Lecture 1 – Learning From Data

DSC 40A, Winter 2024

Agenda

1. Who are we?
2. What is this course about?
3. How will this course run?
4. How do we turn the problem of learning from data into a math problem?

Who are we?

Hi, everyone!

Aobo Li (pronounced obo)

- Assistant Professor with HDSI and Department of Physics
- Undergraduate at UW Seattle, PhD at Boston University, Postdoc at UNC Chapel Hill
- For fun: video game/esports, saxophone, photography

Course Staff

- 1 TA, who will lead the discussion and help run the class.
  - Zhenduo Wen, a MS student in DSC.
- Undergrad tutors, who will hold office hours, grade assignments, and help run the class.
  - Candus Shi, Benjamin Xue, Vivian Lin, Charlie Sun, Yuxin (Emily) Guo, Mert Ozer, Yujia (Joy) Wang, Yosen Lin, Sunan Xu

Course overview

Part 1: Learning from Data (Weeks 1 through 5)

- Summary statistics and loss functions; empirical risk minimization.
- Linear regression (including multiple variables) .
- Clustering.

Part 2: Probability (Weeks 6 through 10)

- Set theory and combinatorics; probability fundamentals.
- Conditional probability and independence.
- Naïve Bayes classifier.

Learning objectives

After this quarter, you'll...

- understand the basic principles underlying almost every machine learning and data science method.
- be better prepared for the math in upper division: vector calculus, linear algebra, and probability.
- be able to tackle the problems mentioned at the beginning.

How will this course run?

Basics

- The course website, **dsc40a.com**, contains all content. **Read the syllabus carefully!**
  - Stay tuned with the update and announcements on course website
  - The course website also contains lecture notes/videos developed by Dr. Janine Tiefenbruck. Use those as a “textbook”.
- We won’t use Canvas. **Campuswire** will be used for announcements and communication. You can sign yourself up with code 3914. **Ask questions here instead of email!**
- Fill out this **Welcome Survey**.

Lectures

- Lectures are held MWF at 1:00pm to 1:50pm in Mandeville B-202.
- Lecture slides will be posted on course website before class.
- Value of lecture: **interaction** and **discussion**.

Discussion

- Discussions on Monday at 5pm to 5:50pm in PCYNH 106.
- Discussion will be used primarily for **groupwork**.
  - Come to the discussion you’re enrolled in, and work on problems in small groups of size 2-4.
  - You may work in a self-organized group outside of the scheduled discussion sections for 80% credit. You may not work alone.
  - Value of attending: **TA/tutor support**.
- Submit groupwork to Gradescope by **11:59pm Monday**.
  - Only one group member should submit and add the other group members.

Assessments and exams

- **Homeworks**: Due **Wednesday at 11:59pm** on Gradescope. Worth 40% of your grade.

- **Groupworks**: Due **Monday at 11:59pm**. Worth 10% of your grade.

- **Exams**: Two midterms and a two-part final exam, which can redeem low scores on the midterms. Exams are Friday, Feb. 9 during lecture, Wednesday, March 13 during lecture, and Friday, March 22.

