

# Lecture 1 – Learning From Data



DSC 40A, Spring 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!

## **Janine Tiefenbruck (call me Janine)**

- ▶ Grew up in NJ. Undergrad Math/CS major at Loyola, MD. PhD in Math (Combinatorics) at UCSD.
- ▶ Lecturer in DSC for 5 years. Developed original versions of DSC 10, 40A, 40B. Mostly teaching 10 and 40A these days.
- ▶ For fun: board games, baking, hiking, crafts

## Course Staff

- ▶ 1 TA, who will lead the discussion and help run the class.
  - ▶ Dylan Stockard, a MS student in DSC.
- ▶ Undergrad tutors, who will hold office hours, grade assignments, and help run the class.
  - ▶ Tunan Li, Pallavi Prabhu, Pranav Rebala, Harshi Saha, Yutian (Skylar) Shi, Aryaman Sinha, Zelong (Alan) Wang, Benjamin Xue, Luran (Lauren) Zhang
  - ▶ All previous students of DSC 40A eager to help!
- ▶ Read about them at [dsc40a.com/staff](https://dsc40a.com/staff).

**What is this course about?**

A top-down view of several avocados scattered on a light green background. Some are whole, showing their dark green, bumpy skin. Others are sliced in half, revealing the bright green flesh and the large, brown, oval-shaped pit. The lighting is bright and even, casting soft shadows.

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?

## 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**

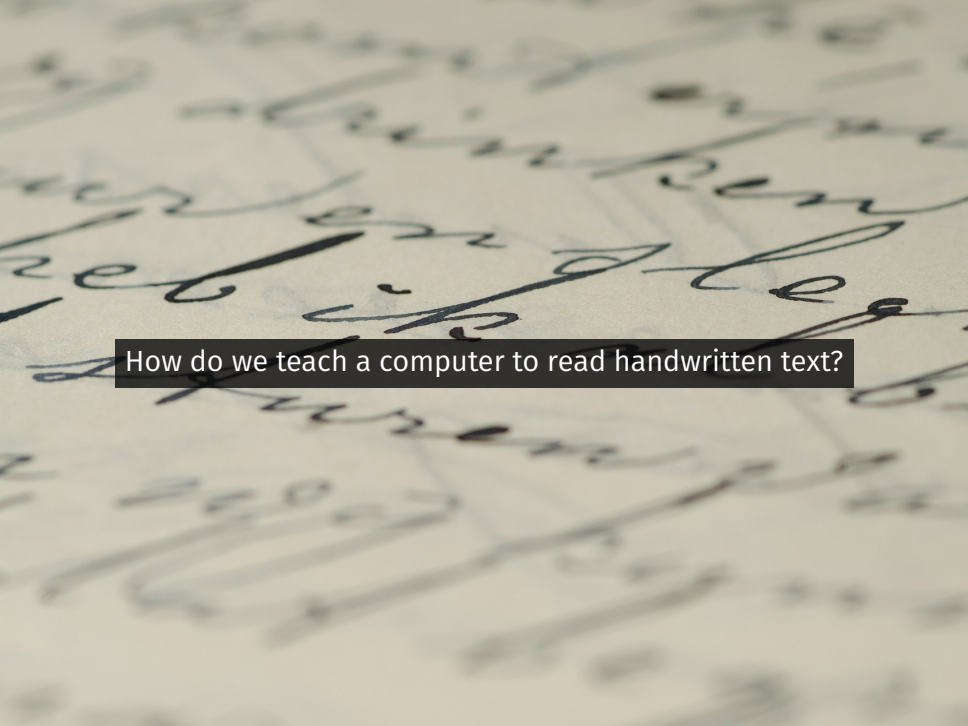
0.45kgf

**GREEN  
SKINS**  
Touch


(Shepard, Wurtz,  
Sharwil, Reed)





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 cursive handwriting. A dark rectangular box is superimposed over 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.

