Assignment 1

Machine, Data and Learning

Team: Poorva Pisal (2020113001), Sneha Raghava Raju (2020101125)

Task 1: Linear Regression

LinearRegression() instantiates an object of the linear regression class and .fit() performs the linear regression on the given data points.

It finds the best value for the intercept and slope, which results in a line that best fits the given data. The equation of a straight line is y = mx + b. Where b is the y-intercept and m is the slope. The linear regression algorithm finds the most optimal value for the intercept and the slope.

There can be multiple straight lines depending upon the values of intercept and slope. The linear regression algorithm fits multiple lines on the data points and returns the line that results in the least error.

Task 2.2.2: Write a detailed report explaining how bias and variance changes as you vary your function classes.

We varied the function classes from low degree (low complexity) to high degree polynomials (high complexity), and the bias and variance changed along with it. For the low degree functions (degree 1 and 2) the bias was very high, the predicted values did not fit the test values well. The model did not model the training data well and is not able to generalize well, so the model was underfit and bias was high. For the higher degree models the model fit the data well, and the bias reduced. However for very large degrees (degree 14 and 15) the bias increased again because now the model is overfit. This means the model is picking up on noise/random fluctuations, so it is not able to generalize well, hence the bias increases.

The variance increased as the degree of the model increased. This is because as the complexity of the model increased it became more sensitive to noise and random fluctuations in the data, i.e overfitting. So the predicted values don’t fit the test values well and variance is high.

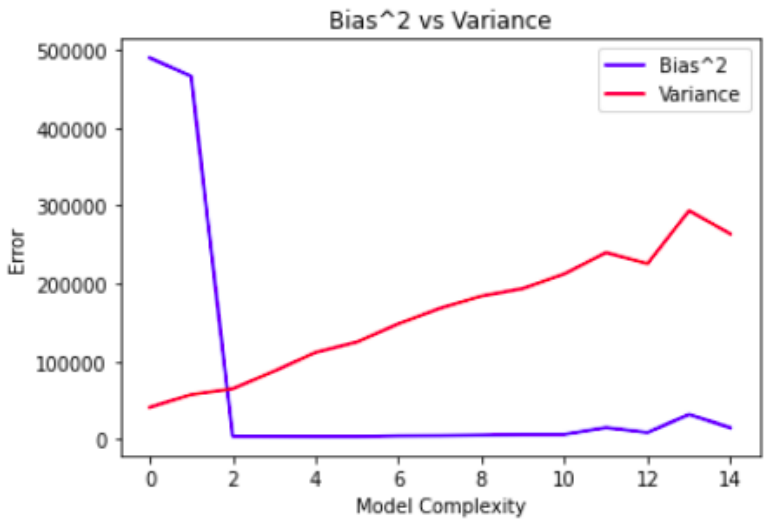
Degree	Bias^2	Variance	Irreducible error
1	489774.535994	41322.989284	1.455192e-11
2	466254.602804	57563.993945	-4.365575e-11
3	4323.197476	65071.932606	-1.455192e-11
4	4176.542958	87636.882093	0.000000e+00
5	3876.961758	111779.926206	0.000000e+00
6	3902.102341	125192.175118	-1.455192e-11
7	4759.094862	148574.152969	0.000000e+00
8	4994.542734	168398.428231	2.910383e-11
9	5604.340519	184115.659393	-2.910383e-11
10	6465.292113	193766.059409	2.910383e-11
11	6595.834313	212608.595803	-2.910383e-11
12	15176.517614	239801.602743	-2.910383e-11
13	8889.286026	225586.400783	0.000000e+00
14	32239.906311	293647.379756	0.000000e+00
15	14965.456466	263676.524918	5.820766e-11

Task 3: Write a detailed report explaining why or why not the value of irreducible error changes as you vary your class function.

As seen in the table above the irreducible error is extremely small across all the models. The irreducible error is an error that we cannot remove with any model. This error is caused by things out of our control like noise. This noise can be due to errors in the data capturing process.

This noise is inherent to the data so cannot be reduced/removed. Therefore changing the model function has no effect on the irreducible error.

Task 4: Plotting $Bias^2$ − Variance graph
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 $Bias^2$ − Variance plot.



As the model complexity increases, bias^2 first decreases, is low for the middle complexities and then increases again for the higher complexities. Variance increases as complexity of the model increases.

For low complexities, the model did not capture all the features of the training data and did not generalize well so it was underfit. As we can see from the graph there is high bias since the model is not able to accurately capture the training data. The variance is low since even though the model is not correct it is still performing consistently.

For high complexities, the model captured the training data too well so it is not able to generalize for new data (testing data), i.e the model is overfit. Because of this variance is high, as is seen in the graph.

From the graph we can see that the error is minimum at model complexity 2, i.e polynomial of degree 2. So from this we can conclude that a degree 2 polynomial best fits the data set.