Lecture 1 – Introduction, Learning From Data

DSC 40A, Fall 2021 @ UC San Diego

Dr. Truong Son Hy, with help from **many others**

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!

Background

- First name Son, last name Hy, middle name Truong. Born & raised in Hanoi, Vietnam.

Education

- PhD in Computer Science, University of Chicago, June 2022
- BSc in Computer Science, University of Budapest (Eotvos Lorand University, Hungary), July 2016

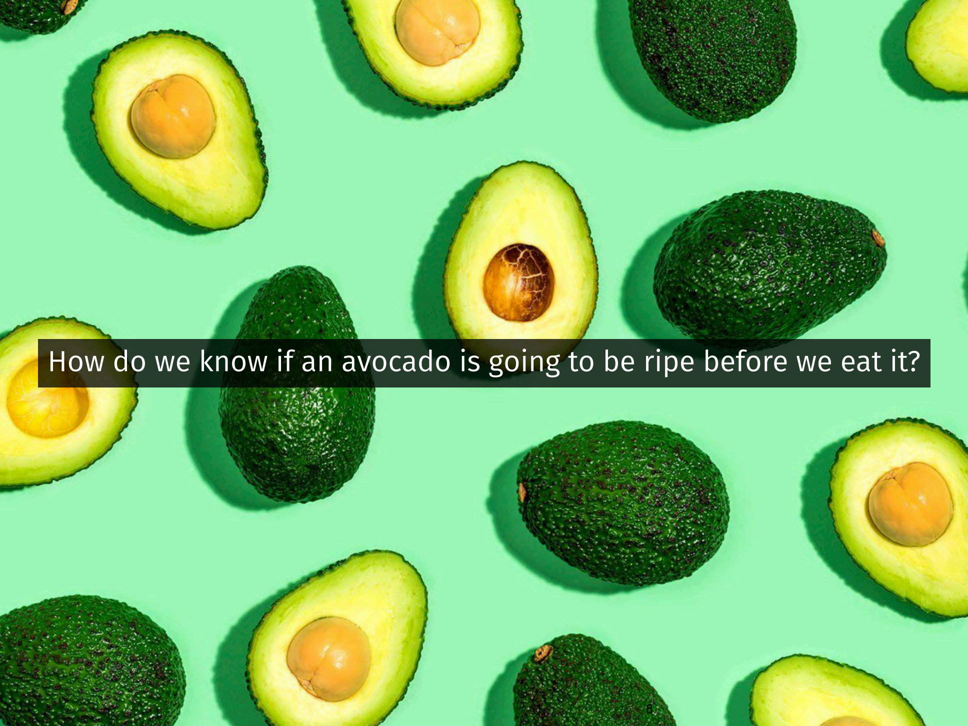
Research

- Graph representation learning & Deep generative models on graphs for drug discovery and material science
- Group/representation theory & Symmetry-preserving, physics-informed Machine Learning
- Multiresolution/multiscale models

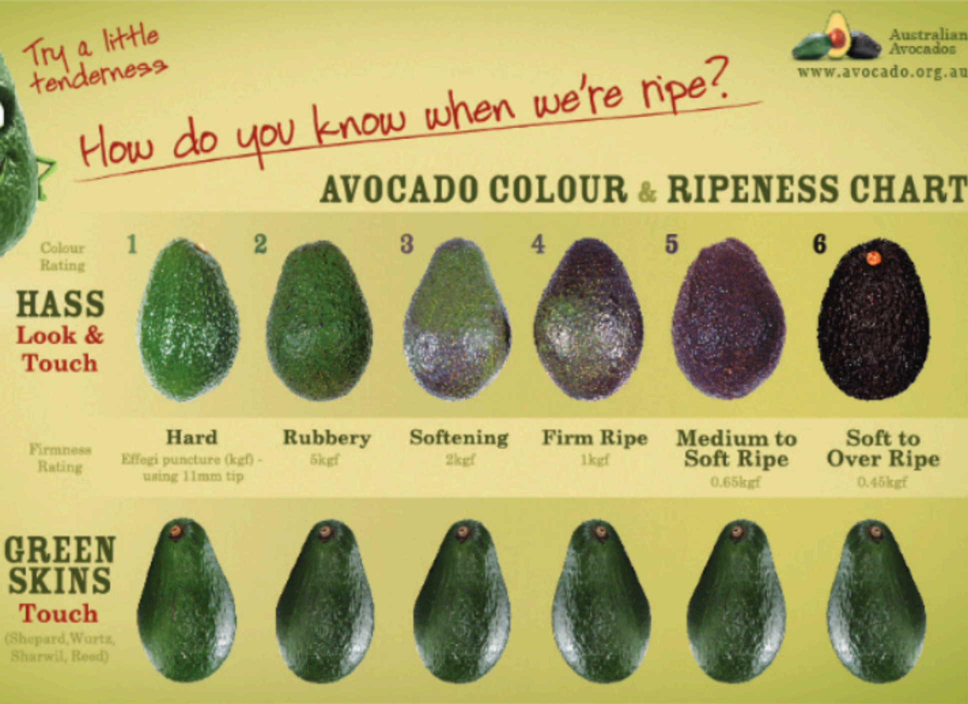
Say hey to course staff!

