New York University Tandon School of Engineering
Computer Science and Engineering

CS-GY 6763 (CS-UY 3943)
Algorithmic Machine Learning and Data Science, Spring 2022

Professor Rajesh Jayaram
Mondays 5:00pm-7:30pm, 2 MetroTech Center 9011

CONTACT INFORMATION

Email: rkj234@nyu.edu

Virtual Zoom Office: Wednesdays 4:30-6:00pm, https://nyu.zoom.us/my/rkjayaram

Ed Discussion Forum (for questions): TBA

TA Office Hours: Weekly times TBA.

Course Webspage: https://rajeshjayaram.com/amlds2022

COURSE DESCRIPTION

This course gives a behind-the-scenes look into the algorithms and computational methods that make machine learning and data science work at large scale. How does a service like Shazam match a sound clip to a library of 10 million songs in under a second? How do scientists find patterns in terabytes of genetic data? How can we efficiently train neural networks with millions of parameters on millions of labeled images? We will address these questions and others by studying advanced algorithmic techniques like randomization, approximation, sketching, continuous optimization, spectral methods, and Fourier methods.

COURSE PREREQUISITES

We require a previous course in machine learning (for example, CS-UY 4563, CS-GY 6923, or ECE-GY 6143), a previous course in algorithm design and analysis (for example, CS-UY 2413, CS-GY 6033, or CS-GY 6043), and a previous course in linear algebra (for example, MA-UY 2034, 3044, or 3054). Experience with probability and random variables is also necessary. In general, *modern algorithm design uses a lot of math!* This is partially what makes it such an interesting subject to study. In this course, we rely most heavily on probability and linear algebra, but we will also learn some approximation theory, Fourier analysis, and high dimensional geometry. Be prepared for theoretically rigorous and proof-based problem solving, which may require you to refresh topics you covered in prior courses. In particular, you should know:

1. **Probability**: Random variables, discrete and continuous probability distributions, expectation, variance, covariance, independence, correlation, conditional and joint probability, Gaussian random variables, law-of-large-numbers, Markov's and Chebyshev's inequalities.

2. **Linear Algebra**: Vector inner and outer products, matrix-vector and matrix-matrix multiplication, vector norms (e.g., Euclidean), matrix norms (e.g., Frobenius, operator), triangle inequality, Cauchy-Schwarz, solving systems of linear equations, linear subspaces, linear independence, projection, matrix rank, column/row span, null space, orthogonal matrices, basics of eigenvectors and eigenvalues.

Finally, you do a small amount of programming in this course. You can use whatever language you like (most students use MATLAB or Python) and will need to use basic linear algebra packages in these languages.

COURSE STRUCTURE AND GRADING

There is one 2.5 hour class meeting per week, divided by a 15 minute break. Grading in the course will consist of several problem sets, one in-class midterm, and either a final exam or a final project. Details follow:

Written problem sets (50% of grade). These assignments are completed at home and involve the analysis and application of methods learned in class, with occasional programming exercises to further explore lecture content. I expect most of your learning to occur while working on these exercises, and investing time on them is the best way to prepare for the midterm exam and final. Assignments and their due dates will be posted on the course webpage. Late assignments will only be accepted if there are extenuating circumstances and you have obtained prior permission from me.

Midterm exam (15% of grade). In person for those who can attend, and we will make remote accommodations for others. The exam will take 1 hour, and we will have a short lecture afterwards. You will be allowed a cheat-sheet (a two-sided piece of paper with whatever information you want on it).

Choice of final project or final exam (25% of grade). You have the option of completing either a final project or final exam for this class. Typically preparing for the final exam is less work, so I suggest this as the "default option". However, any student interested in research, and especially in topics related to this class, might want to choose the project instead. The project can either be applied or theoretical, but must be related to algorithmic aspects of data science or machine learning. Typically, paper's will center around 1 or 2 recent research papers. Expectations and a list of potential topics and papers to read will be posted on the course website.

