CS 59000 / STATS 59800-18 – Advanced Machine Learning (Causal Inference)

Time: MW 4:30pm-5:45pm

Location: Felix Haas Hall G066

Instructor: Elias Bareinboim

Webpage: http://www.cs.purdue.edu/~eb/

Summary

Imagine you run your favorite machine learning algorithm and obtain a strong correlation between two variables, say X and Y. Some weeks later, you implement a new policy that increases the value of X. Surprisingly, nothing happens with the value of Y. Was your first algorithm incorrect? Do you need to collect more data for allowing the learning process to capture the phenomena under scrutiny? Should you use another algorithm? Those are challenging questions that appear in everyday data analysis. This course will try to address this and other common questions that relate different types of data collections and ways for explaining the data, which usually comes under the general name of *causal inference*.

The emergence of causal inference in data-driven fields does not come as a surprise since the interest in knowing that X is (probabilistically) correlated with Y is not rarely devoid of meaningful interpretation or practical implications. There are plenty of real examples demonstrating that correlation does not imply causation, hence not suitable to substantiate causal claims and principled decision making. For instance, no one expects that forcing the rooster to crow in the middle of the night will make the sun rise. Still, these two events are strongly correlated. Likewise, current data suggests that there is a strong correlation between Facebook's growth and the Greek crisis, while no one believes in a conspiracy of Zuckerberg against Greece.

In this course, we will introduce concepts, principles, and algorithms necessary to solve modern, large-scale problems in scientific inference. Emphasis will be given to the tradeoff between assumptions (delineated by current scientific knowledge) and conclusions for standard types of queries, including associational, causal, and counterfactual. In other words, we will consider the problem of providing different types of "explanation" for the vast amount of data collected in different fields of human inquiry, including engineering, medicine, and the empirical sciences. This problem lies at the heart of current discussions in machine learning and statistics.

Prerequisites

In order to be successful in this course, you should have a basic knowledge of:

- Discrete Math (proof techniques, search algorithms, and graph theory)
- Calculus (find min/max of functions)
- Statistics (basic probability, modeling, experimental design)

- Machine Learning (graphical models)
- Some programming experience (with special understanding of complexity analysis)

Material (tentative)

The following is a rough outline of the material we will cover.

Week	Subject	Material	References
1-3	Introduction / Review	Motivation, 3-layer scientific hierarchy, review probability and graphical models (graphoids, bayesian networks, d-separation, i-maps).	PGMs, Chs. 1-3
4	Causal Bayesian Networks	CBNs, functional Models, 3-layers	Causality, Ch. 1
5	Structural Learning	Constraint-based and score-based methods.	TBA
6-7	Causal Inference (The Data-fusion Problem)	Confounding bias, Simpson's paradox, Back-door criterion, Front-door criterion, Do-calculus, Sampling selection bias, Transportability.	TBA
8	Midterm exam		
9	Linear Models	Associational versus structural coefficients. Inference in linear systems.	TBA
10	Bounding	Bounding ATE and ETT.	TBA
11-12	Counterfactual Inference	Axioms of causal inference. Mediation analysis.	TBA
13	Actual Causation		TBA
14-16	Projects' presentations		

References

We will use material selected from different sources, including chapters of the following books:

- Probabilistic Graphical Models: Principles and Techniques (Koller and Friedman, 2009), MIT Press
- Causality, (Pearl, 2000), Cambridge Press
- Causal Inference: A Primer, (Pearl, 2016), Cambridge Press

People

Name	**Email**	**Office Hours**	**Location**
Elias Bareinboim	eb@purdue.edu	Wed 11:30-12:30pm	LWSN 2142L
Daniel Kumor (TA)	dkumor@purdue.edu	TBA	LWSN B116
Junzhe (Justin) Zhang (TA)	junzhez@junzhez.com	TBA	LWSN B116

If you have questions about lectures or material, please ask at Piazza.

Grading

The course will have both a theoretical and practical components. The project will allow students to chose between these two tracks.

- Midterm exam: 25%
- Written notes: 25%
- Homeworks: 25%
    - About 4 homework assignments
- Project: 50%
- Final Grade = Project + $\max_1$(MD, WN, HW) + $\max_2$(MD, WN, HW)
- Attendance and participation: bonus

Please review the Purdue honor code. While working on assignments in small teams is okay, your homework solutions must be your own.

Course Policies

You are expected to attend lectures and participate in class. While taking notes on laptops and *small* snacks are allowed, please make sure you are quiet and respectful of those around you, including those behind you who might be distracted by your snacks/devices.

You are expected to come prepared to class, and to participate in class discussion, *especially* when something is not clear. If you are too shy to ask in class, please post on piazza or attend office hours.