Introduction to Constructing Your Dataset

ng Objectives

When measuring the quality of a dataset, consider reliability, feature representation, and availability at ser time.

Join logs from multiple and complex log sources.

Distinguish between direct and indirect labels.

Explain how a random split of data can result in an inaccurate classifier.

Use downsampling to handle imbalanced data.

Recognize how these sampling and filtering techniques impact your data.

Steps to Constructing Your Dataset

To construct your dataset (and before doing data transformation), you should:

1. Collect the raw data.
2. Identify feature and label sources.
3. Select a sampling strategy.
4. Split the data.

These steps depend a lot on how you've framed your ML problem. Use the self-check below to refresh your memory about problem framing and to check your assumptions about data collection.

Self-check of Problem Framing and Data Collection Concept

For the following questions, click the desired arrow to check your answer:

You're on a brand new machine learning project, about to select your first features. How many features should you pick?

Pick as many features as you can, so you can start observing which features have the strongest predictive power.

Pick 4-6 features that seem to have strong predictive power.

Pick 1-3 features that seem to have strong predictive power.

Your friend Sam is excited about the initial results of his statistical analysis. He says that the data show a positive correlation between the number of app downloads and the number of app review impressions. But he's not sure whether they would have downloaded it anyway without seeing the review. What response would be most helpful to Sam?

Trust the data. It's clear that that great review is the reason users are downloading the app.

You could run an experiment to compare the behavior of users who didn't see the review with similar users who did.

Previous
← The Process (/machine-learning/data-prep/process)

Next
Size and Quality of a Dataset → (/machine-learning/data-prep/construct/collect/data-size-quality)

Last updated 2021-02-05 UTC.