E-mail: yuansi.chen@duke.edu
Web: www2.stat.duke.edu/courses/Fall21/sta521.001/
Class Hours: Tuesday & Thursday 3:30-4:45pm
Class Room: 116 Old Chemistry
Office Hours: after class
Office: 223B Old Chemistry

TAs: Jordan Bryan
      jordan.bryan@duke.edu,
    Michele Caprio
      michele.caprio@duke.edu,
  Michael Christensen
      michael.f.christensen@duke.edu,
    Pritam Dey
      pritam.dey@duke.edu
    Zeki Kazan
      zekican.kazan@duke.edu

Lab Room 1: Social Sciences 311
Lab Hours: Friday 12:00 - 1:15pm
Lab Room 2: Biological Sciences 155
Lab Hours: Friday 1:45 - 3:00pm
TA Office Hours: TBA

Course Description

STA 521: Predictive Modeling and Statistical Learning is a master-level introductory course to statistical learning methods for prediction and inference. This course introduces students to concepts and techniques of modern regression and predictive modelling. The course will blend theory and application using a range of examples. Topics include exploratory data analysis and visualization, linear and generalized linear models, model selection, penalized estimation and shrinkage methods including Lasso, ridge regression and Bayesian regression, decision trees and ensemble methods. Other advanced topics, such as robust estimation, smoothing splines, support vector machines and neural networks, will be briefly discussed. The R programming language and applications are used throughout.

Course objectives

The main objectives of this course are as follows

- Understand the different approaches to formulate statistical prediction and inference problems
• Build a solid foundation for the statistical theory for predictive modelling and inference
• Understand the pros and cons of statistical models for predictive modeling
• Learn to perform data analysis (including exploratory data analysis and visualization) with R and to explain the relevance of the statistical methods chosen
• Learn to communicate statistical results without use of statistical jargon

Books & Materials

Required:

• An Introduction to Statistical Learning: with Applications in R (Second Edition) by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. ISL is our main book reference. We intend to cover most content in this course. The book is freely available on the authors’ website.

• Elements of Statistical Learning (12th printing) by Trevor Hastie, Robert Tibshirani and Jerome Friedman. ESL is more advanced than ISL. It is also freely available on the authors’ website.

Optional:

• Applied linear regression by Sanford Weisberg. In depth coverage of linear regression adn extensions, model checking, and more. The associated Computing Primer for Applied Linear Regression Using R is useful for the companion R package

• A First Course in Bayesian Statistical Methods by Peter Hoff.

Prerequisites/Corequisites

Statistical Science 323D or 523L and Statistical Science 360, 601, or 602L. All students should be comfortable with linear/matrix algebra and mathematical statistics at the level of STA 611 and be familiar with the R programming language and linear regression. Students should be familiar with Bayesian statistics either by taking the introduction to Bayesian inference STA 360/601/602 or currently co-registered in the course. Please see me if you have questions about the pre-requisites/background.

Course topics

• Formulation of a statistical prediction problem
• Exploratory data analysis
• Linear regression
• Shrinkage methods and regularization
• Model assessment and bias-variance trade-off
• Logistic regression and classification
• Support vector machines and the max-margin idea
• Kernel methods
• Classification and regression trees (and forests)
• Ensemble methods
• Neural networks
• Interpretable machine learning

Homework

The objective of the problem sets is to help you develop a more in-depth understanding of the material and help you prepare for exams and projects. Grading will be based on completeness as well as accuracy. The homework assignments are released by weekly, and they to be completely individually.

Submission instructions: You will submit your pdf and code to the Gradescope integrated within Sakai, following the instructions provided in the assignments.

Homework grading policy:

• No late homework will be accepted. You can submit multiple times on Gradescope and only the last submission will be graded. Please submit early.

• The lowest homework score will be dropped.

• Regrade requests must be made within 3 days of when the assignment is returned.

Projects

There will be two data analysis projects. The objective of the Project is to give you real-world data analysis experience using R. You will use all (relevant) techniques learned in this class or explore additional advanced material to solve a problem, explore its properties (either analytically or through simulation) and present it using reproducible methods. Further details will be provided as due dates approach.
Exams

There will be one midterm and one final in this class. See the course website for dates, times and locations of the exams.

You are allowed to use one sheet of notes ("cheat sheet) for each exam. This sheet must be no larger than $8\frac{1}{2} \times 11$, and must be prepared by you. You may use both sides of the sheet and can write as small as you wish.

Exam grading policy:

- There will be no makeup exams. Missing the exam will result in a grade of 0.
- Regrade requests must be made within 5 days of when the graded exam is returned.

Overall Grading Policy

The grade will count the assessments using the following proportions:

- Homework 25%
- Midterm 20%
- Final 20%
- Projects 30%
- Participation 5%

Participation includes lecture attendance, lab attendance, discussion participation on Sakai.

Grades may be curved at the end of the semester. Cumulative numerical averages of 90 - 100 are guaranteed at least an A-, 80 - 89 at least a B-, and 70 - 79 at least a C-, however the exact ranges for letter grades will be determined after the final exam. The more evidence there is that the class has mastered the material, the more generous the curve will be.
Other Course Policies

During Class

I understand that the electronic recording of notes will be important for class and so computers will be allowed in class. Please refrain from using computers for anything but activities related to the class. Phones are prohibited as they are rarely useful for anything in the course. Eating and drinking are allowed in class but please refrain from it affecting the course. Try not to eat your lunch in class as the classes are typically active.

Academic Integrity and Honesty

Duke University is a community dedicated to scholarship, leadership, and service and to the principles of honesty, fairness, respect, and accountability. Citizens of this community commit to reflect upon and uphold these principles in all academic and non-academic endeavors, and to protect and promote a culture of integrity. Cheating on exams and quizzes, plagiarism on homework assignments and projects, lying about an illness or absence and other forms of academic dishonesty are a breach of trust with classmates and faculty, violate the Duke Community Standard, and will not be tolerated. Such incidences will result in a 0 grade for all parties involved as well as being reported to the Office of Student Conduct. Additionally, there may be penalties to your final class grade. Please review the Dukes Academic Dishonesty policies.

Accommodations for Disabilities

Students with disabilities who believe they may need accommodations in this class are encouraged to contact the Student Disability Access Office at (919) 668-1267 as soon as possible to better ensure that such accommodations can be made.