Machine Learning for Social Scientists
PS 132B, Spring 2023
Tuesday, Thursday 12:30pm - 2:00pm Pacific Time
Location: Moffitt 102

Professor: Kirk Bansak
Contact: kbansak@berkeley.edu
Office Hours: Tuesday, 9:30am - 11:30am (sign-up details below)
Location: Social Sciences Building, Room 736

GSI: Alexander Stephenson
Contact: alexander_stephenson@berkeley.edu
Office Hours: TBD
Location: TBD

1 Overview

Social scientists and policymakers increasingly use large quantities of data to make decisions and test theories. For example, political campaigns use surveys, marketing data, and previous voting history to optimally target get out the vote drives. Governments deploy predictive algorithms in an attempt to optimize public policy processes and decisions. And political scientists use massive new data sets to measure the extent of partisan polarization in Congress, the sources and consequences of media bias, and the prevalence of discrimination in the workplace. Each of these examples, and many others, make use of statistical and algorithmic tools that distill large quantities of raw data into useful quantities of interest.

1.1 Objectives

This course introduces techniques to collect, analyze, and utilize large collections of data for social science inferences. The ultimate goal of the course is to introduce students to modern machine learning techniques and provide the skills necessary to apply the methods widely. In achieving this ultimate goal, students will also:

1) Learn about core concepts in machine learning and statistics, developing skills that are transferable to other types of data and inference problems.

2) Develop their programming abilities in `R`.

3) Be introduced to substantive problems and participate in challenges applying the techniques from the course.

1.2 Prerequisites

Students must have taken PS 3 or Data 8 (or have equivalent coursework). If you have any questions regarding whether you're prepared for the class, please talk to the teaching staff.

1.3 Evaluation

Students will be evaluated across the following areas.

Problem Sets 40% of final grade. Students will complete six problem sets. These assignments are intended to expand upon the lecture material and to help students develop the actual skills that will be useful for applied work. Problem sets will be completed and submitted using `R Markdown`, a markup language for producing well-formatted HTML documents with embedded `R` code and outputs. `R Markdown` requires installation of the `knitr` package. We recommend using `RStudio`, an IDE for `R`, which is set up well for the creation of `R Markdown` documents.

More about `RStudio` can be found here:
`http://www.rstudio.com`

`R Markdown` can be found here:
`http://rmarkdown.rstudio.com`

Challenges 25% of final grade. Students will complete two team-based machine learning challenges during the course. The challenges will allow students to apply the techniques learned in the course to real problems. The teaching staff will provide more details and specify the guidelines for each challenge upon assignment. The challenges will be the following:

1. Predicting recidivism. We will provide you with a real-world data set of criminal defendants that has been studied by scholars and journalists in the past. You'll work in teams to build and train models that predict which defendants are likely to commit another crime. In the process of building and evaluating your predictive models, you will also explore the value, risks, and ethics of applying machine learning in public policy areas like criminal justice.

2. Analyzing political text data. We will provide you with a data set of text pertaining to a particular political event or figure. Working in teams, you will employ methods you have learned in the class (supervised and/or unsupervised) to analyze the text data and identify key insights that would have been difficult or impossible to discover without computational methods.

Midterm Exam 15% of final grade. Students will complete a midterm exam, valued at 15% of the final grade.

Final Exam 20% of final grade. Students will complete a cumulative final exam, at the time and date delineated under Berkeley's official exam schedule.

2 Logistics

2.1 Class Meetings

Unless otherwise announced, all class meetings will be conducted in person at our scheduled class time (Tuesday, Thursday 12:30pm - 2:00pm Pacific Time) in Moffitt 102.

2.2 Office Hours

I will hold office hours from 9:30am to 11:30am Pacific Time on Tuesdays in the Social Sciences Building, Room 736. Please make sure to sign up for office hours in advance using the Calendly link: `https://calendly.com/kbansak/officehours`. If you would like to meet but have class during my office hours, please email me to arrange an alternative time.

2.3 Graduate Student Instructor

Alexander Stephenson (`alexander_stephenson@berkeley.edu`) will be the GSI for this course. Alex will be holding office hours from TBD to TBD Pacific Time in TBD.

2.4 Discussion Section

Discussion sections led by our GSI will be held on Fridays at 12:00pm - 1:00pm and 1:00pm - 2:00pm in Wurster 101.

2.5 Course Website

As our primary course website, we will use our bCourses site.

We will distribute course materials, including lecture slides and problem sets, on our course website. There is also a discussion platform (Ed Discussion) that is easy to use and designed to get you answers to questions quickly.

If you have non-personal questions related to course material or logistics, we encourage you to post these questions on Ed Discussion rather than emailing the course instructors. Using Ed Discussion will allow students to see and learn from other students' questions. Course instructors will regularly check the board and answer questions posted, although everyone else is also encouraged to contribute to the discussion and help answer questions. Please be respectful and constructive in your participation on the forum.

