(x)^2$

$E[(x-\mu)^2] = E[x^2] - \mu^2$ where $\mu = E[x]$

$$\text{var} = \frac{f_1(x)^2 + f_2(x)^2 + f_3(x)^2}{3} - \text{mean}(x)^2$$

$$\text{Var}(-3) = \frac{7^2 + 14^2 + 5^2}{3} - \left(\frac{26}{3}\right)^2 = \frac{134}{9}$$

$$\text{Var}(-1) = \frac{1^2 + 2^2 + 1^2}{3} - \left(\frac{4}{3}\right)^2 = \frac{2}{9}$$

$$\text{Var}(0) = \frac{1^2 + 2^2 + 2^2}{3} - \left(\frac{5}{3}\right)^2 = \frac{2}{9}$$

$$\text{Var}(2) = \frac{7^2 + 14^2 + 10^2}{3} - \left(\frac{31}{3}\right)^2 = \frac{74}{9}$$

$$\text{Var}(3) = \frac{13^2 + 26^2 + 17^2}{3} - \left(\frac{56}{3}\right)^2 = \frac{266}{9}$$

$$\text{Var}(4) = \frac{21^2 + 42^2 + 26^2}{3} - \left(\frac{89}{3}\right)^2 = \frac{722}{9}$$

$$\text{Variance} = \left(\frac{134}{9}, \frac{2}{9}, \frac{74}{9}, \frac{266}{9}, \frac{722}{9}\right)$$

~~Variance = (14.83, 0.23, 0.23, 8.29, 29.43, 80.02)~~

$$\text{Avg. variance} = \frac{1}{6}\left(\frac{134}{9} + \frac{2}{9} + \frac{2}{9} + \frac{74}{9} + \frac{266}{9} + \frac{722}{9}\right) = \frac{200}{9}$$

MSE $E_i[(y - \hat{f}(x))^2]$

$$\text{MSE}: \frac{(y - f_1(x))^2 + (y - f_2(x))^2 + (y - f_3(x))^2}{3}$$

$$MSE(-3) = \frac{(10-7)^2 + (10-14)^2 + (10-5)^2}{3} = \frac{3^2+4^2+5^2}{3} = \frac{50}{3}$$

$$MSE(-1) = \frac{(2-1)^2 + (2-2)^2 + (2-1)^2}{3} = \frac{1^2+0+1^2}{3} = \frac{2}{3}$$

$$MSE(0) = \frac{(3-1)^2 + (3-2)^2 + (3-2)^2}{3} = \frac{2^2+1+1}{3} = 2$$

$$MSE(2) = \frac{(8-7)^2 + (8-14)^2 + (8-10)^2}{3} = \frac{1^2+6^2+2^2}{3} = \frac{41}{3}$$

$$MSE(3) = \frac{(18-13)^2 + (18-26)^2 + (18-17)^2}{3} = \frac{5^2+8^2+1^2}{3} = 30$$

$$MSE(4) = \frac{(30-21)^2 + (30-42)^2 + (30-26)^2}{3} = \frac{9^2+12^2+4^2}{3} = \frac{241}{3}$$

~~MSE = (16.67, 0.67, 2, 5.2, 30, 80.33)~~

Bias² + variance = 16.67, 0.67, 2,

$$MSE = \left(\frac{50}{3}, \frac{2}{3}, 2, \frac{41}{3}, 30, \frac{241}{3}\right)$$

$$Bias^2(x) + var(x) = \left(\frac{16}{9}, \frac{4}{9}, \frac{16}{9}, \frac{49}{9}, \frac{4}{9}, \frac{1}{9}\right) + \left(\frac{134}{9}, \frac{2}{9}, \frac{2}{9}, \frac{74}{9}, \frac{266}{9}, \frac{722}{9}\right)$$

$$= \left(\frac{50}{3}, \frac{2}{3}, 2, \frac{41}{3}, 30, \frac{241}{3}\right)$$

$$Avg.\ MSE = \frac{1}{6}\left(\frac{50}{3} + \frac{2}{3} + 2 + \frac{41}{3} + 30 + \frac{241}{3}\right)$$

$$= \frac{215}{9}$$

$$E[Bias^2(x)] + E[var(x)] = \frac{5}{3} + \frac{200}{9} = \frac{215}{9}$$

$$E[MSE(x)] = \frac{215}{9}$$

Bias², variance & MSE are calculated over a single data point and arranged over all the models for a data point, & final bias², var. & MSE are calculated by averaging over all data points.

Task 3

Tabulate the values of bias and variance and also write a detailed analysis and observation explaining how bias and variance change as you vary your function classes from degree 1 to 10.

