STA9890 Statistical Learning: Homework 1 Solution

Soonmo Seong

Feb 18, 2021

1. The performance means accuracy on test dataset, which we can't use when training the statistical learning method.

   (a) In general, both model would work well in this dataset. However, the flexible method can fit the data better and the inflexible one is not enough to fit the data, underfitting and harming the performance.

   (b) The flexible one would overfit this high dimensional data, reducing the performance.

   (c) Given the highly non-linear relationship, the flexible statistical learning method works better than the inflexible one because the inflexible one isn't enough to capture the non-linearity of function that maps predictors to the response. For example, if the non-linear relationship follows the curve of the quadratic function, linear regression is not able to capture the curvature of the function; however, flexible learning method such as polynomial regression can capture the true relationship.

   (d) The flexible method would works worse than the inflexible one because flexible one fits all the highly variable data points, increasing the variance of method and decrease the performance.

2. (a) This is a regression problem because the response is CEO salary, which is quantitative. In addition, we are most interested in inference since we like to understand which predictors influence the salary, meaning that we interpret the regression coefficients.

   - n - 500
   - p - 3 predictors: profit, number of employees, industry

   (b) This is a binary classification problem because the target is a binary categorical variable. And, we more focus on prediction.

   - n - 20
   - p - 13 predictors: price charged for the product, marketing budget, competition price, and ten other variables

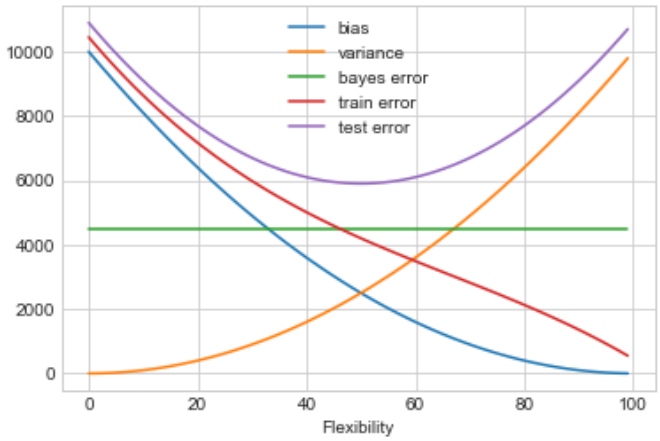
   (c) This is a regression problem for prediction since predicting the % change, which is a continuous variable, is of our interest. Moreover, prediction accuracy is more important than inference.

   - n - 52
   - p - 3 predictors: it% change in the US market, % change in the British market, % change in the German market

3. (a) Figure 1

   (b)
   - squared bias: as the flexibility of method increases, the method more fit the data. Therefore, bias keeps decreasing.

Figure 1: 3.(a) Bias-Variance Decomposition



- variance: variance keeps increasing as the flexibility increase because the method fit data more closely.
- training error: training error continue to decrease as the method become more flexible since the method fit the data well.
- test error: as the method becomes flexible, test error starts to decrease but increases at some point, where overfitting problem takes place. The reason why overfitting happens is that the method too much fits the training data and fails to perform well on test data which is quite different from training data.
- bayes error: bayes error is the lower bound of error in the method. Theoretically, it's constant.

8. Please see the link