Problem Set 4

Data Science 602: Data Analysis and Machine Learning

Spring 2022

Note: For this week's assignment, there is no need to submit code. You may submit your solution as a notebook or document.

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 R^2 score of 0.9 or above, would you: (a) acquire more data points, (b) increase the complexity of the model or replace 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.

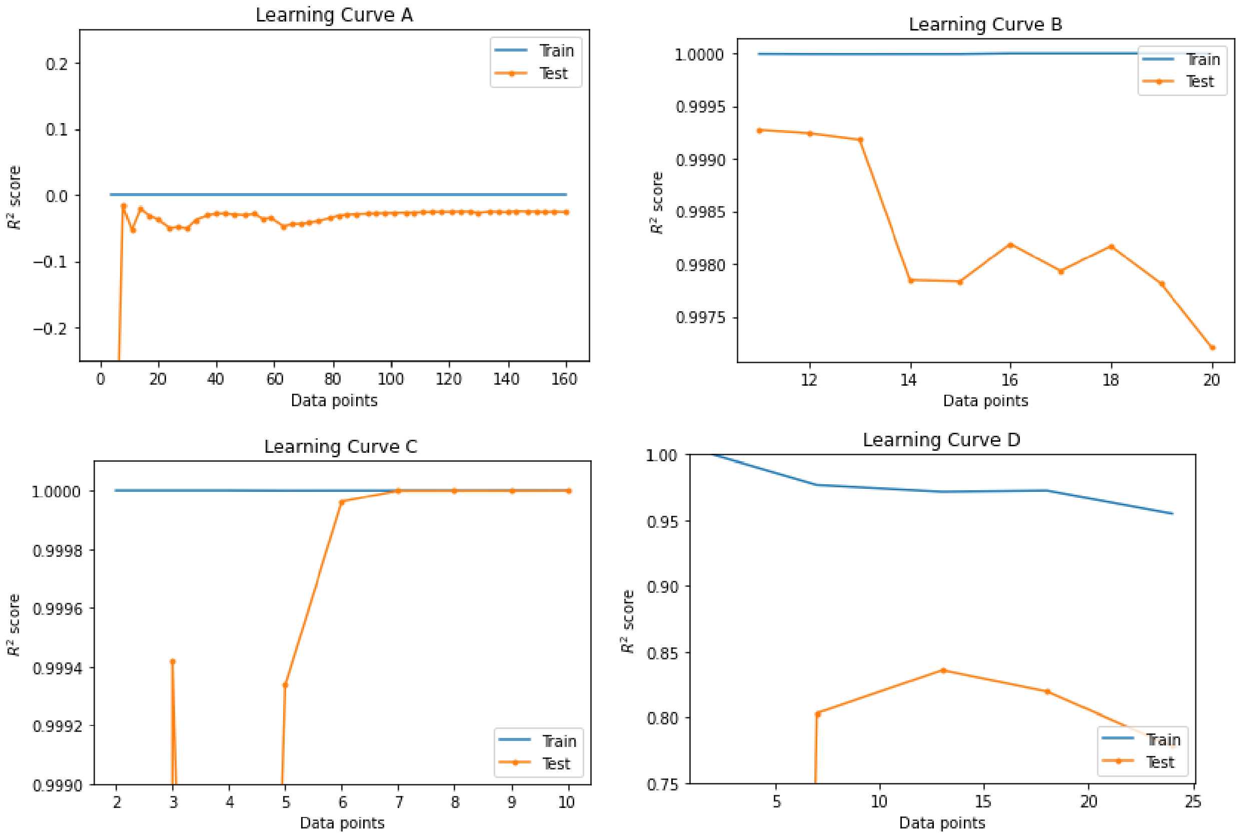

Figure 1: Learning Curves

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 R^2 score, decrease if the parameter is set outside this range?

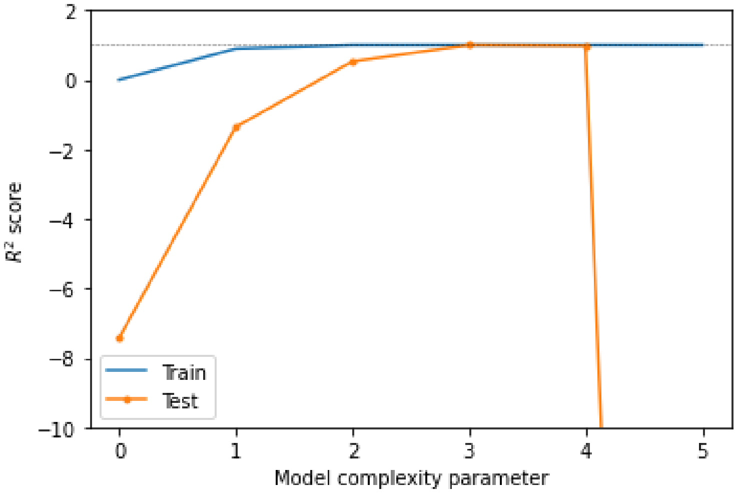

Figure 2: Learning Curves

C	R^2 (training)	R^2 (test)
0	-1.11e-16	-7.42
1	0.895	-1.36
2	1.000	0.531
3	1.000	0.99998
4	1.000	0.9829
5	1.000	-80.34