Lecture 1 – Introduction, Learning From Data

slides at
dsc40a.com

DSC 40A, Fall 2021 @ UC San Diego

Suraj Rampure, 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!

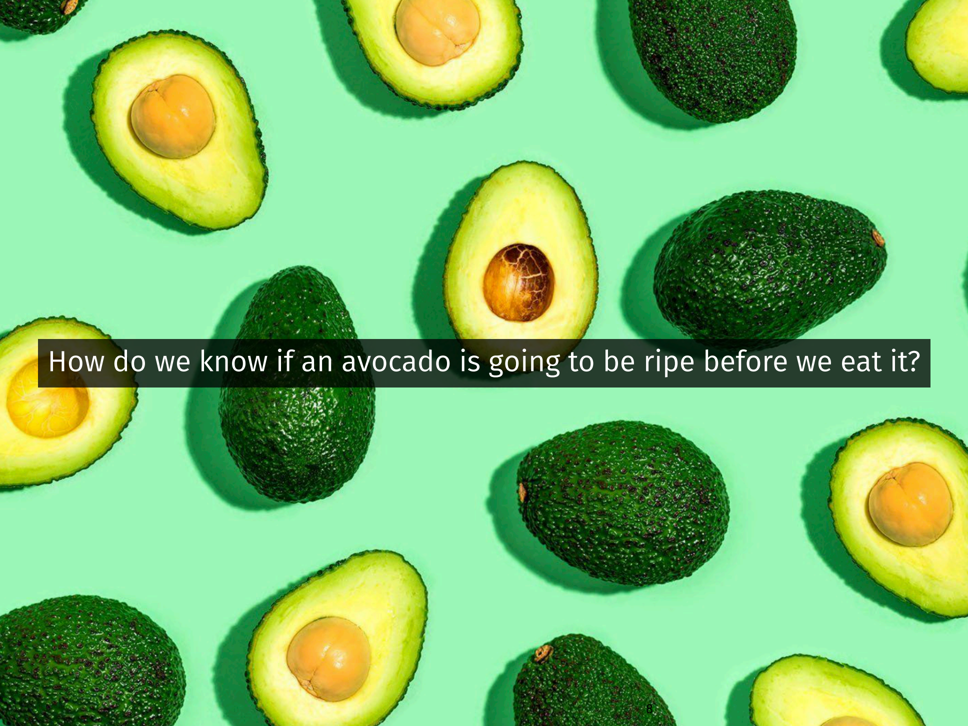
Suraj Rampure (call me Suraj)

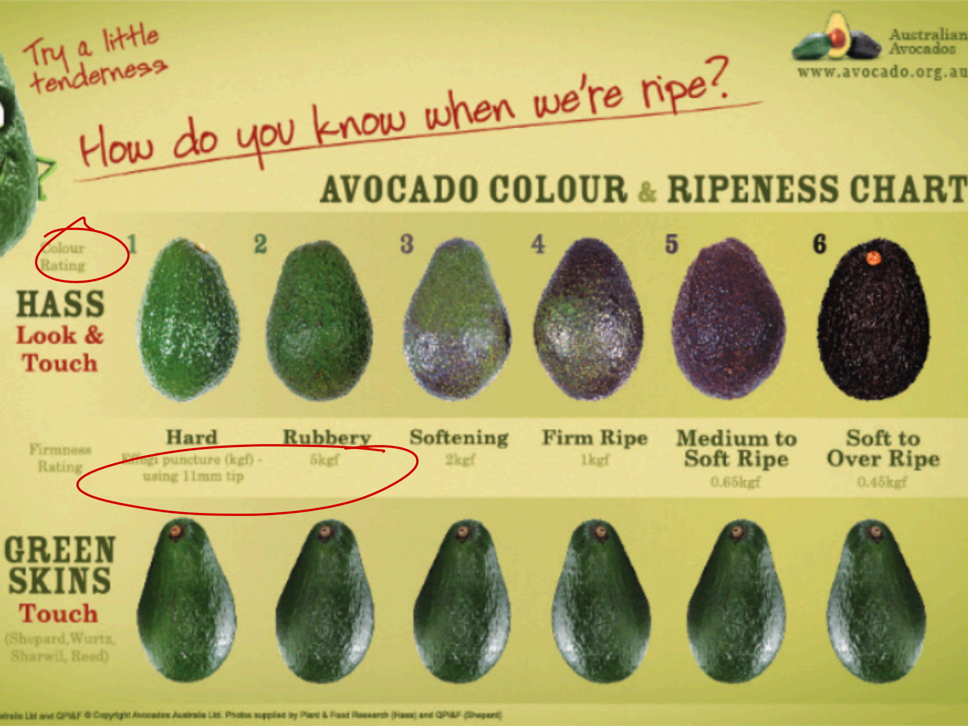
- Originally from Windsor, ON, Canada.
- BS and MS, Electrical Engineering and Computer Sciences, UC Berkeley. Taught and developed multiple DS and CS courses while there.
- DSC Lecturer at UC San Diego starting this fall. (Also teaching DSC 10.)
- For fun: watching basketball, traveling, watching TikTok, FaceTiming my dog,...

