USC Marshall
School of Business

DSO 699 - Advanced Data Science: Modeling, Computing, & Optimization

Professor: Jacob Bien (jbien@usc.edu)

Office hour: Thurs 2:45pm–3:45pm via zoom or by appointment

Website: Please visit our course website on `brightspace.usc.edu`

Units: 3

Schedule: TBD (3 hours/week) (classroom TBD)

Course description:

The field of statistical machine learning has developed a sophisticated framework of methods and models for meeting the complex challenges posed by modern data. These fundamental approaches form the foundation of data science and underpin the AI revolution. This course delves into those fundamentals, including important computational aspects. **A primary goal of this course is to teach students the core data science process: going from data to model; from model to algorithm; from algorithm to output; and from output to insight.**

Prerequisites: This class will make heavy use of linear algebra and multivariable calculus and will assume basic knowledge of linear models, probability, and `R`.

Recommended Textbooks:

All of these are available for free digitally either at the authors' websites or through `library.usc.edu`.

- *The Elements of Statistical Learning* by Trevor Hastie, Robert Tibshirani, and Jerome Friedman.
- *Foundations of Linear and Generalized Linear Models* by Alan Agresti.
- *Statistical Learning with Sparsity: The Lasso and Generalizations* by Trevor Hastie, Robert Tibshirani, and Martin Wainwright.

Grading:

Grades will be based on 3–4 homework assignments, scribing a lecture, and a final:

Category	
Homework	40%
Scribing	20%
R package	20%
Final	20%

Learning Outcomes: Upon successful completion of this course, students will be able to:

1. Determine reasonable modeling strategies for a wide variety of data analysis settings
2. Create code that follows best practices for reproducible research
3. Explain and apply the generalized linear model and regularization framework to statistical modeling in a way that shows deep and practical understanding
4. Develop custom optimization algorithms for computing regularization-based estimators
5. Write an R package with literate programming

Homework: Discussing material and homework problems in study groups can be beneficial for learning and thus is allowed. However, when writing up your homework assignment, please do so on your own (all code should be written independently as well).

To receive full credit, please show all work. A complete answer to a problem includes an explanation or derivation (as appropriate).

Homework assignments should be written in R Markdown. R Markdown allows you to use both LaTeX and R code together. Writing documents with R Markdown is a key skill for reproducible research and data analysis. Reproducibility is an important property of computational research, so developing facility with this workflow will serve you well.[1] I recommend using `RStudio` when writing R Markdown.

`R` offers an impressive collection of statistical functions. That said, as much as possible please write your own implementation of functions. In some cases, it might not be clear how low-level to start. I'll try to make this clear, or you can ask me. Implementing a method yourself is the best way to make sure you fully understand an algorithm. Often when coding an algorithm you will realize that there are additional considerations that have to be made.

Scribing: Each of you will sign up for one class period where you will be the scribe, meaning that you will write up (in R Markdown) lecture notes for that class.

R package: Writing fully functional and well documented R packages is an important research

[1] For example, the *Journal of the American Statistical Association* now requires reproducible code with every published paper: `https://jasa-acs.github.io/repro-guide/`

skill for publishing new methods in the field of statistics. There will be one special assignment in the class in which you are asked to write your own R package. The R package you write could be your own implementation of a statistical method that we have studied in class or that you have read in the literature; or, you can choose to implement an R package for a method you are developing in your research activities outside of this class. More details will be given at the appropriate time in the semester. Your R package will be written using `litr`, `https://jacobbien.github.io/litr-project/`, a framework that uses literate programming for R package development.

Schedule

This course schedule and the topics covered are subject to change.

Week	Topics	Deliverables
Week 1	Linear models (linear algebra + normal theory: projection, QR / Gram Schmidt implications for interpretation of coefficients, SVD + Moore-Penrose)	
Week 2	Dummy coding, interactions, connections to ANOVA	
Week 3	Generalized linear models: Fundamentals... Exponential families + IRLS/Newton-Raphson	
Week 4	Key examples of GLMs	HW 1 due
Week 5	Logistic regression (prospective/retrospective sampling, Bradley-Terry, class imbalance, etc.) + multinomial regression	
Week 6	Count data (Poisson regression, mover-stayer models, connection to multinomial regression via the Poisson trick, overdispersion and the negative binomial)	
Week 7	Designing and solving custom convex regularizers for data modeling (e.g., trend filtering, hierarchical sparse modeling)	HW 2 due
Week 8	Optimization methods for statistical modeling (proximal gradient, coordinate descent)	
Week 9	Optimization methods for statistical modeling (continued; ADMM)	
Week 10	Degrees of freedom and SURE for regularized regression	HW 3 due
Week 11	Nonparametric regression, smoothing splines, generalized additive models	
Week 12	Modeling dependence with covariance estimation	
Week 13	Unsupervised learning: mixture of Gaussians + EM	HW 4 due
Week 14	Unsupervised learning (continued): k-means, hierarchical, and other clustering	
Week 15	Unsupervised learning (continued): PCA and UMAP	R package due

Due dates for scribing will be student-specific and assigned during the first week of class. The final exam timing will be announced during the first week of class.

Class Notes Policy Pursuant to the USC Student Handbook (https://policy.usc.edu/studenthandbook/, page 27), students may not record a university class without the express permission of the in-

structor and announcement to the class. In addition, students may not distribute or use notes or recordings based on USC classes or lectures without the express permission of the instructor for purposes other than personal or class-related group study by individuals registered for the class. This restriction on un-authorized use applies to all information that is distributed or displayed for use in relationship to the class. Violation of this policy may subject an individual or entity to university discipline and/or legal proceedings.

