36-708: The ABCDE of Statistical Methods for Machine Learning

Spring 2023 (Jan 17 to Apr 27), Syllabus

January 20, 2023

1 Basic Course Information

Instructor Aaditya Ramdas, aramdas@cmu.edu
.
TA: James Leiner, jleiner@cmu.edu

[Post-class chat: 3:20-3:40pm, in DH1211, TR]
[Coffee Chat: Wed 11-11:30am, on Zoom]
[Office hours: 2:30-3:30pm M]

Time: 2:00-3:20pm TR
Location: DH 1211
Exceptions: Apr 13 is a university holiday, Spring break is Mar 6-10, see the academic calendar.

Website See https://36708.github.io/ for basic course material.
Announcements All announcements will be made on the above course website and/or Canvas.

Participants This course can be credited by PhD students with appropriately strong mathematical background. The course can also be audited by anyone who is curious about the topic, provided they abide by course rules (eg: about attendance and use of electronic devices). Students who want to sit through the course must officially audit.

Prerequisites Enrolled students are expected to have completed (and mastered, done well in) at least one intermediate statistics course at the PhD level, and at least one course on either machine learning, or linear regression, or related topics. Students must be both mathematically and computationally mature. Specifically, all students should have taken Intermediate Statistics (36705), be proficient at programming in R and/or Python and/or Matlab, and be comfortable with linear algebra, probability, calculus and related topics (see resources below that you should be familiar with). Students who have taken 10701, 10715 or 10716 can still take this course, since there are likely to be many complementary and non-intersecting topics. Apart from the unique angle taken by this course, the smaller size of class will ensure more individual attention and instructor interaction, so attendance (especially for crediting) will be selective.

Textbook We will follow a mixture of (A)"Elements of Statistical Learning (2nd edition)" by Hastie, Tibshirani, Friedman, (B)"Foundations of Machine Learning (2nd edition)" by Mohri, Rostamizadeh and Talwalkar, and (C) "Introduction to Statistical Learing" by James, Witten, Hastie, Tibshirani. However, some material will be directly from well-written papers (eg: by Breiman) or other textbooks and will be linked as required.

2 Course Description

Course philosophy (ABCDE). This course focuses on statistical methods for machine learning, a decades-old topic in statistics that now has a life of its own, intersecting with many other fields. While the core focus of this course is methodology (algorithms), the course will have some amount of formalization and rigor (theory/derivation/proof), and some amount of interacting with data (simulated and real). However, the primary way in which this course complements related courses in other departments is the joint ABCDE focus on

(A) Algorithm design principles (eg: game theory, RKHSs, optimization),

(B) Bias-variance thinking (eg: tradeoffs, cross-validation, hyperparameter tuning),

(C) Computational considerations, conformal and calibration (uncertainy quantification)

(D) Data analysis (restricted to homeworks)

(E) Explainability and interpretability (eg: measures of variable or datapoint importance).

Non-technical blurb. In the instructor's opinion, (B) is the most important — every day, researchers come up with yet another new algorithm/model, scale it up by using distributed computing and stochastic optimization, and throw it at a big real dataset (A, C, D). However, in the era of big data, big bias and big variance is a big issue! Instead of producing just predictions, uncertainty quantification is critical for applications (how sure are we of these predictions?). Blindly throwing lots of data and complex black-box models at a problem might produce initially promising results, but the results may be highly variable and non-robust to minor changes in the data or tuning parameters. Importantly, more data does not eliminate bias — "obvious" bias caused by covariate shift or outliers, and "subtle" bias like selection bias, sample bias, confirmation bias, etc. Understanding the variety of different sources of bias and variance, and the effects they can have on the final outputs, is a critical component of using ML algorithms in practice, and will be a central theme of the course. Of course, (E) is also important and often underemphasized, and we will cover some recent methods for interpreting models such as measures for variable importance and/or data-point importance.

Technical blurb. The course will cover (some) classical and (some) modern methods in statistical machine learning; the field is so vast that the qualifier "some" is critical. These include unsupervised learning (dimensionality reduction, clustering, generative modeling, etc) and supervised learning (classification, regression, etc). Time permitting we might cover dynamic forms of learning (active learning, reinforcement learning, etc). We will assume basic familiarity with linear/parametric methods (linear regression, logistic regression, etc), and dwell more on nonlinear/nonparametric methods (kernels, random forests, boosting, neural nets, etc).

Critical thinking. Unlike other courses, we will not just list one algorithm after another. Instead, we will work on developing some skepticism when using these methods by asking more nuanced questions. When do these methods "work", why do they work, and why might they fail? Can we quantitatively measure if they are "working" or "failing"? Rather than just making a prediction, how can we quantify uncertainty of our predictions? How do we compare different regression methods or classification algorithms? How do we select a model from a nested class of models of increasing complexity? Are prediction algorithms useful for hypothesis testing? How can we interpret complex models, for example: what are measures of variable importance and data-point importance? These questions do not all have easy or straightforward answers, but various attempts at formalization and analysis will nevertheless be discussed (and will naturally lead to course projects, and potentially research projects).

