Sampling and Splitting Data

Introduction to Sampling

It's often a struggle to gather enough data for a machine learning project. Sometimes, however, there is *too much* data, and you must select a subset of examples for training.

How do you select that subset? As an example, consider Google Search. At what granularity would you sample its massive amounts of data? Would you use random queries? Random sessions? Random users?

Ultimately, the answer depends on the problem: what do we want to predict, and what features do we want?

- To use the feature *previous query*, you need to sample at the session level, because sessions contain a sequence of queries.
- To use the feature *user behavior from previous days*, you need to sample at the user level.

Filtering for PII (Personally Identifiable Information)

If your data includes PII (personally identifiable information), you may need to filter it from your data. A policy may require you to remove infrequent features, for example.

This filtering will skew your distribution. You'll lose information in the tail (the part of the distribution with very low values, far from the mean).

This filtering is helpful because very infrequent features are hard to learn. But it's important to realize that your dataset will be biased toward the head queries. At serving time, you can expect to do worse on serving examples from the tail, since these were the examples that got filtered out of your training data. Although this skew can't be avoided, be aware of it during your analysis.

filter PII from your dataset, and in the process you remove the tail, the dataset will be biased toward the he
s. Consider the implications for your project.

Previous

← Check Your Understanding (/machine-learning/data-prep/construct/collect/check-your-understanding)

Next
Imbalanced Data → (/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data)

Last updated 2019-07-11 UTC.