Preprocessing

Datasets are usually some large files with a huge number of rows and columns. While that is generally true , it is not the case — data could be in so many different forms: Structured Tables, unstructured tables, json, Images, Audio files, Videos etc.

Machines don’t always understand free text, image or video data as is, they only understand binary.

In any Machine Learning pipeline, Data Preprocessing is that step in which the data is transformed, Encoded or augmented to bring it to such a shape or state that now the machine can parse it. More likely the features of the data can now be easily interpreted by the sequence of computational steps or algorithms.

A dataset can be referred to as a collection of *data objects*, which are mostly called as records, points, vectors, events, cases, samples, observations, or entities.

Data objects are explained by a number of *features set* , that observe the basic characteristics of an object. Features are usually called as variables, characteristics, fields, attributes, or dimensions, dims.

A feature is an individual quantifiable property or characteristic of a phenomenon being under observation.

Features can be of two types:

* **Categorical :** Features whose values are taken directly from a defined set of values. For example, days in a week : {Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday} is a class because its value is always taken from this set. Another instance could be the Boolean set : {True, False} or {1 , 0}
* **Numerical :** Features whose values are continuous or numeric valued. They are represented by inteteger and possess most of the properties of numbers. For instance, the number of days in a year.

Following are the steps of Data Preprocessing. One thing to note is, not all the steps are necessary for each problem, it is majorly dependent on the data we are working with but Generally they are :

* Data Quality Assessment
* Feature Aggregation
* Feature Sampling
* Dimensionality Reduction
* Feature Encoding

**Data Quality Assessment**

Data is often taken from different sources which are normally not too reliable and that too in different formats, thus it makes a challenge to consume the data for our machine learning.It is sometimes unrealistic to expect that the data will be perfect. Following are the methods to deal with inconsistencies in the dataset.

1.**Missing values** :

Generally there are missing values in the dataset . It could have happened during data collection, or some validation issue, but regardless missing values must be corrected or taken into consideration.

* Removing rows with missing values :  
  Simple and effective strategy. If a feature has a large number of missing values, then that feature itself can be dropped.
* Removing missing values :  
  If only a considerable amount of values are missing, then we can usually run a normal interpolation method to fill in those values. However, the most common method of dealing with missing values is by filling them in with the statistical analysis generally mean, median or mode value of the feature in question.

2. **Inconsistent values** :

Data can contain inconsistent values. Generally it is some trivial misplacing of values on some other fields. It could be due to human error or maybe the data was misinterpreted while being scanned from a handwritten format.

* It is therefore always a good practice to perform a data assessment like knowing what the data type and getting the knowledge of the shape data.

3. **Duplicate values** :

A dataset could include data objects which are copies of one another. The term deduplication is often used to refer to the process of dealing with copies of objects.

* In most cases, the duplicates are removed so as to not give that particular data object an advantage or *bias*, when running machine learning algorithms.

**Feature Aggregation**

Feature Aggregations is the cumulation of values in order to put the data in a better perspective. These aggregate of values can be collected and created as a objects

* As a result memory is saved.
* Aggregations provides a high-level view of the data as the characteristics of groups or aggregates is more stable than individual data objects itself.

**Feature Sampling**

Sampling is a very general method for selecting a subset of the data that we are to analyze. Usually working with the entire dataset can turn out to be too computationally expensive considering the complexities. Using a sampling predefined algorithm can reduce the size of the dataset to a state where it can be used better.

Principle here is that the sampling should be done in such an order that the sample generated should have approx. the same characteristics of the original dataset, thus being a mathematical *representative* of the dataset. Thus having a good sample size and having a good sample strategy is important.

*Simple Random Sampling* dictates that there is an equal probability distribution of selecting any particular object. Following are the two main variations:

* **Sampling without Replacement** :the item selected is removed from the dataset and sample is created.
* **Sampling with Replacement** : The items are not removed from the original dataset and the sample is created.

An Imbalanced dataset is one in which the number of objects of a class(es) are significantly larger than another class(es), thus leading to an imbalance and creating rarer class(es).

**Dimensionality Reduction**

Most datasets in the world have a large number of features. Lets just consider the problem in context, image processing problem, we usually have to deal with millions of features and parameters generally known as dimensions.thus dimensionality reduction helps reduce the number of dimensions.

Mathematically , *dimensions* refers to the number of geometric planes on a coordinate system in which the dataset lies, which could be high to the point that it cannot be visualized with classical tools. Generally the More the number planes, the more complex dataset is.

**The Curse of Dimensionality**

This phenomenon simply refers that the task of analysing the data gets complicated as the dimensions of the dataset increases .

As the dimensionality increases, the number planes in the representation of data increases thus adding more and more sparsity to the data which is difficult to model and analyze.

Following are the major benefits of dimensionality reduction :

* Data representation and analysis algorithms work better if the dimensions of the dataset is lower.
* If Dimensions are reduced the visualization of the data set becomes easy and less computationally heavy.

**Feature Encoding**

The whole purpose of data preprocessing is to encode the data in such a state that a machine could understand it that is in binary format.

Feature Encoding is a process of encoding the data in such a way that the data remains true to the original meaning and can be easily understandable by a computer or a machine.

Following are the general norms of data encoding:

* **Nominal** : permutation of values like one-hot-encoding is generally encoding the data in one-to-one mapping.
* **Ordinal** : The chunks of data output blocks are created and separated in the output. Such as the concept of small, medium and large. Thus increasing the ordanality of the dataset.

**Train / Validation / Test Split**

After the data is being preprocessed the data needs to be in a specific format so that machine learning algorithms could understand it.

**Training data** : This is a subset of a dataset in which the machine learning model is actually trained on. The dataset comprises the actual labels in case of a supervised learning model. This dataset is on which the training occurs and the model learns through it.

**Validation data** : This is the part of our dataset which is used for the validation of model. It is usually a subset of the training dataset.

**Test data** : This is the subset of the dataset which is used to test the hypothesis.

**Split Ratio** : Data split ratio is a hyperparameter that needs to be tuned for case specific tasks. It justifies the split of the original dataset in training data, testing data and validation data.