INTRODUCTION TO MACHINE LEARNING

What is Machine Learning?

Arthur Samuel, a pioneer in AI research, popularized the term Machine Learning in 1959 by defining it as "the field of study that gives computers the ability to learn without being explicitly programmed." This is a vague definition of Machine Learning.

Tom Mitchell gives a modern formal definition. He says, "A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in T, as measured by P, improves with experience E."

Let's look at an example to understand this better. Suppose you want your computer to perform a task. Given an image, tell if it is an image of a cat or a dog. To do this, you take 1000 images of both cats and dogs and show it to your computer one by one, telling it if it is a cat or a dog. After you have done showing all thousand images, you show 100 new images (which the computer has not seen before) to the computer and see how many images the computer correctly identifies as a cat or a dog, and call it the machine's performance. If it performs poorly, you show 1000 more images to the computer and then test it again with new images. If the performance improves, then the machine (computer) is said to have learnt. Here,

E: The experience of looking at 1000 images of cats and dogs.
T: Identifying an image as a cat or a dog.
P: Number of images out of 100 new images that the computer correctly identifies.

After reading the example and defining these terms, correlate it with Tom Mitchell's definition. This should have made the definition clearer.

In regular programming, we have an input and a logic, which we use to program the computer to give an output. However, in Machine Learning, we have an input and a set of outputs (most of the time). We provide the computer the input and output, expecting it to figure out the logic of obtaining the output from the input. This logic is generally called a model.

Types of Machine Learning

Machine Learning can be classified into the following types.

1. **Supervised Learning**
2. **Unsupervised Learning**
3. **Reinforcement Learning**

Supervised Learning

Supervised Learning refers to the type of Machine Learning, where the model is trained on labelled data. In other words, you have a dataset having both input, and output corresponding to the given input. You train your model using this data and expect it to predict the correct output given a new input. This can be further classified into two types

Regression

It is a supervised learning task where the output has continuous values.

Examples:

- Predicting the price of a house based on the size of the house, the number of bedrooms, etc.
- Predicting the salary of a software engineer given other details like years of experience, qualification, company name, etc.

Classification

It is a supervised learning task where the output has discrete values.

Examples:

- Differentiating between the images of a cat and a dog (binary classification).
- Iris flower data: Given features like petal width, petal length, etc. classifying an iris flower into species setosa, versicolor, or virginica (multiclass classification).

Some supervised learning algorithms

- ➔ Linear Regression
- ➔ Logistic Regression
- ➔ Neural Networks
- ➔ K Nearest Neighbours (KNNs)
- ➔ Decision Trees
- ➔ Naive Bayes Classifier
- ➔ Support Vector Machines (SVMs)

Unsupervised Learning

Unsupervised learning refers to the type of Machine Learning where the model is trained on unlabelled data and a minimum of human supervision. In other words, you have a dataset with only inputs. You train your model using this dataset and expect it to find patterns or anomalies in the data.

Examples:

- Place a set of books into different groups based on similarity. (Clustering)
- Discover unusual points in the dataset. (Anomaly detection)

- Discover rules that apply to a large set of data, like people who tend to buy A also tend to buy B. (Association)

Some unsupervised learning algorithms

- ➔ K Means Clustering
- ➔ DBSCAN
- ➔ Autoencoders
- ➔ Principal Component Analysis (PCA)
- ➔ tSNE

Reinforcement Learning

Reinforcement learning is the training of machine learning models to make a sequence of decisions by interacting with the environment. A reinforcement algorithm (or agent) works on a reward system. The agent receives rewards for performing correctly and penalties for performing incorrectly. The agent learns without human intervention by minimizing penalties and maximizing rewards.

Examples:

- Learning how to play chess by treating win as a reward and loss as a penalty.
- Teaching a robot how to walk.

Some reinforcement learning algorithms

- ➔ Deep Q Neural Networks (DQN)
- ➔ Genetic Algorithms

Basic Machine Learning Project Workflow

The sequence of steps involved in a machine learning project are:

1. **Gathering data** from various sources. The performance of the Machine Learning model depends highly on the quality and quantity of data gathered.
2. The **Data Preparation** step involves cleaning (removing missing, duplicate, noisy, unwanted data) and preparing the data for analysis and model fitting.
3. **Exploratory Data Analysis (EDA)** involves understanding the data to reveal patterns and relationships within the data using plots, queries on the dataset, etc.
4. **Feature engineering** step involves choosing useful features, dropping redundant or unwanted features, and creating new features from existing ones to fit the data to the model better.
5. **Choosing the best model** for the given data among the multitude of different algorithms, using the obtained features and the insights gained from EDA is an important step in the project workflow.
6. **Splitting the dataset** into a training set (~80%) and a test set (~20%) randomly.
7. **Training the model** using the train set.

8. **Cross validation** of the model is done using a cross validation set (generally a subset of the training set) to estimate the model's performance and tune the parameters of the model for best fit and performance.
9. **Evaluation of the Model** is done using the test set (unseen) to assess how the model will generalize to unseen data.

Applications of Machine Learning

Machine Learning has applications in almost all fields. Some examples include:

- ➔ Image and speech recognition software.
- ➔ Stock market prediction.
- ➔ Product recommendations.
- ➔ Email spam filtering.
- ➔ Self-driving cars.
- ➔ Training robots for industry use and the list goes on...

Advantages and Disadvantages of Machine Learning

Advantages:

- ❖ It is easier to identify patterns and trends in data and predict them using Machine Learning models.
- ❖ Wherever there is data, there is an opportunity for Machine Learning and hence it has a wide range of applications from Medical to Space Technology.
- ❖ Machine Learning models can improve over time as more data is available for training, unlike traditional programs.
- ❖ Machine Learning can be used to automate various kinds of tasks like driving, social media analysis, virtual assistants, etc.

Disadvantages:

- ❖ Acquiring quality data and huge amounts of them for complicated models becomes difficult.
- ❖ It is a tedious task to run various kinds of algorithms on data to find the best one manually.
- ❖ Some models may consume huge amounts of time and processing power in the training process.
- ❖ It is very easy for errors to sneak into the models, and finding and getting rid of these errors becomes difficult.

Reference:

- ➢ Wikipedia
- ➢ GeeksforGeeks
- ➢ Machine Learning Project Workflow - Data Driven Investor
- ➢ Machine Learning Advantages and Disadvantages