BIOST 546
MACHINE LEARNING FOR BIOMED RESEARCH AND PUBLIC HEALTH
WINTER QUARTER 2019

Instructor: Daniela Witten, PhD, Professor of Statistics & Biostatistics
Office: HSB F-649
Office Hours: See course webpage for up-to-date office hours schedule.
E-mail: dwitten@uw.edu

Course Meeting Times: MW 11:30-12:50 PM
Location: HSK K069
Website: Through `Canvas`

TAs: Arjun Sondhi and Tianyu Zhang
TA E-mail: asondhi@uw.edu and zty@uw.edu
TA Office Hours: See course webpage for up-to-date office hours schedule.

Course Description: Provides an introduction to statistical learning for biomedical and public health data. Intended for graduate students in SPH/SOM. Prerequisite: Basic understanding of statistics, probability, and calculus, and familiarity with `R`.

Evaluation and Grading:

- *Homeworks (100%):* Homework assignments will be due approximately every two weeks, and will contain a combination of analytical and `R`-based exercises. Each assignment will be worth the same amount.
  - You are permitted to brainstorm in groups while working on the assignments, but your final submitted homework assignment must be your own.
  - You may not copy or make use of solutions from the web, other students, or other sources.
  - You will be given two late days that you may use without penalty on your homework assignments over the course of the quarter. After you have used your two late days, each additional late day will result in 20% credit being docked from the assignment.
    * If your homework is late according to the `Canvas` submission timestamp (e.g. if the homework is due at 1:00 PM and the `Canvas` timestamp is 1:01 PM) then you will automatically be docked one late day. If you turn in your homework 24 hours and 1 minute late, then you will automatically be docked two late days, and so forth.
    * **No exceptions will be made to this late-day policy.** Since internet service can be unreliable and `Canvas` might go down at any time, you are encouraged to submit your assignments well in advance of the deadline.

Course Textbook: *Introduction to Statistical Learning, with Applications in R*, by James, Witten, Hastie, and Tibshirani.

- No need to buy it!! Free download at www.statlearning.com.

Course expectations: Though attendance is not required, it is strongly recommended. Students may brainstorm ideas for homework assignments, but may not copy solutions from other students or from other sources.

Computing: We will use the `R` programming language (`www.r-project.org`) throughout this course. Please set up your `R` computing environment right away, and well in advance of the deadline for the first homework assignment.

Communication: The course webpage (through `Canvas`) will serve as an archive of homeworks and other materials. Announcements concerning course logistics will also be placed on the webpage.

Discussion Board: We will be using a `Canvas` discussion board through the course website. Please use this discussion board to ask questions about homework or other course topics. Please do not e-mail the instructor or TAs questions about the homework assignments: such questions will be re-directed to the `Canvas` discussion board.

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 478-121). 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.

Class or TA Concerns: If you have any concerns about the class or your TA, please see the TA about these concerns as soon as possible. If you are not comfortable talking with the TA or not satisfied with the response that you receive, you may contact the Department of Biostatistics Associate Director of Academic Affairs (biostgp@uw.edu). If you are still not satisfied with the response that you receive, you may contact the Department of Biostatistics Chair (bchair@uw.edu). You may also contact the Graduate School at G-1 Communications Building, by phone at 206-543-5139 or by email at raan@uw.edu.

Access and Accommodations: 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.

Rough Sketch of Topics By Week . . . *This is subject to change!*

- *Week 1:* Overview of statistical learning: supervised versus unsupervised learning . . . *ISL* Ch 2.
- *Week 2:* Linear regression . . . *ISL* Ch 3.
- *Week 3:* Linear methods for classification: logistic regression, linear discriminant analysis . . . *ISL* Ch 4.
- *Week 4:* Resampling methods: cross-validation and the bootstrap . . . *ISL* Ch 5.
- *Week 5:* Model selection and regularization, Part I: subset selection, forward and backward stepwise selection . . . *ISL* Ch 6.
- *Week 6:* Model selection and regularization, Part II: ridge regression and the lasso . . . *ISL* Ch 6.
- *Week 7:* Moving Beyond Linearity: polynomial regression, splines, generalized additive models . . . *ISL* Ch 7.
- *Week 8:* Tree-Based Methods: classification and regression trees, bagging . . . *ISL* Ch 8.
- *Week 9:* Support Vector Machines . . . *ISL* Ch 9.
- *Week 10:* Dimension Reduction and Clustering: principal components analysis, k-means clustering, hierarchical clustering . . . *ISL* Ch 10.

LEARNING OBJECTIVES:

Upon completion of this course, a student should be able to:

- characterize the bias-variance trade-off mathematically, and explain it conceptually;
- explain the difference between a supervised and unsupervised learning problem, in terms of the problem formulation and the associated statistical challenges;
- understand the connections between machine learning approaches and classical statistical techniques;
- translate a scientific problem into a statistical model that can be fit using a machine learning method;
- discuss the pros and cons of using a "more complex" or "less complex" statistical model, in terms of the bias-variance trade-off, sample size, and other statistical considerations;
- perform cross-validation in order to estimate generalization error;
- describe the pros and cons of random forests, support vector machines, the lasso, ridge regression, splines, generalized additive models, and other regression and classification techniques; and
- apply the techniques covered in class in `R`.