- 2 Instructors: Dr. Truong Son Hy and Dr. Mahdi Soleymani.
- 1 TA, who will teach discussion and help run the class.
  - Pushkar Bhuse, a MS student in CSE.
- Several tutors, who will hold OH, grade assignments, and help run the class.
  - Aryaman Sinha, Jessica Song, Karthikeya Manchala, Shiv Sakthivel, Vivian Lin, Weiyue Li, Yujia Wang, Yuxin Guo.
  - All undergrads who took DSC 40A before and did well.
- Read about them at **dsc40a.com/staff**.

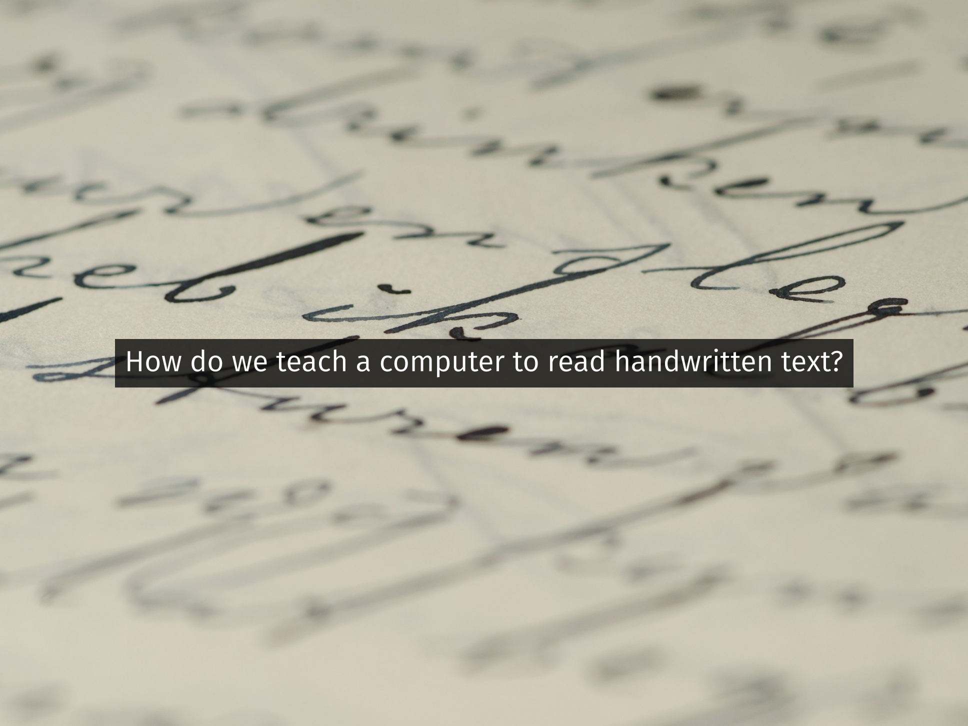
What is this course about?



How do we know if an avocado is going to be ripe before we eat it?



Try a little tenderness
Australian Avocados
www.avocado.org.au
How do you know when we're ripe?
AVOCADO COLOUR & RIPENESS CHART
Colour Rating
1
2
3
4
5
6
HASS
Look & Touch
Firmness Rating
Hard
Effegi puncture (kgf) - using 11mm tip
Rubbery
5kgf
Softening
2kgf
Firm Ripe
1kgf
Medium to Soft Ripe
0.65kgf
Soft to Over Ripe
0.45kgf
GREEN SKINS
Touch
(Shepard, Wurtz, Sharwil, Reed)


How do we teach a computer to read handwritten text?

