Math 465 Introduction to High Dimensional Data Analysis, Fall 2023

Time: WF 11:45AM - 1:00PM
Location: Gross 318

Instructor: Xiuyuan Cheng, Jiajia Yu

Course Overview

The recent developments in the field of data analysis and machine learning have thrived at the intersection of electrical engineering, computer science, statistics, and mathematics. Typical questions are like: how to design an algorithm to identify clustering in the dataset without labels, and how to embed high dimensional data, like images and audio signals, to a low-dimensional space for visualization and downstream tasks, and so on. To answer these questions, we need to set up some fundamental concepts in analysis, probability and computation, so that we can understand the modern data analysis methods as well as to develop new ones.

This course will introduce and explore basic mathematical concepts underlying the practice and theory of various data analysis methodologies and algorithms, and give an introduction to the theoretical and computational tools in the field from a mathematical point of view. The course will start with a review of fundamentals in statistical learning, such as the concept of bias-variance trade-off. Then a sequence of selected topics will be introduced, including dimension reduction, graph-based methods, large random graphs, and neural networks. The course will mix theory and hands-on experience assuming entry-level mathematical background and pave the path to the study of more advanced methods in the field. The course project provides an opportunity to advance the study on a topic of your selection.

Required equipment: a laptop (Windows+WSL / Mac / Linux), and we will give guidance on needed installations in class (mainly Python).

Textbooks

Reading will be assigned as the lecture goes on, and general references are

[HTF09] The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, by Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Springer (2009). ISBN 9780387848570.

[V18] High dimensional probability: An Introduction with Applications in Data Science, by Roman Vershynin. Cambridge University Press (2018). ISBN 9781108231596.

[GBC16] Deep Learning, by Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Cambridge: MIT press (2016). ISBN 9780262035613.

Grading Policy

- Homework (40%)

  Assigned weekly after the lecture and to be handed in the following week. Depending on the content of the lecture, the homework assignment may have a programming part, in which case both codes and a written summary of experimental results (preferably in latex/markdown or any typed-up format) are required. Homework is graded both for accuracy, clarity, and

completeness. For some questions, e.g., those that are broken down into branches, partial credit may be granted depending on the quality and completeness of the handed-in solution. Efforts to give meaningful partial solutions, in homeworks and exams, are always encouraged, but the grader has the right to determine the points to be credited for the handed-in solutions. Late homework may be accepted with discounted credits.

Group work and collaborative efforts for homework are encouraged in this course. However, each handed-in problem set must be independent work. You should name the students or other people with whom you had significant discussions about the problems, if any, on your hand-in solutions for homework. You should present a complete written solution/code to each problem, in your own words, without reference to the written solution of any other person. Any written sources, such as books and online sources other than the course textbook, that contribute significantly to your understanding of the problem should also be cited. Homework or exam credits will not be given in case of violating the policies.

- Two written exams (20%+20%)

  We will have two in-class 75 minutes written exams. The written exams will be closed book and no use of computational aids is allowed.

- Course report (20% + bonus 5%)

  The course project can be either (i) a review report on a selected topic, or (ii) a technical report of a small research project of data analysis. You can select a topic related to the content of the class.

  (i) For the review report, you will summarize the main results in the field of study, organize and cite the related literature, and present the content in a clear, correct, and organized way.

  (ii) For the technical report, you will design the content of your project, and present your results (theoretical, programming, or both) in the report.

  For either type of the report, you will also prepare slides for a short presentation in class. The evaluation will be based on both the in-class presentation and the final written report submitted. We will discuss and provide more guidance in the class about the course report. The bonus 5% points are granted for exceptionally work which are innovative.

  All work in homework assignments and exams must be independent work. Grades will be assigned depending on both the total points and the ranking of the performance among the class. If the total point is 59% and below, a grade of D or F can be expected. If the total points is 90% and above, a grade of A or A- can be expected. The grade A+ will be granted for exceptionally excellent performance.

Tentative schedule

- Week 1- Week 2 till Fri Sep 8

  Module 1: Data analysis fundamentals. Supervised and unsupervised learning. Review of preliminaries in linear algebra and probability.

- Week 3- Week 4 till Fri Sep 22

  Module 2: Principle Component Analysis (PCA), Multi-dimensional Scaling, random walks on graphs, graph Laplacian

- Week 5- Week 7 till Fri Oct 13

  Module 3: Dimension reduction, manifold learning, spectral clustering/embedding

  First Exam is on Fri Oct 6
  Report topic selection due Oct 13
  (Fall break Oct 13-18)

- Week 8- Week 9 till Fri Oct 27

  Module 4: Kernel and spectral methods, concentration of measure, random graph, consistency of spectral clustering

- Week 10- Week 11 till Fri Nov 10

  Module 5: Linear and non-linear classifiers, logistic regression, kNN classifier, neural network classification, Stochastic Gradient Descent

  Report proposal/initial slides due Nov 3

- Week 12- Week 13 till Fri Dec 1

  Module 6: Deep neural networks and generative models (optional); guidance on course report presentation

  Second Exam is on Nov 17
  (Thanksgiving break Nov 22-26)

- Week 14 till Fri Dec 8

  Course wrap-up and review, guidance on course report, student presentation TBD