Mathematical and Computational Foundations of Data Science

Instructor: Mauro Maggioni

Web page: https://mauromaggioni.duckdns.org/teaching/

Office: Wyman Bldg. N438

Synopsis

The course covers several topics in Data Science, focusing on key mathematical and statistical concepts and technical ideas common to many developments in the field, as well as on algorithms and their computational aspects.
The emphasis is on fundamental mathematical ideas (including basic functional analysis and approximation theory, concentration inequalities from a probabilistic and geometric point of view, analysis of and on graphs), core statistical techniques (e.g. linear regression, parametric and non-parametric methods), machine learning techniques for unsupervised (e.g. clustering, manifold learning), supervised (classification, regression), and semi-supervised learning.
Algorithmic and computational aspects of the above and their foundations, including basics of numerical linear algebra, and of linear and nonlinear optimization, to implement solutions to the problems above in a computationally efficient fashion. Applications will include statistical signal processing, imaging, inverse problems, graph processing, and problems at the intersection of statistics/machine learning and physical/dynamical systems (e.g. learning interaction kernels in agent-based models, model reduction for stochastic dynamical systems).

Detailed Topics

- Highlights of Linear Algebra; Eigenvalue and Singular Value Decompositions; Nonnegative Matrix Factorization.
  Applications: function interpolation and fitting; Principal Component Analysis, linear dimension reduction; linear data compression; latent semantic indexing; Multi-Dimensional Scaling.
  Computational aspects: full and sparse matrices; high-level discussion of algorithms for fundamental linear algebra operations and decompositions, and their corresponding computational cost; solution of linear systems.

- Least squares; inverse problems, well-posedness, conditioning; regularized least squares; sparsity.
  Applications: parametric regression; denoising; learning dynamical systems.
  Computational aspects: large matrices, least squares, condition numbers.

- Basics of function spaces; Hilbert spaces; reproducing kernel Hilbert spaces.
  Applications: nonparametric regression; distances between probability measures.

- Highlights of signal processing & approximation: Fourier and wavelet bases; approximation, compression, denoising; sparsity and dictionary learning; regression.

- Limit theorems, concentration inequalities (probabilistic and geometric interpretations).
  Applications: Johnson-Lindenstrauss Lemma; estimation of means and covariance matrices in high-dimensions and for fat-tailed distributions; randomized linear algebra.

- Mathematical Foundations, in depth:
  - high-dimensional phenomena, geometry of convex sets in high dimensions and their low-dimensional projections; connections with sparsity, compressed sensing and matrix completion;
  - parametric and nonparametric statistics: density estimation, regression (with nearest neighbors, kernel methods, tree methods and multiscale methods); reproducing kernel Hilbert spaces (RKHSs), kernel PCA.
  - dimension reduction, linear and nonlinear: PCA, random projections, manifold learning.
  - curse of dimensionality; hyperbolic crosses and sparse grids; structured models for regression, single- and multi-index models.

- Markov chains, random walks; applications to dimension reduction, spectral graph theory, spectral clustering, semisupervised learning; Markov state models, hidden Markov models (HMMs), model reduction for dynamical systems.
  Applications: pageRank; manifold learning with graph-based methods;
  Computational aspects: sparse matrices; eigenvalues and eigenvectors.

- Neural networks: construction, backpropagation; convolutional neural networks.
  Applications: object classification in images; generative models.
  Computational aspects: gradient descent and stochastic gradient descent.

References

Numerical Linear Algebra, L.N. Trefethen and D. Bau.
Introduction to Applied Linear Algebra – Vectors, Matrices, and Least Squares, S. Boyd and L. Vandenberghe.
Finite Dimensional Vector Spaces, Holmes.
Linear Algebra and Learning from Data, G. Strang.
Foundations of Data Science, A. Blum, J. Hopcroft, and R. Kannan.
High Dimensional Probability, An Introduction with Applications in Data Science, R. Vershynin.
A distribution-free theory of nonparametric regression, L. Gyorfi, M. Kohler, A. Krzyzak, H. Walk.
Lectures on Spectral Graph Theory, F.R.K. Chung.
How to Lie with Statistics, D. Huff.
Additional references for specific topics will be added as needed.

Grading
Grade to be based on assignments (20%), exams (40%) and a final project (40%).
Weekly problem sets will include theory, analysis and computational projects.

Prerequisites
Linear algebra will be used throughout the course, as will multivariable calculus and basic probability (discrete random variables). Ability to understand and write basic proofs (e.g. from a course in real analysis). Basic experience in programming in MATLAB/Python/R/C/Julia etc.. will be helpful in understanding demos in class, and in several homework sets.

Additional Information
Target audience: students from all areas of science, engineering, computer science, statistics, economics and quantitative studies that need advanced level skills in solving and creating novel solutions to problems in the general area of data science, signal processing, or statistical modeling are encouraged to enroll.