MATH 191, Fall 2015

Topics in Data Science: Algorithms and Mathematical Foundations

Instructor: Mihai Cucuringu
Time & Location: MWF, 1-1:50pm, GEOLOGY 4645
Course Description: This is a project-based interdisciplinary course covering a number of topics in *Data Science*, that will combine both theoretical and practical approaches. The goal of the course is, on one hand, to understand (at least at a high level) the mathematical foundations behind some of the state-of-the-art algorithms for a wide range of tasks including organization and visualization of data clouds, dimensionality reduction, network analysis, clustering, classification, regression, and ranking. On the other hand, students will be exposed to numerous practical examples drawn from a wide range of topics including social network analysis, finance, statistics, etc. There will be a strong emphasis on research opportunities.

Prerequisites: linear algebra (MATH 33A), elementary probability (MATH 170A) and/or statistics, and computer programming. Familiarity with a programming language such as R or MATLAB is desirable, but students are free to choose their own favorite programming language. Please keep in mind that this course will require that you write a significant amount of code (homework, midterms and final project).

Textbook: The class textbook is

- *An Introduction to Statistical Learning* by James, Witten, Hastie, and Tibshirani,

freely available at http://www-bcf.usc.edu/~gareth/ISL/. Lecture notes and slides will also be posted on the course website. Other readings and publicly available materials will also be made available on the course website.

A list of tentative topics:

1. Introduction & syllabus
2. Review of basic statistics and probability
3. Statistical learning, and introduction to R and several data sets
4. Bias-variance decomposition

Measures of correlation in data:

5. Pearson (sample and population versions), Spearman, Hoeffding's D
6. Maximal correlation, and review of characteristic functions; Distance correlation
7. Information theory (entropy, mutual information), and Maximal Information Coefficient (MIC) (*Detecting Novel Associations in Large Data Sets*, Reshef et al., Science 2011)
8. Simple/multiple linear regression, proof that OLS is BLUE
9. Linear regression - practical considerations
10. Singular Value Decomposition (SVD), rank-k approximation, Principal Component Analysis (PCA)
11. PCA derivation (best d-dimensional affine fit/projection that preserves the most variance)
12. PCA in high dimensions and random matrix theory (Marcenko-Pastur); applications to finance
13. Basics of spectral graph theory

Nonlinear dimensionality reduction methods:

14. Diffusion Maps
15. Multidimensional scaling and ISOMAP
16. Locally Linear Embedding (LLE)
17. Kernel PCA
18. Ranking with pairwise incomplete noisy measurements, and applications; Page-Rank
19. Overview of several ranking algorithms: Serial-Rank, Rank-Centrality, SVD ranking
20. The Angular Synchronization problem and an application to ranking

Clustering:

21. Clustering: K-means and K-medoids, Hierarchical clustering
22. Spectral clustering, isoperimetry, conductance

Modern regression:

23. Ridge regression
24. The LASSO