Report

Artificially Dumb : Team Number: 13

> Khushi Agarwal - 2020101092

> Naimeesh Narayan Tiwari - 2020101074

Task 1: LinearRegression().fit()

It is a function from sklearn.linear_model and it is used to make a predictive ML model on a linear scale and perform linear and polynomial regressions.

When done for a linear classifier, we get a predicted value y, on the input of the training data x and this function predicts a model by giving values a and b(weights) which tries to give the best fit line,

y = ax+b such that we get the lowest MSE [the sum of squares of the difference between predicted value (y) and real value is minimum with the input value (x)]

Task 2: Calculating Bias and Variance

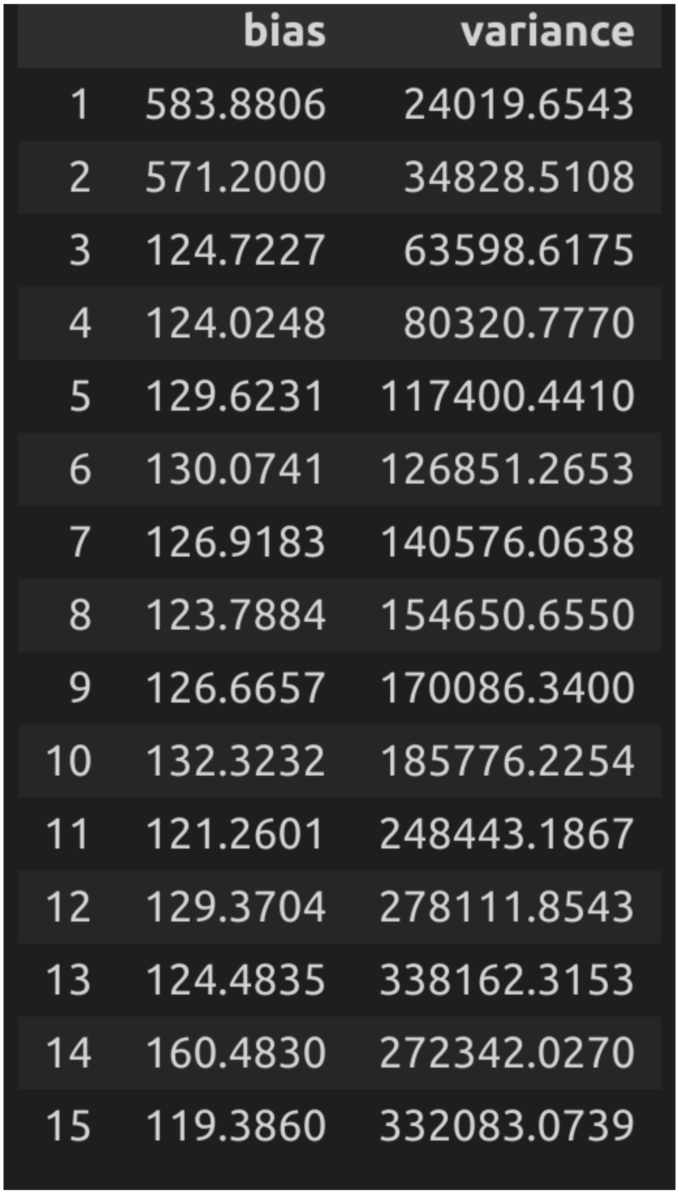
> Write a detailed report explaining how bias and variance change as you vary your function classes.

Bias:- Bias is the difference between the average prediction of our model, that is the predicted y, and the correct value that is the testing data set, which we are trying to predict. A model with high bias pays very little attention to the training data and oversimplifies the model.

Variance:- Variance is the variability of model prediction for a given data point or a value that tells us the spread of our data. A model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before.

Overfitting occurs when the model or the algorithm fits the data too well. Specifically, Overfitting occurs if the model or algorithm shows **low bias** but **high variance**.

