STAT 339: PROBABILISTIC MODELING AND MACHINE LEARNING (SPRING 2020)

REVISED TO REFLECT "VIRTUALIZATION" IN MODULE 2

Times. MWF 10-10:50am EDT
Locations. King 243 (Module 1); Zoom (Module 2)
Instructor. Colin Reimer Dawson (*he/him/his*)
Email. `cdawson@oberlin.edu`
Website. `http://colindawson.net/stat339/`

Course Description. This course is an introduction to the field of machine learning from the particular perspective of probabilistic reasoning. We will discuss fundamental principles of machine learning and probabilistic reasoning, as well as specific models and algorithms used to do classification, prediction and clustering, hidden variable modeling, and sequence learning.

Rather than cover as many different models and algorithms as possible, we will focus on general principles of model-building and probabilistic reasoning as illustrated through a relatively small set of common models and algorithms.

The overarching goal of the course is for you to strengthen your ability to reason probabilistically, and to equip you to learn about and/or develop new models and algorithms to solve problems.

Applications will be selected from a variety of domains.

Learning Goals. By the end of this course, you should be able to:

- Explain the difference between supervised and unsupervised learning, and between generative vs discriminative methods
- Understand the properties, relative strengths, and limitations of some of the most common probabilistic machine learning models used for classification, regression, clustering and segmentation

- Represent probabilistic models using directed graphs, and use these graphs to understand under what conditions two variables in the model are dependent or independent
- Implement and test algorithms to update the parameters of these models using real and synthetic data in a common computer language equipped with standard scientific computing libraries
- Choose appropriate models and machine learning algorithms to solve real problems using real data from a domain of interest

Who This Course is For. The course is designed to appeal to students interested in modern data science from a variety of backgrounds who have a foundation in basic programming and calculus.

This likely includes (among others!) math majors with statistics interest, computer science majors with data science / artificial intelligence interest, natural / social science majors (e.g., psychology, neuroscience, economics, biology, physics) with a methodological/statistical bent, and humanities majors who are interested in quantitative methods (for, e.g., text analysis).

Note that this course is quite different in its approach from the other machine learning course offered at the college (CSCI 374), and it will not be redundant to take both.

Prerequisites. The main prerequisites for the class are familiarity with partial derivatives and integrals of functions of multiple variables (MATH 231), and familiarity with basic programming concepts, such as variables, functions, arguments, loops, and arrays, at the level of CSCI 150 or similar (see the note on Software below).

In terms of additional mathematics, we will rely heavily on probability, and at times linear algebra, but I am not expecting you to have taken courses in these subjects; we will go through the main ideas that we will need.

There is no longer a statistics prerequisite for this class, although previous experience with statistical thinking is certainly helpful.

If you are unsure whether your background is suitable for this course, don't hesitate to talk to me.

Materials

Main Textbook. The main textbook is

- *A First Course in Machine Learning* (2nd edition), by Simon Rogers and Mark Girolami.

You do need the 2nd edition, as it includes several new chapters not in the first edition. We will aim to cover nearly the entire book.

Supplementary Texts. Supplementary readings will come from the following sources, all of which are freely available in pdf online from the authors' websites:

- *Bayesian Reasoning and Machine Learning*, by David Barber (`http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/171216.pdf`)
- Chapter 1 of *Machine Learning: A Probabilistic Perspective* by Kevin Murphy (`https://www.cs.ubc.ca/~murphyk/MLbook/pml-intro-22may12.pdf`)

Most topics in the course are covered by both the Rogers and Girolami (FCML) and the Barber (BRML) books. In most cases, the treatment in BRML is more advanced and technical, and should be considered optional reading if you are interested in going into more detail; however, there are a few places where FCML does not cover a topic at all.

Software. You may use any programming language(s) you like to write code; scientific computing languages with good support for matrix computations and plotting capabilities are strongly recommended. If possible, I recommend Python, with the support of the `matplotlib` library for plotting and the `numpy` and `scipy` libraries for numerical computation. I do not recommend using a compiled language such as Java or C/C++, as the increased time it takes to write code in these languages will outweigh any performance savings you can accrue.

The first (optional) homework assignment is a chance for those whose coding skills are minimal or rusty to get a refresher, and also for those transitioning to using Python with its scientific computing and plotting libraries `numpy` and `matplotlib` to learn some basics.

