Statistical Methods in Machine Learning (36-708)

Spring 2025

Schedule: Tuesday/Thursday, 2:00-3:20pm, WEH 5421

Instructor: Edward Kennedy (edward@stat.cmu.edu)
Instructor Office Hours: 3:30-4:30 Tuesdays at Baker 132J

Teaching Assistant: Siddhaarth Sarkar (siddhaas@andrew.cmu.edu)
TA Office Hours: 3:00-4:00 at Wean Hall 3711

Objectives

The goal of this course is to study some modern machine learning methods, and gain an understanding of the theory underlying the methods. Upon completing this course, you should have a good roadmap in mind of different goals in ML, what methods can be used to achieve these goals, and what guarantees the methods can provide, under what conditions. Some crucial concepts you should be able to explain include: the bias-variance trade-off, risk guarantees, training error optimism, uncertainty quantification, minimax optimality, etc.

Prerequisites

Before enrolling in this course, students should:

- take 36-705 (Intermediate Statistics),
- have proficiency in programming in R (or a similar language),
- be comfortable with linear algebra, probability, and calculus.

Notes & Texts

We will mostly use course notes and readings from the literature. Here are some textbooks that will also play a role, and which are useful references for augmenting your understanding:

- Devroye, Gyorfi, Lugosi: "A probabilistic theory of pattern recognition"
- Gyorfi, Kohler, Krzyzak, Walk: "A distribution-free theory of nonparametric regression"
- Hastie, Tibshirani, Friedman: "Elements of statistical learning"
- Mohri, Rostamizadeh, Talwalkar: "Foundations of machine learning"
- Wasserman: "All of nonparametric statistics"

Course Outline

Some planned topics for the course include:

- k-nearest neighbors & kernel methods
- local polynomials & series regression
- density estimation
- generalized additive models
- random forests
- neural networks
- cross-validation
- aggregation
- classification
- conformal prediction
- inference
- clustering
- causality
- functional estimation
- minimax optimality

Other topics might include high-dimensional regression, variable importance, mixture models, dimension reduction, semisupervised learning, distribution shift, etc.

The course title "The ABCDE of statistical methods in machine learning" is a holdover from a previous semester, but you can think of it as standing for: Algorithm design, Bias-variance thinking, unCertainty quantification, Data analysis, and Explainability and interpretability, which are five core aspects of the course .

Grading

Grades will be weighted as follows:

Homework	50%
Final Project	30%
Recaps	10%
Participation	5%
Quizzes	5%

Homework

Problem sets will be assigned approximately once per month, and should be downloaded and submitted via Canvas/gradescope.

Canvas: https://canvas.cmu.edu/courses/45641/
Gradescope: https://www.gradescope.com/courses/942280

Clear writing and presentation are important and will be weighted highly along with correct solutions. Late homeworks will not be accepted, apart from exceptional circumstances where prior permission is given.

Final Project

There will be a final project due towards the end of the semester. It will consist of a review of a topic or paper(s) of interest; this could be on a generalization/extension of something we covered in class, or a new topic altogether.

Quizzes

There will be short in-class quizzes, approximately once per month. These are partly to assess understanding and partly to help ease gaps between homeworks. They will mostly be multiple-choice or true-false, but you should still make sure to prepare for them, as they will probe important concepts from lecture.

Recaps

Each student will sign up to do a 5-10 minute recap presentation of the previous lecture's material, touching on the most important concepts and refreshing memories on some details. These presentations should just be presented on the board, as mini "chalk talks" (i.e., no slides required).

Participation

Attendance and in-class participation is an important part of learning. You will get the entire 5% if during at least 50% of lectures you ask substantive non-trivial questions. Office hours discussions can also play a role, but it is particularly important for us to have discussions during lectures. Remember: if you have doubts about something we are discussing in class, it is likely that someone else does as well; further, actively engaging with material (as opposed to passively listening and not clarifying questions you have) will improve learning outcomes.

Plagiarism

Plagiarism is a serious offense at CMU. See the policy: www.cmu.edu/academic-integrity/plagiarism/