Machine, Data and Learning Assignment - 1 report

Team: toolazyforaname, team-100
Members:
- Yash Agrawal, 2020114005
- Vanshpreet S Kohli, 2020114014

Task 1: LinearRegression().fit()

`LinearRegression().fit()` from `sklearn.linear_model` creates a predictive ML model using linear regression, by trying to minimize the sum of squares of distances between the predicted value of y and the real value of y. Essentially, it tries to fit

$$y = ax + b$$

into the dataset by finding a and b such that the sum of squares of error is minimum. If used with `PolynomialFeatures().fit_transform()` from `sklearn.preprocessing`, it can be used to fit data not just to a line but an n-degree polynomial in x.

Task 2: Variance and Bias

Here are a table and a graph with the recorded values:

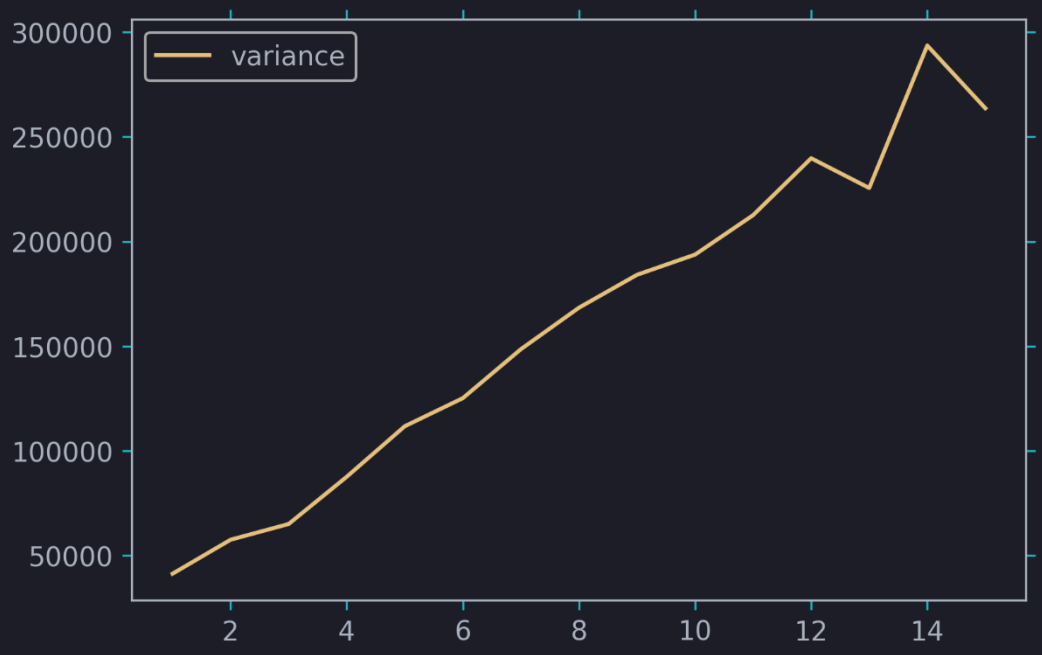
	degree	bias_2	variance	total_error	bias	irreducible
0	1	489774.535994	41322.989284	531097.525278	699.838936	1.455192e-11
1	2	466254.602804	57563.993945	523818.596749	682.828385	-5.820766e-11
2	3	4323.197476	65071.932606	69395.130082	65.751026	-7.275958e-12
3	4	4176.542958	87636.882093	91813.425052	64.626179	0.000000e+00
4	5	3876.961758	111779.926206	115656.887963	62.265253	2.910383e-11
5	6	3902.102341	125192.175118	129094.277459	62.466810	0.000000e+00
6	7	4759.094862	148574.152969	153333.247831	68.986193	-2.910383e-11
7	8	4994.542734	168398.428231	173392.970965	70.672079	0.000000e+00
8	9	5604.340519	184115.659393	189719.999912	74.862143	-2.910383e-11
9	10	6465.292113	193766.059409	200231.351522	80.407040	0.000000e+00
10	11	6595.834313	212608.595803	219204.430116	81.214742	2.910383e-11
11	12	15176.517614	239801.602743	254978.120357	123.193010	0.000000e+00
12	13	8889.286026	225586.400783	234475.686809	94.283010	5.820766e-11
13	14	32239.906311	293647.379756	325887.286067	179.554745	-1.164153e-10
14	15	14965.456466	263676.524918	278641.981383	122.333382	0.000000e+00

