Mathematical and Computational Foundations of Data Science

Instructor: Mauro Maggioni Office: Whitehead 302D

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

Synopsis

The course covers several topics in the mathematical and computational foundations of Data Science. The emphasis is on fundamental mathematical ideas (basic functional analysis, reproducing kernel Hilbert spaces, concentration inequalities), basic statistical modeling techniques (e.g. linear regression, parametric and non-parametric methods), basic machine learning techniques for unsupervised (e.g. clustering, manifold learning), supervised (classification, regression), and semi-supervised learning, and corresponding computational aspects (numerical linear algebra, basic linear and nonlinear optimization to attack the problems above from a computational perspective). 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. model reduction for stochastic dynamical systems).

Detailed Topics

- Highlights of Linear Algebra, including Singular Value Decomposition and Principal Component Analysis, Nonnegative Matrix Factorization, Tucker decomposition of tensors. Function spaces, reproducing kernel Hilbert spaces (RKHSs), kernel PCA.

- Least squares, applications to parameter estimation and regression (RKHS).
  Computational aspects: large matrices, least squares, condition numbers, randomized linear algebra.

- Highlights of signal processing & approximation: Fourier and wavelet bases; approximation, compression, denoising; sparsity and compressed sensing.

- Highlights of optimization: convexity, Newton's method, Lagrange multipliers, gradient descent and stochastic gradient descent.

- Highlights of Probability and Statistics: limit theorems, concentration inequalities; covariance matrices; Kalman filtering.

- Topics in statistics and machine learning: high-dimensional phenomena; geometry of convex sets in high dimensions; connections with compressed sensing and matrix completion; parametric and nonparametric statistics: density estimation, regression, with nearest neighbors, kernel methods, tree methods and multiscale methods. Dimension reduction, linear and nonlinear: random projections, manifold learning.

- Markov chains, random walks; applications to dimension reduction, page rank, dimension reduction, spectral graph theory, spectral clustering, semisupervised learning; Markov state

models, hidden Markov models (HMMs), model reduction for dynamical systems. Computational aspects: sparse matrices; eigenvalues and eigenvectors

- Neural networks: construction, backpropagation; convolutional neural networks.

References

Linear Algebra and Learning from Data, Gilbert Strang.
Numerical Linear Algebra, L.N. Trefethen and D. Bau.
Finite Dimensional Vector Spaces, Holmes.
Foundations of Data Science, Avrim Blum, John Hopcroft, and Ravindran Kannan.
High Dimensional Probability, An Introduction with Applications in Data Science, Roman 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.
Additional references for specific topics will be added.

Grading
Grade to be based on assignments (30%), one midterm (30%) 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 C/Python/MATLAB/R/etc.. will be helpful in several homework sets.

Additional Information
Students from all areas of science, engineering, computer science, statistics, economics and quantitative studies that need advanced level skills in solving problems related to the analysis of data, signal processing, or statistical modeling are encouraged to enroll.