

Day 01

What is Machine Learning

Machine learning is a subset of artificial intelligence (AI) that allows software applications to become more accurate in predicting outcomes without being explicitly programmed to do so.

In machine Learning algorithms we give Data and the corresponding output it should have given as an output and it figures out the underlying Mechanish/procudeure behind it.

Where is ML necessary?

1. Writing a program for everything may not be possible eg: Spam email classifier. In ML we write one algorithm.
 2. In places where the number of cases are very high. eg: Dog, Cat classifier. In this case variety of breeds of dogs and cats exist. writing code for dixtinction is difficult and putting too many conditions might be confusing.
 3. Data Analysis, and Data Mining: Dta pr ML model lagayenge. After applying a ML model on data if some new info can be extracted, then that is data mining.
-

Day 02

AI vs ML vs DL

1. AI, 2. AI ka subset is ML, 3. ML ka subset is DL

Artificial Intelligence (AI)

AI is a broad field that encompasses the development of intelligent agents that can reason, learn, and act autonomously. It aims to create systems that can mimic human intelligence and perform tasks that typically require human cognition.

Machine Learning (ML)

ML is a subset of AI that focuses on the development of algorithms that allow computers to learn from data without being explicitly programmed. It enables systems to automatically improve their performance on a specific task by learning from experience.

Deep Learning (DL)

DL is a subfield of ML that utilizes artificial neural networks with multiple layers to learn complex representations of data. It has achieved remarkable success in various domains,

such as image recognition, natural language processing, and speech recognition.

In ML, after some amount of data, saturation happens, and after some amt of data providing more data, does nothing on the accuracy. but for DL, increasing teh amount of data does incerase the accuracy.

Day 03

Types of ML

1. Supervised: a. Regression b. Classification
2. Unsupervised: a. Clustering b. Dimenstionality Reduction c. Anamoly Detection d. Association
3. Semi-Supervised
4. Reinforcement Learning

ML is Aall about learning from data.

Supervised Learning:

If hamare paas Input and Output explicitly defined hai, basically labelled hai, then for new input, hame output Agar dataset me input and output hai, and hamara kaam input output ke beech relationship shamajhkr, naye input pr prediction karna hai, then that is Supervised Learning.

Data can be: 1. Numerical-age wt, cgpa 2. Categorical-gender, nationality

- Regression: If Output column Numerical hai toh, Regression
- Classification: If output column Categorical hai, then Classification

Unsupervised Machine Learning:

We have only Input.

- Clustering
- Dimentionality Reduction
- Anamoly Reduction
- Association Rule Learning

Clustering:

Clustering is a type of unsupervised machine learning technique. It is used to group similar data points together based on their inherent patterns or similarities, without any pre-defined labels or categories.

Think of it like this: you have a basket of fruits without any labels, and you want to group them. You might group them by color, size or shape. Clustering algorithms do something similar with

data.

Dimensionality Reduction:

Dimensionality reduction is a technique used to reduce the number of features (or dimensions) in a dataset while preserving as much of the important information as possible.

Is used in visualisation of data, where the dimension of data is very high. and thus direct plotting isnt really possible.

Why is this useful?

- Simplifies the data: High-dimensional data can be complex and difficult to visualize or analyze. Reducing the number of dimensions can make the data easier to work with.
- Improves model performance: Sometimes, having too many features can lead to overfitting in machine learning models. Dimensionality reduction can help prevent this by focusing on the most relevant features.

Anamoly Detection:

Anomaly detection, also known as outlier detection, is the process of identifying data points that deviate significantly from the normal or expected behavior within a dataset. These unusual data points are called anomalies or outliers.

Why is this useful?

- Fraud detection: Identifying unusual transactions can help prevent fraudulent activities.
- Network security: Detecting unusual network traffic patterns can indicate potential cyber attacks.

Association Learning:

- Association rule learning is a technique used to discover interesting relationships or associations among variables in large datasets. It aims to find rules that describe how often items or events occur together.
- Think of a supermarket. Association rule learning can be used to analyze customer purchase data and identify patterns like:

"Customers who buy diapers also tend to buy beer." "People who purchase bread often also buy butter."

Semi- Supervised Learning:

Semi-supervised learning is a machine learning approach that combines elements of both supervised and unsupervised learning. It leverages a small amount of labeled data along with a larger amount of unlabeled data to train models.

They have a massive amount of photos uploaded by users, but only a small fraction of those photos have faces manually tagged with names. This is where semi-supervised learning comes in.

Reinforcement Learning:

Reinforcement learning (RL) is a type of machine learning where an agent learns to make decisions by interacting with an environment. The agent receives feedback in the form of rewards or penalties based on its actions, and its goal is to learn a policy that maximizes its cumulative reward over time.

Key Concepts:

- Agent: The entity that interacts with the environment and makes decisions.
- Environment: The world or system that the agent interacts with.
- State: A representation of the current situation in the environment.
- Action: A choice that the agent can make to affect the environment.
- Reward: A numerical signal that the agent receives from the environment, indicating the goodness of its actions. How it Works:
- Observation: The agent observes the current state of the environment.
- Action Selection: Based on its current policy, the agent chooses an action to take.
- Environment Response: The environment transitions to a new state based on the agent's action.
- Reward: The agent receives a reward signal from the environment, indicating the immediate consequence of its action.
- Policy Update: The agent updates its policy based on the reward and the observed transition, aiming to improve its future decision-making.