2.6 Required Readings

The required readings throughout the course are listed in the course outline below. Each reading is associated with a class meeting on a specific date. It is recommended that you complete each reading *in advance of* the associated class meeting.

As our primary reference, we will use the book listed below:

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*, Second Edition, 2021.

Note that we are using the Second Edition of the book! This is important as the page numbers are different from the First Edition.

This book is referred to as *ISLR* in the course outline. The book can be accessed online and downloaded for free here:
`https://www.statlearning.com/`

In addition, readings from other sources are also assigned.

2.7 Other Recommended Reference Books

You might also consider the following books as useful references (but not required).

Murphy, Kevin P. *Machine Learning: A Probabilistic Perspective*. A slightly more advanced text, but an excellent treatment of machine learning methods.

Bishop, Christopher M. *Pattern Recognition and Machine Learning*. A more computer science oriented treatment of machine learning, with more extensive treatment of the estimation techniques used for machine learning methods.

3 Course Outline

Introduction	Week 1	01/17	Introduction
	Week 1	01/19	A Machine Learning Focus on Regression *Read: ISLR pp. 15 - 39*
Unit 1: **R**	Week 2	01/24	Data and Datasets *Read: Kelleher and Tierney (2018); Kaplan (2009)*
	Week 2	01/26	(Re)Introduction to `R` *Do: Install/Update `R` and `RStudio` on your computer*

	Week 3	01/31	Introduction to `R Markdown` *Do: Install the* ***`tidyverse`*** *and* ***`knitr`*** *packages* **HW 1 assigned**
	Week 3	02/02	Diving Deeper into `R`: Core Functionality *Read: Venables et al. (2022), Chapters 2, 3, and 6*
	Week 4	02/07	Diving Deeper into `R`: Data Visualization and Exploration *Read: Wickham and Grolemund (2017), Chapter 3* **HW 1 due, HW 2 assigned**
	Week 4	02/09	Diving Deeper into `R`: Functions and Iteration *Read: Hadley and Grolemund, Chapters 19 and 21*
Unit 2: **Supervised Learning**	Week 5	02/14	The Bootstrap and Linear Regression *Read: ISLR pp. 59 - 92, 209 - 212* **HW 2 due, HW 3 assigned**
	Week 5	02/16	Linear Regression/Classification *Read: ISLR pp. 129 - 133*
	Week 6	02/21	Classification and Logistic Regression *Read: ISLR pp. 129 - 141*
	Week 6	02/23	Training, Testing, Bias-Variance Tradeoff *Read: ISLR pp. 33 - 36, 198 - 200* **HW 3 due, HW 4 assigned**
	Week 7	02/28	LASSO and Ridge Regression *Read: ISLR pp. 237 - 248, 250 - 251*
	Week 7	03/02	Cross-Validation *Read: ISLR pp. 197 - 208* **HW 4 due, HW 5 assigned**
	Week 8	03/07	Classification and Regression Trees *Read: ISLR pp. 327 - 340*
	Week 8	03/09	Random Forests and Boosted Trees *Read: ISLR pp. 340 - 352* **HW 5 due, Challenge 1 assigned**
	Week 9	03/14	Evaluating and Selecting Models *Read: Swalin (2018) Part 1; Swalin (2018) Part 2*
	Week 9	03/16	Machine Learning, Policy, and Ethics *Read: Kleinberg et al. (2016); Buchanan and Miller (2017)*
	Week 10	03/21	Machine Learning, Policy, and Ethics: Case Study

			Read: Obermeyer et al. (2019) **Challenge 1 due**
	Week 10	03/23	**Midterm**
	Week 11	03/28	Spring Recess
	Week 11	03/30	Spring Recess
Unit 3: **Unsupervised Learning**	Week 12	04/04	Intro to Unsupervised Learning & Text as Data *Read: ISLR pp. 497 - 498; Grimmer and Stewart (2013) pp. 1-7*
	Week 12	04/06	Text Analysis and Dictionary Methods *Read: Grimmer and Stewart (2013) pp. 7-14; Loughran and McDonald (2011)*
	Week 13	04/11	Distance, Clustering, and Text Applications *Read: ISLR pp. 516 - 532; Grimmer and Stewart (2013) pp. 14-17* **HW 6 assigned**
	Week 13	04/13	Topic Models for Text Analysis *Read: Mohr and Bogdanov (2013); Grimmer and Stewart (2013) pp. 17-25*
	Week 14	04/18	Principal Components Analysis *Read: ISLR pp. 498 - 510* **HW 6 due, Challenge 2 assigned**
	Week 14	04/20	Other Topics in Dimensionality Reduction *Read: TBD*
Conclusions	Week 15	04/25	Data Wrangling *Read: TBD*
	Week 15	04/27	Review and Next Steps **Challenge 2 due**
	Week 16		Reading, Review, and Recitation Week
	Week 17	05/??	**Final Exam**

3.1 Readings

- **ISLR**: Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Second Edition, 2021.
- **Buchanan and Miller (2017)**: Ben Buchanan and Taylor Miller, "Machine Learning for Policymakers: What It Is and Why It Matters," Belfer Center for Science and International Affairs, 2017.
- **Grimmer and Stewart (2013)**: Justin Grimmer and Brandon M. Stewart, "Text as

Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis* Vol. 21, No. 3 (2013).

