MDL ASSIGNMENT - 2

Name - Varun Gupta
Roll No - 2023101108

Task and Task 2 are at the end of the report.

Task 3

The following table shows the value for bias , variance and MSE for different types of polynomial equation

Bias-Variance Trade-off Metrics:

Degree	Bias²	Variance	MSE
1	2.5705752217	0.0719870009	2.6425622227
2	2.5336615747	0.1204423401	2.6541039148
3	2.0149890221	0.2003840653	2.2153730874
4	1.7492114984	0.2822000902	2.0314115886
5	1.6255842674	0.4992947499	2.1248790173
6	1.6229782024	0.6833792933	2.3063574957
7	1.7129324590	0.9074129359	2.6203453949
8	2.4828464050	2.0154446887	4.4982910937
9	4.5432052556	1.9272324919	6.4704377476
10	5.8334658868	3.0792576905	8.9127235773

Bias emerges when a complex real-world problem is approximated using a simpler model. As the polynomial degree increases, enhancing the model's complexity, bias generally decreases since higher-degree polynomials can better capture the training data's nuances. However, a slight increase in bias from degree 7 to 8 suggests potential overfitting.

Variance measures how much a model's predictions fluctuate for a given data point. As the polynomial degree increases, variance typically rises due to the model's sensitivity to minor variations in the training data. Interestingly, a decrease in variance from degree 9 to 10 indicates an improvement in the model's ability to generalize to unseen data.

In summary, the bias-variance tradeoff highlights that increasing model complexity reduces bias but raises variance, and vice versa. Achieving the right balance is crucial to prevent both underfitting and overfitting.

Task 4

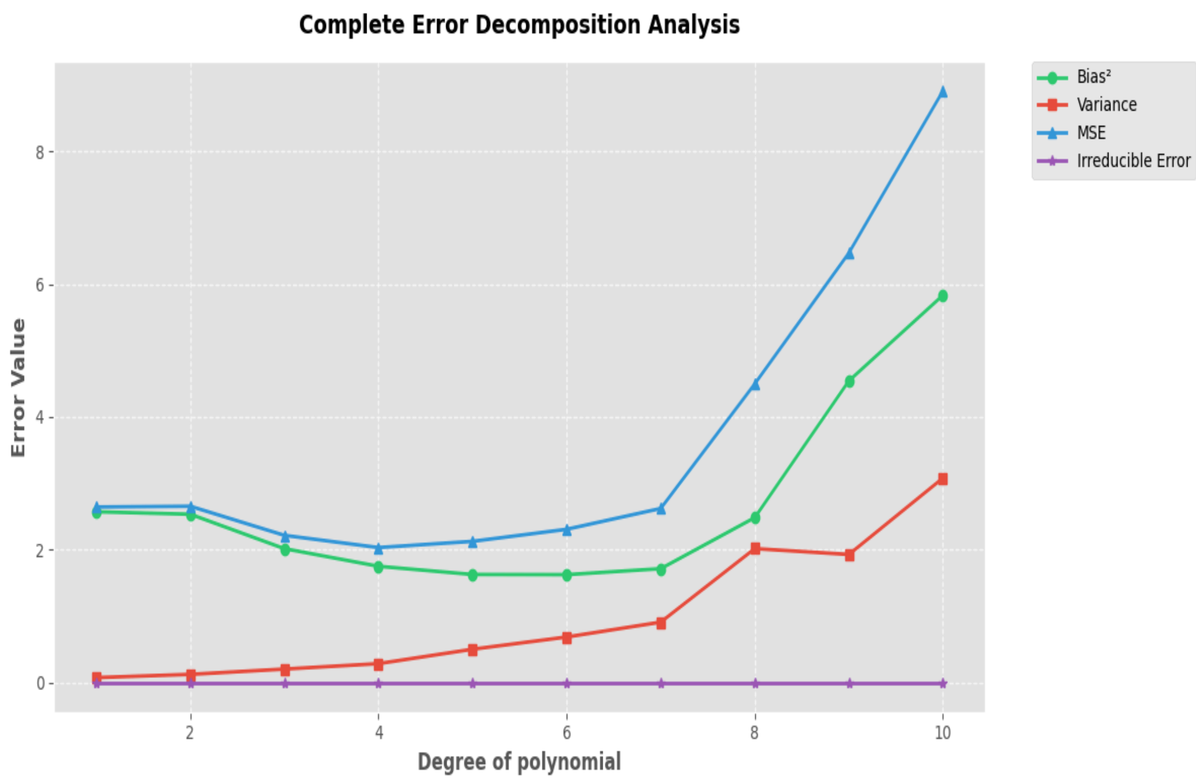
Irreducible Error Analysis:

Degree	Irreducible Error
1	-1.321859e-16
2	1.023660e-15
3	3.665471e-16
4	3.877107e-16
5	-5.107026e-16
6	-6.269291e-16
7	1.148213e-15
8	-1.977585e-17
9	1.448494e-15
10	3.241331e-16

Irreducible Error: The irreducible error remains constant across different polynomial degrees. This stability is expected because the irreducible error reflects the inherent noise or randomness in the data, which no model can eliminate. Regardless of changes in model complexity, as indicated by the polynomial degree, the irreducible error remains unaffected. It is an intrinsic property of the data, persistent even when the modeling approach is adjusted.