Day 04:

Batch Machine Learning- Offline Models

No incremental Training, jo data hai wo pura ka pura initially ji use hojata hai. Train the whole lot eksaath.

1. data
2. model
3. test

4. server

A batch ML model, in the context of machine learning, refers to a model that is trained on a fixed dataset all at once, as opposed to being updated incrementally with new data as it becomes available.

Here's how it works:

- Data Collection: Gather and prepare a complete dataset that will be used for training.
- Model Training: Train the ML model using the entire dataset in one go.
- Deployment: Deploy the trained model to make predictions on new, unseen data.
- Periodic Retraining: As new data becomes available or the environment changes, the model may need to be retrained on an updated dataset. This retraining is typically done offline, in batches.

This makes the model static, but the environment is dynamic and constantly changing, so need for re training the whole model frequently.

Disadvantages:

1. Availability
2. Hardware Limitation
3. Lots of data

Pros:

- Simplicity: Batch training is often easier to implement and manage compared to online or incremental learning.
- Efficiency: Training on the entire dataset at once can be computationally efficient. Stability: Batch models are less prone to fluctuations caused by individual data points, leading to more stable predictions.

Cons:

- Lack of Adaptability: Batch models cannot adapt to changes in the data distribution or environment without being retrained.
- Latency: Retraining the model on new data introduces a delay in incorporating fresh information.

When to Use Batch ML:

- Static Environments: When the underlying patterns in the data are relatively stable and don't change frequently.
- Large Datasets: When you have a large amount of historical data available for training.
- Computational Constraints: When you have limited resources for continuous model

updates.

Online Machine Learning

Incremental in nature. Training can be done on production.

1. use small chunk of data to the model, model learns, we test the model, and then the model is on the server, and then more data is fed to the model, by which it both, 1. learns, and 2. predicts eg chatbots.

- Learning Data Rate: Learning data rate refers to the speed at which new data is fed to the model for learning. It's a crucial parameter that needs to be carefully tuned.

Here's why:

- Too fast: If the learning rate is too high, the model might overreact to new data points and become unstable, leading to fluctuating performance.
- Too slow: If the learning rate is too low, the model might not adapt quickly enough to changes in the data distribution and miss out on capturing important patterns.
- Out of Core Learning: Data is so huge, you cant use batch, so need to use Batch. Convert big data into smaller chunks, and trained in offline/batch type.
- Disadvantage:

1. Tricky
 2. Risky
-

Instance Based Learning VS Model Based Learning

Learning is of two types:

1. Memorization
2. Concept Based

Instance-Based Learning (Memory-Based Learning)

memorization

- No explicit model: Doesn't construct a general model during training. Instead, it stores all the training data.
 - Prediction based on similarity: When a new input is encountered, it finds the most similar instances (data points) from the stored training data and uses them to make predictions.
 - Examples: k-Nearest Neighbors (k-NN), Locally Weighted Regression.
-

Model-Based Learning

- Constructs a model: During training, it builds a generalized model that captures the relationships within the data.
- Prediction using the model: When a new input is presented, the learned model is used to make predictions, without directly referencing the original training data.
- Examples: Linear Regression, Decision Trees, Neural Networks.

1. Preprocessing dono me karne hai
2. model me parameters hote hai, but instance, wale me until a new data point is not received, no calculations are done.
3. instance wale me koi model hota hi nhi hai, but model wale me model ko save karte hai.
4. instance no generalization before scoring, otherwise already generalized.
5. instance wale me training data is a must, in other can remove input data.
6. instance ko lazy bhi boltw hai.

Challenges in ML

1. Gathering Data is difficult
2. Insufficient Data
3. Poor Data Quality
4. Non Representative Data
5. Sampling Noise and Sampling Bias
6. Irrelevant Features
7. Overfitting and Underfitting
8. Feature Engineering
9. Computational Complexity

Application Of ML

1. Retail- BigBasket, Amazon: Personalized Recommendations, Customer Churn Prediction: Predicting which customers are likely to stop shopping with a retailer and taking proactive measures to retain them.
- 2.

Machine Learning Development Life Cycle

Data can be gathered from varying places:

1. csv files
2. APIs se fetch use karke, json file me data

3. Web Scraping, websites like trivago, do data scraping from multiple online hotel websites, same way price analyzer where a products prices are compared across various sides
4. Database: but database isnt directly used kyuki website down hoskti hai, isiliye we make database warehouse
5. present in clusters

Steps:

1. Frame teh Problem
 2. Gathering Data
 3. Data Processing: data is bound to be unclean, can have problems. Remove duplicates, remove misiing avlues, remiove outliers, scaling.
 4. Exploratory Data Analysis: Study teh relationship between input and output, graph plotting, visualising, mean std variation, bivariant analysis, multivariant analysis, outlier database convert to balanced, imbalanced made balanaced, imbalanced means ki data biased na ho.
 5. Feature engineering and Selection: jab features bohot zyada hojaye toh, aise features jo final model me na ya kam required ho toh, removing them ya using them to extract more important and reliant data.
 6. Model Training, Evaluation and Selection: Trying out all possible models and algorithms so as to see which one performs the best for your data. Ensemble Learning: Combining or fusing of algorithm methodologies of multiple algorithms according to the required output.
 7. Model Deployment: Model ko binary file banate hai, and conevert into api, and teh inputs on the website are fetched and given to this binary fiel which then makes the predictions.
 8. Testing: eg:A/B Testing
 9. Optimise: data backup, model backup, rollback and then go live, load balancing, retraining the model often, Model can become rotten.
-