Syllabus for STAT 236 Statistical Learning

Spring 2020

Course Information:

- Time and location: 3pm – 4:15pm Mon. Wed., Science Center 705
- Instructor: Tracy Ke (zke@fas.harvard.edu)
- Teaching Fellow: Yucong Ma (yucongma@g.harvard.edu)

Course Description:

High-dimensional data analysis is a recent interdisciplinary research area of Statistics, Genetics and Genomics, Engineering, and several other scientific areas. It addresses an array of challenging problems of contemporary interest, and research in this area has been very active in the past decade.

This course aims to provide a systematic introduction to various topics in high dimensional data analysis, focusing on large-scale sparse learning, network data analysis and text data analysis.

- Large-scale sparse learning: Sparsity is a universal phenomenon in modern high dimensional data. Sparse structures are observed in many application settings and have many different forms, such as parameter sparsity, graph sparsity, eigenvalue sparsity, and so on. Exploring sparsity has become a common strategy in data analysis and has largely reshaped classical multivariate statistics problems. This course will investigate classical problems such as multiple testing, linear regression, classification and clustering, under the modern sparse settings. For each problem, the course discusses recent statistical methods for taking advantage of sparsity, and introduces the theoretical framework for analyzing these methods.
- Network and text data analysis: Social networks and text documents are unconventional data types. This course introduces statistical models and methods for analyzing such types of data.
  - Topics for network data analysis include community detection, mixed membership estimation, link prediction, and dynamic network modeling.
  - Topics for text data analysis include topic modeling, word embedding, information retrieval, and sentiment analysis.

Who will be interested:

This course is designed for two different groups:

- Students who have little knowledge of high-dimensional statistics or machine learning and are interested in learning the problems, concepts, and methods:

  For example, this course will talk about:

  - Why is exploring sparsity a main strategy in high-dimensional statistics?
  - Three different classes of variable selection methods: penalization, screening & cleaning, and stage-wise algorithm.
  - What are the main methods for covariance matrix estimation? What are the main methods for precision matrix estimation?
  - What is sparse PCA? What is nonnegative matrix factorization?
  - What are the techniques for network data analysis? What is the stochastic block model? What is the modularity method? What is the spectral method?
  - How to use the micro-structures in social networks, such as counts of triangles and short paths, for statistical inference?
  - The topic models for modeling text corpora. What is latent dirichlet allocation? What is the meaning of anchor words?
  - Neural network models for text data.
  - and more ...

- Students or researchers who have learnt these topics but want to have deeper understanding and prepare themselves for research in this area:

  For example, this course will talk about:

  - How to derive oracle inequalities and variable selection consistency for lasso and penalization methods?
  - Are many popular methods (lasso, BH's FDR control) working when the signals are not only sparse but also very weak? The Rare/Weak signal model and phase transitions in multiple testing and variable selection.
  - Large-deviation inequalities (Hoeffding, Bernstein, Azuma, matingale inequalities, etc.).
  - Minimax optimality theory, for testing and for estimation, respectively.
  - Behaviors of empirical eigenvalues of Wishart and Wigner matrices and how to use them in statistics inference. Behaviors of empirical eigenvectors.
  - Non-asymptotic bounds for matrix norms.
  - Phase transitions in Erdos-Renyi models and stochastic block models.
  - How to deal with the severe degree heterogeneity in network data analysis?
  - and more ...

  Not all technical details will be given in lectures, but you will learn the main ideas and know where to find these tools when you need them in research.

Grading:

The grade is based on:

- Attendance and participation: 20%
- Homeworks: 40%
- Final project and presentation: 40%

The homework is posted **weekly**. Each assignment has **two problems**: Problem 1 is True or False (with explanations). Problem 2 is either a theoretical problem or a data analysis problem. **For undergraduates, some theoretical problems are optional.**

The final project is group assignment, for **a group of 3-4 students**. A project can be literature review about a particular research problem, implementation and comparison of existing methods, a new application of existing methods on real data, or an extension of a research paper. Students are required to form the group before Week 7 and get the approval of the project topic from the instructor before Week 10. **Each group is required to submit a report and prepare a 20-minute presentation**. The project should be related to course materials. Students cannot use their own research works as the course project.

Prerequisites:

For graduate students in Statistics, the prerequisites are STAT 211, STAT 213 (students can take the prerequisites simultaneously with this course).

