MACHINE LEARNING I

EN.553.740

Fall 2023

https://www.jhu.edu

Instructor:	James Schmidt	**Time:**	1630–1745 MW	**Office Hours:**
Email:	aschmi40@jhu.edu	**Place:**	Shaffer 300	After class Wednesday

Description:

Machine Learning describes a mishmash of computational techniques for "finding patterns in data." The scope of use, analytic tools, algorithms, and results are almost too numerous to meaningfully batch all such applications under a common appellation. Still, we try. This course, the first in a two-course sequence, focuses on *supervised* machine learning, which roughly deals with using historical labeled data to construct a predictor to accurately label future data.

Formally, we work in joint probability space $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ denotes "input" and $\mathcal{Y}$ denotes "output." While these notions are heuristic, they conveniently describe the scenario wherein data from $\mathcal{X}$ may be readily sampled, whereas data from $\mathcal{Y}$ is difficult or expensive to acquire, and we would like to decision according to how input data x is associated with label y.

In this course, you will learn how to formulate the supervised learning problem in mathematical terms, describe a measure of performance, restrict a search space for constructing models for prediction, optimize performance over the search space, and check for generalization. You will also learn about learning guarantees and of the limits of such results. While this course is primarily theory-centric, there will be no dearth of opportunity for exercising computational techniques.

Prerequisites:

Working knowledge of—or aptitude and willingness to learn as needed—probability theory, real analysis, multivariable calculus, analysis, linear algebra, and python

TAs: Daniel López-Castaño (email: jlopezc1@jhu.edu), Eleanor Belkin (email ebelkin1@jhu.edu)

Office Hours on TBD

Git Repository: https://github.com/schmidttgenstein/fa23-mli.git
Assignments, data, and starter code will be posted on this repository. Please do not push anything to the repo.

Useful References:

Reference 1. *Foundations of Machine Learning, M. Mohri, A. Rostamizadeh, and A. Talwalkar*

Reference 2. *Understanding Machine Learning, S. Shalev-Shwartz and S. Ben-David*

Reference 3. *All of Statistics, L. Wasserman*

Reference 4. *High Dimensional Statistics, M. Mainwright*

Reference 5. *High Dimensional Probability R. Verhynin*

Reference 6. *Real Python (`https://realpython.com/`)*

Reference 7. *Introduction to Statistical Learning, G. James et al.*

Workload and Grading:
The weights, while at the moment indeterminate, will comprise a genuine probability measure.

1. **Homework 60-90%**

   You will have roughly ten assignments. They will include problem sets and programming assignments. Homework should be turned in on time, and late submissions will not be accepted, with the exception of exactly one assignment of your choice which may be turned within two weeks of original due date without penalty. We will separately drop your lowest homework assignment grade. Therefore, any unsolicited rationale regarding non-submission or late submission will be treated as exclusively informational.

2. **Lecture Notes, 10-20%**

   Each student will be assigned at least one lecture to take "notes" for. These notes will not be shared with the class (i.e. are not for scribing purposes). The task requires you to summarize main points of lecture, articulate questions you may have or lingering ambiguities, identify how it fits in with the course narrative. The purpose of this task is twofold: 1. for you to take an active role not merely in following lecture but also synthesizing information and 2. for me to spot check student understanding without giving exams or quizzes to realize the same end

3. **Attendance, 0%**

   Come to class

4. **Final Exam, 0 - 20%**

   There may or may not be a final exam

Expectations:

- Respect: you may and should work together and collaboratively. Ask incisive questions with intellectual curiosity and enthusiasm, challenge assumptions and methods, but do not put down anyone ever.
- Integrity: personal integrity is one of the most important things you can strive to build up, buttress, and maintain. Take pride in that the honesty of your work reflects who you are.
- Know University policies
  - undergraduate students: https://studentaffairs.jhu.edu/policies-guidelines/undergrad-ethics/
  - graduate students: https://homewoodgrad.jhu.edu/academics/policies/

My Philosophy:

- Your educational experience should be rich, joyful, fun, educational, and challenging. I encourage everyone to have fun learning, while maintaining a rigorous sense of responsibility for actively living up to e.g. the above expectations.
- I care dearly about both your academic and human success. Please do not hesitate to reach out to me.

Tentative Schedule of Topics:

- Week 1: Introduction
  - Course overview, outline of machine learning, and probability review (ch 1 reference 1, ch 1 reference 2, ch 2 reference 7)
- Week 2 and 3: Concentration Bounds
  - No class on Monday September 4
  - Hoeffding (ch 1 and 2 reference 5, ch 1 reference 4) and Glivenko-Cantelli (ch 4 reference 4)
- Week 4 and 5: Linear Regression
  - Inner product spaces and Hilbert Projection (notes)
  - Regression $\mathbb{E}(y|x)$ as instance of Hilbert Projection
- Week 5: Beyond Hilbert
  - Overview of generic supervised learning problem
  - Restricting the search space: the Hypothesis Class
  - Empirical risk minimization
- Weak 6 and 7: Expressivity
  - Complexity of Hypothesis Class
  - Partitions of Unity and Universal Approximation Theorem
- Week 7-8: Overfitting and Generalization (Anticipate PAC)
- Week 8-9: Optimization
- Week 9-13: PAC Theory (ch 3-5 reference 1)
- Week 13-14: Reproducing Kernel Hilbert Spaces (ch 6 reference 1)