

DATA SCIENCE & MACHINE LEARNING (part - 1)

techworldthink • February 02, 2022

1. What is data science?

Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data. Data science practitioners apply machine learning algorithms to numbers, text, images, video, audio, and more to produce artificial intelligence (AI) systems to perform tasks that ordinarily require human intelligence. In turn, these systems generate insights which analysts and business users can translate into tangible business value.

Data Science can be defined as the study of data, where it comes from, what it represents, and the ways by which it can be transformed into valuable inputs and resources to create business and IT strategies.

Data science is a deep study of the massive amount of data, which involves extracting meaningful insights from raw, structured, and unstructured data that is processed using the scientific method, different technologies, and algorithms.

It is a multidisciplinary field that uses tools and techniques to manipulate the data so that you can find something new and meaningful.

Data science uses the most powerful hardware, programming systems, and most efficient algorithms to solve the data related problems. It is the future of artificial intelligence.

Data Science is about data gathering, analysis and decision-making.

Data Science is about finding patterns in data, through analysis, and make future predictions.

By using Data Science, companies are able to make:

- Better decisions (should we choose A or B)
- Predictive analysis (what will happen next?)
- Pattern discoveries (find pattern, or maybe hidden information in the data)

Data Science is used in many industries in the world today, e.g. banking, consultancy, healthcare, and manufacturing.

A Data Scientist requires expertise in several backgrounds:

- Machine Learning
- Statistics
- Programming (Python or R)
- Mathematics
- Databases

2. Explain the different types of data.

Almost anything can be turned into DATA. Building a deep understanding of the different data types is a crucial prerequisite for doing Exploratory Data Analysis (EDA) and Feature Engineering for Machine Learning models. You also need to convert data types of some variables in order to make appropriate choices for visual encodings in data visualization and storytelling.

Most data can be categorized into 4 basic types from a Machine Learning perspective:

numerical data

categorical data

time-series data

text data

1.Numerical Data

Numerical data is any data where data points are exact numbers. Statisticians also might call numerical data, quantitative data. This data has meaning as a measurement such as house prices or as a count, such as a number of residential properties in Los Angeles or how many houses sold in the past year.

Numerical data can be characterized by continuous or discrete data. Continuous data can assume any value within a range whereas discrete data has distinct values.

2.Categorical Data

Categorical data represents characteristics, such as a hockey player's position, team, hometown. Categorical data can take numerical values. For example, maybe we would use 1 for the colour red and 2 for blue. But these numbers don't have a mathematical meaning. That is, we can't add them together or take the average.

In the context of super classification, categorical data would be the class label. This would also be something like if a person is a man or woman, or property is residential or commercial.

3.Time Series Data

Time series data is a sequence of numbers collected at regular intervals over some period of time. It is very important, especially in particular fields like finance. Time series data has a temporal value attached to it, so this would be something like a date or a timestamp that you can look for trends in time.

4.Text

Text data is basically just words. A lot of the time the first thing that you do with text is you turn it into numbers using some interesting functions like the bag of words formulation.

3. Differentiate between supervised and unsupervised learning algorithms.

1.unsupervised learning

As the name suggests, unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Instead, models itself find the hidden patterns and insights from the given data. It can be compared to learning which takes place in the human brain while learning new things. It can be defined as:

Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.

Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format.

Why use Unsupervised Learning?

Below are some main reasons which describe the importance of Unsupervised Learning:

- Unsupervised learning is helpful for finding useful insights from the data.
- Unsupervised learning is much similar as a human learns to think by their own experiences, which makes it closer to the real AI.
- Unsupervised learning works on unlabeled and uncategorized data which make unsupervised learning more important.
- In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

Types of Unsupervised Learning Algorithm:

- **Clustering:** Clustering is a method of grouping the objects into clusters such that objects with most similarities remain in a group and have less or no similarities with the objects of another group. Cluster analysis finds the commonalities between the data objects and categorizes them as per the presence and absence of those commonalities.
- **Association:** An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database. It determines the set of items that occurs together in the dataset. Association rule makes marketing strategy more effective. Such as people who buy X item (suppose a bread) also tend to purchase Y (Butter/Jam) item. A typical example of Association rule is Market Basket Analysis.

Unsupervised Learning algorithms:

Below is the list of some popular unsupervised learning algorithms:

- K-means clustering
- Hierarchical clustering
- Anomaly detection
- Neural Networks
- Principle Component Analysis
- Independent Component Analysis
- Apriori algorithm
- Singular value decomposition

Advantages of Unsupervised Learning

- Unsupervised learning is used for more complex tasks as compared to supervised learning because, in unsupervised learning, we don't have labeled input data.

- Unsupervised learning is preferable as it is easy to get unlabeled data in comparison to labeled data.

Disadvantages of Unsupervised Learning

- Unsupervised learning is intrinsically more difficult than supervised learning as it does not have corresponding output.
- The result of the unsupervised learning algorithm might be less accurate as input data is not labeled, and algorithms do not know the exact output in advance.

2.Supervised Machine Learning

Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.

Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).

In the real-world, supervised learning can be used for Risk Assessment, Image classification, Fraud Detection, spam filtering, etc.

How Supervised Learning Works?

In supervised learning, models are trained using labelled dataset, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output.

Steps Involved in Supervised Learning:

- First Determine the type of training dataset
- Collect/Gather the labelled training data.
- Split the training dataset into training dataset, test dataset, and validation dataset.
- Determine the input features of the training dataset, which should have enough knowledge so that the model can accurately predict the output.
- Determine the suitable algorithm for the model, such as support vector machine, decision tree, etc.
- Execute the algorithm on the training dataset. Sometimes we need validation sets as the control parameters, which are the subset of training datasets.
- Evaluate the accuracy of the model by providing the test set. If the model predicts the correct output, which means our model is accurate.

Types of supervised Machine learning Algorithms:

Supervised learning can be further divided into two types of problems:

1. Regression

Regression algorithms are used if there is a relationship between the input variable and the output variable. It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc. Below are some popular Regression algorithms which come under supervised learning:

- Linear Regression
- Regression Trees
- Non-Linear Regression
- Bayesian Linear Regression
- Polynomial Regression

2. Classification

Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.

Spam Filtering,

- Random Forest
- Decision Trees
- Logistic Regression
- Support vector Machines

Advantages of Supervised learning:

- With the help of supervised learning, the model can predict the output on the basis of prior experiences.
- In supervised learning, we can have an exact idea about the classes of objects.
- Supervised learning model helps us to solve various real-world problems such as fraud detection, spam filtering, etc.

Disadvantages of supervised learning:

- Supervised learning models are not suitable for handling the complex tasks.
- Supervised learning cannot predict the correct output if the test data is different from the training dataset.
- Training required lots of computation times.
- In supervised learning, we need enough knowledge about the classes of object.

supervised machine learning algorithms

1. Decision Trees
2. Naive Bayes Classification
- 3.KNN

