Machine Data and Learning

Assignment 1

Team 78
Group members:
Rishabh Khanna (2019113025)
Kshitijaa Jaglan (2019115005)

Task 1: Linear Regression

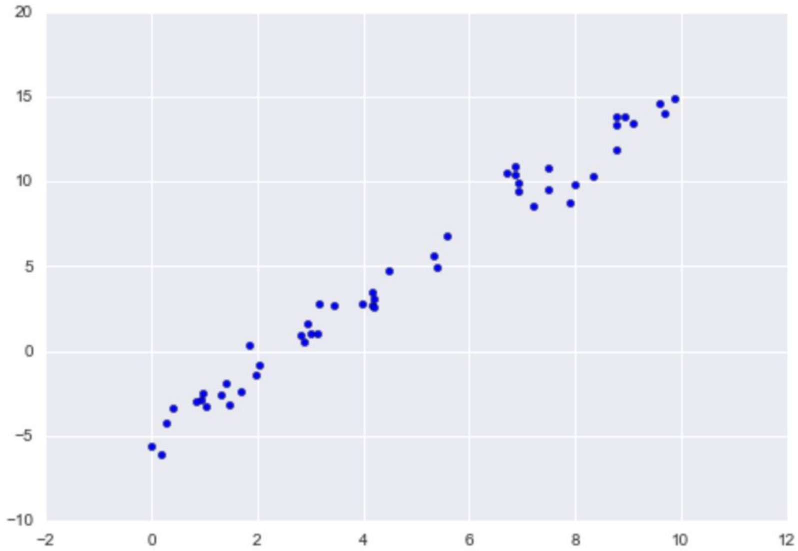
Write a brief about what function does the method, *LinearRegression().fit()* performs

This function *`LinearRegression.fit()`* from *`sklearn.linear_model`* is used to create a predictive machine learning model on a linear scale, by giving us a function *`y = ax + b`* where
`y = predicted value`
`X = input data`

It tries to fit the linear equation ***`y = ax + b`*** with the best value of ***`a`*** and ***`b`*** such that the sum of squares of the difference between **`predicted value (y)`** and **`real value`** is minimum with the **`input value (x)`** .
Here, 'b' is also known as bias coefficient

We can understand this with an example.
Say we generate some data shown below, scattered along a line of slope 2 and intercept -5.

```
r = np.random.RandomState(1)
x = 10 * r.rand(50)
y = 2 * x - 5 + r.randn(50)
plt.scatter(x, y)
```

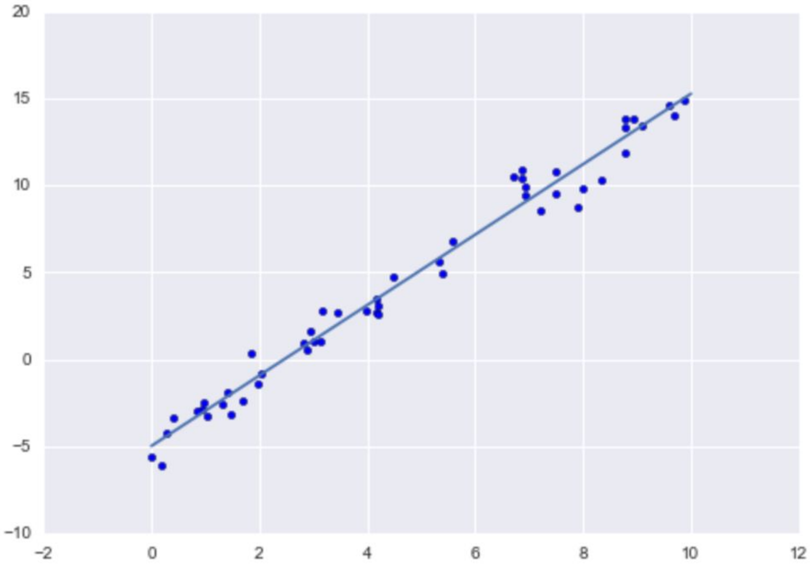


Now fitting this data using LinearRegression,

```
from sklearn.linear_model import LinearRegression
model = LinearRegression(fit_intercept=True)

model.fit(x[:, np.newaxis], y)
xfit = np.linspace(0, 10, 1000)
yfit = model.predict(xfit[:, np.newaxis])

plt.scatter(x, y)
plt.plot(xfit, yfit);
```



