CMPUT 267 Basics of Machine Learning
Winter 2024

Nidhi Hegde

January 9 2024

What is Machine Learning?

What is Machine Learning?

- Mitchell: “The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience."
- Russell & Norvig: “... the subfield of AI concerned with programs that learn from experience.
- Murphy: “The goal of machine learning is to develop methods that can automatically detect patterns in data, and then to use the uncovered patterns to predict future data of other outcomes of interest."

What is this course about?

We want to be able to have good rules (or functions) for predicting outcomes e.g., predict the temperature tomorrow, based on weather over the last five days.

You could construct these rules by hand, or learn them from data:

- But the data are often incomplete:
  - Partial observability: Incomplete knowledge of environment
  - Incomplete knowledge of other agents' actions
- Machine learning algorithms are one way to learn from incomplete data

Course goal

Understand machine learning algorithms by deriving them from the beginning,

- with a focus on prediction on new data.

Example: Predicting House Prices

- Goal: we want to predict house prices, given only the age of the house
  $f(\text{age})$ = price of the house
- Dataset: house sales this year, with attributes age and target value price

$$\{(\text{age}_1, \text{price}_1), (\text{age}_2, \text{price}_2), \ldots, (\text{age}_9, \text{price}_9)\}$$

- Idea: A function that accurately outputs price from age for these specific pairs might also provide good predictions for new houses

Formalizing the problem

Definitions:

Let x be *age* and y be *price*.

Let $D = \{(x_1, y_1), \ldots, (x_9, y_9)\}$ be our dataset.

Let $f(x)$ be our prediction of the house price for age=x.

Objective:

We want the difference between $f(x_i)$ and y_i to be small

$$\min_{f \text{ in function space}} \sum_{i=1}^{9} (f(x_i) - y_i)^2$$

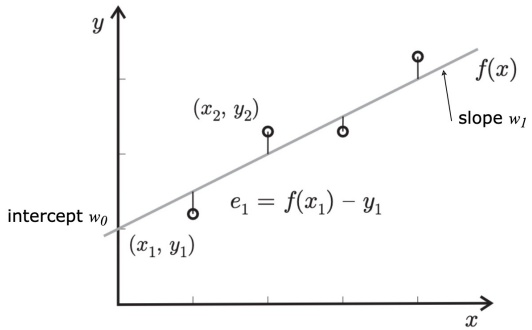
Questions

1. Why are we squaring the difference?
2. Why are we summing up the errors?
3. What could we consider for the function space?

Linear function space

Definition

A function f is a **linear function** of x if it can be written as $f(x) = w_0 + w_1 x$



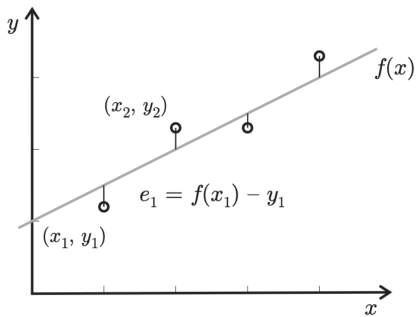
intercept w_0
slope w_1

Solving for the optimal function

Objective then becomes:

$$\min_{f \text{ in function space}} \sum_{i=1}^{9} (f(x_i) - y_i)^2$$

$$= min_{w_0, w_1} \sum_{i=1}^{9} (w_0 + w_1 x_i - y_i)^2$$



Questions:

1. Would you use this to predict the value of a house? Why/why not?
2. Will this predict well? How do we know?
3. What is missing to make these assessments?

Probabilities

- Question: Is it likely that there is a deterministic function from age to price?
  - Many houses will have the same age but different price...
- We can instead use a probabilistic approach:
  - Learn a function that gives a distribution over targets (price) given attributes of the item (age)
- Question: Does this mean that we think the world is stochastic rather than deterministic?
  - Stochasticity can come from partial observability
  - Maybe the outcome really is deterministic if we knew age, and size, and number of rooms, and distance to airport, and whether the queen lives there, and ...