In conclusion, the irreducible error stays unchanged despite variations in the model, underscoring its nature as inherent noise that cannot be reduced by altering the model's complexity.

Task 5



Underfitting occurs when the model is too simplistic to capture the complexity of the data, resulting in high Bias^2 and Variance. This is typically seen in lower polynomial degrees (e.g., 1, 2, 3).

Overfitting occurs when the model closely fits the training data but exhibits high variability in predictions, leading to low Bias^2 and high Variance. This is often observed in higher polynomial degrees (e.g., 8, 9, 10).

The Bias^2-Variance plot is useful for identifying the optimal model complexity, which strikes a balance between bias and variance, minimizing errors and providing the best fit for the dataset.

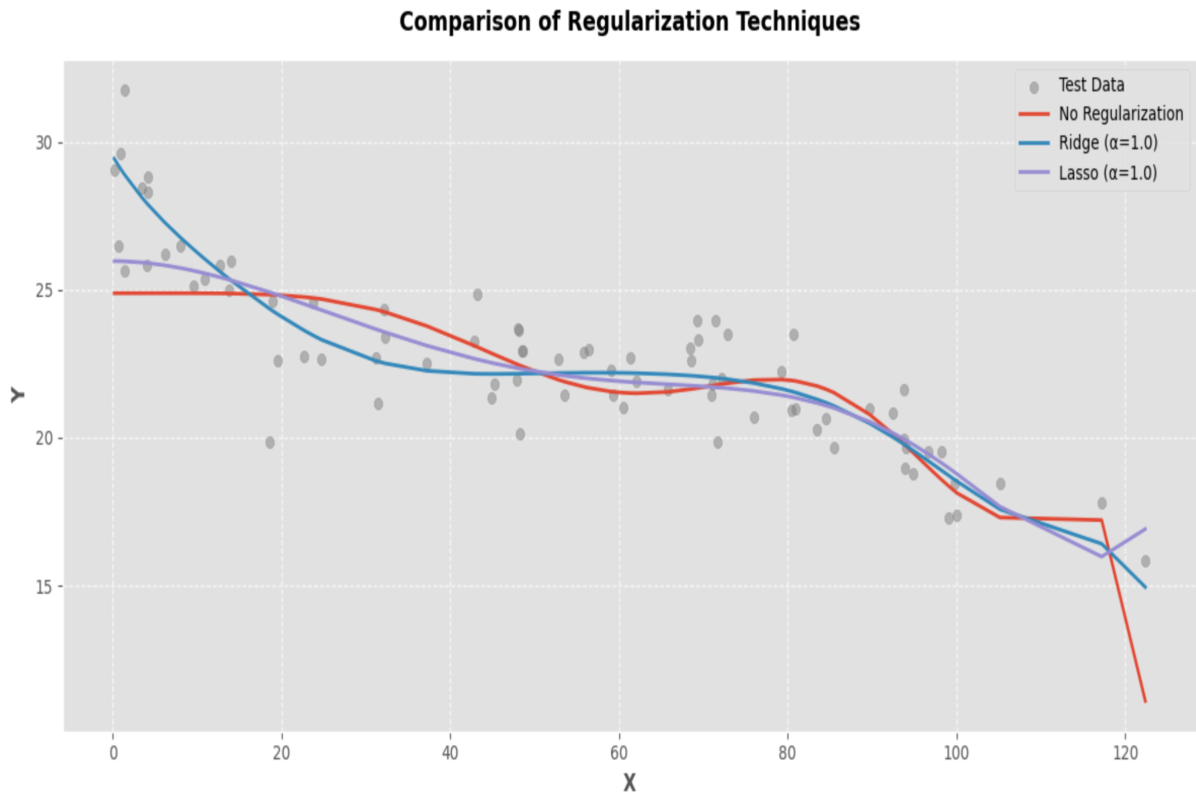
Therefore, it is essential to find an optimal degree where the model fits the data appropriately, avoiding both overfitting and underfitting. An iterative approach should be applied to determine the degree that best suits your dataset.

