Machine Learning for Social Scientists

DACSS 756 (3 Cred.)

University of Massachusetts Amherst

Spring 2025

Instructor

Dr. Omer Yalcin
Bartlett Hall 259
oyalcin@umass.edu

Course Time and Location

Time: Monday & Wednesday, 10:00 AM - 11:15 AM
Location: Machmer W-13 or Zoom

Please find the Zoom link in Canvas.

Office Hours

You can join office hours either in person in my office or on Zoom. Please book an appointment in advance at https://omerfyalcin.youcanbook.me/. If the time slots listed there do not work for you, then please email me for an appointment. I am also happy to arrange evening Zoom slots for those cannot make it during the day.

Course Description & Objectives

This course will provide an overview of machine learning (ML) with special attention to applications for the social and behavioral sciences. Machine learning combines insights from artificial intelligence, probability theory, statistical inference, and information theory to help automate tasks involving pattern recognition, prediction, and classification. "Learning" is loosely analogous to "inference" in statistics and, in fact, the modern statistical toolkit includes various machine learning methods. The course focuses on statistical learning and is a good second or third course in statistical methods for graduate students in the social and behavioral sciences. We will examine key techniques of supervised and unsupervised learning and reflect upon appropriate and inappropriate applications of such approaches for those seeking to understand the social world. We shall also discuss the ethical issues involved in automated analysis and computer-assisted decision-making, including how they may in some cases help overcome human biases and in others only serve to reinforce these tendencies. Topics covered will vary, but should include consideration of key ideas such as the bias-variance tradeoff, the curse of dimensionality, model selection, cross-validation, regularization,

. I would like to thank Justin Gross for permitting me to adopt portions of his syllabus. The syllabus is subject to change with reasonable advance notice.

and dimension reduction. We will look at linear regression from a statistical learning perspective–including extensions such as lasso, ridge regression, and principal components regression–as well as tree-based methods, support-vector machines, and a brief introduction to neural networks. Unsupervised methods will include principal component analysis, and a number of clustering techniques (e.g. k-means clustering, hierarchical clustering).

Prerequisites

Basic knowledge of algebra, sets, functions, and probability is assumed. More advanced mathematics will be introduced as needed for those who may not have encountered such material previously (and as review for others), but we will focus on conceptual understanding and enhancing our ability to read common mathematical notation. Students are expected to have had some experience using R or Python prior to this class, preferably at the level of either DACSS 601 Introduction to Data Science or DACSS 611 Introduction to Python for Data Science, but it is possible to succeed without this background if you are willing to put in extra work toward the beginning of the semester. Additionally, it is highly recommended that students take a statistics or econometrics course that includes linear regression before taking this course, such as DACSS 603 or similar.

Textbook

All required course material is either freely available on the internet or freely available to read for *you* through the UMass Library or Canvas.

The main textbook is available online in two versions: an R version and a Python version. The two books are pretty much the same except for their coding lab sections. You should feel free to study using either of them. (More on R vs Python later in the syllabus). Both versions are available at https://www.statlearning.com/. Even though these are technically two separate books, one in R and one in Python, the syllabus will refer to both of them as a single book, **[ISL]** Introduction to Statistical Learning. Below is a more formal reference:

- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning: with Applications in R.* 2nd ed. 2021 edition. New York, NY: Springer, July. ISBN: 978-1-0716-1417-4. https://www.statlearning.com/

- James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor. 2023. *An Introduction to Statistical Learning: with Applications in Python.* 1st ed. 2023 edition. Cham, Switzerland: Springer, July. ISBN: 978-3-031-38746-3. https://www.statlearning.com/

The Course Schedule section lists the material associated with every week.

Below are some resources that we will not directly use, but that may be useful as further reference both during and beyond this course:

- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2016. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition.* 2nd edition. New York, NY: Springer, January. ISBN: 978-0-387-84857-0 (A more technical treatment of [ISL], by a subset of the same authors. Fulltext pdf at https://hastie.su.domains/ElemStatLearn/)

- Müller, Andreas C., and Sarah Guido. 2016. *Introduction to Machine Learning with Python: A Guide for Data Scientists.* 1st edition. Sebastopol, CA: O'Reilly Media, November. ISBN: 978-1-4493-6941-5 (available online via UMass Library at https://learning.oreilly.com/library/view/~/9781449369880/?ar. Click on "Institution not listed?" and log in via UMass email address)

