Machine, Data and Learning

Assignment -2 , Report

Vyom Goyal* - *2021101099

Task 2.1

Q. Write a brief about what function the method LinearRegression().fit() performs.

Ans) The LinearRegression.fit() is a function present in scikit python library. It Takes as input two arguments , one supplying the value of x which represents the independent variables in dataset and y which represents dependant variables . The function trains the model by finding the best fit line which minimises the least square errors. It returns the intercepts and coefficients which can then be used to predict values. It can also be used to compare accuracy different models and help in analysing the impact of different parameters taken for the model.

Task 2.2

Q. Explain how gradient descent works to find the coefficients. For simplicity, take the case
where there is one independent variable and one dependent variable.

Ans) Finding the coefficients of a Linear Regression model involves an iterative optimisation method called Gradient descent. The main objective is to reduce the difference between the values of dependant variables predicted using our model and their actual values. The algorithm used basically involves first estimating the values of the coefficients and then continually change the values of these coefficients in order to reduce the error . If a stopping requirement is not given it keeps running.

Taking the case given in question (i.e one independant and one dependant variable) the objective is to find slope (M) and intercept (C) which minimises the SSE(Sum of Squared Errors).

SSE is defined as = Σ (y - (Mx + C)) ^2

where y is actual value of dependant variable for independant variable x

In order to minimise the SSE the gradient descent method iteratively modifies M and C. In each iteration the algorithm calculates the value of partial derivatives of SSE by M and C respectively to get descent direction and magnitude which is then used to update the values of M and C. The values are incremented in direction opposite to the gradient . The slope magnitude is also multiplied by a parameter Learning_Rate.

i.e. M = M - (Learning_Rate * d(SSE) / d(M))

and C = C - (Learning_Rate * d(SSE) / d(C))

If the set parameter is reached the algorithm stops running.

Task 2.3

DEGREE	BIAS	BIAS^2	VARIANCE	MSE	IRREDUCIBLE ERROR
1	0.269398	0.114392	0.00868095	0.123073	-1.38778e-17
2	0.0862565	0.0121412	0.00122436	0.0133656	1.73472e-18
3	0.0332718	0.00470811	0.000337339	0.00504545	1.04083e-17
4	0.0242826	0.0042391	0.000366999	0.0046061	2.60209e-18
5	0.0238793	0.00419759	0.000461938	0.00465952	-8.67362e-19
6	0.0239554	0.00419841	0.00058152	0.00477993	-2.60209e-18
7	0.02483	0.00418638	0.000916796	0.00510317	3.46945e-18
8	0.0248874	0.00426172	0.00176092	0.00602264	8.67362e-19
9	0.0304184	0.00485845	0.00827683	0.0131353	1.73472e-18
10	0.0286639	0.00441441	0.00650085	0.0109153	-5.20417e-18
11	0.0365903	0.00639682	0.0326928	0.0390896	-2.34188e-17
12	0.0709172	0.0569339	0.884494	0.941428	4.16334e-17
13	0.0422491	0.00757067	0.054317	0.0618876	2.60209e-18
14	0.156428	0.445784	8.04335	8.48913	2.77556e-15
15	0.0891287	0.0799882	2.66953	2.74952	-6.52256e-16

With Increasing Degree:

Bias

- For degree 1 the value of bias is very high
- This then decreases till degree around degree 7-8. Then it remains almost constant (slight increase) till degree 10. After this the value of of bias Increases steeply as degree increases further.

- This happens because for degree 1 model is underfitting, i.e. it is not able to fit all the trends of the data in the model.
- For the higher degree values the model is overfitting and then catches noise due to which the value bias is high

Variance

- The value is small for smaller degrees and then steeply increases for higher ones.
- This happens because the model is underfitting for small degrees.