Task 6

```
Polynomial Regression (No Regularization):
MSE: 3.476470

Ridge Regression Results:
|         Alpha |          MSE |
|--------------:|-------------:|
|  0.1000000000 | 1.6266983740 |
|  1.0000000000 | 1.6265616301 |
| 10.0000000000 | 1.6252945866 |

Lasso Regression Results:
|         Alpha |          MSE |
|--------------:|-------------:|
|  0.1000000000 | 1.6413537630 |
|  1.0000000000 | 1.6698321931 |
| 10.0000000000 | 2.3777696994 |
```



Analysis of Regularization Impact

1. Polynomial Regression without Regularization

- High-degree polynomials (d=10) lead to overfitting by capturing noise in the training data, resulting in poor test data performance and high MSE.

2. Ridge Regression (L2 Regularization)

- Ridge reduces coefficient magnitudes, stabilizing predictions and improving generalization over unregularized models.
- The performance depends on α:
    - **Low α**: Similar to an unregularized model.
    - **High α**: Simpler model with more regularization.
    - Optimal α balances bias and variance.

3. Lasso Regression (L1 Regularization)

- Lasso performs feature selection by forcing some coefficients to zero, resulting in a sparser model.
- α controls sparsity:
  - **Higher α**: More coefficients become zero.
  - **Lower α**: Closer to unregularized model.

4. Comparative Analysis

- **Model Complexity**: Unregularized is most complex; Ridge retains all features; Lasso eliminates irrelevant features.
- **Generalization**: Both Ridge and Lasso improve test set performance, offering better stability and robustness.
- **Practical Implications**:
  - Use Ridge when all features matter.
  - Use Lasso for feature selection.
  - Select α through cross-validation.

For most cases, it is recommended to start with Ridge regression, as it stabilizes predictions and improves generalization. Cross-validation should be used to find the optimal value of α, which balances bias and variance for the best model performance. If feature selection is a priority, Lasso regression is ideal, as it reduces the model complexity by forcing some coefficients to zero. It is crucial to monitor both training and test errors to assess the model’s generalization ability and prevent overfitting. In conclusion, finding the right balance between model complexity and performance through regularization is key to achieving optimal generalization and avoiding both underfitting and overfitting.

Page No.
Date

MDL Assignment

Name - Varun Gupta
Roll No - 202310108

Task 1 :→

(a) Gradient descent is an iterative optimization algorithm used to minimize a function by adjusting its parameters. In case of simple linear regression with one dep. & one indep. variable it minimize the cost function by finding m & b.

$$y = mx + b$$

y → (Dependendent) ~~Predicted Value~~
m → Slope
x → Independent Variable
b → Intercept

Let $\hat{y}$ be the predicted Value & y be the actual value s.t $\hat{y} = mx + b$ and $(\hat{y} - y)$ is a measure of the error

~~After this Gradie~~
Cost function

$$J(m, b) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \underbrace{(mx_i + b)}_{\hat{y}_i})^2$$

where n is the number of Data Points.

~~After th~~ Now Gradient of Cost is calculated with each coefficient.

Gradients are Partial derivatives of the cost function with respect to m and b.

$$\frac{\partial J}{\partial m} = -\frac{2}{n} \sum_{i=1}^{n} x_i (y_i - \hat{y}_i)$$

$$\frac{\partial J}{\partial b} = -\frac{2}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)$$

Now initially randomly assign values to m & b and using 'α' called as the learning rate, new values of m & b are calculated in following way →

$$m_{new} = m_{old} - \alpha \frac{\partial J}{\partial m}$$

$$b_{new} = b_{old} - \alpha \frac{\partial J}{\partial b}$$

Now the process is repeated until the change in m & b becomes negligible.
↳ Said to Converged

The final m & b are rep by m_{final} & b_{final}

(b) For Multivariate regression model the equation is:

$$y = b + \sum_{i=1}^{n} w_i x_i$$

Now the Error function becomes,

$$E = \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \left(b + \sum_{j=1}^{n} w_j x_{ij}\right)\right)^2$$

where m is the number of Data points
and x_{ij} is the i^{th} input for the j^{th} x variable

Now we find the partial derivatives of the error function wrt b & w_i $\forall i \in \{1, 2, \ldots n\}$:

$$D_b = \frac{-2}{m} \sum_{i=1}^{m} \left(y_i - \left(b + \sum_{j=1}^{n} w_j x_{ij}\right)\right)$$

$$D_{w_K} = \frac{-2}{m} \sum_{i=1}^{m} x_{iK} \left(y_i - \left(b + \sum_{j=1}^{n} w_j x_{ij}\right)\right)$$

Using these partial derivative equations we will find out the next values of our slope and bias as follows:

$$w_i = w_i - L \times D_{w_i}$$
$$b = b - L \times D_b$$

We repeat this iteratively until our error function is very small or ideally 0. The value of w_i & b that are left with now will be the optimum values.

