Categorical data: Common issues

On this page

Numerical data is often recorded by scientific instruments or automated measurements. Categorical data, on the other hand, is often categorized by human beings or by machine learning (ML) models. *Who* decides on categories and labels, and *how* they make those decisions, affects the reliability and usefulness of that data.

Human raters

Data manually labeled by human beings is often referred to as *gold labels*, and is considered more desirable than machine-labeled data for training models, due to relatively better data quality.

This doesn't necessarily mean that any set of human-labeled data is of high quality. Human errors, bias, and malice can be introduced at the point of data collection or during data cleaning and processing. Check for them before training.

Any two human beings may label the same example differently. The difference between human raters' decisions is called **inter-rater agreement**. You can get a sense of the variance in raters' opinions by using multiple raters per example and measuring inter-rater agreement.

Click to learn about inter-rater agreement metrics

The following are ways to measure inter-rater agreement:

- Cohen's kappa and variants
- Intra-class correlation (ICC)
- Krippendorff's alpha

For details on Cohen's kappa and intra-class correlation, see Hallgren 2012. For details on Krippendorff's alpha, see Krippendorff 2011.

Machine raters

Machine-labeled data, where categories are automatically determined by one or more classification models, is often referred to as *silver labels*. Machine-labeled data can vary widely in quality. Check it not only for accuracy and biases but also for violations of common sense, reality, and intention. For example, if a computer-vision model mislabels a photo of a chihuahua as a muffin, or a photo of a muffin as a chihuahua, models trained on that labeled data will be of lower quality.

Similarly, a sentiment analyzer that scores neutral words as -0.25, when 0.0 is the neutral value, might be scoring all words with an additional negative bias that is not actually present in the data. An oversensitive toxicity detector may falsely flag many neutral statements as toxic. Try to get a sense of the quality and biases of machine labels and annotations in your data before training on it.

High dimensionality

Categorical data tends to produce high-dimensional feature vectors; that is, feature vectors having a large number of elements. High dimensionality increases training costs and makes training more difficult. For these reasons, ML experts often seek ways to reduce the number of dimensions prior to training.

For natural-language data, the main method of reducing dimensionality is to convert feature vectors to embedding vectors. This is discussed in the Embeddings module later in this course.

Key terms:

- Inter-rater agreement

Help Center

Previous
Vocabulary and one-hot encoding (10 min)

Next
Feature crosses (5 min)

Was this helpful?

Send feedback

Last updated 2024-10-18 UTC.