MACS 30135: Interpretable and Explainable Machine Learning - from Prediction to Knowledge

Winter 2025

Taught by: Fabricio Vasselai (vasselai@uchicago.edu)
Classoom: room 295, at the MACSS building (1155 E. 60th Street).
Office Hours: by appointment; you can sign up through this appointment page.
Office Hours location: room 213-A, at the MACSS building (1155 E. 60th Street).

Course Overview

ML is already ubiquitous, revolutionizing our lives in all domains. Precisely because of this, its new frontier is occupied with finding ways to ensure that we truly understand and exert control over how our techniques predict what they predict, and generate the results that they generate. This movement beyond being satisfied with just the accuracy of results is critical for a couple of reasons. First, as the wise say, with great power comes great responsibility - since our revolutionary ML techniques only learn from what we feed them, they reproduce (and even exacerbate) biases present in our data. To counter this, we need to be able to interpret and then intervene in how Machine Learning learners are learning what they learn. Second, by being able to interpret, explain and modify how algorithms learn what they learn from our datasets (not merely how algorithms mechanically work), we can uncover relevant latent knowledge hidden in such data. In this course, we start by discussing what Interpretable or Explainable ML even means to different audiences. Next, students will be introduced to several types of state-of-the-art techniques that have been proposed to increase the interpretability/explainability of ML models. Along the way, students will be presented to the discussion on how interpretable or explainable models can help fight biases, improve the fairness, trustworthiness or reliability of predictions, and ensure the ethical use of ML as it continues to change our lives.

Prerequisites

You should only take this course if you fulfill **all** the following prerequisites:

- Prior (introductory) exposure to linear algebra and probability;
- To have taken a graduate-level class specific on Machine Learning and which covered Supervised Machine Learning (that is, a class that covers only Unsupervised Machine Learning, or a class on general AI methods, will not suffice).Evaluate the ethical implications of digital research in the social sciences;
- experience programming in Python.

Course Structure

Class sessions will consist of a mix of lecture, discussion and sometimes going over Python code.

All course content and readings are organized on Canvas under “Modules” on the left-hand side of your screen. You should read all of the readings listed for a given class session ***ahead of class time*** and be prepared for in-class discussion.

While there is no one actual *textbook* for this course, we will often read and discuss parts of a canonical book that happens to be freely available online (made available by the author):

Molnar, Christoph (2022). *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable*. 2nd Edition, Munich - Germany.

Other non-free great books that I will use to prepare classes, but which will either not be required readings or will have required parts scanned, include (in alphabetical order):

Holzinger, Andreas et. al. (2020). *xxAI - Beyond Explainable AI*. Springer, 1st Edition.

Masís, Serg (2023). *Interpretable Machine Learning with Python*. Packt Publishing, Birmingham - UK, 2nd Edition.

Molnar, Christoph (2023). *Introduction To Conformal Prediction With Python*. 1st Edition, Munich - Germany.

Samek, Wojciech et. al. (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 1st Edition.

Besides those, every week there will be papers to be read before coming to class.

Evaluation

DataCamp courses ***(20%)***, *due* ***always before class***.

First Graded Exercise ***(20%)***, *due by* ***02/08/2025, 11:59on CST***.

Second Graded Exercise ***(20%)***, *due by* ***03/01/2025, 11:59on CST***.

Final Project ***(30%)***, *due by* ***03/14/2025, 11:59on CST***.

Participation ***(10%)***.

DataCamp courses

In some weeks, you will have to complete DataCamp courses / course chapters online, that I will assign to all. These must be completed always before class time. They will not be graded for accuracy, just for completion.

First and Second Graded Exercises

There will be 2 graded exercises; in each, I will ask to perform a couple of programming tasks and write a short report.

Final Project

In your final project, you should write a short paper (between 3000 and 4000 words) where you choose 2 techniques (from a list I will provide) learned in this class, and use them to (re)analyze datasets of your interest, to advance your own research or research interests.

Participation

A big part of this class involves in-class discussion of the concepts and challenges discussed in class, as well as of the assigned readings.

Plagiarism and Academic Honesty

- Any plagiarism (of existing work, of your colleagues or even of your past work) will result in a zero grade and will be reported to MACSS and to the University authorities.
- Any copying of others' work will result in a zero grade and will be reported to MACSS and to the University authorities.
- Unless otherwise explicitly stated in my assignment instructions, absolutely all usage of ChatGPT or of other LLMs tools is strictly prohibited. Again, if you use any, you are risking being reported.
- When in doubt, please ask! It is far better to check with us prior to submitting an assignment than waiting.

Submitting Late Assignments

This course has a **very** strict lateness policy for assignment submission:

- Considering their online nature, their short length and their purpose of having you practice code before class, DataCamp courses can never be completed late, under any circumstances.
- In the case of the First and Second graded exercises, assignments submitted:
    - up to 6 hours late will have no penalty;
    - more than 6 but less than 24 hours late will have a -5 points penalty (out of 100);
    - more than 24 but less than 48 hours late will have a -10 points penalty(out of 100);
    - more than 48 but less than 72 hours late will have a -15 points penalty (out of 100);
    - more than 72 but less than 96 hours late will have a -20 points penalty (out of 100);
    - assignments submitted more than 96 hours late **will not be accepted**.
- Given the strict university deadlines for submitting students' final grades, the Final project **cannot be late at all**.

Statement on diversity, inclusion, and disability

The University of Chicago is committed to diversity and rigorous inquiry from multiple perspectives. The MAPSS, CIR, and Computation programs share this commitment and seek to foster productive learning environments based upon inclusion, open communication, and mutual respect for a diverse range of identities, experiences, and positions.

The University of Chicago is committed to ensuring equitable access to our academic programs and services. Students with disabilities who have been approved for the use of academic accommodations by Student Disability Services (SDS) and need a reasonable accommodation(s) to participate fully in this

course should follow the procedures established by SDS for using accommodations. Timely notifications are required in order to ensure that your accommodations can be implemented. Please meet with your instructor to discuss your access needs in this class after you have completed the SDS procedures for requesting accommodations.

- Email: disabilities@uchicago.edu
- Phone: 773-702-6000

Course Schedule

Note: *Schedule is subject to change. Check on Canvas for updates as the course progresses.*

Week	Date	Topics
1	**Jan 09**	Concepts + interpretability challenges
2	**Jan 16**	Intrinsically interpretable methods
3	**Jan 23**	Global Model Agnostic explainability methods
4	**Jan 30**	Local Model Agnostic explainability methods
5	**Feb 06**	Using constraints for interpretability
Due Feb 08		*First graded exercise*
6	**Feb 13**	Counterfactuals and sensitivity analysis
7	**Feb 20**	Interpretable DNNs
8	**Feb 27**	Bias mitigation methods for ML
Due Mar 01		*Second graded exercise*
9	**Mar 06**	Error-bounds estimation methods for ML
Due Mar 14		*Final project*