Introduction to Machine Learning

What is Machine Learning?

- Machine learning is an application (subfield) of AI that provides systems the ability to learn on their own and improve from experiences without being programmed externally.
- Fundamentally, machine learning involves building mathematical models to help understand data. "Learning" enters the fray when we give these models tunable parameters that can be adapted to observed data; in this way the program can be considered to be "learning" from the data. Once these models have been fit to previously seen data, they can be used to predict and understand aspects of newly observed data.

Categories of Machine Learning

- At the most fundamental level, machine learning can be categorized into two main types:
    - Supervised learning
    - Unsupervised learning.

Supervised Machine Learning

- Supervised learning is a type of machine learning that uses labeled data to train machine learning models. In labeled data, the output is already known. The model just needs to map the inputs to the respective outputs.
- Supervised Learning methods need external supervision to train machine learning models. Hence, the name supervised. They need guidance and additional information to return the desired result
- **Models that can predict labels based on labeled training data**.
- For instance, we have images that are labeled a spoon or a knife. This known data is fed to the machine, which analyzes and learns the association of these images based on its features such as shape, size, sharpness, etc. Now when a new image is fed to the machine without any label, the machine is able to predict accurately that it is a spoon with the help of the past data.
- Examples of supervised learning applications are weather prediction, sales forecasting, stock price analysis.
- This is further subdivided into two categories: **classification** and **regression**.

Unsupervised Machine Learning

- In Unsupervised Learning, the machine uses unlabeled data and learns on itself without any supervision. The machine tries to find a pattern in the unlabeled data and gives a response.
- That's why, it is often described as "letting the dataset speak for itself".
- That means, it uses unlabeled data to train machines. Unlabeled data doesn’t have a fixed output variable. The model learns from the data, discovers the patterns and features in the data, and returns the output.
- Unsupervised learning finds patterns and understands the trends in the data to discover the output. So, the model tries to label the data based on the features of the input data.
    - The training process used in unsupervised learning techniques does not need any supervision to build models. They learn on their own and predict the output
- **Models that identify structure in unlabeled data**.
- Let's take a similar example of images (spoon or knife) like before, but this time we do not tell the machine whether it's a spoon or a knife. The machine identifies patterns from the given set and groups them based on their patterns, similarities, etc.
- One of the most common methods/tasks of unsupervised learning is **clustering**.
- Besides, *dimensionality reduction* is another important task of unsupervised learning, which search for more succinct (concise/compact) representations of the data. That is, *models that detect and identify lower-dimensional structure in higher-dimensional data*.

Supervised vs Unsupervised

- The data used in *supervised learning* is labeled. For example, you provide the system with a photo of an apple and let the system know that this is an apple. That is called labeled data. The system learns from the labeled data and makes future predictions. On the other hand, *unsupervised learning* does not require any labeled data because its job is to look for patterns in the input data and organize it.
- Supervised learning is mostly used to predict data, whereas unsupervised learning is used to find hidden patterns or structures in data.

Real-Life Applications of Supervised Learning

- **Risk Assessment**: Supervised learning is used to assess the risk in financial services or insurance domains in order to minimize the risk portfolio of the companies.
- **Image Classification**: Image classification is one of the key use cases of demonstrating supervised machine learning. For example, Facebook can recognize your friend in a picture from an album of tagged photos.
- **Fraud Detection**: To identify whether the transactions made by the user are authentic or not.
- **Visual Recognition**: The ability of a machine learning model to identify objects, places, people, actions, and images.

Real-Life Applications of Unsupervised Learning

- **Market Basket Analysis**: It is a machine learning model based on the algorithm that if you buy a certain group of items, you are less or more likely to buy another group of items.
- **Semantic Clustering**: Semantically similar words share a similar context. People post their queries on websites in their own ways. Semantic clustering groups all these responses with the same meaning in a cluster to ensure that the customer finds the information they want quickly and easily. It plays an important role in information retrieval, good browsing experience, and comprehension.
- **Delivery Store Optimization**: Machine learning models are used to predict the demand and keep up with supply. They are also used to open stores where the demand is higher and optimizing roots for more efficient deliveries according to past data and behavior.
- **Identifying Accident Prone Areas**: Unsupervised machine learning models can be used to identify accident-prone areas and introduce safety measures based on the intensity of those accidents.

Regression

- Regression falls under supervised learning.
- In regression, the labels are continuous quantities.
- Regression is used when the output variable is a real or continuous value. In this case, there is a relationship between two or more variables i.e., a change in one variable is associated with a change in the other variable. For example, salary based on work experience or weight based on height, etc.
- **Models that predict continuous labels**.
- Regression method is used when the predicted data is numerical.
- Let's consider two variables - *humidity* and *temperature*. Here, 'temperature' is the *independent variable* and 'humidity' is the *dependent variable*. If the temperature increases, then the humidity decreases. These two variables are fed to the model and the machine learns the relationship between them. After the machine is trained, it can easily predict the humidity based on the given temperature.
- Example of regression algorithms include: Linear regression.

Classification

- Classification falls under supervised learning.
- Classification is used when the output variable is categorical i.e. with 2 or more classes. For example, yes or no, male or female, true or false, etc.
- **Models that predict labels as two or more discrete categories**.
    - For example, in order to predict whether a mail is spam or not, we need to first teach the machine what a spam mail is by reviewing the content of the mail, reviewing the mail header, and then searching if it contains any false information. Certain keywords and blacklist filters that blackmails are used from already blacklisted spammers. All of these features are used to score the mail and give it a spam score. The lower the total spam score of the email, the more likely that it is not a scam. Based on the content, label, and the spam score of the new incoming mail, the algorithm decides whether it should land in the inbox or spam folder.
- Another example includes, if a business man wants to predict that a particular customer will come back to his shop or not, he will use a classification algorithm.
- Examples of classification algorithms include: Naïve Bayes, K-nearest neighbor (KNN), Decision Trees and Random Forests.

Clustering

- Clustering falls under unsupervised learning.
- Clustering is the method of dividing the objects into clusters that are similar between them and are dissimilar to the objects belonging to another cluster.
- **Models that detect and identify distinct groups in the data**.
- For example, consider a collection of pictures of different fruit. You feed this data to the model, and the model analyzes it to recognize any patterns. In the end, the machine categorizes the photos.
- For instance, Flipkart, Amazon, and other online retailers use clustering for their recommendation systems. Search engines also use clustering to analyze your search history to determine your preferences and provide you the best search results.
- One of the algorithms that fall under clustering is K-means clustering, and Principal Component Analysis.
