Machine, Data Learning Assignment 1
Report

Task 1:
LinearRegression().fit() performs :-

With LinearRegression().fit() you can calculate the optimal values of the weights using the existing input and output i.e. x and y as the arguments.
So we can say .fit() fits the model.

Task 2:
Calculate Bias and Variance :

DEG	BIAS	VARIANCE
1	819.509782	47638.313857
2	809.558968	50092.968913
3	71.315801	74216.059047
4	78.556058	92945.593747
5	79.361688	119539.331613
6	79.318834	142851.889557
7	85.593684	150580.311256
8	91.682391	173075.483881
9	94.168258	212110.912889
10	98.086236	212951.663479
11	93.463245	229661.672184
12	120.322314	250609.447300
13	92.396681	225776.915726
14	125.620583	243409.808228
15	160.157920	238748.047767
16	164.562847	267067.583345
17	231.755494	263276.981361
18	233.519490	291344.679912
19	306.629949	298071.684335

20	304.771690	320826.373903

How bias changes

As we vary the function classes we observe that the bias starts to decrease then after a certain degree the bias again increases.The reason behind this behaviour is that at lower degrees i.e. at degree 1 and 2 we can see that bias was very high (under fitted) but as we increase the the degree we saw that the test data started to show high resemblance with the graph this is because as the degree will increase the number of local maximas and minimas will also increase (as the slope will get 0 at multiple points) therefore with the increased number of curves our graph will try to adapt to the given test values.But after a certain degree the bias again increases as now the complexity is more so to maintain data set compatibility and complexity our model deviates from the values and as a result the bias increases.

How variance changes

From the data attached above we can see as the degree of the function is increasing the variance is also increasing. So the higher the complexity the higher the variance this is because with high complexity the model gives high accuracy to the given data but as the data will vary (even with a slight variation) the accuracy will drop considerably (overfitting).

Conclusion

So we can say that the best possible model is when both bias and variance are low.But to lower the bias we need to increase the complexity and to lower the variance we need to decrease the complexity .From the data attached above we can see at degree 3 both bias and variance are at a reasonable value.So at degree 3 we get the possible model.

Task 3:

Calculate Irreducible Error :-

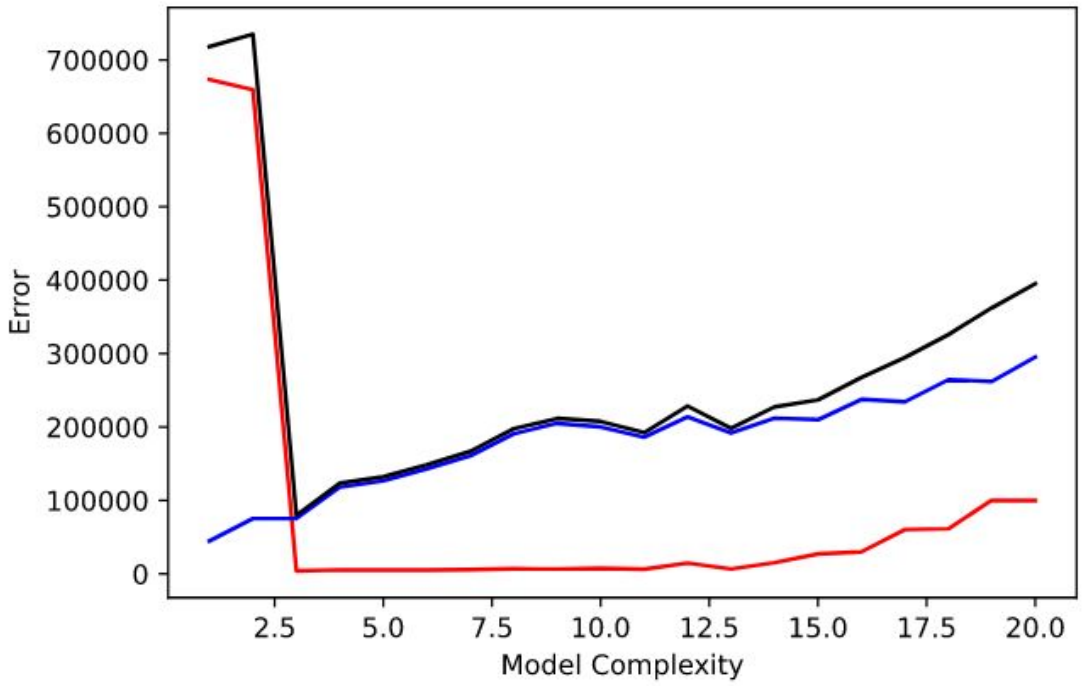
DEG	IRREDUCIBLE ERROR
1	6.91215973e-12
2	1.29602995e-10
3	-3.45607987e-12
4	-8.41282599e-12
5	1.31876732e-12
6	-1.04591891e-11
7	-1.68256520e-11
8	-3.63797881e-13
9	-7.82165444e-12

10	2.72848411e-12
11	-5.27506927e-12
12	-6.82121026e-12
13	1.15505827e-11
14	6.03904482e-11
15	2.27373675e-12
16	1.50976120e-11
17	-3.68345354e-11
18	-5.45696821e-12
19	9.46783985e-11
20	-2.79214873e-11

We know that irreducible error can’t be reduced doesn't matter how good our model is because it’s an error caused by noise in the data,random variations that don’t represent a real pattern in the data, or the influence of variables that are not yet captured as features. From the data attached above we can see unlike bias and variance the irreducible error doesn’t follow any particular trend and is independent of bias and variance.We are even getting negative value of irreducible error because there are factors with are not measurable that’s why we are getting these preposterous values.
Also we can see the values of irreducible error which we are getting are very small and are almost constant because the noise which is contributing in the error and on data shuffling and random sampling the average noise will remain constant.

Task 4:

Bias - Variance TradeOff Graph



Red - Bias^2
Blue - Variance
Black - Total Error

As we have seen above that for degree lesser than three the bias is very high due to less complexity and inability of our model to resemble with the given training data this is the case of underfitting but for these degrees the variance is low as the output values won't deviate much but as the model complexity increases the variance value also increases so that's why we said that the model will be perfectly balanced at degree 3.
But just after degree 3 we can see now both bias and variance are increasing because now with high complexity the model is now more adaptive to the given data but when it's compared to the test data the value vary a lot this is the case of overfitting as there is good performance on the training data but poor generalization to other data.
But also the point to note now is that with increasing complexity the bias is also increasing as now the model is too complicated to fit in the data so to adapt with the given train set our model finds a median of both which results in high bias.
From the graph we can see that the type of data is consistent but with high model complexity due to high variance and bias the total error is more.