Machine Learning Algorithms

- Machine learning algorithms are essentially sets of instructions that allow computers to learn from data, make predictions, and improve their performance over time without being explicitly programmed. Machine learning algorithms are broadly categorized into three types:
- **Supervised Learning**: Algorithms learn from labeled data, where the input-output relationship is known.
- **Unsupervised Learning**: Algorithms work with unlabeled data to identify patterns or groupings.
- **Reinforcement Learning:** Algorithms learn by interacting with an environment and receiving feedback in the form of rewards or penalties.

• Supervised Learning Algorithms

- Supervised learning algos are trained on datasets where each example is paired with a target or response variable, known as the label. The goal is to learn a mapping function from input data to the corresponding output labels, enabling the model to make accurate predictions on unseen data. Supervised learning problems are generally categorized into two main types: Classification and Regression. Most widely used supervised learning algorithms are:

• Linear Regression

- Linear regression is used to predict a continuous value by finding the best-fit straight line between input (independent variable) and output (dependent variable)
- Minimizes the difference between actual values and predicted values using a method called "least squares" to to best fit the data.
- Predicting a person's weight based on their height or predicting house prices based on size.

• Logistic Regression

- Logistic regression predicts probabilities and assigns data points to binary classes (e.g., spam or not spam).
- It uses a logistic function (S-shaped curve) to model the relationship between input features and class probabilities.
- Used for classification tasks (binary or multi-class).
- Outputs probabilities to classify data into categories.
- Example : Predicting whether a customer will buy a product online (yes/no) or diagnosing if a person has a disease (sick/not sick).
- Note : Despite its name, logistic regression is used for classification tasks, not regression.

• Decision Trees

- A decision tree splits data into branches based on feature values, creating a tree-like structure.
- Each decision node represents a feature; leaf nodes provide the final prediction.
- The process continues until a final prediction is made at the leaf nodes
- Works for both classification and regression tasks.
- For more decision tree algorithms, you can explore:
- Iterative Dichotomiser 3 (ID3) Algorithms
- Classification and Regression Trees Algorithms

- **Support Vector Machines (SVM)**
  - SVMs find the best boundary (called a hyperplane) that separates data points into different classes.
  - Uses support vectors (critical data points) to define the hyperplane.
  - Can handle linear and non-linear problems using kernel functions.
  - focuses on maximizing the margin between classes, making it robust for high-dimensional data or complex patterns.

- ## k-Nearest Neighbors (k-NN)
    - KNN is a simple algorithm that predicts the output for a new data point based on the similarity (distance) to its nearest neighbors in the training dataset, used for both classification and regression tasks.
    - Calculates distance between point with existing data points in training dataset using a distance metric (e.g., Euclidean, Manhattan, Minkowski)
    - identifies k nearest neighbors to new data point based on the calculated distances.
    - For classification, algorithm assigns class label that is most common among its k nearest neighbors.
    - For regression, the algorithm predicts the value as the average of the values of its k nearest neighbors.
- ## Naive Bayes
    - Based on Bayes' theorem and assumes all features are independent of each other (hence "naive")
    - Calculates probabilities for each class and assigns the most likely class to a data point.
    - Assumption of feature independence might not hold in all cases (rarely true in real-world data)
    - Works well for high-dimensional data.
    - Commonly used in text classification tasks like spam filtering : Naive Bayes
- ## Random Forest
    - Random forest is an ensemble method that combines multiple decision trees.
    - Uses random sampling and feature selection for diversity among trees.
    - Final prediction is based on majority voting (classification) or averaging (regression).
    - Advantages : reduces overfitting compared to individual decision trees.
    - Handles large datasets with higher dimensionality.
    - For in-depth understanding : What is Ensemble Learning? - Two types of ensemble methods in ML
- **Gradient Boosting (e.g., XGBoost, LightGBM, CatBoost)**
    - These algorithms build models sequentially, meaning each new model corrects errors made by previous ones. Combines weak learners (like decision trees) to create a strong predictive model. Effective for both regression and classification tasks. : Gradient Boosting in ML
    - **XGBoost (Extreme Gradient Boosting)** : Advanced version of Gradient Boosting that includes regularization to prevent overfitting. Faster than traditional Gradient Boosting, for large datasets.
    - **LightGBM (Light Gradient Boosting Machine):** Uses a histogram-based approach for faster computation and supports categorical features natively.
    - **CatBoost:** Designed specifically for categorical data, with built-in encoding techniques. Uses symmetric trees for faster training and better generalization.
    - For more ensemble learning and gradient boosting approaches, explore:
        - **AdaBoost**
        - **Stacking** - ensemble learning
- ## Neural Networks (Including Multilayer Perceptron)
    - Neural Networks, including Multilayer Perceptrons (MLPs), are considered part of supervised machine learning algorithms as they require labeled data to train and learn the relationship between input and desired output; network learns to minimize the error using backpropagation algorithm to adjust weights during training.

