Quiz1

Hyeonjin Lee

3/16/2021

Student ID: 2396602, hyl15102

Problem I.

True or False Questions. This problem is composed of 8 true-or-false statements. You only need to classify these as either true or false. No explanation is required. Remember to mark your answers (T or F) in the bottom table! (See quiz questions for reference)

1) False

2) True

3) True

4) False

5) False

6) True

7) True

8) False

Problem II

Suppose you have regression data generated by a polynomial of degree 3. Characterize the bias-variance of the estimates of the following models on the data with respect to the true model by circling the appropriate entry.

Linear Regression: High Bias, High Variance

Polynomial regression with degree 3: Low Bias, Low Variance

Polynomial regression with degree 10: Low Bias, High Variance

Problem III

Explain briefly (~ 1/2 page) the meaning of the bias-variance tradeoff, i.e., what do we mean by model bias and model variance, and why is there a trade-off between the two?

Bias is the inability for a machine learning method to capture the true relationship. For example, a linear regression fits a straight line to the points. However, if the true relationship was a quadratic curve, we can notice that the linear relationship will never truly fit the data because it cannot be curved. This is a large bias.

Variance is the amount that the method changes given different training data. It is the difference in fits between data sets. For example, if we imagine the quadratic curve once again, a line that connects each of the points will have a low bias because it fits the data perfectly. We can imagine that the sums of squares will be extremely small or 0 in this case. However, if we fit this exact squiggly line into different data sets, it may predict the true relationship completely wrong. There would be vastly different sums of squares for the different datasets. This squiggly line relationship is also known as overfitting.

We can see from these examples that the squiggly line had extremely low bias as it fit the training data very well. However, it has extremely high variance as it leads to very different results based on the data. The linear regression example however had high bias as it did not represent the quadratic relationship completely. However, it had a relatively low variance as it still made good predictions (not great) with different datasets. The ideal algorithm would be one that finds the sweet spot between this tradeoff relationship. We want a low bias that can accurately model the true relationship, but also one that has a low variability so it can produce consistent predictions across different datasets. This is done using different methods to find the sweet spot between a simple model and a complex model.

Problem IV

"In model (P1) a unit increase in X1 results in an increase of the predicted output by 3.2 units, i.e. it is clear that Y is positively correlated with X1. However, in model (P2) a unit increase in X1 instead results in a decrease of 1.4 units in the predicted output, i.e. now X1 appears to be negatively correlated with Y !"
Give a plausible explanation to your friend's dilemma.

Given the change in coefficients between the models, there seems to be evidence of collinearity between the predictors. It is possible that the predictors themselves (or a combination) are highly correlated with each other. This leads to an extreme swing in the coefficients between the models, even to the point where the direction is affected. Adding or removing variables will impact the model greatly. As such, it is difficult to examine the p-values to see if the independent variables are truly statistically significant and also reduces the accuracy of the estimates of the regression coefficients as shown. There may have been an increase in the Rˆ2 value but it is irrelevant as the beta weights are not statistically significant.