PhD Seminar In Analytics: MGMT 69000
Topics in High-dimensional Data Analysis
Fall 2016

Time and Location: Friday 3:15pm - 6:05pm, Jerry S Rawls Hall 2079

Professor: Jiaming Xu, xu972@purdue.edu, Krannert Building 431

Office Hours: TBD

Course Website: http://web.ics.purdue.edu/~xu972/Course_DataAnalysis.html

1 Course Schedule

- Section 1 (70mins): **3:15pm – 4:25pm**
- Break (10mins)
- Section 2 (70mins): **4:35pm - 5:45pm**

Note: on the following dates

- September 2, 2016: Section 1 will be cancelled, and we will meet at 4:35pm for Section 2
- November 25, 2016: No class due to Thanksgiving Vacation.

2 Overview and Objectives

Today we see a surge of online social networks and e-commerce such as Facebook, Amazon, Twitter and Bitcoin, which generate enormous data. Also, with the advent of high-throughput measurement methods in biology, a large amount of biological data has accumulated. There are many other sources of data such as transportation networks, power networks, and sensor networks. This data contains a wealth of information and it is often desirable to extract useful information from it for various reasons, for instance to predict user preferences, discover disease causes or predict traffic patterns. However, this data is often noisy and voluminous; thus extracting useful information from it requires highly efficient algorithms that can process large amount of data and detect tenuous statistical signatures.

This will be a research-oriented course designed for graduate students with an interest in doing research in theoretical aspects of high-dimensional data analysis. Two central questions will be addressed in this course:

1. How shall we characterize the limit above which the task of extracting information is fundamentally possible and below which it is fundamentally impossible?

2. How shall we develop computationally efficient algorithms that attain the fundamental limit, or understand the lack thereof.

This course aims to familiarize students with advanced analytical tools such as concentration of measures, probabilistic methods, information-theoretic arguments, convex duality theory, and random matrix theory, by going over a number of emerging research topics in the area of high-dimensional data analysis such as data clustering, community detection, submatrix localization, sparse PCA, learning graphical models, and fast algorithms for linear algebra.

3 Prerequisites

Maturity of linear algebra and probability is required. Some knowledge of the basic optimization and algorithms is also recommended.

4 Grading

- **30% Homework** (late homework will not be accepted). There will be five homework in total, each counting for 6%.
- **20% Scribe**. There will be two scribes randomly assigned to each student, each counting for 10%.
- **10% Attendance**. You will be expected to attend every lecture.
- **40% Final project**. You will be either presenting a paper or a standalone research project. The potential topics for projects will be sent out by October 14.

5 Tentative Outline

The topics below intersect many disciplines: Mathematics, Computer Science, Electrical Engineering, Statistics, Operations Research, and Statistical Physics.

Lecture notes and additional reading materials will be posted online.

Part I: Clustering

1. **Introduction**: Examples of high-dimensional data analysis and applications
2. **k-means clustering**: Optimization formulation of k-means, convergence of k-means, failure cases of k-means, model-based formulation of k-means, maximum likelihood estimation and EM algorithm for Gaussian mixtures, soft k-means
3. **Review on linear algebra**: eigenvalue decomposition of symmetric matrices, singular value decomposition and best-fit subspace, Frobenius norm, Spectral norm, best low rank matrix approximation, spectral relaxations of k-means, principal component analysis
4. **Concentration inequalities**: Markov inequality, Chebyshev's inequality, Chernoff's bound, Sub-Gaussian random variables, Sub-Exponential random variables, Bernstein inequality, Symmetrization technique, Gaussian isoperimetric inequality
5. **Matrix concentration inequalities**: Wigner semi-circle law, Gaussian comparison inequality, ϵ-net method, Matrix Bernstein inequality
6. **Spectral clustering**: Spectral clustering under Gaussian mixture model, spectral clustering based on Laplacian matrix, perturbation theory for linear operators, Davis-Khan $\sin\theta$ theorem

Part II: Community detection

1. **Random graphs** $\mathcal{G}(n,p)$: giant component, Branching process, connectivity threshold
2. **Panted models**: stochastic block model, planted partition, planted clique

3. **Spectral graph clustering**: spectrum of random graph, failure of naive spectral methods in sparse graph, spectral barrier in planted clique

4. **Information-theoretic tools**: Mutual information, Kullback-Leibler (KL) divergence and operational characterizations, data processing inequality, Fano's inequality, derivation of recovery limits of planted models

5. **First and second moment methods**: Binary hypothesis testing, likelihood ratio test, generalized likelihood ratio test, total variation distance, Hellinger distance, chi-square divergence, first moment method, second moment method, derivation of detection limits, Information-computation gap

6. **Semidefinite relaxations for community detection**: convex duality theory, performance analysis of SDP relaxations, Grothendieck's inequality, SDP in real-world networks

7. **Belief propagation for community detection**: locally tree-like argument, density evolution, fixed-point analysis

8. **Submatrix localization**: Information limits and algorithmic limits

9. **Covariance matrix estimation**: Spiked covariance model, BBP phase transition, sparse PCA

Part III: Graphical models and message passing

1. **Graphical model representation**: Bayesian networks, pairwise graphical models, factor graphs, Markov random fields

2. **Inference via message passing algorithms**: Tree networks, sum-product algorithm, max-product algorithm, Ising model, Hidden Markov model

3. **Variational methods**: Free energy and Gibbs free energy, naive mean field, Bethe free energy

4. **Bayesian inference**: Gaussian mixture and community detection as spin glass

5. **Learning graphical models**: Chow-Liu's algorithm on trees

Part IV: Other selected topics

1. **Randomized linear algebra**: matrix-vector product, matrix multiplication, low-rank approximation

2. **Ranking**: BTL model, Plackett-Luce model, maximum likelihood estimation, EM algorithm

3. **DNA sequencing**: DNA scaffolding, hidden Hamiltonian path problem

6 References

- R. Kannan and S. Vempala, Spectral Algorithms.
`http://web.stanford.edu/class/ee378b/papers/kannan_vempala.pdf`

- Ulrike von Luxburg, A Tutorial on Spectral Clustering.
`http://arxiv.org/pdf/0711.0189v1.pdf`

- Matthias Hein and Ulrike von Luxburg, Graphdemo.
  http://www.ml.uni-saarland.de/code/GraphDemo/GraphDemo.htm
- J. Tropp, User-friendly Tail bounds for Sums of Random Matrices.
  http://web.stanford.edu/class/ee378b/papers/tropp_martingale.pdf
- John Hopcroft and R. Kannan, Foundations of Data Science.
  http://research.microsoft.com/en-US/people/kannan/book-no-solutions-aug-21-2014.pdf
- David J. Mackay, Information theory, inference, and learning algorithms.
  http://www.inference.phy.cam.ac.uk/itila/

7 Related classes

- Inference, Estimation, and Information Processing. Andrea Montanari.
  http://web.stanford.edu/class/ee378b/ee378b.html
- Topics in Mathematics of Data Science. Afonso S. Bandeira.
  http://math.mit.edu/~bandeira/Fall2015.18.S096.html
- Information-theoretic methods in high-dimension. Yihong Wu.
  http://www.ifp.illinois.edu/~yihongwu/teaching/598/