Stat. & C.S. 295: Introduction to Statistical Learning

Fall 2016 – University of Vermont

Instructor:	Samuel V. Scarpino	**Time:**	T/R 13:15 – 14:30
Email:	svscarpi@uvm.edu	**Location:**	Lafayette L307
Office:	Farrell Hall 206	**Office Hours:**	W 10:00 – 12:00

Course Pages: http://scarpino.github.io/teaching/ and Blackboard.

Teaching assistant: Currently, there is no T.A. assigned to this course. If that changes, I will provide their contact information and office hours.

Main Reference:

- James, G., Witten, D., Hastie, T. and Tibshirani, R. 2013. *An introduction to statistical learning.* New York: Springer.[1]

Objectives: In this class, we will explore and discuss statistical learning methods and their application to modern problems in science, industry, and society. We will cover material presented in "An Introduction to Statistical Learning with Applications in R" and examine recent applications of these methods in both scholarly articles and popular applications. Topics will include: linear model selection, lasso and ridge regression, tree-based methods, support vector machines, and unsupervised learning. There will be no traditional in-class exams; instead, the major assignment–a research paper–will ask you to demonstrate your mastery of the methods we will cover in class and their application to real-world problems.

Assignments & Grading Breakdown:
Research paper (50%) – The goal of the research paper is to demonstrate your mastery of the skills we have learned in class and your ability to communicate your work through figures and text. Each student must work on a separate paper, but you are allowed–and encouraged!–to discuss your work and receive feedback from other students. Your grade for this paper will be based on: an initial proposal (20%), a first draft (30%), and the final version (50%). I will provide detailed instructions and a rubric. My hope is that many of your papers will be submitted to a scholarly journal.

Homework (30%) – There will be three homework assignments, which are designed to evaluate your understanding of the skills we will learn in class. You may work together with other students in the course on the homeworks. All homeworks will be due by 5pm on Thursday of the "due by" week.

Final presentation (20%) – During the last week of class you will each give a ten minute presentation on your research paper.

Graduate Students – Graduate students, and any undergraduates who are interested, will have to peer-review two research papers, i.e. the ones being written by your peers in this class. The quality of your reviews will count as 10% of your final paper grade, which will reduce the final version to 40% of the paper grade. I will provide a detailed set of expectations prior to the assignment.

Prerequisites: An advanced undergraduate-level understanding of probability and statistics is expected. The course will be taught in R. Prior experience programming in R and/or the willingness to learn R quickly is required.

[1] Available as a PDF – http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Sixth%20Printing.pdf

Important Dates:

HW1 due	Sept. 15th by 17:00
HW2 due	Sept. 29th by 17:00
HW3 due	Oct. 13th by 17:00
Paper proposal due	Oct. 20th by 17:00
Paper draft due	Nov. 3rd by 17:00
Research paper due	Dec. 1st by 17:00
Final presentations	Dec. 6th & 8th

Course Schedule:

Date(s)	Material	To Do
Aug. 30th & Sept. 1st	Linear regression, classification, & re-sampling	Install R and Rstudio[2]
Sept. 6th & 8th	Regularization & linear model selection	Read AISL[3] Chpt. 6.
Sept. 13th & 15th	Tree-based regression	Read AISL Chpt. 8 & HW1 due
Sept. 20th & 22nd	Regression in practice	Read paper 1 (Tues) & 2 (Thurs)
Sept. 27th & 29th	Support vector machines	Read AISL Chpt. 9 & HW2 due
Oct. 4th & 6th	Unsupervised learning	Read AISL Chpt. 10
Oct. 11th & 13th	Beyond power calculations	Read paper 3 & HW3 due
Oct. 18th & 20th	Regularization	Read paper 4 & Proposal due
Oct. 25th & 27th	The rhetoric of data I - Writing research papers	Read papers 5 & 6
Nov. 1st & 3rd	The rhetoric of data II - Visualization	Read paper 7 & Paper draft due
Nov. 8th & 10th	Understanding bias & Lecture on peer review	Read paper 8
Nov. 15th & 17th	**No class**	Research paper consultations
Nov. 22nd & 24th	Thanksgiving	Enjoy your time off
Nov. 29th & Dec. 1st	The future of data science (presentation primer)	Read paper 9 & Paper due
Dec. 6th & 8th	Final presentations	Present & attend class
Scheduled final period	**No class**	N/A

[2] R: https://cran.r-project.org/ RStudio: https://www.rstudio.com/

[3] "An Introduction to Statistical Learning" – Available as a PDF – http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Sixth%20Printing.pdf

Readings:

1. Naydenova, E., Tsanas, A., Howie, S., Casals-Pascual, C. and De Vos, M., 2016. The power of data mining in diagnosis of childhood pneumonia. Journal of The Royal Society Interface, 13(120).
2. Santillana, M., Nguyen, A.T., Dredze, M., et al., 2015. Combining search, social media, and traditional data sources to improve influenza surveillance. PLoS Comput Biol, 11(10).
3. Gelman, A. and Carlin, J., 2014. Beyond power calculations assessing type s (sign) and type m (magnitude) errors. Perspectives on Psychological Science, 9(6), pp.641-651.
4. Gelman, A. and Shalizi, C.R., 2013. Philosophy and the practice of Bayesian statistics. British Journal of Mathematical and Statistical Psychology, 66(1), pp.8-38.
5. Kallestinova, E.D., 2011. How to write your first research paper. Yale J Biol Med, 84(3), pp.181-90.
6. Wason, P.C., 1970. On writing scientific papers. Physics Bulletin, 21(9), p.407.
7. Cleveland, W.S. and McGill, R., 1984. Graphical perception: Theory, experimentation, and application to the development of graphical methods. Journal of the American statistical association, 79(387), pp.531-554.
8. Torralba, A. and Efros, A.A., 2011, June. Unbiased look at dataset bias. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on (pp. 1521-1528). IEEE.
9. Donoho, D., 2015. 50 years of Data Science. Tukey Centennial workshop, Princeton NJ Sept 18 2015.

Academic assistance: Anyone needing accommodation, e.g., per the ACCESS program, please contact me as soon as possible.

Religious holidays: As per University policy, you have the right to practice the religion of your choice and can make-up missed work due to your religious holidays. For those requesting an accommodation due to a religious holiday, please submit a schedule of your holidays to me by the end of the second full week.

Course policies:
I. Grades – 100–98% (A+), 97–93% (A), 92–90% (A-), 89–87% (B+), 86–83% (B), 82–80% (B-), 79–77% (C+), 76–73% (C), 72–70% (C-), 69–60% (D), <60% (F).

II. Technology – Please silence and put away all electronics before coming to class–there should be zero texting in class. Computers should be used only for course-related work and only when someone isn't addressing the class. Violation of these policies will negatively affect your grade (and your understanding of course material).

III. Turning in assignments – All assignments must be turned in on Blackboard.

IV. Late assignments – Late or missed assignments will be given a score of 0%. Please contact me if you have a documented emergency.

V. Email – I am happy to answer questions via email, but cannot promise to respond same-day. Please remember that email is a professional, mostly-permanent record, so please communicate in a respectful manner.

VI. Academic honesty – As in all UVM classes, academic honesty will be expected and departures will be dealt with appropriately. Lack of knowledge of the academic honesty policy is not a reasonable explanation for a violation, see http://www.uvm.edu/cses/ for guidelines.