Class participation (10% of grade). This grade captures how much you contribute to your own learning and that of your peers. Since different students have different styles, there are *many ways* to earn full credit for this part of the course. You can actively participate in class. You can ask good questions, or answer those of your peers, on the class Ed forum. You can contribute to the reading group. You can attend and actively contribute to office hours.

COURSE POLICIES

Written problem sets: Turned in via NYU Brightspace/Gradescope. While not required, I encourage students to prepare written problem sets in LaTeX. Students will receive 10% extra credit on Problem Set 1 for preparing it in LaTeX.

Homework collaboration policy:

- *Discussion of high-level ideas for solving problems sets or labs is allowed, but all solutions and any code must be written up independently.* This reflects the reality that algorithms and machine learning research are rarely done alone. However, any researcher or practitioner must be able to communicate and work through the details of a solution individually.

- *Students must name any collaborators they discussed problems with at the top of each assignment (list at the top of each problem separately).*

- *Do not write solutions or code in parallel with other students.* If problem ideas are discussed, solutions should be written or implemented at a later time, individually. **I take this policy very seriously. Do not paraphrase, change variables, or in any way copy another students solution.**

Grade changes: Fair grading is important to me. If you feel that you were incorrectly docked points on an assignment or exam, please let me know – it is very possible I or the TAs misunderstood your work. Do not wait until the end of the semester to bring up grading issues! If you notice something off, please ask ASAP.

Questions about performance: If you are struggling in the class, contact me as soon as possible so that we can discuss what's not working, and how we can work together to ensure you are mastering the material. It is difficult to address questions about performance at the end of the semester, or after final grades are submitted: by Tandon policy no extra-credit or makeup work can be used to improve a students grade once the semester closes. So if you are worried about your grade, seek help from me *early*.

READINGS

There is no textbook for this course. Many of the topics covered are new, and not sufficiently addressed in existing textbooks. Course material will consist of my lecture slides, accompany supplementary course

notes, as well as assorted online resources, including papers, notes from other courses, and publicly available surveys. Please refer to the course webpage (https://rajeshjayaram.com/amlds2022/) before and after lecture to obtain links.

COURSE SCHEDULE

The following schedule is tentative and subject to change.

THE POWER OF RANDOMNESS

1. 1/24 Random variables and concentration inequalities, mark and recapture, streaming model, Count-Min Sketch
2. 1/31 Exponential tail bounds (Chernoff + Bernstein), Count-Sketch, load balancing, balls-in-bins and other applications.
3. 2/7 High dimensional geometry and the Johnson-Lindenstrauss Lemma.
4. 2/1 Linear Sketching, Cauchy random variables and L_1-norm estimation, Precision Sampling, L_1 sampling.
5. 2/21 President's Day – No Class
6. 2/28 Nearest neighbor search, locality sensitive hashing

OPTIMIZATION

5. 3/7 Convexity in machine learning, gradient descent and preconditioning
6. 3/14 Spring Break – No Class
7. 3/21 Midterm, then Online and stochastic gradient descent, coordinate descent.
8. 3/28 Discrete optimization: clustering, greedy methods, submodularity

SPECTRAL METHODS AND LINEAR ALGEBRA

9. 4/4, Singular value decomposition, Krylov methods.
10. 4/11 Spectral graph theory, clustering, stochastic block models.
11. 4/18 Randomized linear algebra, sketching for linear regression, ϵ-nets arguments.

FOURIER METHODS

12. 4/25 The Fourier transform and applications, random Fourier features
13. 5/2 Compressed sensing, the restricted isometry property, basis pursuit
14. 5/9 TBD – will vote on a topic.

5/?? Final Exam Schedule, TBD

COURSE OBJECTIVES

1. Students will build experience with the most common algorithmic tools used in machine learning, including optimization, relaxation, spectral methods, Fourier methods, and Monte Carlo algorithms. They will also strengthen their understanding of the mathematical concepts behind these tools.