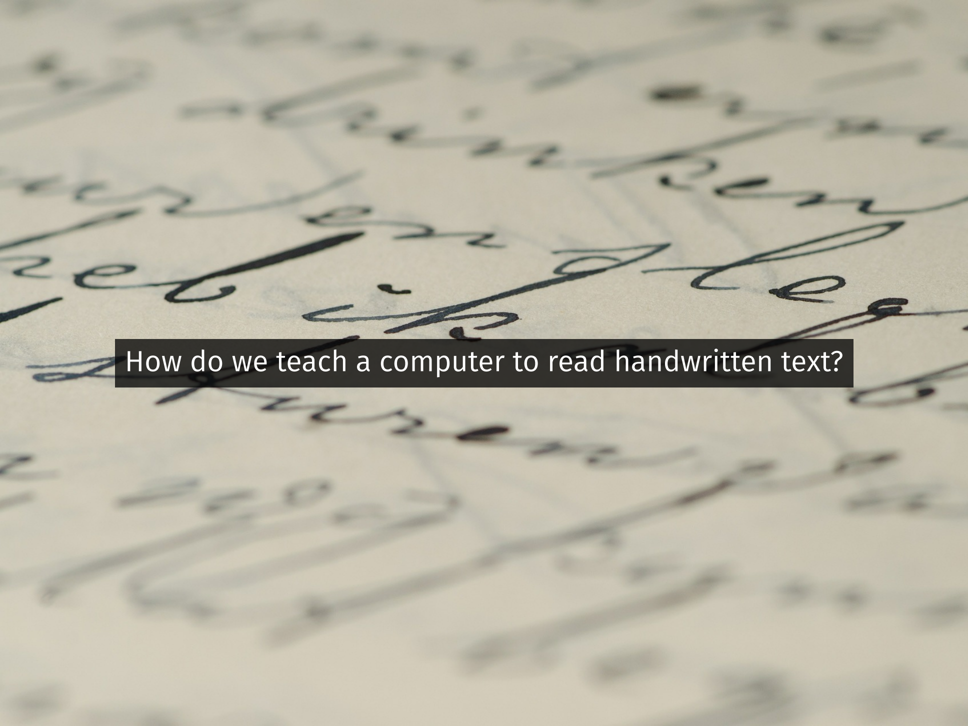
Say hey to course staff!

- 1 TA, who will teach discussion and help run the class.
  - Harpreet Singh, a MS student in CSE.
- 6 tutors, who will hold OH, grade assignments, and help run the class.
  - Jianming Geng, Yujian (Ken) He, Shiv Sakthivel, Aryaman Sinha, Luning Yang, and Sheng Yang.
  - 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
How do you know when we're ripe?
Australian Avocados
www.avocado.org.au
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)
...stralia Ltd and QPI&F © Copyright Avocados Australia Ltd. Photos supplied by Plant & Food Research (Hass) and QPI&F (Shepard)


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.