- Witten, Ian H., Eibe Frank, Mark A. Hall, and Christopher J. Pal. 2016. *Data Mining: Practical Machine Learning Tools and Techniques.* 4th edition. UK USA Canada Ireland Australia India New Zealand South Africa: Morgan Kaufmann, December. ISBN: 978-0-14-198845-0 (available online via UMass Library at https://learning.oreilly.com/library/view/~/9780128043578/?ar Click on "Institution not listed?" and log in via UMass email address. The book is also accompanied by the website https://ml.cms.waikato.ac.nz/weka/ and the Weka software)

- Boehmke, Brad, and Brandon M. Greenwell. 2019. *Hands-On Machine Learning with R.* 1st edition. Boca Raton London New York: Chapman / Hall/CRC, November. ISBN: 978-1-138-49568-5 (book: https://bradleyboehmke.github.io/HOML/, supplementary website: https://koalaverse.github.io/homlr/)

- VanderPlas, Jake. 2017. *Python Data Science Handbook: Essential Tools for Working with Data.* 1st edition. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly Media, January. ISBN: 978-1-4919-1205-8 (fulltext available online at https://jakevdp.github.io/PythonDataScienceHandbook/. Also available online via UMass Library at: https://learning.oreilly.com/library/view/~/9781491912126/?ar)

- Alpaydin, Ethem. 2020. *Introduction to Machine Learning, fourth edition.* 4th edition. Cambridge, Massachusetts: The MIT Press, March. ISBN: 978-0-262-04379-3 (third edition from 2014 available online via UMass Library at http://ebookcentral.proquest.com/lib/uma/detail.action?docID=3339851)

- Raschka, Sebastian, and Vahid Mirjalili. 2019. *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2.* 3rd edition. Birmingham: Packt Publishing, December. ISBN: 978-1-78995-575-0 (available online via UMass Library at https://learning.oreilly.com/library/view/~/9781789955750/?ar)

- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning.* Cambridge, Massachusetts: The MIT Press, November. ISBN: 978-0-262-03561-3 (freely available online, just need to be clicked chapter by chapter: https://www.deeplearningbook.org/)

- Chollet, Francois. 2021. *Deep Learning with Python, Second Edition.* 2nd edition. Shelter Island: Manning, December. ISBN: 978-1-61729-686-4 (available online via UMass Library at https://learning.oreilly.com/library/view/~/9781617296864/?ar)

- Chollet, Francois, Tomasz Kalinowski, and J. J. Allaire. 2022. *Deep Learning with R, Second Edition.* 2nd ed. edition. Shelter Island, NY: Manning, July. ISBN: 978-1-63343-984-9 (available online via UMass Library at https://learning.oreilly.com/library/view/~/9781633439849/?ar)

- Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities.* Cambridge, Massachusetts: The MIT Press, December. ISBN: 978-0-262-04861-3 (fulltext available as pdf: https://fairmlbook.org/)

- Kuhn, Max, and Kjell Johnson. 2013. *Applied Predictive Modeling.* 2013th edition. New York: Springer, May. ISBN: 978-1-4614-6848-6
- Lones, Michael A. 2021. "How to avoid machine learning pitfalls: a guide for academic researchers." *arXiv preprint arXiv:2108.02497* (almost like a mini-book, pdf at https://arxiv.org/pdf/2108.02497)

Learning Management System

Classroom material, including slides, resources, assignments, announcements, deadlines, and more will be posted in our learning management system, Canvas. Canvas can be accessed at https://www.umass.edu/it/canvas. We will also use Canvas for asynchronous discussion. More on this in the participation section.

Kaggle

Throughout the course, we will actively use the Kaggle platform (https://www.kaggle.com/). Kaggle is a platform for machine learning and data science competitions. It allows for users to compete with each other to develop the best models for a given problem. Users can share and collaborate on code, data, and models. Kaggle provides a public leaderboard to track progress, and a forum for participants to communicate and share knowledge. Kaggle competitions are often used by companies and organizations as a way to crowdsource solutions to difficult data science problems. Kaggle is also a great community of people learning data science, with lots of useful tips, resources, code, and data constantly being shared.

Some of the assignments will need to be wholly or partially completed on Kaggle by submitting solutions to a Kaggle competition. The purpose is to make learning more fun and engaging in friendly competition. It will also be more realistic as you will get to work on problems without a known solution. Even beyond the assignments, you are encouraged to try new solutions and submit them to the competition.

You will need to sign up for a Kaggle account to participate in the competitions. Kaggle does not allow one to change their username after signing up. Each competition has a public leaderboard that ranks people based on how good their solution is, based on a criteria such as mean squared error (for regression) or accuracy (for classification). For the purposes of this course, you may choose a username that may or may not be your actual name—that is up to you. It is okay to create a "burner" account for the purposes of this course, if you so prefer. I will need to know your username, but not necessarily anyone else. Kaggle allows submissions in R and Python and students may choose to use either of the languages for their solutions.