Graduate students from other departments (Computer Science, Biostatistics, Economics, etc.) can take this course if they have taken courses in statistics or probability or machine learning or econometrics that has some amount of content in mathematical statistics.

For undergraduate students, the prerequisites are calculus, linear algebra, and statistics. Detailed prerequisites will be given in the first lecture. Grading for undergraduate students is different from that of graduate students.

Text book and Readings:

There is no required text book. The materials will be given in the lecture slides. The slides for each chapter are built on 5-10 research papers. Below is an example list of papers talked about in some chapters:

Example 1: Multiple testing.

- Efron, B. (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. Journal of the American Statistical Association, 99(465), pp.96-104.
- Donoho, D. and Jin, J. (2004) Higher criticism for detecting sparse heterogeneous mixtures. The Annals of Statistics, 32(3), pp.962-994.

- Donoho, D. and Jin, J. (2015) Higher criticism for large-scale inference, especially for rare and weak effects. Statistical Science, 30(1), pp.1-25.
- Fan, J., Han, X. and Gu, W. (2012) Estimating false discovery proportion under arbitrary covariance dependence. Journal of the American Statistical Association, 107(499), pp.1019-1035.
- Hall, P. and Jin, J. (2010) Innovated higher criticism for detecting sparse signals in correlated noise. The Annals of Statistics, 38(3), pp.1686-1732.

Example 2: Variable selection.

- Bickel, P., Ritov, Y. and Tsybakov, A. (2009) Simultaneous analysis of Lasso and Dantzig selector. The Annals of Statistics, 37(4), pp.1705-1732.
- Donoho, D. and Stark, P. (1989) Uncertainty principles and signal recovery. SIAM Journal on Applied Mathematics, 49(3), pp.906-931.
- Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association, 96(456), pp.1348-1360.
- Zhao, P. and Yu, B., 2006. On model selection consistency of Lasso. Journal of Machine learning research, 7(Nov), pp.2541-2563.
- Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. The Annals of statistics, 32(2), pp.407-499.
- Fan, J. and Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(5), pp.849-911.
- Ji, P. and Jin, J. (2012) UPS delivers optimal phase diagram in high-dimensional variable selection. The Annals of Statistics, 40(1), pp.73-103.
- Jin, J. and Ke, Z.T. (2016) Rare and weak effects in large-scale inference: methods and phase diagrams. Statistica Sinica, pp.1-34.
- Zhang, T. (2011) Adaptive forward-backward greedy algorithm for learning sparse representations. IEEE transactions on information theory, 57(7), pp.4689-4708.

Some text books are also helpful but not required:

- Friedman, J., Hastie, T. and Tibshirani, R. (2009). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.
- Wainwright, M.J. (2019). High-dimensional statistics: A non-asymptotic viewpoint (Vol. 48). Cambridge University Press.

Topics covered:

It may change depending on how the course goes.

- Multiple Testing
  - Stein's normal means models, shrinkage estimators, sparsity, thresholding estimators
  - Rare/Weak signal model, phase diagram
  - Global testing (chi-square test, maximum test, higher criticism test)
  - Multiple testing with dependent noise

- Variable Selection
  - Penalization methods (L_0/L_1 methods, non-convex methods)
  - Greedy algorithms (LARS, forward/backward selection)
  - Screen and Clean methods
  - Statistical error bounds for parameter estimation
  - Phase diagram for variable selection
- Covariance and precision matrix estimation (2 lectures)
  - Large covariance matrix estimation (thresholding, banding, statistical error bounds)
  - Precision matrix estimation and graphical models (graphical lasso, regression methods)
- Nonparametric estimation
  - Nonparametric estimators (kernel density estimators, nonparametric regressions)
  - Lower bounds on the minimax risk
- Classification
  - High-dimensional linear discriminant analysis (feature selection, statistical limits of classification)
  - Empirical risk minimization methods, VC theory
- Unsupervised learning
  - Sparse PCA (consistency, trade-off between statistical errors and computational complexity)
  - Spectral clustering
  - Nonnegative matrix factorization
- Network data analysis
  - Review of network models (block models, ERGM, graphons)
  - Stochastic block models and recent mathematical theories
  - Degree-corrected network models, the SCORE methods for community detection and mixed membership estimation
  - Phase transitions for network community analysis
  - Dynamic network modeling
- Text data analysis
  - Topic modeling
  - Word embedding
  - Information retrieval
  - Sentiment analysis