Splitting Your Data

As the news story example (/machine-learning/data-prep/construct/sampling-splitting/example) demonstrates, a pure random split is not always the right approach.

A frequent technique for online systems is to split the data by time, such that you would:

- Collect 30 days of data.
- Train on data from Days 1-29.
- Evaluate on data from Day 30.

For online systems, the training data is older than the serving data, so this technique ensures your validation set mirrors the lag between training and serving. However, time-based splits work best with very large datasets, such as those with tens of millions of examples. In projects with less data, the distributions end up quite different between training, validation, and testing.

Recall also the data split flaw from the machine learning literature project described in the Machine Learning Crash Course (/machine-learning/crash-course/18th-century-literature). The data was literature penned by one of three authors, so data fell into three main groups. Because the team applied a random split, data from each group was present in the training, evaluation, and testing sets, so the model learned from information it wouldn't necessarily have at prediction time. This problem can happen anytime your data is grouped, whether as time series data, or clustered by other criteria. Domain knowledge can inform how you split your data.

ign a split that is representative of your data, consider what the data represents. The golden rule applies t
plits as well: the testing task should match the production task as closely as possible.

For additional review, see these modules in the Machine Learning Crash Course:

- Splitting Data (/machine-learning/crash-course/training-and-test-sets/splitting-data)
- Real-world example of a data splitting flaw in an ML literature project (https://developers.google.com/machine-learning/crash-course/18th-century-literature)

Previous
← Data Split Example (/machine-learning/data-prep/construct/sampling-splitting/example)
Next
Randomization (/machine-learning/data-prep/construct/sampling-splitting/randomization) →

Last updated 2019-07-11 UTC.