Datasets: Transforming data

Machine learning models can only train on floating-point values. However, many dataset features are *not* naturally floating-point values. Therefore, one important part of machine learning is transforming non-floating-point features to floating-point representations.

For example, suppose `street names` is a feature. Most street names are strings, such as "Broadway" or "Vilakazi". Your model can't train on "Broadway", so you must transform "Broadway" to a floating-point number. The Categorical Data module explains how to do this.

Additionally, you should even transform most floating-point features. This transformation process, called **normalization**, converts floating-point numbers to a constrained range that improves model training. The Numerical Data module explains how to do this.

Sample data when you have too much of it

Some organizations are blessed with an abundance of data. When the dataset contains too many examples, you must select a *subset* of examples for training. When possible, select the subset that is most relevant to your model's predictions.

Filter examples containing PII

Good datasets omit examples containing Personally Identifiable Information (PII). This policy helps safeguard privacy but can influence the model.

See the Safety and Privacy module later in the course for more on these topics.

Key terms:

- Normalization

Was this helpful?
Send feedback

Last updated 2024-10-09 UTC.