Machine Learning

Video 6:

Instance-Based Vs Model-Based Learning:

Instance-Based Learning: This approach stores training instances and makes predictions by comparing new data to stored instances. It doesn't build an explicit model but uses proximity-based algorithms (e.g., k-NN) to classify. The learning process happens during prediction, not before.

Example: k-Nearest Neighbors (k-NN), Case-Based Reasoning

Model-Based Learning: Here, a model is created from training data, capturing patterns and generalizing to unseen instances. The model is learned in advance and then used for making predictions (e.g., decision trees, linear regression). It's more efficient for large datasets and makes predictions faster.

Example: Decision Trees, Linear Regression

How they differ?

Aspect	**Instance-Based Learning**	**Model-Based Learning**
Learning Process	Learns during prediction by comparing new instances to stored data.	Builds a model from the data before making predictions.
Memory Usage	Requires storing all training instances.	Stores only the model parameters, not the full data.

Video 7:

Challenges in Machine Learning:

1. Insufficient or Biased Data: Machine learning models require large, diverse datasets for accuracy. Insufficient or biased data can lead to poor generalization and skewed results.
2. Overfitting and Underfitting: Overfitting occurs when the model learns noise in the training data, while underfitting happens when the model is too simple to capture the patterns. Both lead to poor performance on new data.
3. Model Interpretability: Complex models, like deep learning, can be difficult to interpret and explain. Lack of transparency makes it challenging to trust and deploy these models in critical applications.
4. High Computational Cost: Training advanced machine learning models, particularly deep learning, requires substantial computational power and time. This can be a barrier for many organizations with limited resources.
5. Lack of Labeled Data: Many machine learning algorithms require labeled data for training, which can be time-consuming and expensive to obtain. The absence of labeled data limits the effectiveness of supervised learning.

6. Data Privacy and Security: Protecting sensitive data while using it for model training is a major concern. Data privacy issues arise when dealing with personal or confidential information.
7. Model Generalization: Ensuring that models perform well on unseen data is a challenge. Models may struggle to generalize if the training data doesn't adequately represent real-world scenarios.
8. Data Imbalance: Imbalanced datasets, where one class is underrepresented, can lead to biased predictions. This is particularly problematic in classification tasks, resulting in inaccurate or unfair outcomes.
9. Feature Selection and Engineering: Identifying relevant features and transforming them into suitable formats for the model is critical. Poor feature selection can degrade model performance.
10. Scalability of Algorithms: Some machine learning algorithms struggle to handle large datasets or high-dimensional data. Scaling algorithms efficiently for big data is a major challenge in practical applications.

Video 8:

Application of Machine Learning:

1. Healthcare Diagnostics: Machine learning models are used to analyze medical images and patient data to detect diseases like cancer, diabetes, and heart conditions. These models can assist doctors in making faster, more accurate diagnoses.
2. Natural Language Processing (NLP): NLP algorithms power applications like chatbots, voice assistants, and language translation tools. They enable machines to understand, interpret, and generate human language, improving communication between humans and computers.
3. Fraud Detection: Machine learning models are employed by banks and financial institutions to detect fraudulent transactions in real-time. These models learn from transaction patterns to identify unusual behavior, preventing potential fraud.

Video 9:

Machine Learning Development Life Cycle:

The Machine Learning Development Life Cycle is the process of building, deploying, and maintaining machine learning models, including steps like data collection, preprocessing, model training, evaluation, and deployment.

The steps in the Machine Learning Development Life Cycle are:

1. **Problem Definition**: Understand and define the problem to be solved.
2. **Data Collection**: Gather relevant data from various sources.
3. **Data Preprocessing**: Clean and transform raw data into a usable format.
4. **Feature Engineering**: Select and create relevant features from the data.
5. **Model Selection**: Choose the appropriate machine learning algorithm.
6. **Model Training**: Train the model using the prepared data.
7. **Model Evaluation**: Assess the model’s performance using evaluation metrics.
8. **Model Optimization**: Tune the model to improve performance.

9. **Model Deployment**: Deploy the model into a production environment.
10. **Monitoring and Maintenance**: Continuously monitor the model's performance and retrain as needed.

Video 10:

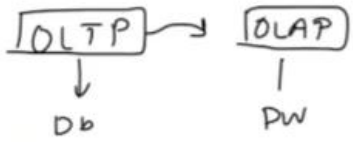
Data Engineer Vs Data Analyst Vs Data Scientist Vs ML Engineer:

1. Data Engineer

Wednesday, March 24, 2021 1:25 PM

Job Roles

- Scrape Data from the given sources.
- Move/Store the data in optimal servers/warehouses.
- Build data pipelines/APIs for easy access to the data.
- Handle databases/data warehouses.



Skills Required

- Strong grasp of algorithms and data structures
- Programming Languages (Java/R/Python/Scala) and script writing
- Advanced DBMS's
- BIG DATA Tools (Apache Spark, Hadoop, Apache Kafka, Apache Hive)
- Cloud Platforms (Amazon Web Services, Google Cloud Platform)
- Distributed Systems
- Data Pipelines

---

2. Data Analyst

Wednesday, March 24, 2021 1:26 PM

Responsibilities of a Data Analyst

- *Cleaning and organizing Raw data.*
- *Analyzing data to derive insights.*
- *Creating data visualizations.*
- *Producing and maintaining reports.*
- *Collaborating with teams/colleagues based on the insight gained.*
- *Optimizing data collection procedures*

Skills

- *Statistical Programming*
- *Programming Languages (R/SAS/Python)*
- *Creative and Analytical Thinking*
- *Business Acumen — Medium to High preferred*
- *Strong Communication Skills.*
- *Data Mining, Cleaning, and Munging*
- *Data Visualization*
- *Data Story Telling*
- *SQL*
- *Advanced Microsoft Excel*

3. Data Scientist

Wednesday, March 24, 2021 1:26 PM

"A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician".

---

4. ML Engineer

Wednesday, March 24, 2021 1:26 PM

Responsibilities

- Deploying machine learning models to production ready environment
- Scaling and optimizing the model for production
- Monitoring and maintenance of deployed models

Skills

- Mathematics
- Programming Languages (R/Python/Java/Scala mainly)
- Distributed Systems
- Data model and evaluation
- Machine Learning models
- Software Engineering & Systems design

	ANALYTICAL SKILLS	BUSINESS ACUMEN	DATA STORYTELLING	SOFT SKILLS	SOFTWARE SKILLS
DATA ANALYST	HIGH	MEDIUM TO HIGH	HIGH	MEDIUM TO HIGH	MEDIUM
DATA ENGINEER	MEDIUM	LOW	LOW	MEDIUM	HIGH
DATA SCIENTIST	HIGH	HIGH	HIGH	HIGH	MEDIUM
ML ENGINEER	MEDIUM TO HIGH	MEDIUM	LOW	HIGH	HIGH
