



MULTIMEDIA INFORMATICS



MMI 712– Machine Learning Systems Design and Deployment

SYLLABUS

Year, Semester: 2022-2023 Fall
Course Conduct: Face-to-face in class II-06 Wednesday@13:40
Lecture videos on YouTube & Lecture notes on ODTUClass

Students are expected to watch the lecture videos and study the course material before attending the weekly face-to-face sessions.

There will be regular quizzes from the content and attendance is expected in these sessions.

Lecturer: Prof. Dr. Alptekin Temizel, atemizel@metu.edu.tr
Teaching Assistant: Ayberk Aydın, aayberk@metu.edu.tr

Course Objective

The course covers several aspects of designing reliable and scalable machine learning systems for real-world deployment. It deals with development of production quality models and introduces the machine learning pipeline, concepts on machine learning system design and data engineering. It provides know-how on model development, and how to scale up the training for large models as well as evaluation, calibration and debugging of these models. Generation of reproducible models via experiment tracking tools and model versioning is also covered. Hardware platforms and frameworks for deployment are introduced, followed by basic deployment concepts, containerized deployment and testing.

Reference Material:

CS 329S: Machine Learning Systems Design - <https://stanford-cs329s.github.io/>

Reading Material:

Rules of Machine Learning: Best Practices for ML Engineering <https://developers.google.com/machine-learning/guides/rules-of-ml>

Challenges in Deploying Machine Learning: a Survey of Case Studies <https://arxiv.org/pdf/2011.09926.pdf>

Grade Distribution:

Assignments	45%
Final project	30%
5x Quizzes	25%

Deliverables

Documents and necessary files of the assignments must be uploaded to ODTUClass by students before the specified due dates.

University Policies

All students are **expected to obey** the university code of integrity and avoid academic dishonesty or plagiarism.

No	Date	
1	5 Oct	Introduction to the Course and Machine Learning Life-Cycle
2	12 Oct	Designing a Machine Learning System – I
		Main Requirements of Machine Learning Systems Reliability, scalability, maintainability, adaptability ML in research vs. in production Traditional software vs. ML software, ML Production Myths
3	19 Oct	Designing a Machine Learning System - II
		Batch vs. online, Edge vs. cloud computing, Offline vs. Online Learning Iterative Development Phases of ML Adoption
4	26 Oct	Versioning and Experiment Tracking
		Experiment tracking tools Data versioning ML pipeline versioning Continuous integration/continuous delivery for ML
5	2 Nov	Hands-on Lab
		Experiment tracking Data versioning
6	9 Nov	Data Engineering
		Data centric approach Data basics and data formats Creating training datasets, labelling Semi-supervised and self-supervised learning
7	16 Nov	Data Engineering-II
		Sampling, Class imbalance problems Data Augmentation Data Leakage Data Analysis with FiftyOne
8	23 Nov	Model Development and Training
		Model Selection AutoML, Neural Architecture Search Optimizers Model Calibration
9	30 Nov	Model Optimization for Deployment
		Model compression Quantization Pruning Knowledge distillation
10	7 Dec	Case Study – Autonomous Driving
		Invited speaker from industry
11	14 Dec	Training and Deployment Platforms and Packaging
		GPUs, TPUs IoT devices and TinyML Packaging
12	21 Dec	Deployment Frameworks
		TensorRT Triton inference server Google Cloud Platform (GCP)
13	28 Dec	Evaluation
		Debugging, System evaluation and testing Data testing, profiling and visualization Benchmarking Perturbation evaluation, ablation study
14	4 Jan	Case Study – Deep Learning Solutions for Retail Stores
		Invited speaker from industry