APMA E4990.001 Topics in Applied Math: Mathematics of Data Science Spring 2024

Vlad Kobzar
vak2116@columbia.edu

January 17, 2024

Description: This course is an application-oriented introduction to mathematical concepts and techniques used in data science, with a balanced combination of theory, algorithms and programming implementations. A provisional list of topics includes:

- *Supervised learning and linear models*
  - Linear models for regression (ordinary least squares and variants) and classification (perceptron, logistic regression and support vector machines) - focus on their geometric and probabilistic interpretations and computational solutions
  - Fitting models to training data, such as empirical risk minimization
  - Performance on unseen (test) data and regularization techniques
  - Sparse models and compressed sensing
- *Unsupervised learning/dimensionality reduction:*
  - Principal component analysis (PCA)
  - Matrix completion and the Netflix problem
  - Clustering and graph-based learning, including diffusion maps, spectral methods and convex relaxations/semidefinite programming

  - Ranking and the PageRank algorithm

- *High-dimensional data*
  - Randomized linear algebra and random projections

- *Nonlinear models and neural networks*
  - Kernel methods and representer theorem
  - Mathematical aspects of deep learning, including universal approximation properties of neural networks, backpropagation and vanishing gradients
  - Convolutional neural networks and neural models for sequential data and graphs

- *Generative models and semi-supervised learning:* linear discriminant analysis, naive Bayes, probabilistic PCA, latent variables, importance sampling, high-dimensional analysis and generalizations to diffusion models

- *From predictions to actions*
  - Sequential decision making and Kalman filtering
  - Interaction with adversarial environments – online learning
  - Incomplete information and exploration-exploitation trade-offs – bandit problems, Markov decision processes and mathematical aspects of reinforcement learning, such as the policy gradient theorem

Mathematical topics will include selected tools of linear algebra (e.g., singular value decomposition and spectral decomposition, randomized linear algebra), probability and statistics, minimax theorems, Fourier transforms and convolutions. We will use these tools to understand the essential character, power, and limitations of techniques used in data science.

Algorithmic topics will include selected techniques for optimization, such as gradient descent, stochastic gradient descent, conjugate gradients, and constrained optimization/duality theory, including linear programming, Lagrange multipliers and semidefinite programming. Applications will emphasize data science, and the required coursework will include programming exercises to implement some of the relevant data science techniques.

Textbook: The main textbook for the course is *Linear Algebra and Learning from Data* by Gilbert Strang (Wellesley-Cambridge Press 2019, ISBN: 978-06921963-8-0). I have requested to reserve a copy of this textbook at the Science & Engineering Library. Freely available supplemental materials may be referenced on the weekly schedule.

Additional materials: Prof. Strang has a collection of video lectures online that cover some of the same material we will cover.[1]

While the following textbooks/courses may be somewhat more advanced than our class and/or provide a different emphasis, they contain possible directions for further study, open problems and research projects in mathematics of data science.

- Hardt and Recht, Patterns, Predictions, and Actions: Foundations of Machine Learning, Princeton University Press (2022)[2]
- Francis Bach, Learning Theory from First Principles (2023), MIT Press forthcoming [3]
- Mathematics of Data Science (2020), Afonso Banderia, ETH Zurich[4]
- Mathematical Tools for Data Science (2020), Carlos Ferndandez-Granda, NYU[5]

Prerequisites: Students are expected to have basic knowledge in multivariable calculus (on the level of APMA E2001), linear algebra (on the level of APMA E3101), and elementary probability (on the level of IEOR E3658). Basic programing skills are required as well. As discussed below, Python is highly recommended for the homework assignments.

Class website and schedule: The lectures will take place on Mondays and Wednesdays 11:40am-12:55pm in 1127 Seeley W. Mudd Building. Students can participate in person or through a zoom meeting link available via Courseworks. The

[1] `https://ocw.mit.edu/courses/mathematics/18-065-matrix-methods-in-data-analysis-signal-processing-and-machine-learning-spring-2018/`

[2] `https://mlstory.org/pdf/patterns.pdf`

[3] `https://www.di.ens.fr/~fbach/ltfp_book.pdf`

[4] `https://people.math.ethz.ch/~abandeira/Spring2020.MathDataScience.html`

[5] `https://cims.nyu.edu/~cfgranda/pages/MTDS_spring20/index.html` and `https://cds.nyu.edu/math-tools/`

course materials and schedule will be available on Courseworks. If you have a foreseeable conflict, you must let me know as soon as possible. We will use Ed Discussion (accessible via Courseworks) for course-related announcements and discussion.

Homework: Homework will be graded strictly because its main purpose is to give you practice of presenting mathematical reasoning clearly and completely. The grading will take into account presentation, as well as the correctness of answers: be sure to show all work in a logical order, and use complete sentences where narrative is called for.

The recommended language for the programming homework is Python; we will often provide ancillary subroutines in Python that can be used in the homework, e.g., for data loading, cleaning and visualization. While students can also submit programming homework solutions in another programming language, such as MATLAB, no ancillary subroutines will be provided in that language, and therefore students using any language other than Python will need to develop them on their own.

Late homework will not be accepted except in the case of a documented medical or similar excuse. Students are encouraged to form study groups and to discuss homework problems with each other. However, for both programming and other problems, each student must implement and present their own solutions. Sharing solutions or copying of another person's solution or other materials are not permitted.

Project Option: Students who are interested in conducting independent research may submit a research proposal for approval to the instructor. The goal of the project is to investigate a specific application of mathematics to data science. The project should be carried out in groups of two (the instructor may approve exceptions to this requirement in case of PhD-related research which can be performed individually). Submitting a project used for academic credit or otherwise in connection with another course at Columbia or elsewhere is not allowed.

The proposal must be *one page long*. It should include the following information (each point will be evaluated separately):

- A description of a specific question that you want to explore. The question can be experimental, i.e. whether a certain technique works for a particular application, or theoretical, whether a certain theoretical tool can be used to analyze a data-analysis method. Be as concrete as you can.

- Context for your topic, including relevant bibliographic references.
- An outline of what you plan to do, including *two major milestones*, and a precise justification of how it relates to the question that you are studying. For experimental projects, this includes a description of the dataset you plan to use.

The project report should be written in Latex and be no more than 5 pages of the main text (not including references and appendices). The contribution, and significance of the report will be evaluated primarily based on the main text (without appendices), and so enough details must be provided in the main text to convince the reader of the report's merits. The report should include the following sections, which will be evaluated separately:

- *Introduction*: Describe the question have you been studying? Why is this question relevant/impactful?
- *State of the art*: Describe the state of the art methods/results for answering this question, with relevant bibliographic references.
- *Methodology*: How did you address the question? Did you modify existing methods? What datasets did you use? What theoretical tools did you apply? If you have deviated from your original proposal, explain why.
- *Results*: What results did you obtain? Do they make sense? Provide a thorough analysis. Negative results are completely fine (they can be very valuable!), but please explain clearly what worked and what did not.
- *Discussion*: What did you find out? How does your work fit into the context of the current state of the art? Do the results suggest any other interesting questions to explore?

Grading: The course grade will determined based on the homework problem set grades.

The project proposal replaces one problem set. If the project proposal is approved, the grades for two project milestone deliverables, the project presentation, and the final project report will replace replace 4 additional problem set grades.

The grade boundaries and extra credit awards will be determined by the instructor at the end of the course.

Academic integrity: Plagiarism and cheating will not be tolerated. Columbia University policies in this area will be followed. See `https://www.college.columbia.edu/academics/academicintegrity`