STA 4241

Fall 2025

Statistical Learning in R

Instructor: Karl Oskar Ekvall
E-mail: k.ekvall@ufl.edu
Phone: (352) 273-3001
Office hours: 10:30 AM – 11:30 AM on Mondays and Wednesdays in Griffin-Floyd Hall (FLO) 205.

Teaching Assistant: Huang, Zhuochao
E-mail: zhuochao.huang@ufl.edu
Office hours: 8:00 AM – 10:00 AM on Wednesdays on Zoom (https://ufl.zoom.us/j/3383284442)

Course Website: e-Learning

Course Prerequisites: STA 4322 & STA 4210 & MAS 4115 (Or linear algebra equivalent)

Course lectures: Mo-We-Fr, 11:45 AM – 12:35 PM, in FLO 0100.

Course Notes/Material: Notes will be on the course website. These should contain nearly all of the material that we cover in class, however, I will present some additional material in the lectures that is not posted on the course website.

Software: We will be using the R software language throughout. R is free and should be easy to download on your personal computer. I highly recommend running R through RStudio. If you have any problems downloading R or RStudio, feel free to talk to myself or the TA. If you do not have access to a computer, please reach out to me via email.

Required Text: James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) An Introduction to Statistical learning with Applications in R. Springer.

ISBN-13: 978-1461471370

Course Description: Overview of the field of statistical learning. Topics include linear regression, classification, resampling methods, shrinkage methods, nonlinear regression, tree-based methods, support vector machines, and neural networks. We

will cover many aspects of these approaches including conceptual, theoretical, and applied aspects Methods will be illustrated and implemented in R.

Course Objectives: The goal of this course is to teach the theoretical underpinnings of several advanced and commonly used statistical learning techniques. We will review classical statistical techniques such as linear and logistic regression before covering more advanced statistical techniques such as classification, regularization, nonlinear regression, and other machine learning approaches. The implementation of all approaches in the R statistical software will be taught throughout. By the end of the course, students should be familiar with a wide range of statistical methodologies that are widely used in practice, and should be able to apply these approaches to data sets.

Homework

There will be a homework assignment roughly every two weeks and it will be due via Canvas submission.

Exams

You will have two in-class exams in the middle of the semester.

Project: Students will be expected to complete a written project at the end of the semester and present their findings to the class. The grade for the final project will consist of two main components: a written report and an oral presentation. The written report will state the objectives of the study, describe data collection, describe the statistical model used, explain any assumptions required by the analysis, and provide conclusions for the main study questions. The oral presentation will be a 10–15-minute presentation that should cover the key components of the written report. Presentations should clearly state the objectives of the project, while using visualizations to illustrate the main results and findings of the project.

Grade Distribution

Homework	20%
Midterms	50%
Project	30%

Letter Grade Assignment: Grades will be assigned as follows: 90-100, A; 87-89.9, A-; 84-86.9, B+; 80-83.9, B; 77- 79.9, B-; 74-76.9, C+; 70-73.9, C; 67-69.9, C-; 64-66.9, D+; 60-63.9, D; 55-59.9, D-; 0- 55, F

The numeric scores will be rounded to the nearest tenth.

This course complies with all UF academic policies: For information on those polices and for resources for students, please see this link.

Approximate weekly breakdown of the material:

Week 1
- Introduction to Statistical Learning

Week 2
- Review of linear regression

Week 3
- Logistic regression
- Linear discriminant analysis

Week 4
- Quadratic discriminant analysis
- Maximal margin classifier

Week 5
- Support vector classifier
- Support vector machines

Week 6
- Cross validation
- The bootstrap

Week 7
- Exam 1 (Date: 9/26)
- Linear model selection and regularization
- Subset selection

Week 8:
- Shrinkage methods
- Ridge regression
- The lasso

Week 9:
- Dimension reduction methods
- Principle components analysis, Principal components regression
- Partial least squares

Week 10
- Nonlinear models
- Polynomial regression - regression splines

Week 11
- Smoothing splines
- Local regression
- Generalized additive models

Week 12
- Exam 2 (Date: 11/3)
- Tree-based methods
- Decision trees

Week 13
- Bagging
- Random forests
- Boosting

Week 14
- Ensemble approaches
- Neural networks

Week 15
- Project presentations

Materials and Supplies Fee
N/A