BIOST 546
MACHINE LEARNING FOR BIOMEDICAL RESEARCH AND PUBLIC HEALTH
SPRING QUARTER 2018

Instructor: Daniela Witten, PhD, Associate Professor of Statistics & Biostatistics
Office: HSB F-649
Office Hours: T/Th 10-10:50 AM in HSB F649
E-mail: dwitten@uw.edu

Course Meeting Times: T and Th 8:30-9:50 AM
Location: HST T639
Website: Through Canvas

TA: Arjun Sondhi
TA E-mail: asondhi@uw.edu
TA Office Hours: Thursday 4/5 1:30-3:30 PM, and then starting on 4/11, Wednesdays 1:30-3:30 PM, all in HSB F643

Course Description: Provides an introduction to statistical learning for biomedical and public health data. Intended for graduate students in SPH/SOM. Prerequisite: BOST 511 or BOST 512 and familiarity with R. Offered: Sp.

Evaluation and Grading:

- **Homeworks (35%)**: 8 weekly homework assignments, due on Fridays, each worth an equal amount. Late homeworks will not be accepted and will receive zero points. Your lowest homework grade will be automatically dropped.

- **Midterm (20%)**: An in-class midterm exam on Tuesday, April 24, 8:30-9:50 AM in HST T639.

- **Final (45%)**: An in-class final exam is tentatively scheduled for Tuesday, June 5, 10:30 AM -12:20 PM.

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](http://www.statlearning.com)

Course expectations: Though attendance is not required, it is strongly recommended. Students may work together on homeworks, but may not copy solutions from other students or from other sources.

Computing: We will be use the R programming language ([www.r-project.org](http://www.r-project.org)) throughout this course.
Communication: The course webpage (through Canvas) will serve as an archive of homework, lecture notes, 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.
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.