Statistical Machine Learning
STAT 427/627

Fall 2021

Instructor:	Zois Boukouvalas	**Time:**	W 5:30 PM - 8:00 PM
Email:	boukouva@american.edu	**Room:**	Don Myers Building - 116
Office:	Don Myers Building - 222		

Office hours: After class, or W 4:00pm - 5:30pm via Zoom, or by appointment. Always feel welcome to stop by my office hours. You are also encouraged to ask me questions online via email. If you are having **ANY** trouble with the class, please come see me about it as soon as possible. **Do not wait until it is too late!**

Course Pages: I will use Canvas (https://american.instructure.com/) to post any supplementary materials, suggested readings/practice exercises, assignments, and announcements. Sometimes I may also use my personal website (https://zoisboukouvalas.github.io/).

Materials:

- **Book** (Required): An Introduction to Statistical Learning with Applications in R, by G. James, D. Witten, T. Hastie, and R. Tibshirani; Springer, 2013, (1st edition)
The second edition of the book is also available (https://web.stanford.edu/~hastie/ISLRv2_website.pdf)
- **Extra Links**: (https://cran.r-project.org/web/packages/ISLR/) (R data sets) (http://www-bcf.usc.edu/~gareth/ISL/), (http://statweb.stanford.edu/~tibs/ElemStatLearn/) (R codes, errata, etc.)
- **Software**: During the course, we'll study statistical machine learning methods and implement them in R, including classroom demonstrations and examples. For all computer assignments, use the language of your choice. Advanced programming skills and advanced computer knowledge are *not required*.
- **Other useful Books**:

1. Kevin Murphy, Machine Learning: A Probabilistic Perspective, MIT Press
2. Christopher Bishop, Pattern Recognition and Machine Learning, Springer
3. Tom Mitchell, Machine Learning, McGraw Hill
4. Sergios Theodoridis, Machine Learning: A Bayesian and Optimization Perspective, Elsevier.

Pre-requisites: STAT 520 "Applied Multivariate Analysis" or STAT 615 "Regression".

Course Plan:

- Introduction, motivation, and examples. Understanding large and complex data sets. Statistical learning. First steps in R.
- Review of regression modeling and analysis; implementation in R.
- Classification problems and classification tools. Logistic regression and review of linear discriminant analysis.
- Resampling methods; bootstrap.
- High-dimensional data and shrinkage. Ridge regression. LASSO. Model selection methods and dimension reduction.
- Regression trees and decision trees.
- Introduction to support vector machines.
- Deep Learning.
- Clustering methods.

Class structure: This class will be a blend of lecture, class discussion, and labs. I want you all to be involved during class and please do not hesitate to ask questions whenever something is unclear to you. You

are expected to attend all class meetings, as I believe that attending class regularly contributes greatly to your performance in the course. It is understandable that you may have to miss class on a rare occasion. You are responsible for any assignments or papers given out during any missed class.
Data scientists must learn to discover solutions for themselves. You should expect to have to research (use Google, stackoverflow, etc) to do your assignments. All you need to do the assignment will NOT have been provided to you in the lectures and course book. This is an essential part of becoming a data scientist!

Assignments & Grading:

Assignments (20%): During the semester I will assign, collect, and grade assignments. You may receive assistance from other students in the class and me, but your submissions must be composed of your own thoughts, coding and words. A typical homework will include a few problems to do by hand, to see how things work, and a few realistic problems to do using R software. **Late submission is accepted at a cost of a 10% deduction for each day**.

Labs (15%): 40-minute labs at the end of each class. Each lab covers the material of the lecture. You will have to submit the solutions of each lab on Canvas the Sunday after each class.

Quizzes (10%): We will have six take home quizzes at the end of our class. Quizzes will cover the material of the preceding week.

Midterm (15%): The midterm covers several chapters of the material and it is take home. No make-up exams will be given unless you have an extremely compelling excuse such as observance of a religious holiday (in which case you need to let me know in advance) or a medical emergency. You will have three days to complete the exam.

Project (20%) (15% Presentation + 5% Project Proposal): You will have to prepare a project using the tools and methods learned in the class. You are expected to submit a mid-semester research proposal in order to get your topic approved. The class projects will be presented as a poster presentation. You should prepare a poster, and be prepared to give a very short explanation (10 minutes), in front of the poster, about your work. At the poster session (online), you'll also have an opportunity to see what everyone else did for their projects. You will also need to submit your poster as a PDF the day before the presentation.

Final Exam (20%): The final exam is cumulative. However, it will mostly cover the last part of the course. You will have three days to complete each exam.

Please schedule a meeting with me if you would like to see or discuss your grade at any point during the semester.

Important dates:

Midterm	October 13, 2021
Project Proposal	October 20, 2021
Project Presentations	December 8, 2021
Final Exam	TBA

Learning Objectives: At the end of this course, you are expected to be able to:

- Identify appropriate statistical learning methods for the given problem involving real data.

- Understand the underlying assumptions, techniques available to verify them, and propose appropriate remedies.

- Use training and testing data to evaluate performance of the chosen regression and classification techniques and compare them.

- Use available empirical tools to find the optimal balance between precision within training data and prediction power.

- Apply cross-validation techniques to find the optimal degree of flexibility - the best subset of predictors or the optimal tuning parameters.

- Show, analytically or empirically, the optimal balance between precision within training data and prediction power.

- Illustrate results with appropriate plots and diagrams.

Emergency preparedness: In the event of an emergency, students should refer to the AU Web site http://www.american.edu/emergency and the AU information line at (202) 885-1100 for general university-wide information. In case of a prolonged closure of the University, I send updates to you by email and will post all announcements on Canvas.

Support services: A wide range of services is available to support you in your efforts to meet the course requirements.

1. Mathematics & Statistics Tutoring Lab (Don Myers Building) provides tutoring in Intermediate Mathematics and Statistics. http://www.american.edu/cas/mathstat/tutoring.cfm

2. Academic Support and Access Center offers study skills workshops, individual instruction, tutor referrals, Supplemental Instruction, writing support, and technical and practical support and assistance with accommodations for students with physical, medical, or psychological disabilities. Writing support is also available in the Writing Center, Battelle-Tompkins 228.

3. Center for Diversity & Inclusion (X3651, MGC 201) is dedicated to enhancing LGBTQ, Multicultural, First Generation, and Women's experiences on campus and to advance AU's commitment to respecting & valuing diversity by serving as a resource and liaison to students, staff, and faculty on issues of equity through education, outreach, and advocacy.

4. The Office of Advocacy Services for Interpersonal and Sexual Violence (X7070) provides free and confidential advocacy services for anyone in the campus community who is impacted by sexual violence (sexual assault, dating or domestic violence, and stalking).

Additional notes:

1. I expect you to be courteous to me and your fellow classmates during the online lectures/meetings.

2. Please let me know during the first week of classes if you have any special needs that require accommodations.

3. Please be sure that you are familiar with AU's Academic Integrity Code, as I am required to report any cases of academic dishonesty to the dean of CAS. For your review: http://www.american.edu/academics/integrity/.