3 Graded Components

There will be several homeworks and in-class quizzes and these will correspond to the majority of the grade.

Homeworks (15 per HW, 60% total) There will be one homework due in Feb, Mar, Apr and May.

All 4 homeworks will be due on some Fri of each month at 5pm: tentatively on **Feb 10, Mar 17, Apr 14, May 5**. The HW will be released around two weeks before the due date, and no less than ten days before the due date. A TOTAL (across all homeworks) of four late days will be tolerated (but not encouraged), but you cannot use more than two late days for any single homework. So, for example, you can submit two homeworks on time and two homeworks on Sun by 5pm if you wish. Or you can use one late day on each homework. But I do not encourage working on the weekends (especially on a Saturday!), so please use the late days only if really necessary.

Homeworks will follow the following broad guideline: the first question will be practice with fundamentals (working with definitions), the second will be a theoretical/mathematical question focusing on conceptual progress, the third will be a programming/computational assignment with a real dataset, and the fourth question will alternate between an extra theoretical question and a more advanced simulation/programming question.

Quizzes (10 per in-class quiz, 30% total) Tentatively on **Feb 23, Mar 28, Apr 25**, to exploit appropriate gaps between homeworks. It will involve multiple-choice or T/F questions only. The questions will carefully probe

concepts that you should know if you attend class and read the relevant chapters/papers, and reflect on the material (in other words, I would prepare hard, even though it's only multiple choice).

Crowd-scribing (5%) Each student (auditing or crediting) will have to scribe one lecture, and you can rely on (meaning, improve, correct, elaborate) an old scribe if you want.

Class participation (5%) I encourage 100% attendance and frequent class participation; every study on education research that I have read concludes that academic performance is negatively affected by skipping class and positively affected by asking questions and engaging with the material (as opposed to passively listening, and not clarifying your doubts). You will get the entire 5% if, during at least 75% of the lectures, and you either ask a nontrivial question or have your video on for most of the attended classes (tracked by the TA). In other words, you will be excused (without needing to provide any reason) for one out of every 4 classes at no loss to your grade. Beyond that, the grade will proportional to the participation.

Projects (optional, bonus, up to 5%) Projects are optional, and can be treated as a bonus. If anyone has lower HW and/or exam grades than they hoped, they can bump up their grade (in borderline cases) by doing an extra project. There are a wide variety of options available for course projects. Typical examples include:

- (Survey) You can survey an area of the literature (that is covered in a textbook, or a set of advanced papers) that is related to the course, and is complementary to what is covered in class.
- (Programming) You can create a set of graphs, plots, or interactive figures, which allow the user to visualize several of the methods covered in the course. For inspiration, check out distill.pub, and specifically, a paper on why momentum works.

Grades will ultimately be awarded based on the instructor's judgment of the amount of work completed in the project. Students will be evaluated on both writing (project reports) and speaking (project presentations).

Teaching (optional, bonus, up to 5%) Collectively, the class is very likely to know much more about statistical learning than the instructor. If anyone is interested in lecturing on a particular topic for they know the literature reasonably well and have good intuition to convey, the instructor is happy to flip the classroom a couple of times. Based on feedback from students/TA/instructor, this can also be used as a bonus to bump up the grade in borderline cases.

At the very least, if someone wants to give a 5-10 minute in-class presentation at the start of class on some topic of their choosing, this can also be used to bump up the grade in borderline cases.

4 Learning Objectives

Upon successful completion of the course, the student will be able to

- Explain how the bias-variance arises in different ML algorithms
- Compare models based on heldout predictive performance
- Implement several nonlinear, nonparametric ML methods
- Quantify generalization error in theory and practice
- Understand the terminology differences in the Stat and ML literatures
- Estimate uncertainty in predictions made by regression algorithms

4.1 Approximate table of contents (approximately ordered)

- K-nearest neighbors: simplest nonparametric method for classification and regression
- Conformal prediction: a generic tool for quantifying uncertainty
- Boosting: including the game-theoretic perspective and the minimax theorem
- VC theory, Rademacher complexity, generalization error, uniform convergence

- Bagging and random forests
- Variable and datapoint importance using Shapley values
- Reproducing Kernel Hilbert Spaces
- Deep neural networks
- Can't choose? Stacking: generic method to combine predictors
- Advanced topics and/or projects, time permitting

5 Course policies

5.1 Attendance

On-time attendance is expected and highly recommended. Every research study on this topic that I have read concludes that academic performance is negatively affected by not showing up to class.

