SCHOOL OF MEDICINE & DENTISTRY
UNIVERSITY of ROCHESTER MEDICAL CENTER

BST432 – High Dimensional Data Analysis

Fall 2017
Saunders 1.402, MW 9:15A-10:45A

Instructor(s): Andrew McDavid
Office Hours: Weds 1-3PM (and by appointment) Saunders 4.206
Course website and discussion board: learn.rochester.edu
Prerequisites: BST 401 and BST 411. BST430 (Introduction to Statistical Computing) or equivalent highly recommended.

Course Description

The exponential growth in computing capacity over the past 20 years has generated dramatic increases in the quantity and breadth of data in the clinical and basic biomedical sciences. Multiple domains have seen this exponential growth, including nucleic acid sequencing, proteomics and data warehoused in the electronic medical record. Fortunately, as statisticians, we are equipped to synthesize and rigorously test hypotheses in these frontiers. This course will provide an overview of modern tools from machine learning, with a particular focus on connecting them to their statistical underpinnings. These tools will be illustrated on biomedical data sources, including high-throughput DNA, RNA, metagenomic, immunologic and electronic medical record data, so students will gain a basic understanding of the challenges associated with these data types. Emphasis will be placed on understanding both benefits and limitations in high-throughput, high-dimensional technologies and how these affect statistical and scientific inference.

Facility with a data-centric programming language (e.g. R/Bioconductor) is crucial to high-dimensional data analysis and will be developed further via class assignments. An intermediate familiarity with basic concepts of computer science (iteration, recursion, computational complexity), numerical linear algebra and statistical computing will be assumed.

Course Aims and Objectives

Prediction (of an unknown label, or a new data point) and **inference** (of the parametric state of a system) are complementary goals in high dimensional data analysis and will be used as organizing themes in the course. **Computational algorithms** are important to accomplish these goals, and will be emphasized. We will address these themes in a series of topics:
1. **Classification and decision theory**. Logistic regression vs LDA vs KNN. Bayes' risk.
2. **Regression**. Model selection and variable screening. AIC/BIC and subset selection. The lasso. Theoretical guarantees. Bias-variance tradeoff. Extensions to the lasso.
3. **Estimating and evaluating classifier performance**. Cross-validation and bootstrapping. AUC, precision-recall, accuracy, and other low-dimensional statistics. Over- and under- sampling of classes. The only place I care about bias.
4. **Multiple and sequential testing procedures**. Union bounds, family-wise error rate (FWER), false discovery rates (FDR) control. Permutation, bootstrapping, current work (knockoffs, etc).
5. **So, how do we fit these things, anyways**? Unconstrained optimization. Local vs global minima. Some convex geometry. Ill-conditioned vs well-conditioned problems. Constrained optimization. KKT conditions. Line searches and descent directions. Coordinate descent, gradient descent, newton's

methods, majorization-minimization, stochastic gradient descent. Connections to gradient boosting and early stopping.
7. Advanced classification/regression—SVM, neural nets, boosting, bagging. MLR.
8. **Clustering and unsupervised classification**. Methods: PCA, Kmeans, hierarchical clustering, finite mixture models, non-parametric density-based. Manifolds. TSNE. Diffusion maps. Evaluation: stability, silhouettes, gaps, information criteria. Application to flow cytometry and scRNAseq.
9. A project applying a clustering or classification algorithm.

Because this is a biostatistics course we will frequently **apply these techniques to high-dimensional data** generated from high-throughput assays, possibly including genomics (DNA), transcriptomics (RNA), cellular assays (flow cytometry) and patient-level electronic medical record data.

Course Policies and Expectations

At the conclusion of the course, students will be able to select and apply an appropriate technique to answer scientific or business questions from a high-dimensional data set. The student will understand some of the standard techniques and limitations of these techniques for analysis of common genomic data sets.

Materials and Access

Required texts

Many readings and assignments will be derived from "The Elements of Statistical Learning" second edition, which is available within the UR network as a freely-downloadable PDF at *https://link.springer.com/book/10.1007%2F978-0-387-84858-7* and in print at the bookstore. We will also reference several journal articles which will be made available on the course website.

Required software

A high-level programming language will be required for homework requiring coding. For some assignments, R 3.4/Bioconductor 3.5, in particular, will be required.

Course Website

Course material and announcements will be posted on Blackboard at learn.rochester.edu. Registered students should already be enrolled in Blackboard.

Assignments and Grading Procedures

You will be evaluated in terms of **homework** (statistical and computational) (40%), **mid-term** (30%), and a **project/presentation** (30%). Homework will be submitted on blackboard. Late homework is penalized at 25% per 24 hours late.

Project

The project will consist of demonstrating understanding and application of two classification or regression techniques to two sets of real data. One of these techniques will be selected from a menu of options I will provide you; the other can also come from this menu, or can be another one you identify. One data set will be supplied by me and used in common by the entire class. You, ideally, will supply another data set, but I can point you towards potential data sets if necessary. You will produce a written report and make several class presentations.

Academic Integrity and Programming Exercises

You are encouraged to work together on mathematical exercises. You should give credit to your collaborators and your final written solution must be your own work. For programming exercises, you **are not** generally permitted to copy and paste your classmates' or the internet's code, except where indicated on assignments. You may consult with your classmates and external resources on algorithmic and implementation details, but the code you submit must have been typed into your editor, with your fingers, the hard way.

Academic integrity is a core value of the University of Rochester. Students who violate the University of Rochester University Policy on Academic Honesty are subject to disciplinary penalties, including the possibility of failure in the course and/or dismissal from the University. Since academic dishonesty harms the individual, other students, and the integrity of the University, policies on academic dishonesty are strictly enforced. For further information on the University of Rochester Policy on Academic Honesty, please visit the following website:
http://www.rochester.edu/college/honesty/docs/Academic_Honesty.pdf

Accommodations for Students with Disabilities

Students needing academic adjustments or accommodations because of a documented disability should contact the Disability Resource Coordinator for the School of Medicine and Dentistry:
School of Medicine and Dentistry, Office of Graduate Studies
Linda Lipani, Administrator
Med G-7522, Box 601
275-7288

Tentative Course Schedule

Week	Monday	Wednesday
August 30	----	- First day of class, syllabus and objectives. Overview.
September 4,6	**Labor Day**	Classification, generative models and risk.
September 11,13:	Classical classification models. Roses and thorns of dimensionality. Regression I.	Generalizability via cross-validation, bootstrap. Over/under sampling. HW1 due
September 18,20:	Right and wrong cross-validation. Cross-validation as black-box optimization. Regression II and model selection.	HW 2 due
September 25,27:	Lasso and friends. Multiple	**Project proposal due**

	testing and inference	
October 2,4:	Lower level methods for transcriptomics.	HW3 due.
October 9,11:	Bioconductor, annotations, data wrangling	HW 4.
October 16, 18	Differential expression, designs, batch effects and meta analysis	**Mid term**
October 23,25:	Other sequencing-based assays (Chip-seq, ATACseq)	Optimization I: Unconstrained, line searches. HW5.
November 6,8	**Project lightning talks**	
November 13,15:	Optimization II, III: stochastic gradient descent, constrained methods, proximal methods	**Rough draft of paper**
November 20,22	Clustering	HW6
November 27,29	Clustering	HW7
December 4,6	Flow cytometry and single cell gene expression.	HW8
December 11,13	**Presentations**	**Presentations**

Other potential topics as time and interest allow and demand:
-markdown, git, debugging, ssh, distributed computing (slurm, Amazon ec2)
-reproducibility, data sharing
-experimental design for high-throughput assays
-databases, out-of-core storage
-graphical models