How do we predict a future data scientist’s salary?

…by **learning** from data.

How do we learn from data?

The fundamental approach:

1. Turn learning from data into a math problem.

2. Solve that problem.

Course overview

Part 1: Learning from Data (Lectures 1-11)

- Summary statistics and loss functions; mean absolute error and mean squared error.
- Linear regression (incl. linear algebra).
- Clustering.

Part 2: Probability (Lectures 12-18)

- Set theory and combinatorics; probability fundamentals.
- Conditional probability and independence.
- Naïve Bayes (mix of both parts of the class).

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?

Technology

- The course website, **dsc40a.com**, is where all content (lectures, **readings**, homeworks, discussions) will be posted. It also contains a calendar of office hours (with Zoom links).

- **Campuswire** is where all announcements will be sent, and where all student-staff and student-student communication will occur. **Ask questions here!**

- **Gradescope** is where all assignments are submitted and all grades live.

- **Zoom** will be used for virtual office hours and discussion.

Lectures

Monday/Wednesday/Friday, Pepper Canyon Hall (PCYNH) room **122**. Two identical sessions:

- 3:00 – 3:50: Dr. Truong Son Hy
- 4:00 – 4:50: Dr. Mahdi Soleymani

What you should do

- Ask questions! Give me and Dr. Mahdi your feedback!
- Learn from everyone including the TA, tutors, classmates.
- Learn from any source including textbooks, online courses, research papers, etc.
- Learn by doing the homeworks!

Discussion

- **Discussion**:
  - Lead by the TA.
  - Monday, Pepper Canyon Hall (PCYNH) room **122**.
  - Two identical sessions: 5:00–5:50 and 6:00–6:50.
  - Come to work on problems in small groups (”groupwork”) of 2-4.
- Worksheets are due to Gradescope by **Monday at 11:59pm**.
  - Only one group member should submit; they should add the rest of the group to the assignment on Gradescope.

Assessments and exams

- **Homeworks**: Released weekly, and usually due **Fridays at 2pm** on Gradescope. Worth 40% of your grade.
- **Groupworks/Discussions**: Due **Monday at 11:59pm**. Worth 10% of your grade.
- **Midterm Exam**: TBD. Worth 20% of your grade.
- **Final Exam**: 12/03/2022, 7:00pm-9:59pm. Worth 30% of your grade.
- Both exams will be held **in-person**. Please resolve your schedule conflicts as soon as possible.

Leniency

We have some leniency built into our grading scheme:

- **Slip days**: 3. Can only be used on homework. Can only use one per homework.
- **Drops**: We will drop your lowest homework and groupwork.

Support

- **Office Hours (starting next week)**: held throughout the week, but concentrated near deadlines. Calendar on course website will be updated with times by the weekend.
  - Some staff OH are remote via Zoom. See Calendar for Zoom links. Others are in-person in the CSE Basement or library (TBD). Put yourself on the queue at autograder.ucsd.edu (”The Autograder”).
- **Campuswire**: Use it! We’re here to help you.
  - Don’t post answers.

How do we turn the problem of learning from data into a math problem?

How do we predict a future data scientist’s salary?

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, how might you predict your future salary?

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 = |(actual future salary) – prediction|

- Then, our goal becomes to **find the prediction with the smallest possible absolute error**.
- There's a problem with this:

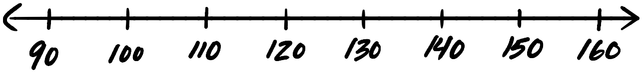
What is good/bad, intuitively?

- The data:

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

- Consider these hypotheses:

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



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:

$$\text{mean} = 112{,}000 \qquad \text{median} = 96{,}000$$

- We can calculate the mean absolute error of each:

$$\text{mean} : 22{,}400 \qquad \text{median} : 19{,}200$$

- 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:

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

- 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.

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

A formula for the mean absolute error

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

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

A formula for the mean absolute error

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

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

$$= \mathbf{111{,}996.8584...}$$

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.