Degree	Bias^2	Variance	MSE
1	2.571206699582909	0.07907299269422811	2.6502796922771363
2	2.533979379020787	0.15923339357643523	2.693212772597222
3	1.989466347400548	0.17976081360132884	2.169227161001877
4	1.7106423637038444	0.24976957330327543	1.960411937007119
5	1.614251373097558	0.3893389114434541	2.003590284541012
6	1.630688275435285	0.5342316719538416	2.1649199473891265
7	1.7350669338688889	0.6014761189405787	2.336543052809467
8	2.180187892591282	1.5221805959939465	3.7023684885852277
9	3.9441253056304206	1.2751830048290507	5.21930831045947
10	4.696250869690388	2.089372093053782	6.785622962744169

Bias-Variance Tradeoff Analysis

1. Low-Degree Polynomials (Degree 1-3)

- High Bias, Low Variance
- The model is too simple to capture complex relationships in data.
- Predictions are stable across different training subsets.
- MSE is relatively high due to poor model complexity.

2. Medium-Degree Polynomials (Degree 4-6)

- Moderate Bias, Moderate Variance
- Bias decreases as the model captures more complexity while avoiding extreme fluctuations.
- Variance increases slightly but remains controlled.
- **Optimal Tradeoff:** These models balance bias and variance effectively.
- MSE reaches its minimum around degree 4, indicating the best trade-off.

3. High-Degree Polynomials (Degree 7-10)

- Low Bias, High Variance
- The model fits training data very well.
- Predictions vary significantly between models.
- The model captures noise rather than actual patterns.
- MSE increases, confirming that overfitting leads to poor generalization.

Discuss underfitting and overfitting behavior observed across different degrees and explain how model complexity affects the bias-variance tradeoff.

Underfitting and Overfitting Behavior

- **Underfitting (Low-Degree Polynomials)**: Underfitting occurs at lower polynomial degrees (1-2), where the model is too simple to capture the underlying pattern of the data, leading to high bias and low variance. These models fail to represent the complexity of the dataset, resulting in poor predictions.
- **Overfitting (High-Degree Polynomials)**: Overfitting is observed at higher polynomial degrees (7-10), where the model becomes excessively complex and starts capturing noise rather than the actual pattern. These models fit training data too closely, capturing noise rather than underlying patterns. This results in high variance, making them sensitive to small changes in data and reducing generalization.
- **Impact of Model Complexity**: Increasing polynomial degree reduces bias but increases variance. An optimal model achieves a balance where bias and variance are minimized to ensure accurate predictions.

Conclusion

- A moderate polynomial degree (around 4-6) provides the best balance between bias and variance.
- Too simple (low-degree) models underfit, while too complex (high-degree) models overfit.
- The findings suggest that moderate complexity models are ideal for Joshita's ad-spending analysis.

Recommendations

- Choose polynomial degrees between 4-6 for best accuracy.
- Regularization techniques can help control overfitting for higher-degree models.
- Further experiments with larger datasets can refine model selection.
- By selecting an appropriate polynomial degree, Joshita can make more accurate sales predictions, optimizing ad-spending strategies effectively.

Task 4

Tabulate the values of irreducible error for the models in Task 3 and also write a detailed report explaining why or why not the value of irreducible error changes as you vary your class function.

Degree	Bias^2	Variance	MSE	Irreducible error	Irreducible error-rounded off
1	2.571206699582909	0.07907299269422811	2.6502796922771363	-8.881784197001252e-16	0
2	2.533979379020787	0.15923339357643523	2.693212772597222	-4.440892098500626e-16	0
3	1.989466347400548	0.17976081360132884	2.169227161001877	0.0	0
4	1.7106423637038444	0.24976957330327543	1.960411937007119	-8.881784197001252e-16	0
5	1.614251373097558	0.3893389114434541	2.003590284541012	0.0	0
6	1.630688275435285	0.5342316719538416	2.1649199473891265	0.0	0
7	1.7350669338688889	0.6014761189405787	2.336543052809467	-6.661338147750939e-16	0
8	2.180187892591282	1.5221805959939465	3.7023684885852277	-8.881784197001252e-16	0
9	3.9441253056304206	1.2751830048290507	5.21930831045947	-8.881784197001252e-16	0
10	4.696250869690388	2.089372093053782	6.785622962744169	-1.7763568394002505e-15	0

Irreducible Error Analysis

The irreducible error remains approximately zero across all polynomial degrees. This suggests that the dataset has minimal inherent noise, meaning that the observed errors are primarily due to bias-variance trade-offs rather than unavoidable noise in the data.

