Machine Data and Learning

Assignment 1

Task 1

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

LinearRegression.fit(X, Y) takes in some X values and its target(Y) values.

Its basic aim is to find a line $y = mx + c$ which will 'fit' the target values for the given X. To accomplish this it need to find the optimal values for the parameters m and c. Fitting a line to target values means minimising the mean squared errors (MSE) that it will produce. MSE is calculated as,

$$MSE = \frac{1}{n} \sum (y(x_i) - y_i)^2$$

After calculating the MSE, it employs gradient descent to identify the weights to minimise MSE. Gradient descent is calculated as,

$$\Delta w = 1/n \sum_{i=1}^{n} (y(x_i) - y_i) * x_i$$

After it gets the gradient descent, it updates the weight every time until it gets the best-fitted value as,

$$W_{new} = W_{old} + \alpha * \Delta w$$

where alpha is some value between 0-1

LinnearRegression().fit() finally returns the best-fitted m and c.

When X has multiple variable as $x_1, x_2, x_3, ...x_n$ the function return $m_1, m_2, m_3, ...m_n$ where the fitted line has the equation,

$$y = m_1x_1 + m_2x_2 + ... + m_nx_n + c$$

Task 2

> 💡 Tabulate the values of bias and variance and also write a detailed report explaining how bias and variance change as you vary your function classes.

Given below is the average variance and bias observed for a random splitting of the given training data set.

Average Bias and Variance

Degree	Average Variance	Average Bias	Average (Bias^2)
1	24449.70837	162.738875	491926.959819
2	44251.482995	160.151502	469163.170835
3	54686.896256	-8.900734	4503.557292
4	64840.262732	-4.465885	4424.074545
5	77650.267035	-3.865008	3883.802336
6	89973.401716	-2.673281	3432.946083
7	131420.325507	-1.979042	3988.11029
8	149993.769428	-5.620938	4141.091897
9	171771.152976	-3.292118	4241.430289
10	207066.11131	-3.054047	5068.024329
11	207995.25187	-2.740466	4780.876735
12	216892.077774	0.820292	10518.559715
13	184684.645924	-8.458282	6724.874227
14	231803.350645	-5.92918	19171.198204
15	206880.496669	-6.738444	11158.198072

In general, as we increase the degree of the polynomial in our function class, the value of bias and the value of average bias^2 reduces. The values of average bias and bias^2 are fairly high for polynomials for degree less than 3 but these values rapidly fall when the degree of polynomial becomes 3. When the degree of the polynomial is greater than 3, bias and bias^2 have and overall gradual decreasing trend, however, values of bias and bias^2 may increase for some degrees.

In general, as we increase the degree of the polynomial in our function classes, the value of average variance increases. The increase is an overall trend, however, average variance for some degree may be less than the average variance for the previous degree.

Task 3

Tabulate the values of irreducible error for the models in Task 2 and also write a detailed report explaining why or why not the value of irreducible error changes as you vary your class function

Given below is the average irreducible error observed for a random splitting of the given training data set.

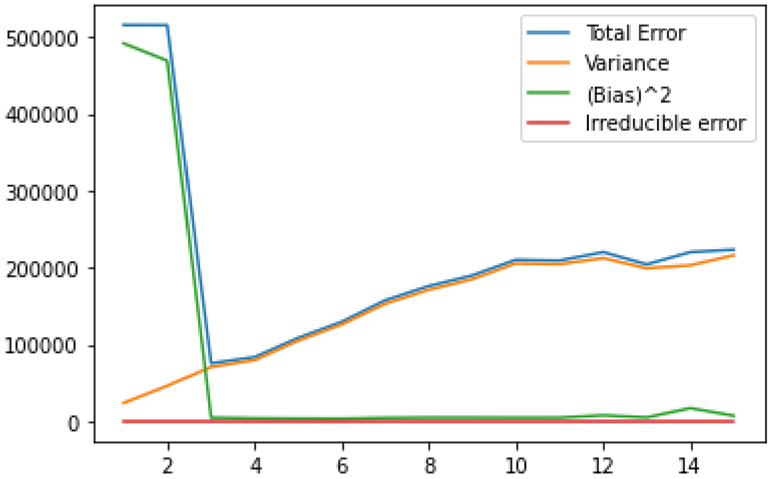
Degree	Irreducible error
1	-5.820766091346741e-11
2	5.820766091346741e-11
3	-5.4569682106375694e-12
4	-4.547473508864641e-12
5	1.4551915228366852e-11
6	-1.000444171950221e-11
7	-1.9099388737231493e-11
8	-8.185452315956354e-12
9	1.8189894035458565e-12
10	4.638422979041934e-11
11	-3.456079866737127e-11
12	2.000888343900442e-11
13	0.0
14	3.637978807091713e-11
15	-3.456079866737127e-11

The calculated values for irreducible errors are extremely close to zero. Our irreducible error values aren't perfect, we encounter floating point errors when computing them. These errors cause a small variance between the irregular error values as our class function is varied. Hence we may infer that the irreducible error is negligible and does not vary outside of acceptable limits of error as the class function is varied.

The irreducible error does not change as it is only dependent on the test set data and the trained models do not influence it in any way. This error is the one that creeps in due to inaccurate reporting. Hence, there is no reason for it to change for different types of models.

Task 4

> 💡 Write your observations in the report with respect to underfitting, overfitting and also comment on the type of data just by analyzing the Bias2 − Variance plot.



For models with the degree of polynomial as 1 and 2 have extremely high bias and low variance resulting in an extremely high MSE. This indicates underfitting as those models are not able to completely capture the data's features.

For models with degree of polynomial greater than and equal to 3 the square of bias is very low as compared to the variance. But at the same time the variance increases steadily as the degree of polynomial increases. For higher degrees we have low bias and high variance indicating overfitting. This indicates the model reads too much into the data and loses its flexibility to vary according to the test data.

We can observe the total error is minimum with the models with the degree of polynomial 3. This indicates that the given data points y is of the form $y = f(x) +$

σ^2 where f(x) is a polynomial of degree 3. We can see that the irreducible error (σ^2) is very close to zero.