Workshop

AI/ML Pipelines with Python

Arijit Saha
VP Data Engineering, Noodle.ai

Sumit Sen
Director, Software Engineering, Noodle.ai

Summary:

In this hands-on workshop, participants will learn how to use Python for building AI/ML Pipelines.

Participants will work on a real-life scenario of building AI/ML pipeline covering the various aspects like

- Ingesting data
- Cleaning & Transforming data
- Perform Exploratory Data Analysis (EDA) on the dataset
- Running ML models
- Analyzing results
- Conclusion: Stitching it all-together as a Pipeline

As part of this workshop participants will be introduced to various useful Python libraries that every AI/ML Engineer should know. The session will cover various other aspects of a robust, scalable AI/ML pipeline.

Prerequisites:

This is an intermediate level hands-on workshop on Python.

To benefit from this course the participants are expected to have

- Basic familiarity with Python programming
- Conceptual knowledge of data pipelines, relational data and big data
- Using Jupyter Python notebook environment

Detail Agenda:

1. Python basics:

   Though participants are expected to have basic familiarity with Python programming, we will start with a quick recap of the following topics to ensure all participants are able to follow-through the hands-on exercise

   a) Variables & examples of some built-in functions
   b) Conditional & Control flow statements
   c) Working with Collections
   d) Handling Strings
   e) Lambda Functions
   f) Classes and objects

2. Environment setup:

   In this module, the environment setup for the rest of the workshop will be covered. Participants will be introduced to the following

   a) Docker
   b) Working with Jupyter notebooks
   c) Python Virtual Environment

3. Ingesting data

   In this module, participants will learn how to read and ingest data using Python. A public data-source will be used for this module. Real-life examples will be provided on how to read data from different sources like FTP, API, Databases and ingest it to the target database.

   Key take-aways from this module are

   a) How to ingest data from wide variety of sources
   b) Considerations & best practices for data ingestion

4. Cleaning & Transforming data

   In this module, participants will learn various techniques to clean & prepare data. Some of the key concepts that will be covered in this module are

   a) Data Formatting – handling various data types, type conversions etc.
   b) Dealing with missing values in the ingested data
   c) Data normalization & Binning
   d) Handling categorical variables
   e) Outlier detection

5. Exploratory Data Analysis

In this module, participants will be introduced to the concept of Exploratory Data Analysis (EDA) through

a) Descriptive Statistics
b) Univariate & Bivariate Statistics
c) Correlation
d) Simple aggregations
e) Plots

6. Running ML models

In this module, participants will be taken through examples of how ML models can be deployed and run as part of an AI/ML Pipeline. This module will not cover the Machine Learning concepts or algorithms and the focus will be primarily on how to integrate the ML models in the pipeline.

Types of models that will be used as an example are

a) A linear regression model
b) A binary classification model

7. Analyzing results

In this module, participants will be introduced to the various techniques for

a) Interpreting the results
b) Measuring ML model accuracy
c) Evaluation alerts

8. Conclusion: Stitching it all together as a Pipeline

In this final module, the various steps taken in the workshop will be stitched together to build the AI/ML Pipeline. Participants will learn from real-world examples on how AI/ML pipelines can be orchestrated in Production environment.

The workshop will conclude with this module.

Code files along with all pre-requisite setup instructions will be provided ahead of the Workshop.