Machine Learning -Curriculum

1. Introduction to Machine Learning

1.1 Introduction

- Machine learning is a subset of artificial intelligence that focuses on developing algorithms that enable computers to learn from data and make predictions or decisions without being explicitly programmed
- It involves techniques like supervised learning, unsupervised learning, and reinforcement learning to identify patterns and improve performance over time
- Machine learning is widely applied in areas such as natural language processing, computer vision, and predictive analytics.

1.2 Applications of Machine learning

- Fraud detection in finance and banking.
- Personalized recommendations in e-commerce and streaming services.
- Image and speech recognition in healthcare and security.
- Predictive maintenance in manufacturing and industrial applications.

1.3 Machine Learning Career

- Machine learning engineer
- Data scientist
- AI researcher
- Deep learning specialist
- Data analyst
- Machine learning consultant

2. Python

2.1 Introduction to Python

2.2 Python Functions, Packages, and Routines.

Functions

- Fuctions are blocks of reusable code that perform a specific task.They are defined using the def keyword, allow parameters, and can return results, making code more modular and organised.

Python Packages

- Packages are collections of modules that group related functions, classes, and routines together.

Routines

- Refers to a series of programmed instructions or functions that can be reused to perform common tasks. They help automate processes, improve efficiency, and minimise code duplication.

Machine Learning -Curriculum

2.3 Data Types, Operators, Variables

Data Types

- Python supports various data types, including integers (int), floating-point numbers (float), strings (str), and complex types like lists, tuples, dictionaries, and sets for managing diverse kinds of data.

Operators

- Python provides operators for performing operations on variables and values, including arithmetic (+, -, *, /), comparison (==, !=, <, >), logical (and, or, not), and assignment (=, +=, -=) operators.

Variables

- Variables are symbolic names assigned to values, acting as containers for storing data. They are dynamically typed in Python, meaning their type can change based on the assigned value.

2.4 Working with Data structure, Arrays, Vectors & Data Frames.

Data structures

- Data structures in Python (e.g., lists, tuples, dictionaries, and sets) are ways to store and organise data efficiently. They allow for easy access, modification, and management of data depending on the structure's properties.

Arrays

- Arrays (using libraries like numpy) and vectors are ordered collections of elements, typically of the same data type. Arrays support fast mathematical operations, while vectors are 1D arrays often used in linear algebra and machine learning.

Data Frame

- It is a two-dimensional, table-like data structure (from libraries like pandas) where data is stored in rows and columns. It's ideal for handling and manipulating structured data, similar to spreadsheets or SQL tables.

2.5 Syntax

- Rules and structure of code in programming.
- Defines correct keyword and symbol usage.
- Ensures code readability and functionality.
- Essential for error-free program execution.

2.6 Working with Numbers & Working with Strings

Working with Numbers

- Arithmetic operations like addition, subtraction, multiplication, and division.
- Handling numeric types like integers, floats, and complex numbers.

Working with Strings

- Manipulating text with functions like concatenation, slicing, and formatting.
- Supporting operations for string comparison, search, and transformation.

2.7 Conditional Statements

- Allow decision-making in programming based on conditions.
- Include if, else if, and else clauses.
- Enable branching logic for different outcomes.
- Support complex conditions with logical operators.

Machine Learning -Curriculum

2.7 For Loop & While Loop

For Loop

- Iterates over a sequence or range of values.
- Commonly used for executing code a specific number of times.

While Loop

- Repeats code while a condition remains true.
- Useful for indeterminate iterations until a condition changes.

2.8 Lists, Tuples, Sets

Lists

- Ordered, mutable collections that can hold mixed data types; defined with [ ]. Supports indexing, slicing, and dynamic modifications.

Tuples

- Ordered, immutable collections that can hold mixed data types; defined with ( ). Ideal for fixed data that should not be altered.

Sets

- Unordered, mutable collections with unique elements; defined with { }. Used for eliminating duplicates and efficient membership testing.

2.9 Dictionaries & Functions

Dictionaries

- Stores data in key-value pairs.
- Allows fast lookup, insertion, and deletion by keys.

Functions

- Encapsulate reusable blocks of code.
- Can accept parameters and return values.

2.10 Pandas, NumPy, Matplotib packages.

Pandas

- Powerful library for data manipulation and analysis, Pandas provides data structures like DataFrames, allowing for easy handling, cleaning, and transformation of structured data.

NumPy

- A fundamental package for numerical computations, NumPy offers support for multi-dimensional arrays and a wide range of mathematical functions for operations on arrays and matrices.

Matplotib

- A popular plotting library used for creating static, interactive, and animated visualisations in Python, Matplotlib allows users to generate a wide variety of charts, including line plots, histograms, and scatter plots.

