Overview
Week 1
Week 2
Week 3
Week 4
Grades
Notes
Discussion Forums
Messages
Course Info

Week 1

Deploying Machine Learning Models in Production

Week 1

Discuss the topic here.

3 threads · Last post 22 days ago

Go to forum

Week 1: Model Serving: Introduction

Learn how to make your ML model available to end-users and optimize the inference process

Learning Objectives

- Identify and contrast the challenges for serving inference requests
- Compare cost, latency and throughput metrics to optimize serving inference requests
- Judge the hardware resources and requirements for your serving models so that your system is reliable and can scale based on demand
- Install and Use TensorFlow Serving to serve inference requests on a simple image classification model

Show Less

A conversation with Andrew Ng, Robert Crowe and Laurence Moroney

Video: Course Overview 4 min Resume

Introduction to Model Serving

Video: Introduction to Model Serving 6 min

Quiz: Introduction to Model Serving 4 questions Due Dec 27, 2:59 AM EST

Reading: Ungraded Labs - Best Practices 5 min

Reading: Ungraded Lab - Introduction to Docker 20 min

Ungraded External Tool: Join us on Discourse! 1h

Introduction to Model Serving Infrastructure

Video: Introduction to Model Serving Infrastructure 5 min

Video: Deployment Options 3 min

Video: Improving Prediction Latency and Reducing Resource Costs 5 min

Video: Creating and deploying models to AI Prediction Platform 2 min

Reading: Optional: Build, train, and deploy an XGBoost model on Cloud AI Platform 45 min

Quiz: Introduction to Model Serving Infrastructure 5 questions Due Dec 27, 2:59 AM EST

Installing TensorFlow Serving

Video: Installing TensorFlow Serving 6 min

Quiz: TensorFlow Serving 2 questions Due Dec 27, 2:59 AM EST

Reading: Ungraded Lab - Tensorflow Serving with Docker 20 min

Reading: Ungraded Lab - Serve a model with TensorFlow Serving 30 min

Lecture Notes (Optional)

Ungraded External Tool: Lecture Notes W1 5 min