# 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](https://dsc40a.com), 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 MWF at 10am and 11am in WLH 2204.
- ▶ 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 7pm and 8pm in FAH 1101.
- ▶ 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 Weds**.
  - ▶ 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 **Wednesdays 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, May 5 during lecture, Monday, June 5 during lecture, and Saturday, June 10 from 9am to 11am.

# 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.
  - ▶ Janine has office hours today 12:30-2:30 if you want to stop by!
- ▶ **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?**



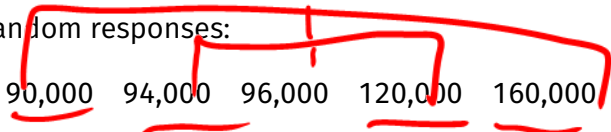
How do we predict a future data scientist's salary?

## 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?

median, ranges, location

## Some common approaches

- ▶ The **mean**:

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

- ▶ The **median**:

90,000   94,000   96,000   120,000   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. =

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

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

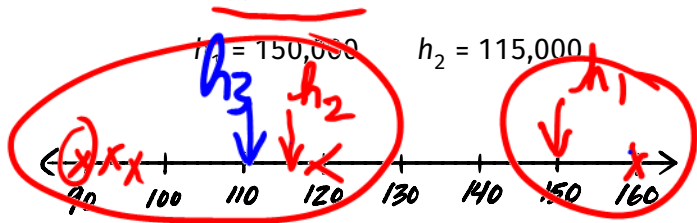
you don't know actual future salary

# What is good/bad, intuitively?

- ▶ The data:

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

- ▶ Consider these hypotheses:



## Discussion Question

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

errors generally smaller  
closer to data  
most

## 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

prediction

Summarize

## Quantifying our intuition

- ▶ Now suppose we had predicted  $h_2 = \underline{115,000}$

salary	absolute error of $h_2$
90,000	25,000
94,000	21,000
96,000	19,000
120,000	5,000
160,000	<del>45,000</del>

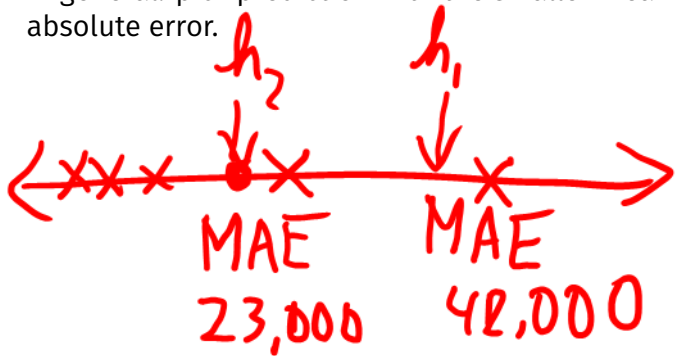
sum of absolute errors: 115,000  
mean absolute error: 23,000

## 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?

inflation  
job supply/demand

## Which is better: the mean or median?

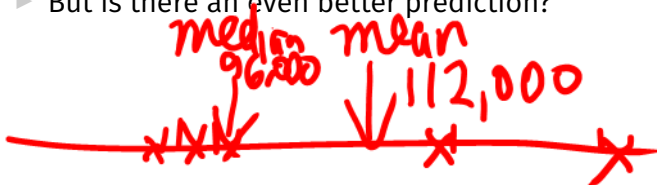
- ▶ Recall:

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**.

the best  
of all  
possible  
 $h$ 's



## 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:

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

minimize  $R = \text{risk}$

## A formula for the mean absolute error

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

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

## A formula for the mean absolute error

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

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

## A formula for the mean absolute error

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

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

~~- 111,996.8584...~~

$R=50$   
 $R=100$

↓ ↓

### Discussion Question

Without doing any calculations, which is correct?

A.  $R(50) < R(100)$

B.  $R(50) = R(100)$

C.  $R(50) > R(100)$

$R$  bigger for worse prediction

## 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:

$n$  = how many data values

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

- ▶ Or, using **summation notation**:

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

note:  $|y_i - h| = |h - y_i|$

## 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.