Theoretical Foundations of Data Science (1)

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.
- **Canvas** will only be used to store lecture/discussion recordings.

Lectures

- Tu/Th 11AM-12:20PM, Center Hall 109. No attendance required; recordings posted.
- Content in the first few weeks will closely follow readings.
- Lecture slides will be posted before class.
- Suggestion: don't write everything down! I'll write definitions, proofs, etc. on the slides.

  hey DSC 40A!

- Value of lecture: **interaction** and **discussion**.

Discussion

- **Discussion**: W 6-6:50PM, Zoom (links on Calendar).
  - Come to work on problems in small groups (”groupwork”) of 2-4.
  - Attendance is highly recommended but not required, however you **must** work on the groupwork problems in a group (whether that’s in discussion or on your own time).
- Groupwork problems must be submitted to Gradescope by **Thursdays 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 **Mondays at 11:59pm** on Gradescope. Worth 40% of your grade.
- **Surveys**: Feedback and conceptual surveys that accompany homeworks. Worth 5% of your grade.
- **Groupworks**: Due **Thursdays at 11:59pm**. Worth 10% of your grade.
- **Midterm Exam**: Th 10/21, 11AM-12:20PM. Remote and synchronous. Worth 15% of your grade.*
- **Final Exam**: W 12/8, 11:30AM-2:30PM. Remote and synchronous. Worth 30% of your grade.*

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, groupwork, and survey score.

- **Exam redemption policy***:
  - The Final Exam will contain a ”Midterm” section.

  - If you do better on the “Midterm” section of the Final Exam than you did on the original Midterm Exam, your score on the “Midterm” section will replace your original Midterm Exam score.

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. Put yourself on the queue at autograder.ucsd.edu (”The Autograder”). not an autograder!
  - Suraj will have one in-person OH (location TBD) and one remote OH per week.
- **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.

  prediction = actual salary

  $$\text{absolute error} = |\underbrace{(\text{actual future salary})} - \underbrace{\text{prediction}}|$$

  |prediction − actual|

- Then, our goal becomes to **find the prediction with the smallest possible absolute error**.

- There's a problem with this:

  we don't know actual future salary!

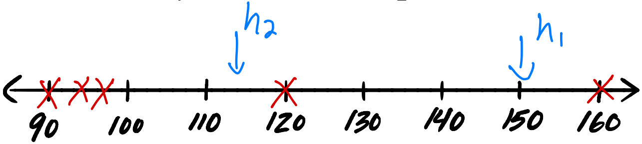
What is good/bad, intuitively?

- The data:

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

- Consider these hypotheses: guess, prediction

  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.

absolute error $= |\text{actual} - \text{predicted}|$

salary	absolute error of h_1
$\|90{,}000 - 150{,}000\|$	60,000
94,000	56,000
96,000	54,000
120,000	30,000
$\|160{,}000 - 150{,}000\|$	10,000

sum of absolute errors: 210,000

mean absolute error: 42,000

on average, prediction was off by 42,000

Quantifying our intuition

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

salary	absolute error of h_2
$\lvert 90{,}000 - 115{,}000 \rvert$	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?

- Supply/demand of data scientist roles might change salaries
- Inflation!

Which is better: the mean or median?

- Recall:

  mean = 112,000 median = 96,000

  mean of abs errors

- We can calculate the mean absolute error of each:

  mean of data! 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.

$h^* =$ the best h

- This is an **optimization problem**.

minimize MAE

$h \in \mathbb{R}^+$

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:

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

function of h

"the mean absolute error of h"

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.

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

Discussion Question

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

50 is a worse prediction
$\implies$ MAE is higher!

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:

$$R(h) = \frac{1}{n}\left(|y_1 - h| + |y_2 - h| + \cdots + |y_n - h|\right)$$

- Or, using **summation notation**:

$$R(h) = \frac{1}{n}\sum_{i=1}^{n} |y_i - h|$$

The best prediction

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

How?

$\Rightarrow$ calculus

$$R(h) = \frac{1}{n} \sum_{i=1}^{n} |y_i - h|$$

function of h only!

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.