Assessing a dataset quality

Tom Rochette <tom.rochette@coreteks.org>

March 22, 2020 — 31842c2f

1 Problem

I've been given a dataset and I need to assess its quality.

2 Solution

Use Pandas Profiling to quickly generate a document that will provide you with a first overview of the data.

Your first step should be to look for warnings and messages at the top of the document. Look for entries about missing values, those will point you to variables that may need attention during the data cleaning and data imputation phases of your machine learning problem. As you are doing an assessment, simply indicate that data is missing in these variables and then see if you can determine why by looking at a few examples by loading the data in a pandas dataframe.

Are there a lot of duplicated rows? Depending on the data you've been provided, this may help you identify whether or not something is wrong with the data you were provided. If all entries are supposed to be unique because they represent a single (entity, timestamp, target) tuple, then you should ask yourself why it isn't the case. Is it possible that the dataset was created by appending a collection of other documents, leading to duplicate lines? If so, you may have to do some dataset preprocessing in order to get rid of duplicate rows.

Look for variables that are indicated as highly correlated with other variables. High correlation means that it may be possible that one variable has exactly (or almost) the same values as the other variable, which would provide little information to a machine learning model. It would also mean that picking one variable out of two correlated variables would avoid the cost of storing both.

Look at each variable in turn and view its details.

Look at the distribution of values. Are they uniformly distributed, normally distributed, binomially distributed, etc.?

If there are only two possible values for a variable, are those values approximately the same or one value is dominant compared to the other? Were you to try and predict this variable, you would have to deal with class imbalance.

Are the values of the variables sensible to you? Are variables composed of multiple information, such as the value and the unit used for the measurement? You would generally prefer composite values to be separated into different variables as it will be easier to process using machine learning models.

When looking at numbers distribution, are there outliers (values that are either a lot smaller or larger than the rest)? It is sometimes important to ask those who provided you with the data if they can explain those outliers. In general you will want to ignore outliers during training as they may skew your model toward them, resulting in less than ideal results for all the other data points.

The quality of a dataset is inversely proportional to the number of operations you need to apply to it to make it a clean dataset. That is to say that if you don't need to do anything on the data provided to you, then it is a good dataset.