# Machine, Data and Learning
## Assignment 1 Report

Keshav Bajaj - 2019115010
Aman Goyal - 2019101097

## Task1

**LinearRegression.fit()** is a function in the scikit-learn library of the LinearRegression model. It sets the coefficients of the linear regression model by minimizing the residual sum of squares between the dataset targets and the targets predicted by the linear regression model.

The method takes in three parameters:

1. **X(required)**:  An array of shape (n samples, n features) - This is the training data.

2. **Y(required)**: An array of shape (*n samples, n targets*) - These are the target values.

3. **sample_weight(optional)**: An array of shape (n samples) - These are the individual weights for each sample. (Default: none)

## Task2

| DEGREE | VARIANCE | BIAS |
|---|---|---|
| 1 | 25999.093010 | 1000.840897 |
| 2 | 39105.833813 | 976.645344 |
| 3 | 56095.893210 | 97.638845 |
| 4 | 114907.291530 | 102.899458 |
| 5 | 151434.027901 | 99.761942 |
| 6 | 174226.745003 | 99.995988 |
| 7 | 198849.502746 | 102.107374 |
| 8 | 221555.662196 | 104.870533 |
| 9 | 232275.805264 | 107.632098 |
| 10 | 232807.771037 | 119.631735 |
| 11 | 238575.678001 | 111.576237 |

| | | |
|---|---|---|
| 12 | 219780.328540 | 175.640419 |
| 13 | 236241.208340 | 126.392050 |
| 14 | 212545.262840 | 198.068473 |
| 15 | 221715.296909 | 250.855063 |
| 16 | 239357.883992 | 264.237671 |
| 17 | 242993.202304 | 339.469769 |
| 18 | 269052.279671 | 347.247096 |
| 19 | 270105.602474 | 438.155766 |
| 20 | 299003.459802 | 444.159465 |

**Observations for Bias:**

- The bias for degree 1 and 2 are large with comparison to the other bias as they are cases of huge underfitting of the dataset
- As we increase the complexity the bias decreases for degree 3 and after degree 11 the bias increases slowly. This result shows us that degree 3 is the best approximation of the data and then the approximation worsens over all the degrees after that and hence the bias increases.

**Observations for Variance:**

- The variance increases from degree 1 to degree 20
- The variance then continues to rapidly increase till the degree is 11.
- This happens as we overfit the data more and more with the increasing polynomial degree and overfitting the data results in high variance as our model adapts to our dataset too precisely and gives inaccurate results for a generalized dataset.

## Task 3

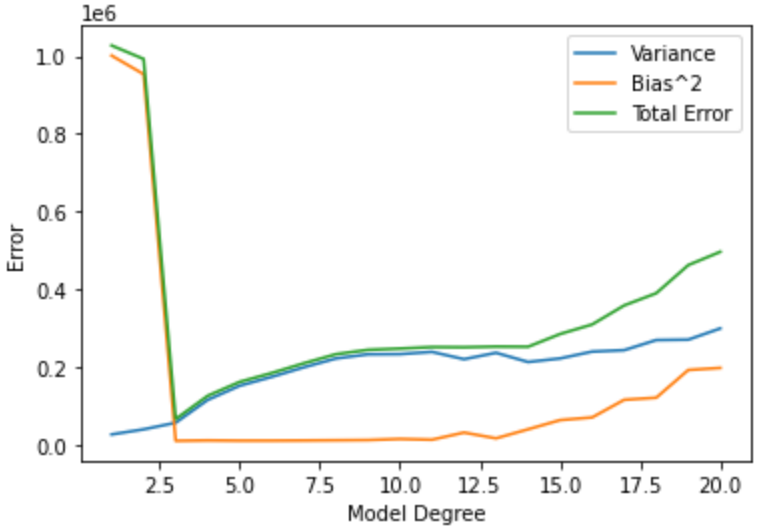
| DEGREE | IRREDUCIBLE ERROR |
|---|---|
| 1 | 1.091394e-10 |
| 2 | 4.365575e-11 |
| 3 | -7.275958e-12 |
| 4 | 0.000000e+00 |
| 5 | 2.910383e-11 |

| | |
|---|---|
| 6 | 0.000000e+00 |
| 7 | 2.910383e-11 |
| 8 | 0.000000e+00 |
| 9 | -2.910383e-11 |
| 10 | -5.820766e-11 |
| 11 | 0.000000e+00 |
| 12 | -5.820766e-11 |
| 13 | 2.910383e-11 |
| 14 | -2.910383e-11 |
| 15 | -2.910383e-11 |
| 16 | -2.910383e-11 |
| 17 | 5.820766e-11 |
| 18 | -5.820766e-11 |
| 19 | 0.000000e+00 |
| 20 | 5.820766e-11 |

**Observations:**

- Irreducible Error does not change much with the changing degree as we see in the listed values above.
- We see such observations because as the irreducible error accounts for the noise in data and without cleaning the noise from our data we cannot remove the error even if we make more complex models.

## Task 4




**Observations:**

- As we see in the model that variance increases with an increasing model degree which means that more and more overfitting to the dataset happens with an increasing degree.
- Also, the bias is very high for degree 1 and 2 as they are the cases of large underfitting of the model to our dataset and higher degree fit the dataset better so they have a low bias as compared to degree 1 and 2.
- The bias is similar from degree 3 to degree 11 and then the bias starts to increase until degree 20
- Hence by looking at the graph, degree 3 will be the optimum degree to generalize the dataset and the noise associated with all of the data is small as an irreducible error if small for almost all of the degrees