Course Topics

1. Probability Background (Ch 2)
2. Estimation with Sample Averages (Ch 3)
   - Concentration inequalities: how confident should we be in our estimates?
   - Sample complexity and convergence rate
3. Optimization (Ch 4)
4. Parameter Estimation (Ch 5)
   - Maximum likelihood and MAP
   - Beyond point estimates: Bayesian estimation
5. Stochastic Gradient Descent and Handling Big Datasets (Ch 6)

Course Topics

6. Prediction (Ch 7)
   - Formalizing the prediction objective
7. Linear and Polynomial Regression (Ch 8)
8. Generalization Error and Evaluating Models (Ch 9)
9. Regularization and Constraining the Function Space (Ch 10)
10. Logistic Regression and Linear Classifiers (Ch 11)
11. Bayesian Linear Regression (Ch 12)

Course Structure:

- Part 1 (Topics 1–5) focused on providing needed background
- Part 2 (Topics 6–) more directly about ML concepts

Course Essentials

- **Course information:** https://nidhihegde.github.io/mlbasics
  - Schedule, readings, lecture notes
  - Getting Started and FAQ
- **Access-controlled course information:** eClass
  - https://eclass.srv.ualberta.ca/course/view.php?id=95783
  - Getting Started and FAQ
  - Assignment release and submission
- My Office hours: Fridays 1:00 pm – 2:00 pm, my office ATH 3-04.
- For questions about coursework (assignments, exams) and other assistance email cmput267@ualberta.ca.
  - Email me only with private questions about health, emergencies, missed work, etc.

Teaching Assistants

Duc Thang Chu
Vedd Kuknur
Guoqing Luo
Olya Mastikhina
Yongchang Hao
Shreya Pekhale
Kailash Seshadri
Aniket Sharma
Vlad Tkachuk

- Office hours: see course webpage, there will be a signup sheet to ease the load
- No lab sessions - seek help during TA office hours.
- Piazza - signup at https://piazza.com/ualberta.ca/winter2024/cmput267.

Readings

- From Basics of ML textbook (see course webpage)
- I will add supplemental material throughout the term.
- Other reference material we will refer to (see course webpage and FAQ document)
  - Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, An Introduction to Statistical Machine Learning.
  - T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning.
  - David Barber, Bayesian Reasoning and Machine Learning.
  - C.M. Bishop, Pattern Recognition and Machine Learning.
  - Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong, Mathematics for Machine Learning.

Lectures

- In person only; whiteboard lectures (my iPad, projected on the screens)
- Lecture slide skeleton will be up on the schedule before class and the annotated lecture slides will be up after class.
- Ask questions!
  - Raise your hand ask - if you're wondering about something, likely many others are too.
  - Ask later or on Piazza.
  - Ask at my or TA office hours.
- **Take your own notes**
  - This is important for you to write things down as you understand them.
  - I may say things I don't write down.
  - There might be questions you want to write down about.
  - My handwriting is not great.
  - Writing things down helps you remember!

Course Discussion

- Piazza
- Piazza - signup at https://piazza.com/ualberta.ca/winter2024/cmput267
- This is the *first* place to go to ask questions:
  - someone else has the question too,
  - your peers might answer,
  - it fosters healthy discussion.
- Details in syllabus and FAQ document

Course Prerequisites

See details on course numbers in the university course catalogue.

- Pre-requisites
  - Calculus 1
  - Computing 1 (CMPUT 174/274)
- Co-requisites (ideally pre-)
  - Probabilities: Stats 1
  - Linear algebra 1
  - CMPUT 175, CMPUT 272
  - We will have short refreshers along the way
- **Motivation** to learn
- Motivation to think **beyond the material**
- I welcome feedback, both during and outside of lecture.

Why is there so much math?