Lectures

We will use synchronous class time to do hands-on activities including coding exercises and other practical applications to reinforce the week's concepts, topics, and algorithms. It is essential that students come to classes having studied the week's material and ready to do the exercises. Please bring your laptop to class if possible, but if you cannot for some reason, please do not let that prevent you from attending (we can find solutions. Just let me know!). All class sessions will be recorded via Zoom and posted on Canvas afterwards. Students who cannot attend class synchronously should watch them later and do the in-class exercises

Studying the week's materials include doing all the readings listed in the Course Schedule, working through the relevant portion of the end-of-chapter code in the [**ISL**] book (i.e. actually run the code line by line, do not just 'read' it), and watching the videos posted under that week in Canvas. Some of the videos will be from the authors of the [**ISL**] book specifically prepared to accompany the book.

Software

Python and R are the officially supported languages of the course. Students can complete the assignments and the final project in R or Python (and it is fine to complete some of them in R and some of them in Python). Being comfortable with at least one of Python or R is a prerequisite. Classroom demonstrations will feature a mix of both languages, but they will be presented in a way that focuses on the concepts and ideas, rather than the specific syntax.

Grading

Grades are calculated as follows:

- Participation (10%)
- Homeworks (50%)
- Final Project (40%)

Participation: It is imperative that students actively and regularly participate in class discussion. Canvas's discussion section will be the primary way of participating in class discussion asynchronously. Students are expected to regularly ask and/or answer questions about code, assignments and so on. You can also participate by sharing insights, material, and other resources on Canvas. Participation does not need to reflect expertise. Synchronous participation during class is another additional opportunity to participate, but online students who may be unable to join synchronously will not be disadvantaged in grading.

Homework Assignments: There will be a homework assignment most weeks, a total of eight. The assignments will include exercises to help you better understand concepts and methods covered during class. Collaboration is acceptable, but please write up your own answers and **your own code**; do **not** hand in identical written responses. Please get started on the final project early and start making progress. See which weeks we have assignments and which of them are related to the final in Course Schedule

Deadlines and Late Work: Most assignments will be due on a Monday at 11:59 pm. (See exceptions at Course Schedule) Each assignment—but not the final project—will have two full days of a grace period following the deadline. That means you can return the assignment penalty-free until two days after the deadline, for example, at Wednesday at 11:59 pm for a deadline on Monday, 11:59 pm. However, if you do make use of the grace period, then you may also receive feedback later than those that return on time. If, without a valid excuse, an assignment is returned even later than the end of the grace period, then it will be penalized by a 10% deduction of the full grade for every day that it is late. ("Day" here means anything that runs into the day. If the grace period ended on Wednesday 11:59 PM, then 20 minutes later, Thursday 12:20 am counts as 1 day late). If you think you will not be able to meet a deadline for a good reason

and you contact me by email at least 24 hours in advance of the deadline, we can work out a new deadline. "Deadline" here means the actual deadline on the syllabus and the Canvas, not the end of the grace period.

Final Project: Please find the details of the final project in Canvas.

Grading Rubric for Final Course Grade:

Final letter grades are assigned using the University's Plus-Minus Grading Scale according to following rubric:

- A (94-100%)
- A- (90-93%)
- B+ (86-89%)
- B (81-85%)
- B- (77-80%)
- C+ (74-76%)
- C (70-73%)
- F (Below 70%)

Email Policy

I expect you to have lots of questions throughout the course. Please ask them! When you have a question about course material, it is likely that others in the class have the same question. It is also likely that someone else in the class can answer the question faster or better. Therefore, students are encouraged to ask their questions on Canvas's discussion section, publicly. For questions that are relevant for you but not for other students or for other types of concerns, please feel free to email me. In most cases, you can expect a response within one business day.

Incomplete Grade Policy

For your reference, a copy of the essential sections of the UMass policy regarding Incomplete grades has been provided below. (More information can be found on page 28 of the following document: https://www.umass.edu/registrar/sites/default/files/academicregs.pdf):

> "Students who are unable to complete course requirements within the allotted time because of severe medical or personal problems may request a grade of Incomplete from the instructor of the course. Normally, incomplete grades are warranted only if a student is passing the course at the time of the request and if the course requirements can be completed by the end of the following semester. Instructors who turn in a grade of "INC" are required to leave a written record of the following information with the departmental office of the academic department under which the course is offered: (1) the percentage

of work completed, (2) the grade earned by the student on the completed work, (3) a description of the work that remains to be completed, (4) a description of the method by which the student is to complete the unfinished work, and (5) the date by which the work is to be completed. In the case of an independent study where the entire grade is determined by one paper or project, the instructor should leave with the department information pertaining to the paper or project, which will complete the course. To avoid subsequent misunderstanding it is recommended that the student also be provided with a copy of this information.

