Q4

Rahul Atre

2023-09-28

Q4 [Statistical Learning]

1. Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.

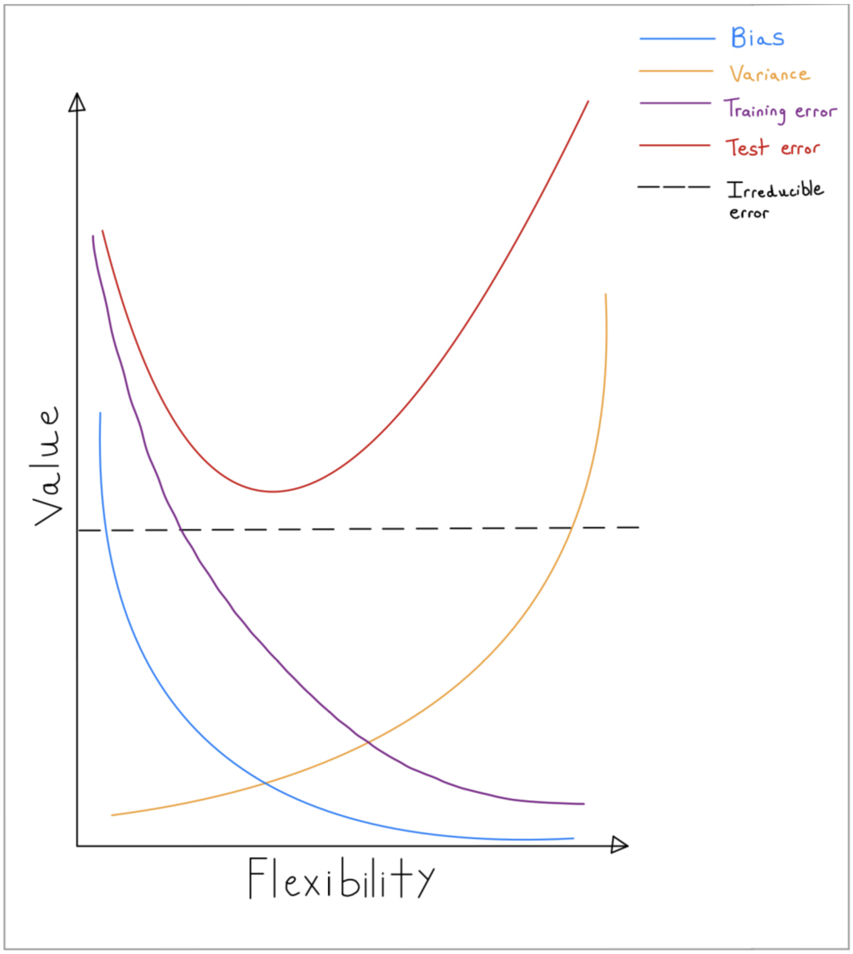

Figure 1: Flexibility vs. Value

2. Explain why each of the five curves has the shape displayed in the answer to the previous part.

- **Bias** decreases with flexibility since, as our model becomes more complex, we make less assumptions about the data. For instance, if we were to use the least squares approach, a simple linear regression model would not be flexible enough to accurately recreate a curve or arc in the data. Using more flexible techniques allows for a realistic representation of the relationship between the data.

- **Variance** increases with flexibility because the model better replicates fluctuations in the data set, making a better fit (possibly even overfitting) of the training data. It is important to note that bias and variance are inversely proportional to the flexibility of the model, due to a phenomenon known as the bias-variance trade-off. If there is too much variance, it could lead to the replication of random noise from the training set.

- **Training error** reduces with an increase in flexibility since the training data is more closely followed and better fitted.

- **Test error** initially declines with an increase in flexibility due to a less biased model, however, as it gets closer to the middle (where the bias and variance line intersect) it reaches a "sweet-spot" where the model is not too biased or overfitting. After this point, it begins increasing due to an increase in variance, leading to overfitting of random noise from the training data. It is most desirable to obtain the minimum point of the test error, one that is closest to the irreducible error line. Also, the test error stays above the irreducible error line regardless of an increase in flexibility since there will always be some error from noise or limitations of the model.

- **Irreducible error** is a constant parallel line that is below the test error. It has no relation to the flexibility of the model as it is always present in a model due to unknown variables that cannot be reduced.