Change in bias with degree:

- Bias represents the accuracy of the model's predictions. The lower the bias, the more accurate the model is.
- When the model goes from n=2 to n=3, the bias decreases significantly, indicating an improvement in the model. This is likely because at n=1 and n=2, the current model was underfitting and thus getting a large bias value.
- For degrees 3-6, the bias remains relatively unchanged and lower than that of the models below or above, indicating that a decent fit has been achieved.
- For models with degree >= 7, the bias is higher and tends to increase with the degree. This is likely because the model is now being overfitted, extracting more patterns than necessary from the training dataset.

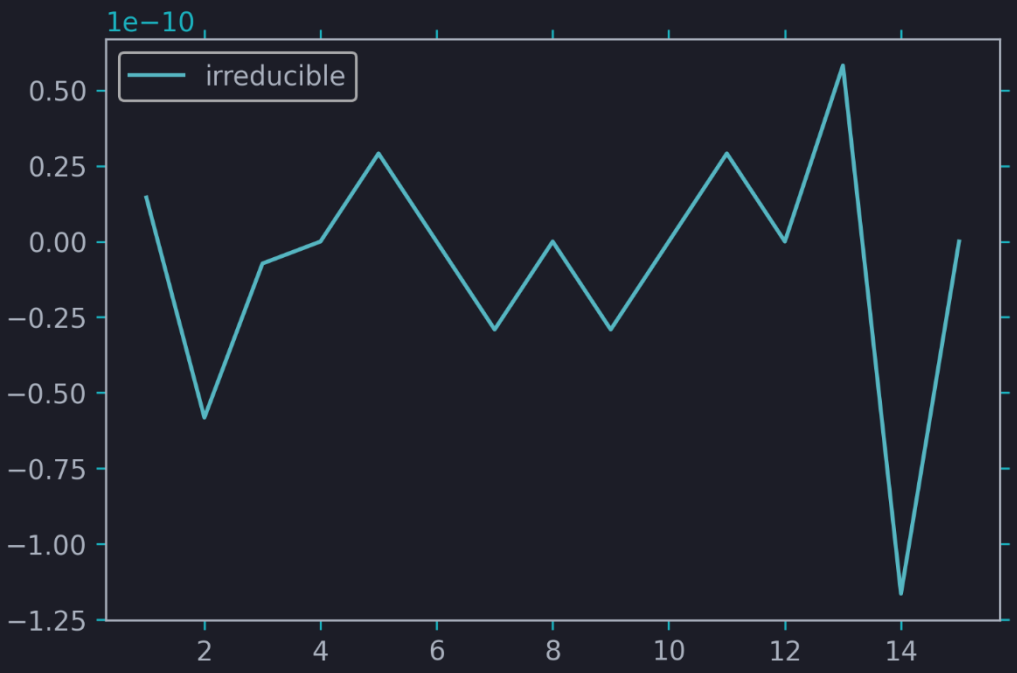
Change in variance with degree:

- Variance represents the spread of the predicted values. It is not a representation of the accuracy, but the precision of the model. If the model captures too much noise from the training data, the variance increases.
- The variance keeps increasing with n, as the model keeps trying to fit the noise that occurs with the the data. As more features are added, more noise is read which scatters the predictions.



Task 3: Irreducible error

Irreducible error does not depend on the degree of the polynomial being fitted, as it is caused by random noise and does not change with models or regression techniques. The data seems to have very little noise, due to which the irreducible error is pretty low.



Task 4: Variance vs Bias^2

As mentioned above, the bias is high at degrees 1 and 2 due to underfitting and slowly starts to increase at higher values of degree due to overfitting. As shown in the graph, the MSE is lowest at degree 3 - i.e. a cubic polynomial best fits the curve of the data. This indicates that the function is cubic.