Report for first and second question

For first cell I’ve imported all required packages {numpy, pandas, matplotlib, sklearn, statistics, mlxtend.evaluate

}

For sklean, I imported the following libraries

{

PolynomialFeatures: to choose model degree

LinearRegression: to create the model

}

For mlxtend.evaluate, I imported the following libraries

{

bias\_variance\_decomp: to get (Mean sequared error, Bias and the variance)

}

----------------------------

For the second cell, I have loaded the data pickle.load() and divid it into(X\_train, y\_train, X\_test, y\_test)

And I created a 3D array with dimension (10, 450, 2)

It will be 10 2D array, each 2D array will be 450\*2

First column -> train dataset

Second column -> test dataset

Also, I have plotted each 450-train sample to read the data

---------------------

In the second question(data already splitted)

For the third cell (the most important one), I have created a function, and it will train the model, it takes 4 parameter (degree, train\_data->array (10,450,2), x\_test, y\_test)

First, for loop it loops through all 10 train\_data and in the second question it will go through all 20 train\_data and it will fit them with the train data, also, it will predict x\_test values and it will save predicted values into list to plot them later with x\_test to see results and compare them with the other degrees, also it will get the avarged MSE, bias, variance and return them to plot them in the final plot.

--------------------

The rest of the cells will recall the function to an appropriate degree.

--------------------

Last cell will plot variance,bias, MSE

SUMMARY in first question

We have noticed that MSE, bias and variance are getting smaller whenever the complexity of the model is becoming higher, but on the other hand the overfitting is becoming higher.

So, from bias and variance data, I think the best model that fits data is model with **degree six,** it has the second smaller bias and the least variance.

MSE, Bias and variance of the data are:

|  |  |  |  |
| --- | --- | --- | --- |
| degree | MSE | Bias | Variance |
| 1 | 30.34395978 | 30.20185675 | 0.14210303 |
| 2 | 6.34810215 | 6.30993467 | 0.03816748 |
| 3 | 5.40426697 | 5.35980992 | 0.04445705 |
| 4 | 3.30188043 | 3.26973681 | 0.03214362 |
| 5 | 3.11033821 | 3.07158918 | 0.03874902 |
| 6 | 2.74694864 | 2.70971188 | 0.03723676 |
| 7 | 2.58063691 | 2.53687822 | 0.0437587 |
| 8 | 2.60351077 | 2.55197226 | 0.05153852 |
| 9 | 2.60835767 | 2.54974972 | 0.05860796 |

Last conclusion:

|  |  |
| --- | --- |
| degree | result |
| 1 | underfitting |
| 2 | Underfitting |
| 3 | Good fitting |
| 4 | Good fitting |
| 5 | Good fitting |
| 6 | Good fitting |
| 7 | Overfitting |
| 8 | Overfitting |
| 9 | Overfitting |

Summary in the second question

SUMMARY in first question

We have noticed that MSE, bias and variance are extremely high, and they are getting smaller sometimes and go higher again in other degrees whenever the complexity of the model is becoming higher, but on the other hand the overfitting is becoming higher.

So, from bias and variance data, I think the best model that fits data is model with **degree three,** it has the second smaller bias and the least variance.

MSE, Bias and variance of the data are:

|  |  |  |  |
| --- | --- | --- | --- |
| degree | MSE | Bias | Variance |
| 1 | 1144382.596074 | 1073241.98908762 | 71140.60698638 |
| 2 | 1203709.89013784 | 1083247.3421319 | 120462.54800594 |
| 3 | 310676.21376801 | 161131.2398695 | 149544.97389851 |
| 4 | 423951.48527679 | 224393.67287055 | 199557.81240624 |
| 5 | 521144.13298451 | 285005.18536366 | 236138.94762085 |
| 6 | 617725.47553042 | 332006.2248253 | 285719.25070511 |
| 7 | 704828.21180317 | 368054.83141904 | 336773.38038413 |
| 8 | 810954.72813213 | 425650.95420392 | 385303.77392821 |
| 9 | 920292.48092935 | 481983.90984617 | 438308.57108318 |
| 10 | 920292.48092935 | 481983.90984617 | 438308.57108318 |
| 11 | 974564.08944789 | 481434.30549826 | 493129.78394963 |
| 12 | 1047628.02698562 | 526198.48826297 | 521429.53872265 |
| 13 | 1103136.09616607 | 528048.04165819 | 575088.05450789 |
| 14 | 1165089.3608428 | 564647.60434134 | 600441.75650146 |

Ok, degree 2 has bias and variance but also you to look at the MSE

|  |  |
| --- | --- |
| degree | result |
| 1 | underfitting |
| 2 | Underfitting |
| 3 | Good fitting |
| 4 | Good fitting |
| 5 | Good fitting |
| 6 | Good fitting |
| 7 | Overfitting |
| 8 | Overfitting |
| 9 | Overfitting |
| 10 | Overfitting |
| 11 | Overfitting |
| 12 | Overfitting |
| 13 | Overfitting |
| 14 | Overfitting |