Syllabus for STAT 236 Statistical Learning

Spring 2019

Course Information:

- Time and location: 1:30pm – 2:45pm Tue. Thu., SC 705
- Instructor: Tracy Ke (zke@fas.harvard.edu)
- Teaching Fellow: Minzhe Wang (mwang@g.harvard.edu)

Course Description and Prerequisites:

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.

This course is designed 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 (CS, Biostatistics, Economics, etc.) can also take this course, if they have taken statistics-related courses which have a significant amount of content in mathematical statistics. Undergraduate students should consult the instructor for prerequisites.

Workload:

There will be three assignments. They are posted in Week 4, Week 7 and Week 11 (approximately), and students have two weeks to complete each assignment. Each assignment includes written problems and data analysis problems. There is no midterm.

- Attendance and participation: 10%
- Homework: 60%
- Final (a three-hour in-class exam): 30%

Topics

Below is a tentative list of topics. It may change depending on how the course goes.

- Multiple Testing (3 lectures)
  - 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 (4 lectures)
  - 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 (2 lectures)
  - Nonparametric estimators (kernel density estimators, nonparametric regressions)
  - Lower bounds on the minimax risk

- Classification (2 lectures)
    - High-dimensional linear discriminant analysis (feature selection, statistical limits of classification)
    - Empirical risk minimization methods, VC theory
- Unsupervised learning (3 lectures)
    - Sparse PCA (consistency, trade-off between statistical errors and computational complexity)
    - Spectral clustering
    - Nonnegative matrix factorization
- Network data analysis (4 lectures)
    - 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 (3 lectures)
    - Topic modeling
    - Word embedding
    - Information retrieval
    - Sentiment analysis
- Reviews, discussions, and open problems (1 lecture)