Teach

DATA

Created at	March 12, 2024, 20:40
Last updated	March 13, 2024, 14:06

DATA

Data is a crucial component in the field of Machine Learning. It refers to the set of observations or measurements that can be used to train a machine-learning model. The quality and quantity of data available for training and testing play a significant role in determining the performance of a machine-learning model. Data can be in various forms such as numerical, categorical, or time-series data, and can come from various sources such as databases, spreadsheets, or APIs.

Machine learning algorithms use data to learn patterns and relationships between input variables and target outputs, which can then be used for prediction or classification tasks.

Data is typically divided into two types:

1. Labeled data
2. Unlabeled data

Labeled data includes a label or target variable that the model is trying to predict, whereas unlabeled data does not include a label or target variable. The data used in machine learning is typically numerical or categorical.

Numerical data includes values that can be ordered and measured, such as age or income.

Categorical data includes values that represent categories, such as gender or type of fruit.

Data can be divided into training and testing sets. The training set is used to train the model, and the testing set is used to evaluate the performance of the model. It is important to ensure that the data is split in a random and representative way.

Data preprocessing is an important step in the machine learning pipeline. This step can include cleaning and normalizing the data, handling missing values, and feature selection or engineering.

DATA, INFORMATION, KNOWLEDGE

DATA: It can be any unprocessed fact, value, text, sound, or picture that is not being interpreted and analyzed. Data is the most important part of all Data Analytics, Machine Learning, and Artificial Intelligence. Without data, we can't train any model and all modern research and automation will go in vain. Big Enterprises are spending lots of money just to gather as much certain data as possible.

Example: Why did Facebook acquire WhatsApp by paying a huge price of $19 billion?

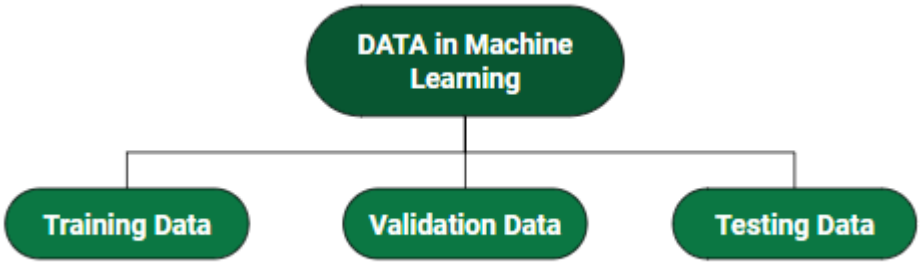
The answer is very simple and logical – it is to have access to the users' information that Facebook may not have but WhatsApp will have. This information about their users is of paramount importance to Facebook as it will facilitate the task of improvement in their services.

INFORMATION: Data that has been interpreted and manipulated and has now some meaningful inference for the users.

KNOWLEDGE: Combination of inferred information, experiences, learning, and insights. Results in awareness or concept building for an individual or organization.

How do we split data in Machine Learning?

- **Training Data:** The part of data we use to train our model. This is the data that your model actually sees(both input and output) and learns from.
- **Validation Data:** The part of data that is used to do a frequent evaluation of the model, fit on the training dataset along with improving involved hyperparameters (initially set parameters before the model begins learning). This data plays its part when the model is actually training.
- **Testing Data:** Once our model is completely trained, testing data provides an unbiased evaluation. When we feed in the inputs of Testing data, our model will predict some values(without seeing actual output). After prediction, we evaluate our model by comparing it with the actual output present in the testing data. This is how we evaluate and see how much our model has learned from the experiences feed in as training data, set at the time of training.



Different Forms of Data

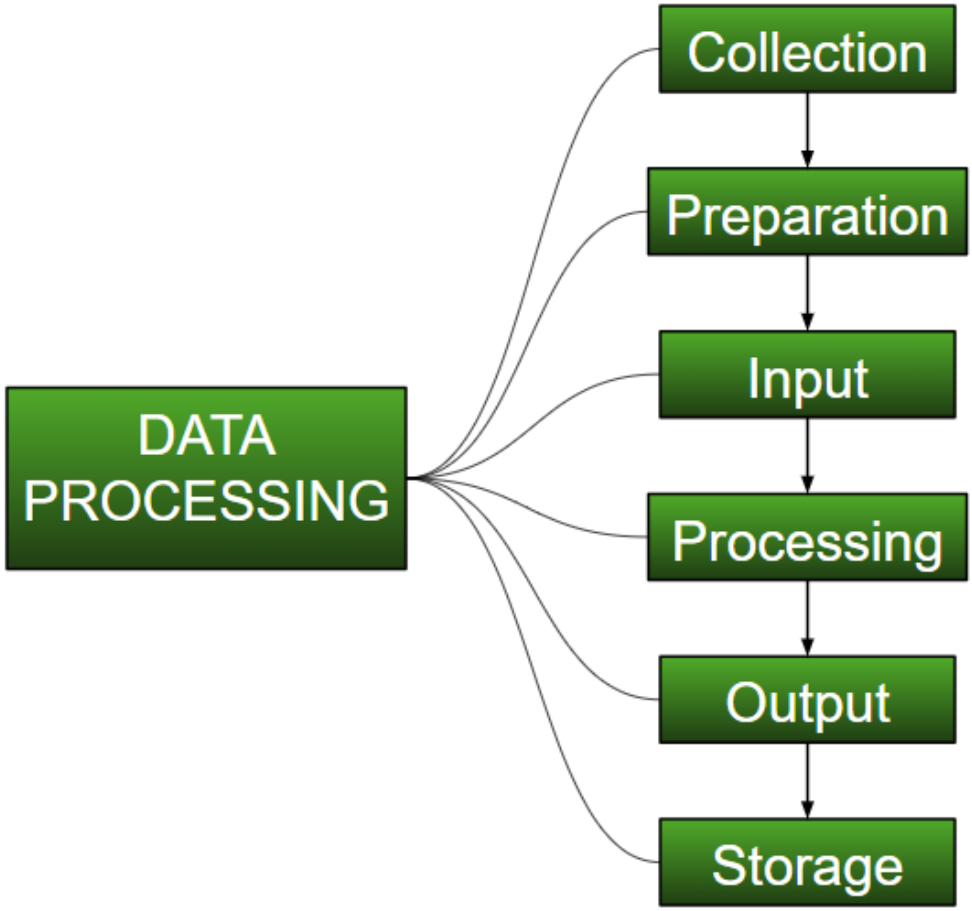
- **Numeric Data** : If a feature represents a characteristic measured in numbers , it is called a numeric feature.
- **Categorical Data :** A categorical feature is an attribute that can take on one of the limited , and usually fixed number of possible values on the basis of some qualitative property . A categorical feature is also called a nominal feature.