Page No.
Date

Task 2 : →

Assumption : $\sigma^2 = 0$ (No Irreducible Error)

Given: →

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

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

True Values

Writing Everything in Vector format

Models Given: →

$\hat{f}_1(x) = x^2 + x + 1$

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

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

Now Calculating Predicted Values;

$\hat{f}_1(x) = [7, 1, 1, 7, 13, 21]$

$\cancel{\hat{f}_2(x)}$

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

$\hat{f}_3(x) = [5, 1, 2, 10, 17, 26]$

$E_i[\hat{f}_i(x)] = \frac{1}{3}[\hat{f}_1(x) + \hat{f}_2(x) + \hat{f}_3(x)]$

$= \frac{1}{3}[26, 4, 5, 31, 56, 89]$

$\text{Bias} = E_i[\hat{f}_i(x_j)] - f(x_j)$

→ Using Bias for the j^{th} datapoint

Ignoring Signs inside Square

$$\text{Bias} = \left[\frac{-4}{3}, \frac{-2}{3}, \frac{-4}{3}, \frac{7}{3}, \frac{2}{3}, \frac{-1}{3}\right]$$

$$\text{Bias}^2 = \left[\frac{16}{9}, \frac{4}{9}, \frac{16}{9}, \frac{49}{9}, \frac{4}{9}, \frac{1}{9}\right]$$

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

$$= \cancel{\left\{\frac{1}{3}\left[\left(\frac{-5}{3}\right)^2 + \left(\frac{-1}{3}\right)^2 +\right.\right.}$$

$$= \frac{1}{3}\cancel{\left[\left[\left(\frac{-5}{3}\right)^2 + \left(\frac{-1}{3}\right)^2 \left(\frac{16}{3}\right)^2 +\right.\right.}$$

For f_1

$$\rightarrow \left[\left(\frac{-5}{3}\right)^2, \left(\frac{-1}{3}\right)^2, \left(\frac{-2}{3}\right)^2, \left(\frac{-10}{3}\right)^2, \left(\frac{-17}{3}\right)^2, \left(\frac{26}{3}\right)^2\right]$$

$$= \left[\frac{25}{9}, \frac{1}{9}, \frac{4}{9}, \frac{100}{9}, \frac{169}{9}, \frac{676}{9}\right]$$

For f_2

$$\rightarrow \left[\left(\frac{16}{3}\right)^2, \left(\frac{2}{3}\right)^2, \left(\frac{1}{3}\right)^2, \left(\frac{11}{3}\right)^2, \left(\frac{22}{3}\right)^2, \left(\frac{37}{3}\right)^2\right]$$

for f_3

$$\rightarrow \left[\left(\frac{11}{3}\right)^2, \left(\frac{1}{3}\right)^2, \left(\frac{1}{3}\right)^2, \left(\frac{1}{3}\right)^2, \left(\frac{5}{3}\right)^2, \left(\frac{11}{3}\right)^2\right]$$

$$\text{Variance} = \frac{1}{3}\left[\text{Var}f_1 + \text{Var}f_2 + \text{Var}f_3\right]$$

$$\cancel{= [}$$

[14.89, 0.22, 0.22, 8.22, 29.56, 80.22]

$$MSE = \frac{1}{3}\sum_{m=1}^{3}\left(y_m - \hat{f}_m(x_j)\right)^2$$

$$= \frac{1}{3}\Big[[3^2, 1^2, 2^2, 1^2, 5^2, 9^2] + [4^2, 0^2, 1^2, 6^2, 8^2, 12^2]$$

$$+ [5^2, 1^2, 1^2, 2^2, 1^2, 4^2]\Big]$$

$$= \cancel{\frac{1}{3}[50, 2,}$$

$$= \frac{1}{3}[9+16+25, 1+0+1, 4+1+1, 1+36+4, 25+64+1, 81+144+16]$$

$$= \frac{1}{3}[50, 2, 6, 41, 90, 241]$$

$$= [16.6\bar{6}, 0.6\bar{6}, 2, 13.6\bar{6}, 30, 80.3\bar{3}]$$

Now, Var + Bias²

$$= [14.88 + 1.78, 0.22 + 0.44, 0.22 + 1.78, 8.22 + 5.44, 29.56 + 0.44, 80.22 + 0.11]$$

$$= [16.6\bar{6}, 0.6\bar{6}, 2, 13.6\bar{6}, 30, 80.\bar{3}]$$

$$= MSE$$

Hence Verified