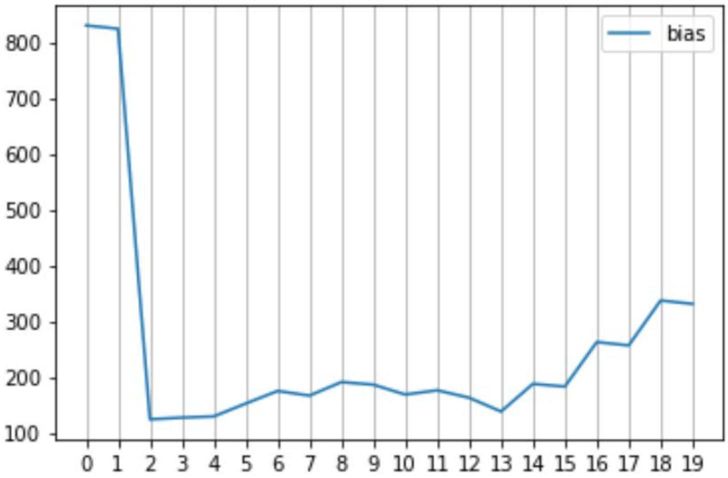
This will give us a model which will try to fit the line accurately around the data such that the average of sum square of distances of the data points from the line (Mean square error) is minimum

Task 2: Calculating Bias and Variance

All the graphs below have value on x-axis = degree - 1

Bias:

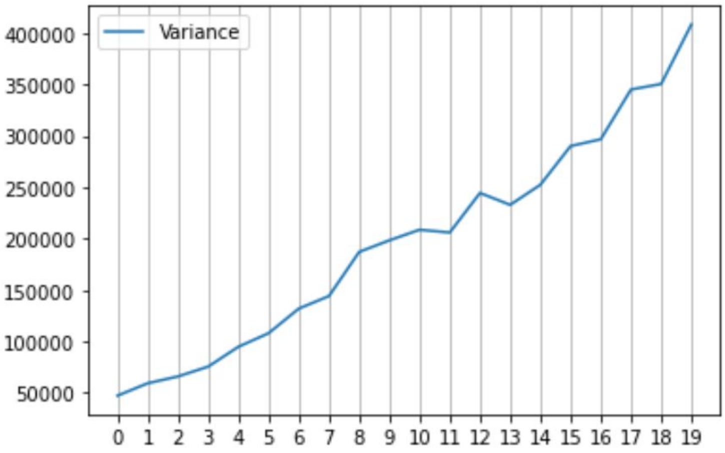
Bias is a measure to measure the accuracy of predictions made by the model. High bias leads to inaccurate predictions as it misses relevant relations between the input and output. In the computation doe, we see there is a sharp decrease till degree 3, followed by a gradual increase.

