New Jersey Institute of Technology
Ying Wu College of Computing
Computer Science Department

Seminar in Computer Science II
Special Topic: Dimensionality and Scalability in AI

Code: SP25-CS786-854
Mode: Synchronous online (Zoom)
Time: Thu 6:00pm–8:50pm

Instructor: Michael Houle
Webpage: https://people.njit.edu/faculty/meh43
Office: GITC 4317D (Newark)
Email: michael.houle@njit.edu (or directly on Canvas)

Note: Your messages will usually be answered by the end of the next day. Grades for all items will generally be posted during the week after their due date. For issues with your grades, contact the instructor directly.

Office Hours: Tue 5:30pm–8:30pm and Thu 2:00pm–5:00pm, online (Zoom).
Reserve an online appointment slot by following this calendar link. Please try to do so at least one day in advance. Other appointment times can also be arranged by email.

Teaching Assistant / Grader: None

Course Description

[From the NJIT catalog]: This seminar course examines in-depth recent research literature in an area of computer science. The selected topic and course prerequisites are announced before the beginning of the semester.

[Instructor's description]: In the AI disciplines of machine learning & deep learning, and in related areas such as data mining and search & indexing, the efficiency and effectiveness of implementations depend heavily on the interplay between data similarity measures and the features representing data objects. The number of features, or *dimensionality*, often fails to capture the true complexity of the data. For this reason, data complexity is usually characterized by some form of *intrinsic dimensionality* (ID), such as the number of latent features needed to represent the full dataset without much loss of information, or the dimension of a manifold that best approximates the dataset. However, these *global* formulations of ID often fail to align with user interests, which may instead focus on specific queries, clusters, or latent-space representations of individual training and test examples. In such cases, the relevant latent

features vary across points and clusters, necessitating a theoretical and practical focus on local characterizations of ID. This course is concerned with emerging research in the area of *local intrinsic dimensionality* characterization and estimation, and its applications in disciplines that rely on similarity information as the basis of data modeling, analysis, and processing. You will become familiar with the theoretical foundations of local ID, as well as the practical issues that surround its estimation and application in machine learning, deep learning, similarity search, and other contexts involving similarity-based analysis.

Prerequisites

This research-oriented course does not have other courses as prerequisites. However, students are assumed to have completed one or more graduate-level courses in any of the following disciplines: machine learning, deep learning, data mining, algorithmics, or mathematics & statistics.

At a minimum, a certain level of mathematical maturity in basic calculus, linear algebra, and probability and statistics is required. Familiarity with the following mathematical concepts and their notation is essential for success in this course.

Calculus	Linear Algebra	Probability and Statistics
Limits and continuity	Vectors and vector notation	Basic probability and selection
Basic notion of derivatives	Lines, planes, hyperplanes	Bayes' theorem*
Differentiation of common functions	Normal vectors	Maximum likelihood*
	Vector norm	Random variables
Differentiation rules: product, quotient, chain	Dot product	Expectation
	Orthogonality and projection	Mean, variance, deviation
L'Hôpital's rule	Linear transformation	Error, bias
Partial differentiation	Matrix notation	Conditional probability & independence
Multivariate chain rule	Matrix addition & multiplication	Standardization of variables
Gradient	Matrix transpose, inverse	Distributions and sampling
Taylor series expansion	Matrix rank	Probability density function
Integral calculus	Eigenvalues and eigenvectors*	Cumulative distribution function
Integration of common functions	Basis vectors*	Uniform distribution
	Eigendecomposition*	Normal distribution (Gaussian)*
Integration by parts		Correlation and covariance*

* = a brief review of these topics will be given, when needed for the discussion.

The following free online materials are recommended for reviewing this background:

- Mathematics for Machine Learning

Course Textbooks

This course covers emerging research trends in an area for which there is as yet no textbook. The course will instead draw on the instructor's notes (in the form of presentation slides), research papers, online resources, and other supplementary materials as needed.

Learning Outcomes

By the end of the course, you will be able to:

a. Identify the issues surrounding dimensionality in AI-related disciplines.
b. Evaluate the quality of online resources related to intrinsic dimensionality.
c. Recognize problems amenable to local intrinsic dimensional (LID) analysis.
d. Describe and explain a wide variety of LID usages in AI-related disciplines.
e. Apply the theory of LID in novel situations.
f. Evaluate the performance of LID-aware models.
g. Modify AI models in order to improve their performance through LID-awareness.

Coursework, Assessment, and Related Outcomes

Participation [10%]
Students are expected to participate in classroom activities throughout the term. They will receive credit for attendance (0.5% per class) and for engagement in discussions in class and on Canvas forums (3%). Special consideration will be given for cases in which the student is absent from class for valid, documented reasons.
[Outcomes: a, c, d, e, g]

Theory Assignments [50%]
There will be five assignments on the concepts, theory, and practice surrounding intrinsic dimensional modeling and its applications. Some assignments may offer the students to choose an empirical investigation (implemented in Python) as an alternative to certain questions. Students will typically have one to two weeks to complete each assignment. Submission will be via Canvas.
[Outcomes: c, d, e, f, g]

Project [40%]
Students will conduct an individual project involving intrinsic dimensional analysis in the broader context of AI and related fields. This can be a theoretical investigation, a practical application, and/or an empirical investigation (implemented in Python). Students are encouraged to choose a topic that relates to their own graduate research, wherever possible. The project will have three milestones: a detailed proposal (worth 10%), a final report (worth 25%), and a video presentation (worth 5%). Submission will be via Canvas. A selection of the best project submissions will be invited for live presentation and Q&A during the Week 14 class.
[Outcomes: b, c, e, f]

Assignment Due Dates

All assignments and milestones are due during the weeks indicated in the table below.

