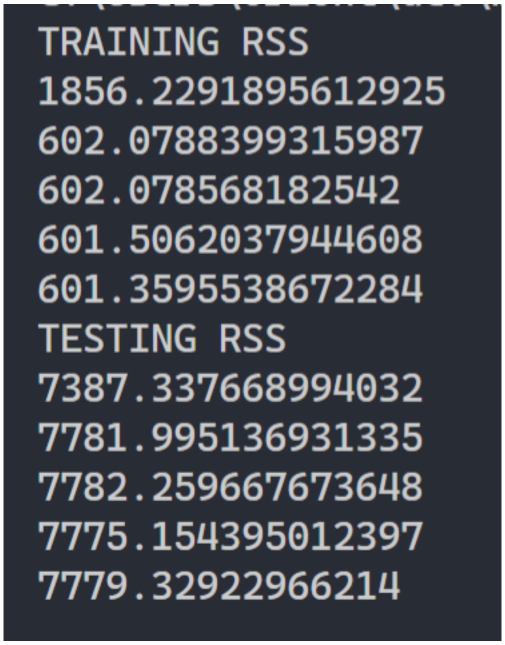
Machine Learning Assignment 62

Elijah Tarr

February 24, 2021

Problem 1

(a) *Gather your data*



```
TRAINING RSS
1856.2291895612925
602.0788399315987
602.078568182542
601.5062037944608
601.3595538672284
TESTING RSS
7387.337668994032
7781.995136931335
7782.259667673648
7775.154395012397
7779.32922966214
```

(b) *Which model is most accurate for residual sums with training data?*

In the picture above, the numbers printed for the "TRAINING RSS" are the linear, quadratic, cubic, quartic, and quintic residual sums of squares, respectiv ly. We can see a pretty obvious pattern where the residual sum starts out large, but then decreases as we add more terms. This happens because there is more complexity to the line we are fitting. For straight lines, we can only fit accurately data in which the true nature of the function is straight. For example, this could be converting inches to centimeters. But for this data, the curve is actually quadratic, as we can see from $f(x)$. However, our data tells us that the quintic regressor, the one with the most terms, has the least error when testing the training data.

(c) *Which model is most accurate for residual sums with testing data?*

In the picture above, the numbers printed for the "TESTING RSS" are also linear, quadratic, cubic, quartic, and quintic residual sums of squares respectively. There is a pattern where the error starts out small, but then gets larger and larger. This is because the more terms we add to our regressor, the more fit it becomes to that exact set of data. However, sometimes we don't want the line to only fit exactly to that data. If there is a degree of error in the data, which there usually is (error in measurement, or manually introduced error like ϵ in the uniform distribution), we want the regressor to fit the data more loosely to account for the slight variation. Regressors with higher order terms are overfit, meaning the best regressor to use in this case would be the linear one.

(d) *Which is the best model overall?*

Based on our observations, we have seen that both high and low numbers of terms can be both beneficial and harmful to the overall accuracy of the regression. Seeing as we want to determine the best one for most cases, I would choose the one in the middle: the cubic regressor (assuming I don't know what the real function is defined as). This is because it maintains a balance of being both slightly accurate with the training data, and slightly accurate with the testing data. However if I did know how the function was defined, I'd want to choose whichever regressor reflected the nature of that function. For this function, the highest order term is $0.5x^2$, which is a quadratic term. This means the true line will be a parabola, with some slight errors, meaning the quadratic regressor should in theory, be most accurate.