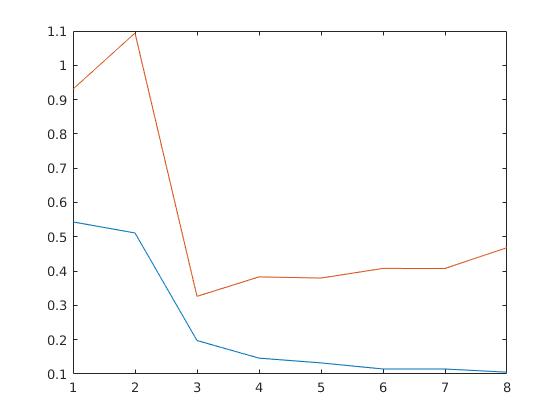
**1. Regression and over-fitting**

13.

a. Which polynomial has the lowest train SSE, which has the lowest validation SSE? (1 mark) (You can use the data cursor on the plot).



The 8th polynomial has the lowest train SSE (blue line)

The 3th polynomial has the lowest validation SSE (orange one)

b. What trend do you observe about the dependence of train and validation SSE on polynomial order? (1 mark)

The train value always lower than validation (underfitting). They both experience the significant drop between 2 and 3 polynomial, after this drop validation SSE stared to rise, in contrast test SSE continue it to decrease, but slower. This has happened because the model after 3 polynomial use become too complicated for this data and the model started to remember the values of the data instead of finding patterns in it: the overfitting of model appears. Demonstrates good performance on training data and bad performance on test data.

**2. Regularization**

7. Compute the sum squared values of the weights for the order 9 model by tying

in the command window: w\_ml’\*w\_ml, or sum(w\_ml.^2).

a) How does this differ for lambda=0.001 versus lambda=1? (1 mark)

Sum(w\_ml.^2) is much more smaller for lambda=1 (6.8684 against 1,2396)

Higher values simplify the complexity of the model and hence, reduce the risk of overfitting.

8.

c. What is the change in test error observed, and why? (1 mark)

|  |  |  |  |
| --- | --- | --- | --- |
| Best Fit  Lambda | Min Train  Error | Corresponding  Lambda | Corresponding Test  Error |
| Order 9  Data = 10 | 0 | 0 | 3492968470.54 |
| Order 9  Data = 100 | 0.84 | 0 and 0.001 | 0.31 for both |

First filled in row: min train error - 0, corresponding Lambda = 0,

Corresponding Test Error = 3492968470.54

The test error drops significantly, two Lambda values helped to gain the best result - 0 and 0.001. This drop happened because we add ten times more data to fit our model.

**3. (Cross)-validation**

5. \*Which is the correct way to complete the script? What is the correct train and

test error? (1 mark)

The correct way to complete script - to use yTest dataset for cross-valudation.

Final Train Error: 0.133, Final Test Error: 0.268

**4. A Real application of multivariate regression**

8.

c. What is the difference between the training and the testing error? Why?

(1 mark)

Trees M5P (training) – MME 2968.1157

Trees M5P (CV 2 folds) – MME 3453.2504

Trees M5P (50% Percentage split) – MME 3469.4293

Training error is the error that you get when you run the trained model back on the training data. This data has already been used to train the model is familiar with this data. Test error is the error when you get when you run the trained model on a set of data that it has previously never been exposed to.