Machine Learning Operations (MLOps)

Course Description

This course aims to give students a comprehensive understanding of Machine Learning Operations (MLOps). MLOps is an emerging discipline that deploys, manages, and optimizes machine learning models in production environments. The course will cover various aspects of MLOps, including model development, version control, testing, deployment, monitoring, and scaling. Students will gain hands-on experience with popular MLOps tools and frameworks, enabling them to manage machine learning workflows effectively and deliver robust and scalable ML solutions.

Prerequisites

- Knowledge of machine learning algorithms and techniques

- Proficiency in a programming language (e.g., Python)

- Familiarity with data preprocessing and feature engineering

- Basic understanding of software development principles

Course Objectives

By the end of the course, students will be able to:

1. Understand the principles and challenges of MLOps.

2. Apply best version control and collaboration practices in machine learning projects.

3. Develop reproducible machine learning pipelines using tools like Git and Docker.

4. Implement testing methodologies to ensure model quality and reliability.

5. Deploy machine learning models in various production environments (local, cloud, etc.).

6. Monitor and evaluate model performance and make necessary adjustments.

7. Scale machine learning workflows to handle large datasets and high traffic.

Course Outline

Module 1: Introduction to MLOps

- Overview of MLOps and its significance
- Key challenges in deploying and managing ML models in production
- Comparison of traditional software development and MLOps
- Key components of MLOps
- Landscape of MLOps tools and technologies
- *Practice*: Build a simple scripts-based automation pipeline for image classification

Module 2: Version Control and Collaboration in MLOps

- Introduction to Git and GitHub/GitLab
- Branching and merging strategies for ML projects
- Managing large datasets using Git LFS (Large File Storage)
- *Practice*: Exercise problems to tackle a) code changes, b) data changes, and c) model changes through Git and Git LFS.

Module 3: Machine Learning Model Development Pipelines

- Overviews of popular machine learning frameworks and libraries
- Data preprocessing and feature engineering
- Introduction to MLflow for model development and monitoring
- *Practice*: Build a language classification system using MLFlow & HFmodels

Module 4: Model Deployment and Serving

- Overview of containerization and orchestration technologies
- Building Docker images for ML applications
- Deployment strategies for different environments (cloud, edge, on-premises)
- Introduction to model serving frameworks (e.g., TensorFlow Serving, FastAPI)
- *Practice*:
    - Build a Chatbot API using Llama2 and Fast API
    - Dockerize the API and deploy it horizontally behind Nginx
    - Build an ONNX API over TF Serving for Seq2Seq classification

Module 5: Monitoring and Performance Optimization

- Techniques for monitoring model performance in production
- Logging and error tracking for ML systems
- Performance optimization and scaling strategies
- Introduction to Prometheus toolkit for alerting and monitoring operations
- Introduction to Grafana for visualization and analytics
- *Practice*: Implement a logging system to track the performance of your model

References

- Introducing MLOps – Mark Traveil & The Dataiku Team – O'Reilly publication
- Practical MLOps – Noah Gift & Alfredo Deza – O'Reilly publication
- Designing Deep Learning Systems – Chi Wang, Donald Szeto – Manning publication
- Designing Machine Learning Systems - Chip Huyen – O'Reilly publication
- MLOps Engineering at Scale – Carl Osipov – Manning publication
- Beginning MLOps with MLFlow – Sridhar Alla & Suman Kalyan Adari – Apress