CSCI 381 Machine Learning

Spring 2025

Prolegomena unit:

- Course introduction (**today**)
- Basich machine learning terminology (Wednesday)
- Lab: Python libraries (Friday)

Today:

- What machine learning is all about
- Syllabus and course details
- Machine learning in context of related fields

One example of where a handcoded approach will fail is in detecting faces in images. Today every smartphone can detect a face in an image. However, face detection was an unsolved problem as recently as 2001. The main problem is that the way in which pixels are "perceived" by the computer is very different from how humans perceive a face. This difference in representation makes it basically impossible for a human to come up with a good set of rules to describe what constitutes a face in a digital image.

Müller and Guido, Introduction to Machine Learning with Python, 2017, pg 2

Purpose of the course

There are several models for what a course in *machine learning* could be. At one extreme are courses that present machine learning as an advanced area of probability and statistics. At the other are courses that teach students to patch together machine learning applications from libraries without needing to understand the mathematics behind them. Both of those models neglect the *algorithms* of machine learning, and neither of them reflect our intention in this course.

Instead, this course is designed to present a balanced approach to machine learning: as we consider a selection of machine-learning techniques, students will competently but not exhaustively explore the *mathematics* of machine learning; they will practice applying machine learning *libraries* to solve real-world problems; but especially they will learn the *algorithms* of machine learning by implementing them from directly.

Intent and philosophy of this course:

- CSCI 381 has minimal overlap with MATH 465 and CSCI 384—the courses are complementary but not redundant or interdependent.
- CSCI 381 requires no prior experience in math besides Calc I and linear algebra—any additional math material (probability, partial derivatives) is self contained.
- CSCI 381 emphasizes the *algorithms* of machine learning—the *mathematical background* and the *applications* are supporting topics.
- CSCI 381's goal is understanding rather than comprehensiveness—it's better to understand the topics covered than to cover more topics.
- CSCI 381's programming assignments involve implementing machine learning algorithms (mostly) from scratch, because that is the way to understand an algorithm.

Practice in the course:

- Frequent, short(-ish), closed-ended programming assignments to implement techniques (almost) from scratch.
- A single, semester-long, open-ended applied project.
- Frequent, lightweight quizzes to summarize the main points/concepts/terms.
- A typical four-day pattern for topics consisting of concepts, applications (in lab), mathematical details, algorithmic details.
- Periodic readings in ethical/social/legal issues throughout the semester, with one week of intense discussion at the end of the semester.

Machine learning. The field of computer science that studies techniques for training algorithms from data.

Artificial intelligence. (Moving target.)

Statistical inference. The area of mathematics that studies building and evaluating statistical models from data.

Data science. An umbrella category for a variety of fields or activities, including data curating, data mining, data analytics, and predictive analytics.

Pattern matching. A field similar to machine learning but from an engineering origin.

Methods of statistical inference help us in estimating the characteristics of a population based upon the data collected from (or the evidence produce by) a sample. Statistical techniques are useful in both the planning of the measurement activities and the interpreation of the collected data.

Trivendi, Probability and Statistics, 1982, pg 469.

The basic problem of statistical inference is the inverse of probability: Given the outcomes, what can we say about the process that generated the data? . . . Data analysis, machine learning, and data mining are various names given to the practice of statistical inference, depending on the context.

Wasserman, All of Statistics, 2004, pg ix.

Machine learning is essentially a form of applied statistics with increased emphasis on the use of computers to statistically estimate complicated functions and a decreased emphasis on proving confidence intervals around these functions.

Goodfellow, Deep Learning, 2016, pg 95.

There's a joke that says a data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician.

Joel Grus, Data Science from Scratch, 2015, pg 1

he problem of searching for patterns in data is a fundamental one and has a long and successful history. For instance, the extensive astronomical observations of Tycho Brahe in the 16th century allowed Johannes Kepler to discover the empirical laws of planetary motion, which in turn provided a springboard for the development of classical mechanics. Similary, the discovery of regularities in atomic spectra played a key role in the development and verification of quantum physics in the early twentieth century. The field of pattern matching [used as a synonym for machine learning] is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories.

Bishop, Pattern Recognition and Machine Learning, 2006, pg 1

Coming up:

Send me your github userid

Learn Python

Take first-day-of-class quiz (due classtime Wednesday)

Do Python warm-up assignment (due Friday)