ML | Introduction to Data in Machine Learning

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: 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?

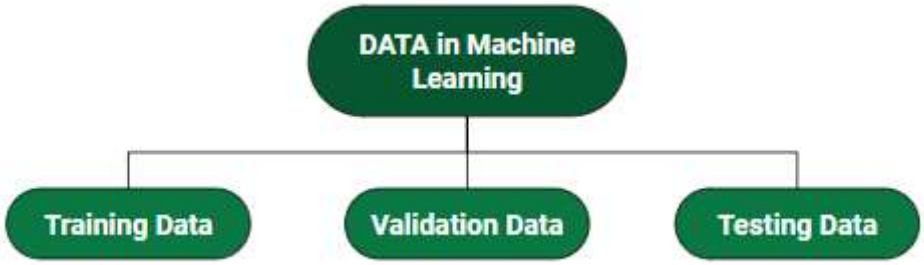
The answer is very simple and logical – it is to have access to the users' information that Facebook may not have but WhatsApp will have. This information about their users is of paramount importance to Facebook as it will facilitate the task of improvement in their services.

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.



Consider an example:

There's a Shopping Mart Owner who conducted a survey for which he has a long list of questions and answers that he had asked from the customers, this list of questions and answers is **DATA**. Now every time when he wants to infer anything and can't just go through each and every question of thousands of customers to find something relevant as it would be time-consuming and not helpful. In order to reduce this overhead and time wastage and to make work easier, data is manipulated through software, calculations, graphs, etc. as per your own convenience, this inference from manipulated data is **Information**. So, Data is a must for Information. Now **Knowledge** has its role in differentiating between two individuals having the same information. Knowledge is actually not technical content but is linked to the human thought process.

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”.

Properties of Data –

1. **Volume:** Scale of Data. With the growing world population and technology at exposure, huge data is being generated each and every millisecond.
2. **Variety:** Different forms of data – healthcare, images, videos, audio clippings.
3. **Velocity:** Rate of data streaming and generation.
4. **Value:** Meaningfulness of data in terms of information that researchers can infer from it.
5. **Veracity:** Certainty and correctness in data we are working on.
6. **Viability**: The ability of data to be used and integrated into different systems and processes.
7. **Security**: The measures taken to protect data from unauthorized access or manipulation.
8. **Accessibility**: The ease of obtaining and utilizing data for decision-making purposes.
9. **Integrity**: The accuracy and completeness of data over its entire lifecycle.
10. **Usability**: The ease of use and interpretability of data for end-users.

Some facts about Data:

- As compared to 2005, 300 times i.e. 40 Zettabytes (1ZB=10^21 bytes) of data will be generated by 2020.
- By 2011, the healthcare sector has a data of 161 Billion Gigabytes
- 400 Million tweets are sent by about 200 million active users per day
- Each month, more than 4 billion hours of video streaming is done by the users.
- 30 Billion different types of content are shared every month by the user.
- It is reported that about 27% of data is inaccurate and so 1 in 3 business idealists or leaders don’t trust the information on which they are making decisions.

The above-mentioned facts are just a glimpse of the actually existing huge data statistics. When we talk in terms of real-world scenarios, the size of data currently presents and is getting generated each and every moment is beyond our mental horizons to imagine.

Example:

Imagine you're working for a car manufacturing company and you want to build a model that can predict the fuel efficiency of a car based on the weight and the engine size. In this case, the target variable (or label) is the fuel efficiency, and the features (or input variables) are the weight and engine size. You will collect data from different car models, with corresponding weight and engine size, and their fuel efficiency. This data is labeled and it's in the form of (weight,engine size,fuel efficiency) for each car. After having your data ready, you will then split it into two sets: training set and testing set, the training set will be used to train the model and the testing set will be used to evaluate the performance of the model. Preprocessing could be needed for example, to fill missing values or handle outliers that might affect your model accuracy.