AI Usage Policy The use of artificial intelligence (AI)-powered programs (such as ChatGPT, Github Copilot, etc.) to help you with assignments is permitted with several caveats:

- The purpose of taking this course and the assignments are for you to learn the material. An over-reliance on AI tools might short-circuit your ability to learn the material, so I would strongly encourage you to use AI sparingly. In particular, please always start by attempting a problem completely on your own.
- You should be aware that AI text generation tools may present incorrect information, biased responses, and incomplete analyses; thus they are not yet prepared to produce text that meets the standards of this course. You are responsible for the correctness of the work you submit.
- To adhere to our university values, you must cite any AI-generated material (e.g., text, images, etc.) included or referenced in your work and provide the prompts used to generate the content. Using an AI tool to generate content without proper attribution will be treated as plagiarism and reported to the Office of Academic Integrity.

Open Expression and Respect for All An important goal of the educational experience at USC Marshall is to be exposed to and discuss diverse, thought-provoking, and sometimes controversial ideas that challenge one's beliefs. In this course we will support the values articulated in the USC Marshall "Open Expression Statement."

Technology requirements We will be using `RStudio` and `R`, which are both freely available online. You are responsible for ensuring that you have the necessary computer equipment and reliable internet access. You are invited to explore what lab or loaner options exist. Contact the Marshall HelpDesk (213-740-3000 or HelpDesk@marshall.usc.edu) if you need assistance.

Academic Integrity

The University of Southern California is foremost a learning community committed to fostering successful scholars and researchers dedicated to the pursuit of knowledge and the transmission of ideas. Academic misconduct is in contrast to the university's mission to educate students through a broad array of first-rank academic, professional, and extracurricular programs and includes any act of dishonesty in the submission of academic work (either in draft or final form).

This course will follow the expectations for academic integrity as stated in the USC Student Handbook. All students are expected to submit assignments that are original work and prepared specifically for the course/section in this academic term. You may not submit work written by others or "recycle" work prepared for other courses without obtaining written permission from the instructor(s). Students suspected of engaging in academic misconduct will be reported to the Office of Academic Integrity.

Other violations of academic misconduct include, but are not limited to, cheating, plagiarism, fabrication (e.g., falsifying data), knowingly assisting others in acts of academic dishonesty, and any act that gains or is intended to gain an unfair academic advantage.

Academic dishonesty has a far-reaching impact and is considered a serious offense against the university. Violations will result in a grade penalty, such as a failing grade on the assignment or in the course, and disciplinary action from the university itself, such as suspension or even expulsion.

For more information about academic integrity see the student handbook or the Office of Academic Integrity's website, and university policies on Research and Scholarship Misconduct.

Please ask your instructor if you are unsure what constitutes unauthorized assistance on an exam or assignment or what information requires citation and/or attribution.

Statement on University Academic and Support Systems

Students and Disability Accommodations:

USC welcomes students with disabilities into all of the University's educational programs. The Office of Student Accessibility Services (OSAS) is responsible for the determination of appropriate accommodations for students who encounter disability-related barriers. Once a student has completed the OSAS process (registration, initial appointment, and submitted documentation) and accommodations are determined to be reasonable and appropriate, a Letter of Accommodation (LOA) will be available to generate for each course. The LOA

must be given to each course instructor by the student and followed up with a discussion. This should be done as early in the semester as possible as accommodations are not retroactive. More information can be found at osas.usc.edu. You may contact OSAS at (213) 740-0776 or via email at osasfrontdesk@usc.edu.

Student Financial Aid and Satisfactory Academic Progress:
To be eligible for certain kinds of financial aid, students are required to maintain Satisfactory Academic Progress (SAP) toward their degree objectives. Visit the Financial Aid Office webpage for undergraduate- and graduate-level SAP eligibility requirements and the appeals process.

Support Systems:

Counseling and Mental Health - (213) 740-9355 – 24/7 on call
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

988 Suicide and Crisis Lifeline - 988 for both calls and text messages – 24/7 on call
The 988 Suicide and Crisis Lifeline (formerly known as the National Suicide Prevention Lifeline) provides free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week, across the United States. The Lifeline consists of a national network of over 200 local crisis centers, combining custom local care and resources with national standards and best practices. The new, shorter phone number makes it easier for people to remember and access mental health crisis services (though the previous 1 (800) 273-8255 number will continue to function indefinitely) and represents a continued commitment to those in crisis.

Relationship and Sexual Violence Prevention Services (RSVP) - (213) 740-9355(WELL) – 24/7 on call
Free and confidential therapy services, workshops, and training for situations related to gender- and power-based harm (including sexual assault, intimate partner violence, and stalking).

Office for Equity, Equal Opportunity, and Title IX (EEO-TIX) - (213) 740-5086
Information about how to get help or help someone affected by harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants.

Reporting Incidents of Bias or Harassment - (213) 740-2500
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office for Equity, Equal Opportunity, and Title for appropriate investigation, supportive measures, and response.

The Office of Student Accessibility Services (OSAS) - (213) 740-0776
OSAS ensures equal access for students with disabilities through providing academic accommodations and auxiliary aids in accordance with federal laws and university policy.

USC Campus Support and Intervention - (213) 740-0411
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity, Equity and Inclusion - (213) 740-2101
Information on events, programs and training, the Provost's Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety - UPC: (213) 740-6000, HSC: (323) 442-1200 – 24/7 on call
Non-emergency assistance or information.

Office of the Ombuds - (213) 821-9556 (UPC) / (323-442-0382 (HSC)
A safe and confidential place to share your USC-related issues with a University Ombuds who will work with you to explore options or paths to manage your concern.

Occupational Therapy Faculty Practice - (323) 442-2850 or otfp@med.usc.edu
Confidential Lifestyle Redesign services for USC students to support health promoting habits and routines that enhance quality of life and academic performance.

Revised June 2024