5.2 Collaboration

Discussion of class material is heavily encouraged. Additionally,

- After submission of a homework, discussion of answers is always encouraged.
- Before submission of a homework, reasonable verbal discussion of homeworks is allowed. Copying in any form is disallowed. The rest of this bullet point is to clarify what "copying" and "reasonable" mean:
  - Most forms/instances of collaboration are not even close to "copying", so my null hypothesis is that most collaborations are well-intentioned and reasonable and there is no need to worry.
  - If there is a group discussion about a problem, in the sense of people trying to solve a problem by brainstorming together around a board or a book, then that is reasonable.
  - If one person has solved the problem, and writes the solution down on a board/book for others to write down (potentially without understanding), then that is unreasonable and counts as copying.
  - If one person has solved the problem, and another person has not solved the problem *after thinking about it for a while*, and the first person explains some key ideas/steps to the second and thus enables them to solve the problem, that is reasonable.
  - If you are stuck at some point in a proof, and ask someone for help and they explain how to get unstuck, that is reasonable.
  - If one person has already solved the problem, and shows a completed Latex-ed PDF solution to someone else for them to read and mimic, that is unreasonable and counts as copying.

  Most students do not copy or enable copying, but if it does happen, both parties may be at fault.
- Litmus test: usually, if the collaboration is reasonable and you explained it to others, most people outside the collaboration would also agree that it was reasonable. However, if you really hesitate to explain to others honestly how the collaboration worked, or if you do explain and your friends are surprised that such collaboration is okay, then you may be misjudging what is expected. (In such situations, ask the TA or instructor.) In short, listen to your own moral compass and you should be fine, and otherwise try to calibrate it using others.
- No matter what discussions have taken place, every homework and cheat sheet and mini-project and self-test (in its entirety) must be written up or coded up alone.

5.3 Academic Integrity

I have a zero tolerance policy for violation of class policies. If you are in any doubt whether a form of collaboration or obtaining solutions is permitted, please clarify it with me before proceeding.

- For each question on each homework, collaborators for that question must be acknowledged. Copying solutions from the internet is explicitly disallowed. You may search for material to help you understand a concept better, but be sure to create your own final solution. If you happen to use results from Wikipedia or textbooks, you must cite the source and are expected to completely understand the result you are citing. However, it is disallowed to copy solutions to exercises from elsewhere on the internet, like other courses or papers. When quoting text from a textbook, paper or website, use the `\begin{quote}` option in Latex.
- Any deviation from the rules will be dealt with according to the severity of the case. For example: evidence of written discussion in a larger group than 3-4 will result in points earned for that question becoming zero for all those relevant students; blindly copying one solution from someone else or online will result in the maximum points that can be earned for that homework becoming zero (maximum eligible grade becomes B); repeat occurrences will result in a failing grade for the course.
- In line with university policy, all instances of cheating/plagiarism will be reported to your academic advisor and the dean of student affairs. See the university policy on academic integrity.

5.4 Use of Mobile Devices and Laptops in Class

The use of mobiles and laptops in class is heavily discouraged. Learning research shows that unexpected noises or movement automatically divert and capture people's attention, meaning that you are affecting everyone's learning experience. For this reason, I ask you turn off your mobile devices and close your laptops during class. If you must use your laptop or mobile, make sure you are sitting at the back of the class.

5.5 Late Assignments

Every student is allowed a total of 4 late days for the course, with a max of two per assignment. Beyond that, the maximum earnable points for that assignment will drop by 20% per day.

6 Additional information

6.1 Global Communication Center

For assistance with the written or oral communication assignments in this class, visit the Global Communication Center (GCC). GCC tutors can provide instruction on a range of communication topics and can help you improve your papers, presentations, and job application documents. It is free, open to all students, and located in Hunt Library. You can make tutoring appointments on the GCC website: http://www.cmu.edu/gcc. Also browse the GCC website to find out about communication workshops offered throug the year.

6.2 Accommodations for Students with Disabilities

If you have a disability and are registered with the Office of Disability Resources, I encourage you to use their online system to notify me of your accommodations and discuss your needs with me as early in the semester as possible. I will work with you to ensure that accommodations are provided as appropriate. If you suspect that you may have a disability and would benefit from accommodations but are not yet registered with the Office of Disability Resources, I encourage you to contact them at access@andrew.cmu.edu.

6.3 Statement of Support for Students' Health & Well-being

Take care of yourself. Do your best to maintain a healthy lifestyle this semester by eating well, exercising, avoiding drugs and alcohol, getting enough sleep and taking some time to relax. This will help you achieve your goals and cope with stress. If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, we strongly encourage you to seek support. Counseling and Psychological Services (CaPS) is here to help: call 412-268-2922 and visit http://www.cmu.edu/counseling/. Consider reaching out to a friend, faculty or family member you trust for help getting connected to the support that can help.

If you or someone you know is feeling suicidal or in danger of self-harm, call someone immediately, day or night (CaPS: 412-268-2922, Resolve Crisis Network: 888-796-8226). If the situation is life threatening, call the police (On-campus CMU Police: 412-268-2323, Off-campus Police: 911).