Course Outline

Machine Learning Engineering

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, as illustrated diagram below.

This course will teach Machine Learning Engineering - the process of productionizing, monitoring and managing ML models.

We will use a cloud environment (Google Cloud or Amazon Cloud or Microsoft Cloud) for our deployment.

Topics

- ML Engineering overview
- Overview of the AI capabilities of the Cloud Platform of choice
- Storing large data in the cloud
- Processing large data in the cloud using distributed tools
- Training models at scale, using GPUs on 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

Audience

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

Prerequisites

- Some knowledge in Machine Learning or Deep Learning is highly recommended
  - You may take one of these courses: *'Machine Learning in Python', 'Deep Learning'*
- Some basic knowledge of Python is highly recommended.
  Our labs utilize Python language. But Python is a very easy language to learn. So even you don't have previous exposure to Python, you will be able to complete the labs.

Duration

Four days

Course Outline

Machine Learning Engineering

Course Outline

I. ML Eng 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 in the 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 into the cloud

III. Cloud Compute
- A. Understanding 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

VI. Containerizing the app
- A. Create a docker file with app artifacts and dependencies
- B. Building a docker container
- C. Test the container locally

VII. Deploying the Container in the Cloud
- A. Publishing the container to registry
- B. Deploy the container using Kubernetees
- C. Testing the deployed application

VIII. Monitoring
- A. Inspecting application logs
- B. Monitoring application metrics
- C. Setting up alerts

IX. Load Testing the application
- A. Setting up load testing clients
- B. Observing application behavior under load
- C. Verifying load balancer
- D. Scaling with the load

X. A/B Testing of Models
- A. Splitting traffic between different models
- B. Observe metrics
- C. Picking the best model

XI. Updating model
- A. Packaging 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 implemnet a solution end to end.
- B. They will ‘ship’ an ML model to the cloud