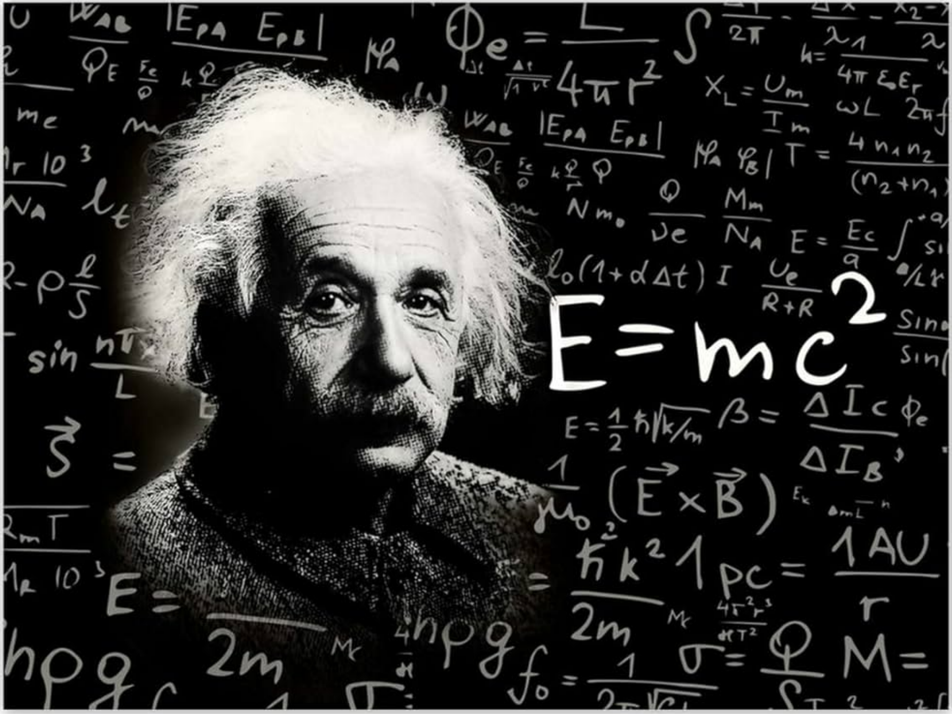
Support

- **Office Hours**: many hours throughout the week to get help on homework problems. Plan to attend at least once a week because the homework is hard!
  - See the calendar on the course website for schedule and location.
- **Campuswire**: Use it! We're here to help you.
  - Don't post answers.

Making predictions

Science

- In a sense, science is about **making predictions**
- On one hand, Nature works in mysterious ways.
  - We can’t see inside it’s head.
  - We only see its inputs and outputs (**data**).
- Very fortunately for us, Nature exhibits patterns.
- We try to understand Nature by building theories, or “models”.
- A model is good when it makes accurate predictions.
- But how do we come up with a model?


$E=mc^2$

Example: predicting energy from mass

- **Given**: a particle's mass, m
- **Predict**: the amount of energy E that the mass is equivalent to
- **Assumption**: Nature behaves in some predictable way (i.e., exhibits a pattern)
- **Einstein predicted:** $E = mc^2$
- He derived this **theoretically**
- Later, verified **empirically**, with data.

Example: predicting salary

- **Goal:** predict the salary of a data scientist
- **Assumption**: “Nature” behaves in some predictable way (i.e., exhibits a pattern)
- **Problem**: There isn’t a formula like $E = mc^2$ that exactly predicts salary
- What do we do?

Idea: use data

- We believe that Nature uses certain factors (years of experience, GPA, degree obtained, etc.) to determine salary
- But we don’t know exactly how it does so
- Like Einstein, we’ll think about what a formula for salary might look like (**theorize**).
- But we’ll leave some parts of the formula **unspecified**.
- Collect **data** about data scientists (name, age, salary, educational degree, …)
- Use that data to **learn** a formula, make predictions.

Learning from data

- Idea: ask a few data scientists about their salary.
  - StackOverflow does this annually.
- Five random responses:

90,000 94,000 96,000 120,000 160,000

Discussion Question

Given this data, what do you predict your future salary will be? How did you come up with this guess?

Some common approaches

- The **mean**:

$$\frac{1}{5} \times (90{,}000 + 94{,}000 + 96{,}000 + 120{,}000 + 160{,}000)$$
$$= 112{,}000$$

- The **median**:

$$90{,}000 \quad 94{,}000 \quad \underbrace{96{,}000}_{\uparrow} \quad 120{,}000 \quad 160{,}000$$

- Which is better? Are these good ways of predicting future salary?

Quantifying the goodness/badness of a prediction

- We want a metric that tells us if a prediction is good or bad.
- One idea: compute the **absolute error**, which is the distance from our prediction to the right answer.

  absolute error = |(your actual future salary) – prediction|
- Then, our goal becomes to **find the prediction with the smallest possible absolute error**.
- There's a problem with this:

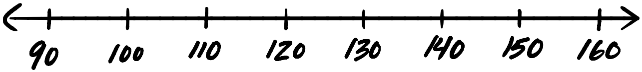
What is good/bad, intuitively?

