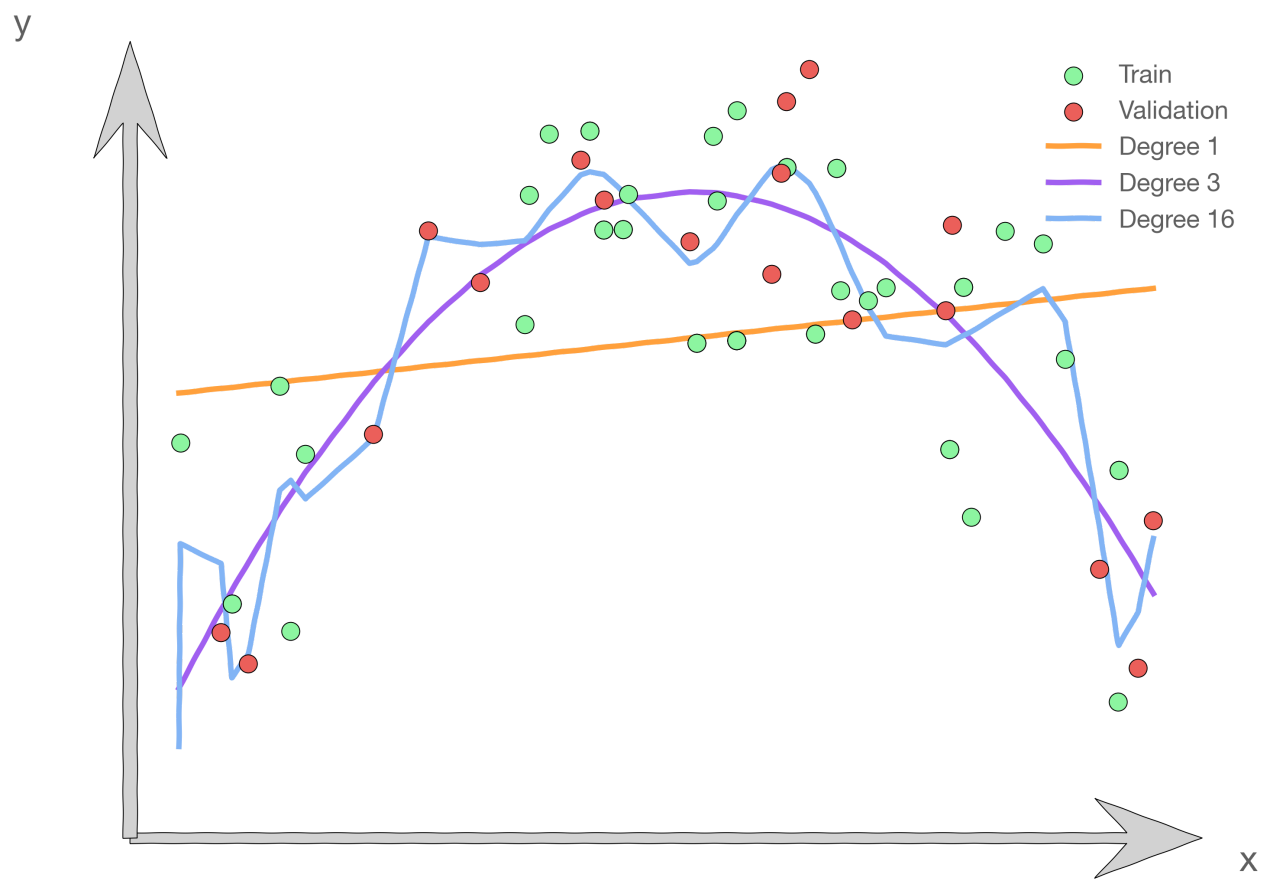
Cross-Validation

Bookmark this page

Motivation

We've seen in the exercise a case where selecting a model's polynomial degree using validation loss gave us very poor results. The plot below shows another example of this.



It is obvious that, out of the choices shown, a degree of 3 is the desired model. But the validation set by chance favors the linear model.

One solution to the problems raised by using a single validation set is to evaluate each model on *multiple* validation sets and average the validation performance. For example, one can randomly split the training set into training and validation multiple times but randomly creating these sets can create the scenario where important features of the data never appear in our random draws.

Cross-Validation

Recall that we use the train split of the data to train a model, the validation split to select the model, and the test split to evaluate the model performance. In the beginning, we always separated a portion of the data from the main dataset which we never touched until the very end. This is the test set we use to evaluate the performance of the final model.

We break down...

Entire Dataset

into...

80% Training	20% Test

In cross validation, the train data is split into k buckets or 'folds'. Then, iteratively, a new fold is used for validation and the remaining for training. This is repeated until all folds have been used as a validation set.

Training 1	Training 2	Training 3	Validation	Testing
Training 1	Training 2	Validation	Training 4	Testing
Training 1	Validation	Training 3	Training 4	Testing
Validation	Training 2	Training 3	Training 4	Testing

Calculator

Validation Score

The validation score for cross validation is the average score across all k validation folds:

$$MSE^{val} = \frac{1}{k}\sum_{i=1}^{k} MSE_i^{val}$$

Choosing number of folds

This is a choice we make (another hyper-parameter!), and it depends on the size of your data. We want every validation fold to have more than 50 or so data points and we want the spread of the validation MSE folds not to be too large. Five folds is a typical number we use.