Grades of Incomplete will be counted as F's until resolved. If not resolved by the end of the following semester, they will automatically be converted to an F if taken before Fall 2004, to an IF if taken thereafter. Faculty wishing to extend this deadline must write to the Registrar's Office stipulating a specific date by which the incomplete will be completed."

AI Policy

AI-powered tools for generating and manipulating text, code, image and alike are fast becoming ubiquitous. ChatGPT is the most recent well-known example. GitHub Copilot is another tool that is especially relevant for coding. There are many more. Learning how to use these tools effectively is likely to become an important skill soon. You can use these tools in your work, but please remember that AI can make mistakes, plagiarize, or veer off the task you gave it. AI is best used as an assistant, rather than as a replacement for yourself. This usually means giving detailed, specific prompts and having AI help with bottlenecks in a project rather than trying to have it do the whole job by itself for you. If I do mention AI use as a problem in feedback on your work, it will not be because of AI use per se (as mentioned, that is fine and allowed) but because of these reasons. With these caveats, feel free to use it.

Academic Honesty Statement

Since the integrity of the academic enterprise of any institution of higher education requires honesty in scholarship and research, academic honesty is required of all students at the University of Massachusetts Amherst. Academic dishonesty is prohibited in all programs of the University. Academic dishonesty includes but is not limited to: cheating, fabrication, plagiarism, and facilitating dishonesty. Appropriate sanctions may be imposed on any student who has committed an act of academic dishonesty. Instructors should take reasonable steps to address academic misconduct. Any person who has reason to believe that a student has committed academic dishonesty should bring such information to the attention of the appropriate course instructor as soon as possible. Instances of academic dishonesty not related to a specific course should be brought to the attention of the appropriate department Head or Chair. Since students are expected to be familiar with this policy and the commonly accepted standards of academic integrity, ignorance of such standards is not normally sufficient evidence of lack of intent http://umass.edu/dean_students/codeofconduct/acadhonesty/.

Accommodation Statement

The University of Massachusetts Amherst is committed to providing an equal educational opportunity for all students. If you have a documented physical, psychological, or learning disability on file with Disability Services (DS), you may be eligible for reasonable academic accommodations to help you succeed in this course. If you have a documented disability that requires an accommodation, please notify me within the first two weeks of the semester so that we may make appropriate arrangements. For further information, please visit Disability Services (https://www.umass.edu/disability/)

Title IX Statement

In accordance with Title IX of the Education Amendments of 1972 that prohibits gender-based discrimination in educational settings that receive federal funds, the University of Massachusetts Amherst is committed to providing a safe learning environment for all students, free from all forms of discrimination, including sexual assault, sexual harassment, domestic violence, dating violence, stalking, and retaliation. This includes interactions in person or online through digital platforms and social media. Title IX also protects against discrimination on the basis of pregnancy, childbirth, false pregnancy, miscarriage, abortion, or related conditions, including recovery. There are resources here on campus to support you. A summary of the available Title IX resources (confidential and non-confidential) can be found at the following link: https://www.umass.edu/titleix/resources. You do not need to make a formal report to access them. If you need immediate support, you are not alone. Free and confidential support is available 24 hours a day / 7 days a week / 365 days a year at the SASA Hotline 413-545-0800.

Course Schedule

Week 1 (Feb 3 and 5): Introduction

- **Required**:
  - [**ISL**] Ch. 1 Introduction and Ch. 2 Statistical Learning

Week 2 (Feb 10 and 12): Supervised Learning - Regression

- **Required:**
  - [**ISL**] Ch. 3 Linear Regression
- **Recommended:**
  - Lones, Michael A. 2021. "How to avoid machine learning pitfalls: a guide for academic researchers." *arXiv preprint arXiv:2108.02497* (https://arxiv.org/abs/2108.02497)

★ **Assignment 1** due on Monday, February 17.

Week 3 (Feb 19 and 20): Supervised Learning - Classification

—Feb 17, Monday, is a holiday (Presidents' Day). Feb 20, Thursday, follows Monday's schedule as per UMass academic calendar. Therefore, our classes this week will be on Feb 19, Wednesday, and Feb 20, Thursday.—

- **Required:**
  - [**ISL**] Ch. 4 Classification
- **Recommended:**
  - Chenoweth, Erica, and Jay Ulfelder. 2017. "Can structural conditions explain the onset of nonviolent uprisings?" *Journal of Conflict Resolution* 61 (2): 298–324

★ **Assignment 2** due on Monday, February 24.