- The data:

  90,000 94,000 96,000 120,000 160,000

- Consider these hypotheses:

  $h_1 = 150{,}000$ $h_2 = 115{,}000$



Discussion Question

Which do you think is better, h_1 or h_2? *Why*?

Quantifying our intuition

- Intuitively, a good prediction is close to the data.
- Suppose we predicted a future salary of $h_1 = 150{,}000$ *before* collecting data.

salary	absolute error of h_1
90,000	60,000
94,000	56,000
96,000	54,000
120,000	30,000
160,000	10,000
	sum of absolute errors: 210,000
	mean absolute error: 42,000

Quantifying our intuition

- Now suppose we had predicted $h_2 = 115{,}000$.

salary	absolute error of h_2
90,000	25,000
94,000	21,000
96,000	19,000
120,000	5,000
160,000	45,000

sum of absolute errors: 115,000

mean absolute error: 23,000

Mean absolute error (MAE)

- Mean absolute error on data:

$$h_1 : 42{,}000 \qquad h_2 : 23{,}000$$

- Conclusion: h_2 is the better prediction.

- In general: pick prediction with the smaller mean absolute error.

We are making an assumption...

- We're assuming that future salaries will look like present salaries.
- That a prediction that was good in the past will be good in the future.

Discussion Question

Is this a good assumption?

Which is better: the mean or median?

- Recall:

  mean = 112,000   median = 96,000

- We can calculate the mean absolute error of each:

  mean : 22,400   median : 19,200

- The median is the best prediction so far!
- But is there an even better prediction?

Finding the best prediction

- Any (non-negative) number is a valid prediction.
- Goal: out of all predictions, find the prediction h^* with the smallest mean absolute error.
- This is an **optimization problem**.

A formula for the mean absolute error

- We have data:

  90,000   94,000   96,000   120,000   160,000

- Suppose our prediction is h.
- The **mean absolute error** of our prediction is:

$$R(h) = \frac{1}{5}\Big(|90{,}000 - h| + |94{,}000 - h| + |96{,}000 - h| + |120{,}000 - h| + |160{,}000 - h| \Big)$$

A formula for the mean absolute error

- We have a function for computing the mean absolute error of **any** possible prediction.

$$R(150{,}000) = \frac{1}{5}\Big(|90{,}000 - 150{,}000| + |94{,}000 - 150{,}000| + |96{,}000 - 150{,}000| + |120{,}000 - 150{,}000| + |160{,}000 - 150{,}000|\Big) = 42{,}000$$

A formula for the mean absolute error

- We have a function for computing the mean absolute error of **any** possible prediction.

$$
\begin{aligned}
R(115{,}000) = \frac{1}{5}\Big(&|90{,}000 - 115{,}000| + |94{,}000 - 115{,}000| \\
&+ |96{,}000 - 115{,}000| + |120{,}000 - 115{,}000| \\
&+ |160{,}000 - 115{,}000|\Big) \\
&= 23{,}000
\end{aligned}
$$

A formula for the mean absolute error

- We have a function for computing the mean absolute error of **any** possible prediction.

$$R(\pi) = \frac{1}{5}\Big(|90{,}000 - \pi| + |94{,}000 - \pi| + |96{,}000 - \pi| + |120{,}000 - \pi| + |160{,}000 - \pi|\Big) = \mathbf{111{,}996.8584\ldots}$$

Discussion Question

Without doing any calculations, which is correct?
A. $R(50) < R(100)$
B. $R(50) = R(100)$
C. $R(50) > R(100)$

A *general* formula for the mean absolute error

- Suppose we collect n salaries, $y_1, y_2, \ldots, y_n$.
- The mean absolute error of the prediction h is:

- Or, using **summation notation**:

The best prediction

- We want the best prediction, h^*.
- The smaller $R(h)$, the better h.
- Goal: find h that minimizes $R(h)$.

Summary

- We started with the learning problem:

  Given salary data, predict your future salary.

- We turned it into this problem:

  Find a prediction h^* which has smallest mean absolute error on the data.

- We have turned the problem of learning from data into a specific type of math problem: an **optimization problem**.

- **Next time**: we solve this math problem.