Hertie School

Fall Semester 2024

Course Syllabus

GRAD-Xoooo: Mathematics for Data Science

1. General information

Course Format	Onsite
Instructor(s)	Asya Magazinnik
Instructor's E-mail	a.magazinnik@hertie-school.org
Assistant	Amanda Slater; slater@hertie-school.org
Instructor's Office Hours	By appointment. Please contact Amanda Slater.

Link to Study, Examination and Admission Rules and MIA, MDS and MPP Module Handbooks

For information on **course room, times and session dates,** please consult the Course Plan on *MyStudies*.

Instructor Information:

Asya Magazinnik is the Professor of Social Data Science at the Hertie School. Her research interests include electoral geography, federalism, local politics, and law enforcement. She also works on political methodology, in particular at the intersection of causal inference and formal theory. Her work has appeared in the American Journal of Political Science, the Journal of Politics, and other outlets. Previously, she was an Assistant Professor of Political Science at MIT. She earned a PhD in Political Science from Princeton University in 2020 and holds an MPP from the Harris School at the University of Chicago.

2. Course Contents and Learning Objectives

Course contents:

This course aims to deliver a compact and tailored introduction to the core mathematical concepts of data science: probability theory, calculus, linear algebra, and statistics. We will establish and reinforce the foundation that will support your further studies in machine learning, deep learning, and causal inference, and which will inform how you conceptualize and analyze data. By the end of the course, students will (re)acquaint themselves with the basic principles of probability theory that structure the process that generates data; with the behavior of functions that describe relationships in data; with the rules of linear algebra that govern how to manipulate data; and, finally, with the tools of statistical learning that we use to extract information from our datasets.

Main learning objectives:

The objective of this course is to establish a strong common foundation of mathematical skills and concepts on which to build the rest of the MDS curriculum. We will place special emphasis on applications to real-world data analysis for public policy throughout the course.

Teaching style:

This course is taught with lectures that cover the main content of the course and small-group lab sessions that reinforce and extend the material. Because the purpose of this class is to build facility with a very large set of mathematical skills that build on one another, there will be regular assignments and hands-on tutorials to ensure ongoing learning.

Prerequisites:

This course assumes some basic familiarity with linear algebra, calculus, probability theory, and statistics, since it moves rather quickly through each of these topics for a first treatment. However, students who are seeing some of this material for the first time (or for the first time in a long time) can still succeed. The only absolute requirement is a **willingness to work hard**. Success in the course requires putting in consistent effort every week with course readings, assignments, and regular attendance of both lectures and lab sessions.

Diversity Statement:

In this classroom and at the Hertie School more broadly, we respect and value differences in race, gender, gender identity, ethnicity, age, physical and language abilities, culture, religion, sexual orientation, and other identities. Such differences contribute to the strength of our community, and upholding an atmosphere of inclusion and mutual respect is of the utmost importance.

This class will also contain a diversity of backgrounds and experiences with mathematics and data science. It is everyone's responsibility to promote and maintain an environment where all students feel comfortable asking questions and voicing their struggle. Mutually supporting one another will be at the heart of everyone's success in this class, and I hope that more advanced students will help their peers.

Students are encouraged to bring up any concerns relating to these policies, or any other hardships that they encounter, with me as early as possible, and I will make every effort to resolve them.

3. Grading and Assignments

Composition of Final Grade:

Assignment 1: 3 Take-Home Problem Sets	Deadlines: 8-10 days after distribution	Submit via Moodle	45% (15% each)

Assignment 2: 5 Short In-Class Quizzes	Deadline: See Moodle for schedule	In lecture, during sessions 3, 5, 7, 9, 11.	25% (5% each)
Assignment 3: Final Exam	Deadline: Final exam week	In-class exam	30%

Assignment Details

Assignment 1: Problem Sets

This class will have three problem sets, distributed evenly over the semester. The assignments consist of different mathematical problems that need to be solved by hand and by writing and running code. By completing the assignments students gain a deeper understanding of the topics covered in class, and experience in solving those problems.

Start assignments early. Please note that the problem sets are designed to take time and thought, and cannot be (successfully) completed the night before they are due. Please read and begin them as soon as they have been assigned in order to give you ample time to struggle with the problems on your own, seek help from each other and the instructor, and write up clear solutions.