Machine Learning -Curriculum

3. Applied Statistics

3.1 Introduction to Statistics

- What is Statistics
- Why do we need Statistics
- Statistical Models

3.2 Descriptive statistics

- Measure of Central Tendency
- Mean
- Outliers
- Median
- Mode
- Bimodal
- Understanding the Bell Curve

3.3 Measure of Spead

- Variance
- Standard Deviation

3.4 Probability

- What is Probability
- How to Calculate Probability with examples
- Union and Intersection

3.5 Conditional Probability

- What is Conditional Probability
- How to Calculate Conditional Probability with examples

3.6 Probability Distributions

- Probability distribution is a statistical function that mathematically defines the probabilities of different outcomes for a random variable.
- Types include discrete (e.g., Binomial, Poisson) and continuous (e.g., Normal, Exponential).
- It has components like the probability mass function (PMF) or probability density function (PDF).
- It is characterized by parameters such as mean, variance, and standard deviation.

3.7 Hypothesis Testing

- Hypothesis testing is a statistical method to evaluate assumptions about a population.
- It involves formulating Standard Erros, Null Hypothesis (H_0) and an Alternative Hypothesis (H_1).
- Types include One Tailed Test, Critical Region, Two Tailed Test
- Common tests include t-tests, chi-square tests, and ANOVA.
- Types of Errors inlcude Type I Errors & Type II Errors

Machine Learning -Curriculum

4. Supervised Learning

4.1 Regression

- Regression analyzes the relationship between dependent and independent variables.
- It predicts numerical values based on input data.
- Types include linear regression (one variable) and multiple regression (multiple variables).
- It helps in forecasting trends and understanding variable dependencies.

4.2 Multiple variable Linear regression

- Multiple variable linear regression models the relationship between one dependent and multiple independent variables.
- It predicts a continuous outcome based on several predictors.
- The model assumes a linear relationship between the variables.
- It is used for more complex predictions and understanding multi-variable dependencies.

4.3 Logistic Regression

- Logistic regression estimates the probability of a binary outcome by applying a logistic function, making it ideal for classification tasks.
- Used in fields like healthcare and finance for binary outcomes (e.g., risk assessment), it assumes no multicollinearity and a linear relationship between predictors and the log-odds.
- It includes binary logistic regression (two outcomes), multinomial logistic regression (multiple categories), and ordinal logistic regression (ordered categories).

4.4 KNN Classification

- KNN is a simple, instance-based algorithm that classifies data points based on the majority class of their nearest neighbors in feature space.
- Commonly used in recommendation systems and image recognition, KNN can be computationally intensive with large datasets and sensitive to irrelevant features.

4.5 Naive Bayes Classifiers

- Naive Bayes is a probabilistic classifier that uses Bayes' Theorem, assuming that features are conditionally independent given the target variable.
- Commonly used for text classification, spam detection, and sentiment analysis, it performs well on high-dimensional data but may struggle with correlated features.

4.6 Support Vector Machines

- Support Vector Machine (SVM) is a supervised learning algorithm used for classification and regression.
- It finds the hyperplane that best separates data into different classes.
- SVM can handle both linear and non-linear data using kernel functions.
- It aims to maximize the margin between data points of different classes for better generalization.

4.7 Decision Trees

- A supervised learning algorithm used for classification and regression.
- Splits data into subsets based on feature values using a tree-like structure.
- Nodes represent decisions, and leaves represent outcomes.
- Simple to interpret and visualize, making it popular for decision-making tasks.

Machine Learning -Curriculum

5. Unsupervised Learning

5.1 Clustering

- Partitioning Clustering
- Density Based Clustering
- Distribution Model Based Clustering
- Hierarchical Clustering
- Fuzzy Clustering

5.2 K-Means clustering

- An unsupervised learning algorithm for clustering data into K groups.
- Uses centroids to represent the center of each cluster.
- Iteratively assigns data points to the nearest centroid and updates centroids.
- Widely used for grouping similar data and pattern recognition.

5.3 Hierarchical clustering

- Top Down Approach
- Bottom Up Approach
- How to linkage affect the Dendrogram
- Single Linkage
- Complete Linkage
- Average Linkage

5.4 DBSCAN

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is an unsupervised clustering algorithm.
- Groups data points based on density, identifying regions of high data point concentration.
- Can handle noise and outliers by marking them as separate clusters.
- Does not require specifying the number of clusters in advance.

5.5 Dimensionality Reduction

- Advantages & Disadvantages
- Feature Selection
- Feature Extraction

5.6 Principal Component Analysis

- A dimensionality reduction technique used to simplify large datasets.
- Identifies the most significant features (principal components) of the data.
- Reduces the number of variables while preserving as much information as possible.
- Commonly used in data visualization and noise reduction.

5.7 Linear Discriminant Analysis

- A supervised dimensionality reduction technique for classification tasks.
- Maximizes class separability by finding a linear combination of features.
- Used to reduce the number of dimensions while retaining class information.
- Commonly applied in pattern recognition and face recognition.

Machine Learning -Curriculum

6. Recommendation Systems

6.1 Recommendation Systems

- Aims to predict user preferences and suggest items accordingly.
- Utilizes collaborative filtering, content-based filtering, or hybrid methods.
- Often used in e-commerce, streaming services, and social media platforms.
- Improves user experience by personalizing content and product suggestions.

6.2 Collaborative Filtering Systems

- A recommendation technique that predicts preferences based on user similarities.
- Relies on past interactions or ratings from other users to make suggestions.
- Can be user-based or item-based, depending on the approach.
- Commonly used in platforms like Netflix and Amazon for personalized recommendations.

