The Hong Kong University of Science and Technology (Guangzhou)

UG Course Syllabus Template

[Course Title] Machine Learning Systems

[Course Code] DSAA 4012

[No. of Credits] 3

[Any pre-/co-requisites] DSAA 2011 Machine Learning, DSAA 2042 Computer Architecture and Systems

Name: Huayi DUAN

Email: huayiduan@hkust-gz.edu.cn

Office Hours: TBD, W4 530

Course Description

This advanced undergraduate course introduces how modern machine learning models are represented, trained, and deployed in real computer systems. It focuses on core ideas needed to make ML work efficiently and reliably at scale: data organization and representation, parallel and distributed computing basics, deep learning system design, distributed training, model serving, and system-level considerations for advanced, large models. The course also covers key aspects of trustworthy ML systems, including fairness, interpretability, privacy and security, and discusses how these issues arise in end-to-end ML workflows.

Instructional methods include weekly lectures, guided in-class discussions, and a small group project that designs and analyzes a modest ML system. Practical exposure is primarily conceptual and design-oriented, with limited but illustrative hands-on components.

Intended Learning Outcomes (ILOs)

By the end of this course, students should be able to:

1. **Explain** core design principles behind modern machine learning systems, including data organization, deep learning system architecture and design, distributed training, and efficient inference.
2. **Organize** and **prepare** data for model training and inference using appropriate storage, preprocessing, and representation techniques.
3. **Apply** basic parallel and distributed programming concepts to reason about how simple ML workloads can be scaled.
4. **Describe** and **analyze** key system-level techniques used in deep learning and advanced, large models, and reason about their performance and resource trade-offs.
5. **Identify** and **discuss** basic techniques for fairness, interpretability, privacy, and security in ML systems, and recognize where they apply in an end-to-end pipeline.
6. **Design** and **justify** an end-to-end ML workflow, including data preparation, training, deployment, and serving, for a given problem scenario.
7. **Collaborat**e in a team to design, partly implement, and evaluate a small ML system, and communicate the design and results clearly.

Assessment and Grading

This course will be assessed using criterion-referencing and grades will not be assigned using a curve. Detailed rubrics are provided below, outlining the criteria used for evaluation.

Assessments:

Assessment Task	Contribution to Overall Course grade (%)	Due date
Mid-Term	20%	dd/mm/yyyy *
In-course essay	10%	dd/mm/yyyy *
Group Project	10%	dd/mm/yyyy *
Final examination	60%	dd/mm/yyyy

* Assessment marks for individual assessed tasks will be released within two weeks of the due date.

Mapping of Course ILOs to Assessment Tasks

Assessed Task	Mapped ILOs	Explanation
Assignments	ILO 1, ILO 2, ILO 3, ILO 4, ILO 5	Assignments assess understanding of ML system principles and components (ILO 1), reasoning about data organization and preparation (ILO 2), basic parallel and distributed concepts (ILO 3), system-level techniques and trade-offs (ILO 4), and introductory reasoning about fairness, interpretability, privacy, and security (ILO 5).
Group Project	ILO 2, ILO 3, ILO 4, ILO 6, ILO 7	The project evaluates the ability to design an ML pipeline including data preparation (ILO 2), consider scaling aspects (ILO 3), analyze system-level techniques in context (ILO 4), justify end-to-end design choices (ILO 6), and work effectively in a team to partly implement and evaluate a small ML system (ILO 7).
Final Examination	ILO 1, ILO 2, ILO 3, ILO 4, ILO 5, ILO 6	The final exam provides a holistic assessment of students' ability to explain ML system design principles (ILO 1), reason about data and pipelines (ILO 2), discuss parallel and distributed concepts (ILO 3), analyze system-level techniques for deep learning and advanced models (ILO 4), incorporate trustworthy ML considerations (ILO 5), and design or critique end-to-end ML workflows (ILO 6).

Grading Rubrics

Detailed rubrics for assignments, the group project, and the final examination will be provided on Canvas. These rubrics will specify criteria such as:

- **Conceptual Understanding**
  Accuracy and depth of explanations; appropriate use of ML systems concepts and terminology.

- **Technical and Analytical Quality**
  Quality of reasoning about system design, trade-offs, and constraints; where applicable, clarity and correctness of any pseudo-code, design diagrams, or small-scale implementation.

- **Reasoning and Justification**
  Ability to compare alternatives, articulate trade-offs, and justify design decisions in writing and/or presentations.

- **Communication**
  Organization, clarity, and coherence of written work and oral presentations.

