



Bookmarks

- ▶ [Module 1: The Basics of R and Introduction to the Course](#)
- ▶ [Entrance Survey](#)
- ▶ [Module 2: Fundamentals of Probability, Random Variables, Distributions, and Joint Distributions](#)
- ▶ [Module 3: Gathering and Collecting Data, Ethics, and Kernel Density Estimates](#)
- ▶ [Module 4: Joint, Marginal, and Conditional Distributions &](#)

Module 11: Intro to Machine Learning and Data Visualization > Machine Learning II > Econometrics and Machine Learning - Quiz

## Econometrics and Machine Learning - Quiz

🔖 Bookmark this page

### Question 1

1.0/1.0 point (graded)

True or False: Machine learning algorithms decide the optimal training-tuning data split.

☐ a. True

☒ b. False ✓

### Explanation

As Prof. Mullainathan mentioned, how to split your data between test and training data is usually decided arbitrarily. However, one could apply the “power calcs” framework to take a more structured approach in making this decision.

Submit

You have used 1 of 1 attempt

## Functions of Random Variable

- ▶ Module 5: Moments of a Random Variable, Applications to Auctions, & Intro to Regression
- ▶ Module 6: Special Distributions, the Sample Mean, the Central Limit Theorem, and Estimation
- ▶ Module 7: Assessing and Deriving Estimators - Confidence Intervals, and Hypothesis Testing
- ▶ Module 8: Causality, Analyzing Randomized Experiments, & Nonparametric Regression
- ▶ Module 9: Single and Multivariate Linear

## Question 2

0/1 point (graded)

True or False: In machine learning, cross validation is used to find the model that maximizes the out of sample prediction.

☒ a. True ✖

☐ b. False

## Explanation

This is an important point. Cross validation (or this type of tuning) is used to find the complexity that does best out of sample. It is not used to find the model that does best out of sample. This is precisely how machine learning solves “the curse of dimensionality”. We reduce the problem of selecting the best model (which usually is very high dimensional, if we have a large number of features) to a very low-dimensional problem to find the complexity that does best out of sample.

Here’s an example to illustrate these concepts:


Suppose you have data on amazon book sales, and **100** variables that contain different information about the book, and all the corresponding details of the book’s page on amazon over time. You are interested in predicting the demand for books, but don’t have any theory of the determinants.

You want to fit a regression tree model to this data. To this goal, you load your data into R. Now, you don’t know how deep that tree should be (how many splits it should have). Following the tuning procedure, you tell R to fit the regression tree of depth  $d$  that best fits your training data for each possible number of splits  $x$  ranging over some sufficiently wide range.


## Models

- ▶ Module 10: Practical Issues in Running Regressions, and Omitted Variable Bias
- ▼ Module 11: Intro to Machine Learning and Data Visualization


### Machine Learning I

Finger Exercises due Dec 12, 2016  
05:00 IST 

### Machine Learning II

Finger Exercises due Dec 12, 2016  
05:00 IST 

### Visualizing Data

Finger Exercises due Dec 12, 2016  
05:00 IST 

- ▶ Module 12: Endogeneity, Instrumental Variables, and Experimental Design
- ▶ Exit Survey

R automatically computes measures of fit (R squared is the default) for each of these models. Now, this graph would be obtained by plotting your measure of fit resulting from fitting a regression tree model with  $x$  splits. So each of these models might weight different features differently, and may or may not included the same feature set. But you have now reduced the problem of selecting the best model among all possible models ( $2^d$  possible models), to just finding the optimal complexity (the optimal number of splits), which is just **1** dimensional. At this point, we use cross-validation to choose the level of complexity that does best out of sample, not to select the model.

Submit

You have used 1 of 1 attempt

✘ Incorrect (0/1 point)

## Discussion

**Topic:** Module 11 / Econometrics and Machine Learning - Quiz

Show Discussion



© 2016 edX Inc. All rights reserved except where noted. EdX, Open edX and the edX and Open EdX logos are registered trademarks or trademarks of edX Inc.