6.3 Content Based Recommendation System

- Recommends items based on the features of the items themselves.
- Analyzes item attributes such as genre, keywords, or descriptions.
- Personalizes recommendations by matching user preferences with item features.
- Commonly used in platforms like movie or music streaming services.

6.4 Hybrid Recommendation System

- Combines multiple recommendation techniques, like collaborative and content-based filtering.
- Aims to improve accuracy and overcome limitations of individual methods.
- Can use weighted, switching, or mixed strategies to generate recommendations.
- Commonly used in e-commerce and media platforms for better personalization.

7. Semi-Supervised Learning

7.1 Overview

- Semi-supervised learning combines a small amount of labeled data with a large amount of unlabeled data to improve learning accuracy without extensive labeling costs.
- Common techniques include self-training, co-training, and graph-based methods, which iteratively use unlabeled data to enhance model predictions
- Often used in areas like image and speech recognition, where labeling is expensive, allowing models to learn patterns with minimal labeled input.
- Balances label scarcity and data availability but may suffer from lower accuracy if the unlabeled data is noisy or not representative of the task.

Machine Learning -Curriculum

7.2 Applications

- **Text Classification:** Used for categorizing documents, emails, or news articles by leveraging a small set of labeled examples to classify a larger set of unlabeled text.
- **Image Recognition:** Enhances performance in tasks like object detection and image classification by combining few labeled images with numerous unlabeled ones, reducing the need for extensive manual labeling.
- **Speech Recognition:** Improves models by using a limited number of transcribed audio clips along with vast amounts of untranscribed audio data for better speech recognition accuracy.
- **Web Content Classification:** Helps categorize web pages by using small labeled samples from a large pool of unlabeled pages, useful in search engine optimization and content recommendation systems.

7.3 Assumptions

- **Smoothness Assumption:** Assumes that similar inputs should have similar outputs, meaning that if two points are close in the input space, they are likely to belong to the same class.
- **Cluster Assumption:** Suggests that data points tend to form clusters, and points within the same cluster are more likely to share the same label than those in different clusters.

8. Reinforcement Learning

8.1 Intro to Reinforcement Learning

- A learning method where agents learn by interacting with an environment to maximize rewards.
- Uses feedback from actions to adjust and improve future behavior.
- Key elements include states, actions, rewards, and policies.
- Commonly applied in robotics, gaming, and autonomous systems.

8.2 Features

- Trial and Error Learning
- Reward System
- Policy
- Value Function
- Exploration vs. Exploitation
- Markov Decision Process (MDP

8.3 Approaches

- **Value-Based:** Focuses on estimating the value of actions to maximize cumulative reward (e.g., Q-learning).
- **Policy-Based:** Directly optimizes the policy that maps states to actions (e.g., REINFORCE algorithm).
- **Model-Based:** Builds a model of the environment to plan actions (e.g., Dynamic Programming).
- **Actor-Critic:** Combines value-based and policy-based approaches with separate actor and critic networks.

8.4 Types and Algorithms

- Positive Reinforcement
- Negative Reinforcement
- Exploratory Learning
- Exploitation Learning

Machine Learning -Curriculum

CAPSTONE PROJECTS

1 **Predict Diabetes with Machine Learning**

- **Data Preprocessing:** Handling missing values and scaling improved the model's performance, underscoring the importance of data preprocessing in machine learning.
- **Model Comparison:** Testing KNN, Decision Tree, and MLP classifiers revealed distinct strengths, highlighting the value of multiple models in robust prediction.
- **Hyperparameter Tuning:** Adjusting parameters, like neighbors in KNN and max depth in Decision Trees, refined accuracy, illustrating the impact of model tuning.

2 **Number of Orders Prediction**

- **Data Visualization Insight:** Pie charts helped understand the distribution of store types, location, and discounts, emphasizing the role of visual analysis in data comprehension.
- **Data Transformation:** Mapping categorical variables to numerical values streamlined the machine learning process, reinforcing the need for data preparation.
- **Model Implementation:** Using LightGBM for order prediction demonstrated how ensemble methods can efficiently handle structured data for accurate predictions.

3 **Human Action Detection**

- **Data Processing and Annotation:** Learn the importance of accurate data preprocessing and labeling to ensure reliable model training and high detection precision.
- **Feature Engineering Skills:** Develop skills in feature extraction and engineering, improving the model's ability to distinguish between various activities.
- **Machine Learning Model Selection:** Gain insights into selecting appropriate models for time-series data, enhancing accuracy in detecting human activities.

LIVE PROJECT

1 **Bike Sharing Demand Prediction**

- **Data Transformation:** Converting numeric codes into readable labels (e.g., seasons, months) enhances dataset interpretability, aiding in clearer visual analysis.
- **Feature Selection:** RFE and VIF analysis were essential for choosing impactful variables, emphasizing the importance of removing redundant features in regression.
- **Model Evaluation:** The high R2 score on test data validates the model's accuracy, showing that refined feature selection can lead to reliable predictions.