- **Ordinal Data** : This denotes a nominal variable with categories falling in an ordered list . Examples include clothing sizes such as small, medium , and large , or a measurement of customer satisfaction on a scale from “not at all happy” to “very happy”.

Understanding Data Processing

Data Processing is the task of converting data from a given form to a much more usable and desired form i.e. making it more meaningful and informative. Using Machine Learning algorithms, mathematical modeling, and statistical knowledge, this entire process can be automated. The output of this complete process can be in any desired form like graphs, videos, charts, tables, images, and many more, depending on the task we are performing and the requirements of the machine. This might seem to be simple but when it comes to massive organizations like Twitter, Facebook, Administrative bodies like Parliament, UNESCO, and health sector organizations, this entire process needs to be performed in a very structured manner. So, the steps to perform are as follows:

Data processing is a crucial step in the machine learning (ML) pipeline, as it prepares the data for use in building and training ML models. The goal of data processing is to clean, transform, and prepare the data in a format that is suitable for modeling.



1. **Collection :**

The most crucial step when starting with ML is to have data of good quality and accuracy. Data can be collected from any authenticated source like data.gov.in, Kaggle or UCI dataset repository. For example, while preparing for a competitive exam, students study from the best study material that they can access so that they learn the best to obtain the best results. In the same way, high-quality and accurate data will make the learning process of the model easier and better and at the time of testing, the model would yield state-of-the-art results.

A huge amount of capital, time and resources are consumed in collecting data. Organizations or researchers have to decide what kind of data they need to execute their tasks or research.

Example: Working on the Facial Expression Recognizer, needs numerous images having a variety of human expressions. Good data ensures that the results of the model are valid and can be trusted upon.

2. **Preparation :**

   The collected data can be in a raw form which can't be directly fed to the machine. So, this is a process of collecting datasets from different sources, analyzing these datasets and then constructing a new dataset for further processing and exploration. This preparation can be performed either manually or from the automatic approach. Data can also be prepared in numeric forms also which would fasten the model's learning.

   Example: An image can be converted to a matrix of N X N dimensions, the value of each cell will indicate the image pixel.

3. **Input :**

   Now the prepared data can be in the form that may not be machine-readable, so to convert this data to the readable form, some conversion algorithms are needed. For this task to be executed, high computation and accuracy is needed.

   Example: Data can be collected through the sources like MNIST Digit data(images), Twitter comments, audio files, video clips.

4. **Processing :**

   This is the stage where algorithms and ML techniques are required to perform the instructions provided over a large volume of data with accuracy and optimal computation.

5. **Output :**

   In this stage, results are procured by the machine in a meaningful manner which can be inferred easily by the user. Output can be in the form of reports, graphs, videos, etc

6. **Storage :**

   This is the final step in which the obtained output and the data model data and all the useful information are saved for future use.

Advantages of data processing in Machine Learning:

1. Improved model performance: Data processing helps improve the performance of the ML model by cleaning and transforming the data into a format that is suitable for modeling.
2. Better representation of the data: Data processing allows the data to be transformed into a format that better represents the underlying relationships and patterns in the data, making it easier for the ML model to learn from the data.

3. Increased accuracy: Data processing helps ensure that the data is accurate, consistent, and free of errors, which can help improve the accuracy of the ML model.

Disadvantages of data processing in Machine Learning:

1. Time-consuming: Data processing can be a time-consuming task, especially for large and complex datasets.
2. Error-prone: Data processing can be error-prone, as it involves transforming and cleaning the data, which can result in the loss of important information or the introduction of new errors.
3. Limited understanding of the data: Data processing can lead to a limited understanding of the data, as the transformed data may not be representative of the underlying relationships and patterns in the data.

There are many tools and libraries available for data processing in ML, including pandas for Python, and the Data Transformation and Cleansing tool in RapidMiner. The choice of tools will depend on the specific requirements of the project, including the size and complexity of the data and the desired outcome.

Read more about the source.