STA 325: Machine Learning and Data Mining
Duke University, Fall 2023

Instructor: Simon Mak, Assistant Professor of Statistical Science, e-mail: sm769@duke.edu
Lecture: Tuesdays / Thursdays 3:05 – 4:20pm (Old Chem 116)
Office Hours: Tuesdays / Thursdays 4:30 – 5:30pm (Old Chem 112A)

Head TA: Yi (Irene) Ji, e-mail: yj136@duke.edu
Office Hours: Wednesdays / Fridays 10 – 11am (https://duke.zoom.us/j/93517035541)

TA 1: Yinting Zhong, e-mail: yz790@duke.edu
Lab: Wednesdays 3:05 – 4:20pm (Gross Hall 104)
Office Hours: Mondays 10 – 11am (in-person, Old Chem 203B), Fridays 4:30 – 5:30pm (https://duke.zoom.us/j/93517035541)

TA 2: Olivia Fan, e-mail: zf59@duke.edu
Lab: Wednesdays 4:40 – 5:55pm (Social Sciences 105)
Office Hours: Tuesdays 11:30am – 12:30pm, Thursdays 11:45am – 12:45pm (https://duke.zoom.us/j/93517035541)

Course Outline: The increasing availability of data and computing power has created immense opportunities for machine learning. This course introduces a comprehensive toolbox of machine learning and data mining methods, including regularized regression, nonlinear regression, cross-validation, random forests, support vector machines, and clustering. Course evaluation will be based on not only knowledge of these methods, but also its selection and justification in applications, its implementation in R, and effective communication of data analysis for solving real-world problems.

Labs: Labs will be led by Yinting and Olivia, and will cover supplemental problems and coding exercises for course material, as well as review sessions for quizzes. Lab attendence is *not* mandatory, but you are highly encouraged to actively participate. You are required to submit a short completion assignment after each lab, which will count for participation marks.

Course Resources:

- *An Introduction to Statistical Learning, with Applications in R*. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2013). https://static1.squarespace.com/static/5ff2adbe3fe4fe33db902812/t/6009dd9fa7bc363aa822d2c7/1611259312432/ISLR+Seventh+Printing.pdf.

  This is the required course textbook. We will call this text "ISL" throughout the course.

- *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Trevor Hastie, Robert Tibshirani, Jerome Friedman (2009). http://statweb.stanford.edu/~tibs/ElemStatLearn/.

  A good supplemental textbook for students seeking a more advanced perspective on course topics.

- *The R Cookbook*, http://www.cookbook-r.com/.

An excellent resource on how to use R for coding, data analysis, and data visualization.

Prior Knowledge: Students are expected to have a solid background in regression analysis (STA 210) and elementary probability (STA 230). These will be building blocks for course topics, and little review will be provided during class time. If you are unsure what prior knowledge is expected, please refer to the following past syllabi:

- *STA 210*: https://www2.stat.duke.edu/courses/Spring19/sta210.001/
- *STA 230*: https://www2.stat.duke.edu/courses/Fall18/sta230/

Course Grading: This course will be graded on the following grade breakdown:

Participation	5%
Homework	25%
Quiz 1	10%
Quiz 2	10%
Quiz 3	10%
Case Study	15%
Project (Proposal + Presentation + Report)	25% (5% + 10% + 10%)

Grades may be curved at the end of the semester. Cumulative 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 quiz. The more evidence there is that the class has mastered the material, the more generous the curve will be.

Course Logistics:

- *Participation*: Participation marks will be given for *active* participation (e.g., asking and answering questions, seeking feedback) in lectures, labs, office hours and Piazza (https://piazza.com/duke/fall2023/sta325). If you cannot attend lectures or labs, make sure to actively participate in office hours and Piazza. These are *easy* marks, so don't lose out!

- *Homework*: You should expect around 5 homework assignments. Assignments will be posted on Sakai, and will be due in roughly two weeks time. Assignments will include both conceptual and data analysis questions. For conceptual questions, all answers should either be written *legibly* and scanned as a pdf, or typed out. For data analysis, all code should be done in R and should be well-documented. All files (pdf and code) should be submitted as a single .zip file to Sakai by the due date.