- **Teamwork**
  Evidence of shared contribution, coordination, and professional collaboration.

Students are encouraged to consult these rubrics before submitting work.

Final Grade Descriptors:

Grades	Short Description	Elaboration on subject grading description
A	Excellent Performance	Demonstrates a strong, well-connected understanding of ML systems concepts and methods. Provides accurate analyses, designs sensible small-scale systems, and clearly explains trade-offs. Work often shows insight beyond the minimum requirements.
B	Good Performance	Shows good understanding of the main ideas and techniques. Applies concepts correctly in typical situations and provides mostly clear explanations. There may be minor gaps in detail or reasoning, but overall performance is solid.
C	Satisfactory Performance	Possesses adequate knowledge of core material and can handle standard tasks with guidance. Explanations and designs meet the basic expectations of the course but may lack depth, completeness, or consistency.
D	Marginal Pass	Demonstrates only threshold understanding of key concepts and struggles to apply them without substantial support. Work meets minimum expectations but shows limited readiness for further study in this area.
F	Fail	Shows insufficient understanding of course concepts or frequent serious errors. Unable to carry out basic analyses or system design tasks at the required level. Does not meet the minimum standards for passing.

Course AI Policy

Students **may** use generative AI tools (e.g., ChatGPT, Copilot) for:

- Clarifying concepts (e.g., “Explain data parallelism in simple terms”).
- Getting help debugging small, specific pieces of code.
- Brainstorming or outlining ideas, followed by the student’s own development and verification of the final content.

Students **should not**:

- Use generative AI to produce full or near-complete assignment solutions, project reports, or exam answers.
- Submit AI-generated content without meaningful modification and understanding.
- Use AI tools to circumvent learning or misrepresent authorship.

Any use of generative AI must be **openly acknowledged** in an appendix or comment (e.g., “Used ChatGPT to clarify concept X; prompt: …”). Work that appears to rely substantially on AI without understanding may be treated as academic misconduct in accordance with university policy.

Communication and Feedback

- Assessment marks and feedback will be released via **Canvas** within two weeks of the submission or examination date.
- Feedback will normally include:
  - Main strengths
  - Key areas for improvement
  - Concrete suggestions for further study or refinement

Students who have questions about their marks or feedback should contact the instructor within **five working days** after feedback is released.

Resubmission Policy

- Resubmission of assessed work is **not normally permitted**.
- In exceptional, documented circumstances (e.g., serious personal issues), the instructor may allow resubmission or an alternative assessment in line with departmental and university regulations.
- Where resubmission is allowed, revised work must be submitted within **one week** of approval, and a cap on the maximum attainable grade may apply (e.g., capped at a pass mark).

Recommended Texts and Materials

- Dive into Deep Learning: https://d2l.ai
- Machine Learning Systems: Principles and Practices of Engineering Artificially Intelligent Systems: https://www.mlsysbook.ai

Academic Integrity

Students are expected to adhere to the university’s academic integrity policy. Students are expected to uphold HKUST(GZ)’s Academic Honor Code and to maintain the highest standards of academic integrity. The University has zero tolerance of academic misconduct. Please refer to Regulations for Academic Integrity and Student Conduct for the University’s definition of plagiarism and ways to avoid cheating and plagiarism.

Tentative Schedule

Lecture	Topics	ILOs
1	Machine Learning and Data Systems Fundamentals	ILO 1, ILO 2, ILO 6
2	Data Preparation and Representation	ILO 2, ILO 6
3	Computer Systems, Parallel and Distributed Programming	ILO 1, ILO 3
4	Deep Learning System I: Overview and Overall Design	ILO 1, ILO 4, ILO 6
5	Deep Learning System II: Gradient Computation and Optimization	ILO 1, ILO 3, ILO 4
6	Deep Learning System III: Hardware Acceleration	ILO 1, ILO 3, ILO 4

7	Distributed Training and Communication Protocols	ILO 1, ILO 3, ILO 4
8	Model Serving and Efficient Inference	ILO 1, ILO 2, ILO 4, ILO 6
9	Efficient Training and Inference of GNNs	ILO 1, ILO 3, ILO 4
10	Efficient Training and Inference of Large Models	ILO 1, ILO 3, ILO 4
11	Trustworthy ML Systems I: Fairness, Interpretation, and Debugging	ILO 1, ILO 5, ILO 6
12	Trustworthy ML Systems I: Privacy and Security	ILO 1, ILO 5
13	Course Review	ILO 1, ILO 4, ILO 5, ILO 6