Week 4 (Feb 24 and 26): Resampling Methods

- **Required:**
  - [**ISL**] Ch. 5 Resampling Methods

★ **Assignment 3** due on Monday, March 3.

Week 5 (Mar 3 and 5): Unsupervised Learning I: Principal Components Analysis

- **Required:**
  - [**ISL**] Ch 12.1 The Challenge of Unsupervised Learning, 12.2 Principal Components Analysis, Lab: 12.5.1
- **Recommended:**
  - Michaud, Kristy EH, Juliet E Carlisle, and Eric RAN Smith. 2009. "The relationship between cultural values and political ideology, and the role of political knowledge." *Political Psychology* 30 (1): 27–42

Week 6 (Mar 10 and 12): Unsupervised Learning II: Clustering Approaches

- **Required:**
  - [**ISL**] Ch 12.3 Missing Values and Matrix Completion , 12.4 Clustering Methods, Lab: 12.5.2,3,4
- **Recommended:**
  - Harris, Rebecca C. 2015. "State responses to biotechnology: Legislative action and policy-making in the US, 1990–2010." *Politics and the Life Sciences* 34 (1): 1–27
  - Balakrishnan, Krishnachandran, and Shriya Anand. 2015. "Sub-cities of Bengaluru: Urban heterogeneity through empirical typologies." *Economic and Political Weekly,* 63–72

★ **Assignment 4** due on Saturday, March 15.

Spring Recess March 16 - March 23

Week 7 (Mar 24 and 26): Model Selection and Regularization

- **Required:**
  - [**ISL**] Ch. 6 Linear Model Selection and Regularization
- **Recommended:**
  - Mellon, Jonathan, and Christopher Prosser. 2024. "Regularized Regression Can Reintroduce Backdoor Confounding: The Case of Mass Polarization." *American Political Science Review,* 1–9

★ **Assignment 5** due on Monday, March 31.

Last day to Drop with "DR" is April 3, Thursday.

Week 8 (Mar 31 and April 2): Moving Beyond Linearity

- **Required:**
  - [**ISL**] Ch. 7 Moving Beyond Linearity

★ **Assignment 6** due on Monday, April 7.

Week 9 (April 7 and 9): Decision Trees & Ensembles (Random Forests, Bagging, Boosting)

- **Required:**
  - [**ISL**] Ch. 8 Tree-Based Methods
- **Recommended:**
  - Streeter, Shea. 2019. "Lethal force in black and white: Assessing racial disparities in the circumstances of police killings." *The Journal of Politics* 81 (3): 1124–1132

★ **Assignment 7** due on Monday, April 14.

Week 10 (April 14 and 16): Support Vector Machines

- **Required:**
  - [**ISL**] Ch. 9 Support Vector Machines
- **Recommended:**
  - Pan, Jennifer. 2019. "How Chinese officials use the Internet to construct their public image." *Political Science Research and Methods* 7 (2): 197–213
  - Hindman, Matthew. 2015. "Building better models: Prediction, replication, and machine learning in the social sciences." *The Annals of the American Academy of Political and Social Science* 659 (1): 48–62

★ **Assignment 8** due on Monday, April 21.

Week 11 (April 18 and 23): Deep Learning

—April 21, Monday, is a holiday (Patriot's Day). April 18, Friday, will follow Monday's class schedule as per UMass academic calendar. So, even though April 18 technically is part of "last" week, we will consider it part of "this" week (Week 11) for course content purposes.—

- **Required:**
  - [**ISL**] Ch. 10 Deep Learning

  Recommended:
  - Zhao, Qingyuan, and Trevor Hastie. 2021. "Causal interpretations of black-box models." *Journal of Business & Economic Statistics* 39 (1): 272–281
  - Davidson, Thomas. 2019. "Black-box models and sociological explanations: Predicting high school grade point average using neural networks." *Socius* 5:2378023118817702

Week 12 (April 28 and 30): Ethics

- **Required:**
  - Rudin, Cynthia. 2019. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature machine intelligence* 1 (5): 206–215
  - Corbett-Davies, Sam, Johann D Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. 2023. "The measure and mismeasure of fairness." *The Journal of Machine Learning Research* 24 (1): 14730–14846
  - https://www.vox.com/future-perfect/22916602/ai-bias-fairness-tradeoffs-artificial-intelligence
  - https://www.aclu.org/news/privacy-technology/wrongfully-arrested-because-face-recognition-cant-tell-black-people-apart

Week 13 (May 5 and 7): Catch-up & Review

★ **Final Project** due on Monday, May 12.