I ask that all solutions be typed. Not only does this make the instructor's job of grading much easier, but the exercise of writing up your work clearly and understandably is an important skill in its own right for any practicing data scientist. **I strongly recommend using LaTeX**, but will not penalize you for using other tools (e.g., if you want to struggle with Word's equation editor). If you are still learning how to use LaTeX, please account for some extra time to write up your solutions. **I promise it gets faster with practice, and is worth your time.**

Please show all work. I should be able to trace every step of your process, whether it's a mathematical proof, a solution to a problem, or code. **I should be able to successfully run any of your code on my machine; replication is key!**

You will be assigned small groups to work on the problem sets and you may type up one solution set per group.

Assignment 2: Short In-Class Quizzes

In lieu of a midterm exam, we will have five small, low-stakes assessments, conducted at the beginning of lecture. These will assess your knowledge of the mathematical concepts and will be done with pen and paper, not code.

Assignment 3: Final Exam

This class will have a final exam. The exam will cover the most important topics discussed in class or covered in core readings until that date. The exam may include mathematical exercises, multiple choice questions and questions that require written answers.

Late submission of assignments: For each day the assignment is turned in late, the grade will be reduced by 10% (e.g. submission two days after the deadline would result in 20% grade deduction).

Attendance: Students are expected to be present and prepared for every class session. Active participation during lectures and seminar discussions is essential. If unavoidable circumstances arise which prevent attendance or preparation, the instructor should be advised by email with as much advance notice as possible. Please note that students cannot miss more than two out of 12 course sessions. For further information please consult the Examination Rules §10.

Academic Integrity: The Hertie School is committed to the standards of good academic and ethical conduct. Any violation of these standards shall be subject to disciplinary action. Plagiarism, misuse of AI, free riding in group work, and other deceitful actions are not tolerated. See Examination Rules §16, the Hertie Plagiarism Policy, and the Hertie Guidelines for Artificial Intelligence Tools.

Compensation for Disadvantages: If a student furnishes evidence that he or she is not able to take an examination as required in whole or in part due to disability or permanent illness, the Examination Committee may upon written request approve learning accommodation(s). In this respect, the submission of adequate certificates may be required. See Examination Rules §14.

Extenuating Circumstances: An extension can be granted due to extenuating circumstances (i.e., for reasons like illness, personal loss or hardship, or caring duties). In such cases, please contact the course instructor and Examination Office *in advance* of the assignment deadline.

4. General Readings

All readings are either available online, or the relevant portions will be uploaded to the Moodle site.

- Blitzstein, Joseph K. and Jessica Hwang. *Introduction to Probability: Second Edition.* Taylor & Francis Group, 2019. URL: https://drive.google.com/file/d/1VmkAAGOYCTORq1wxSQqy255qLJjTNvBI/edit. (Called **Blitzstein and Hwang** in course readings.)
    - If you would like to watch some video lectures by the author to supplement your learning, they are available on YouTube here: https://projects.iq.harvard.edu/stat110. You will also find practice problems and solutions at this site.
- Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis: Third Edition.* Boca Raton, FL: CRC Press, 2014. (Called **Gelman** in course readings.)
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep Learning.* Cambridge: MIT Press, 2016. URL: https://www.deeplearningbook.org/. (Called **GBC** in course readings.)
- Herman, Edwin "Jed" and Gilbert Strang. *Calculus Volume 1.* Houston: OpenStax, 2017. URL: https://openstax.org/details/books/calculus-volume-1. (Called **Herman and Strang, Vol. 1** in course readings.)

- Herman, Edwin "Jed" and Gilbert Strang. *Calculus Volume 2.* Houston: OpenStax, 2017. URL: https://openstax.org/details/books/calculus-volume-2. (Called **Herman and Strang, Vol. 2** in course readings.)
- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning.* URL: https://www.statlearning.com/. (Called **ISL** in course readings.)
- Murphy, Kevin P. *Machine Learning: A Probabilistic Perspective.* Cambridge, MA: MIT Press, 2012. (Called **Murphy** in course readings.)
- Savov, Ivan. *No Bullshit Guide to Linear Algebra.* Montreal: Minireference Co., 2016. (Called **Savov** in course readings.)
- Wasserman, Larry. *All of Statistics.* New York: Springer, 2004. (Called **Wasserman** in course readings.)

Other References:

- Bishop, Christopher M. *Pattern Recognition and Machine Learning.* New York: Springer, 2006.
- Savov, Ivan. *No Bullshit Guide to Math & Physics.* Montreal: Minireference Co., 2016. (Chapters 1,3, and 5)
- Boyd, Stephen and Lieven Vandenberghe. *Convex Optimization.* New York: Cambridge University Press, 2004.
- Strang, Gilbert. *Linear Algebra and Learning from Data.* Cambridge: Wellesley-Cambridge Press, 2019.

