BIOST 546
Machine Learning for Biomedical and Public Health Big Data
Syllabus

Instructor: Ali Shojaie, PhD, Associate Professor of Biostatistics
Office: HSB F642
Phone: 616-5323 (Biostatistics)
Email: ashojaie@u.washington.edu

TA: Asad Haris
Email: aharis@u.washington.edu

Time and Place: Tuesdays, Thursdays 1:30-2:50pm HSB RR-134

Office hours:

Instructor: Tuesdays 12:30-1:30pm or by appointment

TA: Wednesday 10:30-11:50am at South Campus Center (SCC) 303, except the following dates:

- On 4/5, 5/3 & 5/31 TA OH will be held in Foege S060 (Genome Sciences)

Class Web Pages: http://faculty.washington.edu/ashojaie/teaching/ML.html

Course Objectives
This course provides an introduction to statistical machine learning methods for analysis of Biomedical Big Data, including high dimensional regression and classification methods, variable selection techniques, high dimensional inference, clustering and dimension reduction methods.

Learning Objectives
By the end of this class, students should be able to:

1. Identify the appropriate statistical model for analysis of Biomedical Big Data – including *omics* data (gene expression, RNA-seq, ChIP-seq, …), electronic health records (EHR) and imaging data – based on the properties of the data and scientific hypotheses and/or research questions.
2. Identify challenges and potential pitfalls corresponding to the application of statistical models in high-dimensional settings.
3. Apply appropriate dimension reduction techniques for data visualization and pattern discovery.
4. Choose and apply appropriate clustering methods to identify subgroups of patients and/or components of biological systems with orchestrated activity.
5. Apply graphical modeling tools to discover novel patterns in the data, and interpret the identified associations.
6. Choose and apply appropriate high dimensional (generalized) linear regression methods for flexible prediction of biomedical phenotypes based on different biomedical data.
7. Choose and apply appropriate statistical models to classify subjects into categories (e.g. patient vs. healthy) using diverse biomedical data.
8. Apply regularization and variable selection techniques for biomarker discovery and improved prediction/classification performance.
9. Assess the model fit in high dimensional settings using cross-validation and measures of model complexity.
10. Develop and execute strategies to prevent over-fitting and control false positives.

Prerequisites: Knowledge of statistical inference at the level of BIOST 511-12 and familiarity with computing or permission of the instructor.

Computing Software: The in-class examples and labs will use the R programming language; students are welcome to use any other computing software/language of their choice, but only R will be supported.

Text: There are a number of good textbooks on statistical machine learning, which can be used as reference for the course. We will use material from the following texts (in the order of relevance):

1. Introduction to Statistical Learning by James et al (2013)
2. Elements of Statistical Learning by Hastie et al (2009)
3. Machine Learning by Murphy (2012)

Grading: Graded (3 credits)

Assessment:
Homework (5-7 total): 30%
Exam: 30%
Project: 40%
The projects should ideally involve applications of statistical machine learning to biomedical big data, and be related to students' research areas. Each team will consist of a group of 2 (at most 3) students working on the same project, and will complete the following steps:

a. project abstract (proposal & progress report): 5%
b. proposals: 5%
c. project presentations: 10%
d. final report: 20%

Important Notes:

1. Class material, including lecture notes, homework assignments, and other course-related information will be posted on the webpage. Printed course material will not be provided by the instructor. Please check the webpage regularly for updated class material.
2. Questions and discussions are welcome, and encouraged throughout the class; keep in mind that if there is something that is not clear to you, it most likely is unclear to others as well.

Tentative Schedule and Topics Covered:

1. Overview of machine learning and its applications in biomedical sciences
2. Practical considerations (data sources, sampling, missing values)
3. Bias-variance tradeoff and cross validation
4. Penalized regression for high dimensional biomedical data
5. High dimensional classification (penalized logistic regression, support vector machines, naïve Bayes classifiers and tree-based methods, KNN)
6. Ensemble learning (bagging, boosting and the bootstrap)
7. Penalized generalized linear models (GLMs) and survival analysis for Biomedical Big Data
8. Multiple comparison adjustment and high dimensional hypothesis testing
9. Dimension reduction (PCA, MDS) and applications to biomedical Big Data
10. Clustering (hierarchical, k-means and model-based clustering) for Biomedical Big Data

Academic Integrity:
Students at the University of Washington (UW) are expected to maintain the highest standards of academic conduct, professional honesty, and personal integrity. The UW School of Public Health (SPH) is committed to upholding standards of academic integrity consistent with the academic and professional communities of which it is a part. Plagiarism, cheating, and other misconduct are serious violations of the University of Washington Student Conduct Code (WAC 478-120). We expect you to know and follow the university's

policies on cheating and plagiarism, and the SPH Academic Integrity Policy. Any suspected cases of academic misconduct will be handled according to University of Washington regulations. For more information, see the University of Washington Community Standards and Student Conduct website.

Access and Accommodation:
Your experience in this class is important to me. If you have already established accommodations with Disability Resources for Students (DRS), please communicate your approved accommodations to me at your earliest convenience so we can discuss your needs in this course. If you have not yet established services through DRS, but have a temporary health condition or permanent disability that requires accommodations (conditions include but not limited to; mental health, attention-related, learning, vision, hearing, physical or health impacts), you are welcome to contact DRS at 206-543-8924 or uwdrs@uw.edu or disability.uw.edu. DRS offers resources and coordinates reasonable accommodations for students with disabilities and/or temporary health conditions. Reasonable accommodations are established through an interactive process between you, your instructor(s) and DRS. It is the policy and practice of the University of Washington to create inclusive and accessible learning environments consistent with federal and state law.

Learning Environment:
To provide a supportive learning environment, I ask your commitment to showing respect to each other and to your instructors both inside and outside of class by avoiding behavior that might be offensive or distracting to others. Students with concerns about the instructor or teaching assistant (TA) should discuss these concerns with the course instructor and/or TA. If the student is not satisfied with the response, s/he may contact the Biostatistics Department Chair at heagerty@uw.edu. If concerns are not satisfactorily resolved, s/he may also contact the Graduate School at G1 Communications Building by phone at (206) 543-5139.

NOTE: The instructor maintains the right to modify/update the syllabus when necessary.