# Fall 2024
# Learning from Data

Instructor: Yang Li

TA: Weida Wang, Yuabo Tang, Tong Wu, Chengfeng Wu

## Description

This introductory course gives an overview of many concepts, techniques, and algorithms in machine learning, beginning with topics such as logistic regression and SVM and ending up with more recent topics such as deep neural networks and reinforcement learning. The course will give the student the basic ideas and intuition behind modern machine learning methods as well as a bit more formal understanding of how, why, and when they work. The underlying theme in the course is statistical inference as it provides the foundation for most of the methods covered.

The course includes 2.5 hours of **lectures** (by the Instructor) each Friday from 9:50 am to 12:15 pm, and 40 mins of **optional recitation session** (by the TAs) on Friday from 17:00 pm to 17:40 pm online. Homework includes both written exercises and programming exercises. There will be a written midterm exam and a final course project includes a proposal, final presentation, and report.

## Intended Students

The course is geared towards students who are interested in understanding machine learning algorithms, and to prepare them for research in this field. One of the objectives of the course is to understand the fundamental perspectives and develop solid connections between mathematical theory and learning systems.

## Prerequisites

Basic concepts in vector calculus, probability theory, and linear algebra. Basic knowledge of Python programming.

## Problem Sets

There will be a total of 5 written and 4 programming problem sets, due roughly every two weeks. The content of the problem sets will vary from theoretical questions to more applied problems. You are encouraged to collaborate with other students while solving the problems

but you will have to turn in your own solutions. **Homework plagiarism will not be tolerated. If you collaborate with others in any part of the homework, you must acknowledge them in your submission. If you get homework help from an online resource (e.g. github, ChatGPT), you must also give credit to the source.**

**Final Project**

The final project for the course will involve using applied techniques on learning related applications or theoretical explorations of machine learning. The instructor will provide a list of suggested datasets for students to chose from, but students are encouraged to find their own dataset or topic, subject to the approval of the instructor.

The final project will be done in groups of two. Each group will submit a written report and optionally present in class[1].

**Grading**

Your overall grade will be determined roughly as follows:

| ACTIVITIES | PERCENTAGES |
|---|---|
| Midterm | 15 % |
| Final Project | 25 % |
| Continuous performance assessment | 3 % |
| Problem sets (written & programming) | 57 % |

Each person will have **two free chances** to turn in a late homework assignment (except for the final project). Each late homework must be turned in within three days after its original deadline. Beyond the three-day grace period, late homework will not be graded. (Talk to the instructor if you have special circumstances.) **Continuous assessment** includes the regular evaluation of a student's performance in class, such as participation, engagement, and contribution during lessons. It reflects the student's consistent effort and understanding throughout the course, rather than relying solely on final exams, problem sets, or major projects.

**Textbook**

The primary course materials are the lecture slides. You can also read the following reference materials:
- (Recommended) Machine Learning Lecture Notes by Andrew Ng: `https://cs229.stanford.edu/main_notes.pdf`
- Pattern Recognition and Machine Learning, 2nd Edition, by Christopher Bishop

---

[1]We may not have enough time for all groups to present during the final week. The final presentation method will be up to discussion later.

**Course Syllubus**

Note: PA stands for "programming assignment"; WA stands for "written assignment".

| Date | Topic | Homework release |
|------|-------|------------------|
| 9/12 | Review Session (optional) | WA0 (don't need to hand in) |
| 9/13 | Introduction | |
| 9/20 | Supervised Learning I<br>• Linear regression<br>• Logistic regression<br>• Generalized linear model | WA1 |
| 9/27 | Supervised Learning II<br>• Support vector machines | PA1 |
| 10/4 | National Day | |
| 10/11 | Supervised Learning III<br>• Generative model: GDA<br>• Generative model: naive Bayesian model | WA2 |
| 10/18 | Supervised Learning IV<br>• Deep neural networks | PA2 |
| 10/25 | Supervised Learning V<br>• Bias and variance trade off<br>• Model selection | WA3 |
| 11/1 | Midterm | |
| 11/8 | Unsupervised Learning I<br>• K-means clustering<br>• Principal component analysis | WA4, Final Project |
| 11/15 | Unsupervised Learning II<br>• Independent component analysis<br>• Canonical component analysis | PA3 |

| | | |
|---|---|---|
| 11/22 | Unsupervised Learning III<br>• Spectral Clustering<br>• Laplacian Eigenmaps | WA5 |
| 11/29 | Reinforcement Learning<br>• MDP, value and policy iterations<br>• Deep Q-learning | PA4 |
| 12/6 | Learning Theory Primer<br>• Generalization error<br>• VC dimension | |
| 12/13 | History of deep learning (I): Image processing and computer vision<br>• Coming of the data-driven age<br>• AlexNet, CNN, ResNet, Transformer | |
| 12/20 | History of deep learning (II)<br>• Large language models, BERT, GPT, scaling law<br>• Multimodal learning<br>• Transfer learning | |
| 12/27 | Final Project Presentations | |