Introduction to Statistical Learning, Data Analysis and Signal Processing - Spring 2017

Instructor	Mauro Maggioni	Office: Krieger Hall 405
Web page	www.math.jhu.edu/~mauro	Course: AS.110.446, EN.550.416

Syllabus

- Introduction to high dimensional data sets. Key problems in statistics and machine learning.

- Review of linear algebra: vector spaces, linear operators. Norms, inner products. Basic function spaces. The Fourier basis, Fourier series, Parseval's theorem. Haar wavelet basis, multiscale decompositions. Examples from approximation theory with the Fourier and Haar bases. Applications to sound and image compression, denoising.

- Review of basic probability: probability spaces, discrete and continuous random variables and their distributions. Overview of law of large numbers and central limit theorem. Mean and covariance. Median(s). Introduction to Brownian motion and its multiscale construction, connections with Fourier and wavelet analysis.

- Singular value decomposition (SVD), low-rank matrices. First applications of SVD: dimension reduction, variances and covariances. Data and matrix compression; applications to computation.

- Curse of dimensionality. Histograms and binning. Estimating probability measures and densities. Concentration of measure phenomena.

- Classification: Nearest neighbor classification, Linear Discriminant Analysis, Classification Trees. Cross-validation.

- Dimension reduction. Random projections; Johnson-Lindenstrauss Lemma. Random matrices, basic concentration inequalities. Introduction to metric spaces. Mappings between metric spaces, Lipschitz maps, distortion; embeddability of metric spaces into Euclidean space, and by trees.

- Regression: problem statement, examples. Curse of dimensionality in regression. Linear regression: least squares; another application of SVD. Regularization. Nonparametric methods. Regression of Lipschitz functions. Fourier and multiscale methods.

- Going nonlinear: kernel methods and their applications to dimension reduction and regression.

- Manifold learning: introduction to manifolds; Isomap, LLE, diffusion maps.

- Graphs and networks: random walks, diffusion. Basic spectral graph theory. Pagerank. Random graphs, Erdös-Rényi graphs, stochastic block models. Clustering: K-means, K-flats, vector quantization, spectral clustering. Semi-supervised learning.

- Algorithmic and computational aspects of the above will be consistently in focus, as will be computational experiments on synthetic and real data. In particular, computational aspects of: SVD; eigenvalues and eigenvectors; Fast Fourier Transform; sparse and full matrices; nearest-neighbors; the kernel trick; multiscale representations of functions and operators.

Grading

Grade to be based on weekly assignments (20%), one midterm and a final exam (20+30%), and a final project (30%). The final project includes a report on the topic of the project, typically involving analyzing a data set with an algorithm, either from class or from the literature, and each project will be conducted by a small group.

Weekly homework problem sets will include theory, analysis and computational exercises. Homework is due at the beginning of class, with name written at the top of the first page, stapled, written legibly, on one side of each page only. Otherwise, it will be returned ungraded. Some problems from the homework may reappear on exams. The lowest homework score will be dropped. No late homework, exams, final projects will be accepted. Johns Hopkins policies apply with no exceptions to cases of incapacitating short-term illness, or for officially recognized religious holiday.

You may, and are encouraged to, discuss issues raised by the class or the homework problems with your fellow students and both offer and receive advice. However all submitted homework must be written up individually without consulting anyone else's written solution.

Prerequisites

Linear algebra will be used throughout the course, as will multivariable calculus and basic probability (discrete random variables). Basic experience in programming in C or MATLAB or R or Octave.

Recommended: More than basic programming experience in Matlab or R; document preparation: LaTeX; some more advanced probability (e.g. continuous random variables), some signal processing (e.g. Fourier transform, discrete and continuous); basic functional analysis.

Resources

The *Instructor* is available during his office hours, in Krieger 405, on Tuesdays from 2pm to 3pm.

Our *Teaching Assistant*, Mr. Zachary Lubberts, is available during his office hours, in Whitehead 212 on Thursdays from 4:30-6:30pm

The *Math help room*, in Krieger 213 I believe, is staffed daily, roughly from 9am-9pm. I found a schedule for last semester at http://www.math.jhu.edu/helproomschedule.pdf: there you may ask questions there about any math topic related to the course.