2021 Call for Code Awards: Live from New York, with SNL's Colin Jost!

Learn more

Article

Data preprocessing in detail

Learn how to create better models and predictions using data preprocessing

By Sana Mushtaq

Published June 14, 2019

The probability of anomalous data has increased in today's data due to its large size and its origin for heterogenous sources. Considering the fact that high-quality data leads to better models and predictions, data preprocessing has become vital, and the fundamental step in the data science/machine learning/AI pipeline. In this article, learn about the need to process data and discuss different approaches to each step in the process.

While gathering data, you might come across three main factors that would contribute to the quality of data:

1. **Accuracy:** Erroneous values that deviate from the expected. The causes for inaccurate data can vary, but include:

   - Human/computer errors during data entry and transmission

- Users deliberately submitting incorrect values (called disguised missing data)
- Incorrect formats for input fields
- Duplication of training examples

2. **Completeness**: Lacking attribute/feature values or values of interest. The data set might be incomplete due to:

- Unavailability of data
- Deletion of inconsistent data
- Deletion of data deemed irrelevant initially

3. **Consistency**: Aggregation of data is inconsistent.

Some other features that also affect the data quality include timeliness (the data is incomplete until all relevant information is submitted after certain time periods), believability (how much the data is trusted by the user) and interpretability (how easily the data is understood by all stakeholders).

To ensure high-quality data, it's crucial to preprocess it. To make the process easier, data preprocessing is divided into four stages: data cleaning, data integration, data reduction, and data transformation.

Data cleaning

Data cleaning refers to techniques to 'clean' data by removing outliers, replacing missing values, smoothing noisy data, and correcting inconsistent data. Many techniques are used to perform each of these tasks, where each technique is specific to a user's preference or problem set. Below, I explain each task in terms of the techniques used to overcome it.

Missing values

You can use multiple approaches to deal with missing data. Let's look at some of them.

1. **Removing the training example**: You can ignore the training example if the output label is missing (if it is a classification problem). This is usually discouraged as it leads

Site feedback

to loss of data because you are removing the attribute values that can add value to data set as well.

2. **Filling in missing value manually**: This approach is time consuming, and is not recommended for large data sets.
3. **Using a standard value to replace the missing value**: The missing value can be replaced by a global constant such as ‘N/A’ or ‘Unknown.’ This is a simple approach, but not foolproof.
4. **Using central tendency (mean, median, mode) for attribute to replace the missing value**: Based on data distribution, mean (in the case of normal distribution) or median (for non-normal distribution) can be used to fill in for the missing value.
5. **Using central tendency (mean, median, mode) for attribute belonging to same class to replace the missing value**: This is the same as method 4, except that the measures of central tendency are specific to each class.
6. **Using the most probable value to fill in the missing value**: Using algorithms like regression and decision trees, the missing values can be predicted and replaced.

Site feedback

Noisy data

Noise is defined as a random variance in a measured variable. For numeric values, box plots and scatter plots can be used to identify outliers. To deal with these anomalous values, data smoothing techniques are applied, which are described below.

1. **Binning:** Using binning methods smooths sorted values by using the values around it. The sorted values are then divided into bins. There are various approaches to binning. Two of them are smoothing by bin means, where each bin is replaced by the mean of the bin’s values, and smoothing by bin medians, where each bin is replaced by the median of the bin’s values.
2. **Regression:** Linear regression and multiple linear regression can be used to smooth the data, where the values are conformed to a function.
3. **Outlier analysis:** Approaches such as clustering can be used to detect outliers and deal with them.

Data integration

Because data is being collected from multiple sources, data integration has become a vital part of the process. This might lead to redundant and inconsistent data, which could result in poor accuracy and speed of a data model. To deal with these issues and

maintain the data integrity, approaches such as tuple duplication detection and data conflict detection are sought after. The most common approaches to integrate data are explained below.

1. **Data consolidation:** The data is physically brought together to one data store. This usually involves Data Warehousing.
2. **Data propagation:** Copying data from one location to another using applications is called data propagation. Data propagation can be synchronous or asynchronous and is event-driven.
3. **Data virtualization:** An interface is used to provide a real-time and unified view of data from multiple sources. The data can be viewed from a single point of access.

Site feedback

Data reduction

The purpose of data reduction is to have a condensed representation of the data set that is smaller in volume, while maintaining the integrity of the original data set. This results in efficient, yet similar, results. A few methods to reduce the volume of data are:

1. **Missing values ratio**: Attributes that have more missing values than a threshold are removed.
2. **Low variance filter**: Normalized attributes that have variance (distribution) less than a threshold are also removed because little changes in data means less information.
3. **High correlation filter**: Normalized attributes that have correlation coefficients more than a threshold are also removed because similar trends means similar information is carried. A correlation coefficient is usually calculated using statistical methods such as Pearson's chi-square value.
4. **Principal component analysis**: Principal component analysis, or PCA, is a statistical method that reduces the numbers of attributes by lumping highly correlated attributes together. With each iteration, the initial features are reduced to principal components, with greater variance than the original set on the condition that they are uncorrelated with the preceding components. However, this method only works for features with numerical values.

Data transformation

The final step of data preprocessing is transforming the data into a form appropriate for data modeling. Strategies that enable data transformation include:

1. **Smoothing**: Eliminating noise in the data to see more data patterns.
2. **Attribute/feature construction**: New attributes are constructed from the given set of attributes.
3. **Aggregation**: Summary and aggregation operations are applied on the given set of attributes to come up with new attributes.
4. **Normalization**: The data in each attribute is scaled between a smaller range, for example, 0 to 1 or -1 to 1.
5. **Discretization**: Raw values of the numeric attributes are replaced by discrete or conceptual intervals, which can be further organized into higher-level intervals.
6. **Concept hierarchy generation for nominal data**: Values for nominal data are generalized to higher-order concepts.

Site feedback

Summary

Despite having multiple approaches to preprocessing data, it's still an actively researched field due to the amount of incoherent data being generated each day. IBM Cloud provides a platform for data scientists called IBM Watson Studio, which includes services that enable data scientists to preprocess data using drag-and-drop services in addition to the conventional method of programming. Learn how Watson Studio can help with the data science lifecycle.

This content also appears in:

Legend

Categories

Data science Machine learning Watson Studio

Table of Contents