Course Outline

Machine Learning Operations

Course Summary

Description

Machine Learning is all the rage today. Most ML courses focus on building models. However, taking the ML models to production involves quite a bit of extra work. This course will teach **Machine Learning Engineering** – the process of productionizing, monitoring, and managing ML models.

Objectives

At the end of this course, students will learn:

- ML Engineering Overview
- Overview of the AI capabilities in the cloud
- Storing large data in the cloud
- Processing large data in the cloud using distributed tools
- Training models at scale, using GPUs in the cloud
- Deploying models as webservices
- Logging and tracing of model runtime
- Model metrics
- Setting up alerts
- A/B testing different models
- Updating newer model versions

Topics

- Machine Learning Engineering Overview
- Cloud Storage
- Cloud Compute
- Training in the cloud with GPUs
- Creating a Model Service
- Containerizing the app
- Deploying the Container in the Cloud
- Monitoring
- Load-testing the application
- A/B Testing of Models
- Updating your model
- Final Workshop (Time Permitting)

Audience

This course is intended for Data Scientists, DevOps, and Data Engineers.

Prerequisites

Students for this course should possess the following:

- Basic knowledge of Machine Learning or Deep Learning
- Python basic programming recommended

Duration

Three days

Machine Learning Operations

Course Outline

I. *Machine Learning Engineering Overview*
   a. Machine Learning workflow
   b. Going from notebooks to production
   c. Understanding what is involved in ML Eng
   d. Lab: Getting up and running cloud environment

II. *Cloud Storage*
   a. Bringing data into the cloud
   b. Data storage options in the cloud
   c. Ingesting data
   d. Lab: Ingesting data in the cloud

III. *Cloud Compute*
   a. Understanding the different types of compute resources
   b. Using GPU instances
   c. Customizing a cloud VM
   d. Lab: Using cloud VMs

IV. *Training in the cloud with GPUs*
   a. GPU options in the cloud
   b. Training with a GPU
   c. Monitoring training using Tensorboard
   d. Early stopping training when desired accuracy is reached
   e. Lab: Various labs on training

V. *Creating a Model Service*
   a. Creating a simple web service for serving predictions
   b. Loading the saved model
   c. Serving incoming requests
   d. Error handling
   e. Running and testing on local environment
   f. Lab: Training and deploying a model

VI. *Containerizing the app*
   a. Create a Dockerfile with app artifacts and dependencies
   b. Building a Docker container
   c. Test the container locally
   d. Lab: Docker basics

VII. *Deploying the Container in the Cloud*
   a. Publishing the container to registry
   b. Deploy the container using Kubernetes
   c. Testing the deployed application
   d. Lab: Kubernetes deployment

VIII. *Monitoring*
   a. Inspecting application logs
   b. Monitoring application metrics
   c. Setting up alerts
   d. Lab: Monitoring in the cloud

IX. *Load-testing the application*
   a. Setting up load testing clients
   b. Observing application behavior under load
   c. Verifying the load balancer
   d. Scaling with the load
   e. Lab: Stress-testing an application

X. *A/B Testing of Models*
   a. Splitting traffic between different models
   b. Observe metrics
   c. Picking the best model
   d. Lab: Google Anthos deployment

XI. *Updating your model*
   a. Packing newer model into a container
   b. Performing a rolling update on running containers
   c. Exercise: Rolling back in case of failures

XII. *Final Workshop (Time Permitting)*
   a. Attendees will work in groups to implement a solution end to end
   b. They will 'ship' an ML model to the cloud