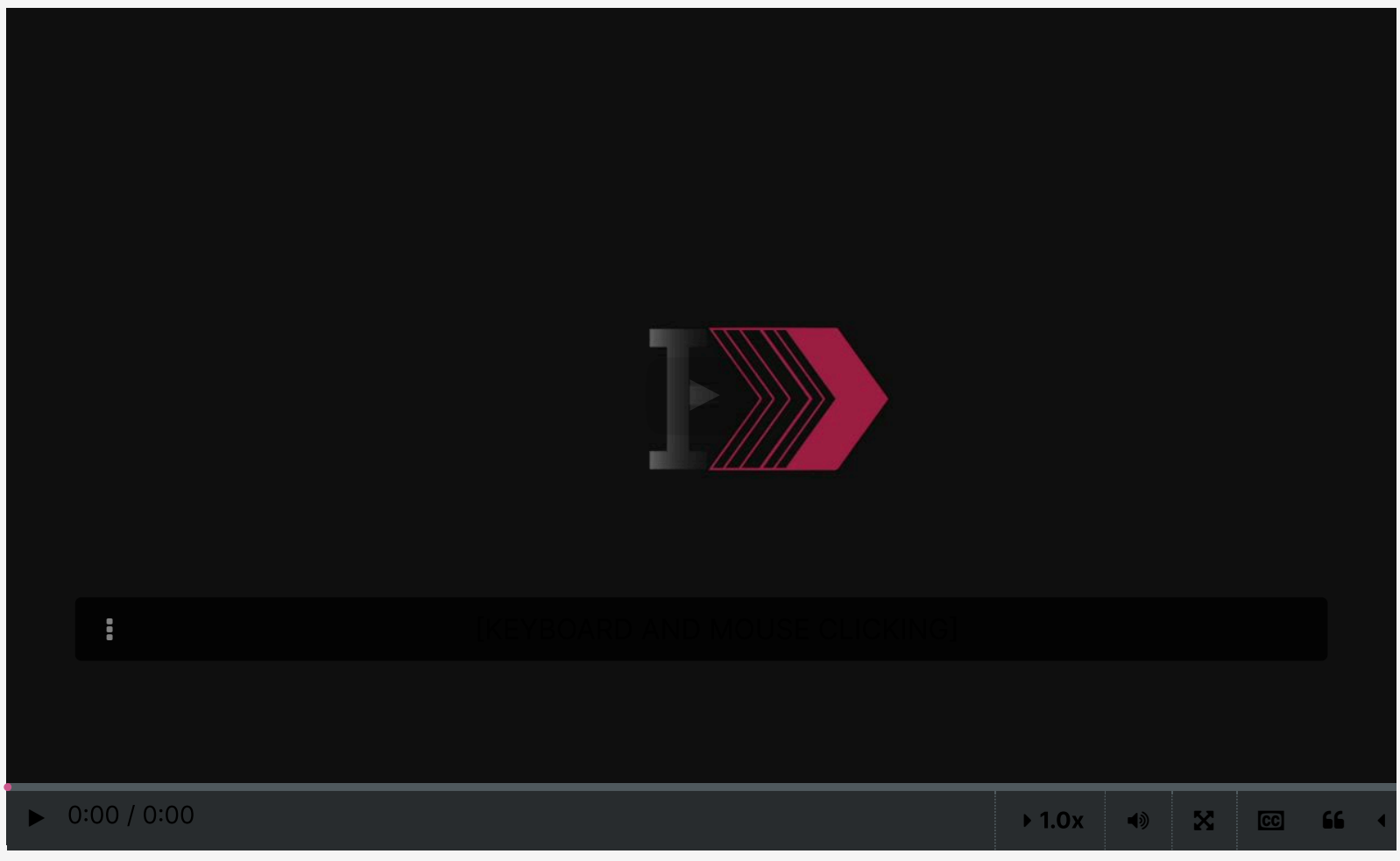
When to use cross-validation

We may choose to use a single validation set rather than cross-validation if training takes too long as in the case of a very complex model like a deep neural network or when the dataset is large. Generally, the larger your dataset, the less sensitive your model selection process will be to your choice of validation set.

Missing Python Functions?

What happens when you're trying to do a train-test-validation split... but you can't find the right functions for it? Varshini ran into that issue herself. Here's how she solved it.

Video



Video

Descargar archivo de video

Transcripciones

Descargar SubRip (.srt) file

Descargar Text (.txt) file

Discussion Board (External resource)

Haga clic en Aceptar para que su nombre de usuario y dirección de correo electrónico se envíen a una aplicación de terceros.

Calculator

Previous Next

edX

About
Affiliates
edX for Business
Open edX
Careers
News

Legal

Terms of Service & Honor Code
Privacy Policy
Accessibility Policy
Trademark Policy
Sitemap
Cookie Policy
Your Privacy Choices

Connect

Idea Hub
Contact Us
Help Center
Security
Media Kit

Calculator