Miscellany

Communicating With Me. I have created a Slack workspace for the class, at `stat339s2020.slack.com`, which you have received an invitation to join. This workspace serves as a convenient way to organize asynchronous communication between you, me, and your peers. When contacting me outside class time or office hours, sending me a PM on Slack is the best way to ensure a prompt response. I generally try to respond to Slack messages the same day if they are sent before 8pm, but it may sometimes be the following day (especially on Thursdays and Saturdays). Email may be somewhat slower than that, but is preferable for less time-sensitive correspondence. If you don't receive a reply from me within 24 hours for a Slack message, or 72 hours for an email, don't hesitate to follow up as it may have slipped through the cracks.

Accommodations. If you require accommodations of any kind in order for you to do your best work in this class, please let me know as early as possible, and consult as well with the Office of Disability Services (ODS). By college policy, **all requests for accommodation require documentation from ODS.**

Edit: With the transition to a virtual format for the second half of the semester, I am cognizant of the fact that many of us may be trying to accomplish our work under circumstances that may be significantly less conducive to academic study than our usual on campus community. Please do not hesitate to talk to me if you have needs that may require additional flexibility.

Honor Code. The Oberlin College Honor Code formalizes the idea that all work that you submit is your own and that you have given credit to the ideas and work of others when you incorporate them. You will be asked to write and sign the honor pledge on each graded assignment that you hand in. The honor pledge reads: "I have adhered to the Honor Code in this assignment."

What it means to adhere to the honor code depends on context. I describe what it means to follow the honor for each assignment type below.

I take the Honor Code extremely seriously. This means two things: First, I presume a relationship of mutual trust will not try to police your behavior. However, if I do have reason to believe that this trust has been betrayed, I treat this as a very serious matter, and will not hesitate to involve the Honor Code committee in the matter. I will do my best to remind you of the concrete expectations for each assignment as they come up, but it is ultimately your responsibility to make sure you understand them.

More information about the honor code can be found on the web at the Dean of Students site:

`https://www.oberlin.edu/dean-of-students/student-conduct/academic-integrity/students`

Graded Work (REVISED)

With the virtualization of the second half of the course and the increased challenge associated with collaborative work, **I have increased the weight of the daily reflections during the second half, as it becomes even more important to stay engaged with the content outside regular class meetings, and alternatives to a final project** to add flexibility:

Option 1: Problem Set Focused.

- Daily reflections via Slack during Module 1: 5%
- Daily reflections via Slack during Module 2: 15%
- Exam 1: 20%
- Eight problem sets: 60%, weighted as follows
  - First four (pre-break): 5% each
  - Last four (post-break):
    - Highest one: 20%
    - Next two: 10% each
    - Next one: 5%
  - The lowest grade across all eight problem sets will be dropped, resulting in a total weight of 60%

Option 2: Exam or Project Focused.

- Daily reflections via Slack during Module 1: 5%
- Daily reflections via Slack during Module 2: 15%
- Exam 1: 20%
- Seven problem sets: 40%, weighted as follows
  - Highest two: 10% each
  - Next four: 5% each
  - Lowest: Dropped

- Exam 2 or Final Project: 20%

Problem Sets (40-60%). Problem sets consist of a mix of written (conceptual) questions, mathematical calculations and derivations, and computer implementation. In some cases I will provide you with partial code in Python; if you choose to use a different language to complete your assignment, you can either translate this scaffolding code, or find a library that does the same thing.

Note: The problem sets will require *significant* time to complete; expect to spend around 10 hours per week on them, though you may be able to do them in less time if your background exceeds the minimum prerequisites for the course. In any case, you will need to space out your work over the course of the week.

I encourage you to work in pairs or threes on the problem sets. Each group should turn in one copy of the assignment.

I will most likely not be able to grade every problem on every assignment even with paired work, but will "spot check" them, and grade and give detailed feedback on one or two problems.

Honor Code: You may freely discuss homework assignments with each other, whether or not you are working together; however **each pair must turn in their own original code and writeups; you may not copy or "paraphrase" anyone else's work.** In addition, it is expected that both/all members of a group contribute substantially to all pieces of an assignment, namely, the math, coding, and writing for each problem. **The intent is *not* for you to split the assignment in half, each doing part of it, but rather to discuss and collaborate with each other as you work on it**.

Participation and Engagement (20%).

Module 1 (In Person) (5%). You are expected to attend all classes except when some urgent circumstance prevents you from doing so. After each class I will ask you to write a short (∼2 sentence) reflection on that day's material (both reading and lecture) via a Slack message. These are intended to be very much open-ended, and can consist of questions, comments, observations, or a mix of these. The idea is just to get your mind to revisit and consolidate the ideas.