Task 2.4

IRREDUCIBLE ERROR
-1.38778e-17
1.73472e-18
1.04083e-17
2.60209e-18
-8.67362e-19
-2.60209e-18
3.46945e-18
8.67362e-19
1.73472e-18
-5.20417e-18
-2.34188e-17
4.16334e-17
2.60209e-18
2.77556e-15
-6.52256e-16

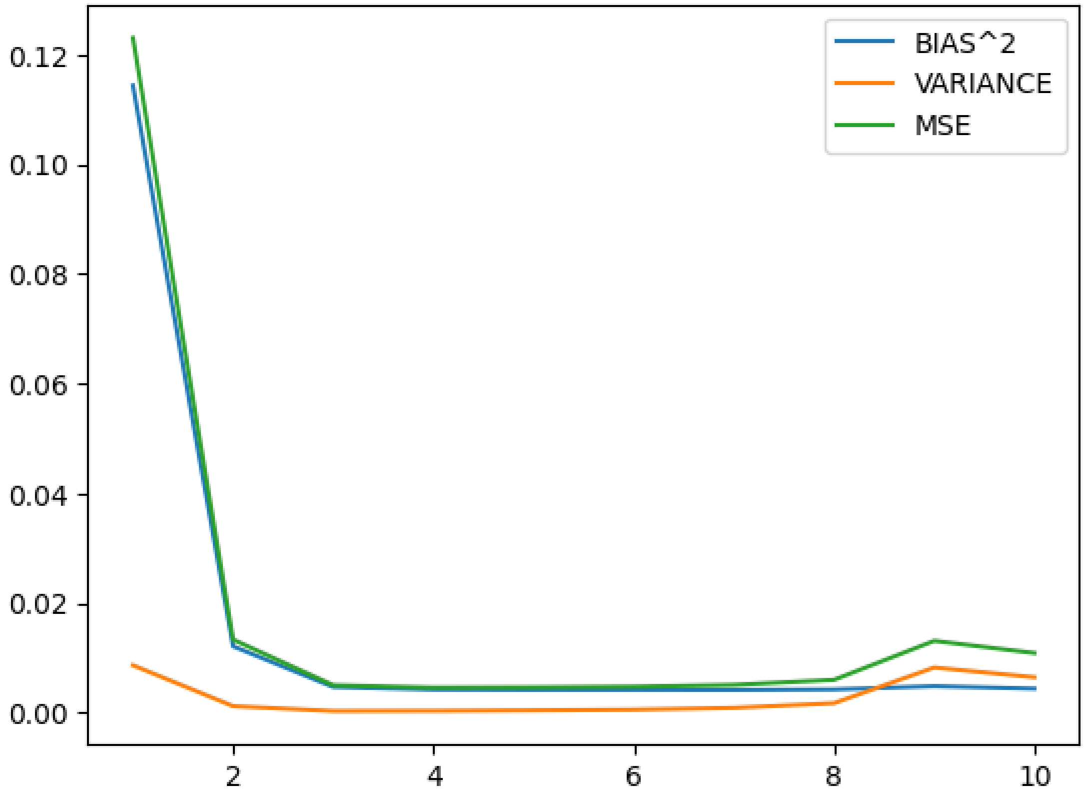
Observation

The values of irreducible errors are almost same for all the values of degrees and it follows no general trens in their values. Also they are very small in magnitude (of the order 1e-15 to 1e-19)

Reasons

This error appears due to several unknown variables involved in calculation of the values. These values are neither dependant variables nor independant variables used to design the model (i.e. noise in data). These errors can occur due to factors like rounding errors. Thus these values are independant of the model.

Task 2.5



As indicated by the high bias and low variance of values, a model such as Linear model is not able to fit the data properly. This is called underfitting and causes high error in prediction thus we must prevent underfitting and thus must not oversimplify the model.

Similar to the above, we must also ensure that the model's variance is minimal as we need the predictions to be similar for a particular value across different datasets. We must ensure that a model does not rely on an incorrect characteristic. Overfitting is the term used to describe a model that performs well with the train dataset but generates a lot of mistakes when applied to real-world data i.e. low bias and high variance So, it is crucial to keep the variance low in order to keep the model efficient in the longer real world while predicting values.

Inferences

We can see that the MSE is least around degree 4-5 and therefore the optimal model is a ploynomial of degree close to 4-5 .Lower degrees are underfitting and higher ones are overfitting.

Thus this data approximately forms a curve of 4th degree on 2-D plane.

Also the irreducible error is very low thus data is well structured.

Bonus

- Capacitance= 4.999999999999485e-05
- Resistance= 100000.00000001038