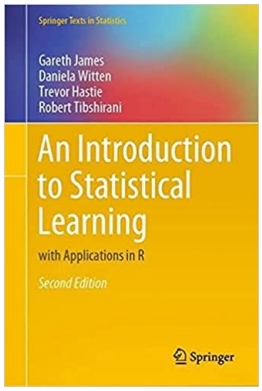
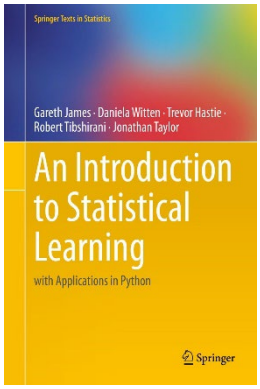
STAT 627/427 **Fall 2024**

STATISTICAL MACHINE LEARNING

Lectures	Tuesday 5:30 – 8:00 pm in DMTI-217 (section 001) Wednesday 5:30 – 8:00 pm in DMTI-114 (section 002)	**Instructor**	Michael Baron
		Office	DMTI 106-D
Course web site	http://fs2.american.edu/baron/www/627/	Phone	202-885-3130
Office hours	Tuesday and Wednesday 4:00-5:15 pm in DMTI 106D	Email	baron@american.edu
Assistant	Roland Abi	TA Email	ra0033a@american.edu

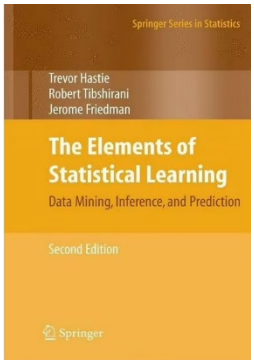
Textbooks





- [Our main text, with R] *An Introduction to Statistical Learning with Applications in R,* by G. James, D. Witten, T. Hastie, and R. Tibshirani. 2nd edition, 2021. ISBN 1071614177. The book is available for a free download at https://www.statlearning.com/

- [Same, with Python] *An Introduction to Statistical Learning with Applications in Python,* by G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, 2023. ISBN 3031387465. This book is also available for a free download at https://www.statlearning.com/



- [Supplementary; more technical; contains advanced explanations and mathematical proofs] *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, by T. Hastie, R. Tibshirani, and J. Friedman, 2nd Edition; Springer, 2009. ISBN 0387848576. Available on Hastie's page at https://web.stanford.edu/~hastie/Papers/ESLII.pdf

Materials:
https://www.statlearning.com/ - the main author's page for our textbook
https://cran.r-project.org/web/packages/ISLR/ - ISLR package for R
https://pypi.org/project/ISLP/ - ISLP package for Python
https://www.statlearning.com/resources-second-edition - R codes, data sets, slides (which I don't use)
https://www.statlearning.com/resources-python - ISLP instructions and manual

Course plan:

1. Introduction, motivation, and examples. Main principles of statistical machine learning. Regression and classification, bias and variance, training and testing, prediction and inference. [Chap. 1-2].
2. Review of regression modeling and analysis; implementation in R and Python. [Chap. 3].
3. Classification problems and classification tools. Logistic regression. Linear and quadratic discriminant analysis. K-nearest neighbor classification. Thresholds and tuning. [Chap. 4]
4. Resampling and cross-validation methods; jackknife, bootstrap. [Chap. 5 and lecture notes].
5. High-dimensional data and shrinkage. Ridge regression. LASSO. Model selection methods and dimension reduction. Principal components. Partial least squares. [Chap. 6; also 12.2]
6. Nonlinear trends and splines. [Chap. 7; 7.4-7.5]
7. Regression trees and decision trees. Bagging. Random forests, BART, and other tree ensembles. [Chap. 8]
8. Support vector machines [Chap. 9]
9. Deep learning, and introduction to neural networks [Chap. 10]
10. Clustering methods [Chap. 12]

Software: During the course, we'll study statistical machine learning methods and implement them in R and Python, including classroom frequent demonstrations and examples. Mostly R will be used in class; materials will be developed and published for both R and Python on the course web site. For all computer assignments, use the language of your choice. Advanced programming skills and advanced computer knowledge are *not* required.

To use R, install it from https://cran.r-project.org/, free of charge.

To use Python via Jupiter notebook, install Anaconda from https://www.anaconda.com/, start it, and click "Launch" in the Jupiter Notebook box. Here is a nice installation video guide.

Pre-requisites: STAT 415/615 "Regression" or STAT 520 "Applied Multivariate Analysis". These courses have their own pre-requisites, so an intermediate-level knowledge of statistical methods is assumed.

