Module 1 – Learning From Data

DSC 40A, Summer 2023

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!

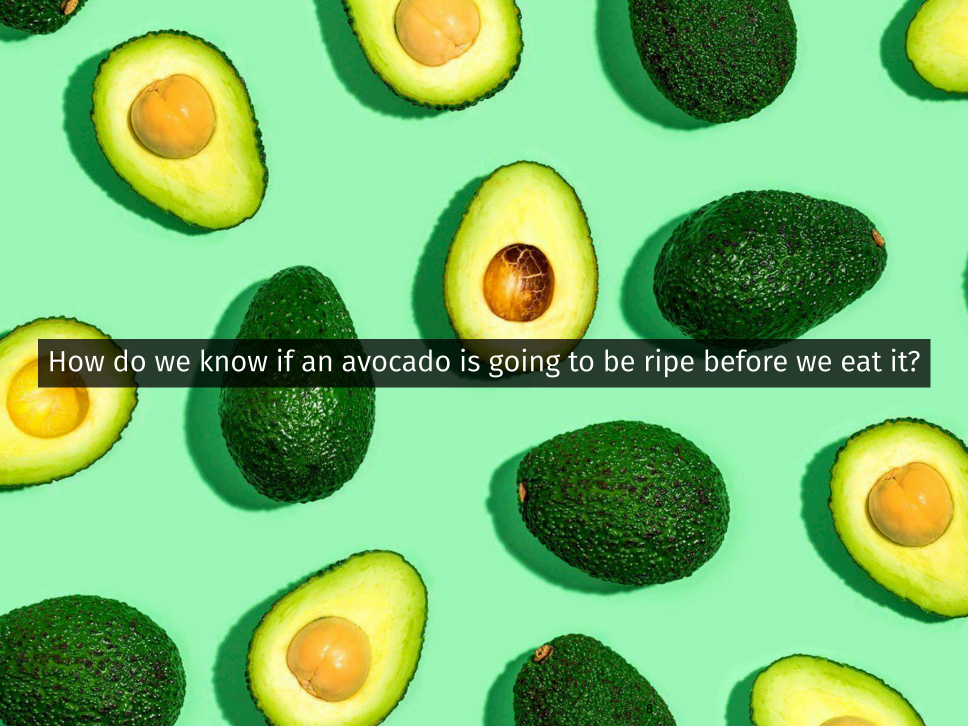
Rod Albuyeh (call me Rod)

- Grew up in CA. PhD in Political Science (quantitative) at USC.
- Data Scientist and Machine Learning Architect in Silicon Valley FinTech and San Diego HealthTech scenes, currently running a full-stack consultancy.
- Intermittent lecturer at UCSD, teaching DSC 102, joining USD AAI faculty in fall teaching CNNs.
- For fun: martial arts, dabbling musician, outdoorsy things

