APMA E4990.002 Topics in Applied Math: Mathematics of Data Science Spring 2022

Vlad Kobzar
vak2116@columbia.edu

November 15, 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-based techniques and regularization, nonnegative matrix factorization, compressed sensing and matrix completion, problems on graphs and convex relaxations, PageRank algorithm, clustering, statistical learning theory, classification methods, such as support vector machines and kernel-based methods, methods based on prior knowledge, and neural networks.

Mathematical topics will include selected tools of linear algebra (e.g., singular value decomposition and spectral decomposition, randomized linear algebra), probability, duality theory, 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 linear programming, gradient descent, stochastic gradient descent, conjugate gradients, and alternating direction method of multipliers (ADMM). 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 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 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 Wednesday 11:40am-12:55pm. The course materials and schedule will be available on Courseworks.

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.

[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/`

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.

Grading: Grading will be based on homework, quizzes and/or projects. The grade boundaries 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`