Assignments and Grading:

Weekly homework assignments and mini-projects	15%	Homework will be assigned weekly and submitted via **Canvas**. A steady effort to work out all the assigned problems is essential for learning statistical methods and for the successful performance in this course. Complete homework solutions will be posted after the homework deadline. 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 software, R or Python. Late submissions may be accepted with a 40% deduction plus 10% per day.
Weekly quizzes	20%	15-minute quizzes at the end of each class. Each quiz covers the material of the preceding week and the latest homework. During a quiz, you may use **one cheat sheet**.
Midterm Test	20%	The midterm covers several chapters of the material. Taken in class; **notes and our course materials are allowed**. Time is limited – 1 hour 30 minutes.
Final project	15%	Using statistical machine learning methods, you will be asked to do the necessary modeling, data analysis, tuning, and cross-validation and either write a report or make a slide presentation (your choice) summarizing results and answering specific questions of your project. Work in groups of up to four; each group submits one report or makes one presentation.
Final Test	30%	The final test covers the 2nd part of the course, but it is cumulative indirectly because the 2nd part is heavily based on the 1st part. Taken in class, during our regular class time on December 10-11, 5:30 - 8:00. **Notes and our course materials are allowed,** but of course, serious preparation is essential for getting a good grade. Time: 2.5 hours.

90 – 100 % = A	87 – 90 % = A-	84 – 87 % = B+	80 – 84 % = B
77 – 80 % = B-	74 – 77 = C+	70 – 74 % = C	60 – 70 % = C-

Tips

- Collaboration on homework is ok. Even encouraged! Quizzes and exams are to be done individually.
- On quizzes and exams, *show your work*. I will grade your solutions, not your answers.
- *No late assignments*. However, it is possible to take an exam or quiz early, for a good reason, for example, a business trip or a religious holiday. So, plan ahead.
- A steady effort to review material and work out all the assigned problems is your best chance to succeed in this course. Always *keep up* with the course because material is built upon the previously covered concepts.
- Use your absolute right to ask questions in class and during office hours. For example, any homework problem can be discussed.
- For each exam and quiz, review all the new concepts, methods, formulas, etc. Try to understand the methods rather than to memorize them.
- For each quiz, it may be useful to prepare a brief *summary* of important formulas and methods that you may need. Arrange it on a sheet of paper in the most convenient way. Do the same for the exams! Such summaries will help you use your exam time efficiently.

Learning outcomes

Graduate students (STAT 627)	Undergraduate students (STAT 427)
Students will be able to: • Identify appropriate statistical learning methods for the given problem involving real data. • Analyze the underlying assumptions, verify them, and propose appropriate remedies for invalid assumptions. • Identify other possible problems with messy data, such as multicollinearity, understand their consequences, and propose solutions. • Create and use training and test data to evaluate the performance of the chosen regression and/or classification techniques and analyze the results. • Show, analytically or empirically, 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. • Illustrate results with appropriate plots and diagrams. • Assess ethical implications of the application of statistical machine learning for a given problem. • Communicate statistical machine learning methods, findings, results, and implications in oral presentations or written reports.	Students will be able to: • Identify appropriate statistical learning methods for the given problem involving real data. • Analyze the underlying assumptions, techniques available to verify them, and propose appropriate remedies for invalid assumptions. • Create and use training and test data to evaluate the performance of the chosen regression and/or classification techniques and analyze the results. • Use available empirical tools to find the optimal balance between precision within training data and prediction power. • Illustrate results with appropriate plots and diagrams. • Assess ethical implications of the application of statistical machine learning for a given problem. • Communicate statistical machine learning methods, findings, results, and implications in oral presentations or written reports.

Students will demonstrate competence in using different statistical learning methods involving large, messy, and multi-dimensional numeric and categorical data. Methods include linear, logistic, and polynomial regression with proper variable selection, linear and quadratic discriminant analysis, K-nearest neighbor classifier, jackknife, bootstrap, ridge regression, lasso, principal components regression, partial least squares, splines, regression and classification trees, support vector machines, clustering, artificial neural networks, and related methods. In addition, graduate students (STAT 627) will demonstrate competency in the analytic justification of the chosen methods, tuning of the algorithms, and evaluating their prediction power.

Support Services

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

Mathematics & Statistics Tutoring Lab. Tutors should be able to help you with Calculus, Algebra, and basic Statistics, may be statistical software, but you should not count on getting homework solutions for advanced Statistics courses! The Math & Stats Lab offers both one-on-one and drop-in tutoring. Fall hours will be determined; usually the lab works Monday through Thursday 11^{00} am-8^{00} pm; Friday: 11^{00} am-3^{00} pm; and Sunday: 3^{00}–8^{00} pm in DMTI room 103. Online tutoring is available on https://american.mywconline.net/. Visit https://www.american.edu/provost/academic-access/mathstat.cfm or email tutoring@american.edu for more information. This service is *free* for all our students.

Software support - CTRL Connect, ctrl@american.edu, 202-885-2117, Help with Python, Help with R.