Underfitting occurs when the model or the algorithm does not fit the data well enough. Specifically, underfitting occurs if the model or algorithm shows **low variance** but **high bias**



	bias	variance
1	583.8806	24019.6543
2	571.2000	34828.5108
3	124.7227	63598.6175
4	124.0248	80320.7770
5	129.6231	117400.4410
6	130.0741	126851.2653
7	126.9183	140576.0638
8	123.7884	154650.6550
9	126.6657	170086.3400
10	132.3232	185776.2254
11	121.2601	248443.1867
12	129.3704	278111.8543
13	124.4835	338162.3153
14	160.4830	272342.0270
15	119.3860	332083.0739

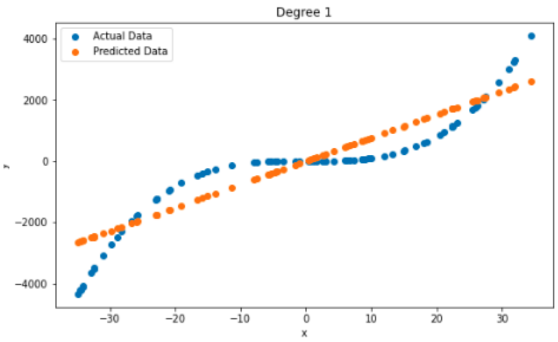
The above table shows the varying bias and variance for different degree polynomials for the data set provided to us.

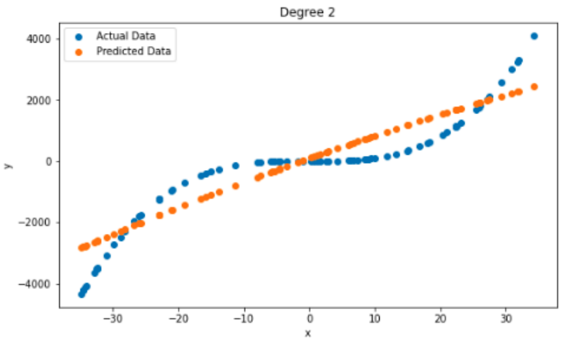
The above table shows the varying bias and variance for the data set provided to us.

Observations

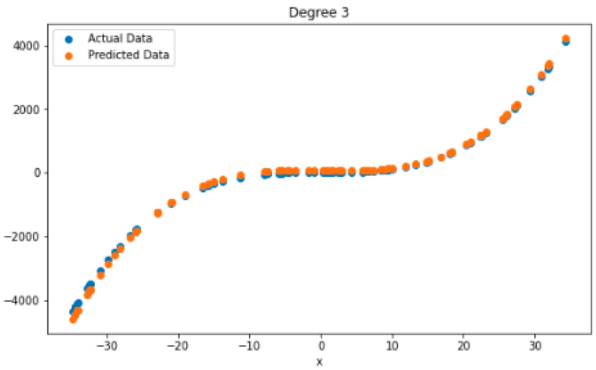
- We observe that there is a very high bias for degrees 1 and 2 and as we move to degree 3, the bias decreases dramatically, and then for further higher powers, it remains almost constant.

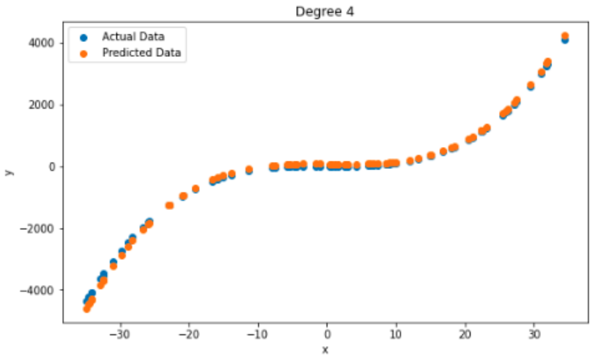
- In the case of variance, we observe that it increases gradually from degree 1 to 10 with a very low value and increases suddenly and from degree 10 onwards with comparatively large values. But overall the variance increases as we increase the degree.
- Based on the observations, we see that degrees 1-2 have high bias and low variance. And hence we can conclude that the models in these cases produce **underfitting models**, which is clearly visible from the graphs.



