MITx: 14.310x Data Analysis for Social Scientists

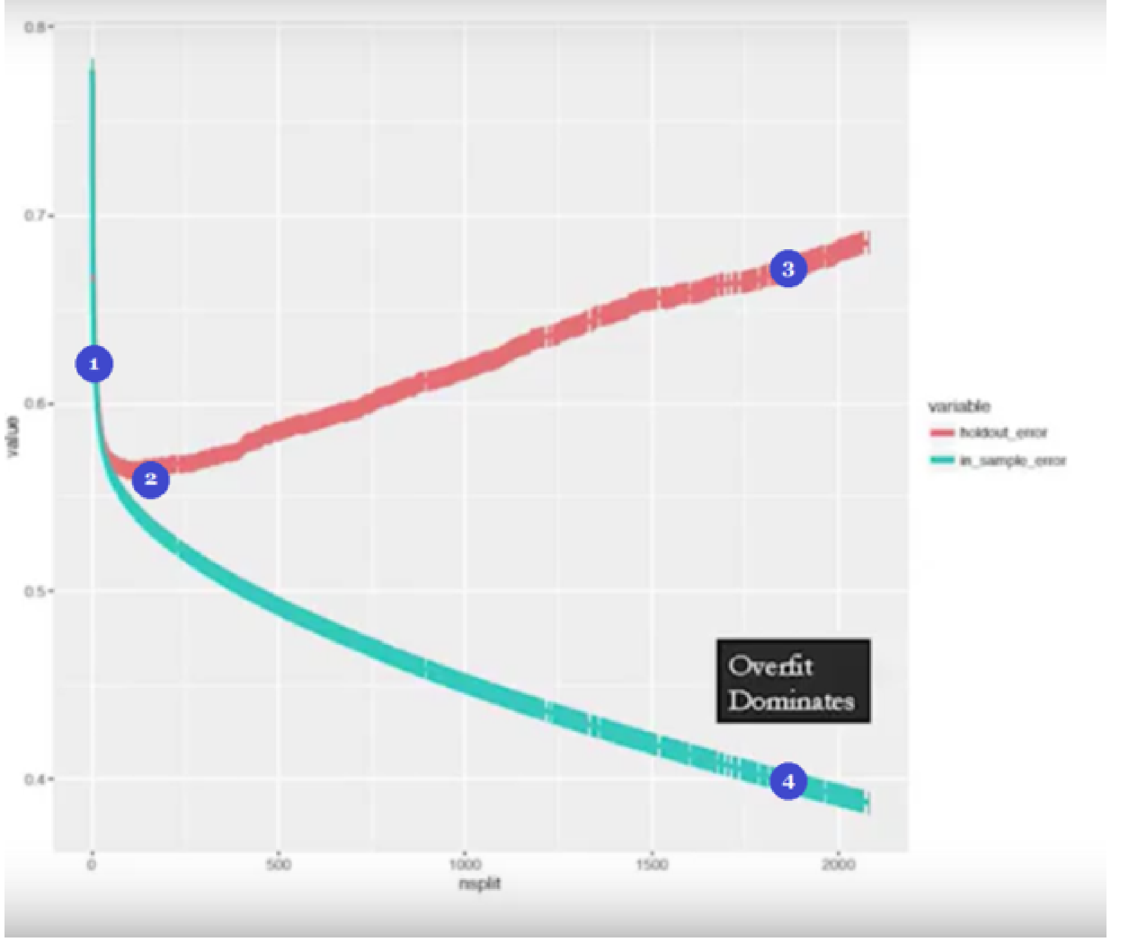
Help

Bookmarks

Observability of Prediction - Quiz

Bookmark this page

Each mark in the following graph represents a level of complexity for a decision tree and the corresponding prediction error. The y-axis represents the value of the empirical loss function, and the x-axis denotes the number of splits in a decision tree. As in lecture, the blue curve estimates this relationship in the training data, and the red curve is constructed using test data:



Using the graph above, answer the following questions:

Question 1

1.0/1.0 point (graded)

Which of the points on the plot has the lowest in-sample prediction error?

4

Answer: 4

4

Explanation

As we can see, point 4 is the point with lowest prediction error in the training sample. Note, that at that number of splits, the out-of-sample prediction error is also very high. This illustrates the over-fitting problem.

Submit

You have used 1 of 2 attempts

Question 2

1.0/1.0 point (graded)

Which of the points on the plot has the highest prediction error in the hold-out (tuning) data?

3

Answer: 3

3

Explanation

As we can see, point 3 is the highest point in our set of four points, representing the worst out-of-sample prediction. At this point, the over-fitting due to the high complexity implies it is objectively a worse predictor than point 1.

Exit Survey

Submit You have used 1 of 2 attempts

Question 3

1.0/1.0 point (graded)

Which of the points on the plot is most likely to correspond to the optimal level of complexity?

2

Answer: 2

2

Explanation

As we can see, point 2 is the point at which the sum of the prediction errors in both samples is the smallest.

Submit You have used 1 of 2 attempts

Question 4

1.0/1.0 point (graded)

In the graph above, Point 2 shifts _________ as the sample size increases.

- a. upward
- b. downward
- c. to the right ✔
- d. to the left

Explanation

Increasing the sample size implies that it will take more splits to fit the data perfectly. So intuitively, you can "split" the tree more times to improve both your in-sample and out-of-sample fits, before you start overfitting the data. (Note: it takes more splits to overfit the data if you have more data).

The important point to note here is that sample size matters a lot in machine learning. As you can see, increasing the sample size allows you to fit more complex models without sacrificing your prediction error both in your training and tuning datasets. As Prof. Mullainathan mentioned, assuming you have a reasonably complex model, the quality of your predictions can really only be improved by adding more observations. Because then you can fit a more complex model, without the overfitting prediction error dominates.

So the quality of your algorithm can only take you so far, your ability to predict is bounded by your number of observations.

Submit You have used 1 of 2 attempts

Discussion

Topic: Module 11 / Observability of Prediction - Quiz

Show Discussion