SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Artificial Intelligence and Data Science

Course Name – 19AD501 Big Data Analytics

III Year / V Semester

Unit 4 – Data Preparation

Topic 2- Data Pipeline and ML

Data Pipeline

- A machine learning pipeline is used to help automate machine learning workflows.
- They operate by enabling a sequence of data to be transformed and correlated together in a model that can be tested and evaluated to achieve an outcome, whether positive or negative.
- Machine learning (ML) pipelines consist of several steps to train a model.
- Machine learning pipelines are iterative as every step is repeated to continuously improve the accuracy of the model and achieve a successful algorithm.

Data Pipeline

The main objective of having a proper pipeline for any ML model is to exercise control over it.

A well-organised pipeline makes the implementation more flexible.

A typical machine learning pipeline would consist of the following processes:

- Data collection
- Data cleaning
- Feature extraction (labelling and dimensionality reduction)
- Model validation
- Visualization

Data Pipeline

The main objective of having a proper pipeline for any ML model is to exercise control over it.

A well-organised pipeline makes the implementation more flexible.

A typical machine learning pipeline would consist of the following processes:

- Data collection
- Data cleaning
- Feature extraction (labelling and dimensionality reduction)
- Model validation
- Visualization

The machine learning data pipeline helps identify patterns in given data, which leads businesses to better decision-making.

The machine learning pipeline boosts the machine learning model's performance leading to more efficient model deployment and better management of the models.

Data Pipeline

ML Pipeline Architecture

There are various stages in a machine learning pipeline architecture, mainly-

- Data preprocessing,
- Model training,
- Model evaluation, and
- Model deployment.

Each stage of the data pipeline passes processed data to the next step.

Data Pipeline

Data Preprocessing

- This step entails collecting raw and inconsistent data selected by a team of experts.
- The pipeline processes the raw data into an understandable format.
- Data processing techniques include feature extraction, feature selection, dimensionality reduction, sampling, etc. The final sample used for training and testing the model is the output of data preprocessing.

Model Training

- Selecting an appropriate machine learning algorithm for model training is crucial in a machine learning pipeline architecture.
- A mathematical algorithm specifies how a model will detect patterns in data.

Data Pipeline

Model Evaluation

The sample models are trained and tested on historical data to make predictions and choose the best-performing model for the next step.

Model Deployment

The final step is to deploy the machine learning model to the production line. Ultimately, the end-user can obtain predictions based on real-time data.

Data Pipeline

How do Machine Learning Pipeline Tools Benefit Businesses?

Accurate Machine Learning Models

It creates better machine learning models that will generate more accurate predictions.

Faster Deployment

Data pipeline automation accelerates the process of training, testing, and refining machine learning models, allowing you to deploy them sooner in the market.

Enhanced Business Forecasting

You may improve your business forecasting abilities by using data pipeline technologies that help you construct a better machine learning model.

Improved business forecasting enables you to stay ahead of the competition, provide a better client experience, and reap business profits.

Data Pipeline

Popular tools used in building an end-to-end machine learning pipeline-

MLFlow

MLflow is a free and open-source tool for managing machine learning workflow, including experimentation, production, deployment, and a centralized model repository.

DVC

Data Version Management, or DVC, is an experimental tool that helps define your pipeline irrespective of the programming language used.

Neptune

Neptune is a machine learning metadata repository designed for monitoring various experiments by research and production teams.

Polyaxon

Polyaxon is a Kubernetes machine learning platform for recreating and managing machine learning workflows.

THANK YOU