Data Preprocessing and Model Validation

Madson L. D. Dias

Huawei / IFCE

March 4, 2021

Agenda

1. Data Preprocessing
   - Missing Data
   - Outliers
   - Duplicate values
   - Feature encoding
   - Feature selection
   - Normalization

Agenda

1. Data Preprocessing
   - Missing Data
   - Outliers
   - Duplicate values
   - Feature encoding
   - Feature selection
   - Normalization

Data Preprocessing

- Data could be in so many different forms: Structured Tables, Images, Audio files, Videos etc..
- In general, machines models don't understand free text, image or video data as it is, they understand arrays of values.

In any Machine Learning process, Data Preprocessing is that step in which the data gets transformed, or Encoded, to bring it to such a state that now the machine can easily parse it. In other words, the features of the data can now be easily interpreted by the algorithm.

Data Preprocessing

- A data set can be viewed as a collection of data objects, which are often also called as a records, points, vectors, patterns, events, cases, samples, observations, or entities.
- Data objects are described by a number of features, that capture the basic characteristics of an object, such as the mass of a physical object or the time at which an event occurred, etc.. Features are often called as variables, characteristics, fields, attributes, or dimensions.

Data Preprocessing

Features can be:

- **Categorical**: Features whose values are taken from a defined set of values. For instance, days in a week: (*Monday*, *Tuesday*, *Wednesday*, *Thursday*, *Friday*, *Saturday*, *Sunday*) is a category because its value is always taken from this set. Another example could be the Boolean set: (*True*, *False*).
- **Numerical**: Features whose values are continuous or integer-valued. They are represented by numbers and possess most of the properties of numbers. For instance, number of steps you walk in a day, or the speed at which you are driving your car.

Data quality

- Because data is often taken from multiple sources which are normally not too reliable and that too in different formats, more than half our time is consumed in dealing with data quality issues when working on a machine learning problem.
- It is simply unrealistic to expect that the data will be perfect. There may be problems due to human error, limitations of measuring devices, or flaws in the data collection process.

Agenda

1. Data Preprocessing
   - Missing Data
   - Outliers
   - Duplicate values
   - Feature encoding
   - Feature selection
   - Normalization

Missing Data

- It is very much usual to have missing values in your data set. It may have happened during data collection, or maybe due to some data validation rule, but regardless missing values must be taken into consideration.
- Eliminate rows with missing data
- Eliminate columns with missing data
- Estimate missing values
  - → mean, median, mode
  - → simple interpolation
  - → classification/regression models

Agenda

1. Data Preprocessing
   - Missing Data
   - Outliers
   - Duplicate values
   - Feature encoding
   - Feature selection
   - Normalization

Outliers

- Extreme values that are outside the range of what is expected and unlike the other data.
- Standard Deviation Method
- Interquartile Range Method
- Automatic Outlier Detection

Agenda

1. Data Preprocessing
   - Missing Data
   - Outliers
   - Duplicate values
   - Feature encoding
   - Feature selection
   - Normalization

Duplicate values

- A dataset may include data objects which are duplicates of one another. It may happen when say the same person submits a form more than once.
- 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.

Agenda

1. Data Preprocessing
   - Missing Data
   - Outliers
   - Duplicate values
   - Feature encoding
   - Feature selection
   - Normalization

Feature encoding

- Feature encoding is basically performing transformations on the data such that it can be easily accepted as input for machine learning algorithms while still retaining its original meaning.
- **Nominal**: Any one-to-one mapping can be done which retains the meaning. For instance, a permutation of values like in One-Hot Encoding.
- **Ordinal**: An order-preserving change of values. The notion of small, medium and large can be represented equally well with the help of a new function, that is, <new_value = f(old_value)> - For example, 0, 1, 2 or maybe 1, 2, 3.

Agenda

1. Data Preprocessing
   - Missing Data
   - Outliers
   - Duplicate values
   - Feature encoding
   - Feature selection
   - Normalization

Feature selection

- Select a subset of input features from the data set.
  - **Unsupervised**: Do not use the target variable (e.g. remove redundant variables).
  - **Supervised**: Use the target variable (e.g. remove irrelevant variables).
    - → Wrapper: Search for well-performing subsets of features.
    - → Filter: Select subsets of features based on their relationship with the target.
    - → Intrinsic: Algorithms that perform automatic feature selection during training.
  - **Dimensionality Reduction**: Project input data into a lower-dimensional feature space.

Agenda

1. Data Preprocessing
   - Missing Data
   - Outliers
   - Duplicate values
   - Feature encoding
   - Feature selection
   - Normalization

Normalization

- Normalization of ratings means adjusting values measured on different scales to a notionally common scale, often prior to averaging.
  - → z-score

$$x' = \frac{x - \mu}{\sigma} \tag{1}$$

  - → Min-Max

$$x' = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \tag{2}$$