Syllabus

Stat89a: Linear Algebra for Data Science: Spring 2019 Semester

Instructor: Michael Mahoney (mmahoney@stat.berkeley.edu)
Important. Include "stat89a" on the subject line of email correspondence.

Office Hours: TBA

GSI: William Krinsman (krinsman@berkeley.edu), Xinlei Pan (xinleipan@berkeley.edu)

Office Hours: TBA

UGSIs: Ivon Liu (ivon.liu@berkeley.edu)

Office Hours: TBA

Time and Location: Main class: Tue-Thu 12:30-2:00am, Evans 60. First meeting is Tuesday, January 22, 2018. Additional Required Lab Sessions: Four Lab sections at various times on Friday.

Prerequisite or corequisite: Foundations of Data Science (COMPSCI C8 / INFO C8 / STAT C8). One year of calculus.

Course Description: An introduction to linear algebra for data science. The course will cover introductory topics in linear algebra, starting with the basics; discrete probability and how probability can be used to understand high-dimensional vector spaces; matrices and graphs as popular mathematical structures with which to model data (e.g., as models for term-document corpora, high-dimensional regression problems, ranking/classification of web data, adjacency properties of social network data, etc.); and geometric approaches to eigendecompositions, least-squares, linear equations, principal components analysis, etc.

Data Science Connector: During Spring 2016 and Spring 2017, this course was a two-unit connector. During Spring 2018, this course was expanded, covering similar topics in a more methodical manner. In particular, going forward, it will serve as a comprehensive introduction to linear algebra, but presented in a way more appropriate for students of data science.

Course Requirements: The course evaluation will consist primarily of homeworks (ca. 50%)—split between pencil-and-paper (ca 30%) and computational (ca. 20%) homeworks—lab sessions (ca. 5% to 10%), and a midterm (ca. 25%), and a final (ca. 25%).

Piazza: Sign up for this course on Piazza. We will distribute information there, and it's a great place to ask/answer questions.

Academic Honesty: This course will follow the requirements from the main data science course.

Textbook: Online teaching materials from previous iterations of the class, expanded and supplemented by modified and updated computational notebooks.

Syllabus: The course will provide an introduction to linear algebra for data science. As such, the presentation and emphasis will be quite different than typical linear algebra classes that are more geared toward engineering, scientific computing, and related areas. The challenge is to convey the basic insights early in the course without getting delayed by extensive coverage of concepts that matter strongly in traditional numerical presentations of linear algebra, but that matter less for data science. To do this, we will focus on quadratic form structure, rather than the subspace structure, of the linear algebra, as well as connections with probability. The former is easier at the introductory level than subspaces, QR decompositions, etc., and is more intuitive as to how the results will generalize. The order of topics is also different. Linear equations will be covered, but at the end. More central is the role of least-squares approximations and principal components analysis, which pervade data science. After introducing basic concepts of Euclidean spaces, matrix multiplication, linear combination, span, etc. (to show that intuitive notions from low-dimensional Euclidean spaces can be generalized to high dimensions and can be used to model high-dimensional data), the course introduces the basics of discrete probability (to provide a model for high-dimensional spaces and to illustrate how high-dimensional spaces are very different than low-dimensional spaces). Then quadratic forms, eigenvalue decompositions of symmetric matrices, least squares regression, and principal components analysis are covered. Depending on time, the use of these methods for spectra clustering and ranking, solving linear equations, etc. will be covered.

Tentative week-by-week outline: Here is a tentative week-by-week outline.

- Week 1. Introduction, motivation, and overview: Representing data as flat tables versus matrices and graphs; Different ways probability/randomness/noise interacts with data; Probability and matrices/graphs in data science versus other areas; Trying to quantify the inference step.

- Week 2. Introduction to matrices, vectors, and $\mathbb{R}^n$: Ways to label points, elements, vectors, matrices, etc.; Basic properties of $\mathbb{R}^n$; Norms and balls; Operations on matrices; Vector addition and scalar multiplication.

- Week 3. Vector spaces, matrices and operations on matrices, linear functions, and examples: Introduction to vector spaces and subspaces; Matrices and operations on matrices, including matrix multiplication; Special types of matrices; Functions, linear functions, and linear transformations; Examples of matrices as transformations.

- Week 4. Geometry of $\mathbb{R}^n$, linear combinations, spans, etc.: dot products, angles, and perpendicularity; Linear combinations, span, and linear independence; Bases, orthonormal bases, and projections.

- Week 5. Probability, Random Variables, etc.: Flipping coins, rolling dice, and throwing darts; Sample spaces and events; Some basic set theory; Basics of probability; Mass/Volume as an intuition; Conditional Probability; Independence.

- Week 6. Probability, Random Variables, etc., cont.: Bayes' Theorem; Computing more complex probabilities; Random Variables; Mean, variance, and moments of random variables; Properties of high dimensions versus low dimensions; Concentration in Flipping coins; Concentration in throwing darts; Two basic properties of expectations of random variables.

- Week 7. Probability, Random Variables, etc., cont.: Conditional expectation; More complex combinations: Covariances and correlations; Connections between probability theory and linear algebra; Large, small, and typical variability; Bounding deviations from the mean (weak bounds); A normal baseline to aim for; Bounding deviations from the mean (strong bounds); Examples and applications.

- Week 8. Eigendecompositions: Eigenvectors and Eigenvalues: Introduction to eigenvectors and eigenvalues; Some simple examples; Computing eigenvectors and eigenvalues; Basic properties of determinants; Using determinants to understand small eigendecompositions; Expressing matrices in terms of their eigendecompositions; A larger example.

- Week 9. Eigendecompositions, cont.: The Quadratic Forms Perspective: Two-dimensional conic sections; Quadratic forms and matrices; Some simple examples; Symmetric bi-linear functions; Connections with conic sections; Definiteness, indefiniteness, and quadratic forms as a sum/difference of squares; Two other topics.

- Week 10. The spectral decomposition of a symmetric matrix: Orthogonal subspaces; Finding the spectral decomposition; Many uses of the spectral decomposition.

- Week 11. Least-squares (LS) regression and Principal component analysis (PCA); Computing LS; Statistical aspects of LS; Regression diagnostics with LS; Computing PCA; Statistical aspects of PCA; Connections between PCA and LS; Extension to the SVD.

- Week 12. Solving linear equations: Linear equations; Geometry of linear equations; Gaussian elimination; Row exchanges; Networks and incidence matrices; The four fundamental subspaces.

- Week 13. Solving linear equations, cont.: Basis transformations; Orthogonal bases; Gram-Schmidt Orthogonalization; Direct versus indirect methods; Numerical issues.

- Week 14. PageRank: Definition and history; Linear equation perspective; Eigenvector perspective; Probability and random walk perspective; Classification and ranking perspective.