APMA E4990.002 Topics in Applied Math: Mathematics of Data Science Fall 2021

Vlad Kobzar
vak2116@columbia.edu

August 24, 2021, updated September 12, 2021*

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: principal component analysis, regression, sparsity, compressed sensing and matrix completion, clustering, linear programming, game theory and duality, problems on graphs and convex relaxations, statistical learning theory, methods based on prior knowledge, gradient descent and proximal methods, stochastic gradient descent and neural networks.

Mathematical topics will include selected tools of linear algebra (e.g., singular value decomposition and spectral decomposition), probability, duality, and minimax theorems. 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 and stochastic gradient descent. Applications will emphasize data science (e.g., supervised and unsupervised learning, classification, dimensionality reduction, and convex-relaxation-based approaches to inverse problems). 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*

*The section entitled "Class website, schedule and Piazza" and the subsequent sections have been added or updated.

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 courses are more advanced than our class, they contain possible directions for further study, open problems and research projects in mathematics of data science.

- Mathematics of Data Science, Prof. Afonso Banderia, ETH Zurich[2]
- Mathematical Tools for Data Science, Prof. Carlos Ferndandez-Granda, NYU[3]

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 (MATLAB, Python, etc.) are required as well.

Class website, schedule and Piazza: The lectures will take place on Tuesdays and Thursdays 11:40am-12:55pm in Hamilton 705. Students can participate in person or through a zoom meeting link available via Courseworks. The course materials, office hour details and the dates of the quizzes are also available on Courseworks. If you have a foreseeable conflict with the time of the quizzes, you must to let me know by September 21. Please sign up for a Piazza page of this course (the sign-up link has been shared on Courseworks) to receive course-related communications and participate in the discussion of the course material.

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:

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

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

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

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 sometimes provide ancillary subroutines in Python that can be used in the homework, e.g., for data cleaning. While students can also submit programming homework solutions in MATLAB, no ancillary subroutines will be provided in MATLAB, and therefore students using MATLAB will need to develop them on their own.

In fairness to other students in the course, late homework will not be accepted except in the case of a documented medical or similar excuse. Students are encouraged to discuss homework problems with each other, and to seek and provide help with homework on Piazza. However, for both programming and other problems, each student must implement and present their own solutions. Copying of another person's solution or other materials is not permitted.

Quizzes: There will be three quizzes. In preparing for them, make sure you understand the correct solutions to as many of the homework problems as possible, as well as the statements and proofs of theorems and the definitions of key terms covered in class. Questions in the quizzes will be based on a selection of the homework questions and the material covered in the lectures. No collaboration is permitted on the quizzes.

Grading: Grading will be based on the following combination of homework and quizzes: 70% homework and 30% quizzes. Extra credit may be given to students who regularly and correctly answer their classmates' questions on Piazza and otherwise provide high-quality contributions to Piazza discussions. 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 has policies in this area, which will be followed. See `https://www.college.columbia.edu/academics/academicintegrity`