- **Kaplan (2009)**: Daniel T. Kaplan, "Data: Cases, Variables, Samples" in *Statistical Modeling: A Fresh Approach*, Project MOSAIC Books, 2009.

- **Kelleher and Tierney (2018)**: John D. Kelleher and Brendan Tierney, "What are Data, and What is a Data Set?" in *Data Science*, MIT Press, 2018.

- **Kleinberg et al. (2016)**: Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, "A Guide to Solving Social Problems with Machine Learning," *Harvard Business Review*, December 8, 2016. Available at: https://hbr.org/2016/12/a-guide-to-solving-social-problems-with-machine-learning

- **Loughran and McDonald (2011)**: Tim Loughran and Bill McDonald, "When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *The Journal of Finance* Vol. 66, No. 1 (2011).

- **Mohr and Bogdanov (2013)**: John W. Mohr and Petko Bogdanov, "Introduction–Topic Models: What They Are and Why They Matter," *Poetics* Vol. 41, No. 6 (2013).

- **Obermeyer et al. (2019)**: Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science* Vol. 366, No. 6464 (2019).

- **Swalin (2018) Part 1**: Alvira Swalin, "Choosing the Right Metric for Evaluating Machine Learning Models: Part 1," Towards Data Science, *Medium*, April 6, 2018. Available at: https://medium.com/usf-msds/choosing-the-right-metric-for-machine-learning-models-part-1-a99d7d7414e4

- **Swalin (2018) Part 2**: Alvira Swalin, "Choosing the Right Metric for Evaluating Machine Learning Models: Part 2," Towards Data Science, *Medium*, May 2, 2018. Available at: https://medium.com/usf-msds/choosing-the-right-metric-for-evaluating-machine-learning-models-part-2-86d5649a5428

- **Venables et al. (2022)**: W. N. Venables, D. M. Smith, and the R Core Team, *An Introduction to R*, Available at: https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

- **Wickham and Grolemund (2017)**: Hadley Wickham and Garrett Grolemund, "R for Data Science," O'Reilly, 2017, Available at: https://r4ds.had.co.nz/index.html

4 Policies

4.1 Academic Integrity

You are a member of an academic community at one of the world's leading research universities. Universities like Berkeley create knowledge that has a lasting impact in the world of ideas and on the lives of others; such knowledge can come from an undergraduate paper as well as the lab of an internationally known professor. One of the most important values of an academic community is the balance between the free flow of ideas and the respect for the intellectual property of others. Researchers don't use one another's research without permission; scholars and students always use proper citations in papers; professors may not circulate or publish student papers without the writer's permission; and students may not circulate or post materials (handouts, exams, syllabi–any class materials) from their classes without the written permission of the instructor.

Any test, paper or report submitted by you and that bears your name is presumed to be your own original work that has not previously been submitted for credit in another course unless you obtain prior written approval to do so from your instructor. In all of your assignments, including your homework or drafts of papers, you may use words or ideas written by other individuals in publications, web sites, or other sources, but only with proper attribution. If you are not clear about the expectations for completing an assignment or taking a test or examination, be sure to seek clarification from your instructor or GSI beforehand. Finally, you should keep in mind that as a member of the campus community, you are expected to demonstrate integrity in all of your academic endeavors and will be evaluated on your own merits. The consequences of cheating and academic dishonesty—including a formal discipline file, possible loss of future internship, scholarship, or employment opportunities, and denial of admission to graduate school—are simply not worth it.

Finally, note that UC Berkeley's honor code states "As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others." As a tool to promote academic integrity in this course, written work submitted via bCourses may be checked for originality using Turnitin. Turnitin compares student work to a database of books, journal articles, websites, and other student papers. This creates an opportunity for students to improve their academic writing skills, by ensuring that other sources have been properly cited and attributed. For more information about Turnitin at UC Berkeley, visit `http://ets.berkeley.edu/academic-integrity`.

4.2 Student Accessibility

Please see me as soon as possible if you need particular accommodations, and we will work out the necessary arrangements. Students requesting accommodations for this course due to a disability must provide a current Letter of Accommodation (LOA) issued by the Disabled Students' Program (DSP) (`https://dsp.berkeley.edu/`).

4.3 Scheduling Conflicts

Please notify me in writing by the second week of the term about any known or potential extracurricular conflicts (such as religious observances, graduate or medical school interviews, or team activities). I will try my best to help you with making accommodations, but cannot promise them in all cases. In the event there is no mutually-workable solution, you may be dropped from the class.