JOHNS HOPKINS
UNIVERSITY

Machine Learning: Data to Models (EN.601.476/676)

Instructor: Rohit Battacharya (rbhattacharya@jhu.edu)
Assistants: Jaron Lee (jaron.lee@jhu.edu)
Baichuan Jiang (baichuan@jhu.edu)

Course Meeting Times And Links

Meeting Times: MW 3pm – 4:15pm
Piazza: piazza.com/jhu/spring2021/en601476676 (copy and paste link into your browser)
Zoom: https://wse.zoom.us/j/97301275462?pwd=NnRtcFFudStYTFN5TjA3YTQ4YWd3QT09
Zoom passcode: mldata
Gradescope entry code: 5VKPEN

Course Description

Graphical models offer compact visual representations that bridge the gap between statistical jargon and domain scientific knowledge. This makes them very powerful for performing data analyses in fields like genomics, health/social science, and economics, where communication between the analyst and domain experts is crucial. In this course, we will cover statistical models and algorithms for undirected graphs, directed acyclic graphs, chain graphs, and acyclic directed mixed graphs. We will study these models and their usage through bi-weekly case studies demonstrating the use of graphical reasoning for various inferential tasks, including detecting and quantifying bias in scientific studies. The final project will be "hands on," where students will apply techniques learned in class to derive data-driven insights into a scientific question of their choosing, and write up their results.

Course Objectives

At the end of this course you should be able to

- Make judicious choices about the model class appropriate to a given problem
- Understand the semantics and limitations of different model choices
- Perform inference and learning tasks for multiple model classes
- Analyze real data using probabilistic graphical models

Textbooks

There are no required textbooks for this course. Here are some textbooks for supplementing your understanding if you wish to dig deeper.

- *Probabilistic Graphical Models* by Daphne Koller and Nir Friedman (2009)
- *Graphical Models* by Steffen L. Lauritzen (1996)
- *Machine Learning: A Probabilistic Perspective* by Kevin Murphy (2012)

Grading Overview

This course focuses more on learning rather than assessment. In the long run, the knowledge and skills you acquire are far more important than your final grade. The grading policy and guidelines are made to reflect this course philosophy.

- Homework: HW 1: 6% and HW 2-4: 39% ($3 \times 13\%$)
- Project proposal: 5%
- Peer review: 5% (details to be announced in class)
- Final project: 45%

Course Schedule

This is a rough schedule, and subject to change.

- Week 1: Jan 25, 27
  - Course overview and introduction.
  - Review of probability and statistics.
  - Intro to directed acyclic graphical models a.k.a Bayesian networks.
  - **HW 1 is assigned Jan 27.**
- Week 2: Feb 1, 3
  - DAG models continued – d-separation and parameter estimation.
  - Case study: hormone replacement therapy and endometrial cancer.
  - **HW 1 is due Feb 3, 11:59pm EST.**
  - **HW 2 is assigned Feb 3.**
- Week 3: Feb 8, 10
  - Undirected graphical models a.k.a Markov random fields.

  - Links between UG and DAG models – u-separation and augmentation criterion.

- Week 4: Feb 15, 17
  - Chain graphical models.
  - Case study: Diffusion and equilibriation of opinions in social networks.
  - **HW 2 is due Feb 17, 11:59pm EST**.
  - **HW 3 is assigned Feb 17**.

- Week 5: Feb 22, 24
  - Acyclic directed mixed graphical models.
  - The "triforce hierarchy" of graphical models, and intuition via Gaussian models.

- Week 6: Mar 1, 3
  - Beyond conditional independence – generalized equality constraints.
  - Case study: testing a causal null hypothesis via generalized equality constraints.
  - **HW 3 is due Mar 3, 11:59 pm EST**.
  - **HW 4 is assigned Mar 3**.

- Week 7: Mar 8, 10
  - Introduction to structure learning.
  - Constraint-based and score-based learning for DAGs.

- Week 8: Mar 15, 17
  - Constraint-based learning for ancestral ADMGs.
  - Continuous optimization schemes for structure learning.
  - Case study: inferring gene regulatory networks from data.
  - **HW 4 is due Mar 17 11:59pm, EST.**

- Week 9: Mar 24 (Mar 22 is part of the scattered "Spring break")
  - Introduction to messy data: missing data, selection bias, and generalization.
  - **Final project proposals due Mar 24 11:59pm, EST.**

- Week 10: Mar 29, 31
  - Dealing with missing data.
  - Dealing with selection bias.

