Label-based Bootstrap Sampling

Estimated time: 4 minutes

One important way to deal with a dataset that has imbalance classes is to use bootstrap sampling, like we talked about in the last lesson, to artificially enhance the dataset so that the classes are more balanced.

There are two main methods that can be used to do this:

You can increase instances from the minority class. This is called over-sampling (or sampling with replacement).

You can remove instances from the majority class, called under-sampling.

There are also methods that combine these two methods.

Over sampling works better if you have a smaller dataset and cannot afford to remove any samples, while undersampling can work if you have a large dataset.

One thing to note is that you don't necessarily need to achieve a perfect 1 to 1 ratio between the classes. You can try out different methods and see what works to get better model performance.

Besides these sampling methods, there are also ways to generate synthetic data samples by randomly sampling the attributes from instances in the minority class. There are algorithms that are specifically designed to generate such synthetic samples. The most commonly used is called SMOTE or the Synthetic Minority Over-sampling Technique.

SMOTE works by oversampling, where it creates synthetic samples from the minority class instead of creating copies like traditional oversampling methods.