Machine Data and Learning

Assignment 1 Report

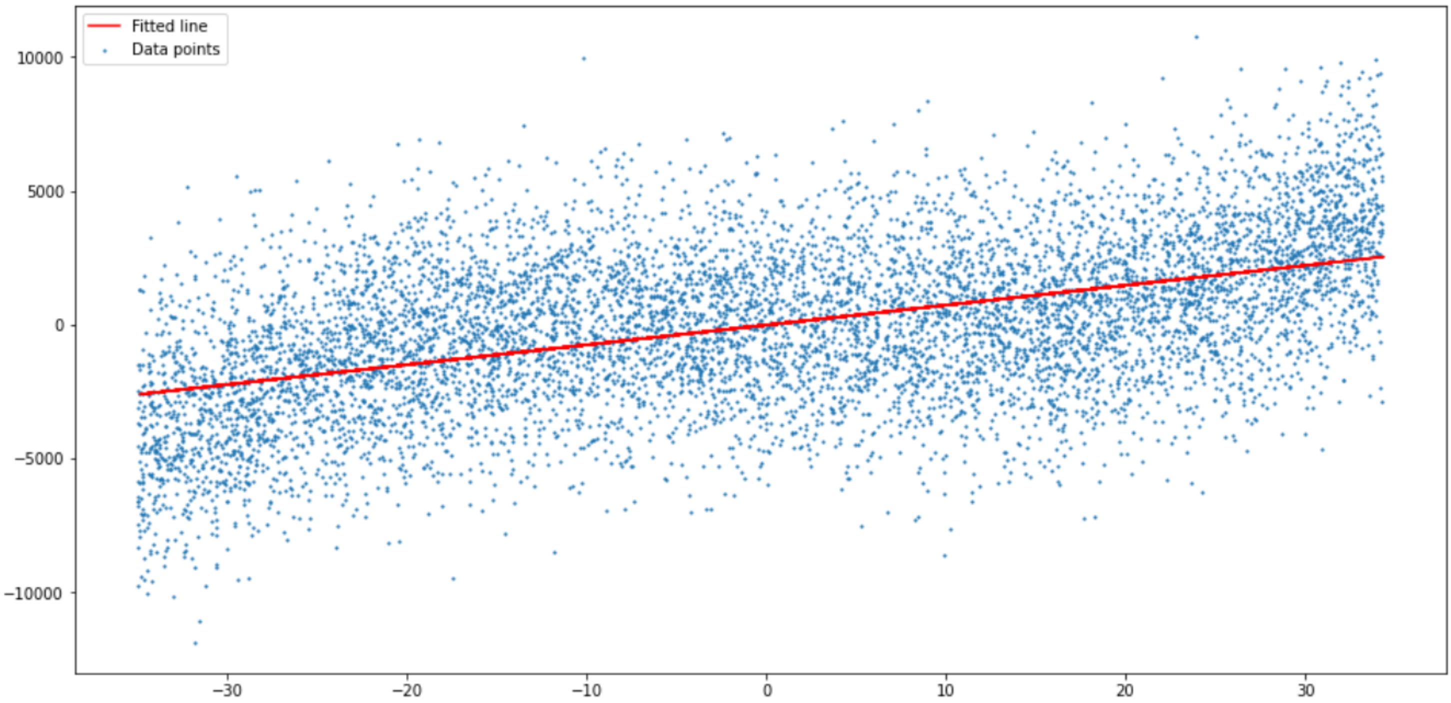
Name: Naval Surange & Aryamaan Basu Roy

Roll: 2020113018 (CND) & 2020101128 (CSE)

TASK 1

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

- The skLearn function LinearRegression().fit() has the following function declaration:
  - *LinearRegression().fit(X , Y)*
    Where,
  - X is a list of input features for the model and
  - Y is a list of output labels to which model has to fit data.
- When the function *fit()* is called it tries to find fit a curve of degree 1 through the input data X and Y such that the cost is as minimum as possible.



The function internally uses an algorithm called gradient descent which attempts to find the coefficients of the fitted linear line in such a way that the cost is as small as possible.
Mathematically, for simple linear regression, in which the label is dependent on only one feature, we have to fit a line

$$\hat{Y} = W_1 . X + b$$

Where, $\hat{Y}$ is the predicted value of label, X is the input features and W_1 and b are the weights and biases of the fitted curve such that the cost function defined as the MSE of $\hat{Y}$ and Y as,

$$MSE = \frac{1}{N} (\sum_{i=1}^{n} (y_i - \hat{Y}_i)^2),$$

where, N is the number of data points,
y_i is the actual value of observation,
And $\hat{y}_i$ is the predicted value.

Is as small as possible.

In order to find such values of W_1 and b we use a algorithm called **gradient decent** which finds a local minimum/maximum of a function(in our case it finds minima in Cost function).

