Building Datasets

Written entirely by
Sreetam Ganguly
sreetamneverrests@gmail.com

Dated: October 10, 2020

Preparation of datasets is the first step in any Machine Learning Project. In general, the following steps should be followed:

1. Data must be gathered from reliable sources and stored in the desired format. The data may be in Excel sheets, ODS or CSV, or it can be stored in a server or a database. It can also be spread out across several devices. All these data must be brought together and combined/compiled into a single dataset. The dataset should then be converted to a format that is easy to handle by software (preferably in CSV format). CSV is preferred since it can be easily imported into a Pandas data frame in Python.
2. If personal data, identity information or any such information is considered, sensitive information like phone numbers, zip codes, names, primary keys and other information which might lead to a data biased towards or against a particular group of people must be removed. These can result in data which might train a Machine Learning model to be unconsciously biased.
3. The dataset must now be standardised and cleaned. Numeric values, like currency, distance, time, etc. must be converted into a single unit and a single format. This pre-processing step can be avoided by ensuring restrictions during data entry.
4. String values are generally categorical values. For non-categorical string values, data processing depends on the application and type of data. A frequency count of individual tokens in each string is a good approach. However, such kinds of processing are computationally intensive, and therefore, must be avoided.
5. After the initial pre-processing steps, NaN (not a number) values, empty rows, invalid entries and columns with too many missing values must be removed. The removal of columns is a drastic step and must be avoided at this stage.
6. Categorical data must be one-hot encoded or label encoded, depending on the situation. If the categorical data do not have too many categories, one-hot encoding is a better approach. However, if there are too many categories, label encoding is the way to go.
7. Next, highly correlated features must be removed. This is an essential feature, which often saves a lot of computational resources. There several methods to detect highly correlated features. One such measure is the Pearson correlation coefficient or PCC. It can be expressed by $\rho = \frac{cov(X,Y)}{\sigma X \times \sigma Y}$. X and Y are two random variables. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. If the absolute value of the PCC between two columns is near 1 or is sufficiently high, one of them can be removed, according to the context of the problem.

After these steps, the data can be stored in CSV format for further processing. It can also be kept as a Pandas Data Frame (in Python) for immediate processing or in a database for long term storage.