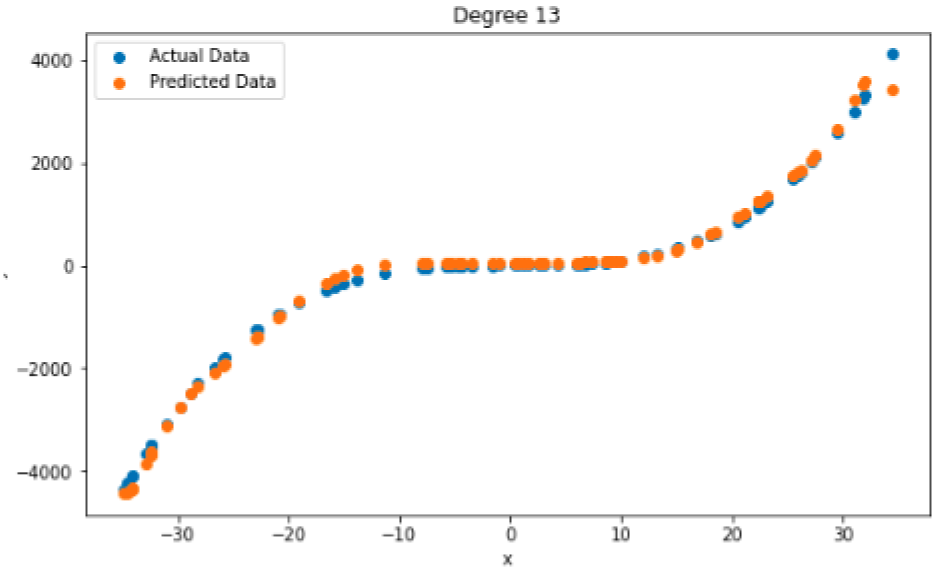


- We see that for degrees 3 and 4, the bias, as well as variance, is low and so these are the **best fit models**. Also, degrees 5-10 somewhat produce best-fit models though there is a large difference b/w the variance of degrees 3 and 10 that difference is attained gradually, whereas, in the case of degrees 10 and 11, there is a drastic change observed.





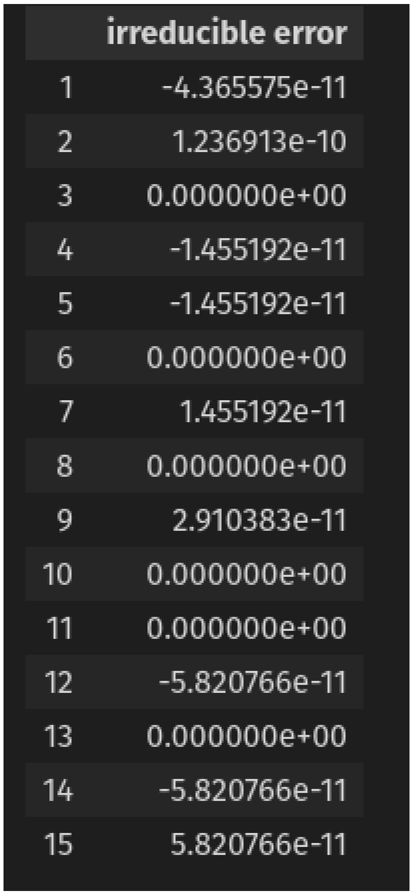
- And for all other models, the bias is low but the variance is very high which is clearly visible from the table pasted above and also the graph, so we can conclude that this results in **overfitting**.



Task 3: Calculating Irreducible Error

> Write a detailed report explaining why or why not the value of irreducible error changes
> as you vary your class function.

