STAT 435
INTRODUCTION TO STATISTICAL MACHINE LEARNING
SPRING QUARTER 2018

Instructor: Daniela Witten, PhD, Associate Professor of Statistics & Biostatistics
Office: B-323 Padelford
Office Hours: T 1:30-3:20 PM, in B-323 Padelford
E-mail: dwitten@uw.edu

Course Meeting Times: T and Th 11:30-12:50; W 1:30-2:20
Location: T/Th in SMI 304; W in MGH 234 or MGH 271, depending on what section you are in
Website: Through `Canvas`

TA: Hannah Director & Gang Cheng
TA E-mail: direch@uw.edu & gangc@uw.edu
TA Office Hours: Wed 3-5 PM in CMU B023 (Hannah) & TBD (will be updated on `Canvas`) (Gang)

Course Description: Introduces the theory and application of statistical machine learning. Topics may include supervised versus unsupervised learning; cross-validation; the bias-variance trade-off; regression and classification; regularization and shrinkage approaches; non-linear approaches; tree-based methods; and support vector machines. Includes applications in `R`. Prerequisite: either STAT 341, STAT 390/MATH 390, or STAT 391; recommended: MATH 308. Offered: Sp. 4 credits.

Prerequisites: Single-variable calculus and linear algebra (MATH 308 or equivalent). Previous coursework in statistics and probability (STAT 341 OR STAT 390 OR STAT 391). Previous programming experience in any programming language.

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, 11:30 AM -12:50 PM in SMI 304.
- *Final (45%):* An in-class final exam on Wednesday, June 6, 2018, 4:30-6:20 PM in SMI 304.

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 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`) 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`.