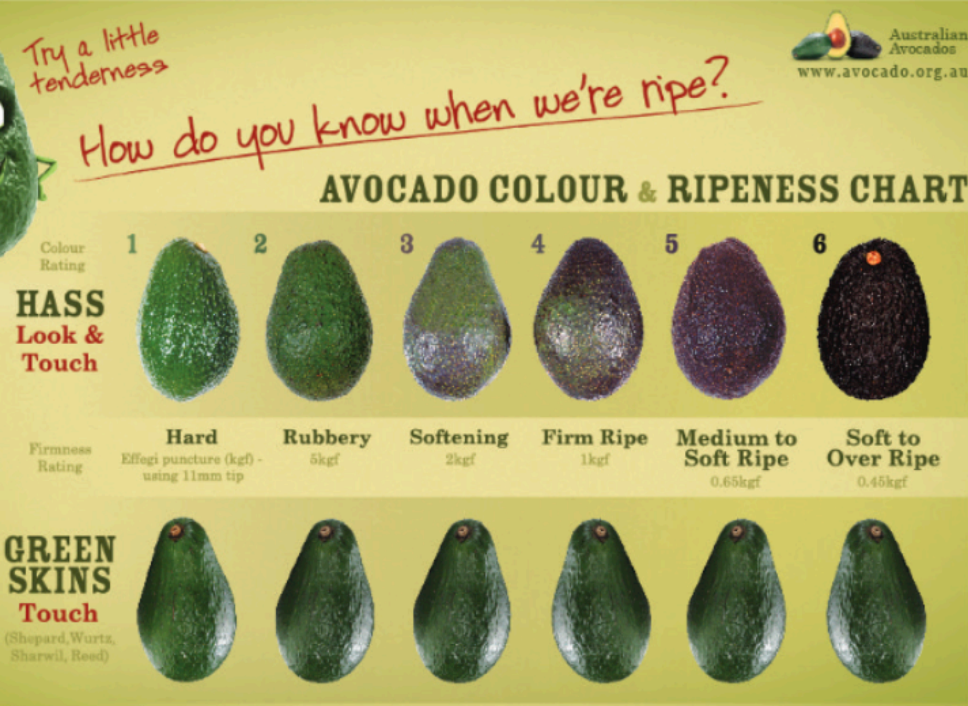
Course Staff

- Two TAs, who will lead the discussion and help run the class.
  - Yingyu (Anna) Lin, PhD student in DSC.
  - Fatameh Asgarinejad, PhD student in DSC.
- Three undergrad tutors, Daniel, Vivian, and Yujia (Joy), who will hold office hours, grade assignments, and help run the class.
  - All previous students of DSC 40A eager to help!
- Read about them at **rodalbuyeh.github.io/dsc40a-su23/staff/**.

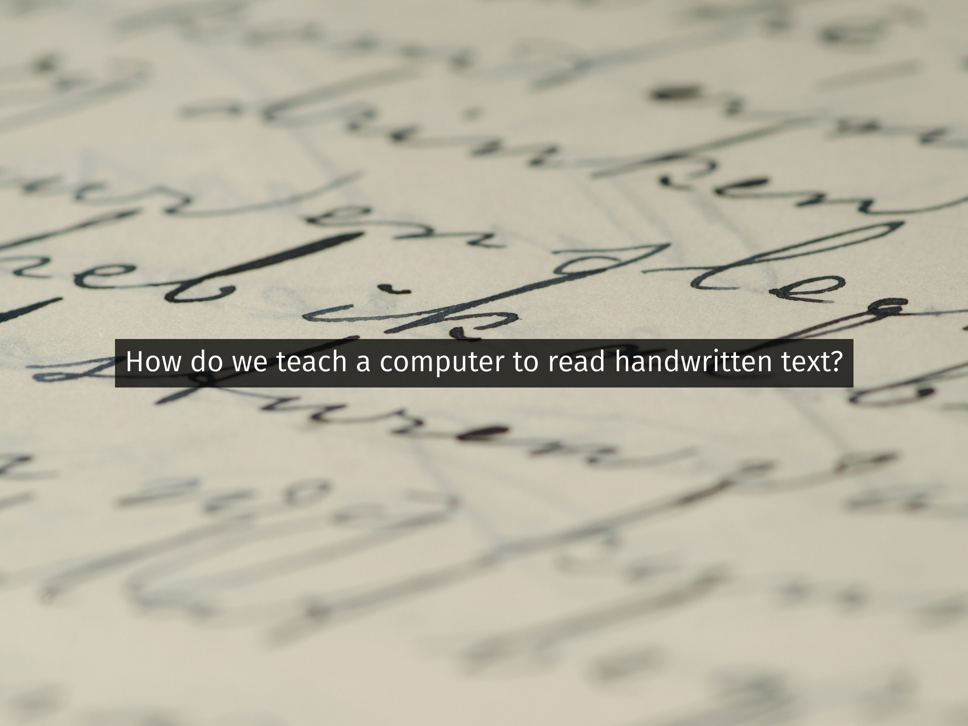
What is this course about?



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 (Modules 1 through 5)

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

Part 2: Probability (Modules 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, **rodalbuyeh.github.io/dsc40a-su23/**, contains all content. **Read the syllabus carefully!**

- We won't use Canvas.

- **Campuswire** will be used for announcements and communication. You can sign yourself up. **Ask questions here instead of email!**

- Fill out this **Welcome Survey**.

Lectures

- Lectures are held MW at 8am in RWAC 0115.
- Lecture slides will be posted on course website before class.
- Suggestion: don't write everything down! I'll post my annotated slides after class.
- Value of lecture: **interaction** and **discussion**.

Discussion

- Discussions on Weds at 12pm in RWAC 0115.
- Discussion will be used primarily for **groupwork**.
  - 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 95% credit. You may not work alone.
  - Value of attending: **TA/tutor support**.
- Submit groupwork to Gradescope by **11:59pm Fri**.
  - Only one group member should submit and add the other group members.

Assessments and exams

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

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

- **Exams**: One midterm and a two-part final exam, which can redeem low scores on the midterms. Exams are Wednesday, July 19 during lecture, and Friday, August 4th at 8am (location TBD).

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.
  - Rod has office hours Tuesdays and Thursdays at 8am via Zoom, open drop ins.
- **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:

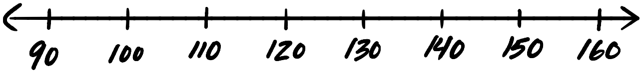
What is good/bad, intuitively?

- The data:

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

- Consider these hypotheses:

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



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.

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

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.

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

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, $y_1, y_2, \dots, y_n$.
- 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**: we solve this math problem.