Irreducible error: The irreducible error is the error that we can not remove from any model. It is present there due to noise.



	irreducible error
1	-4.365575e-11
2	1.236913e-10
3	0.000000e+00
4	-1.455192e-11
5	-1.455192e-11
6	0.000000e+00
7	1.455192e-11
8	0.000000e+00
9	2.910383e-11
10	0.000000e+00
11	0.000000e+00
12	-5.820766e-11
13	0.000000e+00
14	-5.820766e-11
15	5.820766e-11

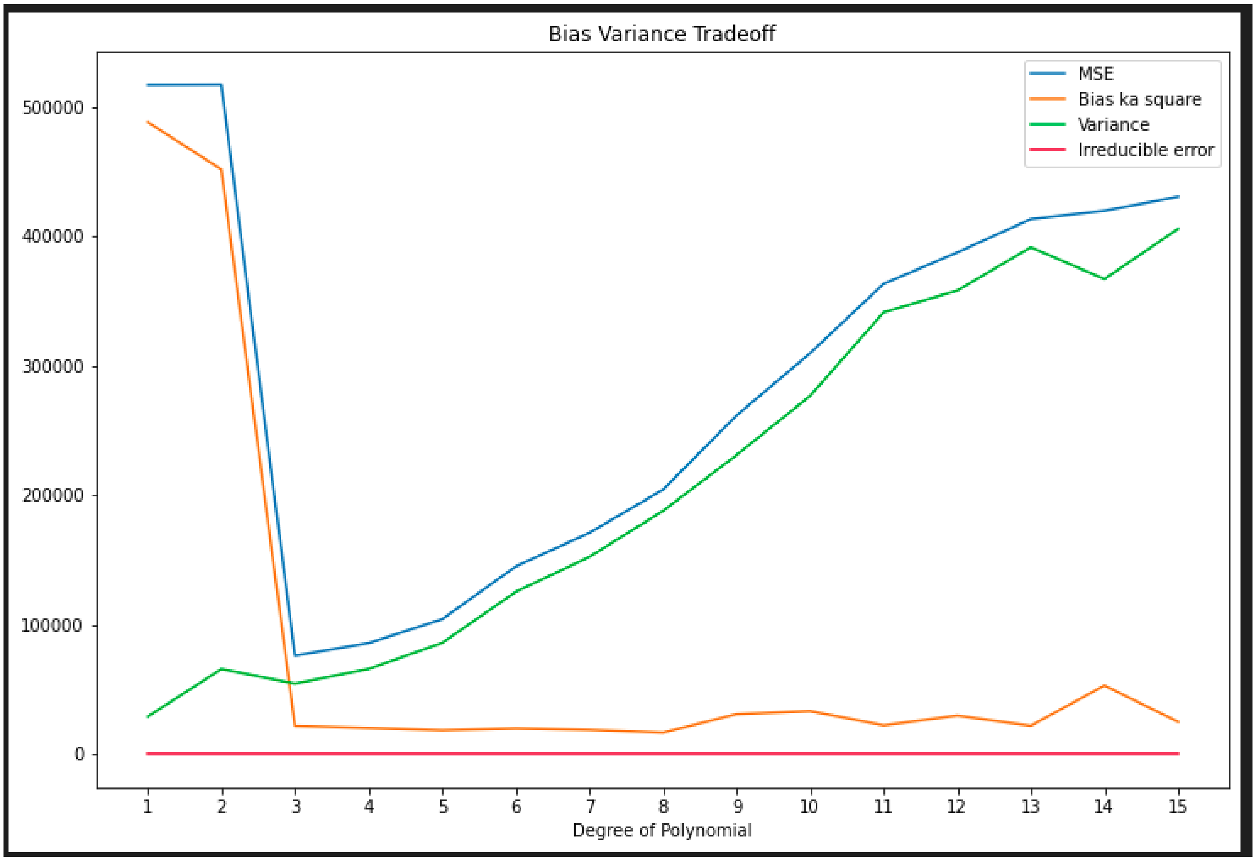
The irreducible error in the provided data set was very low and sometimes even so low that it got reported as 0.

