Foundations of Data Science

Instructor: Mihai Cucuringu

Course Description: This is a 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.

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. Review of basic statistics and probability; introduction to statistical learning
2. Bias-variance decomposition

Measures of correlation in data:

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

Nonlinear dimensionality reduction methods:

11. Diffusion Maps
12. Multidimensional scaling and ISOMAP
13. Locally Linear Embedding (LLE)
14. Kernel PCA
15. Ranking with pairwise incomplete noisy measurements, and applications; Page-Rank, Serial-Rank, Rank-Centrality, SVD ranking

Clustering:

16. Clustering: K-means, Hierarchical clustering
17. Spectral clustering, isoperimetry, conductance

Modern regression:

18. Ridge regression
19. The LASSO