

# 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~~ ECE.
- ▶ 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/](https://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

target €

How do you know when we're ripe?

Independent var

## 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**

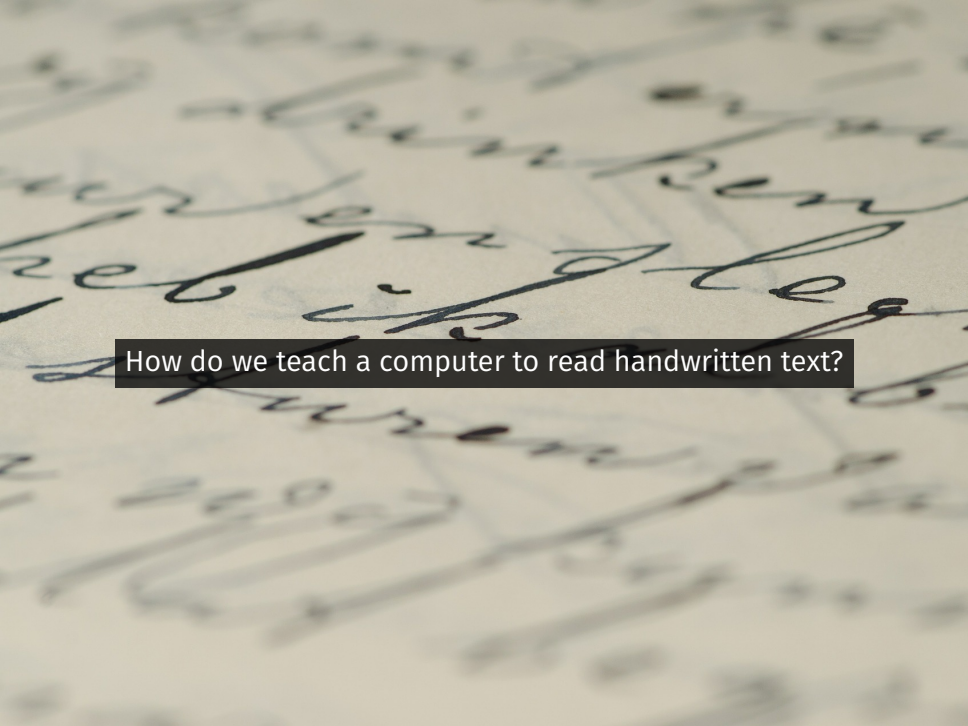
0.45kgf

**GREEN  
SKIN**  
Touch


(Shepard, Wurtz,  
Sharwil, Reed)





A close-up, slightly blurred image of a document with handwritten text in a cursive script. The ink is dark, and the paper has a light, aged tone. The text is written in a fluid, connected style, typical of 18th or 19th-century handwriting. A dark, semi-transparent rectangular box is overlaid on the center of the image, containing white text.

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.

← 1. closed form solution  
2. algorithmic approach

# 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.

↳ precision and recall

# 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/](https://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~~ <sup>Fridays</sup> 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

*is Data Scientist*

- ▶ The **mean**:

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

- ▶ The **median**:

90,000    94,000    96,000    120,000    ~~160,000~~ *900,000*

↑

- ▶ 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.

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

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

When making prediction, salary is unknown.

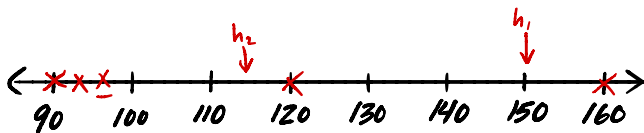
# What is good/bad, intuitively?

- ▶ The data:

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

- ▶ Consider these hypotheses:

$$h_1 = 150,000 \quad h_2 = 115,000$$



## 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: <u>210,000</u>	
<b>mean absolute error:</b> <u>42,000</u>	

## 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:

$$h_1 : 42,000 \quad h_2 : 23,000$$

- ▶ 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:

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

## A formula for the mean absolute error

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

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

## A formula for the mean absolute error

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

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

## A formula for the mean absolute error

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

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

*Handwritten notes in red: "50-100" above the first term, and "90k 94k 96k 120k 160k" above the terms in the absolute value expression.*

### 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:

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

---

- ▶ 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  $R(h)$ .

## 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.