        - Multilayer Perceptron (MLP): Neural network with multiple layers of nodes.
        - Used for both classification and regression (Examples: image classification, spam detection, and predicting numerical values like stock prices or house prices)
    - For in-depth understanding : Supervised multi-layer perceptron model - What is perceptron?

• Unsupervised Learning Algorithms

- Unsupervised learning algos works with unlabeled data to discover hidden patterns or structures without predefined outputs. These are again divided into three main categories based on their purpose: Clustering, Association Rule Mining, and Dimensionality Reduction. First we'll see algorithms for Clustering, then dimensionality reduction and at last association.

- **Clustering**
    - Clustering algorithms group data points into clusters based on their similarities or differences. The goal is to identify natural groupings in the data. Clustering algorithms are divided into multiple types based on the methods they use to group data. These types include Centroid-based methods, Distribution-based methods, Connectivity-based methods, and Density-based methods. For resources and in-depth understanding, go through the links below.
- **Centroid-based Methods**: Represent clusters using central points, such as centroids or medoids.
    - K-Means clustering: Divides data into k clusters by iteratively assigning points to nearest centers, assuming spherical clusters.
    - K-Means++ clustering
    - K-Mode clustering
    - Fuzzy C-Means (FCM) Clustering
- **Distribution-based Methods**
    - Gaussian mixture models (GMMs) : Models clusters as overlapping Gaussian distributions, assigning probabilities for data points' cluster membership.
    - Expectation-Maximization Algorithms
    - Dirichlet process mixture models (DPMMs)
- **Connectivity based methods**
    - Hierarchical clustering : Builds a tree-like structure (dendrogram) by merging or splitting clusters, no predefined number.
    - Agglomerative Clustering
    - Divisive clustering
    - Affinity propagation
- **Density Based methods**
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) : Forms clusters based on density, allowing arbitrary shapes and detecting outliers, with distance and point parameters.
- OPTICS (Ordering Points To Identify the Clustering Structure)
- Dimensionality Reduction
- Dimensionality reduction is used to simplify datasets by reducing the number of features while retaining the most important information.
- **Principal Component Analysis (PCA):** Transforms data into a new set of orthogonal features (principal components) that capture the maximum variance.
- **t-distributed Stochastic Neighbor Embedding (t-SNE): Reduces** dimensions for visualizing high-dimensional data, preserving local relationships.
- **Non-negative Matrix Factorization (NMF) :** Factorizes data into non-negative components, useful for sparse data like text or images.
- **Independent Component Analysis (ICA)**

- Isomap : Preserves geodesic distances to capture non-linear structures in data.
- Locally Linear Embedding (LLE) : Preserves local relationships by reconstructing data points from their neighbors.
- Latent Semantic Analysis (LSA) : Reduces the dimensionality of text data, revealing hidden patterns.
- Autoencoders : Neural networks that compress and reconstruct data, useful for feature learning and anomaly detection.

- **Association Rule**
- Find patterns (called association rules) between items in large datasets, typically in market basket analysis (e.g., finding that people who buy bread often buy butter). It identifies patterns based solely on the frequency of item occurrences and co-occurrences in the dataset.
- Apriori algorithm : Finds frequent itemsets by iterating through data and pruning non-frequent item combinations.
- FP-Growth (Frequent Pattern-Growth) : Efficiently mines frequent itemsets using a compressed FP-tree structure without candidate generation.
- ECLAT (Equivalence Class Clustering and bottom-up Lattice Traversal) : Uses vertical data format for faster frequent pattern discovery through efficient intersection of itemsets.
- Reinforcement Learning Algorithms
- Reinforcement learning involves training agents to make a sequence of decisions by rewarding them for good actions and penalizing them for bad ones. Broadly categorized into Model-Based and Model-Free methods, these approaches differ in how they interact with the environment.
- Model-Based Methods
- These methods use a model of the environment to predict outcomes and help the agent plan actions by simulating potential results.
- Markov decision processes (MDPs)
- Bellman equation
- Value iteration algorithm
- Monte Carlo Tree Search
- Model-Free Methods
- These methods do not build or rely on an explicit model of the environment. Instead, the agent learns directly from experience by interacting with the environment and adjusting its actions based on feedback. Model-Free methods can be further divided into Value-Based and Policy-Based methods:
- Value-Based Methods: Focus on learning the value of different states or actions, where the agent estimates the expected return from each action and selects the one with the highest value.
- Q-Learning
- SARSA
- Monte Carlo Methods
- Policy-based Methods: Directly learn a policy (a mapping from states to actions) without estimating values where the agent continuously adjusts its policy to maximize rewards.
- REINFORCE Algorithm
- Actor-Critic Algorithm
- Asynchronous Advantage Actor-Critic (A3C)