- **This course is very mathematical**, with detailed derivations.
  - This is absolutely necessary.
- "But I just want to use machine learning to solve Problem X!"
  1. **Applying algorithms correctly** is much easier when you understand their development and assumptions.
     - You will be more effective at solving Problem X if you **understand the algorithms** that you apply.
     - This means understanding their derivation.
  2. **Formalizing the problem** is often half the battle to solving it effectively!
     - Comfort with math is an important part of being a computer scientist.

Problem Solving

- CS is about problem solving through the medium of computing
  - Not about becoming an expert programmer
- Primary goal is carefully designing solutions to problems, by:
  - **Formalizing** the problem,
  - **Understanding** different potential approaches,
  - **Evaluating** the solution.
- Comfort with mathematical concepts enables clarity through logical thinking

Marking

Component	Weight	Date/Time
Participation and Reading Questions	10%	per eClass
Assignments (8, top 7 counted)	30%	per eClass
Midterm Exams (2, 15% each)	30%	per eClass
Final Exam	30%	per eClass

- Participation and Reading Questions: 1 per week (starting next week)
  - based on lectures and readings
  - quick questions, answered on eClass.
- Assignments: 8, best 7 out of 8 count towards grade.
- Letter grade determined at end of course.

Marking

- Participation and reading questions: 24 hour window; No late submissions will be accepted.
- Assignments: no late policy
  - submitted on eClass at 11:59 pm on deadline date
  - 48 hour grace period, no late submissions after that.
- Missed work
  - Assignments considered as incomplete if not submitted.
  - Note: you can miss one and incur no penalty (best 7 of 8)
- Midterm Exams
  - No extensions or re-examinations for midterm exams.
  - If you cannot complete an exam due to incapacitating illness, severe domestic affliction or other compelling reasons, you must contact the instructor as soon as possible, to request an excused absence.
  - If an excused absence is granted, then the weight of that missed exam is transferred to the final exam.
  - An excused absence is a privilege and not a right. There is no guarantee that an absence will be excused.

Collaboration Policy

- Consult the syllabus
- Midterm exams and Final exam:
  - Individual work / solo effort: **no** collaboration of any kind
  - Have to be able to explain your submission upon request from TA/instructors
  - TAs are on a look-out for signs of cheating
- Assignments. You are encouraged to discuss assignments with other students:
  - You must list everyone you talked with about the assignment.
  - You may not share or look at each other's written work or code.
  - You must write up your solutions individually
- Violations result in disciplinary action and a file record

Assignments

- 8 assignments
- Marks are binned:
  - $81 - 100 \rightarrow 100$
  - $61 - 80 \rightarrow 80$
  - $41 - 60 \rightarrow 60$
  - $0 - 40 \rightarrow 0$

Exams

- Giving clear answers to short answer questions is a skill
  - It takes practice!
- Practice questions will be available.
- The Midterm and Final Exams will be in person.

Academic Conduct

- Submitting someone else's work as your own is plagiarism.
- So is helping someone else to submit your work as their own.
- We report all cases of academic misconduct to the university.
- The university takes academic misconduct very seriously. Possible consequences:
  - Zero on the assignment or exam (virtually guaranteed)
  - Zero for the course
  - Permanent notation on transcript
  - Suspension or expulsion from the university

Readings

- It is critical that you do the readings
- in-class lectures follow the notes quite closely
- If you read and understand the notes, you have learned a lot about ML
- Ideally, you read the notes before I lecture on it, to help you understand them better (seeing it a second time)
- Marked Participation and Reading questions encourage you to actually do the readings

Participation and Readings questions

- Once a week, starting Week 2 (Jan 15)
- Based on lectures and readings exercises
- On eClass
- Open for 48 hours, you have 30 minutes to complete.
- Goal:
  - relatively simply questions
  - help to keep up with material

Julia instead of Python

- Julia is focussed on numerical computing, a bit like Matlab
- It can be a steep learning curve
- We have provided resources and a tutorial notebook.
- Tutorial sessions by TAs