5. Course Sessions and Readings

All course readings can be accessed on the course Moodle page.

Session 1: Probability Theory	
Required Readings	**Blitzstein and Hwang** - Chapter 1 (p. 1-28) - Chapter 2 (p. 45-66) - Example 2.7.1, The Monty Hall Problem

Session 2: Random Variables and Their Distributions	
Required Readings	**Blitzstein and Hwang** - Chapter 3 (p. 103-133)

Session 3: Expectation and Joint Distributions	

Required Readings	**Blitzstein and Hwang**
	- Chapter 4
	    - Sections 4.1 to 4.3 (p. 149-163)
	    - Sections 4.5 to 4.7 (p. 170-181)
	- Chapter 7
	    - Section 7.1 (p. 303-312)
	    - Section 7.4 (p. 332-337)

Session 4: Continuous Random Variables and Limit Theorems	

Required Readings	**Blitzstein and Hwang**
	- Chapter 5
	    - Sections 5.1-5.2 (p. 213-223)
	    - Sections 5.4-5.5 (p. 231-244)
	- Chapter 7
	    - p. 312-320 (stop before Example 7.1.25)
	    - Sections 7.2-7.3 (p. 324-332)
	    - Section 7.5 (p. 337-343)
	- Chapter 10
	    - Sections 10.2-10.3 (p. 467-476)

Recommended Readings	If you need a refresher on differential calculus, now is the time.
	Herman and Strang, Vol. 1:
	- Chapter 3.3 (Differentiation Rules)
	- Chapter 3.4 (Derivatives as Rates of Change)
	    - p. 270-272 (sections titled "Population Change" and "Changes in Cost and Revenue")
	- Chapter 3.6 (The Chain Rule)
	    - p. 287-289
	    - Example 3.54 on p. 292
	- Chapter 3.9 (Derivatives of Exponential and Logarithmic Functions)
	    - Theorem 3.14 (p. 323) and Example 3.75 (p. 323-324)
	    - Theorem 3.15 (p. 324-325) and Examples 3.77 (p. 325-326) and 3.79 (p. 327)
	- Chapter 4.3 (Maxima and Minima)
	    - p. 366-375
	- Chapter 4.5 (Derivatives and the Shape of a Graph)

Session 5: Fitting Models to Data I: Maximum Likelihood Estimation	
Required Readings	**Wasserman** - Chapter 6 - Chapter 9 - Focus on p. 19 as well as sections 9.1, 9.3, and 9.4 Orloff and Bloom notes: https://math.mit.edu/~dav/05.dir/class10-prep.pdf
Recommended Readings	**ISL**, Chapter 2 (Statistical Learning)

Session 6: Fitting Models to Data II: Bayesian Analysis	
Required Readings	**Gelman** - Chapter 1 (p. 1-13) - Chapter 2

Session 7: Linear Algebra Fundamentals, Eigendecomposition, and SVD	
Required Readings	**GBC**, Chapter 2 (Linear Algebra) - p. 29-43 **Savov**, Chapter 7 - Section 7.1 (p. 257-268) (stop at Discussion) - Section 7.6 (p. 292-295) (stop at LU decomposition)
Recommended Readings	For additional review, see: https://www.cs.cmu.edu/~zkolter/course/15-884/linalg-review.pdf For another treatment of SVD, see: https://www-users.cse.umn.edu/~lerman/math5467/svd.pdf

Session 8: Optimization I: Matrix Calculus and PCA	
Required Readings	Matrix Calculus (Section 4, p. 20-26) https://www.cs.cmu.edu/~zkolter/course/15-884/linalg-review.pdf Lecture Notes on Lagrange Multipliers **ISL**, Chapter 12 - Section 6.3.1 (p. 253-256) - Section 12.2 (p. 496-508) - Section 12.5.1 (will do together in lab)

Session 9: Optimization II: Linear Regression and Regularization	
Required Readings	**ISL** - Chapter 3 (p. 59-83; stop at Section 3.3) - Chapter 6.2 (p. 237-251)

Session 10: Optimization III: Taylor Series Approximation and Gradient Descent	
Required Readings	**Herman and Strang, Vol. 2** - Chapter 6.1 (Power Series and Functions) - Chapter 6.3 (Taylor and Maclaurin Series) **GBC** - Chapter 4 (Numerical Computation)

Session 11: Modeling Latent Data Structures	
Required Readings	**ISL** - Section 12.4 (Clustering Methods), p. 514-519 **Murphy** - Chapter 11 (Mixture Models and the EM Algorithm), p. 337-365

Session 12: Catch-Up and Review	
Required Readings	None