Assignments [50%] Due on Sundays at 23:59	Project [40%] Due on Wednesdays at 23:59
Week 4: #1 [10%] Week 7: #2 [10%] Week 9: #3 [10%] Week 12: #4 [10%] Week 15: #5 [10%]	Week 10: Proposal [10%] Week 14: Final Report [25%] Week 14: Video Presentation [5%]

Note that the week numbering does not count the Study Break. Also, note that there is no class in Week 10 (Thursday April 3 being Wellness Day).

Course Topic Schedule

Week 1	Features, Similarity, and Search (Introduction)
Week 2	Dimensionality Reduction and the Manifold Model of Data
Week 3	The Curse of Dimensionality and Intrinsic Dimensionality
Week 4	Generalized Expansion Dimension and Similarity Search
Week 5	The Theory of Local Intrinsic Dimensionality
Week 6	LID and Extreme Value Theory
Week 7	Estimation of LID
Week 8	Anomaly Detection
Week 9	Data Perturbation and Dimensionality
Week 10	NO CLASS (Wellness Day)
Week 11	Entropy, Divergence, and Alignment of Learned Representations
Week 12	Estimation of Convergence Order
Week 13	LID, Dynamical Systems, and Deep Learning
Week 14	Summary and Future Directions
Week 15	Selected Student Project Presentations

Grading Policies

Letter Grades
In accordance with the graduate grade legend, the raw total percentage assessment score will be converted to a final letter grade that will appear on your transcript. The conversion table for this course is:

Letter Grade	Percentage Range
A	90 — 100
B+	80 — <90
B	70 — <80
C+	65 — <70
C	60 — <65
F	<60

In cases where the project outcomes are judged to be original contributions meeting the publication standard of reputable international research conferences or journals, the letter grade may be raised one level higher than what is suggested by the above conversion table.

Incomplete

A grade of **I** (incomplete) is given in rare cases where work cannot be completed during the semester due to documented long-term illness or unexpected absence for other serious reasons. A student requesting special consideration should be in good standing (i.e. with a passing grade on coursework submitted before the absence). When special consideration is granted, the student receives a provisional **I** if there is no other way to make up for the documented lost time; in such cases, an email with a timeline for makeup work will be sent to the student. Note that according to NJIT regulations, an **I** must always be resolved by the end of the following semester.

Late Submission Policy

Generally speaking, assignments, project milestones, and participation exercises will be accepted late without penalty, but only up until the time grading has begun, or solutions or other feedback have been released to members of the class. At that time, the class will be informed that submissions have closed, that no further submissions will be accepted, and that any missing work will be given a mark of zero. Students should be aware that grades, solutions, and/or feedback may be released at any time after the deadline, at the sole discretion of the lecturer and without prior warning. *The only way to ensure that work will be accepted without penalty is to submit it before the deadline*. No individual extensions will be granted. However, special consideration may be given in rare cases when a student is unable to complete an assignment for serious, unavoidable reasons — these must be communicated and documented promptly.

Grading Feedback

Assignment marks will in most cases be accompanied with class discussion of the solutions. Individual grading feedback will be given where appropriate. Further clarifications can be provided by contacting the instructor.

Grade Corrections

Check the grades in course work and report errors promptly. Please try and resolve any issue within one week of the grade notification.

Other Course Policies

Email
Use of your NJIT email or Canvas inbox is strongly encouraged.

Requesting Accommodations
If you need an accommodation due to a disability, please contact Marsha Williams-Nicholas, Associate Director of the Office of Accessibility Resources and Services, Kupfrian Hall 201 to discuss your specific needs. A Letter of Accommodation Eligibility from the office authorizing student accommodations is required.

NJIT Services for Students, Including Technical Support
Please follow this link.

Canvas Accessibility Statement
Please follow this link.

Collaboration on Assignments
You are expected to tackle all the problems **on your own**. However, some of the assignment problems may be quite challenging! For difficulties that persist, you are welcome to raise questions in the Canvas Discussion Forum, or talk to the instructor during office hours. In consulting with others, you are allowed to exchange general ideas and approaches only: unless you are given explicit permission to do so in the assignment statement, the full solutions themselves must be worked out by you alone.

Generative AI Tools and Other External Resources
Sometimes you may come across code, text or other helpful information online, or you may be able to generate it using AI tools such as ChatGPT or other Large Language Models (LLMs). In most cases, you will be allowed to integrate this information into your solution. However, if you do, you must always give the appropriate credit and citations (e.g. links) for the material you use (especially when you use the code and text you found online). In the case you use an LLM, you must say that you did so, and present the entire transcript of your 'conversation' with it, which should show what you asked and how you guided it, or were guided by it to the delivered solution. Your 'conversation' with it must be entirely yours, and sufficiently different from that of other students. Failure to give appropriate credit when using the work of others (whether human or AI) is considered plagiarism, and may lead to disciplinary action under NJIT's Academic Integrity policy (see below).

Statement on Academic Integrity

"Academic Integrity is the cornerstone of higher education and is central to the ideals of this course and the university. Cheating is strictly prohibited and devalues the degree that you are working on. As a member of the NJIT community, it is your responsibility to protect your educational investment by knowing and following the academic code of integrity policy that is found at: http://www5.njit.edu/policies/sites/policies/files/academic-integrity-code.pdf .

Please note that it is my professional obligation and responsibility to report any academic misconduct to the Dean of Students Office. ***Any student found in violation of the code by cheating, plagiarizing or using any online software inappropriately will result in disciplinary action. This may include a failing grade of F, and/or suspension or dismissal from the university.*** *If you have any questions about the code of Academic Integrity, please contact the Dean of Students Office at dos@njit.edu ."*