STAT 905: High Dimensional Data and Statistical Learning (Fall 2014)

Instructor
Weixing Song, Ph.D.
Associate Professor
Department of Statistics
Kansas State University

Office: 108F Dickens Hall
Phone: 785-532-0524
Email: weixing@ksu.edu
Web: http://weixingsong.weebly.com/

Course Time & Location
2:30 PM – 3:45 PM, Tuesday & Thursday, 302 Dickens Hall.

Office Hours
10:00 AM – 11:00 AM, Monday & Wednesday, or by appointment.

References

- Published journal articles.
- Hastie, Tibshirani and Friedman (2009). *The Element of Statistical Learning*. Springer. (Free electronic version available through K-State library)
- Bühlmann, P., van de Geer, S. (2011). *Statistics for High-dimensional Data*. Springer.
- Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques*. Springer.
- Efron, B. (2010). *Large-Scale Inference*. Cambridge.
- Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. Springer.

Prerequisites
STAT 713 and STAT 771, plus one introductory course in statistical computing (e.g. STAT 726 or equivalent background).

Course Materials and Website
Announcements, lecture notes and other course information will be posted on K-State Online. You are expected to bring your copy of the lecture notes before each class, and read the assigned reading materials on each topic as we proceed.

Course Objectives
Vast amounts of data are being generated in many fields as science and technology advance. The challenges we are facing in learning from large-scale/high-dimensional data

have been reshaping statistical thinking, methodological development, and theoretical studies. Such has led to an expanding field of "statistical learning and data mining". The main goal of the course is to equip students with a good understanding of many key techniques and theoretical ideas in statistical learning developed at the interface of statistics, computer science and artificial intelligence. The emphasis will be on the statistical aspects of these methods and their integration with many standard statistical methodologies. The course will be suitable for graduate students and practitioners from many disciplines, and will prepare students in Statistics to do research in this dynamic and fast-growing area.

Course Description

This course covers methodological development, theoretical analysis and computation implementation of several important statistical learning methods for large scale data and high dimensional settings. Some cutting-edge developments on sparsity-inducing regularization methods will be highlighted. General topics include supervised and unsupervised statistical learning techniques for prediction and pattern recognition; methods for model selection, multiple testing control, and estimation in high-dimensions; applications in various fields of science and engineering.

A tentative list of topics to be covered is as follow.

- Introduction to high dimensional statistics.
  - Framework and overview of high-dimensional inference
  - Curses and blessings of high dimensionality
- Statistical decision theory.
  - Framework of statistical learning
  - Local methods in high dimensions
  - Empirical risk minimization and penalized ERM
- Linear regression, shrinkage estimation and L_0 penalization.
  - Classical linear regression
  - Wiener filter and L_2 shrinkage estimation
  - Subset selection and L_0 penalization
- Variable selection via L_1 regularization.
  - Introduction to Lasso
  - Orthogonal design and geometry
  - Computation via coordinate descent

  - Least angle regression
- Theory of L_1 regularization.
  - Shrinkage and thresholding estimators
  - Unveiling the mystery of "oracle inequality"
  - Estimation and prediction properties
  - Selection consistency and irrepresentable condition
- Variants and extensions of sparse estimation.
  - Adaptive Lasso, bridge regression and elastic net
  - Group/bilevel selection
  - Concave penalties
- Model selection and assessment.
  - Error rate estimation
  - Derivation of AIC and BIC
  - Cross validation and bootstrap
- Multivariate models in high dimensions.
  - Reduced-rank regression and canonical correlation analysis.
  - Nuclear norm penalization
  - Low rank and sparse estimation
- Classification models and high dimensional classification methods.
- Matrix decomposition methods, including SPCA, SSVD, etc.
- Large-scale hypothesis testing and false discovery rate control.
- Other selected topics, e.g., boosting, neural network, SVM, clustering and more (if time permits).

Computing
We will mainly use the free statistical computing environment R.

Attendance
Attendance is required. The penalty for absence is at the discretion of the instructor and may include deduction of the final grade and failure of the course.

Grading (tentative)

Homework	40%
Midterm Exam	25%
Project and Presentation	35%
Total	100%

Grades for the course are assigned totally at the instructor's discretion. As a **rough** guide: >90%, A; 80–90%, B; 70–80%, C; 60–70%, D; <60%, F.

- Exams: There will be one in-class midterm exam. More details will be announced later.
- Homework: Homework will be assigned throughout the semester. You may discuss the homework assignments with each other, but the submitted work must be entirely you own. Unless prior arrangements are made for reasons judged to be acceptable by me, late homework will receive ZERO credits.
- Final Project: Each student enrolled in this course is expected to complete a class project and give an in-class presentation. More details will be announced later.

Disclaimer
This syllabus is tentative. The instructor reserves the right to change the syllabus as necessitated by circumstances.

Updated: 08/20/2014

KSU Required Statements on Course Syllabus

Academic Honesty

Kansas State University has an Honor System based on personal integrity, which is presumed to be sufficient assurance that, in academic matters, one's work is performed honestly and without unauthorized assistance. Undergraduate and graduate students, by registration, acknowledge the jurisdiction of the Honor System. The policies and procedures of the Honor System apply to all full and part-time students enrolled in undergraduate and graduate courses on-campus, off-campus, and via distance learning. The honor system website can be reached via the following URL: www.ksu.edu/honor. A component vital to the Honor System is the inclusion of the Honor Pledge which applies to all assignments, examinations, or other course work undertaken by students. The Honor Pledge is implied, whether or not it is stated: "On my honor, as a student, I have neither given nor received unauthorized aid on this academic work." A grade of XF can result from a breach of academic honesty. The F indicates failure in the course; the X indicates the reason is an Honor Pledge violation.

Accommodations for Students with Disabilities

Any student with a disability who needs a classroom accommodation, access to technology or other assistance in this course should contact Disability Support Services (dss@k-state.edu) and/or the instructor.

Classroom Conduct

All student activities in the University, including this course, are governed by the Student Judicial Conduct Code as outlined in the Student Government Association By Laws, Article VI, Section 3, number 2. Students that engage in behavior that disrupts the learning environment may be asked to leave the class.