UDC 004.8

DATA PREPROCESSING FOR MACHINE LEARNING

MIKHAIL SHAUTSOU, YURI PASTUHOF
Polotsk State University, Belarus

Pre-processing and cleaning data are important tasks that typically must be conducted before dataset can be used effectively for machine learning. Raw data is often noisy and unreliable, and may be missing values. Using such data for modeling can produce misleading results. These tasks are part of the Team Data Science Process and typically follow an initial exploration of a dataset used to discover and plan the pre-processing required.

Reason to Data Pre-Processing

Real world data is gathered from various sources and processes and it may contain irregularities or corrupt data compromising the quality of the dataset. The typical data quality issues that arise are:

- **Incomplete:** data lacks attributes or containing missing values.
- **Noisy**: Data contains erroneous records or outliers.
- **Inconsistent**: Data contains conflicting records or discrepancies.

Quality data is a prerequisite for quality predictive models. To avoid "garbage in, garbage out" and improve data quality and therefore model performance, it is imperative to conduct a data health screen to spot data issues early and decide on the corresponding data processing and cleaning steps.

Standard Monitoring Methods

We can check the general quality of data by checking:

- The number of records.
- The number of attributes (or features).
- The attribute data types (nominal, ordinal, or continuous).
- The number of missing values.
- Well-formedness of the data.

When you find issues with data, **processing steps** are necessary which often involves cleaning missing values, data normalization, discretization, text processing to remove and/or replace embedded characters which may affect data alignment, mixed data types in common fields, and others.

Major Tasks In Data Pre-Processing

1. **Data cleaning** – fill in or missing values, detect and remove noisy data and outliers.

To deal with missing values, it is best to first identify the reason for the missing values to better handle the problem. Typical missing value handling methods are:

- Deletion: remove records with missing values
- Dummy substitution: replace missing values with a dummy value: e.g, *unknown* for categorical or 0 for numerical values.
- Mean substitution: If the missing data is numerical, replace the missing values with the mean.
- Frequent substitution: If the missing data is categorical, replace the missing values with the most frequent item
- Regression substitution: use a regression method to replace missing values with regressed values.

2. **Data transformation** – normalize data to reduce dimensions and noise.

Data normalization re-scales numerical values to a specified range. Popular data normalization methods include:

- Min-Max Normalization: linearly transform the data to a range, say between 0 and 1, where the min value is scaled to 0 and max value to 1.
- Z-score Normalization: scale data based on mean and standard deviation: divide the difference between the data and the mean by the standard deviation.
- Decimal scaling: scale the data by moving the decimal point of the attribute value.

3. **Data reduction –** sample data records or attributes for easier data handling.

There are various methods to reduce data size for easier data handling. Depending on data size and the domain, the following methods can be applied:

- Record Sampling: sample the data records and only choose the representative subset from the data.

- Attribute Sampling: select only a subset of the most important attributes from the data.
- Aggregation: divide the data into groups and store the numbers for each group.

4. **Data discretization** — convert continuous attributes to categorical attributes for ease of use with certain machine learning methods.

Data can be discretized by converting continuous values to nominal attributes or intervals. Some ways of doing this are:

- Equal-Width Binning: divide the range of all possible values of an attribute into N groups of the same size, and assign the values that fall in a bin with the bin number.
- Equal-Height Binning: divide the range of all possible values of an attribute into N groups, each containing the same number of instances, then assign the values that fall in a bin with the bin number.

5. **Text cleaning** — remove embedded characters which may cause data misalignment, for e.g., embedded tabs in a tab-separated data file, embedded new lines which may break records, etc.

Text fields in tabular data may include characters which affect columns alignment and/or record boundaries. For e.g., embedded tabs in a tab-separated file cause column misalignment, and embedded new line characters break record lines. Improper text encoding handling while writing/reading text leads to information loss, inadvertent introduction of unreadable characters, e.g., nulls, and may also affect text parsing. Careful parsing and editing may be required in order to clean text fields for proper alignment and/or to extract structured data from unstructured or semi-structured text data.

REFERENCES

1. Задачи по подготовке данных для расширенного машинного обучения [Electronic resource]. – Mode of access: https://docs.microsoft.com/ru-ru/azure/machine-learning/team-data-science-process/prepare-data.– Date of access: 10.02.2018.
2. Data Mining: Concepts and Techniques / Kaufmann M. [et al.]. – Third Edition. – 2011.