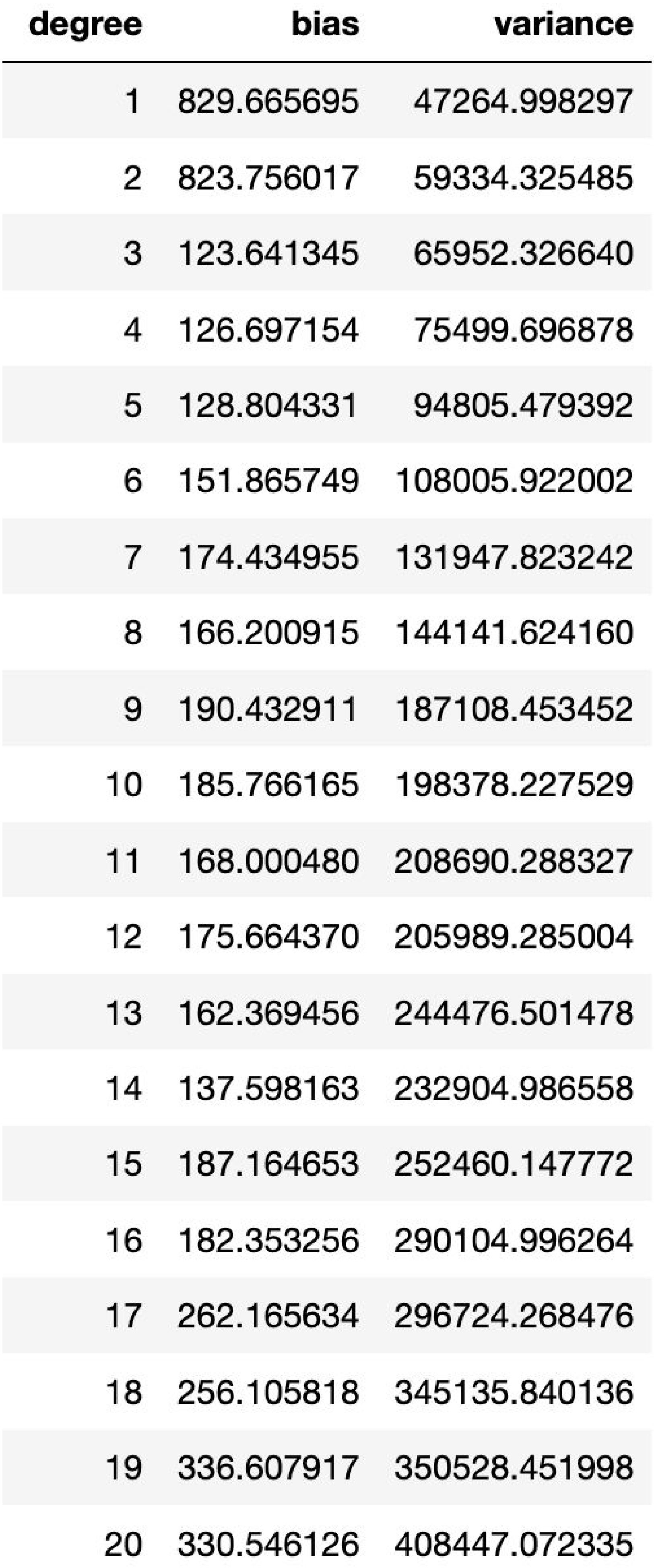


For lower values (< 3 degree), it is expected that there was a case of underfitting, leading to a high bias as the model cannot extract the required data properly.

Variance:

Variance is a measure of precision of the predictions made. A high value of variance leads to the algorithm modelling the noise from the input data. In the computations made, we see there continuous gradual increase in variance as the degree increases. As complexity increases, the variance of the data increases.



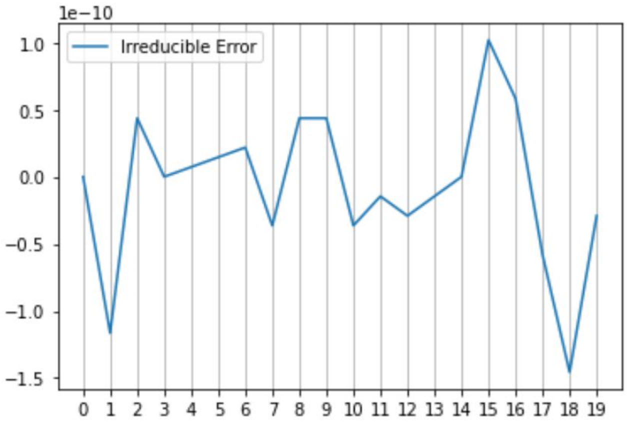


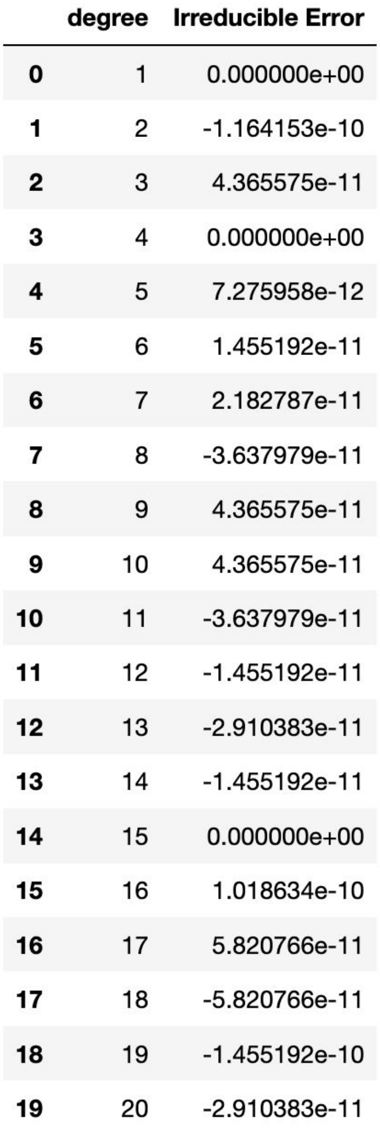
degree	bias	variance
1	829.665695	47264.998297
2	823.756017	59334.325485
3	123.641345	65952.326640
4	126.697154	75499.696878
5	128.804331	94805.479392
6	151.865749	108005.922002
7	174.434955	131947.823242
8	166.200915	144141.624160
9	190.432911	187108.453452
10	185.766165	198378.227529
11	168.000480	208690.288327
12	175.664370	205989.285004
13	162.369456	244476.501478
14	137.598163	232904.986558
15	187.164653	252460.147772
16	182.353256	290104.996264
17	262.165634	296724.268476
18	256.105818	345135.840136
19	336.607917	350528.451998
20	330.546126	408447.072335

Task 3: Calculating Irreducible Error

Irreducible is a measure of noise exhibited in the data used. Ideally it should be zero but due to the presence of noise existing in virtually every data, there is a very small amount of error which arises.