Implementation:

Example: 1

Python3

```
from  sklearn.linear_model  import  LogisticRegression

X  =  [[``1``,  2``], [``2``,  3``], [``3``,  4``], [``4``,  5``], [``5``,  6``]]

y  =  [``0``,  0``,  1``,  1``,  1``]

model  =  LogisticRegression()

model.fit(X, y)

prediction  =  model.predict([[``6``,  7``]])[``0``]

print``(prediction)
```

Output:

```
0,1
```

If you run the code I provided, the output will be the prediction made by the model. In this case, the prediction will be either 0 or 1, depending on the specific parameters learned by the model during training.

For example, if the model learned that input data with a high second element is more likely to have a label of 1, then the prediction for [6, 7] would be 1.

Advantages Or Disadvantages:

Advantages of using data in Machine Learning:

1. Improved accuracy: With large amounts of data, machine learning algorithms can learn more complex relationships between inputs and outputs, leading to improved accuracy in predictions and classifications.
2. Automation: Machine learning models can automate decision-making processes and can perform repetitive tasks more efficiently and accurately than humans.
3. Personalization: With the use of data, machine learning algorithms can personalize experiences for individual users, leading to increased user satisfaction.
4. Cost savings: Automation through machine learning can result in cost savings for businesses by reducing the need for manual labor and increasing efficiency.

Disadvantages of using data in Machine Learning:

1. Bias: Data used for training machine learning models can be biased, leading to biased predictions and classifications.
2. Privacy: Collection and storage of data for machine learning can raise privacy concerns and can lead to security risks if the data is not properly secured.
3. Quality of data: The quality of data used for training machine learning models is critical to the performance of the model. Poor quality data can lead to inaccurate predictions and classifications.
4. Lack of interpretability: Some machine learning models can be complex and difficult to interpret, making it challenging to understand how they are making decisions.

**Use of Machine Learning 😗*

Machine learning is a powerful tool that can be used in a wide range of applications. Here are some of the most common uses of machine learning:

- **Predictive modeling:** Machine learning can be used to build predictive models that can predict future outcomes based on historical data. This can be used in many applications, such as stock market prediction, fraud detection, weather forecasting, and customer behavior prediction.
- **Image recognition:** Machine learning can be used to train models that can recognize objects, faces, and other patterns in images. This is used in many applications, such as self-driving cars, facial recognition systems, and medical image analysis.
- **Natural language processing:** Machine learning can be used to analyze and understand natural language, which is used in many applications, such as chatbots, voice assistants, and sentiment analysis.

- **Recommendation systems:** Machine learning can be used to build recommendation systems that can suggest products, services, or content to users based on their past behavior or preferences.
- **Data analysis:** Machine learning can be used to analyze large datasets and identify patterns and insights that would be difficult or impossible for humans to detect.
- **Robotics:** Machine learning can be used to train robots to perform tasks autonomously, such as navigating through a space or manipulating objects.

Issues of using data in Machine Learning:

- **Data quality:** One of the biggest issues with using data in machine learning is ensuring that the data is accurate, complete, and representative of the problem domain. Low-quality data can result in inaccurate or biased models.
- **Data quantity:** In some cases, there may not be enough data available to train an accurate machine learning model. This is especially true for complex problems that require a large amount of data to accurately capture all the relevant patterns and relationships.
- **Bias and fairness:** Machine learning models can sometimes perpetuate bias and discrimination if the training data is biased or unrepresentative. This can lead to unfair outcomes for certain groups of people, such as minorities or women.
- **Overfitting and underfitting:** Overfitting occurs when a model is too complex and fits the training data too closely, resulting in poor generalization to new data. Underfitting occurs when a model is too simple and does not capture all the relevant patterns in the data.
- **Privacy and security:** Machine learning models can sometimes be used to infer sensitive information about individuals or organizations, raising concerns about privacy and security.
- **Interpretability:** Some machine learning models, such as deep neural networks, can be difficult to interpret and understand, making it challenging to explain the reasoning behind their predictions and decisions.