Problem Set 4

Data Science 602: Data Analysis and Machine Learning

Spring 2022

1. Learning Curves To evaluate the performance of four candidate models performance, you produce the learning curves shown below. For each model, supposing you want an R2 score of 0.9 or above, would you:
(a) acquire more data points .
(b) increase the complexity of the model or replace it with a more complex model.
(c) decrease the complexity of the model, or
(d) accept the model?
You can indicate more than one approach. Explain your reasoning.

a. From Learning Curve A, we can observe that the R squared value for the training model is 0. So we cannot consider this model as there will not be accurate results with such less R squared value. Increasing the complexity of the model could help with better accuracy.
b. Learning Curve B has an R squared value of 0.99, but we can observe that there is a huge difference between the training and test outputs. This will make this model over accurate which would not be considered a perfect model. However we can reduce the Model complexity to get more accuracy.
c. Learning Curve C can be considered as the most accurate model as the training and test results are converging with an R squared value of 1. Increasing or decreasing model complexity might affect accuracy so we can accept this model.
d. In Learning Curve D, we can observe that the R squared value of the test data is decreasing and there is a huge difference between the training result and test results. So having more data points could increase the accuracy of this model.

2. Validation curves You are evaluating a complexity hyperparameter C on a candidate model, and produce the validation curve shown on the following page. Given this curve, what is the optimal setting for this parameter? Why does accuracy, as measured by the R2 score, decrease if the parameter is set outside this range?

From Figure 2, we can observe that the model is under predictive at the initial values of C and we can also observe the huge gap between the curves of training and test results. As we keep on increasing at C=3 we can observe that the training and test results are converging with an R squared value of about 0.99 of the training results. But by looking at C=4 to C=4.3 we can consider it as the realistic value of more accuracy because the R squared value is decreasing after this point which would make the model more predictive.