Reflections are due by midnight on each class day. You can miss up to three reflections without affecting your grade. Reflections submitted after the deadline but before the following class receive half credit. If you miss a class, indicate that in your message.

Up to three absences will be considered "excused" (i.e., not count toward your three "free" missed reflections) if you submit a reflection on the reading for that day.

Honor Code: You must accurately represent your attendance or non-attendance, and may not submit a reflection on anyone else's behalf.

Module 2 (Virtual Classes via Zoom) (15%). We will continue to meet via Zoom at the regularly scheduled times (MWF, 10-11am EDT), however the class format and nature of the Slack assignments will change. I will post video lectures at least 24 hours before each Zoom class meeting (likely broken into multiple segments) covering the scheduled content for that day. **You are expected to view these and write down questions and comments (include slide numbers and timestamps wherever applicable) prior to the start of each class, and post these to Slack.** We will then spend the class period on discussion, clarifications, additional examples, and in "breakout sessions" where you can work together on homework problems (I will virtually "circulate" around the breakout rooms)

Exam (20%). There will be one required written exam covering the foundational material through Fundamentals of Bayesian Inference and Belief Networks. Due to students having to make travel plans on short notice when this exam was originally scheduled to occur, **this has been rescheduled to be handed out on Friday 4/10 and handed in by Wednesday 4/15**.

The questions will consist of a combination of conceptual and mathematical components, but will not involve coding.

Honor Code: The exam must be done individually, without consulting any other individuals apart from me, but you may consult your notes, or other course materials I provide.

End of Semester Options (20%). The original plan was that everyone would do a project in small groups of 2 or 3; however, due to the increased difficulty associated with group work in a virtual class setting, **I have defined two alternative assessment options that can replace a project if desired, though a project remains an option (albeit slightly revised to reflect changes in planned course content).**

Project Option (Revised). If you choose the final project option, you will work in a small group of your choosing (2-3 people), defining your own classification, clustering, and/or prediction question(s), and implementing and comparing at least three different machine learning approaches to your problem. At least two of the three should be probabilistic methods covered in the course. The third may or may not be a probabilistic method and may (optionally) be a technique we did not discuss in the course.

Your writeup should be in the form of a conference-style paper as might be submitted to a machine learning conference such as the International Conference on Machine Learning (ICML), the Conference on Neural Information Processing Systems (NeurIPS), or Uncertainty in Artificial Intelligence (UAI).

These are typically 6-10 pages in length (single spaced), including figures, and are laid out as follows:

(1) An **introduction** section in which you describe your problem

(2) A **methods** or **models** section in which you describe the techniques you are using

(3) A **results** section, including graphical and numerical summaries of the performance of your chosen methods

(4) A **discussion** section, reflecting on the relative strengths and weaknesses of your chosen methods as revealed by your results and perhaps theoretical/practical considerations.

You will need to decide on ways to measure performance that supply "apples-to-apples" comparisons across methods.

Honor Code: Free collaboration with your team members is of course required. Describing the work that other people did before you is an important part of intellectual inquiry, and **you must give credit and cite sources for any data, code, or ideas that did not originate within your team**. This includes paraphrases as well as direct quotations. It is fine (and indeed you are encouraged) to call other people's open-source code in your own code, but you need to give attribution.

All members of the group must make approximately equal overall contributions to the project, though it is possible, for example, for one person to do more coding and another to do more writing, etc.

Problem Set Option (Revised). In lieu of a project you may choose instead to complete one extra problem set, which will be based on material we discuss in the last two weeks of the semester. If you choose this option, the 20% weight associated with the final project will be redistributed across **all problem sets in the second module** as described above.

Exam Option (Revised). Alternatively you can replace the project with a second exam, which will be in the same format as the first exam, but which will cover material not covered on the first exam.

Pass/No Pass Option. With the college extendeding the option to convert a traditional letter grade to a Pass/No Pass grade until the last day of the semester, I have defined concrete criteria that you may use to lock in a "pass" mark for the course in case you would like to reduce your workload.

Students who complete problem set 3 and one of problem sets 4 or 5, as well as keeping up with Slack posts through at least the unit on clustering (approximately Friday 4/24) **will be guaranteed a grade of "pass" if they have at least a "B" average on that material**.

Approximate Topic Outline

See the course website: `http://colindawson.net/stat339/schedule` for a tentative (and periodically updated) schedule of topics and associated readings (substantially revised in light of moving to a virtual class in Module 2).