- Week 11: Apr 5, 7

  - Generalization and transportability of machine learning models.
  - Being a good citizen of academia – scientific writing and peer review.
- Week 12: Apr 12 (Apr 14 is part of the scattered "Spring break")
  - Causal inference – Simpson's paradox and backdoor adjustment.
  - Causal inference – beyond backdoor adjustment.
  - **Final projects due for "peer review" Apr 16 11:59pm, EST.**
- Week 13: Apr 19, 21
  - Case study: quantifying the causal effect of smoking on developing lung cancer.
  - Algorithmic fairness.
- Week 14: Apr 26, 28
  - Discussion of final project questions.
  - Special topics and course wrap-up.
  - **Peer review due Apr 29, 11:59pm EST.**
- Finals week: **Final project is due the day our exam is scheduled by registrar**.

Homework And Final Project Schedule

There will be 4 homeworks in total, the first of which is shorter and meant to familiarize you with Gradescope. You will be given one week to complete HW 1 and roughly two weeks to complete each subsequent homework assignment. Homeworks are due on the date assigned in the above schedule. Homework assignments must be completed individually, not in groups. For the final project, we are considering allowing students to work individually or in pairs. More details and expectations for the final project will be released as the course progresses.

Late Homework Policy

Homeworks submitted 0-24 hours late will be penalized 10%, 24-48 hours late by 20%, 48-72 hours late by 30%, and later than 72 hours by 100%. Students are granted **4 late days** to be used at their discretion to turn in late assignments without incurring any penalty. Late days can be used to postpone any assignment deadline **except** those related to submitting the final project. This exception is made to allow enough time for others to peer review your final project, and prevent any delay in entering the final grades on SIS (for which there exists a hard deadline from the university.)

If there are extenuating circumstances, such as a family or medical emergency, that prevent you from completing an assignment on time (even after having used all four late days), please let us know as soon as possible – we will do our best to accommodate your request.

Programming Language

Homework assignments will require some programming, and the final project will require real data analysis which may either require novel programming or the use of available software. All homework assignments must be completed in Python. You may use whatever programming languages and software you like for the final project.

On Group/Individual Work

All assignments in this course are individual, not group, assignments. You may freely discuss homework assignments with your fellow classmates. The final solutions, however, must be written entirely on your own. This includes programming: you must implement any programming task on your own. Copying someone else's code (and then subsequently making minor changes) constitutes plagiarism. So, if you need to discuss programming assignments, you may discuss general strategy but should write the code by yourself.

Students With Disabilities

Any student with a disability who may need accommodations in this class must obtain an accommodation letter from Student Disability Services, 385 Garland, (410) 516–4720, `studentdisabilityservices@jhu.edu`.

Ethics

The strength of the university depends on academic and personal integrity. In this course, you must be honest and truthful, abiding by the *Computer Science Academic Integrity Policy*:

Cheating is wrong. Cheating hurts our community by undermining academic integrity, creating mistrust, and fostering unfair competition. The university will punish cheaters with failure on an assignment, failure in a course, permanent transcript notation, suspension, and/or expulsion. Offenses may be reported to medical, law or other professional or graduate schools when a cheater applies. Violations can include cheating on exams, plagiarism, reuse of assignments without permission, improper use of the Internet and electronic devices, unauthorized collaboration, alteration of graded assignments, forgery and falsification, lying, facilitating academic dis- honesty, and unfair competition. Ignorance of these rules is not an excuse. Academic honesty is required in all work you submit to be graded. Except where the instructor specifies group work, you must solve all homework and programming assignments without the help of others. For example, you must not look at anyone else's solutions (including program code) to your homework problems. However, you may discuss assignment specifications (not solutions) with others to be sure you understand what is required by the assignment. If your instructor permits using fragments of source code from outside sources, such as your textbook or on-line resources, you must properly cite the source. Not citing it constitutes plagiarism. Similarly, your group projects must list everyone who participated. Falsifying program output or results is prohibited. Your instructor is free to override parts of this policy for particular

assignments. To protect yourself: (1) Ask the instructor if you are not sure what is permissible. (2) Seek help from the instructor, TA or CAs, as you are always encouraged to do, rather than from other students. (3) Cite any questionable sources of help you may have received. On every exam, you will sign the following pledge: "I agree to complete this exam without unauthorized assistance from any person, materials or device. [Signed and dated]". Your course instructors will let you know where to find copies of old exams, if they are available. Please report any violations you witness to the instructor.

You can find more information about university misconduct policies on the web at these urls:

- Undergraduates: `https://studentaffairs.jhu.edu/policies-guidelines/undergrad-ethics/`
- Graduate students: `http://e-catalog.jhu.edu/grad-students/graduate-specific-policies/`