Since the irreducible error is theoretically the lowest possible error that any model can achieve, its near-zero values indicate that our dataset is relatively clean and does not contain significant random noise. Therefore, variations in model performance are mostly due to underfitting and overfitting behaviors rather than intrinsic data limitations.

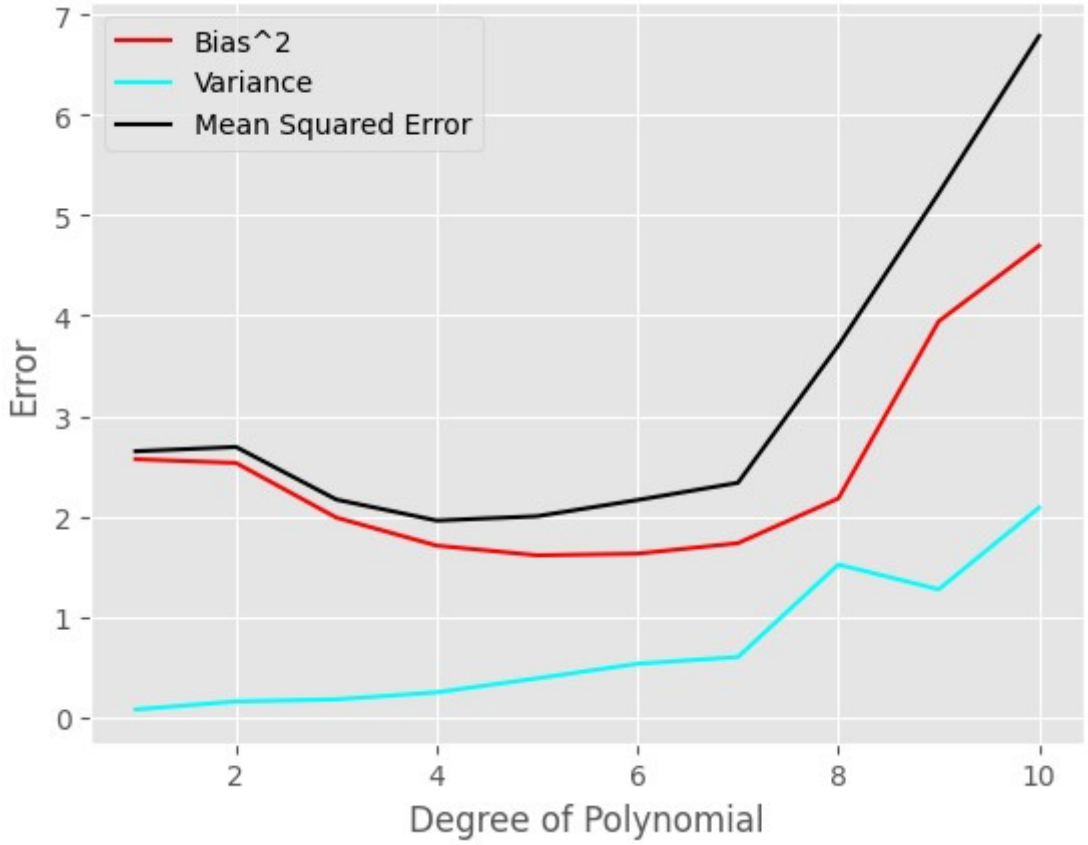
This study highlights the importance of selecting an appropriate model complexity to minimize generalization error and achieve the best trade-off between bias and variance.

Task 5

Based on the variance, bias, and total error calculated in earlier tasks, plot the Bias^2-Variance tradeoff graph and write your observations in the report with respect to underfitting, overfitting, and also comment on the type of data just by analyzing the Bias2-Variance plot.

Bias-Variance Tradeoff Graph and Observations

The Bias-Variance tradeoff is visualized in the graph below, showing how bias, variance, and MSE change as the degree of the polynomial increases.



Key Observations:

- **Bias^2 (Red Line):** Decreases initially as model complexity increases, reaching a minimum at moderate polynomial degrees. However, it starts to rise again at higher degrees due to overfitting.
- **Variance (Cyan Line):** Gradually increases with model complexity, rising significantly for higher-degree polynomials.
- **MSE (Black Line):** Initially decreases, reaches a minimum at an optimal degree, and then increases due to overfitting.

From the plot, we observe:

- **Underfitting at low degrees (1-2):** High bias, low variance, and high MSE indicate that the model is too simple.
- **Optimal performance at moderate degrees (3-6):** Bias and variance are well-balanced, leading to the lowest MSE.
- **Overfitting at high degrees (7-10):** Bias increases slightly, but variance grows significantly, leading to high MSE.

