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

Spring 2020 (Jan 13 to May 6), Syllabus

January 15, 2020

1 Basic Course Information

Instructor Aaditya Ramdas, aramdas@stat.cmu.edu [Office hours: 12-12:30pm MW]
TA: Pratik Patil, pnp1@andrew.cmu.edu [Office hours: 2-3pm T]

Time: 10:30-11:50am MW
Location: DH 1211
Exceptions: Jan 20 (MLK day) is a university holiday, 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.

Participants This course can be credited by PhD students with good mathematical background, but it can be audited by anyone who is curious about the topic. Students who want to sit through the course must officially audit.

Prerequisites Enrolled students are expected to have completed at least one intermediate statistics course, 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.

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,

(B) Bias-variance thinking,

(C) Computational considerations

(D) Data analysis

(E) Explainability and interpretability.

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, 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) Approximately, there will be one homework due at the end of Jan, Feb, Mar and Apr.

All 4 homeworks will be due (tentatively) on the last Fri of each month at 5pm: Jan 31, Feb 28, Mar 27, Apr 24. A TOTAL (across all homeworks) of two late days will be tolerated (but not encouraged). So, for example, you can submit two homeworks on time and two homeworks on Sat by 5pm if you wish. Or you could submit three homeworks on time, and one on Sun by 5pm if you wish. Some aspects of the homeworks will be discussed in class on Monday morning, so no submissions beyond Sun 5pm will be accepted.

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 12, Mar 18, Apr 15. It will involve multiple-choice or T/F questions. Basically testing concepts that you should know if you attend class.

Crowd-scribing (10%) Each student (auditing or crediting) will have to scribe one lecture.

Projects (optional) 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 (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) 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.

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 encouraged.
- Before submission of a homework, reasonable verbal discussion of homeworks is allowed. An example of unreasonable verbal discussion: one person reciting formulae orally while another one writes them down. Written discussion (in any form) is permitted in groups smaller than 3 (or in rare exceptions 4) students.
- 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 2 late days per mini. 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. The GCC is a free service, open to all students, and located in Hunt Library. You can make tutoring appointments directly on the GCC website: http://www.cmu.edu/gcc. You may also browse the GCC website to find out about communication workshops offered throughout the academic 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).