Using the plot from the task 4 section below, or using the table above we can easily say that the irreducible error remained almost constant with an increase in degrees.

The reason for that is that the error is present in the test set provided to us and can not be resolved by using any model.

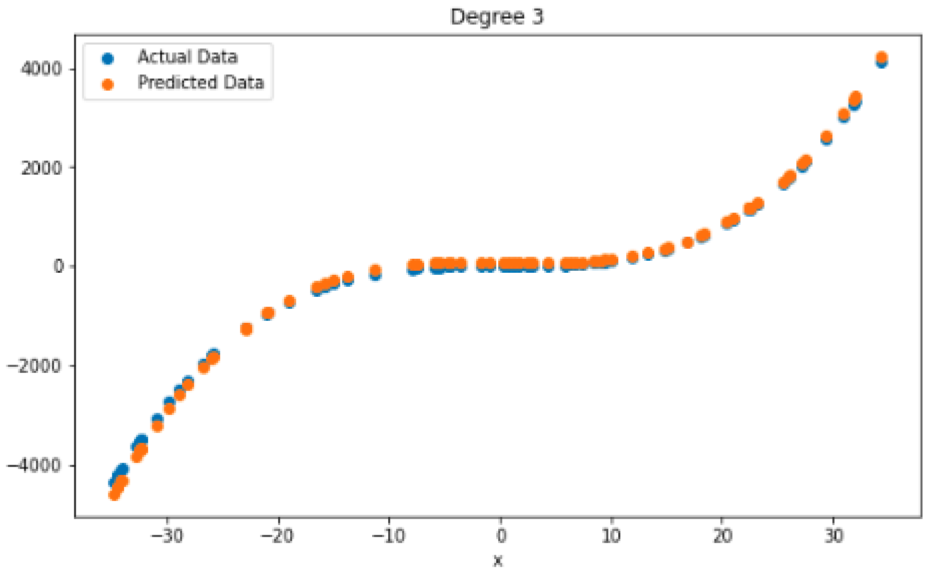
Task 4: Plotting Bias^2 − Variance graph

> Write your observations in the report with respect to underfitting, overfitting, and also comment on the type of data just by analyzing the Bias^2 − Variance plot.



the MSE comes out to be the minimum at x=3, indicating that the function is either a cubic or close to the highest power 3 if not polynomial.

We can see that in the cubic model we get a very good fit for the test data.



- We observe that there is a very high bias for degrees 1 and 2 and as we move to degree 3, the bias decreases dramatically, and then for further higher powers, it remains almost constant. Moreover, the bias is initially high because of underfitting, drops to the best value, and increases again as the model conforms too closely with the test data and loses its generality, causing it to perform poorly.
- In the case of variance, we observe that it increases gradually from degree 1 to 10 with a very low value and increases suddenly and from degree 10 onwards with comparatively large values. But overall the variance increases as we increase the degree. The variance is continuously increasing (after degree 3) because the curve of best fit overfits the training data, resulting in an inaccurate representation of the test data while also decreasing the precision of the model.
- Based on the observations, we see that degrees 1-2 have high bias and low variance. And hence we can conclude that the models in these cases produce underfitting models, which is clearly visible from the graphs.
- A good model is one with a good balance between bias and variance or with a good tradeoff between bias and variance that it minimizes the total error. We see that for degrees 3 and 4, the bias, as well as variance, is low and so these are the best fit models. Also, degrees 5-10 somewhat produce best-fit models though there is a large difference b/w the variance of degrees 3 and 10 that difference is attained

gradually, whereas in the case of degrees 10 and 11, there is a drastic change observed.

- And for all other models, the bias is low but the variance is very high which is clearly visible from the table pasted above and also the graph, so we can conclude that this results in **overfitting**.