By analyzing this plot, we can infer that the dataset has a well-defined pattern with minimal noise, making it ideal for polynomial regression with an appropriate degree. Choosing an optimal polynomial degree (around 4-6) helps achieve the best generalization performance.

This study highlights the importance of selecting an appropriate model complexity to minimize generalization error and achieve the best trade-off between bias and variance.

Task 6

(a) Fit a polynomial regression model of degree 10 on the given training data.Compute the Mean Squared Error (MSE) on the test set and report the result.

(b) Implement a regularized regression model using either Ridge Regression or Lasso Regression. Train the model with a polynomial of degree 10 using regularization.

(c) Compute and report the MSE on the test set after applying regularization.

(d) Compare the MSE values obtained in parts (a) and (c). Discuss the impact of regularization on overfitting and model performance.

A polynomial regression model of degree 10 was trained on a given dataset. The Mean Squared Error (MSE) was calculated on the test set to assess the impact of overfitting. To mitigate overfitting, regularization techniques were implemented:

1. **Polynomial Regression (No Regularization)**
2. **Ridge Regression (α=1.0)**
3. **Lasso Regression (α=0.1)**

Each model's performance was evaluated by computing the test set MSE.

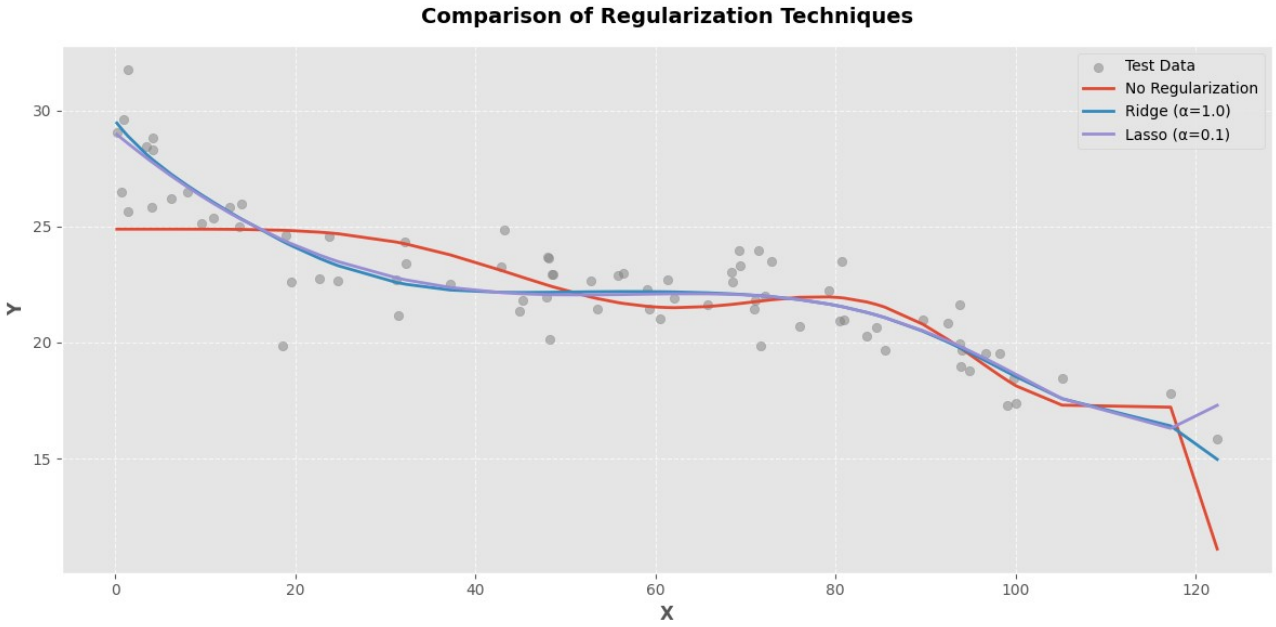
Results

```
Comparison of MSE values:
Polynomial Regression MSE: 3.476470277813317
Ridge Regression MSE: 1.62656233148536
Lasso Regression MSE: 1.6413537629867156
```

Observations and Analysis

1. **Polynomial Regression (No Regularization)**
    - Results in high MSE due to overfitting.
    - The model captures noise rather than the true pattern.
2. **Ridge and Lasso Regression**
    - Both methods significantly reduce MSE, improving generalization.

- Ridge Regression provides smoother solutions by reducing coefficient magnitudes.
- Lasso Regression can drive some coefficients to zero, performing feature selection.



Conclusion

Regularization effectively mitigates overfitting in polynomial regression models. Ridge and Lasso Regression help balance model complexity by preventing excessive variance, leading to improved generalization and lower test errors.
