STAT 718: High-Dimensional Data

Spring 2023

Instructor:	Ray Bai	**Time:**	MWF 1:10 PM – 2:00 PM
Email:	RBAI@mailbox.sc.edu	**Place:**	LeConte 103

Course Page:

https://blackboard.sc.edu/ (Check regularly for announcements and homework assignments)

Office Hours: By appointment.

Course Description: Modern data mining problems often deal with large and complex data sets, otherwise known as "big data." Typically, these models contain a large number of parameters (possibly exceeding sample size) and/or a very large number of samples. In this course, we will study these models for both supervised and unsupervised machine learning. Topics covered include penalized regression and classification, convex and nonconvex optimization, recommender systems, graphical models, and deep learning.

We will examine key ideas from optimization theory (e.g. convexity and duality) and standard optimization algorithms, including coordinate descent and gradient descent. Practical issues such as the bias-variance trade-off, evaluation of predictive accuracy, and the selection of tuning parameters are also discussed. Statistical computing and programming are key components of the course.

Prerequisites: Probability and statistics at the level of STAT 511-512 and linear algebra at the level of MATH 544. Students should also be comfortable with a programming language such as R, Python, MATLAB, or C++. These are **strict** requirements. Students without this prerequisite knowledge and/or who do not have sufficient programming experience may struggle in this class.

Tentative Schedule of Topics:

- **Week 1-3**: methods for evaluating predictive accuracy, bias-variance trade-off, linear regression, regression trees, random forests, penalized regression (ridge regression, lasso, nonconvex penalties), coordinate descent algorithm, cross-validation
- **Week 4-5**: generalized linear models (GLMs) and penalized GLMs, classification trees, binary and multiclass classification
- **Week 6-9**: convex and nonconvex optimization (unconstrained and constrained), gradient descent, momentum-based accelerated methods, proximal gradient descent, stochastic gradient descent, variance reducing gradient, ADMM algorithm, Frank-Wolfe algorithm
- **Week 10-11**: dimension reduction, sparse and robust PCA, and recommender systems
- **Week 12**: graphical models
- **Week 13-14**: deep learning, feed-forward neural networks, backpropagation algorithm, generative adversarial networks (GANs)
- **Week 15**: group project presentations

Computing: Students may use one (or more) of the following languages: Python, R, MATLAB, or C/C++. Using a more obscure language is discouraged because I will not be as familiar with it and thus may be less able to help you debug your code.

Course Materials: We will use typed handouts prepared by the instructor. Parts of these lecture notes are *not* complete and will be filled in during lecture. Thus, it is in your best interest to attend every lecture.

Learning Outcomes:

1. Demonstrate the ability to analyze high-dimensional data and compare the performance of different methods.
2. Understand the methodology and theory underlying high-dimensional data analysis.
3. Be able to implement algorithms for analyzing high-dimensional data.
4. Be able to communicate effectively through writing scientific reports and making/delivering presentations.

Homework: There will be five or six homework sets. Students will work in small groups, and each group will submit a single, typed report for each homework assignment. The homework will consist of both conceptual/theoretical exercises and questions that involve programming and data analysis.

All homework reports should be typed, including answers to math exercises and data analysis portions that may include figures, tables, etc. Homework reports **and** code are to be submitted through Blackboard. The reports should also briefly summarize the individual contributions of each team member.

The first few weeks of the class, groups should scour the Internet for a homework template to use. It is expected that students will be able to learn Latex on their own by following existing templates and using the Internet.

Final Project: Students will work in small groups to read one or two papers on a topic of their choosing, prepare a 15-20 minute presentation, and write a short report in the style of a journal article: abstract, introduction, method, simulations and data analysis, and a bibliography. The last week of the semester will be devoted to project presentations.

Students are encouraged to pursue projects that are relevant to their current research or their research interests. Projects must be approved in advance by the instructor, and no two groups may do the same topic for their project. If you have an idea of what you want to do for your project, please "claim" it early. Detailed instructions for the presentation and report will be given at a later date.

Grading Policy: Homework (70%), group project and presentation (30%). The grading scale is as follows: 90-100 for an A, 80-89 for a B+, 70-79 for a B, 60-69 for a C+, 0-59 for a C.

Honor Code: See the Carolinian Creed in the *Carolina Community: Student Handbook and Policy Guide*. The *minimum* punishment for violations of the USC Honor Code is a grade of zero for the work in question. In accordance with university policy, there may be other punishments including an automatic F in the class and/or expulsion from the university.

Accommodation: If you need special accommodations for examinations or any other aspects of the course, please contact me before or during the first week of the semester. Note that reasonable accommodations are available for students with a documented disability. If you have a disability and may need accommodations to fully participate in this class, contact the Office of Student Disability Services by phone (803-777-6142), email sasds@mailbox.sc.edu, or stop by LeConte College Room 112A. All accommodations must be approved through the Office of Student Disability Services.