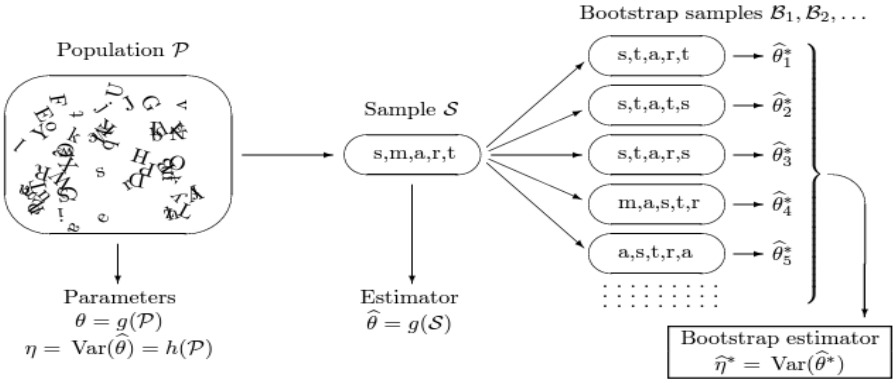
Counseling Center (x3500, https://www.american.edu/ocl/counseling/) offers counseling and consultations regarding personal concerns, self-help information, and connections to off-campus mental health resources.

Academic Support and Access Center (x3360) 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.

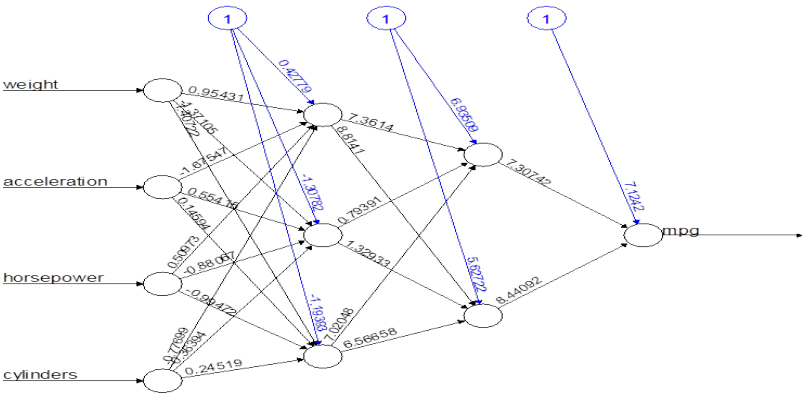
Religious Holidays. Students may receive accommodation in the course for the observance of a religious and/or cultural holiday. The student should notify the professor as soon as possible should such a need exist. More information about accommodations for religious and/or cultural holidays can be found at www.american.edu/ocl/kay/request-for-religious-accommodation.cfm.

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 the course web site.

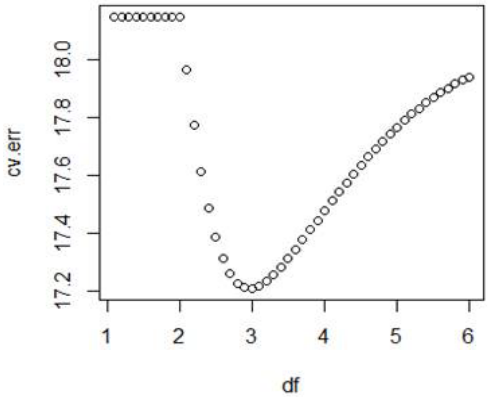
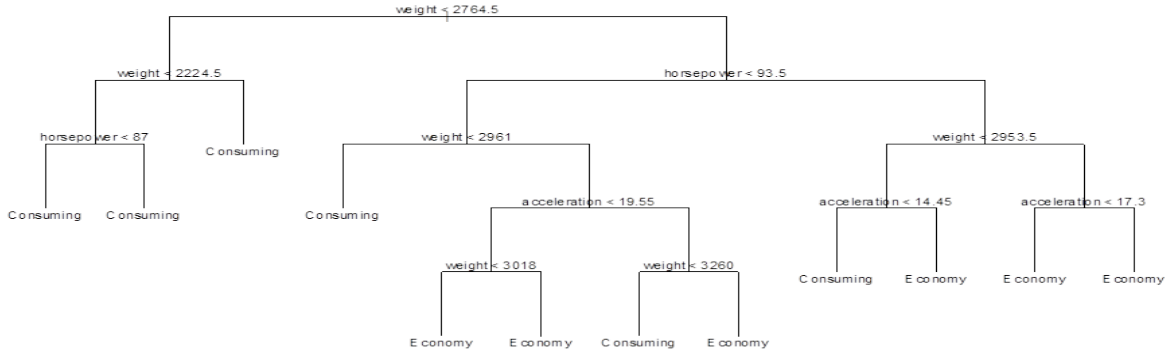
BOOTSTRAP



ARTIFICIAL NEURAL NETWORK



TUNING



TREE