TASK 2

We briefly followed the following steps-

- **Task 2.2.1**
  - Steps
    - Firstly we loaded the test and training data using pickle.load function and then extracted features and labels out of them.
    - Then we plotted the training and testing data for visualization and analysis purposes

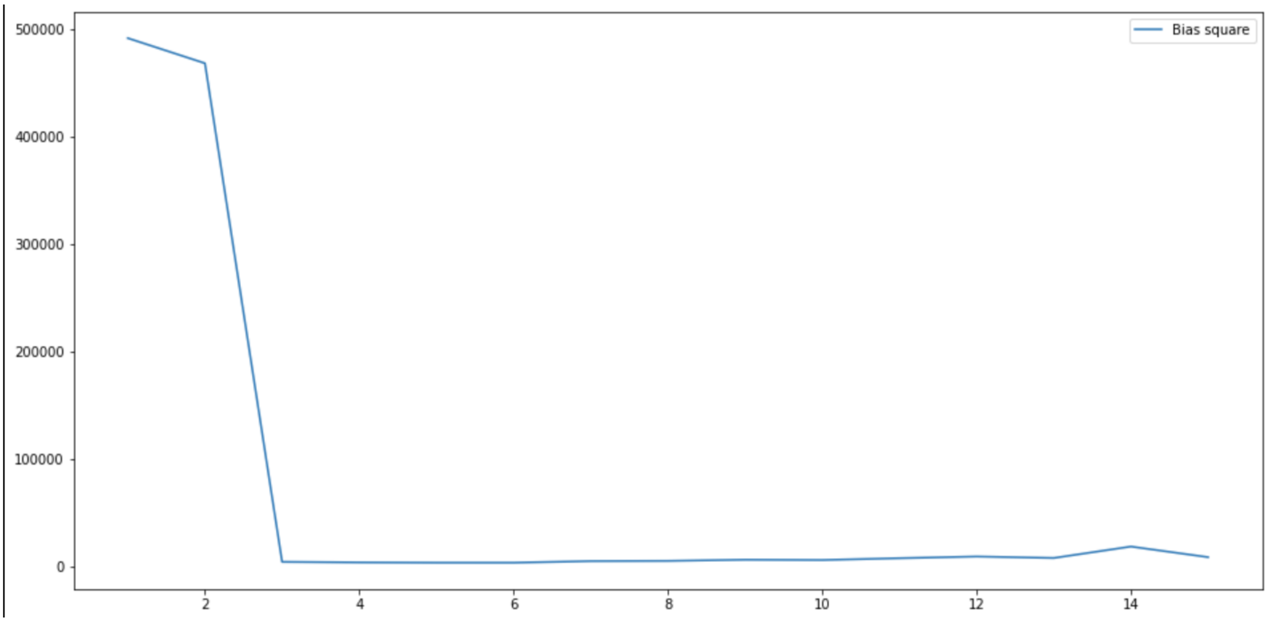
- **Task 2.2.2**
  - Steps
    - We randomly shuffled the training data using sklearn's inbuilt function **shuffle**
    - And then splitted it into 16 different training sets.
    - Now finally we trained a linear regression model over all of the above obtained 16 training sets and fitted a polynomial of degree 1 to 15 in each of them.
    - Next then we predicted the values of $\hat{y}$ using the training datasets for each of the 16×15 models.
    - Then used those predicted values to calculate the Bias using the formula $Bias = E[\hat{f}(x)] - f(x)$ and the averaged this value across all the 16 models with different datasets and plotted the average value with respect to the power of the fitted curves.
    - Similarly, We calculated the variance using the relation $Variance = E[(\hat{f}(x) - E[\hat{f}(x)])^2]$ and averaged this value across all the 16 models trained on different datasets and plotted the average with respect to the power of the fitted curves.

➢ Conclusion

- The average bias squared that we obtained were :

