



Week 2

Deploying Machine Learning Models in Production

Overview

- Week 1
- Week 2
- Week 3
- Week 4

Grades

Notes

Discussion Forums

Messages

Course Info

Week 2

Discuss the topic here.

Go to forum

2 threads · Last post a month ago

Week 2: Model Serving: Patterns and Infrastructure



Learn how to serve models and deliver batch and real-time inference results by building scalable and reliable infrastructure

Learning Objectives

- Serve models and deliver inference results by building scalable and reliable infrastructure.
- Contrast the use case for batch and realtime inference and how to optimize performance and hardware usage in each case
- Implement techniques to run inference on both edge devices and applications running in a web browser
- Outline and structure your data preprocessing pipeline to match your inference requirements
- Distinguish the performance and resource requirements for static and stream based batch inference

Show Less



Model Serving Architecture

Video: Model Serving Architecture 4 min

Resume

Video: Model Servers: TensorFlow Serving 3 min

Video: Model Servers: Other Providers 5 min

Reading: Documentation on model servers 10 min

Practice Quiz: Model serving architecture 3 questions

Reading: Ungraded Lab - Deploy a ML model with FastAPI and Docker 1h

Scaling Infrastructure

Video: Scaling Infrastructure 10 min

Reading: Learn about scaling with boy bands 10 min

Reading: Explore Kubernetes and KubeFlow 10 min

Practice Quiz: Scaling Infrastructure 3 questions

Reading: Ungraded Lab: Intro to Kubernetes 1h 10m

Online Inference

Video: Online Inference 6 min

Practice Quiz: Online Inference 3 questions

Reading: Ungraded Lab - Latency testing with Docker Compose and Locust 45 min



Data Preprocessing

Video: Data Preprocessing 4 min





Reading: Data preprocessing 10 min

Practice Quiz: Data Preprocessing 3 questions


Batch Inference Scenarios

-  **Video:** Batch Inference Scenarios 5 min
-  **Practice Quiz:** Batch inference scenarios 3 questions

Batch Processing with ETL

-  **Video:** Batch Processing with ETL 3 min
-  **Reading:** Ungraded Lab (Optional): Machine Learning with Apache Beam and TensorFlow 45 min
-  **Practice Quiz:** Batch Processing with ETL 3 questions
-  **Graded External Tool:** Autoscaling TensorFlow model deployments with TF Serving and Kubernetes 2h Due Jan 3, 2:59 AM EST

Lecture Notes (Optional)

-  **Ungraded External Tool:** Lecture Notes W2 5 min