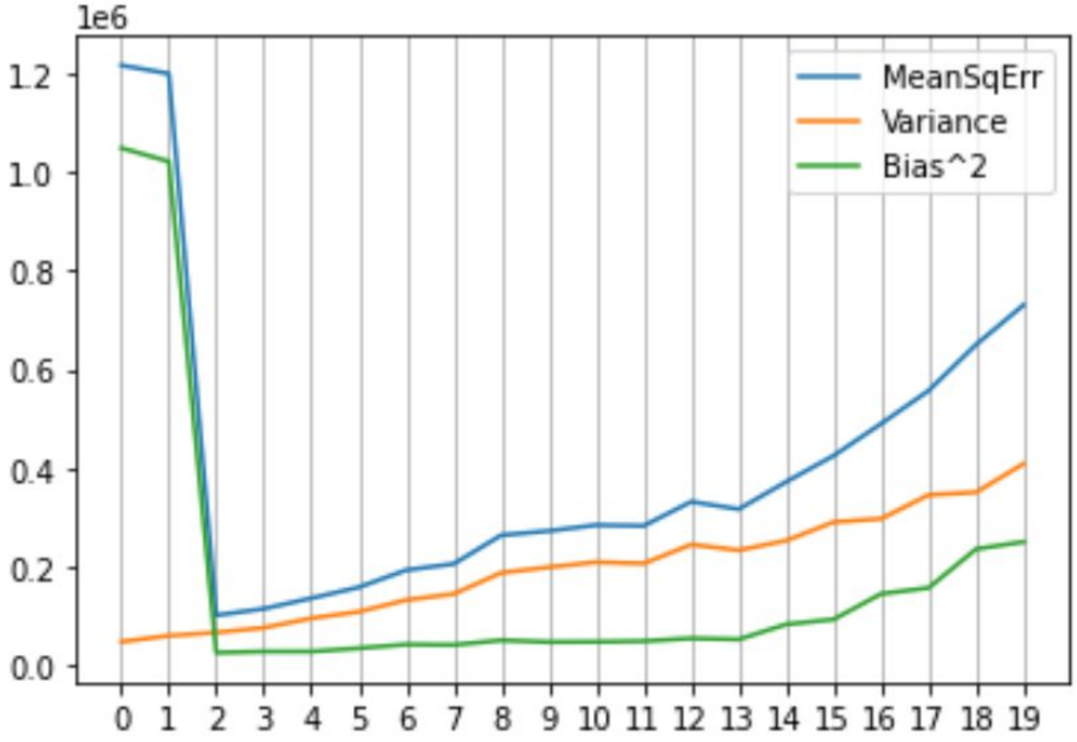
The values for the same have been shown at the right and the graph is shown below





	degree	Irreducible Error
0	1	0.000000e+00
1	2	-1.164153e-10
2	3	4.365575e-11
3	4	0.000000e+00
4	5	7.275958e-12
5	6	1.455192e-11
6	7	2.182787e-11
7	8	-3.637979e-11
8	9	4.365575e-11
9	10	4.365575e-11
10	11	-3.637979e-11
11	12	-1.455192e-11
12	13	-2.910383e-11
13	14	-1.455192e-11
14	15	0.000000e+00
15	16	1.018634e-10
16	17	5.820766e-11
17	18	-5.820766e-11
18	19	-1.455192e-10
19	20	-2.910383e-11

Task 4: Plotting Bias^2 - Variance Graph



As you can see in the graph above, the MSE and Bias^2 are the least at x = 2, which is degree = 3, indicating that the function might be a cubic one and a cubic polynomial fits the data the best out of the given options.
Moreover, the bias is initially high because of underfitting, drops to the best value, and increases again as the model conforms too closely with the test data and loses its generality, causing it to perform poorly.
The variance is continuously increasing (after degree 3) because the curve of best fit overfits the training data, resulting in an inaccurate representation of the test data while also decreasing the precision of the model.