Math 465 Introduction to High Dim Data Analysis, Fall 2024

Class: WF 4:40PM – 5:55PM @ Gross Hall 103
Instructor: Jiajia Yu
E-mail:jiajia.yu@duke.edu
Office hours: WF 6:00PM – 7:00PM or by appointment

Course Overview

This course introduces tools in high dimensional data analysis, including linear and nonlinear dimension reduction, utilizing sparsity and low-rank structure, and theories in probability.

Learning Objectives

By the end of this course, students are expected to:

- (General) know the difficulty of high dimensional data related tasks, and common concepts and techniques to deal with them
- (Dimension reduction) know and understand the mechanisms of PCA, MDS, k-means clustering, isomap, spectral methods, diffusion map and t-SNE
- (Sparsity) know how to enforce sparsity, how to relax the original problem to a convex one and how to solve the convex problem
- (Kernel method) know the difference between parametric and non-parametric models, why and how to use kernel methods
- (High dimensional statistics) know why and how to develop concentration inequalities and its applications

Textbooks

There are no required textbooks. As necessary, links to papers and other reading materials will be provided. Some general references are:

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

[WM22] High-dimensional data analysis with low-dimensional models: principles, computation, and applications, by John Wright and Yi Ma. Cambridge University Press (2022). ISBN 9781108489737
free pre-publication version available at https://book-wright-ma.github.io/

Prerequisites

Math: multivariable calculus, linear algebra, and probability.
We will discuss most of the necessary mathematical techniques as we proceed through the semester. But basic knowledge on these three courses is needed to understand the lecture.

Coding: python.
There will be programming questions in homework and/or reports. The instructor and TA/grader are not responsible for helping debugging students' programming.

Attendance

Attendance in class is a vital part of the learning process. Regular class attendance is strongly encouraged. It is the student's responsibility to keep informed of any announcements, syllabus adjustments, or policy changes made during scheduled classes.

Grading Policy

Letter grades will be computed from the semester average. Maximum lower bound cutoffs for A, B, C and D grades are 93%, 83%, 73% and 60% respectively. These bounds may be moved lower at the instructor's discretion.

- Homework (40%)

  Homework will be assigned weekly. 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 are always encouraged, but the grader has the right to determine the points to be credited for the handed-in solutions.

- 1 midterm exam and 1 in-class final exam (15%+15%)

  We will have two in-class 75 minutes written exam. The written exams will be closed book and no use of computational aids is allowed. There is NO make-up exams except for university approved excuses.

- Course project (30%)

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

  - 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.
  - 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 report, you will also prepare one slide for a lightning presentation in class. The evaluation will be based on topic selection (2%), proposal (5%), the in-class presentation (8%) and the final written report (15%). We will discuss and provide more guidance in the class about the course report.

  - topic selection: 1 sentence or 1 paragraph, describe the topic you want to study
  - proposal: 1-2 pages, describe the background/motivation of your study, the specific problem you want to study and a rough plan
  - presentation: details TBD
  - final report: no longer than 10 pages, a formal technical report describing your study

- Collaboration policy

  Group work and collaborative efforts for homework and projects are encouraged in this course. However, each handed-in problem set, and report 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 and report credits will not be given in case of violating the policies.

- Late policy

  Students have three free "late" days they can use on homework throughout the semester, with *at most one used for any one homework*. A late day is defined as any whole or partial day after the submission deadline. These free late days are to be used for minor illnesses, balancing other course work, conflicts associated with traveling, problems with your computer, etc. After using up the three free late days, late homework will be penalized 10% per day, and no credit will be awarded once solutions are posted (which can be as soon as the next class). Homework submitted on the due date but after the time specified will be penalized 5%. Late submissions of project reports may result in a complete forfeiture of credit for the report section.

- AI tools policy

  Students are encouraged to explore AI tools in homework and projects while maintaining academic integrity. The use must be properly documented and credited. For example, text or answers generated using ChatGPT-3 should include a citation such as: "Chat-GPT-3. (YYYY, Month DD of query). "Text of your query." Generated using OpenAI. https://chat.openai.com/" Material generated using other tools should follow a similar citation convention. Please be aware that AI does not always give the right answer or properly cite the references. Students using AI tools have the responsibility to make sure their final hand in homework and project report do not violate the academic integrity and include students' independent understanding and intellectual contributions.

Academic Integrity

Students are responsible for informing themselves of Duke's policies regarding academic integrity. Students found in violation of the code are subject to penalties ranging from loss of credit for work involved to a grade of F in the course, and possible risk of suspension or probation. The academic integrity policy will be enforced in all areas of the course, including homework, exams, and projects. If you have any questions concerning this policy before submitting homework or project reports, please ask for clarification.

Topic and Schedule (Tentative)

Dimension reduction and manifold learning
- SVD and PCA
- MDS, graph data and ISOMAP
- spectral clustering and spectral embedding
- diffusion map
- t-SNE

Regularization and Sparsity
- Linear regression, ridge regression and LASSO
- sparse signal recovery
- low-rank matrix recovery

Kernel methods

Randomness
- concentration inequality
- coordinate descent
- stochastic gradient descent

Week, date	Wed	Thur	Fri	Note
1, 0828, 0830	Examples of HD data, difficulties, and strategies for analysis		SVD and PCA	0826 fall begin
2, 0904, 0906	Metric and non-metric MDS Classical MDS and PCA Graph data, classical MDS and ISOMAP	HW1 due	Spectral clustering and spectral embedding	0906 drop/add ends
3, 0911, 0913	diffusion map and PageRank	HW2 due	t-SNE	
4, 0918, 0920	Linear regression, ridge regression and LASSO Regularization	HW3 due	Sparsity, relaxation and l_1 regularization Convexity Proximal gradient method	
5, 0925, 0927	Sparse signal models and examples Noisy observations and approximate sparsity	HW4 due Project topic selection due	Recitation HW1-3, project proposal guidance Correctness results for sparse signal models	
6, 1002, 1004	low-rank modeling and examples convex relaxation and nuclear norm	HW5 due	Sparse PCA and robust PCA	

7, 1009, 1011	Limitation of convex relaxation and some nonconvex methods Kernel methods	HW6 due	Recitation HW4-5 midterm review	1012-1015 fall break
8, 1016, 1018	Concentration inequality	Project proposal due	Concentration inequality	
9, 1023, 1025	Midterm cover HW1-5		TBD	
10, 1030, 1101	Concentration of norm	HW7 due	The Johnson-Lindenstrauss Lemma	
11, 1106, 1108	Random projection, RIP of Gaussian random matrices	HW8 due	Optimization algorithms, convergence rate, and gradient descent	1108 last day to withdraw with W
12, 1113, 1115	Coordinate descent	HW9 due	Stochastic gradient descent	
13, 1120, 1122	TBD		Recitation 7-9, presentation and report guidance Final review	
1127, 1129				1127-1201 Thanksgiving recess
14, 1204, 1206	presentation		presentation	1206 last day of class 1208 project report due
				1211-1216 Final cover HW 1-9

*Note: This syllabus is subject to change as the semester progresses.