2. Through problem sets and in-class assignments, students will practice applying, combining, and modifying the methods learned in lectures to specific problems in machine learning and data analysis. The goal is to prepare students to use these tools in industrial or academic positions where they are developing computationally efficient machine learning methods.

3. Students will learn how theoretical questions (involving asymptotic complexity, communication complexity, convergence rate, approximation quality, etc.) can guide the design of new algorithmic methods in machine learning. We will practice framing and rigorously answering such questions.

4. The course will prepare students to contribute to research projects on computational methods for machine learning and data science.

5. Through assigned papers and discussions, students will improve their ability to read, summarize, and extract value from contemporary research papers in machine learning and data mining conferences like NeurIPS, ICML, KDD, ICLR and AAAI.

ADDITIONAL INFORMATION

MOSES CENTER STATEMENT OF DISABILITY.

If you are student with a disability who is requesting accommodations, please contact New York University's Moses Center for Students with Disabilities (CSD) at 212-998-4980 or mosescsd@nyu.edu. You must be registered with CSD to receive accommodations. Information about the Moses Center can be found here.

NYU SCHOOL OF ENGINEERING POLICIES AND PROCEDURES ON ACADEMIC MISCONDUCT.

The complete Student Code of Conduct can be found here.

A. Introduction: The School of Engineering encourages academic excellence in an environment that promotes honesty, integrity, and fairness, and students at the School of Engineering are expected to exhibit those qualities in their academic work. It is through the process of submitting their own work and receiving honest feedback on that work that students may progress academically. Any act of academic dishonesty is seen as an attack upon the School and will not be tolerated. Furthermore, those who breach the School's rules on academic integrity will be sanctioned under this Policy. Students are responsible for familiarizing themselves with the School's Policy on Academic Misconduct.

B. Definition: Academic dishonesty may include misrepresentation, deception, dishonesty, or any act of falsification committed by a student to influence a grade or other academic evaluation. Academic dishonesty also includes intentionally damaging the academic work of others or assisting other students in acts of dishonesty. Common examples of academically dishonest behavior include, but are not limited to, the following:

   1. Cheating: intentionally using or attempting to use unauthorized notes, books, electronic media, or electronic communications in an exam; talking with fellow students or looking at another person's work during an exam; submitting work prepared in advance for an in-class examination; having someone take an exam for you or taking an exam for someone else; violating other rules governing the administration of examinations.
   2. Fabrication: including but not limited to, falsifying experimental data and/or citations.
   3. Plagiarism: intentionally or knowingly representing the words or ideas of another as one's own in any academic exercise; failure to attribute direct quotations, paraphrases, or borrowed facts or information.
   4. Unauthorized collaboration: working together on work meant to be done individually.
   5. Duplicating work: presenting for grading the same work for more than one project or in more than one class, unless express and prior permission has been received from the course instructor(s) or research adviser involved.
   6. Forgery: altering any academic document, including, but not limited to, academic records, admissions materials, or medical excuses.

NYU SCHOOL OF ENGINEERING POLICIES AND PROCEDURES ON EXCUSED ABSENCES.

The complete policy can be found here.

A. Introduction: An absence can be excused if you have missed no more than 10 days of school. If an illness or special circumstance has caused you to miss more than two weeks of school, please refer to the section labeled Medical Leave of Absence.

B. Students may request special accommodations for an absence to be excused in the following cases:

1. Medical reasons
2. Death in immediate family
3. Personal qualified emergencies (documentation must be provided)
4. Religious Expression or Practice

Deanna Rayment, deanna.rayment@nyu.edu, is the Coordinator of Student Advocacy, Compliance and Student Affairs and handles excused absences. She is located in 5 MTC, LC240C and can assist you should it become necessary.

NYU SCHOOL OF ENGINEERING ACADEMIC CALENDAR

The full calendar can be found here. Please pay attention to notable dates such as Add/Drop, Withdrawal, etc. For confirmation of dates or further information, please contact Susana Garcia: sgarcia@nyu.edu.