- *Quizzes*: There will be a total of three quizzes. For preparation, you will be provided a test bank of problems a few weeks prior to each quiz. It is in your interest to prepare *well* before the quiz by studying lecture slides & notes, reading the textbook, and practicing on problems – cramming will be difficult! The lowest quiz score of the three quizzes will be waived, and the remaining quizzes will be worth 15%.

- *Case Study*: There will also be a case study on applying predictive modeling methods to a real-world scientific problem on turbulence prediction, which has important applications in astrophysics, climatology, and aerospace engineering. You will be required to attend (or watch

online) a workshop hosted by the Pratt School of Engineering, who will introduce the scientific problem at hand. You will then work in groups to develop predictive models for addressing this problem, and will submit a short report and presentation justifying your model and interpreting scientific results.

- *Project*: You will work in groups of 3-5 on a comprehensive data analysis project. Project topics are entirely up to you, as long as it tackles an important real-world problem, and meaningful conclusions & decisions can be made from data analysis. The project will be evaluated in three parts: a proposal (before Quiz 2), a 20-minute oral presentation (at end of semester), and a final report (at end of semester). Further details will be provided as the semester progresses.
- *Regrades*: Regrade requests must be made within *one week* of when the assignment or exam is returned, and must be e-mailed to myself and the TAs. There will be no grade changes after the final exam.

Course Schedule: The course will tentatively follow the schedule below:

Week	*Topics*	*Readings*
1	Introduction, Statistical Learning	ISL 1-2
2	Statistical Learning, Linear Regression	ISL 2, ISL 3
3	Classification, Model Selection	ISL 4, ISL 6.1
4	Cross-Validation, Model Selection	ISL 5.1, ISL 6.2
5	Model Selection, Review	ISL 6.2
6	Quiz 1, Nonlinear Regression	ISL 7.1-7.4
7	FALL BREAK, Nonlinear Regression	ISL 7.1-7.4
8	Nonlinear Regression	ISL 7.5-7.7
9	Trees	ISL 8.1-8.2
10	Trees, Support Vector Machines	ISL 8.2, 9.1-9.2
11	Support Vector Machines	ISL 9.3, 9.4-9.5
12	Review, Quiz 2	N/A
13	Multiclass Regression, THANKSGIVING	N/A
14	Multiclass Regression	Notes
15	Gaussian Processes	None

Missing class / homework / exams: Students who miss graded work due to a scheduled varsity trip, religious holiday or short-term illness should fill out an online NOVAP, religious observance notification or short-term illness notification form, respectively. If you are faced with a personal or family emergency or a long-range or chronic health condition that interferes with your ability to attend or complete classes, you should contact your academic dean's office. See more information on policies surrounding these conditions at https://trinity.duke.edu/undergraduate/acadeamic-policies/personal-emergencies, and your academic dean can provide more information as well.

Academic 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, plagiarism on homework assignments, projects, and code, 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 University Judicial Board. Additionally, there may be penalties to your final class grade. Please review Duke's Standards of Conduct. For more information on the Duke honor code (known as Duke Community Standard), please go to http://integrity.duke.edu/faq/faq1.html. All work turned in for grading must be entirely your own. This particularly relates to homework. You are encouraged to talk to each other regarding homework problems; however, the write up, solution, and code *must* be entirely your own work.

Students with Disabilities: Students who require special accommodations in class or during exams should follow the procedures outlined by the Disability Management Program http://access.duke.edu/students. 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. If you have a special accommodation outlined by the Disability Management Program, I encourage you to please meet with me during the first week of class. Please email me to set up an appointment or speak with me in person.

Academic Resource Center: The Academic Resource Center (ARC) offers free services to all students during their undergraduate careers at Duke. Services include Learning Consultations, Peer Tutoring and Study Groups, ADHD/LD Coaching, Outreach Workshops, and more. Because learning is a process unique to every individual, we work with each student to discover and develop their own academic strategy for success at Duke. Contact the ARC to schedule an appointment. Undergraduates in any year, studying any discipline can benefit!
Location: 211 Academic Advising Center Building, East Campus – behind Marketplace.
Contact: Website: arc.duke.edu, e-mail: theARC@duke.edu, phone: (919) 684-5917.