Cross validation (StatQuest)

<https://www.youtube.com/watch?v=fSytzGwwBVw>

There are many ways you can split a data set into training and test. Cross validation is a method wherein we decide on some initial parameters and try every way of splitting the data using those parameters.

n-fold validation splits the data set into n chunks, then tries every possible split wherein n-1/n blocks constitute the training set, and the 1/n blocks constitute the test set.

leave one out validation treats every observation in the set as a single block.

Bootstrapping (StatQuest)

https://www.youtube.com/watch?v=Xz0x-8-cgaQ

Bootstrapping versus cross validation

<https://datascience.stackexchange.com/questions/32264/what-is-the-difference-between-bootstrapping-and-cross-validation>

Creates a surrogate data set using sampling with replacement.

Training error versus test error

<https://www.youtube.com/watch?v=vBeg-HySamE>

Split your data set into a training set and a test set. If we test the predictive accuracy of a model on the training set, the resulting average error rate is called the training error.

If we test the predictive accuracy of a model on NEW data, the resulting average error rate is called the test error.

Training error is not a good estimate of test error.

We want to minimize test error.

If you make a model too flexible, it will overfit the training data and as a result, your test error will be high.

If you make the model too stiff, it will underfit the training data and as a result, your test error will be high.

Find the goldilocks zone in between.

Scikit Learn pipeline

<https://www.youtube.com/watch?v=jzKSAeJpC6s>

Scikit learn pipeline (if you already get the general idea)

https://www.youtube.com/watch?v=1Y6O9nCo0-I