STAT/BIOST 527
NON-PARAMETRIC REGRESSION AND CLASSIFICATION
SPRING QUARTER 2021

Instructor: Daniela Witten, PhD, Professor of Statistics & Biostatistics
Office Hours: W 10-11AM, Zoom
E-mail: dwitten@u.washington.edu

TA: Anna Neufeld
Office Hours: Friday 11:30-1:30, Zoom
E-mail: aneufeld@u.washington.edu

Course Meeting Times: One lecture M 10-11:20 AM and posted on Canvas. Another asynchronous lecture to be posted on Canvas weekly, typically on Mondays
Location: Zoom
Website: Through Canvas

Course Description: This course will explore a number of flexible nonparametric models including splines, generalized additive models, regression trees, random forests, neural networks, and more. We will address both theoretical and practical aspects of these methods, and will implement and/or apply them in R.

Official Course Description: Covers techniques for smoothing, regression, and classification including spline models, tree-based methods, generalized additive models, and the averaging of multiple models. Describes measures of predictive performance, along with methods for balancing bias and variance.

Prerequisites: Either STAT 502 and STAT 504, or BIOST 514 and BIOST 515. Solid understanding of linear algebra and multivariable calculus. Experience with R.

Synchronous Versus Asynchronous Lecture Times: There will be a synchronous Zoom lecture on Mondays at 10-11:20 AM each week. I hope you will attend this lecture synchronously! However, in case you aren’t able to, it will be recorded and posted on Canvas afterwards. There will also be a 1 hour 20 minute asynchronous lecture each week, which will be posted directly on Canvas, typically on Mondays. The Wednesday 10AM lecture time that appears in the time schedule will be used for a professor office hour (Wednesday 10-11 AM).

Zoom Link: The Zoom link that will be used for Monday’s synchronous lecture and Wednesday’s office hour is [XXXX]. The Zoom link for Friday TA office hours is [XXXX].

Grading Policy: 100% Homeworks. There will be 5 homework assignments, assigned approximately every two weeks, and due electronically. If the Canvas time stamp for your submission indicates that the submission is late, then the submission
will be considered late. So, please plan ahead, so that your homework is not put
at risk by last minute internet/computer/etc issues. Each student is allowed TWO
LATE DAYS over the course of the quarter. Students may discuss homework in small
groups, but may not copy solutions from each other or from other sources. Solutions
and code must be prepared individually.

**Texts:** The following textbook is required:

- **Elements of Statistical Learning** (ESL), by Hastie, Tibshirani, and Friedman.
  - Most of the lectures will correspond to (parts of) chapters in ESL.
  - Many of the figures used in the lecture slides are borrowed from ESL.

You may also find the following textbook helpful:

- **Introduction to Statistical Learning, with Applications in R** by James, Witten, Hastie, and Tibshirani.
  - Contains R code for most of the topics covered in this class. Helpful for homeworks.
  - No need to buy it!! Free download at [www.statlearning.com](http://www.statlearning.com).
  - Some of the figures used in the lectures slides are borrowed from this book.

**Course expectations:** Students may work together on homeworks, but may not
copy solutions from other students or from other sources.

**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. Occasionally, time-sensitive announce-
ments will be e-mailed to the class through the class list-serve provided by the reg-
istrar’s office.

**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. There will be a designated discussion thread for each homework. Please
do not e-mail the professor or TA about homework or course material: instead, post
it on the discussion board.

**Acknowledgments:** The course materials rely very heavily on the textbooks listed
above. Many figures and examples are taken directly from those textbooks.

**Academic Integrity:** Students at the University of Washington (UW) are expected
to maintain the highest standards of academic conduct, professional honesty, and per-
sonal 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.

**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.

**Religious Accommodations:** Washington state law requires that UW develop a policy for accommodation of student absences or significant hardship due to reasons of faith or conscience, or for organized religious activities. The UW’s policy, including more information about how to request an accommodation, is available at Religious Accommodations Policy (https://registrar.washington.edu/staffandfaculty/religious-accommodations-policy/). Accommodations must be requested within the first two weeks of this course using the Religious Accommodations Request form (https://registrar.washington.edu/students/religious-accommodations-request/).

**Classroom Climate:** UW seeks to ensure that all students are fully included in each course. We strive to create an environment that reflects community and mutual caring. We encourage students with concerns about classroom climate to talk to your instructor, your advisor, a member of the departmental or SPH/A&S Diversity Committee and/or the program director. vg@uw.edu is a resource for Biostat students with classroom climate concerns.
Rough Sketch of Topics by Week: *This is subject to change!*

- **Week 1:** Welcome and overview of supervised learning . . . *ESL* Ch 2.
- **Week 2:** Linear methods for regression and classification . . . *ESL* Ch 3 and 4.
- **Week 3:** Basis expansions and regularization . . . *ESL* Ch 5.
- **Week 4:** Kernel smoothing methods . . . *ESL* Ch 6.
- **Week 5:** Model assessment and selection . . . *ESL* Ch 7.
- **Week 6:** CART, bagging, and random forests . . . *ESL* Ch 8.7 & 9.2 & 15
- **Week 7:** Boosting and ensemble learning . . . *ESL* Ch 10 & 16.
- **Week 8:** Support vector machines . . . *ESL* Ch 12.
- **Week 9:** Unsupervised learning *ESL* Ch 14 & 17.
- **Week 10:** Deep learning *ISL* 2nd edition preview chapter.
LEARNING OBJECTIVES:

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

- characterize the bias-variance trade-off mathematically, and explain it conceptually;
- apply cross-validation to perform model selection;
- describe the pros and cons of generalized additive models, random forests, support vector machines, the lasso, ridge regression, kernels, splines, neural networks, and other non-parametric regression and classification techniques; and
- apply the techniques covered in class in R.