Introduction to machine learning

Rahee d s.

Reason for introduction of machine learning:

- Machine learning was introduced because it can solve problems that cannot be solved using typical approaches. The reason for this might be the high amount of data production by applications, the increase of computation power in the past few years and the development of better algorithms.

- Machine learning learns from data and also feeds upon it.It is a powerful tool for implementing artificial intelligence technologies.

Types of Machine Learning Based on the methods and way of learning:

Machine learning is divided into mainly four types, which are:

1. Supervised Machine Learning:

- It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately. Supervised learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox.

2. Unsupervised Machine Learning:

- It uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention.

3. Semi-Supervised Machine Learning :

- It is a type of machine learning that falls in between supervised and unsupervised learning. It is a method that uses a small amount of labeled data and a large amount of unlabeled data to train a model.

4. Reinforcement Learning:

- It is about taking suitable action to maximize reward in a particular situation.

All about train and test data:

- In machine learning, a dataset is usually split into two or more subsets to train and evaluate a model. The two most common subsets are the training set and the test set.

Train Dataset: Used to fit the machine learning model.
Test Dataset: Used to evaluate the fit machine learning model.

- The training set is used to train the model, while the test set is used to evaluate the model's performance after it has been trained. The idea is to test the model on data that it has not seen during the training process to see how well it generalizes to new data.

- The amount of data allocated to each set depends on the size and complexity of the dataset, as well as the algorithm being used. As a general rule of thumb, it is common to allocate 60% to 80% of the data to the training set, with the remaining 20% to 40% allocated to the test set.

- In addition to the training and test sets, it is also common to allocate a smaller portion of the data to a validation set. The validation set is used to tune hyperparameters and prevent overfitting. A common split is to allocate 60% to 70% of the data to the training set, 10% to 15% to the validation set, and the remaining 15% to 20% to the test set. However, the exact split can vary depending on the specific problem and dataset.

Stages of model building in machine learning:

The three stages of building a model in machine learning are:

1. Data Preprocessing: This stage involves collecting and cleaning the data, selecting relevant features, and transforming the data to make it suitable for modeling. This stage is critical because the quality of the data and the features selected can greatly impact the performance of the model.

2. Model Building: This stage involves selecting an appropriate algorithm or model architecture, training the model on the preprocessed data, and evaluating its performance using metrics such as accuracy, precision, recall, and F1 score. This stage also involves tuning hyperparameters to improve the performance of the model.

3. Model Deployment: This stage involves deploying the model in a production environment, where it can be used to make predictions on new data. This stage also involves monitoring the model's performance and updating it as needed to ensure that it continues to perform well.

Applications of Supervised Machine Learning in Modern Businesses:

Supervised machine learning algorithms have a wide range of applications in modern businesses, including:

1. Predictive analytics: Supervised machine learning algorithms can be used to predict future outcomes based on historical data, such as predicting customer churn, sales forecasting, or demand forecasting.

2. Classification: Supervised machine learning algorithms can be used to classify data into different categories, such as identifying spam emails, detecting fraud, or identifying customer segments.

3. Recommendation systems: Supervised machine learning algorithms can be used to build recommendation systems that suggest products or services to customers based on their previous purchases or preferences.

4. Natural language processing: Supervised machine learning algorithms can be used for natural language processing tasks, such as sentiment analysis, text classification, or named entity recognition.

5. Computer vision: Supervised machine learning algorithms can be used for computer vision tasks, such as object detection, image recognition, or facial recognition.

6. Personalization: Supervised machine learning algorithms can be used to personalize products or services based on customer preferences, such as personalized product recommendations or personalized marketing campaigns.

Overall, supervised machine learning algorithms can help businesses make better decisions, improve customer experiences, and increase efficiency by automating tasks and processes.

Semi-supervised Machine Learning:

- Semi-supervised machine learning is a type of machine learning in which an algorithm learns from both labeled and unlabeled data to improve its performance on a specific task. In contrast to supervised learning, where the algorithm is trained on labeled data, and unsupervised learning, where the algorithm learns from unlabeled data, semi-supervised learning combines the two.

Difference between machine Learning and Deep Learning:

S.NO	Deep Learning	Machine Learning
1.	To be qualified for deep learning, there has to be at least three layers	Can be defined as a shallow neural network which consists one input and one output, with barely one hidden layer
2.	Requires large amount of unlabelled training data	Requires small amount of data
3.	Performs automatic feature extraction without the need for human intervention	Cannot perform automatic feature extraction, requires labelled parameters
4.	High-performance hardware is required	High-performance hardware is not required
5.	Can create new features	Needs accurately identified features by human intervention
6.	Offers end-to-end problem solution	Tasks are divided into small portions and then forms a combined effect
7.	Takes a lot of time to train	Takes less time to train

Difference between supervised, unsupervised and reinforcement learning:

Criteria	Supervised ML	Unsupervised ML	Reinforcement ML
Definition	Learns by using labelled data	Trained using unlabelled data without any guidance.	Works on interacting with the environment
Type of data	Labelled data	Unlabelled data	No – predefined data
Type of problems	Regression and classification	Association and Clustering	Exploitation or Exploration
Supervision	Extra supervision	No supervision	No supervision
Algorithms	Linear Regression, Logistic Regression, SVM, KNN etc.	K – Means, C – Means, Apriori	Q – Learning, SARSA
Aim	Calculate outcomes	Discover underlying patterns	Learn a series of action
Application	Risk Evaluation, Forecast Sales	Recommendation System, Anomaly Detection	Self Driving Cars, Gaming, Healthcare

Unsupervised Machine Learning Techniques:

Unsupervised machine learning techniques are a class of machine learning algorithms that are used to identify patterns and relationships in data without explicit supervision or guidance from a human.

In unsupervised learning, the algorithm is trained on a dataset with no pre-existing labels or categories, and it is left to find patterns or groupings on its own.
There are several unsupervised learning techniques, including:

1. Clustering: Clustering algorithms group similar data points together based on their characteristics. Clustering can be used for customer segmentation, image or text grouping, and anomaly detection.

2. Dimensionality Reduction: Dimensionality reduction algorithms reduce the number of features in a dataset by identifying the most important ones. This can be useful for data visualization, feature selection, and speeding up training times.

3. Association Rules: Association rule mining algorithms identify relationships between variables in a dataset. This can be used to identify product bundles, market basket analysis, and customer behavior analysis.

4. Anomaly Detection: Anomaly detection algorithms identify unusual data points or outliers in a dataset. This can be useful for fraud detection, network intrusion detection, and fault detection.

Unsupervised learning is useful when there is no pre-existing knowledge or labeling of the data, or when the goal is to uncover hidden patterns or relationships in the data. Unsupervised learning is commonly used in fields such as data mining, computer vision, and natural language processing.

Difference between Inductive learning and Deductive learning:

Attribute	*Deductive approach*	*Inductive approach*
Direction	'Top-down'	'Bottom-up'
Focus	Prediction changes, validating theoretical construct, focus in 'mean' behaviour, testing assumptions and hypotheses, constructing most likely future	Understanding dynamics, robustness, emergence, resilience, focus on individual behaviour, constructing alterative futures
Spatial scales	Single (one landscape, one resolution)	Multiple (multiple landscapes, one resolution)
Temporal scales	Multiple (deterministic)	Multiple (stochastic)
Cognitive scales	Single (homogeneous preferences)	Multiple (heterogeneous preferences)
Aggregation scales	Single (core aggregation scale)	Single or multiple (one or more aggregation scales)
Predictive vs. stochastic accuracy	High-low (one likely future)	Low-high (many likely futures)
Data intensity	Low (group or partial attributes)	High (individual or group attributes)

Type I vs Type II error:

In hypothesis testing and statistical analysis, Type I and Type II errors are two types of errors that can occur when making decisions about a hypothesis.

- Type I error, also known as a false positive, occurs when a null hypothesis is rejected when it is actually true. In other words, the test detects an effect that is not actually present. Type I error is often denoted by the Greek letter alpha (α), and is typically set at a predetermined level, such as 0.05 or 0.01.

- Type II error, also known as a false negative, occurs when a null hypothesis is not rejected when it is actually false. In other words, the test fails to detect an effect that is actually present. Type II error is often denoted by the Greek letter beta (β), and is influenced by factors such as sample size, effect size, and statistical power.

The relationship between Type I and Type II errors is inversely related. As the probability of Type I error decreases (e.g., by lowering the alpha level), the probability of Type II error increases, and vice versa. Therefore, the goal of hypothesis testing is to strike a balance between these two types of errors based on the specific research question and context.

In summary, Type I error occurs when a null hypothesis is rejected when it is actually true, while Type II error occurs when a null hypothesis is not rejected when it is actually false. The goal of hypothesis testing is to minimize both types of errors based on the specific research question and context.

Ensemble learning:

Ensemble learning is a machine learning technique that involves combining multiple models to improve the overall performance and accuracy of a predictive model.
The basic idea behind ensemble learning is that by combining multiple models that are individually weak, it is possible to create a stronger and more accurate model.

There are several different types of ensemble learning techniques, including:

1. Bagging: Bootstrap Aggregating (bagging) involves training multiple models on different subsets of the training data, and then combining their predictions by averaging or taking a majority vote.
2. Boosting: Boosting involves training multiple models sequentially, with each model being trained on the errors of the previous model. The idea is to focus on the data points that are misclassified by the previous models and try to correctly classify them in subsequent models.
3. Stacking: Stacking involves training multiple models and then combining their predictions using a meta-model. The meta-model takes the predictions of the individual models as inputs and then outputs the final prediction.

Ensemble learning is a powerful technique that can improve the accuracy and robustness of machine learning models, especially when dealing with complex and noisy data. It is widely used in various applications, including image recognition, natural language processing, and predictive analytics.