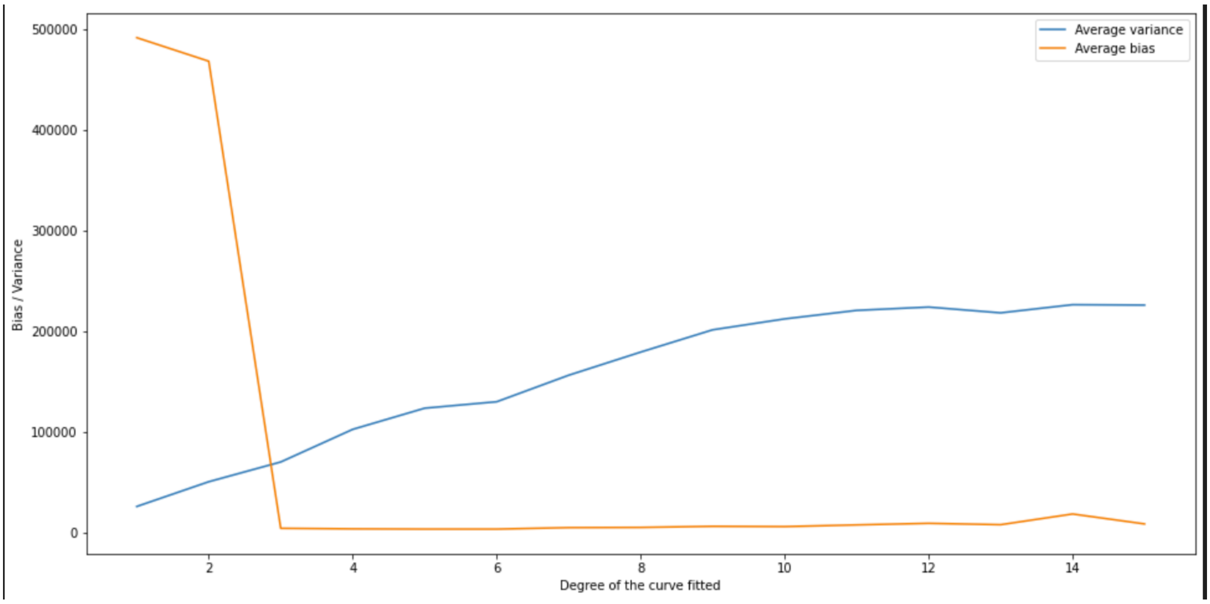
	Average bias squared	Power of fitted curve
0	491887.673133	1
1	468499.943715	2
2	4451.057676	3
3	3903.213490	4
4	3672.589815	5
5	3663.497705	6
6	5134.918871	7
7	5327.473591	8
8	6413.803773	9
9	6146.919934	10
10	7834.634013	11
11	9416.682394	12
12	8035.559776	13
13	18666.852066	14
14	8808.096175	15

- Now we plot the above values to obtain the following graph of bias square w.r.t power of fitted curve:



- We have obtained the average variance of each power of the fitted curve :

	Average Variance	Power of fitted curve
0	26134.194977	1
1	50709.739020	2
2	70350.492578	3
3	102753.301453	4
4	123782.502925	5
5	130173.990812	6
6	156526.656333	7
7	179516.396144	8
8	201637.602428	9
9	212483.623568	10
10	220954.517680	11
11	224245.882474	12
12	218501.920211	13
13	226609.164254	14
14	226230.941575	15



- The graph of the above values along with the bias square is as follows:

- **Task 2.3**
  - Steps
    - Firstly we calculated the mean squared error(MSE) for all the models.
    - The we calculated the irreduacable error by using the relation $\sigma^2 = MSE - (Bias^2 + Variance)$ and then tabulated and plotted it.
  - Conclusion
    - Irreducible error is error that cannot be minimized regardless of how excellent the model is, meaning that irreducible error is independent of the model. It is the amount of noise in the model, that is, the error that occurs when X cannot supply information for Y; this might be due to unmeasured factors that were valuable in forecasting Y, unmeasurable variance, and so on.
    - Y is equal to f(X) + e.
    - Where X is the input variable, Y denotes the response variable, and e denotes the error.
    - A good model should have a good balance of bias and variance for a low total error.
    - $Bias^2$ + Variance + Irreducible error = Total error
    - Across all function classes, the irreducible error remains about the same, i.e. near to zero.

- **Task 2.4**
  - Steps
    - In this task we plotted the already calculated values of $Bias^2$, $variance$, MSE and $Irreducabe\ error$ on a same graph and analysed the curves which is given in the next section.

  ➢ Conclusion:
    - Between the values 1 and 3, the value of bias reduces while the value of variance increases.
    - The graph below also shows that for X = 3, the error is the smallest.
    - The objective of every machine learning algorithm is to have as little bias and variation as possible.
    - As a result, a good model is one that has a good balance of bias and variance, or one that has a good tradeoff between bias and variance that minimizes overall error.
    - Bias is much higher before X = 3 and reduces abruptly from roughly X = 2 to its lowest at X = 3. Because we have a large bias and low variance until X = 3 (variance is lowest at X = 1), this suggests to an underfitting model, as expected in Task 2.
    - After X = 3, the variance gradually increases, whereas the bias remains constant until X = 11 and then increases. As a result, after X = 3, the bias is minimal and the variance is very significant until X = 19, indicating an overfitting graph.
    - The graph below corresponds to our findings in Task 3, namely, the model transitions from an underfitting to an overfitting model. Since the error is lowest at X = 3, this means that a fair balance between bias and variance exists at X = 3.
    - Finally, the data is best suited to a model of degree 3.

The graph is as follows: