

Table of Contents

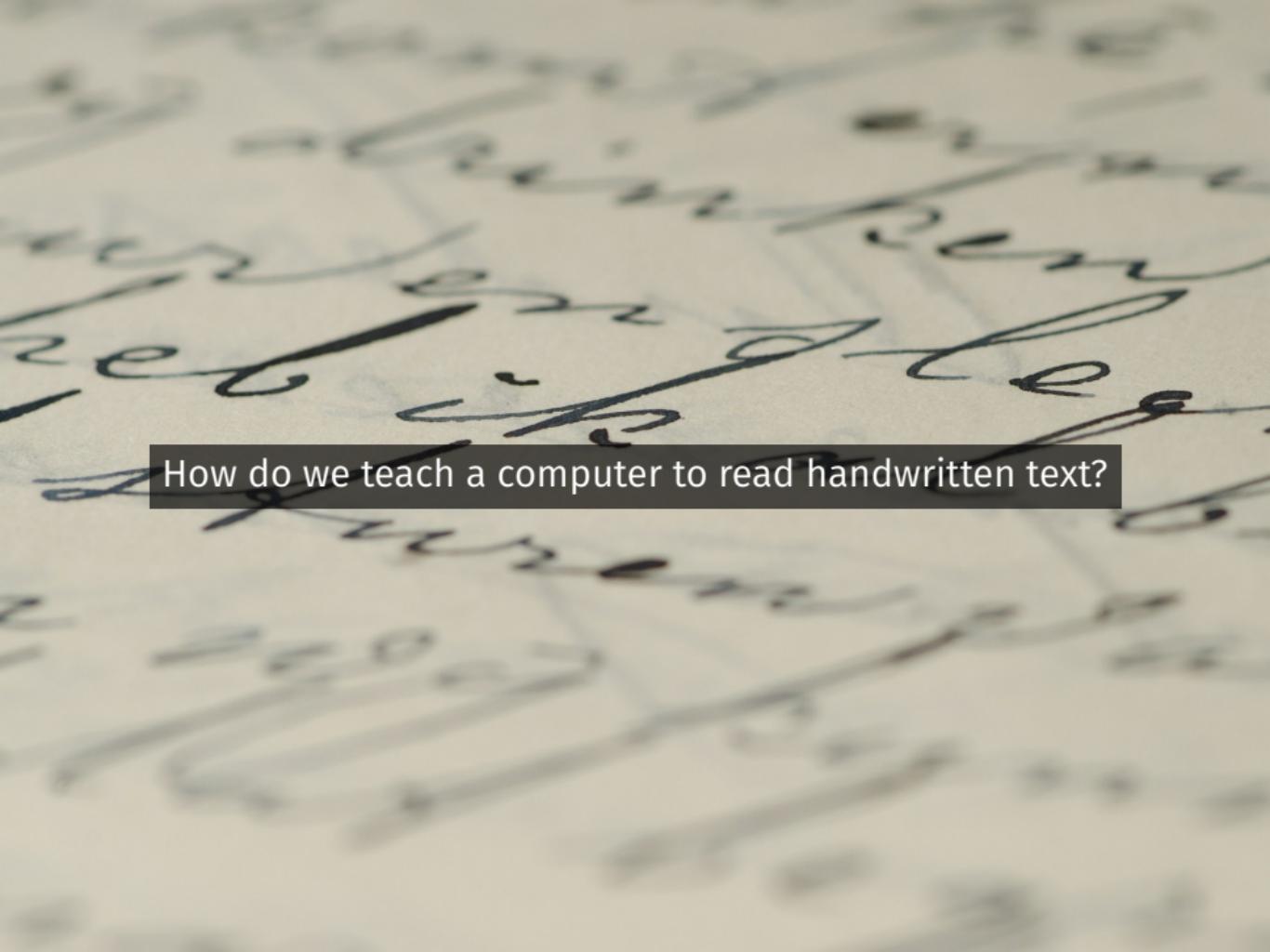
- Lecture 01 - intro
- Lecture 02 - median
- Lecture 03 - mean
- Lecture 04 - gradient - descent
- Lecture 05 - convexity
- Lecture 06 - regression - I
- Lecture 07 - regression - II
- Lecture 08 - linear - algebra
- Lecture 09 - gradient
- Lecture 10 - multiple - regression
- Lecture 11 - perceptron
- Lecture 12 - probability
- Lecture 13 - combinatorics
- Lecture 14 - conditional
- Lecture 15 - independence
- Lecture 16 - naive - bayes - I
- Lecture 17 - naive - bayes - II

DSC 40A

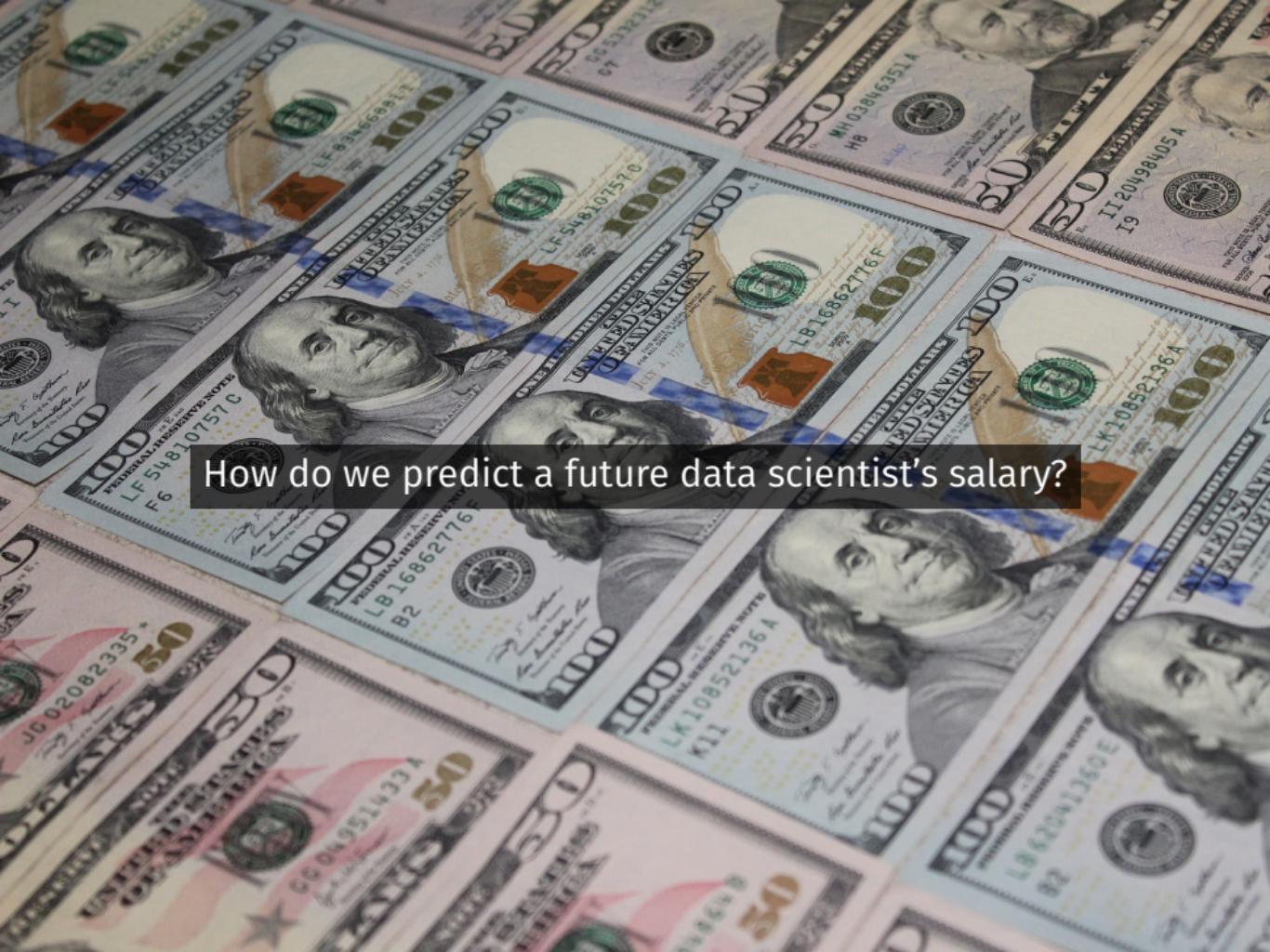
Theoretical Foundations of Data Science I

A photograph taken from space, showing the curvature of the Earth's horizon against a dark void. The atmosphere is filled with various types of clouds, from wispy cirrus to more dense cumulus and cumulonimbus formations. The lighting suggests the sun is low on the horizon, casting long shadows and highlighting the texture of the clouds.

How do we decide who needs the most help before a hurricane hits?

A close-up, slightly blurred photograph of handwritten cursive text on lined paper. The text is written in black ink and appears to be in English. A dark rectangular box is overlaid on the image, containing the question "How do we teach a computer to read handwritten 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 into a math problem.
- 2) Solve that problem.

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

www.dsc40a.com



DSC 40A
Lecture 01
Learning via Optimization, pt I.

Lecture Format

- ▶ Lecture slides will be posted before class.
 - ▶ Suggestion: don't write everything down!
 - ▶ I'll write definitions, proofs, etc. on the slides.
-
-
-

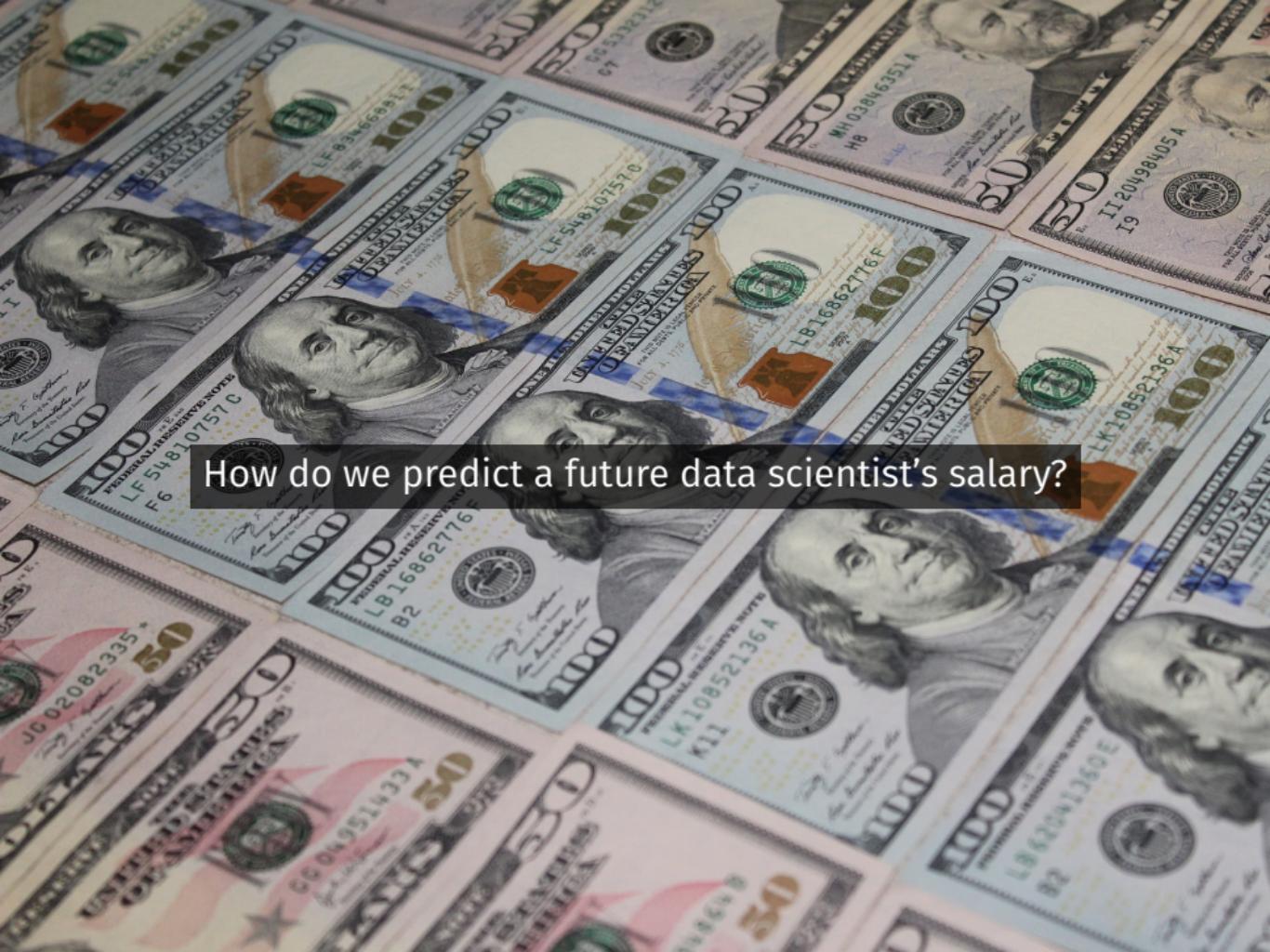
- ▶ Value of lecture: **interaction** and **discussion**.

Today's Question

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

Recommended Reading

Chapter 01, Section 01



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

Learning from Data

- ▶ Idea: ask a few data scientists about their salary.
- ▶ StackOverflow survey.
- ▶ 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?

The Mean and the Median

- ▶ The **mean**:

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

- ▶ The **median**:

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

- ▶ Which is better? Are these good ways of predicting future salary?

Quantifying goodness/badness of a prediction

- ▶ The **error**: distance from prediction to the right answer.

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

- ▶ Find prediction with smallest possible error.
 - ▶ There's a problem with this:
-

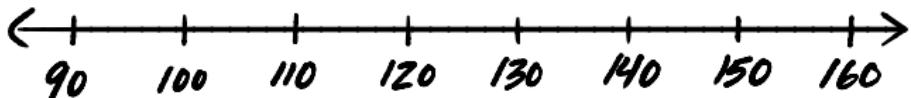
What is good/bad, intuitively?

- ▶ The data:

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

- ▶ Consider these hypotheses:

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



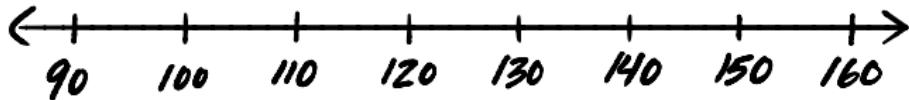
What is good/bad, intuitively?

- ▶ 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	error of h_1
90,000	60,000
94,000	56,000
96,000	54,000
120,000	30,000
160,000	10,000

total error: 210,000
mean error: 42,000

Quantifying our intuition

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

salary	error of h_2
90,000	25,000
94,000	21,000
96,000	19,000
120,000	5,000
160,000	45,000

total error: 115,000
mean error: 23,000

Mean Errors

- ▶ Mean 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 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 better: the mean or median?

- ▶ Recall:

mean = 112,000 median = 96,000

- ▶ We can calculate the average 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 error.
- ▶ This is an **optimization problem**.

Status Update

- ▶ We started with the learning problem:

Given salary data, predict your future salary.

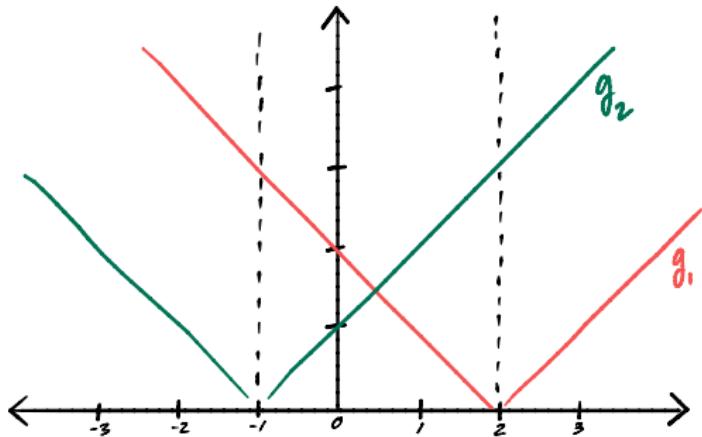
- ▶ We turned it into this problem:

Find a prediction h^ which has smallest mean error on the data.*

- ▶ We have turned the problem of learning into a specific type of math problem: an **optimization problem**.

What's Left?

- ▶ We need to solve this math problem.
- ▶ Next time: math.



DSC 40A
Lecture 02
Learning via Optimization, pt II

Announcements

- ▶ Extension students: join Gradescope/Campuswire using codes found on www.dsc40a.com
- ▶ Need iClicker starting next week for tokens.

Last Time

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

Last Time

- ▶ What will be your future salary?

- ▶ Collect data:

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

- ▶ Could use the **mean** or the **median** as a prediction.
- ▶ But why?
- ▶ What is the best prediction?

Last Time: The Mean Error of a Prediction

- ▶ Suppose we predicted a future salary of $h_1 = 150,000$ before collecting data.

salary	error of h_1
90,000	60,000
94,000	56,000
96,000	54,000
120,000	30,000
160,000	10,000
total error: 210,000	
mean error: 42,000	

- ▶ A good prediction is one with small mean error.

Last Time: The Best Prediction

- ▶ Any (non-negative) number is a valid prediction.
- ▶ Goal: out of all possible predictions, find the prediction h^* with the smallest **mean error**.
- ▶ This is an **optimization problem**.

Today

We've turned learning into an **optimization problem**.
How do we solve it?

A Formula for the Mean Error

- ▶ We have data:

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

- ▶ Suppose our prediction is h .
- ▶ The **mean error** of our prediction is:

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

A Formula for the Mean Error

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

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

A Formula for the Mean Error

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

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

A Formula for the Mean Error

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

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

A General Formula for the Mean Error

- ▶ Suppose we collect n salaries, y_1, y_2, \dots, y_n .
 - ▶ The mean 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 $R(h)$.

Discussion Question

Can we use calculus to minimize R ?

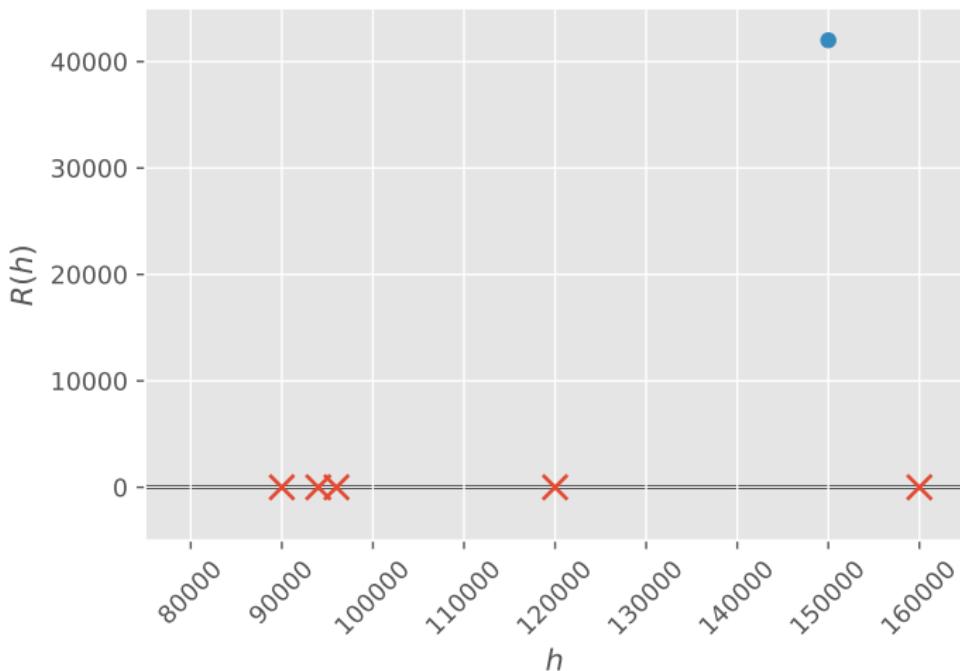
Minimizing with Calculus

- ▶ Calculus: take derivative, set equal to zero, solve.

Uh oh

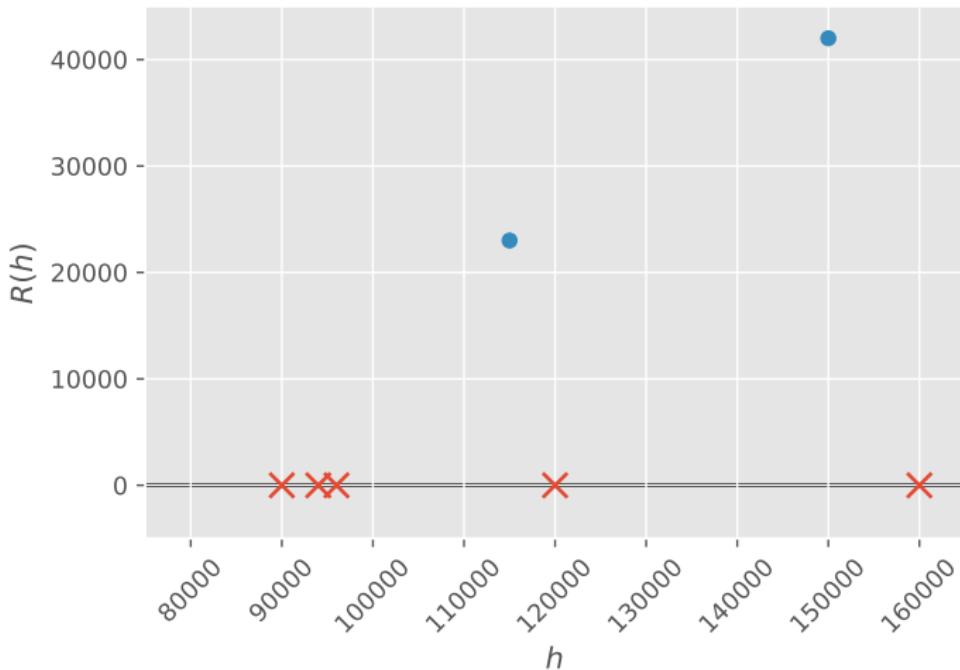
- ▶ R is **not differentiable**.
- ▶ We can't use calculus to minimize it.
- ▶ Let's try plotting $R(h)$

Plotting the Mean Error



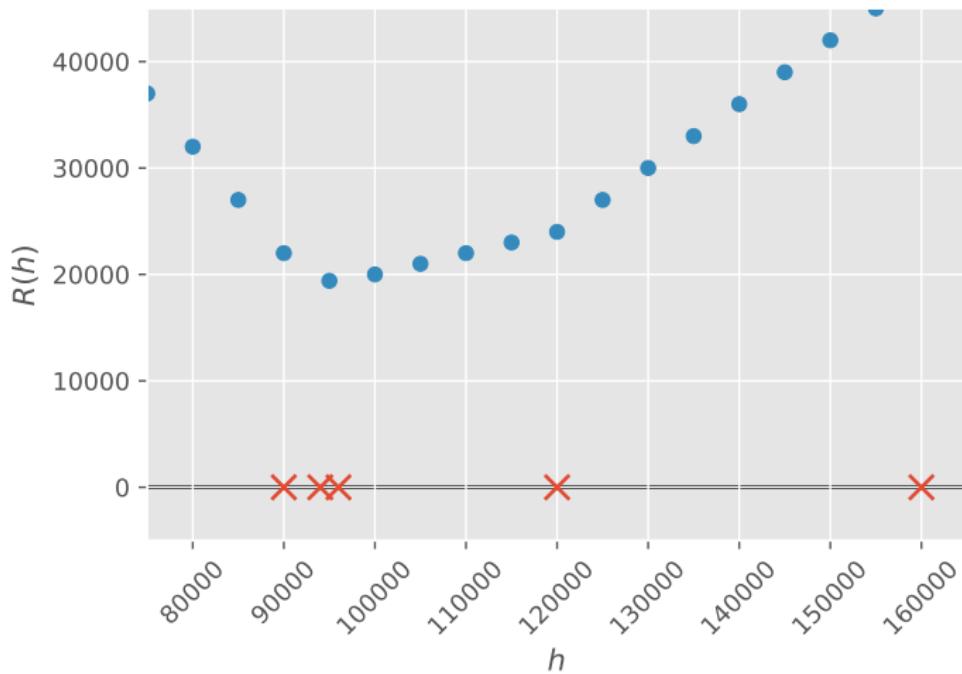
Recall: $R(150,000) = 42,000$

Plotting the Mean Error

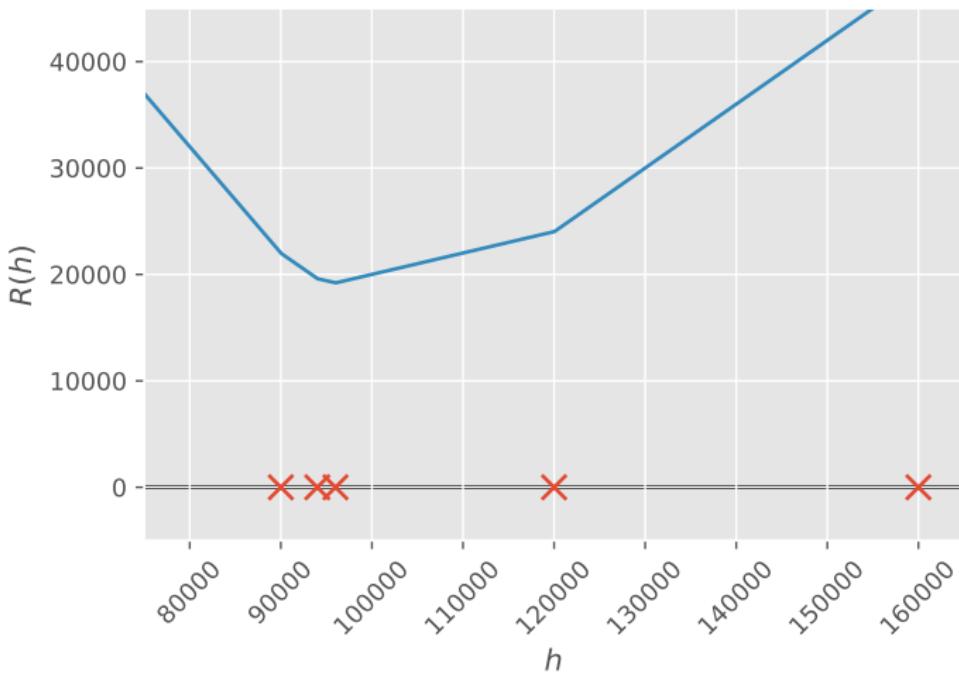


Recall: $R(115,000) = 23,000$

Plotting the Mean Error



Plotting the Mean Error

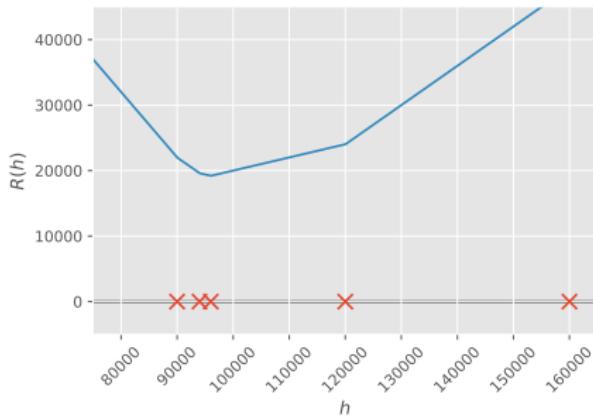


Discussion Question

A local minimum occurs when the slope goes from _____ . Select all that apply.

- A) positive to negative
- B) negative to positive
- C) positive to zero.
- D) negative to zero.
- E) zero to zero.

Goal



- ▶ Find where slope of R goes from negative to non-negative.
- ▶ Want a formula for the slope of R at h .

Sums of Linear Functions

- ▶ Let $f_1(x) = 3x + 2$
- ▶ Let $f_2(x) = 5x + 1$
- ▶ What is the slope of $f(x) = f_1(x) + f_2(x)$?

Sums of Absolute Values

- ▶ Let

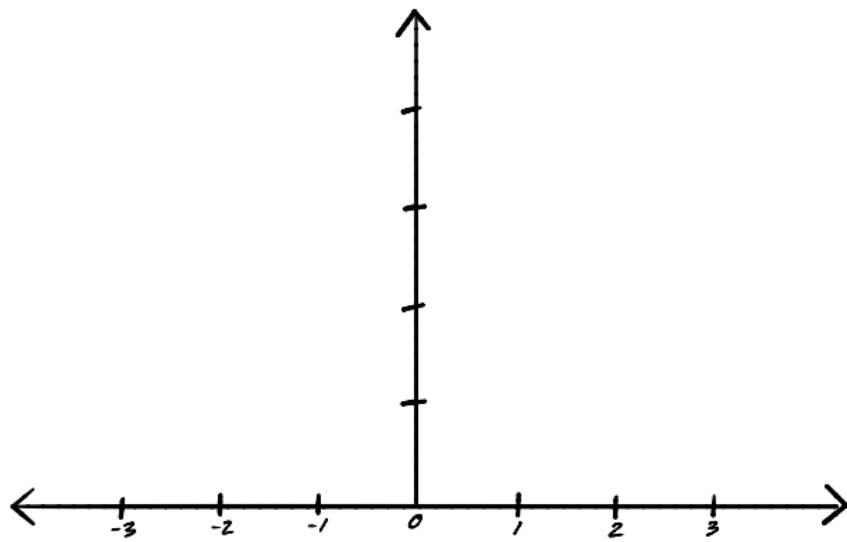
$$g_1(x) = |x - 2| \quad g_2(x) = |x + 1|$$

- ▶ Let $g(x) = g_1(x) + g_2(x)$.

Discussion Question

What is the slope of g at $x = 1$?

Answer



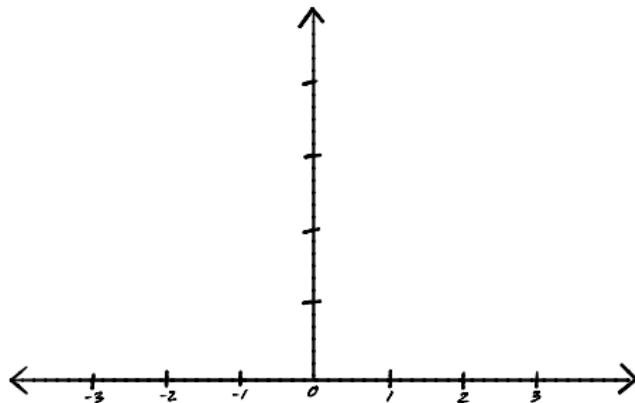
Sums of More Absolute Values

- ▶ Let $y_1 < y_2 < y_3$

$$h_1(x) = |x - y_1| \quad h_2(x) = |x - y_2| \quad h_3(x) = |x - y_3|$$

- ▶ Let $h(x) = h_1(x) + h_2(x) + h_3(x)$.
- ▶ The slope changes at y_1, y_2, y_3 .

Sums of More Absolute Values



- ▶ Slope when $x < y_1$:
- ▶ Slope when $y_1 < x < y_2$:
- ▶ Slope when $y_2 < x < y_3$:
- ▶ Slope $x > y_3$:

Slope at $x = (\# \text{ of } y_i's \text{ } \underline{\text{---}} \text{ } x) - (\# \text{ of } y_i's \text{ } \underline{\text{---}} \text{ } x)$

The Slope of Error Function

- ▶ R is the sum of absolute value functions (times $\frac{1}{n}$):

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

- ▶ So:

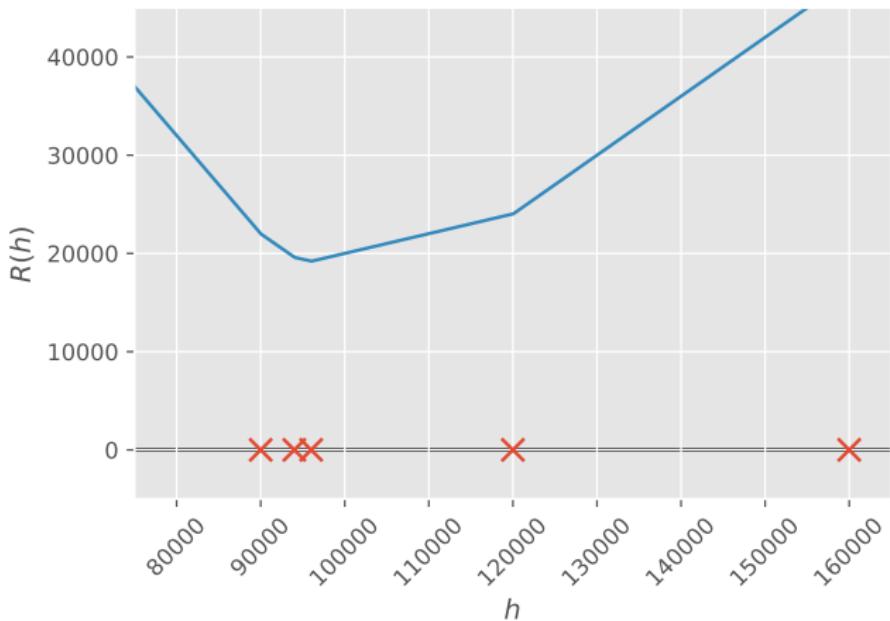
$$\text{Slope at } h = \frac{1}{n} \cdot [(\# \text{ of } y_i's \text{ } \leq \text{ } h) - (\# \text{ of } y_i's \text{ } > \text{ } h)]$$

Discussion Question

Suppose that n is odd. At what value of h does the slope go from negative to positive?

- A) $h = \text{mean of } y_1, \dots, y_n$
- B) $h = \text{median of } y_1, \dots, y_n$
- C) $h = \text{mode of } y_1, \dots, y_n$

Where the Slope's Sign Changes



The Median Minimizes the Mean Error

- ▶ Our problem was: find h^* which minimizes the mean error, $R(h) = \frac{1}{n} \sum_{i=1}^n |y_i - h|$.
- ▶ The answer is: $\text{Median}(y_1, \dots, y_n)$.
- ▶ The **best prediction**¹ is the **median**.

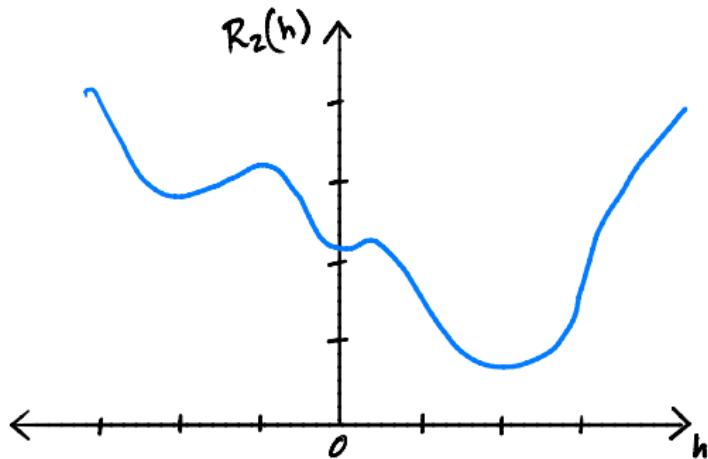
¹in terms of mean error

Status Update

- ▶ Last time, we turned predicting salary into a math problem: minimize the mean error.
- ▶ Today: we solved it. The **median** minimizes the mean error.

What's Left?

- ▶ We did all this because $R(h)$ isn't differentiable.
- ▶ What if we tried to minimize a *different* measure of error that *is* differentiable?



DSC 40A
Lecture 03
Learning via Optimization, pt III

Announcements

- ▶ First homework posted; due Friday at 5:00 pm.
- ▶ First discussion tonight! Will help with homework.
- ▶ My office hours: Tuesday, 5:30 – 6:30 pm and Thursday, 12 – 1 pm

Last Time

- ▶ To predict future salary:
 - ▶ Gather salaries y_1, y_2, \dots, y_n .
 - ▶ Find a prediction h^* which minimizes the **mean error**:

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

- ▶ We saw that $R(h)$ is minimized by $\text{Median}(y_1, \dots, y_n)$.
- ▶ We turned learning into a math problem and solved it.

Two things we don't like

1. **Minimizing** the mean error wasn't so easy.
2. Actually **computing** the median isn't so easy, either.

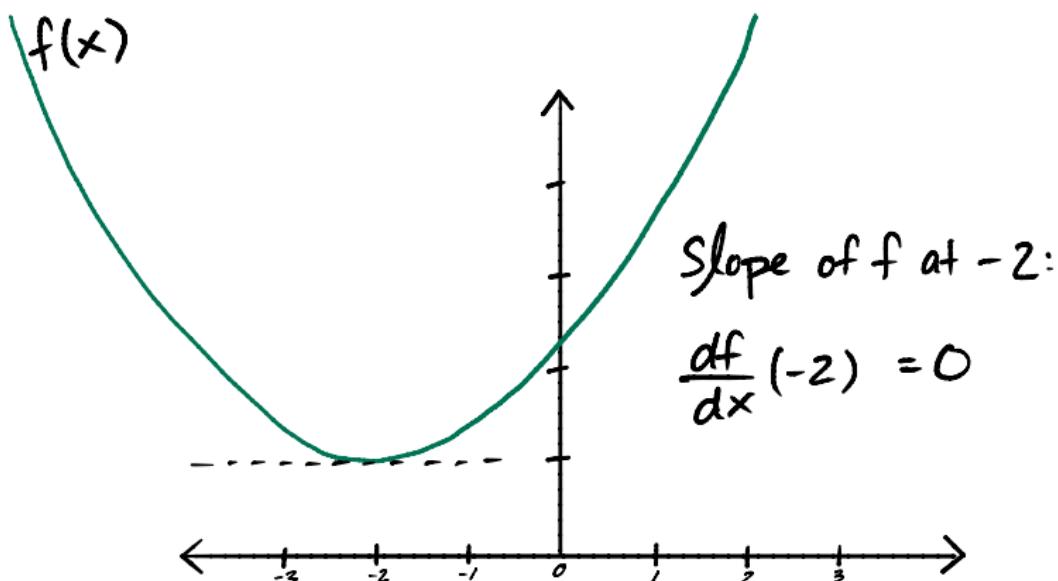
Today

Is there another way to measure the quality of a prediction that avoids these problems?

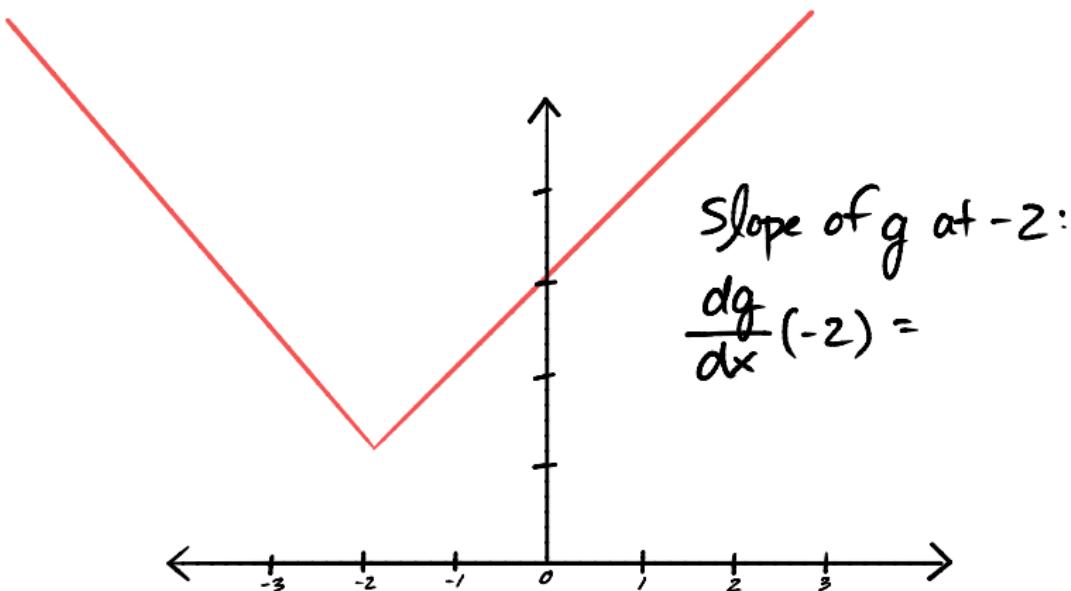
Minimizing via calculus

- ▶ Strategy: take derivative, set to zero, solve.
- ▶ Finding where the slope is zero.
- ▶ Only works if the function is **differentiable**.

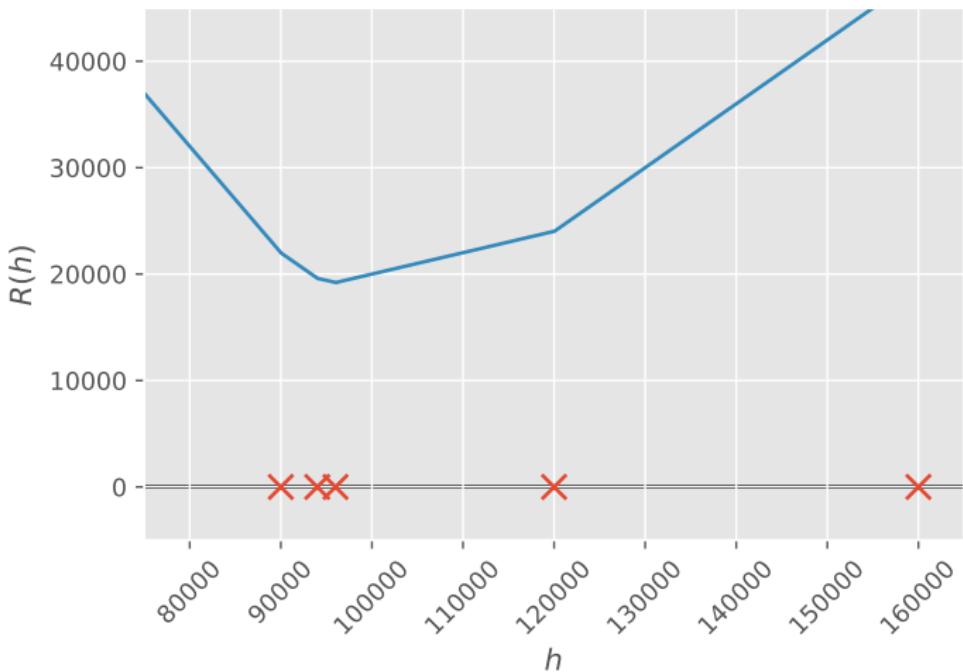
Example: differentiable



Example: not differentiable



The mean error is **not** differentiable



The mean error is not differentiable

$$\frac{dR}{dh}(h) = \frac{d}{dh} \left[\frac{1}{n} \sum_{i=1}^n |y_i - h| \right]$$

The core issue

- ▶ We can't compute $\frac{d}{dh} |y_i - h|$; it is **not differentiable**.
- ▶ Remember: $|y_i - h|$ measures how far h is from y_i .
- ▶ Is there something besides $|y_i - h|$ which:
 1. Measures how far h is from y_i ; *and*
 2. is **differentiable**?

The core issue

- ▶ We can't compute $\frac{d}{dh} |y_i - h|$; it is **not differentiable**.
- ▶ Remember: $|y_i - h|$ measures how far h is from y_i .
- ▶ Is there something besides $|y_i - h|$ which:
 1. Measures how far h is from y_i ; and
 2. is **differentiable**?

Discussion Question

Which of these would work?

a) $e^{|y_i - h|}$

b) $|y_i - h|^2$

c) $|y_i - h|^3$

d) $\cos(y_i - h)$

The Squared Error

- ▶ Let h be a prediction and y be the right answer. The **squared error** is:

$$|y - h|^2 = (y - h)^2$$

- ▶ Like error, measures how far h is from y .
- ▶ But unlike error, the squared error is **differentiable**:

$$\frac{d}{dh}(y - h)^2 =$$

The Mean Squared Error

- ▶ Suppose we predicted a future salary of $h_1 = 150,000$ before collecting data.

salary	error of h_1	squared error of h_1
90,000	60,000	$(60,000)^2$
94,000	56,000	$(56,000)^2$
96,000	54,000	$(54,000)^2$
120,000	30,000	$(30,000)^2$
160,000	10,000	$(10,000)^2$

total squared error: 1.0652×10^{10}

mean squared error: 2.13×10^9

- ▶ A good prediction is one with small **mean squared error**.

The Mean Squared Error

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

salary	error of h_2	squared error of h_2
90,000	25,000	$(25,000)^2$
94,000	21,000	$(21,000)^2$
96,000	19,000	$(19,000)^2$
120,000	5,000	$(5,000)^2$
160,000	45,000	$(45,000)^2$

total squared error: 3.47×10^9
mean squared error: 6.95×10^8

- A good prediction is one with small mean squared error.

The New Idea

- ▶ Make prediction by minimizing the **mean squared error**:

$$R_{\text{sq}}(h) = \frac{1}{n} \sum_{i=1}^n (y_i - h)^2$$

- ▶ Strategy: Take derivative, set to zero, solve for minimizer.

The New Idea

- ▶ Make prediction by minimizing the **mean squared error**:

$$R_{\text{sq}}(h) = \frac{1}{n} \sum_{i=1}^n (y_i - h)^2$$

- ▶ Strategy: Take derivative, set to zero, solve for minimizer.

Discussion Question

Which of these is dR_{sq}/dh ?

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

b) 0

c) $\sum_{i=1}^n y_i$

d) $\frac{2}{n} \sum_{i=1}^n (h - y_i)$

Solution

$$\frac{dR_{\text{sq}}}{dh} = \frac{d}{dh} \left[\frac{1}{n} \sum_{i=1}^n (y_i - h)^2 \right]$$

Set to zero and solve for minimizer

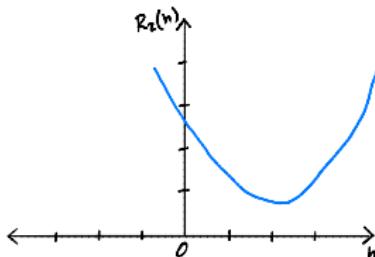
The mean minimizes the mean squared error

- ▶ That is:

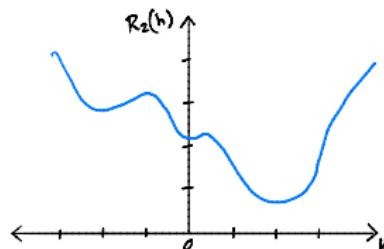
$$\arg \min_{h \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n (y_i - h)^2 = \text{Mean}(y_1, \dots, y_n)$$

Discussion Question

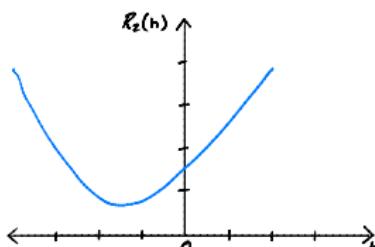
Suppose y_1, \dots, y_n are salaries. Which plot could be $R_{\text{sq}}(h)$?



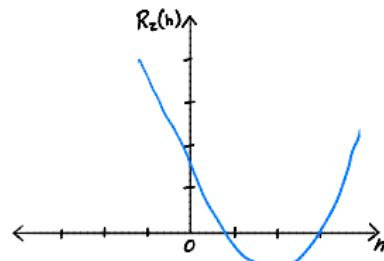
(a)



(b)



(c)



(d)

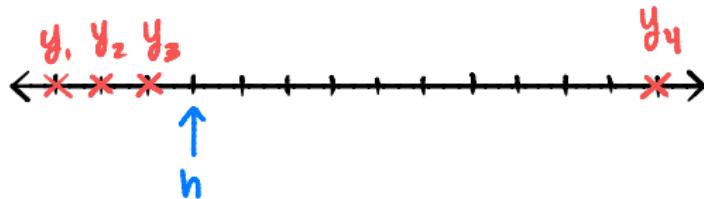
The mean is easy to compute

```
def mean(numbers):
    total = 0
    for number in numbers:
        total = total + number
    return total / len(numbers)
```

- ▶ Time complexity: $\Theta(n)$
- ▶ Median by sorting: $\Theta(n \log n)$
- ▶ But there's a $\Theta(n)$ way to find median: quickselect.
- ▶ DSC 40B.

Outliers

- The mean is quite **sensitive** to outliers.

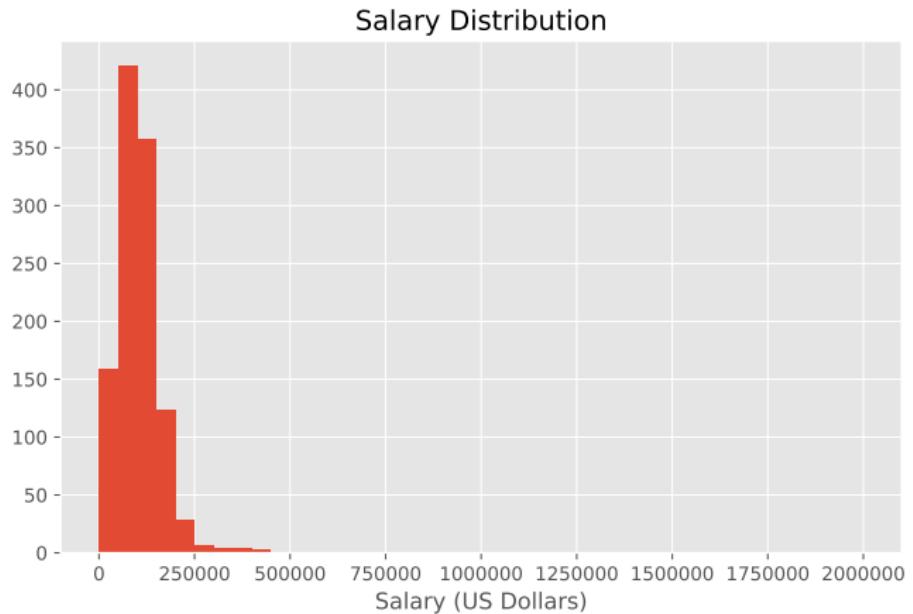


- $|y_4 - h|$ is 10 times as big as $|y_3 - h|$.
- $(y_4 - h)^2$ is 100 times as big as $(y_3 - h)^2$.
- Squared error can be dominated by outliers.

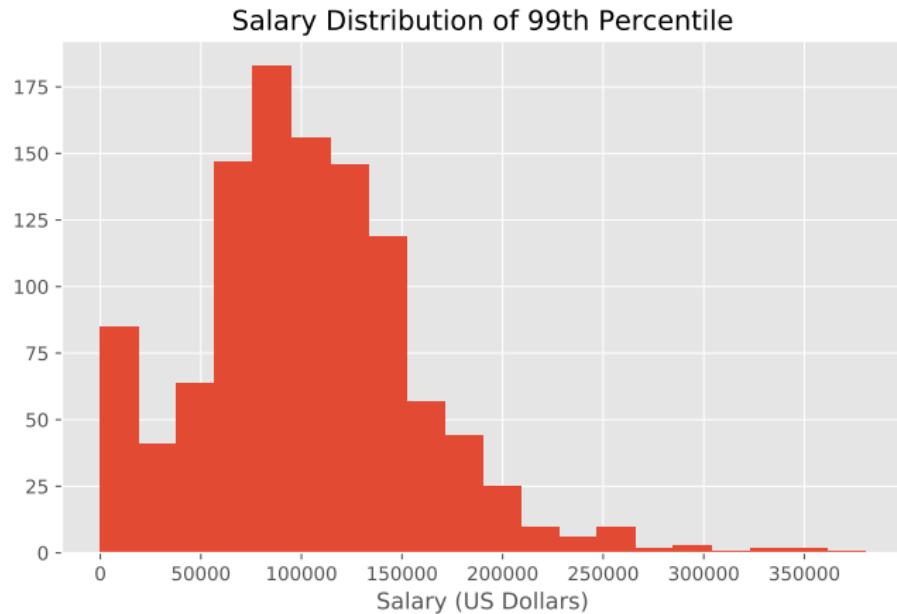
Example: Data Science Salaries

- ▶ Data set of 1121 self-reported data science salaries in the United States from the 2018 StackOverflow survey.
- ▶ Median = \$100,000
- ▶ Mean = \$111,032
- ▶ Max = \$2,000,000
- ▶ Min = \$52
- ▶ 95th Percentile: \$200,000

Example: Data Science Salaries



Example: Data Science Salaries



Example: Income Inequality

Average vs median income

Median and mean income between 2012 and 2014 in selected OECD countries, in USD; weighted by the currencies' respective purchasing power (PPP).

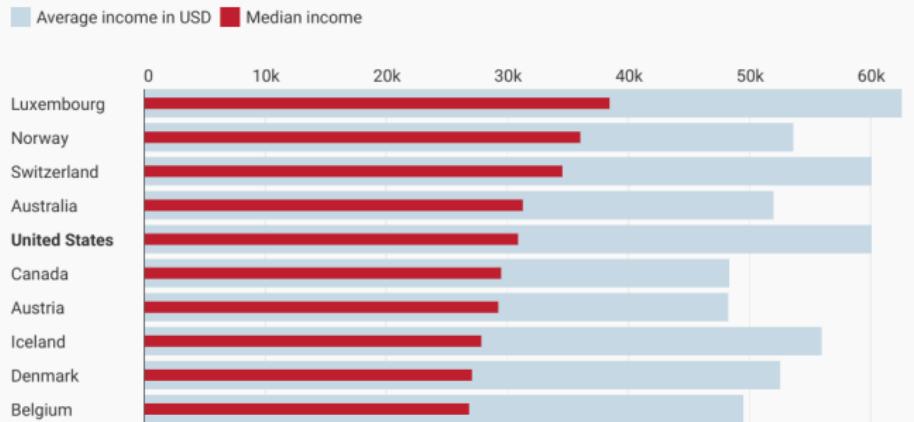
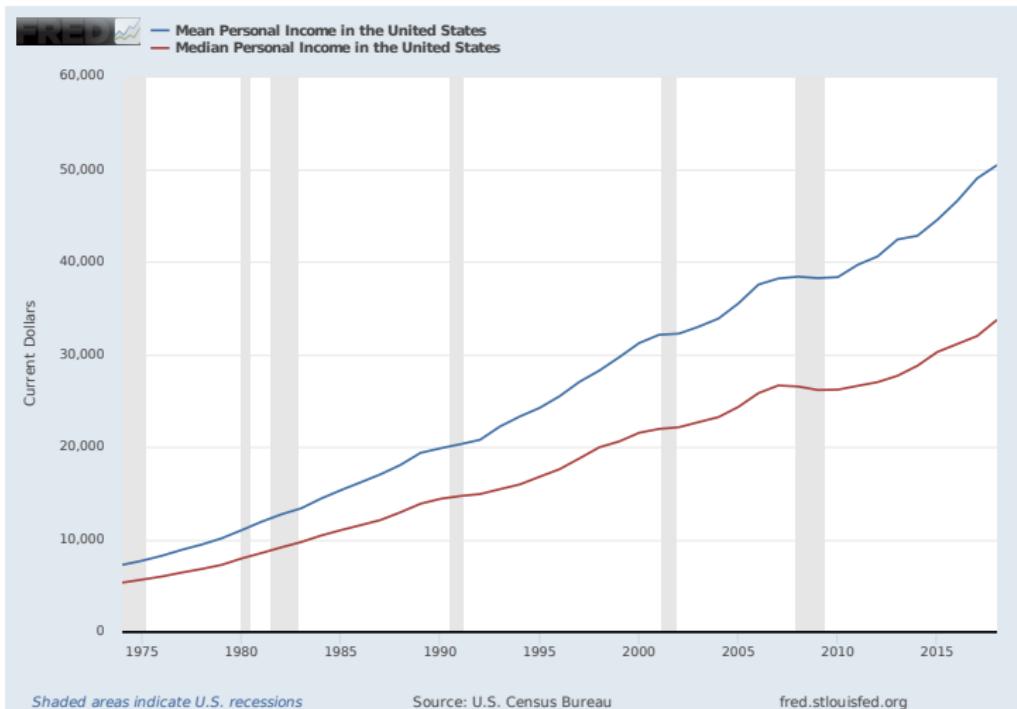


Chart: Lisa Charlotte Rost, Datawrapper

Example: Income Inequality



A General Framework

- ▶ We started with the **mean error**:

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

- ▶ Today, we introduced the **mean squared error**:

$$R_{\text{sq}}(h) = \frac{1}{n} \sum_{i=1} (h - y_i)^2$$

- ▶ They have the same form: average difference between h and data.

A General Framework

- ▶ Definition: A **loss function** $L(h, y)$ takes in a prediction h and a right answer, y , and outputs a number measuring how far h is from y (bigger = further).
- ▶ The **absolute loss**:

$$L_{\text{abs}}(h, y) = |y - h|$$

- ▶ The **square loss**:

$$L_{\text{sq}}(h, y) = (y - h)^2$$

A General Framework

- ▶ Suppose that y_1, \dots, y_n are some data points, h is a prediction, and L is a loss function. The **empirical risk** is the average loss:

$$R_L(h) = \frac{1}{n} \sum_{i=1}^n L(h, y_i)$$

- ▶ The goal of learning: find h that minimizes R_L . This is called **empirical risk minimization (ERM)**.

Designing a learning algorithm using ERM

1. Pick a loss function.
 2. Pick a way to minimize the average loss (empirical risk)
-
- **Key Idea:** The choice of loss function determines the properties of the result and the difficulty of finding it.

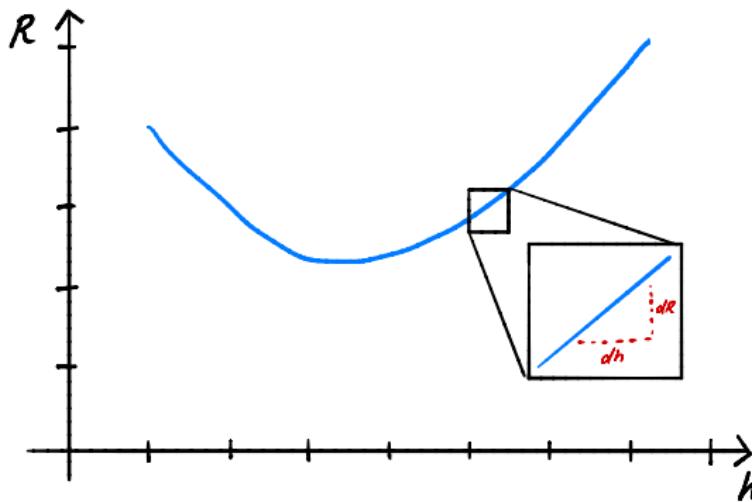
Loss	Minimizer	Outliers	Differentiable	Algorithm
L_{abs}	median	insensitive	no	not simple
L_{sq}	mean	sensitive	yes	simple, fast

Status Update

- ▶ We introduced the **mean squared error** because it is differentiable.
- ▶ The minimizer of the mean squared error is the **mean**.
- ▶ The mean error and the mean squared error fit into a general framework of **empirical risk minimization**.

Next Time

- ▶ We'll design our own loss function.
- ▶ We'll develop a general way of solving minimization problems: **gradient descent**.



DSC 40A
Lecture 04
Learning via Optimization, pt IV

Announcements

- ▶ Remember: homework due tomorrow @ 5 pm.

Last Time: Empirical Risk Minimization

- ▶ To learn, pick a **loss function** L and minimize the **empirical risk**:

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

- ▶ Absolute loss: $L_{\text{abs}}(h, y) = |h - y|$ (gives the **median**)
- ▶ Square loss: $L_{\text{sq}}(h, y) = (h - y)^2$ (gives the **mean**)
- ▶ **Key Point:** Tradeoffs to each loss function.

Today

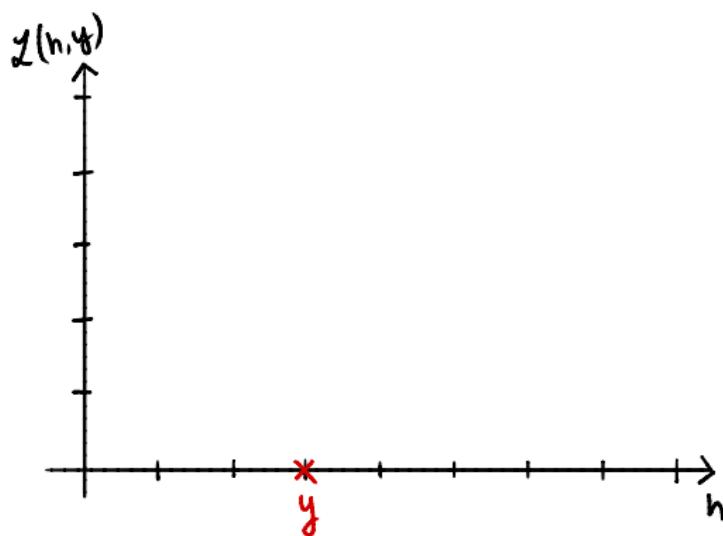
- ▶ We'll design our own loss function.
- ▶ We'll get stuck when trying to minimize.
- ▶ We'll invent **gradient descent** as a general approach to minimizing functions.

Loss Functions

- ▶ A loss function $L(h, y)$ quantifies how “bad” a prediction is.
- ▶ Example: take $h = 4$ and $y = 6$.
- ▶ Absolute loss: $L_{\text{abs}}(h, y) = |4 - 6| = 2$
- ▶ Square loss: $L_{\text{sq}}(h, y) = (4 - 6)^2 = 4$

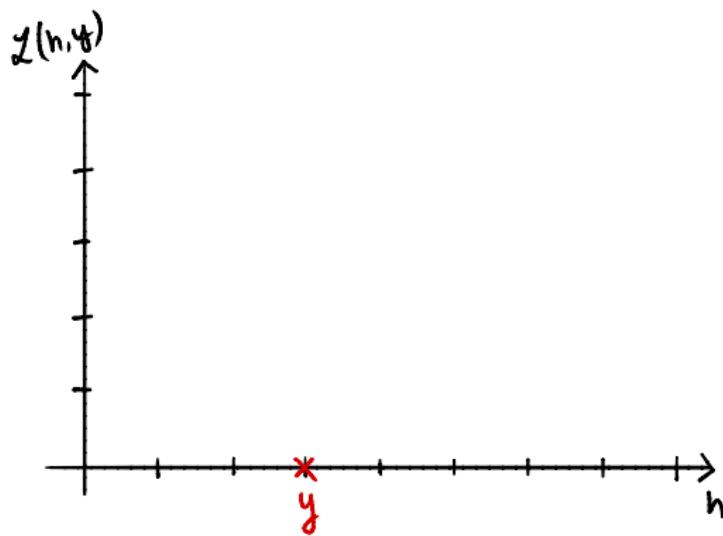
Plotting a Loss Function

- ▶ The plot of a loss function tells us how it treats outliers.
- ▶ Consider y fixed. Plot $L_{\text{abs}}(h, y) = |h - y|$:



Plotting a Loss Function

- ▶ The plot of a loss function tells us how it treats outliers.
- ▶ Consider y fixed. Plot $L_{\text{sq}}(h, y) = (h - y)^2$:

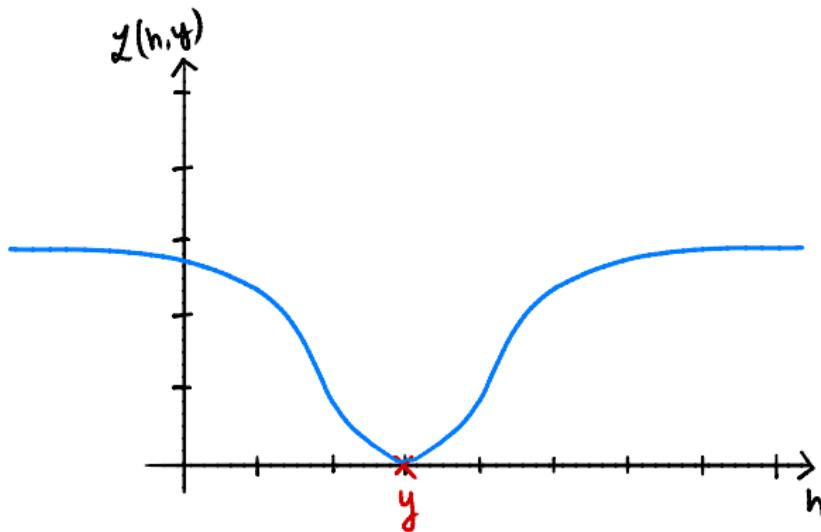


Discussion Question

Suppose L considers all outliers to be equally as bad.
What would it look like far away from y ?

- a) flat
- b) rapidly decreasing
- c) rapidly increasing

A very insensitive loss



- We'll call this loss L_{ucsd} because it doesn't have a name.

Discussion Question

Which of these could be $L_{\text{ucsd}}(h, y)$?

- a) $e^{-(h-y)^2}$
- b) $1 - e^{-(h-y)^2}$
- c) $1 - (h - y)^2$
- d) $1 - e^{-|h-y|}$

Adding a scale parameter

- ▶ Problem: L_{ucsd} has a fixed scale.
- ▶ Won't work for all data sets (e.g., salaries).
- ▶ Fix: add a **scale parameter**, σ :

$$L_{\text{ucsd}}(h, y) = 1 - e^{-(h-y)^2/\sigma^2}$$

Empirical Risk Minimization

- ▶ We have salaries y_1, \dots, y_n .
- ▶ To find prediction, ERM says to minimize the mean loss:

$$\begin{aligned} R_{\text{ucsd}}(h) &= \frac{1}{n} \sum_{i=1}^n L_{\text{ucsd}}(h, y_i) \\ &= \frac{1}{n} \sum_{i=1}^n \left[1 - e^{-(h-y_i)^2/\sigma^2} \right] \end{aligned}$$

Let's plot R_{ucsd}

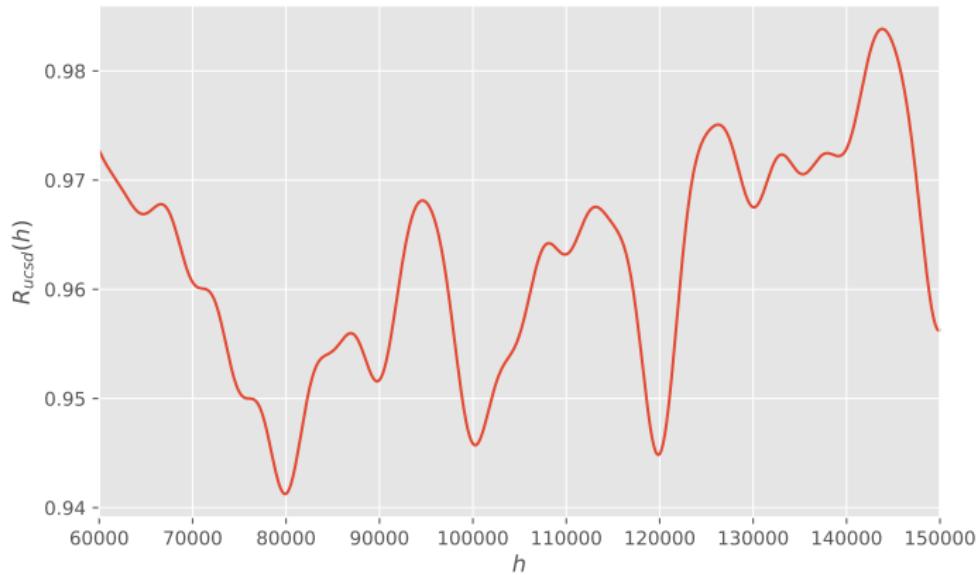
- ▶ Recall:

$$R_{\text{ucsd}}(h) = \frac{1}{n} \sum_{i=1}^n \left[1 - e^{-(h-y_i)^2/\sigma^2} \right]$$

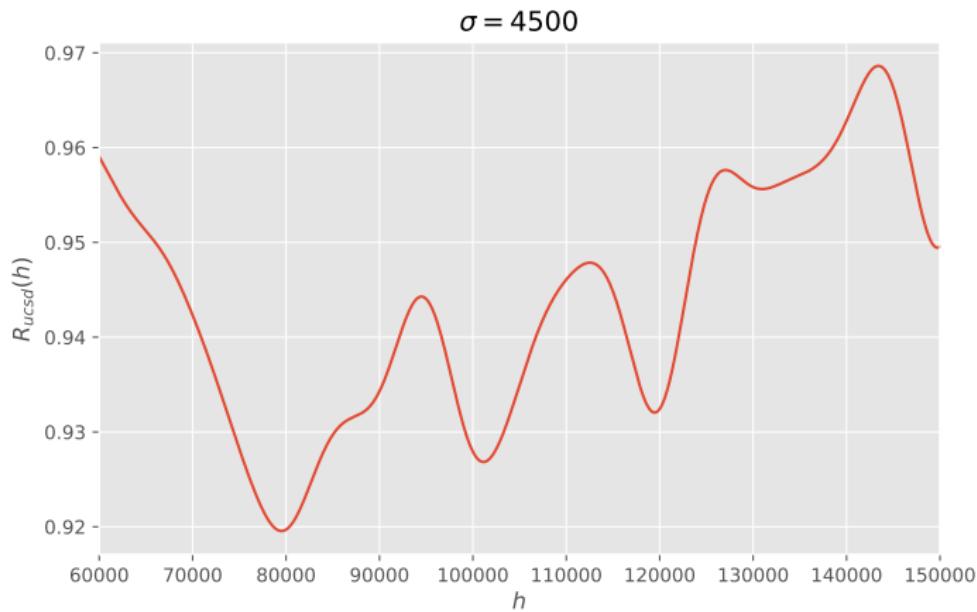
- ▶ Once we have data y_1, \dots, y_n and a scale σ , we can plot $R_{\text{ucsd}}(h)$
- ▶ We'll use full StackOverflow data ($n = 1121$)
- ▶ Let's try several scales, σ .

Plot of R_{ucsd}

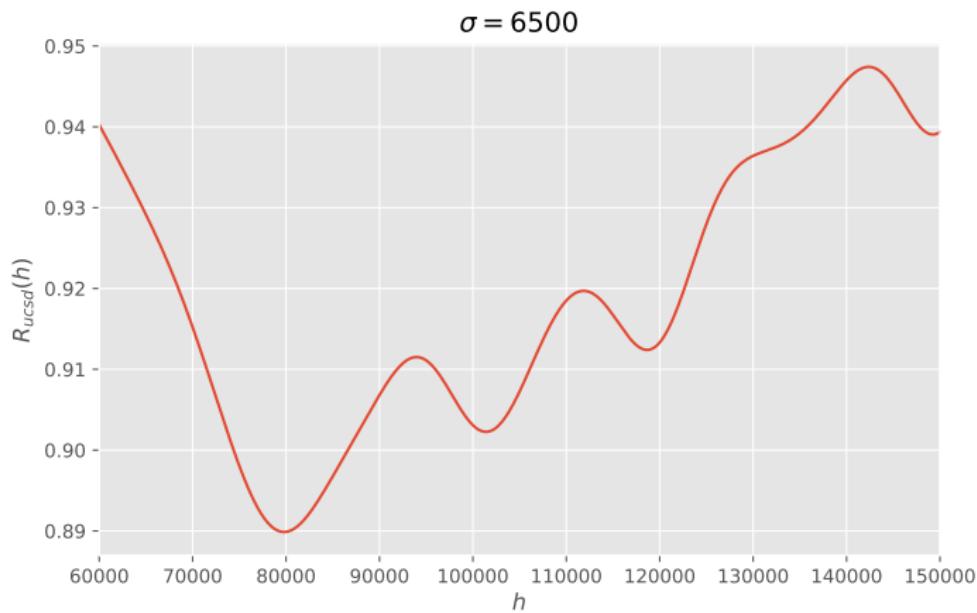
$\sigma = 3000$



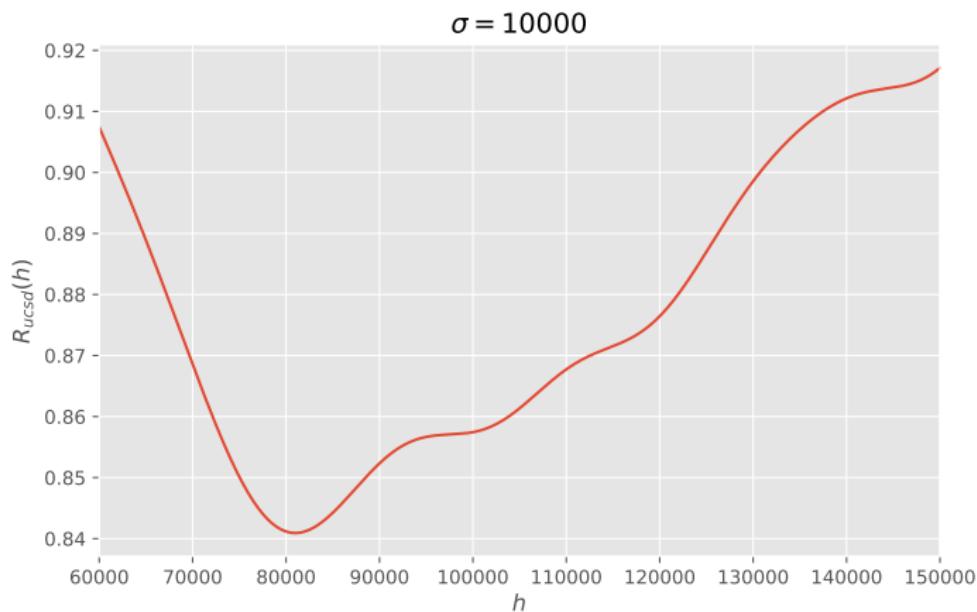
Plot of R_{ucsd}



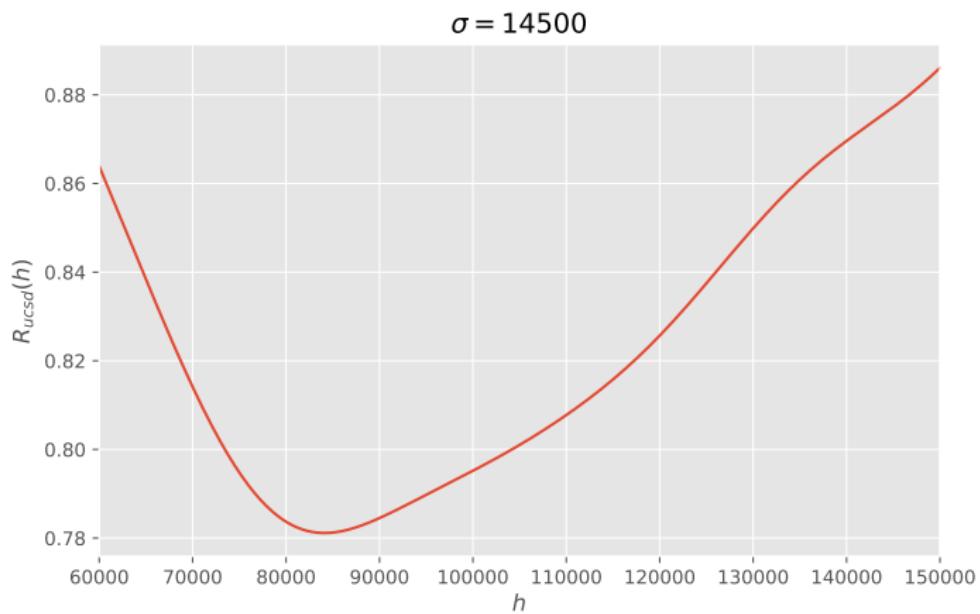
Plot of R_{ucsd}



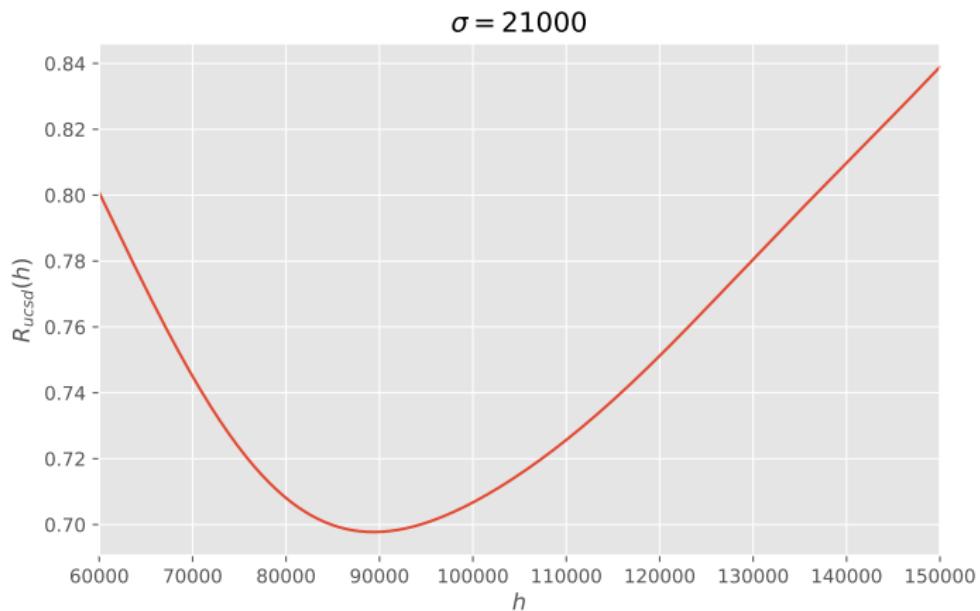
Plot of R_{ucsd}



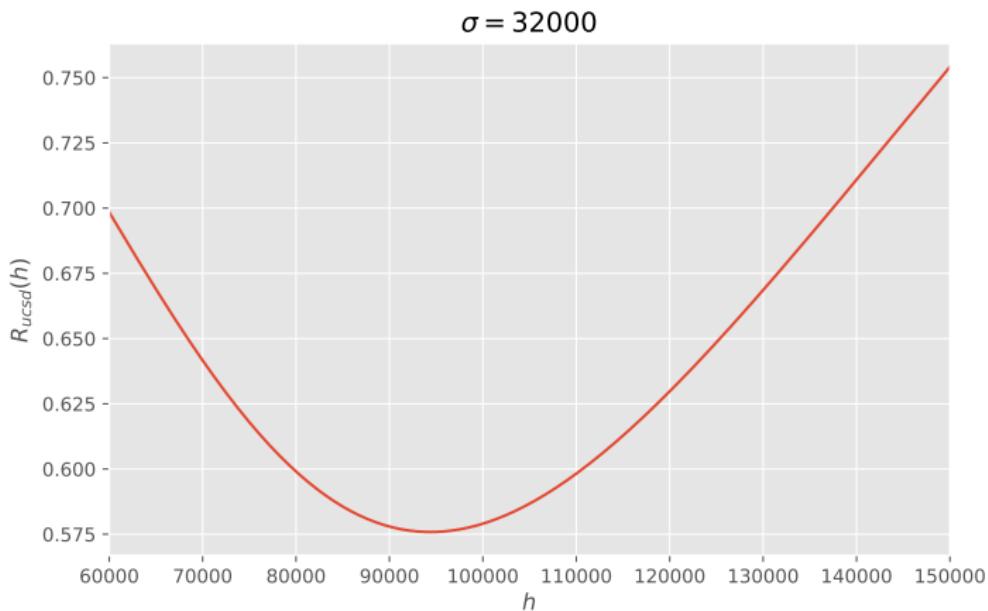
Plot of R_{ucsd}



Plot of R_{ucsd}



Plot of R_{ucsd}



Minimizing R_{ucsd}

- ▶ To make prediction, we find h^* minimizing $R_{\text{ucsd}}(h)$.
- ▶ R_{ucsd} is differentiable (no cusps).
- ▶ To minimize: take derivative, set to zero, solve.

Step 1) Taking the derivative

$$\frac{dR_{\text{ucsd}}}{dh} = \frac{d}{dh} \left(\frac{1}{n} \sum_{i=1}^n \left[1 - e^{-(h-y_i)^2/\sigma^2} \right] \right)$$

Step 2) Setting to zero and solving

- ▶ We found (hopefully):

$$\frac{dR_{\text{ucsd}}}{dh}(h) = \frac{2}{n\sigma^2} \sum_{i=1}^n (h - y_i) \cdot e^{-(h-y_i)^2/\sigma^2}$$

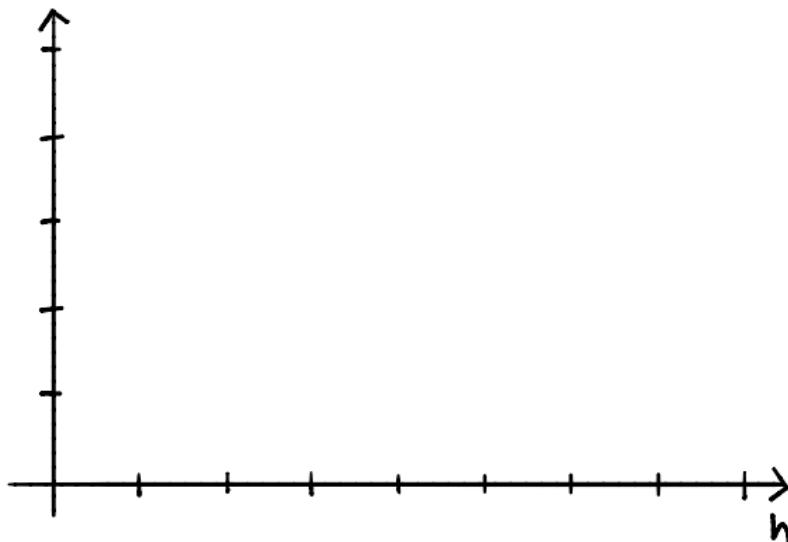
- ▶ Now we just set to zero and solve for h :

$$0 = \frac{2}{n\sigma^2} \sum_{i=1}^n (h - y_i) \cdot e^{-(h-y_i)^2/\sigma^2}$$

- ▶ We **can** calculate derivative, but we **can't** solve for h ; we're stuck again.

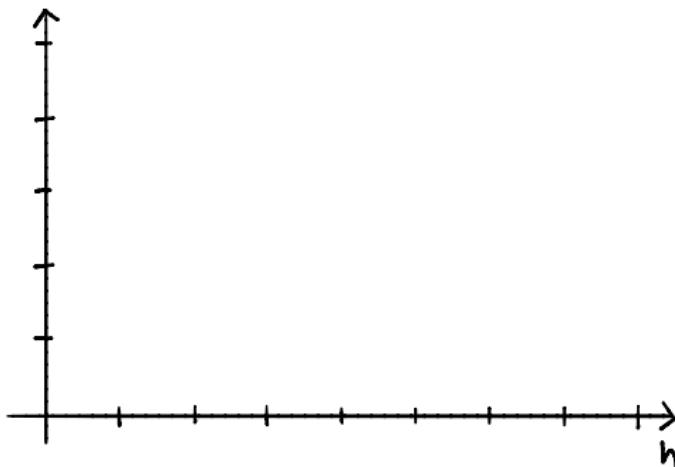
Meaning of the Derivative

- ▶ We have the derivative; can we use it?
- ▶ $\frac{dR}{dh}(h)$ is a function; it gives the **slope** at h .



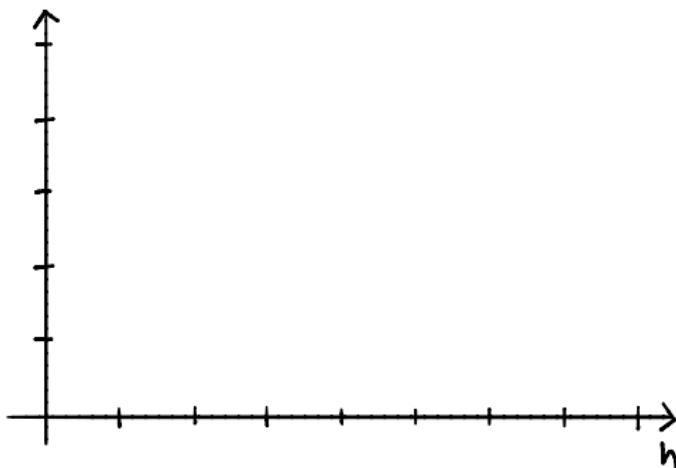
Key Idea Behind Gradient Descent

- ▶ If the slope of R at h is **positive** then moving to the **left** decreases the value of R .
- ▶ i.e., we should **decrease** h



Key Idea Behind Gradient Descent

- ▶ If the slope of R at h is **negative** then moving to the **right** decreases the value of R .
- ▶ i.e., we should **increase** h



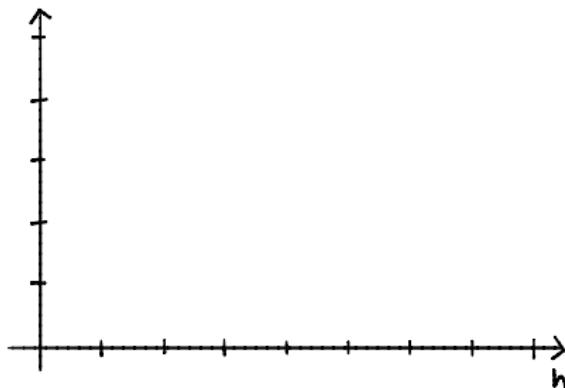
Key Idea Behind Gradient Descent

- ▶ Pick a starting place, h_0 . Where do we go next?
- ▶ Slope at h_0 negative? Then increase h_0 .
- ▶ Slope at h_0 positive? Then decrease h_0 .
- ▶ This will work:

$$h_1 = h_0 - \frac{dR}{dh}(h_0)$$

Gradient Descent

- ▶ Pick α to be a positive number. It is the **learning rate**.
- ▶ Pick a starting prediction, h_0 .
- ▶ On step i , perform update $h_i = h_{i-1} - \alpha \cdot \frac{dR}{dh}(h_{i-1})$
- ▶ Repeat until convergence (when h doesn't change much).



```
def gradient_descent(derivative, h, alpha, tol=1e-12):
    """Minimize using gradient descent."""
    while True:
        h_next = h - alpha * derivative(h)
        if abs(h_next - h) < tol:
            break
        h = h_next
    return h
```

Example: Minimizing Mean Squared Error

- Recall the mean squared error and its derivative:

$$R_{\text{sq}}(h) = \frac{1}{n} \sum_{i=1}^n (h - y_i)^2 \quad \frac{dR_{\text{sq}}}{dh}(h) = \frac{2}{n} \sum_{i=1}^n (h - y_i)$$

Discussion Question

Let $y_1 = -4, y_2 = -2, y_3 = 2, y_4 = 4.$

Pick $h_0 = 4$ and $\alpha = 1/4$. What is h_1 ?

- a) -1
- b) 0
- c) 1
- d) 2

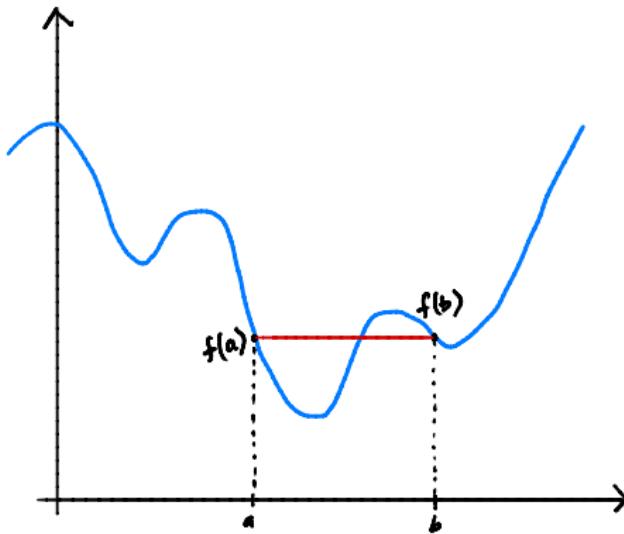
Example

Status Update

- ▶ We introduced the UCSD loss and got stuck trying to minimize.
- ▶ In response, we invented **gradient descent**.

What's Left?

- ▶ When does gradient descent work?
- ▶ When does it fail?



DSC 40A
Lecture 05
Learning via Optimization, pt IV

Last Week: Empirical Risk Minimization

- ▶ To learn, pick a **loss function** L and minimize the **empirical risk**:

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

- ▶ Absolute loss: $L_{\text{abs}}(h, y) = |h - y|$ (gives the **median**)
- ▶ Square loss: $L_{\text{sq}}(h, y) = (h - y)^2$ (gives the **mean**)

Last Week: The UCSD Loss

- ▶ We defined the “UCSD Loss”:

$$L_{\text{ucsd}}(h, y) = 1 - e^{-(h-y)^2/\sigma^2}$$

- ▶ Goal: minimize the “UCSD Risk”,

$$R_{\text{ucsd}}(h, y) = \frac{1}{n} \sum_{i=1}^n \left[1 - e^{-(h-y_i)^2/\sigma^2} \right]$$

- ▶ We tried taking a derivative and solving, but we couldn’t solve for h .

Last Week: Gradient Descent

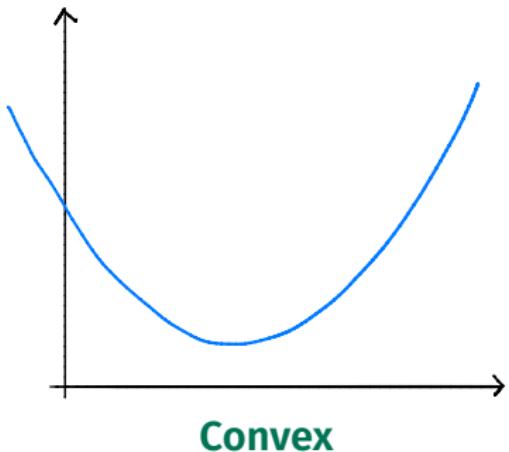
- ▶ Pick α to be a positive number. It is the **learning rate**.
- ▶ Pick a starting prediction, h_0 .
- ▶ On step i , perform update $h_i = h_{i-1} - \alpha \cdot \frac{dR}{dh}(h_{i-1})$
- ▶ Repeat until convergence (when h doesn't change much).

Demo notebook on [DataHub](#)

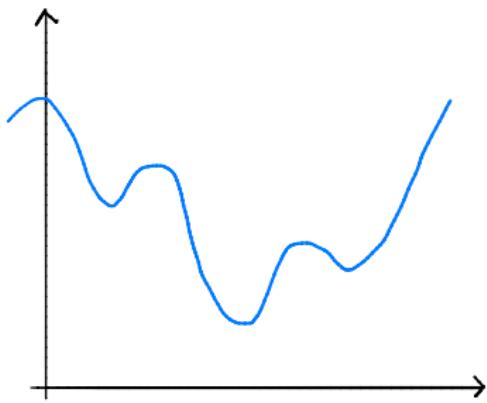
Today

When is gradient descent guaranteed to work?

Convex Functions



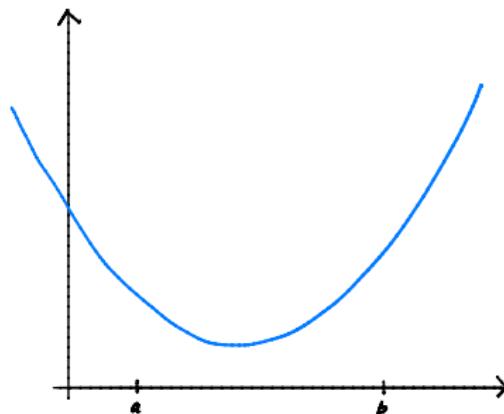
Convex



Non-convex

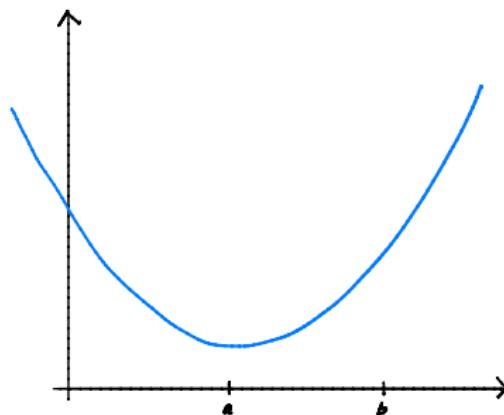
Convexity: Definition

- ▶ f is **convex** if for **every** a, b the line segment between
$$(a, f(a)) \quad \text{and} \quad (b, f(b))$$
does not go below the plot of f .



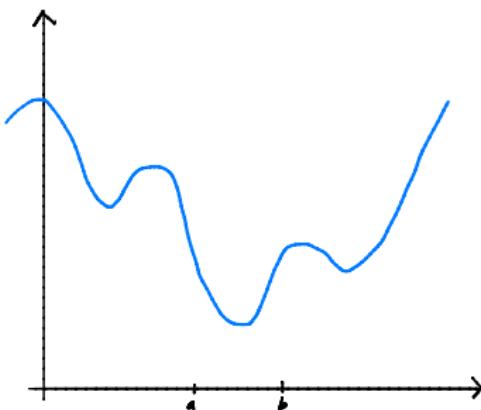
Convexity: Definition

- ▶ f is **convex** if for **every** a, b the line segment between
$$(a, f(a)) \quad \text{and} \quad (b, f(b))$$
does not go below the plot of f .



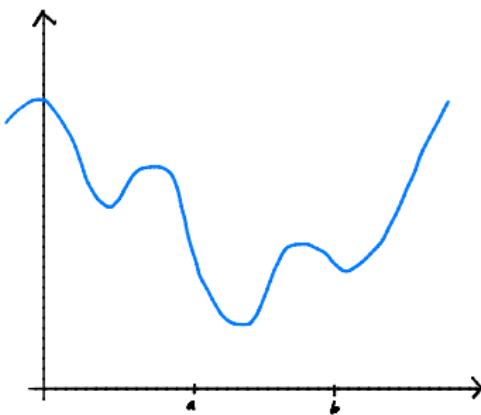
Convexity: Definition

- ▶ f is **convex** if for **every** a, b the line segment between
$$(a, f(a)) \quad \text{and} \quad (b, f(b))$$
does not go below the plot of f .



Convexity: Definition

- ▶ f is **convex** if for **every** a, b the line segment between
$$(a, f(a)) \quad \text{and} \quad (b, f(b))$$
does not go below the plot of f .



Deriving a More Useful/Formal Definition

- ▶ Walk from a at time $t = 0$ to b at time $t = 1$.
- ▶ Let $\text{height}_f(t)$ be height of f at time t .
- ▶ Let $\text{height}_{\text{line}}(t)$ be height of line segment at time t .
- ▶ If f is convex, then for every $t \in [0, 1]$:

$$\text{height}_{\text{line}}(t) \geq \text{height}_f(t)$$

Position at time t

- ▶ Let $x(t)$ be horizontal position at time t .
- ▶ At time $t = 0$, we're at a , so $x(0) = a$.
- ▶ At time $t = 1$, we're at b , so $x(1) = b$.
- ▶ This formula works:

$$x(t) =$$

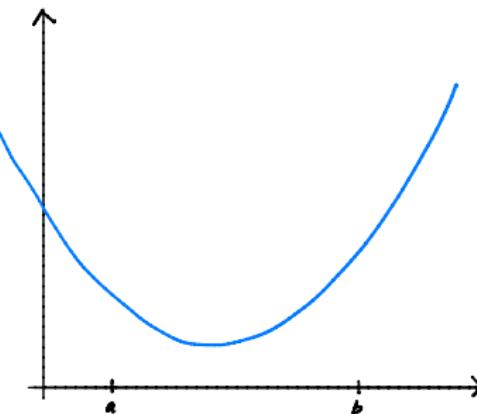
=

Height of f at time t

- ▶ We want a formula for $\text{height}_f(t)$
- ▶ Remember $x(t) = (1 - t)a + bt$. So:

$$\text{height}_f(t) =$$

=



Height of line segment at time t

- ▶ We want a formula for $\text{height}_{\text{line}}(t)$
- ▶ It is a linear function: $\text{height}_{\text{line}}(t) = w_1 t + w_0$
- ▶ We know $\text{height}_{\text{line}}(0) = f(a)$ and $\text{height}_{\text{line}}(1) = f(b)$.

Height of line segment at time t

- ▶ We want a formula for $\text{height}_{\text{line}}(t)$
- ▶ It is a linear function: $\text{height}_{\text{line}}(t) = w_1 t + w_0$
- ▶ We know $\text{height}_{\text{line}}(0) = f(a)$ and $\text{height}_{\text{line}}(1) = f(b)$.

Discussion Question

What is the formula for $\text{height}_{\text{line}}(t)$?

- a) $at + (1 - b)t$
- b) $(1 - t)f(a) + tf(b)$
- c) $(a \cdot f(t) + b \cdot f(t))/2$
- d) $t[f(b) - f(a)]$

Height of line segment at time t

$$\text{height}_{\text{line}}(t) = w_1 t + w_0$$

$$\text{height}_{\text{line}}(0) = f(a) \quad \text{height}_{\text{line}}(1) = f(b)$$

Convexity: Formal Definition

$$\text{height}_{\text{line}}(t) \geq \text{height}_f(t)$$

$$(1 - t)f(a) + tf(b) \geq f((1 - t)a + tb)$$

Convexity: Formal Definition

$$\text{height}_{\text{line}}(t) \geq \text{height}_f(t)$$

$$(1 - t)f(a) + tf(b) \geq f((1 - t)a + tb)$$

- ▶ A function $f : \mathbb{R} \rightarrow \mathbb{R}$ is **convex** if for every choice of $a, b \in \mathbb{R}$ and $t \in [0, 1]$:

$$(1 - t)f(a) + tf(b) \geq f((1 - t)a + tb).$$

Convexity: Formal Definition

$$\text{height}_{\text{line}}(t) \geq \text{height}_f(t)$$

$$(1 - t)f(a) + tf(b) \geq f((1 - t)a + tb)$$

- ▶ A function $f : \mathbb{R} \rightarrow \mathbb{R}$ is **convex** if for every choice of $a, b \in \mathbb{R}$ and $t \in [0, 1]$:

$$(1 - t)f(a) + tf(b) \geq f((1 - t)a + tb).$$

- ▶ A function f is **nonconvex** if it is not convex.

Discussion Question

Is $f(x) = |x|$ convex?

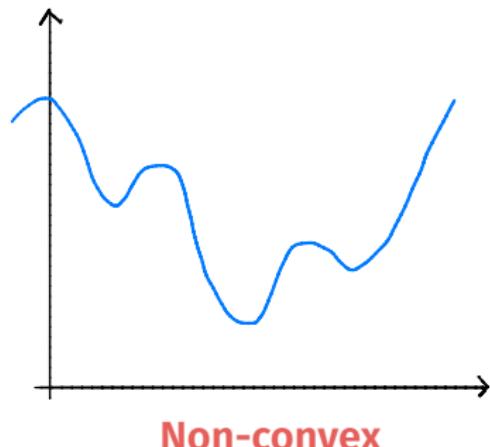
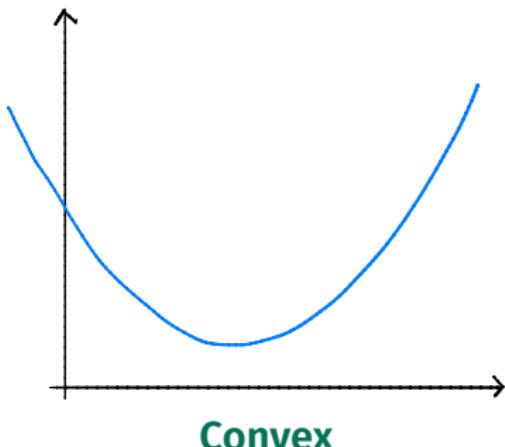
- a) Yes.
- b) No.
- c) Maybe.

Example: Prove that $f(x) = |x|$ is convex

Hint: remember triangle inequality, $|\alpha + \beta| \leq |\alpha| + |\beta|$.

Proving Convexity: Second Derivative Test

- ▶ If $\frac{d^2f}{dx^2}(x) \geq 0$ for all x , then f is convex.
- ▶ Example: $f(x) = x^4$ is convex.
- ▶ Only works if f is twice differentiable!



Proving Convexity: Using Properties

Suppose that $f(x)$ and $g(x)$ are convex. Then:

- ▶ $w_1 f(x) + w_2 g(x)$ is convex, provided $w_1, w_2 \geq 0$
 - ▶ Example: $3x^2 + |x|$ is convex
- ▶ $g(f(x))$ is convex, provided g is non-decreasing.
 - ▶ Example: e^{x^2} is convex
- ▶ $\max\{f(x), g(x)\}$ is convex
 - ▶ Example: $\begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$ is convex (max of 0 and x)

Convex Losses

- ▶ If $L(h, y)$ is a convex function (when y is fixed) then

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

is convex.

- ▶ Proof: sums of convex functions are convex.

Convexity and Gradient Descent

- ▶ Convex functions are (relatively) easy to optimize.
- ▶ **Theorem:** if $R(h)$ is convex and differentiable¹ then gradient descent converges to a **global optimum** of R *provided* that the step size is small enough².

¹and it's derivative is not too wild

²step size related to steepness.

Convexity and Gradient Descent

- ▶ Convex functions are (relatively) easy to optimize.
- ▶ **Theorem:** if $R(h)$ is convex and differentiable¹ then gradient descent converges to a **global optimum** of R *provided* that the step size is small enough².
- ▶ We can even modify GD to work with convex, non-differentiable functions.

¹and it's derivative is not too wild

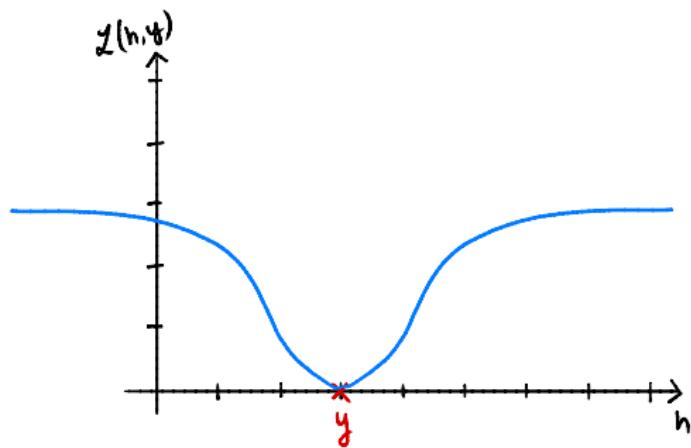
²step size related to steepness.

Nonconvexity and Gradient Descent

- ▶ Nonconvex functions are (relatively) hard to optimize.
- ▶ Gradient descent can still be useful.
- ▶ But not guaranteed to converge to a global minimum.

Convexity of Losses

- ▶ Is $L_{\text{sq}}(h, y) = (h - y)^2$ convex? Yes or No.
- ▶ Is $L_{\text{abs}}(h, y) = |h - y|$ convex? Yes or No.
- ▶ Is $L_{\text{ucsd}}(h, y)$ convex? Yes or No.



Convexity of UCSD Risk

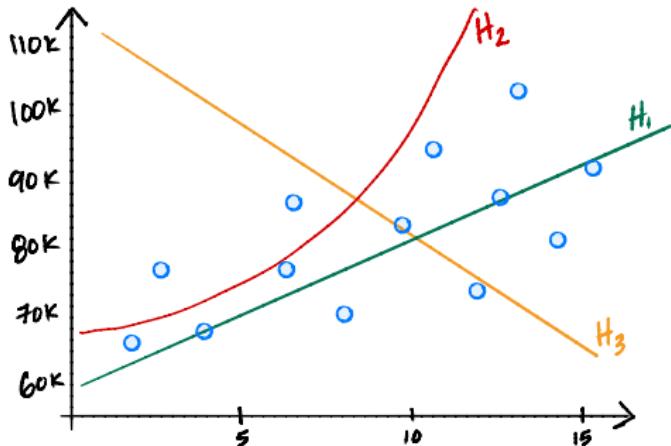
- ▶ A function can be convex in a region.
- ▶ If σ is large, $R_{\text{ucsd}}(h)$ is convex in a big region around data.
- ▶ If σ is small, $R_{\text{ucsd}}(h)$ is convex in only small regions.

Status Update

- ▶ We learned what it means for a function to be **convex**.
- ▶ Convex functions are (relatively) **easy** to optimize with gradient descent.
- ▶ We like **convex loss functions**, like the square loss and absolute loss.

What's Left?

- ▶ We've been predicting salary without using any information about the individual.
- ▶ Making predictions using some information.



DSC 40A

Lecture 06
Least Squares Regression, pt. I

How do we predict someone's salary?

- ▶ Gather salary data, find prediction that minimizes risk.
- ▶ So far, we haven't used any information about the person.
- ▶ How do we incorporate, e.g., years of experience into our prediction?

Features

A **feature** is an attribute – a piece of information.

- ▶ **Numerical**: age, height, years of experience
- ▶ **Categorical**: college, city, gender
- ▶ **Boolean**: knows Python?, had internship?

We'll start with just one feature (years of experience).

Today

- ▶ **Goal:** Predict salary from years of experience.
- ▶ How do we turn this into a math problem and solve it?

Prediction Rules

- ▶ We believe that salary is a function of experience.
- ▶ I.e., there is a function H so that:

$$\text{salary} \approx H(\text{years of experience})$$

- ▶ H is called a **hypothesis function** or **prediction rule**.
- ▶ **Our goal:** find a good prediction rule, H .

Example Prediction Rules

$$H_1(\text{years of experience}) = \$50,000 + \$2,000 \times (\text{years of experience})$$

$$H_2(\text{years of experience}) = \$60,000 \times 1.05^{(\text{years of experience})}$$

$$H_3(\text{years of experience}) = \$100,000 - \$5,000 \times (\text{years of experience})$$

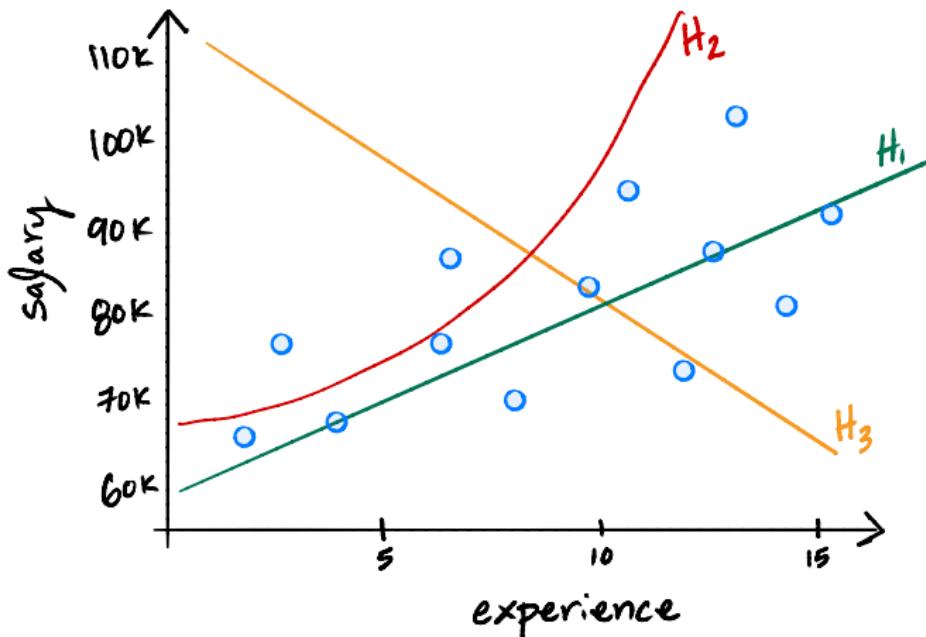
Comparing predictions

- ▶ How do we know which is best: H_1, H_2, H_3 ?
- ▶ We gather data from n people. Let x_i be experience, y_i be salary:

$$\begin{array}{ll} (\text{Experience}_1, \text{Salary}_1) & (x_1, y_1) \\ (\text{Experience}_2, \text{Salary}_2) & (x_2, y_2) \\ \dots & \dots \\ (\text{Experience}_n, \text{Salary}_n) & (x_n, y_n) \end{array} \rightarrow$$

- ▶ See which rule works better on data.

Example



Quantifying the error of a prediction rule H

- ▶ Our prediction for person i 's salary is $H(x_i)$
- ▶ The **absolute error** in this prediction:

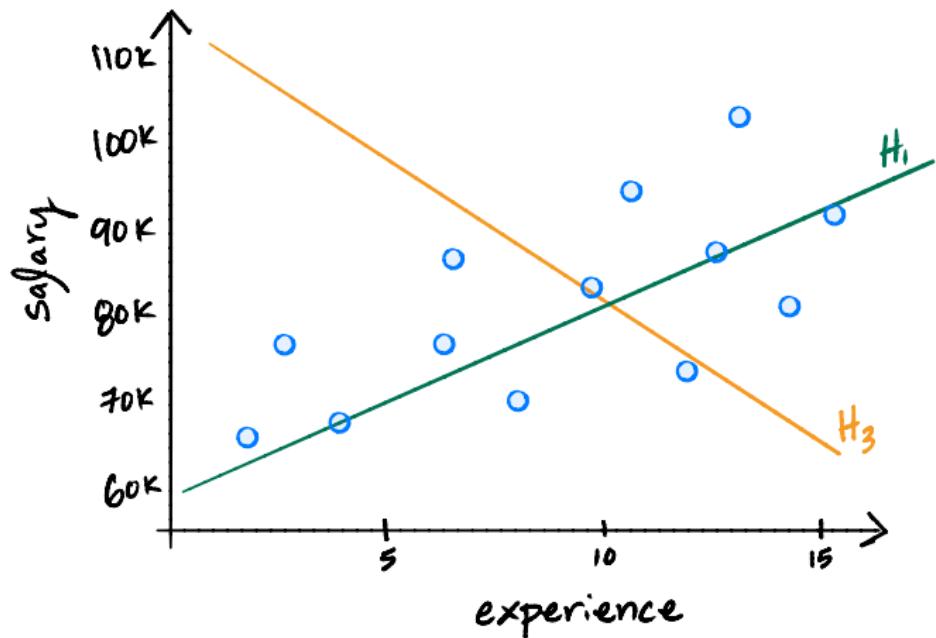
$$|H(x_i) - y_i|$$

- ▶ The **mean absolute error** of H :

$$R_{\text{abs}}(H) = \frac{1}{n} \sum_{i=1}^n |H(x_i) - y_i|$$

- ▶ Smaller the mean absolute error, the **better** the prediction rule.

Mean Absolute Error



Finding the best prediction rule

- ▶ **Goal:** out of all functions $\mathbb{R} \rightarrow \mathbb{R}$, find the function H^* with the smallest mean absolute error.
- ▶ That is, find:

$$H^* = \arg \min_H \frac{1}{n} \sum_{i=1}^n |H(x_i) - y_i|$$

Finding the best prediction rule

- ▶ **Goal:** out of all functions $\mathbb{R} \rightarrow \mathbb{R}$, find the function H^* with the smallest mean absolute error.
- ▶ That is, find:

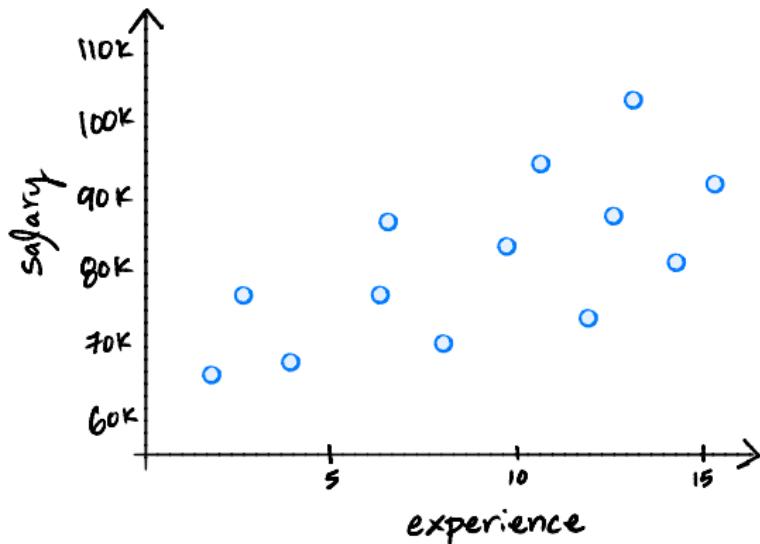
$$H^* = \arg \min_H \frac{1}{n} \sum_{i=1}^n |H(x_i) - y_i|$$

- ▶ **There are two problems with this.**

Discussion Question

Given the data below, is there a prediction rule H which has **zero** mean absolute error?

- a) yes
- b) no



Problem #1

- ▶ We can make mean absolute error very small, even zero!
- ▶ But the function will be weird.
- ▶ This is called **overfitting**.
- ▶ Remember our real goal: make good predictions on data **we haven't seen**.

Solution

- ▶ Don't allow H to be just any function.
- ▶ Require that it has a certain form.
- ▶ Examples:
 - ▶ Linear: $H(x) = w_1x + w_0$
 - ▶ Quadratic: $H(x) = w_2x^2 + w_1x + w_0$
 - ▶ Exponential: $H(x) = w_0e^{w_1x}$
 - ▶ Constant: $H(x) = w_0$

Finding the best linear prediction rule

- ▶ **Goal:** out of all **linear** functions $\mathbb{R} \rightarrow \mathbb{R}$, find the function H^* with the smallest mean absolute error.
- ▶ That is, find:

$$H^* = \arg \min_{\text{linear } H} \frac{1}{n} \sum_{i=1}^n |H(x_i) - y_i|$$

Finding the best linear prediction rule

- ▶ **Goal:** out of all **linear** functions $\mathbb{R} \rightarrow \mathbb{R}$, find the function H^* with the smallest mean absolute error.
- ▶ That is, find:

$$H^* = \arg \min_{\text{linear } H} \frac{1}{n} \sum_{i=1}^n |H(x_i) - y_i|$$

- ▶ **There is still a problem with this.**

Problem #2

- ▶ It is hard to minimize the mean absolute error:¹

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

- ▶ Not differentiable!
- ▶ What can we do?

¹Though it can be done with linear programming.

Quantifying the error of a prediction rule H

- ▶ Instead of absolute error, use the **squared error** of a prediction:

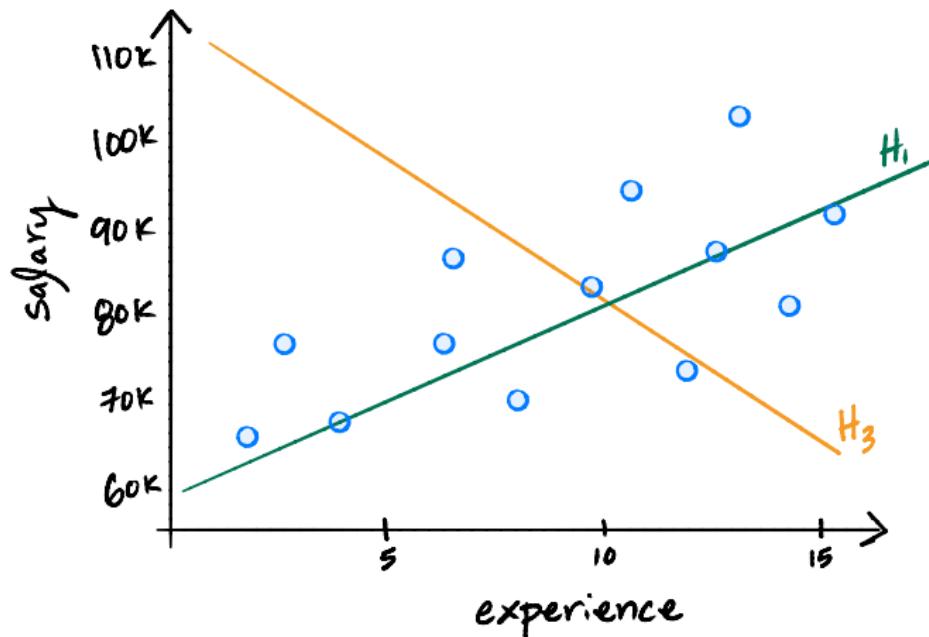
$$(H(x_i) - y_i)^2$$

- ▶ The **mean squared error** (MSE) of H :

$$R_{\text{sq}}(H) = \frac{1}{n} \sum_{i=1}^n (H(x_i) - y_i)^2$$

- ▶ **Is differentiable!**

Mean Squared Error



Our Goal

- ▶ Out of all **linear** functions $\mathbb{R} \rightarrow \mathbb{R}$, find the function H^* with the smallest **mean squared error**.
- ▶ That is, find:

$$H^* = \underset{\text{linear } H}{\arg \min} \frac{1}{n} \sum_{i=1}^n (H(x_i) - y_i)^2$$

- ▶ This problem is called **least squares regression**.

Minimizing the MSE

- ▶ The MSE is a function of a function:

$$R_{\text{sq}}(H) = \frac{1}{n} \sum_{i=1}^n (H(x_i) - y_i)^2$$

- ▶ But since H is linear, $H(x) = w_1 x + w_0$.

$$R_{\text{sq}}(w_1, w_0) = \frac{1}{n} \sum_{i=1}^n ((w_1 x_i + w_0) - y_i)^2$$

- ▶ Now it's a function of w_1, w_0 .

Updated Goal

- ▶ Find slope w_1 and intercept w_0 which minimize the MSE,
 $R_{\text{sq}}(w_1, w_0)$:

$$R_{\text{sq}}(w_1, w_0) = \frac{1}{n} \sum_{i=1}^n ((w_1 x + w_0) - y_i)^2$$

- ▶ Strategy: multivariate calculus.

Recall: the gradient

- If $f(x, y)$ is a function of two variables, the **gradient** of f at the point (x_0, y_0) is a **vector** of **partial derivatives**:

$$\nabla f(x_0, y_0) = \begin{pmatrix} \frac{\partial f}{\partial x}(x_0) \\ \frac{\partial f}{\partial y}(y_0) \end{pmatrix}$$

- Key Fact #1:** derivative : tangent line :: gradient : tangent plane
- Key Fact #2:** points in direction of biggest increase
- Key Fact #3:** if the gradient is zero at critical points.

Strategy

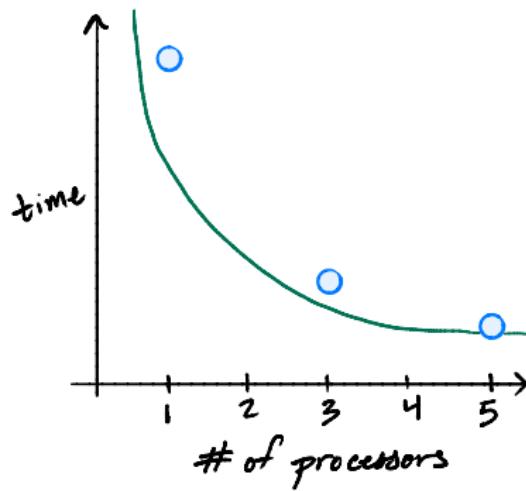
To minimize $R(w_1, w_0)$: compute the gradient, set equal to zero, solve.

$$R_{\text{sq}}(w_1, w_0) = \frac{1}{n} \sum_{i=1}^n ((w_1 x + w_0) - y_i)^2$$

$$\frac{\partial R_{\text{sq}}}{\partial w_1} =$$

$$R_{\text{sq}}(w_1, w_0) = \frac{1}{n} \sum_{i=1}^n ((w_1 x + w_0) - y_i)^2$$

$$\frac{\partial R_{\text{sq}}}{\partial w_0} =$$



DSC 40A

Lecture 07
Least Squares Regression, pt. II

Last Time

- ▶ **Goal:** Find prediction rule $H(x)$ for predicting salary given years of experience.
- ▶ Minimize mean absolute error?

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

- ▶ **Not differentiable**, instead minimize **mean squared error**:

$$\frac{1}{n} \sum_{i=1}^n (H(x_i) - y_i)^2$$

- ▶ To avoid **overfitting**, use linear prediction rule:

$$H(x) = w_1 x + w_0$$

Last Time

- ▶ **Goal:** find w_1 and w_0 which minimize MSE:

$$R_{\text{sq}}(w_1, w_0) = \frac{1}{n} \sum_{i=1}^n ((w_1 x + w_0) - y_i)^2$$

- ▶ **Strategy:** Take derivatives $\partial R_{\text{sq}} / \partial w_1$ and $\partial R_{\text{sq}} / \partial w_0$, set to zero, solve.
- ▶ We found:

$$\frac{\partial R_{\text{sq}}}{\partial w_1}(w_1, w_0) = \frac{2}{n} \sum_{i=1}^n ((w_1 x_i + w_0) - y_i) x_i$$

$$\frac{\partial R_{\text{sq}}}{\partial w_0}(w_1, w_0) = \frac{2}{n} \sum_{i=1}^n ((w_1 x_i + w_0) - y_i)$$

Today

- ▶ Solve these equations to find the **least squares solutions**.
- ▶ See how to easily fit non-linear trends, too.

Strategy

$$0 = \frac{2}{n} \sum_{i=1}^n ((w_1 x_i + w_0) - y_i) x_i \quad 0 = \frac{2}{n} \sum_{i=1}^n ((w_1 x_i + w_0) - y_i)$$

1. Solve for w_0 in second equation.
2. Plug solution for w_0 into first equation, solve for w_1 .

Solve for w_0

$$0 = \frac{2}{n} \sum_{i=1}^n ((w_1 x_i + w_0) - y_i)$$

Solve for w_0

$$0 = \frac{2}{n} \sum_{i=1}^n ((w_1 x_i + w_0) - y_i)$$

Key Fact

- ▶ Define

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

- ▶ Then

$$\sum_{i=1} (x_i - \bar{x}) = 0 \quad \sum_{i=1} (y_i - \bar{y}) = 0$$

Solve for w_1

$$0 = \frac{2}{n} \sum_{i=1}^n ((w_1 x_i + w_0) - y_i) x_i \quad w_0 = \bar{y} - w_1 \bar{x}$$

Solve for w_1

$$0 = \frac{2}{n} \sum_{i=1}^n ((w_1 x_i + w_0) - y_i) x_i \quad w_0 = \bar{y} - w_1 \bar{x}$$

Least Squares Solutions

- The least squares solutions for the slope w_1 and intercept w_0 are:

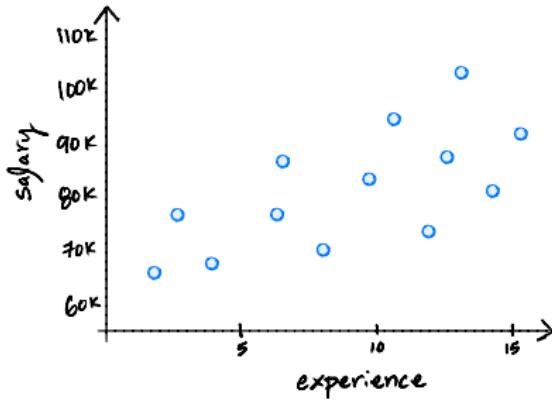
$$w_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad w_0 = \bar{y} - w_1 \bar{x}$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

Interpretation of Slope

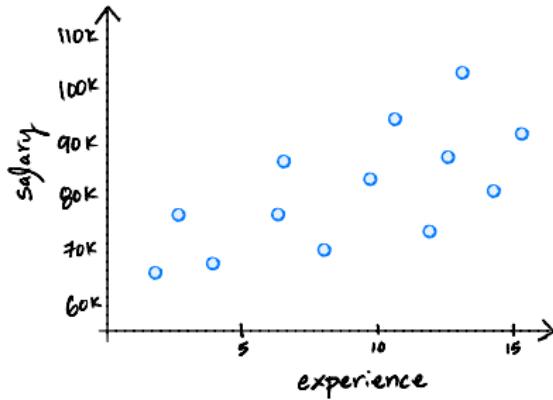
$$w_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$



- ▶ What is the sign of $(x_i - \bar{x})(y_i - \bar{y})$ when:
 - ▶ $x_i > \bar{x}$ and $y_i > \bar{y}$?
 - ▶ $x_i < \bar{x}$ and $y_i < \bar{y}$?
 - ▶ $x_i > \bar{x}$ and $y_i < \bar{y}$?
 - ▶ $x_i < \bar{x}$ and $y_i > \bar{y}$?

Interpretation of Intercept

$$w_0 = \bar{y} - w_1 \bar{x}$$



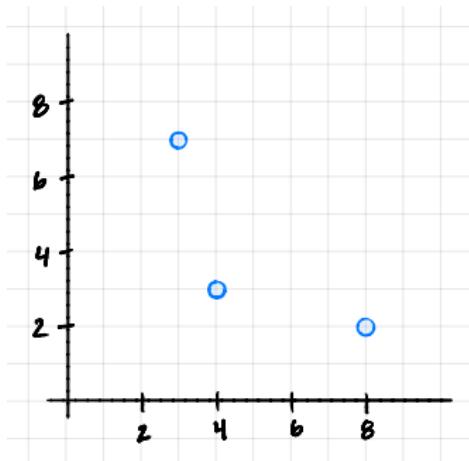
- ▶ What is $H(\bar{x})$?

Discussion Question

We fit a linear prediction rule for salary given years of experience. Then everyone gets a \$5,000 raise. Which of these happens?

- a) slope increases, intercept increases
- b) slope decreases, intercept increases
- c) slope stays same, intercept increases
- d) slope stays same, intercept stays same

Example



$$\bar{x} =$$

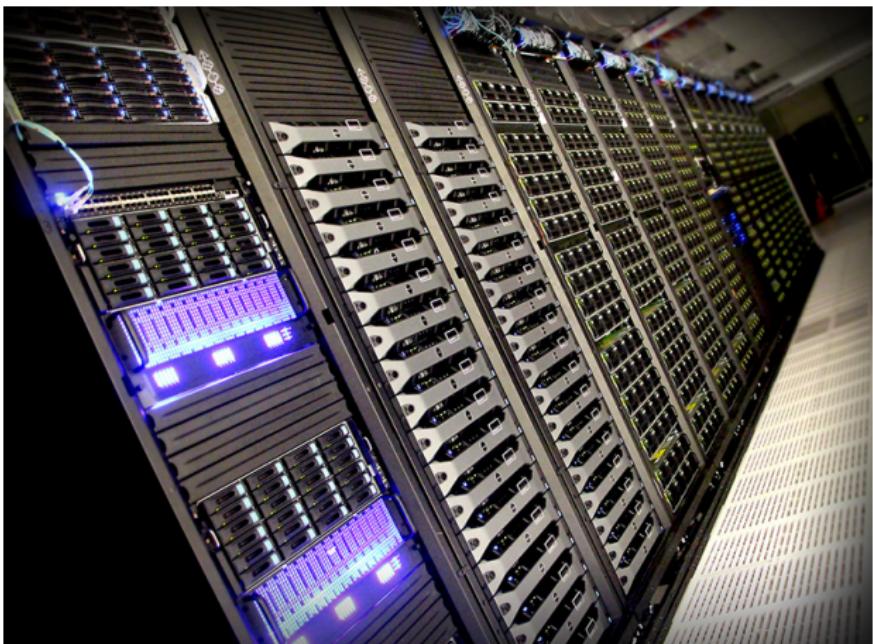
$$\bar{y} =$$

$$w_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} =$$

$$w_0 = \bar{y} - w_1 \bar{x}$$

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
3	7	-1	4	-4	1
4	3	0	-4	0	0
8	2	5	-5	-25	25

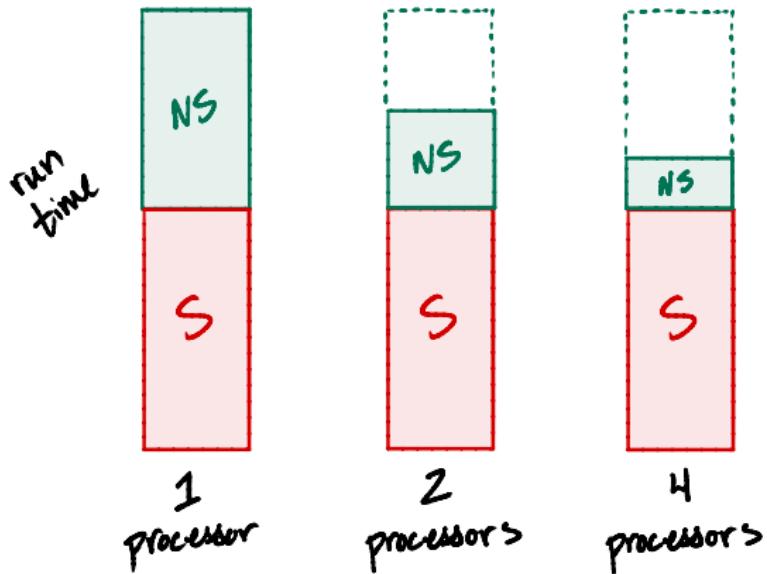
Example: Parallel Processing



Problem

- ▶ Some parts of a program are necessarily **sequential**.
- ▶ E.g., downloading the data must happen before analysis.
- ▶ More processors do not speed up **sequential** code.
- ▶ But they do speed up **non-sequential** code.

Speedup



Amdahl's Law

The time T it takes to run a program on p processors is:

$$T(p) = t_S + \frac{t_{NS}}{p}$$

where t_S and t_{NS} are the time it takes the sequential and non-sequential parts to run on one processor, respectively.

Amdahl's Law

The time T it takes to run a program on p processors is:

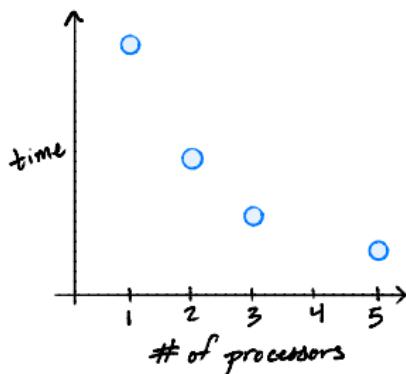
$$T(p) = t_S + \frac{t_{NS}}{p}$$

where t_S and t_{NS} are the time it takes the sequential and non-sequential parts to run on one processor, respectively.

Problem: we don't know t_S and t_{NS} .

Fitting Amdahl's Law

- ▶ **Solution:** we will learn t_S and t_{NS} from data.
- ▶ Run with varying number of processors, record total time:



- ▶ Find decision rule $H(p) = \frac{t_{NS}}{p} + t_S$ by minimizing MSE.

General Problem

- ▶ Given data $(x_1, y_1), \dots, (x_n, y_n)$.
- ▶ Fit a **non-linear** rule $H(x) = w_1 \cdot \frac{1}{x} + w_0$ by minimizing MSE:

$$R_{\text{sq}}(H) = \frac{1}{n} \sum_{i=1}^n (H(x_i) - y_i)^2$$

Using definition of H :

Minimizing MSE

- ▶ Take derivatives, you'll find:

$$\frac{\partial R_{\text{sq}}}{\partial w_1}(w_1, w_0) = \frac{2}{n} \sum_{i=1}^n \left[\left(w_1 \cdot \frac{1}{x_i} + w_0 \right) - y_i \right] \frac{1}{x_i}$$

$$\frac{\partial R_{\text{sq}}}{\partial w_0}(w_1, w_0) = \frac{2}{n} \sum_{i=1}^n \left[\left(w_1 \cdot \frac{1}{x_i} + w_0 \right) - y_i \right]$$

Minimizing MSE

- ▶ Set to zero, solve. You'll find:

$$w_1 = \frac{\sum_{i=1}^n \left(\frac{1}{x_i} - \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i} \right) (y_i - \bar{y})}{\sum_{i=1}^n \left(\frac{1}{x_i} - \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i} \right)^2}$$

$$w_0 = \bar{y} - w_1 \cdot \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}$$

Minimizing MSE

- ▶ Set to zero, solve. You'll find:

$$w_1 = \frac{\sum_{i=1}^n \left(\frac{1}{x_i} - \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i} \right) (y_i - \bar{y})}{\sum_{i=1}^n \left(\frac{1}{x_i} - \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i} \right)^2}$$
$$w_0 = \bar{y} - w_1 \cdot \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}$$

- ▶ Define $z_i = \frac{1}{x_i}$, $\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i = \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}$. Then:

$$w_1 =$$

$$w_0 =$$

Fitting Non-Linear Trends

To fit a prediction rule of the form $H(x) = w_1 \cdot \frac{1}{x} + w_0$:

1. Create a new data set $(z_1, y_1), \dots, (z_n, y_n)$, where $z_i = \frac{1}{x_i}$.
2. Fit $H(z) = w_1 z + w_0$ using familiar least squares solutions:

$$w_1 = \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n (z_i - \bar{z})^2} \quad w_0 = \bar{y} - w_1 \cdot \bar{z}$$

3. Use w_1 and w_0 in original decision rule, $H(x)$.

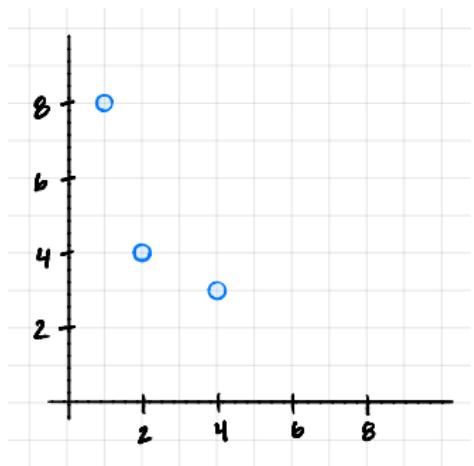
Example: Amdahl's Law

- ▶ We have timed our program:

Processors	Time (Hours)
1	8
2	4
4	3

- ▶ Fit prediction rule: $H(p) = \frac{t_{NS}}{p} + t_S$

Example: fitting $H(x) = w_1 \cdot \frac{1}{x_i} + x_0$



$$\bar{z} =$$

$$\bar{y} =$$

$$w_1 = \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n (z_i - \bar{z})^2} =$$

$$w_0 = \bar{y} - w_1 \bar{z}$$

x_i	z_i	y_i	$(z_i - \bar{z})$	$(y_i - \bar{y})$	$(z_i - \bar{z})(y_i - \bar{y})$	$(z_i - \bar{z})^2$
3	7					
4	3					
8	2					

Example: Amdahl's Law

- ▶ We found: $t_{NS} = \frac{48}{7} \approx 6.88$, $t_S = 1$
- ▶ Our prediction rule:

$$H(p) = \frac{t_{NS}}{p} + t_S$$

$$= \frac{6.88}{p} + 1$$

Fitting Non-Linear Trends

To fit a prediction rule of the form $H(x) = w_1 \cdot f(x) + w_0$:

1. Create a new data set $(z_1, y_1), \dots, (z_n, y_n)$, where $z_i = f(x_i)$.
2. Fit $H(z) = w_1 z + w_0$ using familiar least squares solutions:

$$w_1 = \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n (z_i - \bar{z})^2} \quad w_0 = \bar{y} - w_1 \cdot \bar{z}$$

3. Use w_1 and w_0 in original decision rule, $H(x)$.

Fitting Non-Linear Trends

- We can fit rules like:

$$w_1x + w_0 \quad w_1 \cdot \frac{1}{x} + w_0 \quad w_1x^2 + w_0 \quad w_1e^x + w_0$$

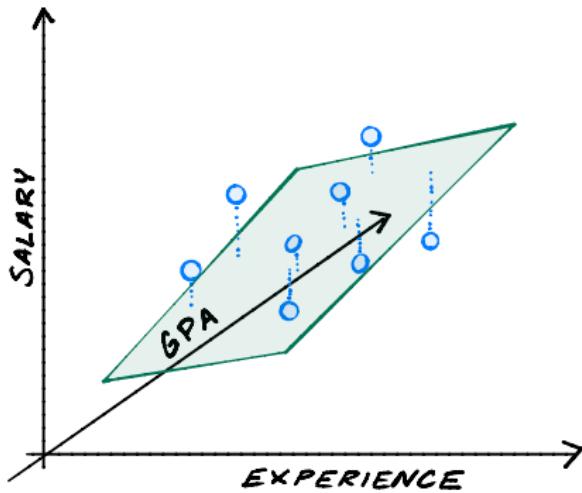
- We can't fit rules like:

$$w_0e^{w_1x} \quad \sin(w_1x + w_0)$$

- Can fit as long as **linear** function of w_1, w_0 .

What's Left?

- ▶ How do we make predictions with lots of features?
- ▶ E.g., experience, age, GPA, number of internships, etc.



DSC 40A

Lecture 08
Least Squares Regression, pt. III

Announcements

- ▶ The midterm is Tuesday, in lecture.
- ▶ Covers Lectures 01 through 07 (this Tuesday).
- ▶ Concepts:
 - ▶ loss functions and ERM, gradient descent, convexity, least squares regression, etc.
- ▶ Core Skills:
 - ▶ partial derivatives, working with summations, chains of inequalities, etc.
- ▶ Best study device: homeworks and discussion worksheets.

Last Time

- ▶ **Goal:** Find prediction rule $H(x)$ for predicting salary given years of experience.
- ▶ To avoid **overfitting**, use linear prediction rule:

$$H(x) = w_1 x + w_0$$

- ▶ We want w_1 and w_0 to minimize the mean squared error:

$$R_{\text{sq}}(w_1, w_0) = \frac{1}{n} \sum_{i=1}^n ((w_1 x + w_0) - y_i)^2$$

Last Time

- ▶ Take derivatives, set to zero, solve:

$$w_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$w_0 = \bar{y} - w_1 \bar{x}$$

Today

- ▶ How do we predict salary given **multiple** features?
 - ▶ years of experience, number of internships, GPA, etc.
- ▶ We'll need to use some linear algebra...

Basic Linear Algebra Review

Matrices

An $m \times n$ **matrix** is a table of numbers with m rows, n columns:

- ▶ Example: 2×3 matrix:

$$\begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \end{pmatrix}$$

- ▶ Example: 3×3 “square” matrix:

$$\begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{pmatrix}$$

- ▶ Example: 3×1 “column”:

$$\begin{pmatrix} m_{11} \\ m_{21} \\ m_{31} \end{pmatrix}$$

Matrix Notation

- We use upper-case letters for matrices.

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$$

- Sometimes use subscripts to denote particular elements:
 $A_{13} = 3, A_{21} = 4$
- A^T denotes the transpose of A:

$$A^T = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}$$

Matrix Addition and Scalar Multiplication

- ▶ We can add two matrices only if they are the same size.
- ▶ Addition occurs elementwise:

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} + \begin{pmatrix} 7 & 8 & 9 \\ -1 & -2 & -3 \end{pmatrix} = \begin{pmatrix} 8 & 10 & 12 \\ 3 & 3 & 3 \end{pmatrix}$$

- ▶ Scalar multiplication occurs elementwise, too:

$$2 \cdot \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} = \begin{pmatrix} 2 & 4 & 6 \\ 8 & 10 & 12 \end{pmatrix}$$

Matrix-Matrix Multiplication

- ▶ We can multiply two matrices A and B only if # cols in A is equal to # rows in B
- ▶ If $A = m \times n$ and $B = n \times p$, the result is $m \times p$.
 - ▶ This is **very useful**. Remember it!
- ▶ The low-level definition. the ij entry of the product is:

$$(AB)_{ij} = \sum_{k=1}^n A_{ik}B_{kj}$$

Matrix-Matrix Multiplication Example

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 3 & 4 & 5 \end{pmatrix} \quad B = \begin{pmatrix} 3 & 6 \\ 1 & 3 \\ 4 & 8 \end{pmatrix}$$

- ▶ What is the size of AB ?
- ▶ What is $(AB)_{12}$?

Matrix-Matrix Multiplication Properties

- ▶ Distributive: $A(B + C) = AB + AC$
- ▶ Associative: $(AB)C = A(BC)$
- ▶ **Not commutative in general:** $AB \neq BA$

Identity Matrices

- The $n \times n$ **identity matrix** I has ones along the diagonal:

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

- If A is $n \times m$, then $IA = A$.
- If B is $m \times n$, then $BI = B$.

Vectors

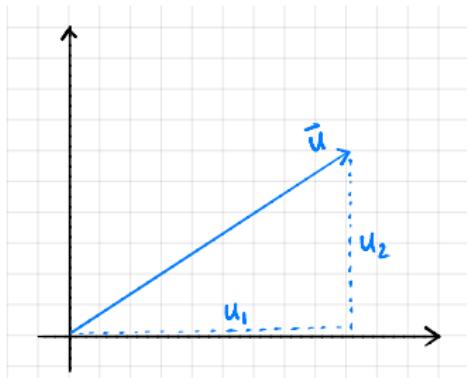
- ▶ An d -vector is an $d \times 1$ matrix.
- ▶ Often use arrow, lower-case letters to denote: \vec{x} .
- ▶ Often write $\vec{x} \in \mathbb{R}^d$ to say \vec{x} is a d vector.
- ▶ Example. A 4-vector:

$$\begin{pmatrix} 2 \\ 1 \\ 5 \\ -3 \end{pmatrix}$$

- ▶ Vector addition and scalar multiplication are also elementwise.

Geometric Meaning of Vectors

- ▶ A vector $\vec{u} = (u_1, \dots, u_d)^T$ is an arrow to the point (u_1, \dots, u_d) :



- ▶ The length, or **norm**, of \vec{u} is $\|\vec{u}\| = \sqrt{u_1^2 + u_2^2 + \dots + u_d^2}$.
- ▶ A **unit vector** is a vector of norm 1.

Dot Products

- ▶ The **dot product** of two d -vectors \vec{u} and \vec{v} is:

$$\vec{u} \cdot \vec{v} = \vec{u}^T \vec{v}$$

- ▶ Using low-level matrix multiplication definition:

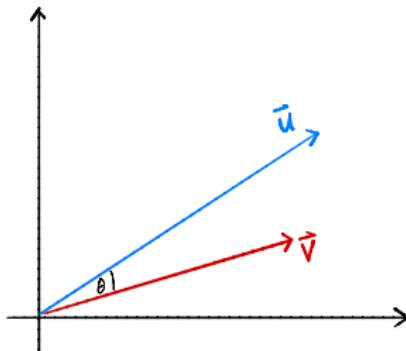
$$\begin{aligned}\vec{u} \cdot \vec{v} &= \sum_{i=1}^n u_i v_i \\ &= u_1 v_1 + u_2 v_2 + \dots + u_n v_n\end{aligned}$$

Dot Product Example

$$\vec{u} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \quad \vec{v} = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} \quad \vec{u} \cdot \vec{v} =$$

Geometric Interpretation of Dot Product

► $\vec{u} \cdot \vec{v} = \|\vec{u}\| \|\vec{v}\| \cos \theta.$



Discussion Question

Which of these is another expression for the norm of \vec{u} ?

- a) $\vec{u} \cdot \vec{u}$
- b) $\sqrt{\vec{u}^2}$
- c) $\sqrt{\vec{u} \cdot \vec{u}}$
- d) \vec{u}^2

Properties of the Dot Product

- ▶ Commutative: $\vec{u} \cdot \vec{v} = \vec{v} \cdot \vec{u}$
- ▶ Distributive: $\vec{u} \cdot (\vec{v} + \vec{w}) = \vec{u} \cdot \vec{v} + \vec{u} \cdot \vec{w}$
- ▶ Linear: $\vec{u} \cdot (\alpha\vec{v} + \beta\vec{w}) = \alpha\vec{u} \cdot \vec{v} + \beta\vec{u} \cdot \vec{w}$

Matrix-Vector Multiplication

- ▶ Special case of matrix-matrix multiplication.
- ▶ Result is always a vector with same number of rows as the matrix.
- ▶ One view: a “mixture” of the columns.

$$\begin{pmatrix} 1 & 2 & 1 \\ 3 & 4 & 5 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = a_1 \begin{pmatrix} 1 \\ 3 \end{pmatrix} + a_2 \begin{pmatrix} 2 \\ 4 \end{pmatrix} + a_3 \begin{pmatrix} 1 \\ 5 \end{pmatrix}$$

Matrices and Functions

- ▶ Matrix-vector multiplication takes in a vector, outputs a vector.
- ▶ An $m \times n$ matrix is an encoding of a function mapping \mathbb{R}^m to \mathbb{R}^n .
- ▶ Matrix multiplication evaluates that function.

For more, see www.dsc40a.com

Today

- ▶ How do we predict salary given **multiple** features?
 - ▶ years of experience, number of internships, GPA, etc.

Using Multiple Features

- ▶ We believe salary is a function of experience *and* GPA.
- ▶ I.e., there is a function H so that:

$$\text{salary} \approx H(\text{years of experience}, \text{GPA})$$

- ▶ Recall: H is a **prediction rule**.
- ▶ **Our goal:** find a good prediction rule, H .

Example Prediction Rules

$$H_1(\text{experience}, \text{GPA}) = \$40,000 \times \frac{\text{GPA}}{4.0} + \$2,000 \times (\text{experience})$$

$$H_2(\text{experience}, \text{GPA}) = \$60,000 \times 1.05^{(\text{experience}+\text{GPA})}$$

$$H_3(\text{experience}, \text{GPA}) = \sin(\text{GPA}) + \cos(\text{experience})$$

Linear Prediction Rule

- ▶ We'll restrict ourselves to **linear** prediction rules:

$$H(\text{experience}, \text{GPA}) = w_0 + w_1 \times (\text{experience}) + w_2 \times (\text{GPA})$$

- ▶ Can add more features, too¹:

$$H(\text{experience}, \text{GPA}, \# \text{ internships}) =$$

$$w_0 + w_1 \times (\text{experience}) + w_2 \times (\text{GPA}) + w_3 \times (\# \text{ of internships})$$

- ▶ Interpretation of w_i : the **weight** of feature x_i .

¹In practice, might use tens, hundreds, even thousands of features.

Feature Vectors

- In general, if x_1, \dots, x_d are d features:

$$H(x_1, \dots, x_d) = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_d x_d$$

- Nicer to pack into a **feature vector** and **parameter vector**:²

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \quad \vec{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \\ w_d \end{pmatrix}$$

²This slide originally had an error; see beginning of Lecture 09.

Augmented Feature Vectors

- The **augmented feature vector** $\text{Aug}(\vec{x})$ is the vector obtained by adding a 1 to the front of \vec{x} :

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \quad \text{Aug}(\vec{x}) = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \quad \vec{w} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{pmatrix}$$

- Then:

$$\begin{aligned} H(x_1, \dots, x_d) &= w_0 + w_1 x_1 + w_2 x_2 + \dots + w_d x_d \\ &= \text{Aug}(\vec{x}) \cdot \vec{w} \end{aligned}$$

Example

- ▶ Recall the prediction rule:

$$H_1(\text{experience}, \text{GPA}) = \$40,000 \times \frac{\text{GPA}}{4.0} + \$2,000 \times (\text{experience})$$

- ▶ This is linear. If x_1 is experience, x_2 is GPA, then:

$$\vec{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 2,000 \\ 10,000 \end{pmatrix}$$

- ▶ Prediction for someone with 2 years experience, 3.0 GPA:

$$\text{Aug}(\vec{x}) = \begin{pmatrix} & & \end{pmatrix} \quad H(\vec{x}) = \text{Aug}(\vec{x}) \cdot \vec{w} =$$

The Data

- ▶ For each person, collect 3 features, plus salary:

Person #	Experience	GPA	# Internships	Salary
1	3	3.7	1	85,000
2	6	3.3	2	95,000
3	10	3.1	3	105,000

- ▶ We represent each person with a **data vector**:

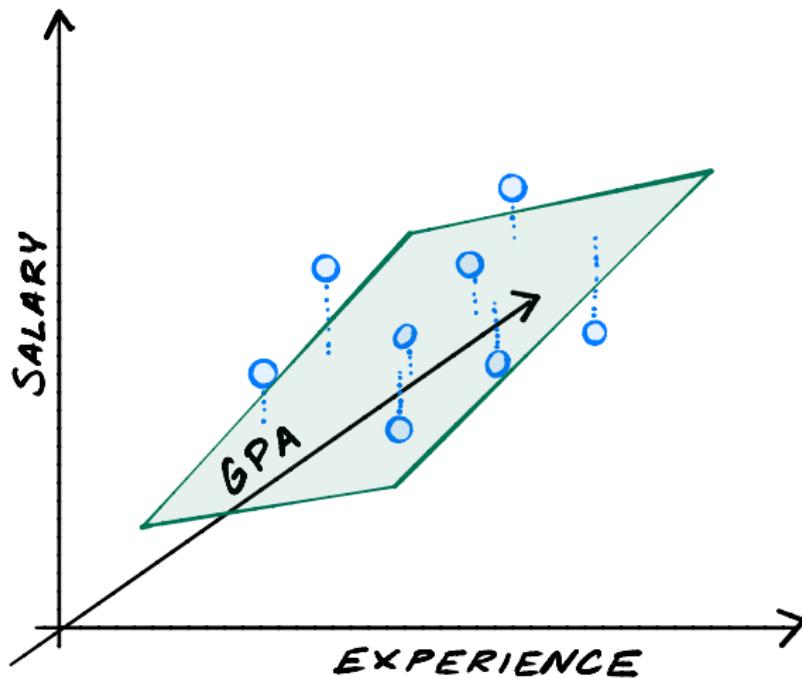
$$\vec{x}^{(1)} = \begin{pmatrix} 3 \\ 3.7 \\ 1 \end{pmatrix}, \quad \vec{x}^{(2)} = \begin{pmatrix} 6 \\ 3.3 \\ 2 \end{pmatrix}, \quad \vec{x}^{(3)} = \begin{pmatrix} 10 \\ 3.1 \\ 3 \end{pmatrix}$$

Notation

- ▶ $\vec{x}^{(i)}$ is the i th data vector.
- ▶ $x_j^{(i)}$ is the j th feature in the i th data vector.
- ▶ If there are d features:

$$\vec{x}^{(i)} = \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_d^{(i)} \end{pmatrix}$$

Geometric Interpretation



The General Problem

- ▶ We have n data points (or **training examples**):
 $(\vec{x}^{(1)}, y_1), \dots, (\vec{x}^{(n)}, y_n)$
- ▶ We want to find a good linear prediction rule:

$$H(\vec{x}) = \vec{w} \cdot \text{Aug}(\vec{x})$$

- ▶ To do so, we'll minimize the mean squared error:

$$\begin{aligned} R_{\text{sq}}(\vec{w}) &= \frac{1}{n} \sum_{i=1}^n (H(\vec{x}^{(i)}) - y_i)^2 \\ &= \frac{1}{n} \sum_{i=1}^n ((\vec{w} \cdot \text{Aug}(\vec{x}^{(i)})) - y_i)^2 \end{aligned}$$

The Risk

- ▶ With d features, we have $d + 1$ parameters: w_0, w_1, \dots, w_d .
- ▶ The risk $R_{\text{sq}}(\vec{w})$ is a function from \mathbb{R}^{d+1} to \mathbb{R}^1 .
- ▶ It is a $(d + 1)$ -dimensional hypersurface.
- ▶ **No hope of visualizing it directly when $d \geq 2$.**

Rewriting the Mean Squared Error

- Let \vec{e} be such that e_i is the (signed) error on i th example:

$$e_i = (\vec{w} \cdot \text{Aug}(\vec{x}^{(i)})) - y_i$$

- Then:

$$\begin{aligned} R_{\text{sq}}(\vec{w}) &= \frac{1}{n} \sum_{i=1}^n [(\vec{w} \cdot \text{Aug}(\vec{x}^{(i)})) - y_i]^2 \\ &= \frac{1}{n} \sum_{i=1}^n e_i^2 \end{aligned}$$

Rewriting the Mean Squared Error

- Let \vec{e} be such that e_i is the (signed) error on i th example:

$$e_i = (\vec{w} \cdot \text{Aug}(\vec{x}^{(i)})) - y_i$$

- Then:

$$\begin{aligned} R_{\text{sq}}(\vec{w}) &= \frac{1}{n} \sum_{i=1}^n [(\vec{w} \cdot \text{Aug}(\vec{x}^{(i)})) - y_i]^2 \\ &= \frac{1}{n} \sum_{i=1}^n e_i^2 \\ &= \end{aligned}$$

Rewriting the Mean Squared Error

- ▶ Define $\vec{y} = (y_1, \dots, y_n)^T$. Then:

$$\vec{e} = \begin{pmatrix} (\vec{w} \cdot \text{Aug}(\vec{x}^{(1)})) - y_1 \\ (\vec{w} \cdot \text{Aug}(\vec{x}^{(2)})) - y_2 \\ \vdots \\ (\vec{w} \cdot \text{Aug}(\vec{x}^{(n)})) - y_n \end{pmatrix} =$$

- ▶ \vec{h} is the vector of predictions.

Rewriting the Mean Squared Error

- ▶ So far: $R_{\text{sq}}(\vec{w}) = \frac{1}{n} \|\vec{e}\|^2$, and $\vec{e} = \vec{h} - \vec{y}$.

- ▶ Therefore:

$$R_{\text{sq}}(\vec{w}) = \frac{1}{n} \|\vec{h} - \vec{y}\|^2$$

- ▶ \vec{w} is hidden inside of \vec{h} , let's pull it out.

Rewriting the Mean Squared Error

- ▶ Define the **design matrix** X :

$$X = \begin{pmatrix} \text{Aug}(\vec{x}^{(1)}) \\ \text{Aug}(\vec{x}^{(2)}) \\ \vdots \\ \text{Aug}(\vec{x}^{(n)}) \end{pmatrix} = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_d^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \dots & x_d^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & \dots & x_d^{(n)} \end{pmatrix}$$

- ▶ Then $\vec{h} = X\vec{w}$.

Rewriting the Mean Squared Error

- ▶ The mean squared error is:

$$R_{\text{sq}}(\vec{w}) = \frac{1}{n} \|X\vec{w} - \vec{y}\|^2$$

where X is the **design matrix** containing the data, \vec{w} is the **parameter vector**, and \vec{y} is the vector of **observations** (or right answers).

- ▶ To minimize MSE: take derivative (gradient), set to zero, solve.

$$\begin{aligned} R_{\text{sq}}(\vec{w}) &= \|X\vec{w} - \vec{y}\|^2 \\ \nabla_{\vec{w}} R_{\text{sq}}(\vec{w}) &= \frac{d}{d\vec{w}} R_{\text{sq}}(\vec{w}) \\ &= 2X^T X \vec{w} - 2X^T \vec{y} \\ (X^T X) \vec{w} &= X^T \vec{y} \end{aligned}$$

DSC 40A

Lecture 08
Least Squares Regression, pt. IV

Last Time

- ▶ How do we make predictions using multiple features?
- ▶ Assume a linear decision rule:

$$H(\text{experience, GPA, \# internships}) =$$

$$w_0 + w_1 \times (\text{experience}) + w_2 \times (\text{GPA}) + w_3 \times (\#\text{ of internships})$$

- ▶ In general:

$$H(x_1, \dots, x_d) = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_d x_d$$

Feature Vectors

- Nicer to pack into a **feature vector** and **parameter vector**:

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \quad \vec{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix}$$

- Then: $H(\vec{x}) = w_0 + \vec{w} \cdot \vec{x}$

Feature Vectors

- Nicer to pack into a **feature vector** and **parameter vector**:

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \quad \vec{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix}$$

- Then: $H(\vec{x}) = w_0 + \vec{w} \cdot \vec{x}$
- Actually, we should include w_0 in \vec{w} ...

Augmented Feature Vectors

- The **augmented feature vector** $\text{Aug}(\vec{x})$ is the vector obtained by adding a 1 to the front of \vec{x} :

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \quad \text{Aug}(\vec{x}) = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \quad \vec{w} = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix}$$

- Then:

$$\begin{aligned} H(x_1, \dots, x_d) &= w_0 + w_1 x_1 + w_2 x_2 + \dots + w_d x_d \\ &= \text{Aug}(\vec{x}) \cdot \vec{w} \end{aligned}$$

Last Time

- ▶ We want to fit a decision rule of the form $H(\vec{x}) = \text{Aug}(\vec{x}) \cdot \vec{w}$.
- ▶ Minimize **mean squared error**:

$$R_{\text{sq}}(\vec{w}) = \frac{1}{n} \sum_{i=1}^n \left[(\vec{w} \cdot \text{Aug}(\vec{x}^{(i)})) - y_i \right]^2$$

Rewriting the Mean Squared Error

- ▶ Define the **design matrix**:

$$X = \begin{pmatrix} \text{Aug}(\vec{x}^{(1)}) \\ \text{Aug}(\vec{x}^{(2)}) \\ \vdots \\ \text{Aug}(\vec{x}^{(n)}) \end{pmatrix} = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_d^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \dots & x_d^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & \dots & x_d^{(n)} \end{pmatrix}$$

- ▶ And the vector of **observations**: $\vec{y} = (y_1, \dots, y_n)^T$

Rewriting the Mean Squared Error

- ▶ Then:

$$\begin{aligned} R_{\text{sq}}(\vec{w}) &= \frac{1}{n} \sum_{i=1}^n \left[(\vec{w} \cdot \text{Aug}(\vec{x}^{(i)})) - y_i \right]^2 \\ &= \frac{1}{n} \|X\vec{w} - \vec{y}\|^2 \end{aligned}$$

- ▶ Today's goal: find the \vec{w} that minimizes the MSE.

Minimizing the Mean Squared Error

- Our goal: minimize the function:

$$R_{\text{sq}}(\vec{w}) = \frac{1}{n} \|X\vec{w} - \vec{y}\|^2$$

- Strategy:

1. Take partial derivatives,

$$\frac{\partial R_{\text{sq}}}{\partial w_0}(\vec{w}), \quad \frac{\partial R_{\text{sq}}}{\partial w_1}(\vec{w}), \quad \frac{\partial R_{\text{sq}}}{\partial w_2}(\vec{w}), \quad \dots \quad \frac{\partial R_{\text{sq}}}{\partial w_d}(\vec{w})$$

2. Set each equal to zero and solve for w_0, w_1, \dots, w_d .

Minimizing the MSE: Gradient Edition

- The vector of partial derivatives is called the **gradient**:

$$\left(\frac{\partial R_{\text{sq}}}{\partial w_0}(\vec{w}), \quad \frac{\partial R_{\text{sq}}}{\partial w_1}(\vec{w}), \quad \frac{\partial R_{\text{sq}}}{\partial w_2}(\vec{w}), \quad \dots, \quad \frac{\partial R_{\text{sq}}}{\partial w_d}(\vec{w}) \right)^T$$

- Written: $\nabla_{\vec{w}} R_{\text{sq}}(\vec{w})$ or $\frac{dR_{\text{sq}}}{d\vec{w}}(\vec{w})$
- Strategy:
 - Compute the gradient of $R_{\text{sq}}(\vec{w})$.
 - Set it to zero and solve for \vec{w} .

Gradients Review

Computing Gradients

When computing $\frac{df}{d\vec{x}}(\vec{x})$:

- ▶ Before: make sure that f takes in vectors, outputs scalars.
 - ▶ **Example:** $\frac{d}{d\vec{x}} [A\vec{x}]$
 - ▶ **Example:** $\frac{d}{d\vec{x}} [\vec{x} \cdot \vec{x}], \frac{d}{d\vec{x}} [\vec{x}^T A^T A \vec{x}]$
- ▶ After: make sure your result is a vector.

Finding the Gradient: Strategy #1

Example: Find $\frac{d}{d\vec{x}} [\vec{a} \cdot \vec{x}]$ where \vec{x} and \vec{a} have d elements.

1. “Unpack” all matrix multiplications/dot products
► $\vec{a} \cdot \vec{x} =$

Finding the Gradient: Strategy #1

Example: Find $\frac{d}{d\vec{x}} [\vec{a} \cdot \vec{x}]$ where \vec{x} and \vec{a} have d elements.

1. “Unpack” all matrix multiplications/dot products

$$\triangleright \vec{a} \cdot \vec{x} = a_1x_1 + a_2x_2 + \dots + a_dx_d$$

2. Take partial derivatives (perhaps with arbitrary index):

$$\frac{\partial}{\partial x_1} [a_1x_1 + a_2x_2 + \dots + a_dx_d] =$$

$$\frac{\partial}{\partial x_2} [a_1x_1 + a_2x_2 + \dots + a_dx_d] =$$

$$\vdots$$

$$\frac{\partial}{\partial x_d} [a_1x_1 + a_2x_2 + \dots + a_dx_d] =$$

Finding the Gradient: Strategy #1

3. Pack partial derivatives into a gradient vector:

$$\frac{d}{d\vec{x}} [\vec{a} \cdot \vec{x}] = (a_1, a_2, \dots, a_d)^T$$

4. Simplify:

$$(a_1, a_2, \dots, a_d)^T = \vec{a}$$

- ▶ So $\frac{d}{d\vec{x}} [\vec{a} \cdot \vec{x}] = \vec{a}$
- ▶ Check: **result is a vector.**

Finding the Gradient: Strategy #1

- ▶ **Pro:** Always works, straightforward
- ▶ **Con:** Unpacking everything can get messy

Example

Show that $\frac{d}{d\vec{x}} [\vec{x}^T A^T A \vec{x}] = 2A^T A \vec{x}$, where A is $n \times d$ and \vec{x} is $n \times 1$.

- ▶ Check: **it is a scalar**

1. After unpacking: $\vec{x}^T A^T A \vec{x} = \sum_{i=1}^n \left(\sum_{j=1}^d A_{ij} x_j \right)^2$
2. Take partial derivatives:

$$\frac{\partial}{\partial x_1} \left[\sum_{i=1}^n \left(\sum_{j=1}^d A_{ij} x_j \right)^2 \right] = \sum_{i=1}^n \sum_{j=1}^d A_{i1} A_{ij} x_j$$

Example

3. Pack into a gradient vector:

$$\frac{d}{d\vec{x}} [\vec{x}^T A^T A \vec{x}] = \begin{pmatrix} \sum_{i=1}^n \sum_{j=1}^d A_{i1} A_{ij} x_j \\ \sum_{i=1}^n \sum_{j=1}^d A_{i2} A_{ij} x_j \\ \vdots \\ \sum_{i=1}^n \sum_{j=1}^d A_{id} A_{ij} x_j \end{pmatrix}$$

4. Somehow simplify this to $A^T A \vec{x}$...

Finding the Gradient: Strategy #2

Chain Rule: If $f : \mathbb{R} \rightarrow \mathbb{R}$, and $g : \mathbb{R}^d \rightarrow \mathbb{R}$, then:

$$\frac{d}{d\vec{x}} f(g(\vec{x})) = \frac{df}{dg} \frac{dg}{d\vec{x}}$$

Example: What is $\frac{d}{d\vec{x}} [(\vec{a} \cdot \vec{x})^2]?$

- ▶ $f(g) =$
- ▶ $g(\vec{x}) =$
- ▶ $\frac{d}{d\vec{x}} [(\vec{a} \cdot \vec{x})^2] =$

Finding the Gradient: Strategy #2

1. Unpack until we can use chain rule, but no more.
2. Use the chain rule.
3. Simplify.

Recall

Suppose A is $n \times d$.

Let \vec{A}_{i*} denotes its i th row. Then:

$$A\vec{x} = \begin{pmatrix} \vec{A}_{1*} \cdot \vec{x} \\ \vec{A}_{2*} \cdot \vec{x} \\ \vdots \\ \vec{A}_{n*} \cdot \vec{x} \end{pmatrix}$$

Let \vec{A}_{*j} denotes its j th column, then:

$$A\vec{x} = \vec{A}_{*1}x_1 + \vec{A}_{*2}x_2 + \dots + \vec{A}_{*d}x_d$$

Finding the Gradient: Strategy #2

Show that $\frac{d}{d\vec{x}} [\vec{x}^T A^T A \vec{x}] = 2A^T A \vec{x}$, where A is $n \times d$ and \vec{x} is $n \times 1$.

1. Unpack $\vec{x}^T A^T A \vec{x} =$

Finding the Gradient: Strategy #2

Show that $\frac{d}{d\vec{x}} [\vec{x}^T A^T A \vec{x}] = 2A^T A \vec{x}$, where A is $n \times d$ and \vec{x} is $n \times 1$.

2. Use chain rule:

Finding the Gradient: Strategy #2

Show that $\frac{d}{d\vec{x}} [\vec{x}^T A^T A \vec{x}] = 2A^T A \vec{x}$, where A is $n \times d$ and \vec{x} is $n \times 1$.

3. Show that this = $2A^T A \vec{x}$.

Back to Regression...

Minimizing the MSE

- We want to compute:

$$\frac{d}{d\vec{w}} \left[R_{\text{sq}}(\vec{w}) \right] = \frac{d}{d\vec{w}} \left[\|X\vec{w} - \vec{y}\|^2 \right]$$

- Step 1: Rewrite squared norm using dot product. Recall:

$$(A + B)^T = A^T + B^T$$

$$(AB)^T = B^T A^T$$

$$\vec{u} \cdot \vec{v} = \vec{v} \cdot \vec{u}$$

$$(\vec{u} + \vec{v}) \cdot (\vec{w} + \vec{z}) = \vec{u} \cdot \vec{w} + \vec{u} \cdot \vec{z} + \vec{v} \cdot \vec{w} + \vec{v} \cdot \vec{z}$$

$$\|\vec{u}\|^2 = \vec{u} \cdot \vec{u}$$

Step 1: Rewriting squared norm

$$\|X\vec{w} - \vec{y}\|^2 =$$

=

=

=

Step 2: Take gradients

$$\frac{d}{d\vec{w}} \left[R_{\text{sq}}(\vec{w}) \right] = \frac{d}{d\vec{w}} \left[\vec{w}^T X^T X \vec{w} - 2\vec{y}^T X \vec{w} + \vec{y}^T \vec{y} \right]$$

=

The Normal Equations

- ▶ To minimize $R_{\text{sq}}(\vec{w})$, set gradient to zero, solve for \vec{w} :

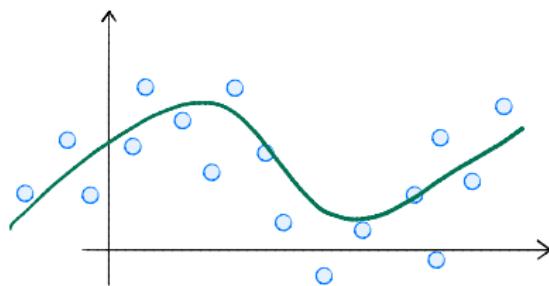
$$2X^T X \vec{w} - 2X^T \vec{y} = 0 \implies X^T X \vec{w} = X^T \vec{y}$$

- ▶ This is a system of equations in matrix form, called the **normal equations**.
- ▶ Solution¹: $\vec{w} = (X^T X)^{-1} X^T \vec{y}$.

¹Don't actually compute inverse! Use Gaussian elimination.

Regression with Multiple Features

- ▶ We want to find \vec{w} which minimizes $\|X\vec{w} - \vec{y}\|^2$.
- ▶ The answer: $\vec{w} = (X^T X)^{-1} X^T \vec{y}$.



DSC 40A
Lecture 10
Least Squares Regression, pt. II

Last Time

- ▶ How do we make predictions using multiple features?
- ▶ Assume a linear prediction rule:

$$\begin{aligned}H(x_1, \dots, x_d) &= w_0 + w_1 x_1 + w_2 x_2 + \dots + w_d x_d \\&= \text{Aug}(\vec{x}) \cdot \vec{w}\end{aligned}$$

- ▶ We found the normal equations:

$$X^T X \vec{w} = X^T \vec{y}$$

- ▶ Solving the normal equations for \vec{w} gives the best-fitting prediction rule.

Today

- ▶ Interpreting the results.
- ▶ How do we fit prediction rules like $H(x) = w_2x^2 + w_1x + w_0$?
- ▶ Least squares **classification**.

Interpreting \vec{w}

- ▶ With d features, \vec{w} has $d + 1$ entries.
- ▶ w_0 is the **bias**.
- ▶ w_1, \dots, w_d each give the **weight** of a feature.

$$H(\vec{x}) = w_0 + w_1 x_1 + \dots + w_d x_d$$

- ▶ Sign of w_i tells us about relationship between i th feature and outcome.

Example: Predicting Sales

- ▶ For each of 26 stores, we have:
 - ▶ net sales,
 - ▶ size (sq ft),
 - ▶ inventory,
 - ▶ advertising expenditure,
 - ▶ district size,
 - ▶ number of competing stores.
- ▶ Goal: predict net sales given size, inventory, etc.
- ▶ To begin:

$$H(\text{size, competitors}) = w_0 + w_1 \times \text{size} + w_2 \times \text{competitors}$$

Discussion Question

What will be the sign of w_1 and w_2 ?

- A) $w_1 = +, w_2 = -$
- B) $w_1 = +, w_2 = +$
- C) $w_1 = -, w_2 = -$
- D) $w_1 = -, w_2 = +$

$$H(\text{size, competitors}) = w_0 + w_1 \times \text{size} + w_2 \times \text{competitors}$$

(DEMO)

Discussion Question

Which has the greatest effect on the outcome?

- A) size: $w_1 = 16.20$
- B) inventory: $w_2 = 0.17$
- C) advertising: $w_3 = 11.53$
- D) district size: $w_4 = 13.58$
- E) competing stores: $w_5 = -5.31$

Which features are most “important”?

- ▶ Not necessarily the feature with largest weight.
- ▶ Features are measured in different units, scales.
- ▶ We should **standardize** each feature.

Standard Units

- ▶ To standardize (z-score) a feature, subtract mean, divide by standard deviation.
- ▶ Example: 10, 20, -30, 5, 15
 - ▶ Mean: 4
 - ▶ Standard Dev: $\sqrt{\frac{1}{5} \sum (x_i - \bar{x})^2} \approx 17.7}$
 - ▶ Standardized:

$$\frac{10 - 4}{17.7} = 0.34, \quad \frac{20 - 4}{17.7} = 0.90, \quad \frac{-30 - 4}{17.7} = -1.92,$$

$$\frac{5 - 4}{17.7} = 0.06, \quad \frac{15 - 4}{17.7} = 0.62$$

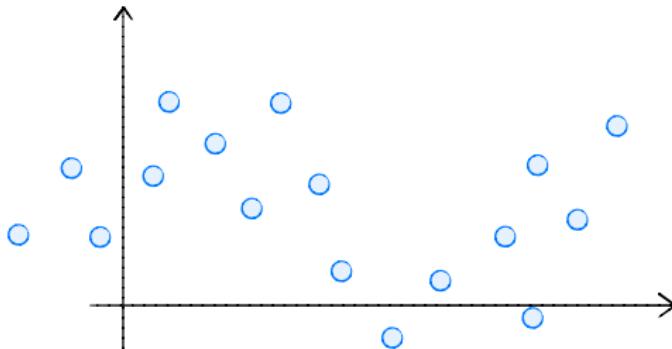
Standard Units

- ▶ Standardize each feature (store size, inventory, etc.) separately.
- ▶ No need to standardize outcome (net sales).
- ▶ Solve normal equations. The resulting w_0, w_1, \dots, w_d are called the **standardized regression coefficients**.
- ▶ They can be directly compared to one another.

(DEMO)

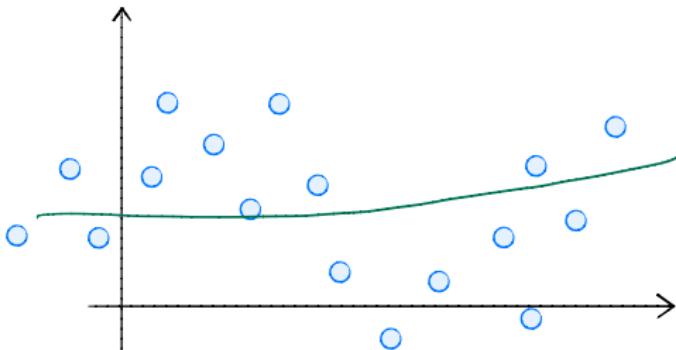
Fitting Non-Linear Patterns

- ▶ Fit a 4th-order polynomial to the data:



- ▶ We know how to fit rules of the form $H(x) = w_1 x^4 + w_0$.
 - ▶ Define $z_i = x_i^4$.
 - ▶ Use $w_1 = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})}{\sum(z_i - \bar{z})^2}$ and $w_0 = \bar{y} - w_1 \bar{z}$.

The Result



- ▶ The rule $H(x) = w_1 x^4 + w_0$ **underfits** the data.
- ▶ We need a more complicated rule:

$$H(x) = w_4 x^4 + w_3 x^3 + w_2 x^2 + w_1 x + w_0$$

The Trick

- ▶ Treat x, x^2, x^3, x^4 as different features.
- ▶ Create design matrix:

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 & x_1^4 \\ 1 & x_2 & x_2^2 & x_2^3 & x_2^4 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 & x_n^4 \end{pmatrix}$$

- ▶ Solve $X^T X \vec{w} = X^T \vec{y}$ for \vec{w} , as usual.
- ▶ Works for more than just polynomials.

(DEMO)

Polynomial Regression

- ▶ More complicated patterns can be fit with higher-order polynomials.
- ▶ If there are n points, a $n + 1$ degree polynomial can fit them exactly.
- ▶ But for high-order polynomials, it becomes **very hard** to solve the normal equations (numerical accuracy).

Polynomial Regression with Multiple Features

- ▶ Suppose we want to fit a rule of the form:

$$\begin{aligned}H(\text{size, competitors}) &= w_0 + w_1 \text{size} + w_2 \text{size}^2 \\&\quad + w_3 \text{competitors} + w_4 \text{competitors}^2 \\&= w_0 + w_1 s + w_2 s^2 + w_3 c + w_4 c^2\end{aligned}$$

- ▶ Make design matrix:

$$X = \begin{pmatrix} 1 & s_1 & s_1^2 & c_1 & c_1^2 \\ 1 & s_2 & s_2^2 & c_2 & c_2^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & s_n & s_n^2 & c_n & c_n^2 \end{pmatrix}$$

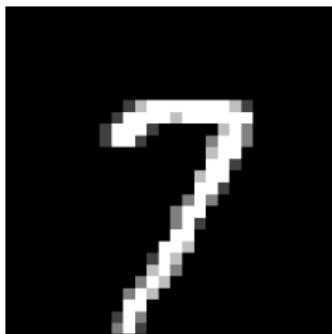
Where c_i and s_i are the competitors and size of the i th store.

Regression vs. Classification

- ▶ **Regression**: predict a number
 - ▶ Examples: salary, store sales, height of child
- ▶ **Classification**: predict a *class*, or *group label*.
 - ▶ is this person at high risk of disease (yes/no)?
 - ▶ what type of tree is in this (pine, elm, oak, etc.)?

Binary Classification

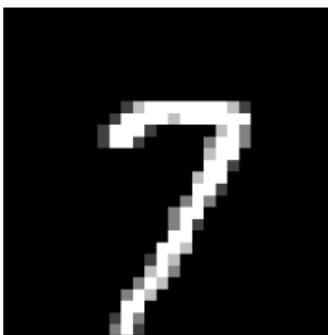
- ▶ There are two possible classes.
- ▶ Example: handwritten digits. Is image a 7, or a 3?



- ▶ Data: images $\vec{x}^{(i)}$, **labels** $y_i = 1$ if a seven, $y_1 = 0$ if a three.

Images as Feature Vectors

- ▶ We can pack an image into a feature vector.
- ▶ Each feature is the intensity of a particular pixel.
- ▶ Example: a 28×28 image has 784 pixels, becomes a vector in \mathbb{R}^{784} .



Decision Rule

- ▶ We want a rule $H(\vec{x})$ that takes in images and outputs:
 - ▶ 1 if image is a seven
 - ▶ 0 if image is a three
- ▶ We'll use a linear decision rule:

$$\begin{aligned}H(\text{image}) &= w_0 + w_1 \times (\text{pixel 1}) + \dots + w_{784} \times (\text{pixel 784}) \\&= \text{Aug}(\vec{x}) \cdot \vec{w}\end{aligned}$$

- ▶ Minimize MSE, same solutions:

$$R_{\text{sq}}(\vec{w}) = \frac{1}{n} \sum_{i=1}^n \left(\text{Aug}(\vec{x}^{(i)}) \cdot \vec{w} - \vec{y}_i \right)^2 \quad X^T X \vec{w} = X^T \vec{y}$$

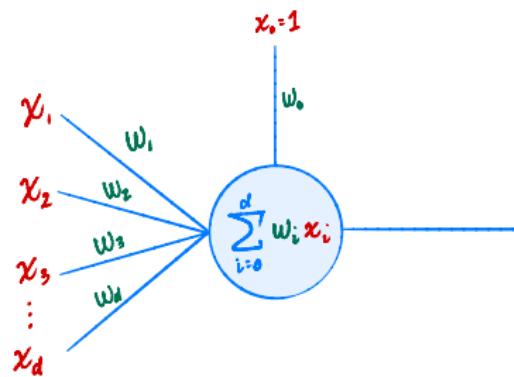
Least Squares Classification

- ▶ Our prediction $H(\vec{x})$ will not be 0 or 1 exactly.
- ▶ If $H(\vec{x}) > \frac{1}{2}$, we'll claim it is a 1; else, a 0.

(DEMO)

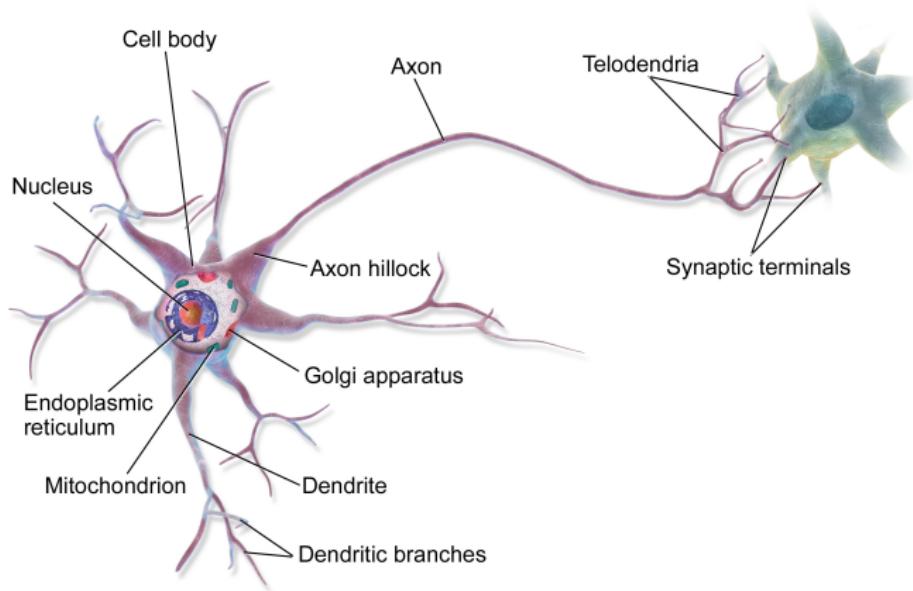
Least Squares Classification

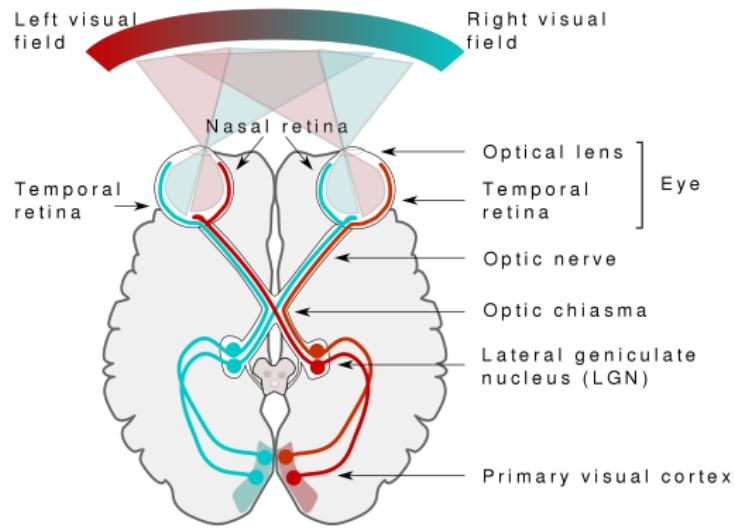
- ▶ Square loss is good for regression: want $H(\vec{x})$ close to right answer.
- ▶ Not great for classification.
- ▶ If real class is 1, and $H(x) = 10$, great!
- ▶ If real class is 1, and $H(x) = -1$, not great.
- ▶ Better loss functions: hinge loss, logistic loss, etc.

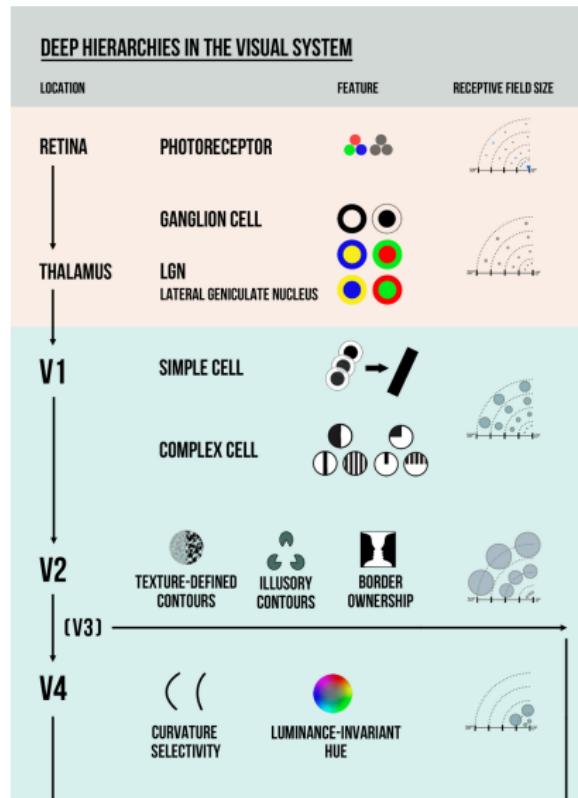


DSC 40A

Lecture 11
Perceptrons



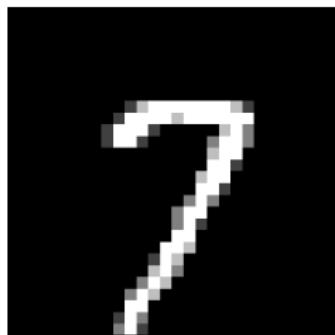




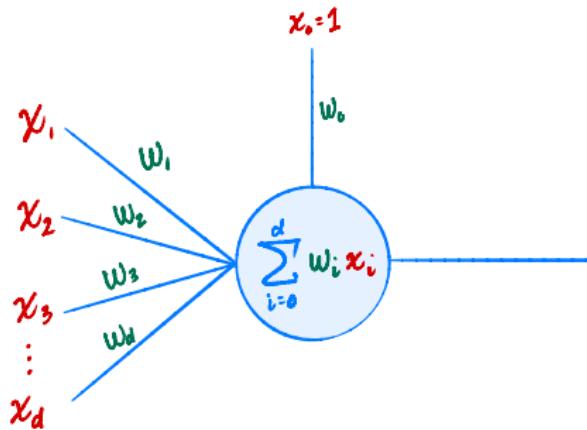


Today

- ▶ Design an artificial “neuron”, a **perceptron**.
- ▶ Train it to recognize images (handwritten digits).

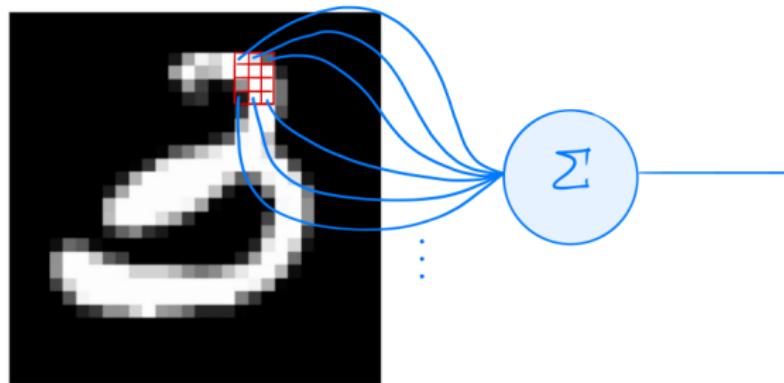


Modeling a Neuron



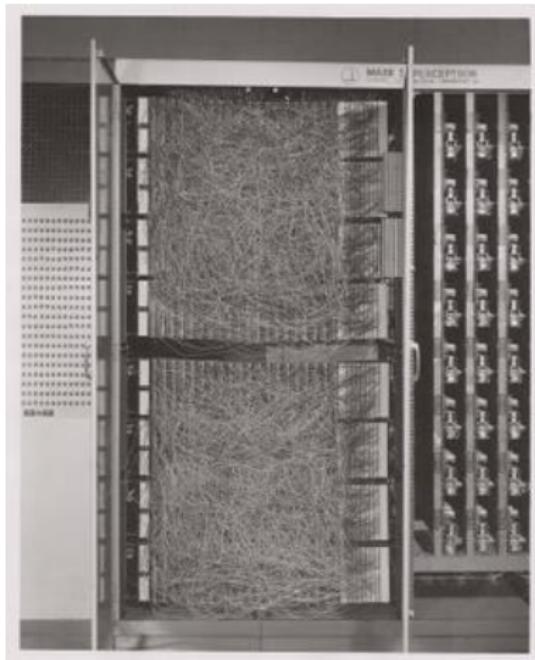
- ▶ Input: $\vec{x} = (x_1, \dots, x_d)^T$
- ▶ **Synapse weights:** $\vec{w} = (w_0, w_1, \dots, w_d)^T$
- ▶ Output: $\sum_{i=0}^d w_i x_i = \text{Aug}(\vec{x}) \cdot \vec{w}$
- ▶ This model is called a **perceptron**.

Example: Image Recognition



- ▶ Binary classifier:
 - ▶ If output > 0 , predicted digit is a three
 - ▶ If output < 0 , predicted digit is not a three

Rosenblatt's Perceptron (1958)



NEW NAVY DEVICE LEARNS BY DOING

Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser

WASHINGTON, July 7 (UPI)—The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

The embryo—the Weather Bureau's \$2,000,000 "704" computer—learned to differentiate between right and left after fifty attempts in the Navy's demonstration for newsmen.

The service said it would use this principle to build the first of its Perceptron thinking machines that will be able to read and write. It is expected to be finished in about a year at a cost of \$100,000.

Dr. Frank Rosenblatt, designer of the Perceptron, conducted the demonstration. He said the machine would be the first device to think as the human brain. As do human be-

ings, Perceptron will make mistakes at first, but will grow wiser as it gains experience, he said.

Dr. Rosenblatt, a research psychologist at the Cornell Aeronautical Laboratory, Buffalo, said Perceptrons might be fired to the planets as mechanical space explorers.

Without Human Controls

The Navy said the perceptron would be the first non-living mechanism "capable of receiving, recognizing and identifying its surroundings without any human training or control."

The "brain" is designed to remember images and information it has perceived itself. Ordinary computers remember only what is fed into them on punch cards or magnetic tape.

Later Perceptrons will be able to recognize people and call out their names and instantly translate speech in one language to speech or writing in another language, it was predicted.

Mr. Rosenblatt said in principle it would be possible to build brains that could reproduce themselves on an assembly line and which would be conscious of their existence.

1958 New York Times...

In today's demonstration, the "704" was fed two cards, one with squares marked on the left side and the other with squares on the right side.

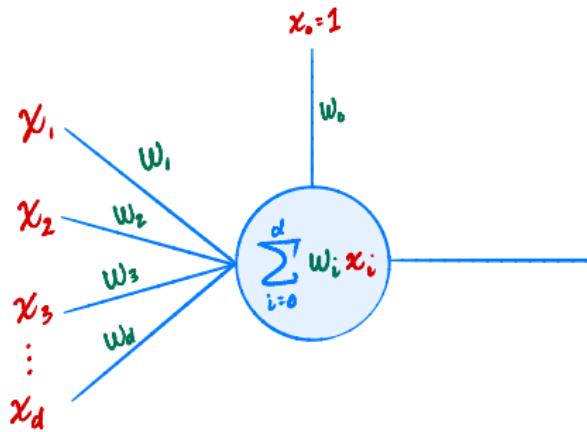
Learns by Doing

In the first fifty trials, the machine made no distinction between them. It then started registering a "Q" for the left squares and "O" for the right squares.

Dr. Rosenblatt said he could explain why the machine learned only in highly technical terms. But he said the computer had undergone a "self-induced change in the wiring diagram."

The first Perceptron will have about 1,000 electronic "association cells" receiving electrical impulses from an eye-like scanning device with 400 photo-cells. The human brain has 10,000,000,000 responsive cells, including 100,000,000 connections with the eyes.

Training a Perceptron



- ▶ A perceptron is trained by adjusting its weights, w_0, \dots, w_d .

Example: Predicting Survival

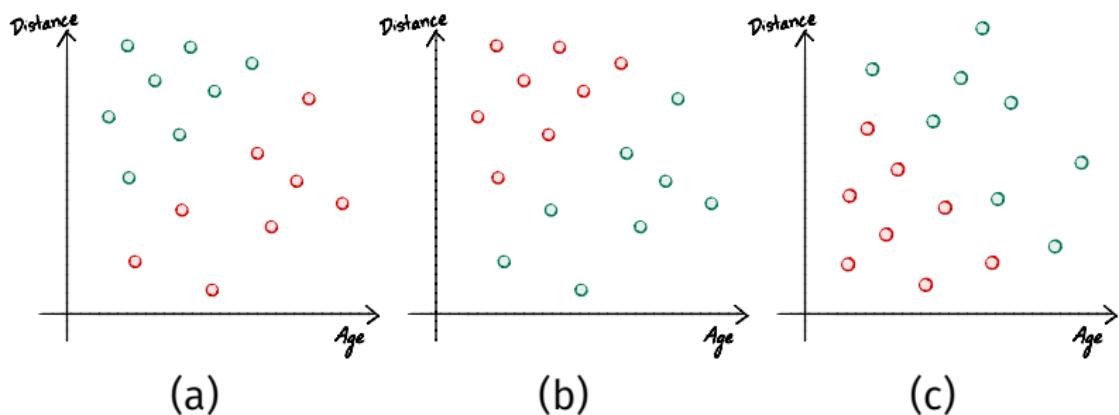
- ▶ We have a data set of hurricane survivors:

Age	Distance From Coast	Survived
21	30	1 (Yes)
84	2	-1 (No)
47	10	-1 (No)
30	15	1 (Yes)
:	:	:

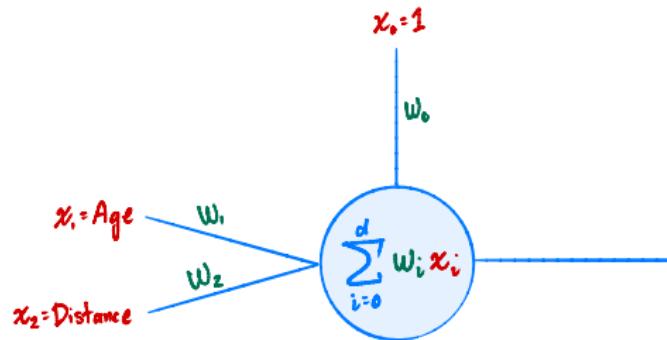
- ▶ Goal: train perceptron to predict if someone will survive.
 - ▶ If output is positive, prediction is “yes”
 - ▶ If output is negative, prediction is “no”

Discussion Question

Suppose we plot the data, with green points for people who survived, and red points for those who did not. Which do we see?



Example: Predicting Survival

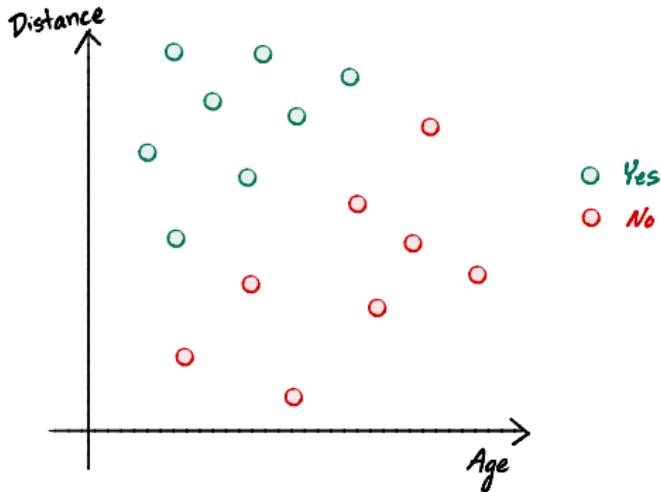


- ▶ Prediction rule:

$$H(\text{Age}, \text{Distance}) = w_0 + w_1 \times \text{Age} + w_2 \times \text{Distance}$$

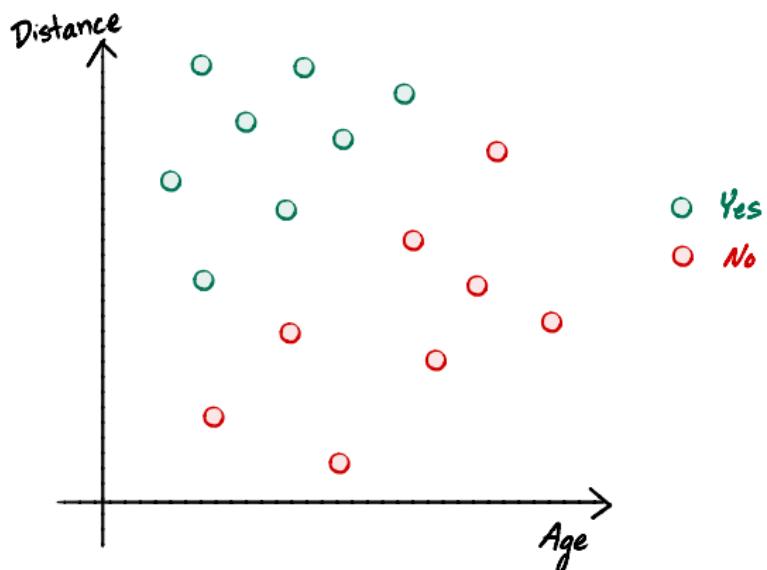
- ▶ If $H(x_1, x_2) > 0$, prediction is “yes”
- ▶ If $H(x_1, x_2) < 0$, prediction is “no”

The Decision Boundary



- ▶ The **decision boundary** is where $H = 0$.
- ▶ On one side of boundary, $H > 0$; prediction is **yes**.
- ▶ On other side, $H < 0$; prediction is **no**.
- ▶ **Important:** $|H(\vec{x})| \propto$ distance from boundary.

A Good Decision Boundary



Learning a Linear Decision Boundary

- ▶ Given feature vectors $\vec{x}^{(1)}, \dots, \vec{x}^{(n)}$ and labels y_1, \dots, y_n .
- ▶ Goal: find \vec{w} such that each point is **classified** correctly; i.e., it is on the correct side of boundary.
- ▶ We'll use empirical risk minimization.
- ▶ What is an appropriate loss function?

Loss #1: The 0-1 Loss

- ▶ How do we measure the error of a prediction?
- ▶ Four possibilities:
 - ▶ Correct label is -1 , and $H(\vec{x}) < 0$, classified **correctly**.
 - ▶ Correct label is -1 , and $H(\vec{x}) > 0$, classified **incorrectly**.
 - ▶ Correct label is 1 , and $H(\vec{x}) < 0$, classified **incorrectly**.
 - ▶ Correct label is 1 , and $H(\vec{x}) > 0$, classified **correctly**.
- ▶ The **0-1 loss**:

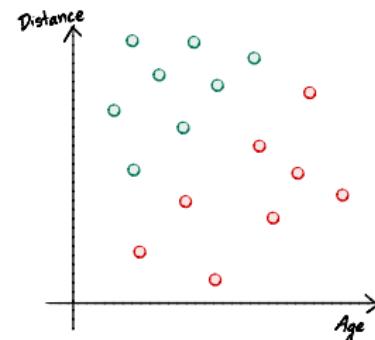
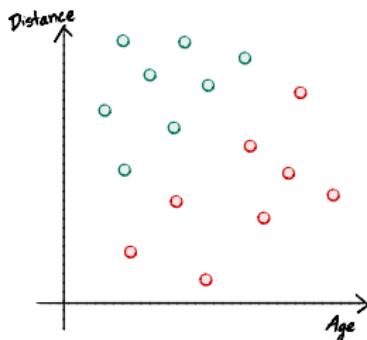
$$L_{01}(H(\vec{x}), y) = \begin{cases} 1, & \text{sign}(H(\vec{x})) \neq \text{sign}(y) \\ 0, & \text{sign}(H(\vec{x})) = \text{sign}(y) \end{cases}$$

Loss #1: The 0-1 Loss

- ▶ The **0-1 risk** is then:

$$R_{01}(\vec{w}) = \frac{1}{n} \sum_{i=1}^n \begin{cases} 1, & \text{sign}(H(\vec{x})) \neq \text{sign}(y) \\ 0, & \text{sign}(H(\vec{x})) = \text{sign}(y) \end{cases}$$

- ▶ Counts proportion of points which are misclassified.
- ▶ Example:



Goal: Minimize the 0-1 Risk

- ▶ Find \vec{w} which results in fewest # of misclassified points.
- ▶ That is, minimize

$$R_{01}(\vec{w}) = \frac{1}{n} \sum_{i=1}^n \begin{cases} 1, & \text{sign}(H(\vec{x}^{(i)})) \neq \text{sign}(y_i) \\ 0, & \text{sign}(H(\vec{x}^{(i)})) = \text{sign}(y_i) \end{cases}$$

- ▶ Gradient descent?

Gradient Descent for Functions of a Vector

- ▶ Pick α to be a positive number. It is the **learning rate**.
- ▶ Pick a starting parameter vector, $\vec{w}^{(0)}$.
- ▶ On step i , perform update $\vec{w}^{(i)} = \vec{w}^{(i-1)} - \alpha \cdot \frac{dR}{d\vec{w}}(\vec{w}^{(i-1)})$
- ▶ Repeat until convergence (when $\vec{w}^{(i)}$ doesn't change much).

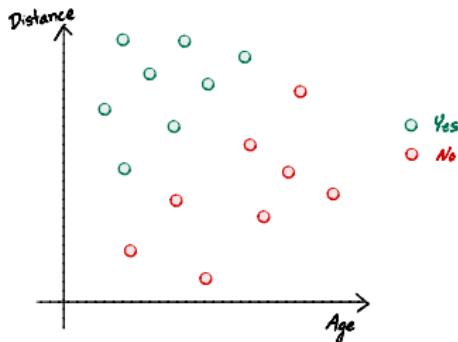
Discussion Question

Is

$$R_{01}(\vec{w}) = \frac{1}{n} \sum_{i=1}^n \begin{cases} 1, & \text{sign}(H(\vec{x}^{(i)})) \neq \text{sign}(y_i) \\ 0, & \text{sign}(H(\vec{x}^{(i)})) = \text{sign}(y_i) \end{cases}$$

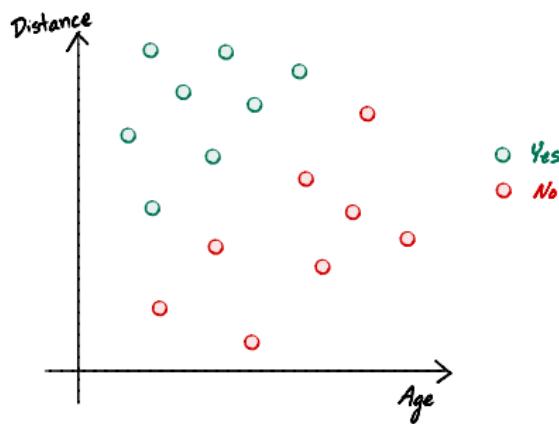
continuous? Differentiable?

- A) Continuous, not differentiable
- B) Continuous, differentiable
- C) Neither



The Problem

- ▶ R_{01} is flat.
- ▶ The gradient gives no information.



Loss #2: The Perceptron Loss

- ▶ We need a loss function that isn't flat.
- ▶ If prediction is wrong, size of $|H(\vec{x})|$ measures how wrong.
- ▶ The **perceptron loss**:

$$L_{\text{tron}}(H(\vec{x}), y) = \begin{cases} 0, & \text{sign}(H(\vec{x})) = \text{sign}(y) \\ |H(\vec{x})|, & \text{sign}(H(\vec{x})) \neq \text{sign}(y) \end{cases}$$

Loss #2: The Perceptron Loss

- ▶ The **perceptron risk**:

$$R_{\text{tron}}(\vec{w}) = \frac{1}{n} \sum_{i=1}^n \begin{cases} 0, & \text{sign}(H(\vec{x}^{(i)})) = \text{sign}(y_i) \\ |H(\vec{x}^{(i)})|, & \text{sign}(H(\vec{x}^{(i)})) \neq \text{sign}(y_i) \end{cases}$$

=

where M is the set of misclassified points.

- ▶ **Continuous, not differentiable, not flat.**

Gradient Descent for Perceptron Learning

- ▶ Compute gradient of:

$$R_{\text{tron}}(\vec{w}) = \frac{1}{n} \sum_{i \in M} |H(\vec{x}^{(i)})|$$

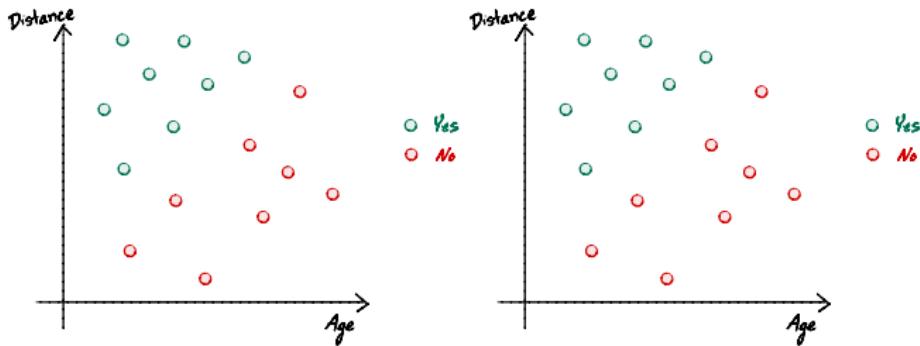
The Perceptron Algorithm

- ▶ Pick an initial $\vec{w}^{(0)}$.
- ▶ On iteration t :
 - ▶ Construct set M of misclassified points using $\vec{w}^{(t-1)}$
 - ▶ If M is empty, break (no points misclassified).
 - ▶ Otherwise, perform update:

$$\vec{w}^{(t)} = \vec{w}^{(t-1)} - \frac{\alpha}{n} \sum_{i \in M} \begin{cases} \text{Aug}(\vec{x}^{(i)}), & \vec{w}^{(t-1)} \cdot \text{Aug}(\vec{x}^{(i)}) \geq 0 \\ -\text{Aug}(\vec{x}^{(i)}), & \vec{w}^{(t-1)} \cdot \text{Aug}(\vec{x}^{(i)}) < 0 \end{cases}$$

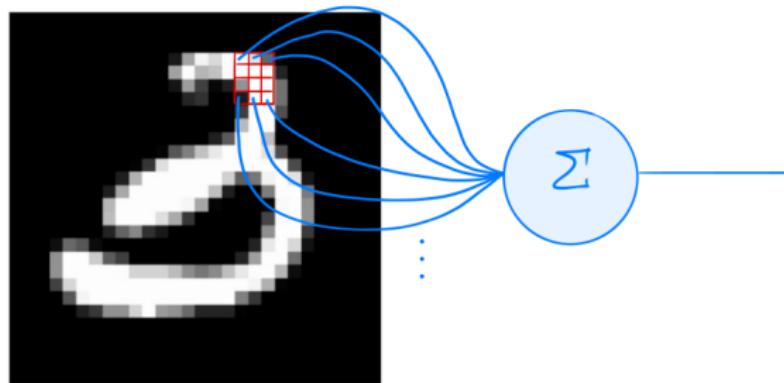
The Perceptron Algorithm

- ▶ Data is **linearly separable** if classes can be separated by a line (plane):



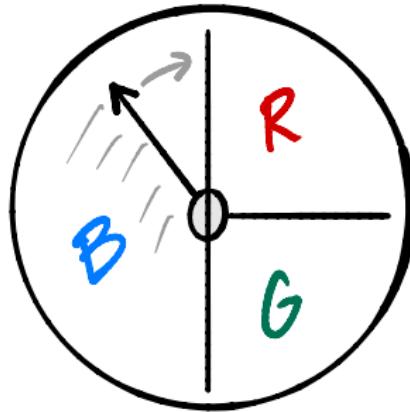
- ▶ If linearly separable, perceptron algorithm will terminate; classify all points correctly.

Example: Image Recognition



- ▶ Binary classifier:
 - ▶ If output > 0 , predicted digit is a three
 - ▶ If output < 0 , predicted digit is not a three

(DEMO)



DSC 40A

Lecture 12

Probability

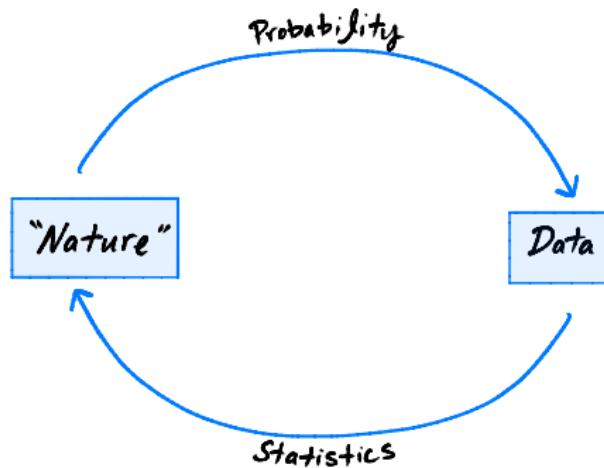
Suggested Reading

Chapter 1.2 of Grinstead and Snell

Why Probability

- ▶ We use data to make decisions.
- ▶ But the data could have been different.
- ▶ **Probability:** how different?

Probability vs. Statistics



The Language of Probability: Set Theory

- ▶ A **set** is a collection of distinct items.
 - ▶ Example: the six colleges. **Finite.**
 $\{ \text{Marshall, Roosevelt, Warren, Muir, Revelle, Sixth} \}$
 - ▶ Example: positive integers. **Discrete, infinite.**
 $\{ 1, 2, 3, 4, \dots \}$
 - ▶ Example: all real numbers. **Continuous, infinite.**

Sets

- ▶ Sets are **unordered**.
- ▶ They do not contain **duplicates**.

The Empty Set

- ▶ The **empty set** is the set with nothing in it.
- ▶ Written {} or \emptyset .

Elements

- ▶ The things in a set are called **elements**.
- ▶ Use $x \in A$ to denote that x is an element of A :
 - ▶ $3 \in \{1, 2, 3, 4\}$
 - ▶ $1.7 \notin \{1, 2, 3, 4\}$
- ▶ The **size** of a set A , written $|A|$, is the number of elements it contains.
 - ▶ $|\{1, 2, 3\}| = 3$

Subsets

- ▶ If every element of set A is in set B , then A is a **subset** of B .
- ▶ Written $A \subset B$ (or sometimes $A \subseteq B$).
- ▶ Examples:
 - ▶ $\{1, 4\} \subset \{1, 2, 3, 4\}$.
 - ▶ $\{1, 2, 3, 4\} \subset \{1, 2, 3, 4\}$.
- ▶ If $A \subset B$ and $B \subset A$, then $A = B$.

Discussion Question

Let $S = \{1, 2, 3, 4\}$. Which of these is true?

- A) $\emptyset \not\subset S$ and $\emptyset \in S$.
- B) $\emptyset \not\subset S$ and $\emptyset \notin S$.
- C) $\emptyset \subset S$ and $\emptyset \in S$.
- D) $\emptyset \subset S$ and $\emptyset \notin S$.

Intersection

- ▶ The **intersection** of sets A and B is the set containing all elements that are in **both** A and B .
- ▶ Written $A \cap B$.
- ▶ Examples:
 - ▶ $\{1, 2, 4\} \cap \{2, 3, 4\} =$
 - ▶ $\{1, 2\} \cap \{3, 4\} =$
- ▶ If $A \cap B = \emptyset$, A and B are said to be **disjoint**.

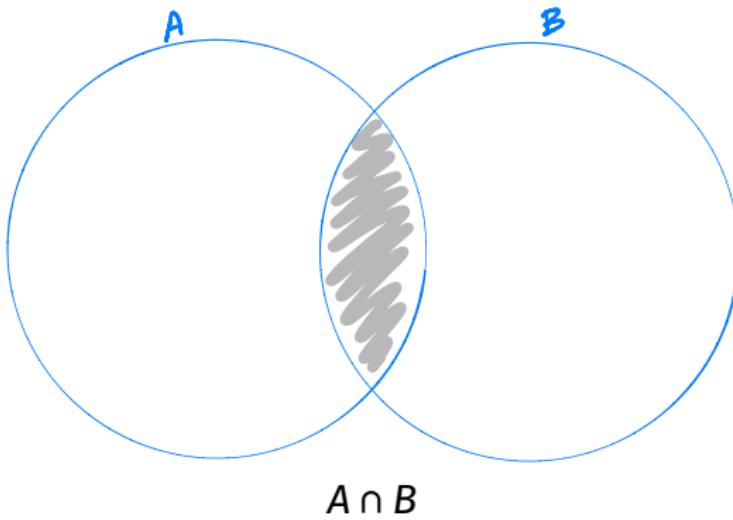
Union

- ▶ The **union** of sets A and B is the set containing all elements that are in **at least one** of A or B .
- ▶ Written $A \cup B$.
- ▶ Examples:
 - ▶ $\{1, 2\} \cup \{2, 3, 4\} =$
 - ▶ $\{1\} \cup \{2\} \cup \{3\} \cup \emptyset =$

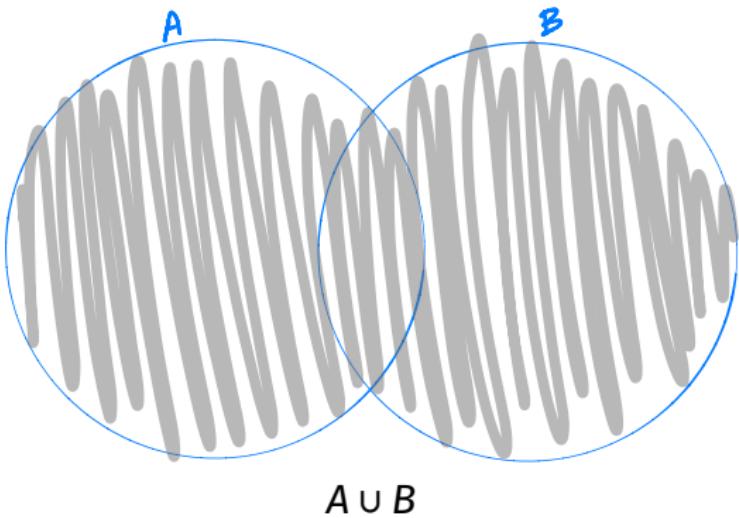
Difference

- ▶ The **difference** $A - B$ is the set of all elements that are in A and not in B .
- ▶ Examples:
 - ▶ $\{1, 2, 3, 4\} - \{3, 5, 6\} =$
 - ▶ $\emptyset - \{1, 2, 3\} =$

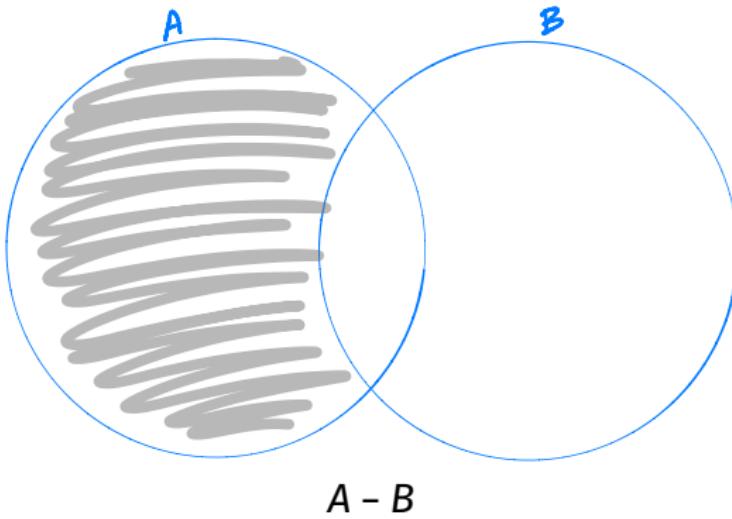
Venn Diagrams



Venn Diagrams



Venn Diagrams



Tuples

- ▶ A **tuple** is an ordered sequence.
 - ▶ A 2-tuple is an **ordered pair**.
- ▶ Example: result of flipping coin four times.
 $(\text{Heads}, \text{Tails}, \text{Heads}, \text{Heads})$
- ▶ Example: a point in three dimensions.
 $(3, -1, 2)$

Tuples

- ▶ Tuples are **ordered**.
- ▶ Duplicates **are allowed**.

Products of Sets

- ▶ Options for dinner: {sushi, tacos}
- ▶ Options for dessert: {ice cream, milk tea, espresso}
- ▶ Set of all possibilities for dinner/dessert:

(sushi, ice cream)
(sushi, milk tea)
(sushi, espresso)
(tacos, ice cream)
(tacos, milk tea)
(tacos, espresso)

Products of Sets

- ▶ The **Cartesian Product** of sets A and B , written $A \times B$, is the **set** of all ordered pairs (**2-tuples**) whose:
 - ▶ first element is in A
 - ▶ second element is in B
- ▶ Example: $\{1, 2\} \times \{a, b, c\}$.
- ▶ Example: $\{1, 2\} \times \{1, 2\}$.

Discussion Question

Which of these correctly gives the size of the Cartesian product of A and B ?

- A) $|A \times B| = |A| + |B|$
- B) $|A \times B| = |A| \cdot |B|$
- C) $|A \times B| = |A|^{|B|}$
- D) $|A \times B| = |B|^{|A|}$

Experiments

- ▶ An **experiment** is something whose outcome appears to be random.
- ▶ Examples:
 - ▶ Rolling a die.
 - ▶ Flipping a coin, twice.
 - ▶ Asking someone what college they're in.
 - ▶ Looking for an open parking spot in Hopkins Parking Structure.

Outcomes

- ▶ An **outcome** is the result of an experiment.
- ▶ The **sample space**, Ω , is the set of all outcomes of an experiment.
 - ▶ Experiment: Rolling a die.
Possible outcomes: {1, 2, 3, 4, 5, 6}
 - ▶ Experiment: Flipping a coin, twice.
Possible outcomes:
$$\{H, T\} \times \{H, T\} = \{(H, H), (H, T), (T, H), (T, T)\}$$
 - ▶ Experiment: Looking for parking in Hopkins.
Possible outcomes: {Spots, No Spots}

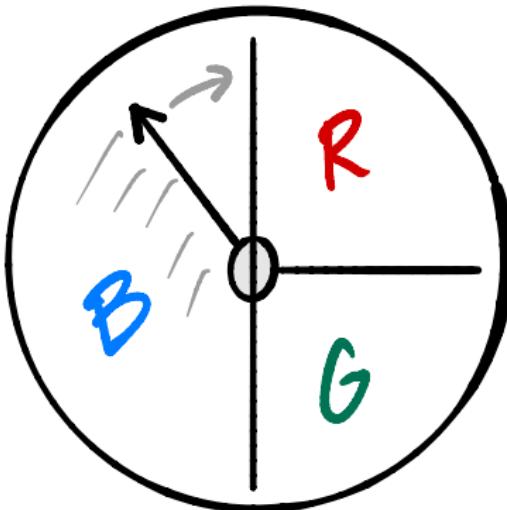
Discrete vs. Continuous Probability

- ▶ The sample space can be discrete or continuous.
- ▶ Discrete: rolling a die.
- ▶ Continuous: measuring temperature.
- ▶ We'll focus on **discrete** setting.

Probability

- ▶ The **probability** of an outcome is the proportion of times it happens if the experiment is repeated an infinite number of times.
- ▶ Example: probability of seeing Heads is $1/2$.
- ▶ Example: probability of rolling a 3 is $1/6$.
- ▶ Outcomes need not be equally-probable!

Example



- ▶ Outcomes: {R, G, B}
- ▶ Probability of B: $1/2$. Probability of R and G: $1/4$, each.

Probability Distribution Function

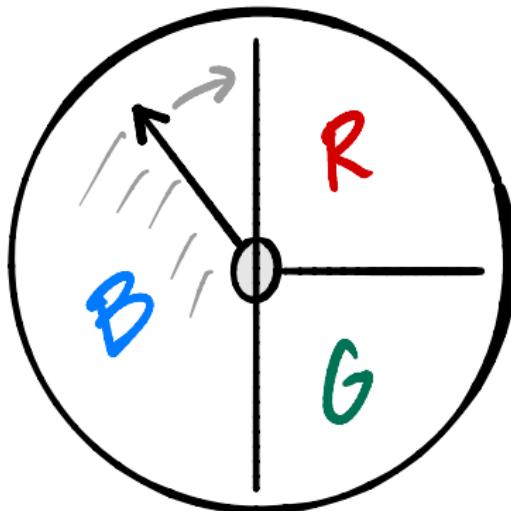
- ▶ A **probability distribution function** $m(\omega)$ assigns a probability to every outcome $\omega \in \Omega$.
- ▶ Requirement #1: probabilities are ≥ 0 .

$$m(\omega) \geq 0$$

- ▶ Requirement #2: probabilities sum to 1.

$$\sum_{\omega \in \Omega} m(\omega) = 1$$

Example



- ▶ $m(B) = 1/2, \quad m(R) = 1/4, \quad m(G) = 1/4.$

Events

- ▶ An **event** is a set of outcomes.
- ▶ An event “happens” if the result of the experiment is contained in the event.
- ▶ Example:
 - ▶ Experiment: rolling a die.
 - ▶ Sample space: $\{1, 2, 3, 4, 5, 6\}$.
 - ▶ Event: $\{2, 4, 6\}$ (i.e., rolling an even number).

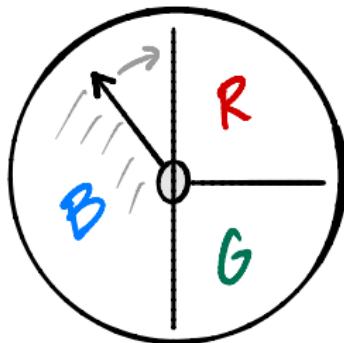
Probability of an Event

- ▶ The **probability** of an event E , written $P(E)$ is the sum of the probabilities of the elements of E :

$$P(E) = \sum_{\omega \in E} m(\omega)$$

Example

- ▶ What is the probability of spinning either a **G** or a **B**?



- ▶ $E =$
- ▶ $P(E) =$

Equally-Probable Outcomes

- ▶ If all of the outcomes are equally-probable, then

$$P(E) = \frac{|E|}{|\Omega|}$$

- ▶ Proof:

$$P(E) = \sum_{\omega \in E} m(\omega) = \sum_{\omega \in E} \frac{1}{|\Omega|} = \frac{|E|}{|\Omega|}$$

Example

- ▶ what is the probability of rolling an even number?
 - ▶ $E =$
 - ▶ $|E| =$
 - ▶ $|\Omega| =$
 - ▶ $P(E) =$

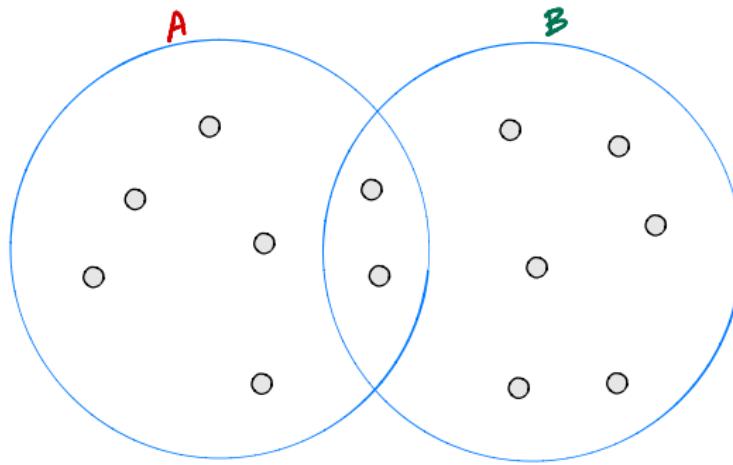
Combining Events

- ▶ The event that “**A or B**” happens = $A \cup B$.
- ▶ The event that “**A and B**” happens = $A \cap B$.
- ▶ The event that “**A but not B**” happens = $A - B$.
- ▶ The event that “**A doesn’t**” happen = $\Omega - A$.

Example

- ▶ What is the probability of rolling an even number ≤ 3 ?
 - ▶ $A =$
 - ▶ $B =$
 - ▶ $|A \cap B| =$
 - ▶ $P(A \cap B) =$

Probability of a Union



- ▶ $P(A \cup B) = P(A) + P(B)$?
- ▶ $P(A \cup B) =$



DSC 40A

Lecture 13

Combinatorics

Example

- ▶ What is the probability of seeing exactly 2 heads in 3 flips of a fair coin?
- ▶ **Sample space:** ordered triples.

(T, T, T), (T, T, H), (T, H, T), (H, T, T),
(H, H, T), (H, T, H), (T, H, H), (H, H, H).

- ▶ **Event:** {(H, H, T), (H, T, H), (T, H, H)}
- ▶ All outcomes equally-likely, so:

$$P(E) = \frac{|E|}{|\Omega|} = \frac{3}{8}$$

Example

- ▶ What is the probability of seeing exactly 40 heads in 100 flips of a fair coin?
- ▶ All outcomes equally-likely, so:

$$P(E) = \frac{|E|}{|\Omega|} = \frac{?}{?}$$

Today

How do we count the number of outcomes, besides enumerating them all?

- ▶ How many outcomes are possible if a die is rolled 100 times?
- ▶ How many different ways are there to shuffle 52 cards?
- ▶ How many ways are there to choose a jury of 12 people from a panel of 100?

The area of math concerned with counting is called **combinatorics**.

Sampling

- ▶ Many experiments involve choosing things from a set P called the **population**.
- ▶ Examples: drawing cards from a deck, selecting people for a survey, rolling a die.
- ▶ Two decisions to make:
 - ▶ With or without **replacement**?
 - ▶ Does the **order** in which things are selected matter?

Sequences, Permutations, and Combinations

- ▶ # of **sequences**: with replacement, order matters
- ▶ # of **permutations**: without replacement, order matters
- ▶ # of **combinations**: without replacement, order doesn't matter

Sequences

- ▶ A ***k*-sequence** is a **tuple**¹ obtained by selecting things from P **with replacement**.
- ▶ Example: draw a card, put it back, repeat four more times.

$(A\heartsuit, 2\clubsuit, 6\spadesuit, A\heartsuit, 3\diamondsuit)$

- ▶ Example: flip a coin 100 times.

$(H, T, T, H, \dots, H, T, T, T)$

¹tuples are ordered!

Example: Flip a coin three times

- ▶ Possible outcomes:

- ▶ Two choices for first item.
- ▶ For each choice of first item, two choices for second.
- ▶ For each choice of first two items, two choices for third.
- ▶ In total: $2 \cdot 2 \cdot 2 = 2^3$

Counting Sequences

- ▶ How many sequences of length k are there?
 - ▶ Remember: P is the **population**.
- ▶ $|P|$ choices for first item.
- ▶ For each choice of first item, $|P|$ choices for second.
- ▶ ...
- ▶ For each choice of first $k - 1$ items, $|P|$ choices for k th.
- ▶
$$\underbrace{|P| \cdot |P| \cdots |P|}_{k \text{ times}} = |P|^k$$

Counting Sequences (Another View)

- ▶ A sequence of length k is the Cartesian product of P with itself, k times.
- ▶ there are $|\underbrace{P \times P \times \dots \times P}_{k \text{ times}}| = |\underbrace{|P| \cdot |P| \dots |P|}_{k \text{ times}}| = |P|^k$.

Example

- ▶ Draw a card, put it back, repeat four more times.
- ▶ $|P| =$
- ▶ $k =$
- ▶ Number of possible outcomes:

Discussion Question

How many possible outcomes are there if a coin is flipped 100 times?

- A) 2^{100}
- B) 100^2
- C) $2 \cdot 100$
- D) 4^{100}

Example

- ▶ Flip a coin 100 times.
- ▶ $|P| =$
- ▶ $k =$
- ▶ Number of possible outcomes:

Exponential growth

- ▶ There are a lot of possible sequences.

n	# of Sequences of Length n
5	$2^5 = 32$

Exponential growth

- ▶ There are a lot of possible sequences.

n	# of Sequences of Length n
5	$2^5 = 32$
10	$2^{10} = 1024$

Exponential growth

- ▶ There are a lot of possible sequences.

n	# of Sequences of Length n
5	$2^5 = 32$
10	$2^{10} = 1024$
15	$2^{15} = 32,758$

Exponential growth

- ▶ There are a lot of possible sequences.

n	# of Sequences of Length n
5	$2^5 = 32$
10	$2^{10} = 1024$
15	$2^{15} = 32,758$
20	$2^{15} \approx 1 \text{ million}$

Exponential growth

- ▶ There are a lot of possible sequences.

n	# of Sequences of Length n
5	$2^5 = 32$
10	$2^{10} = 1024$
15	$2^{15} = 32,758$
20	$2^{15} \approx 1 \text{ million}$
50	$2^{50} \approx \# \text{ of grains of sand on Earth}$

Permutations

- ▶ A ***k*-permutation** is a **tuple** obtained by selecting ***k*** things from ***P* without replacement**.
- ▶ Example: draw a card, **don't** put it back, repeat four more times.

(A♥, 2♣, 6♠, 7♥, 3♦)

- ▶ Example: rank all 6 colleges by preference.
(Warren, Sixth, Muir, Roosevelt, Marshall, Revelle)
- ▶ Example: rank top four movies from a list of 250.
(Fistful of Dollars, Parasite, Psycho, Hot Rod)

Example: Ranking top two cities

- ▶ How many ways are there to rank top two out of {LA, SD, SF, SJ}?

- ▶ Four choices for first city.

- ▶ For each choice of first city, three choices for second.

- ▶ $4 \cdot 3$ possible rankings.

Counting Permutations

- ▶ $|P|$ choices for first item.
- ▶ For each choice of first item, $|P| - 1$ choices for second.
- ▶ For each choice of first two items, $|P| - 2$ choices for second.
- ▶ ...
- ▶ For each choice of first $k - 1$ items, $|P| - (k - 1)$ choices for k th.
- ▶ $|P| \cdot (|P| - 1) \cdot (|P| - 2) \cdots (|P| - k + 1)$

Another Formula for Counting Permutations

- ▶ The number of k -permutations is $|P| \cdot (|P| - 1) \cdots (|P| - k + 1)$.
- ▶ Equivalently:

$$\frac{|P|!}{(|P| - k)!}$$

Special Case

- ▶ Suppose $k = |P|$.
- ▶ How many permutations are there of $|P|$ items?
- ▶ $|P|! = |P| \cdot (|P| - 1) \cdot (|P| - 2) \cdots 3 \cdot 2 \cdot 1$

Example

- ▶ Rank all 6 colleges by preference.
- ▶ $|P| =$
- ▶ $k =$
- ▶ Number of possible rankings:

Example

- ▶ Rank top four movies out of a list of 250.
- ▶ $|P| =$
- ▶ $k =$
- ▶ Number of possible rankings:

Combinations

- ▶ A ***k*-combination** is a **set** obtained by selecting *k* things from *P* without replacement.
- ▶ Example: draw a hand of five cards from a deck of 52.

{A♥, 2♣, 6♣, 7♥, 3♦}

- ▶ Example: make a group of 5 people from a class of 100.

{Clint, Zelda, Alfred, Andy, Yvonne}

Counting Combinations

- ▶ How many ways are there to choose 2 unique things from $P = \{a, b, c\}$?
- ▶ Step 1) Enumerate all $3!/2!$ 2-permutations:

$(a, b), (a, c), (b, a), (b, c), (c, a), (c, b)$

- ▶ Step 2) Group the k -permutations which have same elements.

$(a, b), (a, c), (b, a), (b, c), (c, a), (c, b)$

- ▶ Step 3) Count number of groups: $(3!/2!) / 2! = 3!/(2! \cdot 2!)$

Example: Choose two cities

- ▶ How many ways are there of choosing two cities from {LA, SD, SF, SJ}?

- ▶ ?

Counting Combinations

- ▶ How many ways are there to choose k unique things from P ?
- ▶ Step 1) Enumerate all $|P|!/(|P| - k)$ **k -permutations**:
- ▶ Step 2) Group the k -permutations which have same elements.
- ▶ Step 3) Count number of groups:

$$\# \text{ of } k\text{-combinations} = \frac{|P|!}{k!(|P| - k)!}$$

Counting Combinations

- ▶ The number of ways of choosing k items from n possibilities is often called **n choose k**, written:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- ▶ Also called the **binomial coefficients**.

Example

- ▶ How many different hands of five cards are there?
- ▶ $|P| =$
- ▶ $k =$
- ▶ # of hands:

Example

- ▶ How many different ways are there to select a group of 5 people from 100?
- ▶ $|P| =$
- ▶ $k =$
- ▶ # of ways

Counting and Probability

- ▶ When outcomes are equiprobable, $P(E) = |E|/|\Omega|$
- ▶ To find $|E|$, we often need to count sequences, permutations, or combinations.
- ▶ Must decide if order matters.
- ▶ **Pro tip:** think about the sample space first!

Example: Groups from Warren

- ▶ A class of 100 contains 30 people from Warren college.
- ▶ What is the probability that a group of 5 randomly-selected people are all from Warren?
- ▶ Does order matter?
- ▶ What is the sample space? That is, what is an outcome?

Example: Groups from Warren

- ▶ An outcome is a **set** of five people.
- ▶ Sample space, Ω : all possible sets of five people.

$$|\Omega| =$$

- ▶ Event, E : all possible sets of five people from Warren.

$$|E| =$$

- ▶ $P(E) = |E|/|\Omega| =$

Example: 40 Heads

- ▶ What is the probability of seeing exactly 40 heads in 100 flips of a fair coin?
- ▶ Does order matter? No. But also yes.
- ▶ **Decide on your sample space!**

Example: 40 Heads

- ▶ An outcome is a sequence of 100 flips.
- ▶ Sample space, Ω : all possible sequences of 100 flips.

$$|\Omega| =$$

- ▶ Event, E : all sequences with exactly 40 heads.

$$|E| =$$

Example: 40 Heads

$$|E|$$

=

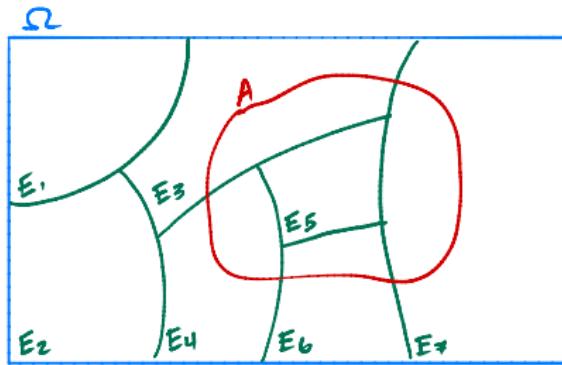
of sequences of 100 flips with exactly 40 heads

=

of ways of choosing here the 40 heads appear in 100 flips

=

$$P(E) = |E|/|\Omega| =$$



DSC 40A

Lecture 14

Conditional Probability

Getting to Campus

- ▶ 100 people were surveyed.
- ▶ How did you get to campus today? Walk, bike, or drive?
- ▶ Were you late or on-time?
- ▶ Results:

(Walk, Late)	6%
(Walk, Not Late)	24%
(Bike, Late)	3%
(Bike, Not Late)	7%
(Drive, Late)	36%
(Drive, Not Late)	24%

Example

- ▶ What is the probability that a randomly-selected person was late?

(Walk, Late)	6%
(Walk, Not Late)	24%
(Bike, Late)	3%
(Bike, Not Late)	7%
(Drive, Late)	36%
(Drive, Not Late)	24%

Example

- ▶ What is the probability that a randomly-selected person drove?

(Walk, Late)	6%
(Walk, Not Late)	24%
(Bike, Late)	3%
(Bike, Not Late)	7%
(Drive, Late)	36%
(Drive, Not Late)	24%

Getting to Campus

- ▶ Suppose we no longer have this table:

(Walk, Late)	6%
(Walk, Not Late)	24%
(Bike, Late)	3%
(Bike, Not Late)	7%
(Drive, Late)	36%
(Drive, Not Late)	24%

- ▶ Instead, we are told only that:
 - ▶ 30% of people walk; 20% of them are late.
 - ▶ 10% of people bike; 30% of them are late.
 - ▶ 60% of people drive; 60% of them are late.
- ▶ Can we recover the table?

Conditional Probabilities

- ▶ Of those who walked, 20% were late.
- ▶ We say the **conditional probability** of being late **given** walking is 20%.
- ▶ Written: $P(\text{Late} \mid \text{Walk}) = 0.20$
- ▶ We saw:

$$P(\text{Walk} \cap \text{Late}) = P(\text{Walk}) \cdot P(\text{Late} \mid \text{Walk})$$

- ▶ So:

$$P(\text{Late} \mid \text{Walk}) = \frac{P(\text{Walk} \cap \text{Late})}{P(\text{Walk})}$$

Conditional Probability

- ▶ Let A and B be events, with $P(B) > 0$.
- ▶ The **conditional probability** of A given B , written $P(A | B)$, is defined by:

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

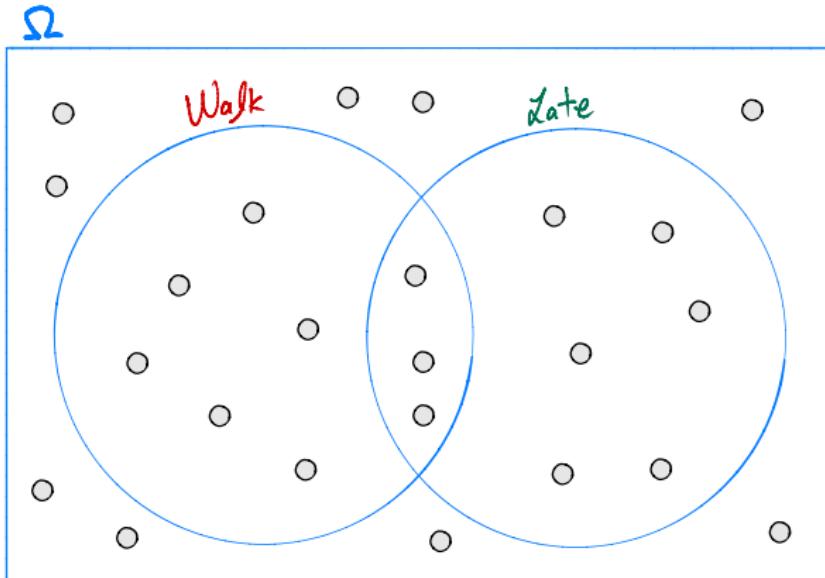
- ▶ Useful: $P(A \cap B) = P(A | B) \cdot P(B)$

Example

- ▶ Suppose someone **tells you** that they walked. What is the probability that they were late?
- ▶ That is, what is $P(\text{Late} \mid \text{Walk})$?

(Walk, Late)	6%
(Walk, Not Late)	24%
(Bike, Late)	3%
(Bike, Not Late)	7%
(Drive, Late)	36%
(Drive, Not Late)	24%

Venn Diagram: Late given walk



Conditional Probability

- ▶ Use the definition:

$$P(\text{Late} \mid \text{Walk}) = \frac{P(\text{Late} \cap \text{Walk})}{P(\text{Walk})}$$

(Walk, Late)	6%
(Walk, Not Late)	24%
(Bike, Late)	3%
(Bike, Not Late)	7%
(Drive, Late)	36%
(Drive, Not Late)	24%

Discussion Question

The probability of driving is 30%. The probability of being late, given that they drove, is 50%. What is the probability that a randomly-selected person drove **and** was late?

- A) 20%
- B) 30%
- C) 6%
- D) 15%

Tree Diagrams

- ▶ In what ways can a person arrive on campus?
 - ▶ $P(\text{Walk}) = 30\%;$ $P(\text{Late} \mid \text{Walk}) = 20\%.$
 - ▶ $P(\text{Bike}) = 10\%;$ $P(\text{Late} \mid \text{Bike}) = 30\%.$
 - ▶ $P(\text{Drive}) = 60\%;$ $P(\text{Late} \mid \text{Drive}) = 60\%.$

Law of Total Probability

- ▶ What is $P(\text{Late})$?

(Walk, Late)	6%
(Walk, Not Late)	24%
(Bike, Late)	3%
(Bike, Not Late)	7%
(Drive, Late)	36%
(Drive, Not Late)	24%

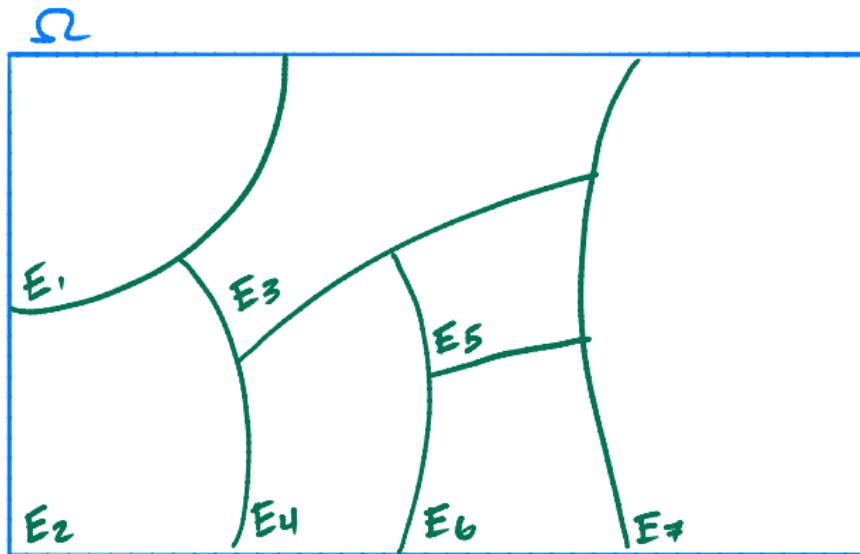
Law of Total Probability

- ▶ What is $P(\text{Late})$?
 - ▶ $P(\text{Walk}) = 30\%; P(\text{Late} \mid \text{Walk}) = 20\%.$
 - ▶ $P(\text{Bike}) = 10\%; P(\text{Late} \mid \text{Bike}) = 30\%.$
 - ▶ $P(\text{Drive}) = 60\%; P(\text{Late} \mid \text{Drive}) = 60\%.$
- ▶ Remember: $P(A \cap B) = P(A \mid B) \cdot P(B)$

Partitions

- ▶ Suppose events E_1, \dots, E_k are events such that, whatever the outcome, **exactly one** of the events is satisfied.
- ▶ That is:
 - ▶ No two events can happen simultaneously; they are mutually disjoint.
 - ▶ One of the events must happen. $P(E_1) + \dots + P(E_k) = 1$.
- ▶ We say that E_1, \dots, E_k **partition** the outcome space.

Partitions



Example: Partitions

- ▶ Examples of events which partition the outcome space:
 - ▶ In getting to campus, the events Walk, Bike, Drive.
 - ▶ In getting to campus, the events Late, On-Time.
 - ▶ In rolling a die, the events Even, Odd.
 - ▶ In rolling a die, the events ≤ 3 , > 3 .
 - ▶ In drawing a card, the events Spades, Clubs, Diamonds, Hearts.

Law of Total Probability

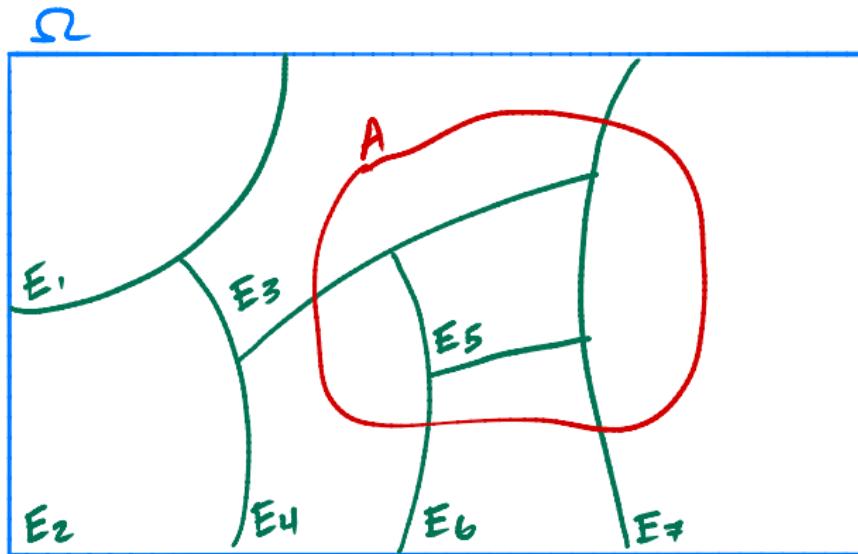
- ▶ Let A be an event, let E_1, \dots, E_k be events partitioning Ω .
- ▶ Then:

$$\begin{aligned} P(A) &= P(A \cap E_1) + P(A \cap E_2) + \dots + P(A \cap E_k) \\ &= \sum_{i=1}^k P(A \cap E_i) \end{aligned}$$

- ▶ And since $P(A \cap E) = P(A | E) \cdot P(E)$:

$$\begin{aligned} P(A) &= P(A | E_1) \cdot P(E_1) + \dots + P(A | E_k) \cdot P(E_k) \\ &= \sum_{i=1}^k P(A | E_i) \cdot P(E_i) \end{aligned}$$

Law of Total Probability



Bayes' Theorem

- ▶ Someone tells you that they were late. What is the probability that they drove to campus?
- ▶ We know: $P(\text{Late}) = 45\%$; $P(\text{Late} \mid \text{Drive}) = 60\%$.
- ▶ We want: $P(\text{Drive} \mid \text{Late})$.
- ▶ Using the definition:

Bayes' Theorem

- ▶ Let A and B be events (with $P(A) > 0$ and $P(B) > 0$).
- ▶ Then:

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)}$$

Example

- ▶ A certain disease occurs in only 1% of the population.
- ▶ A test for the disease is 95% accurate.
- ▶ You've tested positive for the disease; what is the probability that you actually have it?

Bayes' Theorem: Alternate Form

- ▶ Let A be an event.
- ▶ Let E_1, \dots, E_k be events partitioning Ω .
- ▶ Then, using the law of total probability:

$$\begin{aligned} P(E_1 | A) &= \frac{P(A | E_1) \cdot P(E_1)}{P(A)} \\ &= \frac{P(A | E_1) \cdot P(E_1)}{P(A \cap E_1) + \dots + P(A \cap E_k)} \\ &= \frac{P(A | E_1) \cdot P(E_1)}{P(A | E_1) \cdot P(E_1) + \dots + P(A | E_k) \cdot P(E_k)} \end{aligned}$$

Example

In a collection of 65 coins, one has two heads (the rest are fair). You select a coin at random and flip it six times, seeing Heads each time. What is the probability that the coin you selected is Unfair?

Example

A deck of five cards is numbered: 2, 4, 6, 8, 10. Three cards are drawn, one at a time with replacement; the sum of their values is 12. What is the probability that 2 was drawn twice?

♥: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A

♦: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A

♣: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, A

♠: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A

DSC 40A

*Lecture 15
Independence*

Last Time

- ▶ $P(A | B)$ = probability of A given that we know B has occurred.
- ▶ **Bayes' Theorem:**

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

The Bayesian View

- ▶ Bayesian view: probabilities quantify level of **belief**.
 - ▶ $P(A) = 1$; absolutely certain it will happen
 - ▶ $P(A) = 0$; absolute certain it will not
 - ▶ $P(A) = .75$; about 75% sure
- ▶ Bayes' Theorem allows us to “update” our beliefs when given new information.

Example

- ▶ In San Diego: 3,752 burglaries per year.
- ▶ Roughly 10 burglaries per night.
- ▶ 1.5 million people in San Diego.
- ▶ On any given night:

$$P(\text{Burglary}) = \frac{10}{1.5 \text{ million}} \approx 6 \times 10^{-7}$$

Example

- ▶ You hear your burglar alarm going off.
- ▶ How worried should you be?
- ▶ Assume:
 - ▶ If there is a burglary, there is a 95% of alarm.
 - ▶ If there isn't a burglary, there is a 1% chance of alarm.

Example

$$P(\text{Burglary}) = 6 \times 10^{-7} \quad P(\text{Alarm} \mid \text{Burglary}) = 0.95 \quad P(\text{Alarm} \mid \text{No Burglary}) = 0.01$$

$$P(\text{Burglary} \mid \text{Alarm}) =$$

Prior and Posterior Probabilities

- ▶ **Before** hearing the alarm, the probability of a burglary is

$$P(\text{Burglary}) = 6 \times 10^{-7}$$

- ▶ We call this the **prior** probability.
- ▶ **After** hearing the alarm, the probability increases:

$$P(\text{Burglary} \mid \text{Alarm}) = 5.6 \times 10^{-5}$$

- ▶ We call this the **posterior** probability.

Discussion Question

Now suppose $P(\text{Alarm} \mid \text{No Burglary}) = 10^{-5}$ instead of 0.01. What happens to the **posterior probability**, $P(\text{Burglary} \mid \text{Alarm})$?

- A) It goes up.
- B) It goes down.
- C) Nothing; it stays the same.

Example

- ▶ Suppose $P(\text{Alarm} \mid \text{No Burglary}) = 10^{-5}$.
- ▶ Then $P(\text{Burglary} \mid \text{Alarm}) = 0.054 \approx 5\%$

“Updating” Probabilities

- ▶ $P(A)$ is our prior belief that A happens.
- ▶ $P(A | B)$ is our updated belief that A happens, now that we know B happens.
- ▶ Sometimes knowing that B happens doesn't change anything.

Example

- ▶ We flip a fair coin twice.
- ▶ $P(\text{Second Flip} = \text{Heads}) =$
- ▶ $P(\text{Second Flip} = \text{Heads} \mid \text{First Flip} = \text{Heads}) =$

Independence

- ▶ We say that A and B are **independent** if knowing that B happens doesn't change our belief that A happens (and *vice versa*).
- ▶ Formally, A and B are independent if¹:

$$P(A | B) = P(A)$$

- ▶ Equivalently: $P(B | A) = P(B)$.
- ▶ Equivalently, A and B are independent if:

$$P(A \cap B) = P(A) \cdot P(B)$$

¹Assuming $P(B) > 0$

Example #1

Discussion Question

You throw two dice. A is the event that the first is a 6, B is the event that the sum is 10. Are these independent?

- A) Yes.
- B) No.

Example #2

Discussion Question

You draw two cards, one-at-a-time, **with** replacement. A is the event that the first card is a heart, B is the event that the second card is a club. Are these independent?

- A) Yes.
- B) No.

Example #3

Discussion Question

You draw two cards, one-at-a-time, **without** replacement. A is the event that the first card is a heart, B is the event that the second card is a club. Are these independent?

- A) Yes.
- B) No.

Example #4

Discussion Question

You draw one card from a deck of 52 cards. A is the event that the card is a heart, B is the event that the card is a face card (J,Q,K,A). Are these independent?

♥: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A

♦: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A

♣: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A

♠: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A

- A) Yes.
- B) No.

Assuming Independence

- ▶ Sometimes we **assume** that events are independent to make calculation easier.
- ▶ Real-world events are almost never exactly independent, but may be “close”.
- ▶ Example: A is event that a student is a data science major, B is event that they bike to campus.

Example

2% of UCSD students are data science majors. 20% of UCSD students bike to campus. Assuming that biking to campus and being a DSC major are independent:

- ▶ What percentage of data science majors bike to campus?

- ▶ What percentage of students are data science majors who bike to campus?

Conditional Independence

- ▶ Sometimes events A and B might not be independent.
- ▶ But they **become** independent upon learning some new information.

Example

Discussion Question

We've lost the King of Clubs! You draw one card from this deck of 51 cards. A is the event that the card is a heart, B is the event that the card is a face card (J,Q,K,A). Are these independent?

♥: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A

♦: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A

♣: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, A

♠: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A

- A) Yes.
- B) No.

Example

We've lost the King of Clubs! You draw one card from this deck of 51 cards. A is the event that the card is a heart, B is the event that the card is a face card (J, Q, K, A).

Now suppose you know that the card is red. Are A and B independent **given** this information?

♥: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A

♦: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A

♣: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, A

♠: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A

Conditional Independence

- ▶ Let A, B, C be events. A and B are **conditionally independent** given C if

$$P(A \cap B | C) = P(A | C) \cdot P(B | C).$$

Example

Discussion Question

A box contains two coins: one is fair, and the other is not – both sides are Heads. A coin is selected at random and flipped ten times. Let A be the event that the first nine flips are Heads, and let B be the event that the tenth flip is Heads. Are A and B **independent**?

- A) Yes.
- B) No.

Example

Discussion Question

A box contains two coins: one is fair, and the other is not – both sides are Heads. A coin is selected at random and flipped ten times. Let A be the event that the first nine flips are Heads, and let B be the event that the tenth flip is Heads. Let C be the event that the coin is fair. Are A and B **conditionally independent** given C ?

- A) Yes.
- B) No.

Relationship Between Independence and Conditional Independence

- ▶ There is none.

Assuming Conditional Independence

- ▶ Sometimes we **assume** that events are conditionally independent to make calculation easier.
- ▶ Real-world events are almost never exactly conditionally independent, but may be “close”.
- ▶ Example: A is the event that a person knows Python. B is the event that someone knows Bayes Theorem. C is the event that they are a data science major.

Example

Suppose 80% of data science majors know Python, and 70% know Bayes' Theorem. What is the probability that a randomly-selected major knows both, assuming that the events are conditionally independent given that they are a data science major?



Barack Obama

@BarackObama

Follow



I love data science!!!

10:05 PM - 28 Feb 2018

21,976 Retweets 15,772 Likes



212



21K



15K



DSC 40A
Lecture 16
Naïve Bayes, pt I

Last Time

- ▶ $P(A | B)$ = probability of A given that we know B has occurred.
- ▶ **Bayes' Theorem:**

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

- ▶ **Independence:** $P(A|B) = P(A)$, or $P(A \cap B) = P(A) P(B)$.
- ▶ **Conditional Independence:** $P(A \cap B | C) = P(A|C) P(B|C)$

Computing Probabilities

Two ways:

1. With math (combinatorics).
2. With data.

Example

You draw a card from a deck of 52 cards. What is the probability that it is red?

$$\frac{|E|}{|\Omega|} = \frac{26}{52} = \frac{1}{2}$$

Example

A person is chosen at random from DSC 40A. What is the probability that they can play piano?

$$\frac{|E|}{|\Omega|} = \frac{\text{\# in class who can play piano}}{\text{\# in class}}$$

Example

A person in the United States is chosen at random. What is the probability that they can play piano?

$$\frac{|E|}{|\Omega|} = \frac{\# \text{ in US who can play piano}}{\# \text{ in US}} = \frac{?}{\# \text{ in US}}$$

Estimating Probabilities

- ▶ We can estimate probabilities by randomly **sampling**.
- ▶ Example: ask 1000 people if they can play piano.
- ▶ If the sample is **representative**:

$$\frac{\# \text{ in US who can play piano}}{\# \text{ in US}} \approx \frac{\# \text{ in sample who can play piano}}{\# \text{ in sample}}$$

Example: Survey

Relationship Status	Favorite Subject	Favorite Car Brand	Favorite Cuisine
Married	English	Nissan	Mexican
Divorced	Physics	Daimler-Benz	Lebanese
Married	Chemistry	Daimler-Benz	Italian
Unmarried	History	Toyota	Chinese
Married	Biology	Ford	Thai
Unmarried	Chemistry	Audi	Lebanese
Unmarried	Biology	Honda	Chinese
Married	Commerce	Ford	Thai
Married	Maths	Daimler-Benz	Mediterranean
Married	Chemistry	Toyota	Thai

Example: Survey

- ▶ What is the probability that someone is married?

Relationship Status	Favorite Subject	Favorite Car Brand	Favorite Cuisine
Married	English	Nissan	Mexican
Divorced	Physics	Daimler-Benz	Lebanese
Married	Chemistry	Daimler-Benz	Italian
Unmarried	History	Toyota	Chinese
Married	Biology	Ford	Thai
Unmarried	Chemistry	Audi	Lebanese
Unmarried	Biology	Honda	Chinese
Married	Commerce	Ford	Thai
Married	Maths	Daimler-Benz	Mediterranean
Married	Chemistry	Toyota	Thai

Law of Large Numbers

- ▶ As the sample size grows, the estimate becomes more accurate.
- ▶ Example:
 - ▶ Flip coin $n = 10$ times, 5 repetitions. Proportions of heads:
0.4, 0.3, 0.8, 0.5, 0.5
 - ▶ Flip coin $n = 1000$ times, 5 repetitions. Proportions of heads:
0.52, 0.50, 0.51, 0.49, 0.48

Estimating Joint Probabilities

- ▶ To estimate $P(A \cap B)$, count number satisfying both A and B .
- ▶ Divide by size of sample.

Example

What is the probability that someone is married and their favorite car brand is Daimler-Benz?

Relationship Status	Favorite Subject	Favorite Car Brand	Favorite Cuisine
Married	English	Nissan	Mexican
Divorced	Physics	Daimler-Benz	Lebanese
Married	Chemistry	Daimler-Benz	Italian
Unmarried	History	Toyota	Chinese
Married	Biology	Ford	Thai
Unmarried	Chemistry	Audi	Lebanese
Unmarried	Biology	Honda	Chinese
Married	Commerce	Ford	Thai
Married	Maths	Daimler-Benz	Mediterranean
Married	Chemistry	Toyota	Thai

Estimating Conditional Probabilities

- ▶ To estimate $P(A | B)$, count number satisfying both A and B .
- ▶ Divide by number satisfying B .

Example

What is the probability that someone is married given that their favorite car brand is Daimler-Benz?

Relationship Status	Favorite Subject	Favorite Car Brand	Favorite Cuisine
Married	English	Nissan	Mexican
Divorced	Physics	Daimler-Benz	Lebanese
Married	Chemistry	Daimler-Benz	Italian
Unmarried	History	Toyota	Chinese
Married	Biology	Ford	Thai
Unmarried	Chemistry	Audi	Lebanese
Unmarried	Biology	Honda	Chinese
Married	Commerce	Ford	Thai
Married	Maths	Daimler-Benz	Mediterranean
Married	Chemistry	Toyota	Thai

Estimating Conditional Probabilities

- ▶ To estimate $P(A | B \cap C)$, count number satisfying $A \cap B \cap C$
- ▶ Divide by number satisfying $B \cap C$.

Example

What is the probability that someone's favorite car brand is Toyota given that they are married and their favorite food is Thai?

Relationship Status	Favorite Subject	Favorite Car Brand	Favorite Cuisine
Married	English	Nissan	Mexican
Divorced	Physics	Daimler-Benz	Lebanese
Married	Chemistry	Daimler-Benz	Italian
Unmarried	History	Toyota	Chinese
Married	Biology	Ford	Thai
Unmarried	Chemistry	Audi	Lebanese
Unmarried	Biology	Honda	Chinese
Married	Commerce	Ford	Thai
Married	Maths	Daimler-Benz	Mediterranean
Married	Chemistry	Toyota	Thai

Example

What is the probability that someone's favorite car brand is Toyota given that they are married and their favorite food is Thai and their favorite subject is physics?

Relationship Status	Favorite Subject	Favorite Car Brand	Favorite Cuisine
Married	English	Nissan	Mexican
Divorced	Physics	Daimler-Benz	Lebanese
Married	Chemistry	Daimler-Benz	Italian
Unmarried	History	Toyota	Chinese
Married	Biology	Ford	Thai
Unmarried	Chemistry	Audi	Lebanese
Unmarried	Biology	Honda	Chinese
Married	Commerce	Ford	Thai
Married	Maths	Daimler-Benz	Mediterranean
Married	Chemistry	Toyota	Thai

Estimating Conditional Probabilities

- ▶ We might not have enough data to estimate conditional probabilities with very specific conditions.
- ▶ Does **assuming** conditional independence help?
- ▶ Example: $P(A | C_1 \cap C_2 \cap C_3)$.
- ▶ Assume C_1, C_2, C_3 conditionally independent given A.
Then:

Sentiment Analysis

- ▶ Goal: given a tweet, determine if it is **positive**, **negative**, or **neutral**.

Sentiment Analysis

- ▶ Goal: given a tweet, determine if it is **positive**, **negative**, or **neutral**.



Barack Obama 
@BarackObama

I love data science!!!

10:05 PM - 28 Feb 2018

21,976 Retweets 15,772 Likes

212 21K 15K

Informative Words

- ▶ Some words are very informative in sentiment analysis:
 - ▶ “love”, “fantastic”, “enjoy”, etc., are **positive**.
 - ▶ “hate”, “terrible”, “angry”, etc., are **negative**.
- ▶ But “love” doesn’t automatically make tweet positive:

Informative Words

- ▶ Some words are very informative in sentiment analysis:
 - ▶ “love”, “fantastic”, “enjoy”, etc., are **positive**.
 - ▶ “hate”, “terrible”, “angry”, etc., are **negative**.
- ▶ But “love” doesn’t automatically make tweet positive:

Donald J. Trump 
@realDonaldTrump

Follow

The Fake News Media loves saying “so little happened at my first summit with Kim Jong Un.” Wrong! After 40 years of doing nothing with North Korea but being taken to the cleaners, & with a major war ready to start, in a short 15 months, relationships built, hostages & remains....

5:21 AM - 24 Jan 2019

21,780 Retweets 105,627 Likes

16K 21K 105K

Sentiment Analysis and Probability

- ▶ How likely is it that a tweet containing “love” is **positive**?
- ▶ In other words, what is:

$P(\text{tweet is positive} \mid \text{it contains “love”})$

- ▶ From the definition:

$P(\text{positive} \mid \text{contains “love”})$

$$= \frac{\# \text{ tweets which are positive and contain “love”}}{\# \text{ tweets containing “love”}}$$

Estimating Probabilities

- ▶ Gathering all tweets ever tweeted is not feasible.
- ▶ Instead, we gather a sample and *approximate* these probabilities.

$P(\text{positive} \mid \text{contains "love"})$

$$= \frac{\# \text{ tweets which are positive and contain "love"}}{\# \text{ tweets containing "love"}}$$

$$\approx \frac{\# \text{ tweets in sample which are positive and contain "love"}}{\# \text{ tweets in sample containing "love"}}$$

- ▶ Law of Large Numbers says: bigger the sample, better the approximation.

Estimating Probabilities

- ▶ We sample n tweets at random, label each as **positive**, **negative**, or **neutral** by hand.
- ▶ Mark whether each tweet contains “love”.
- ▶ The result is a table:

Sentiment	Contains “Love”
positive	yes
positive	no
negative	no
negative	no
neutral	yes
neutral	no
positive	yes

Estimating Probabilities

Sentiment	Contains “Love”
positive	yes
positive	no
negative	no
negative	no
neutral	yes
neutral	no
positive	yes

$P(\text{positive} \mid \text{contains “love”})$

$$\approx \frac{\# \text{ tweets in sample which are positive and contain “love”}}{\# \text{ tweets in sample containing “love”}}$$

=

Using Bayes' Theorem

- We could've used Bayes Theorem, too:

$$P(\text{positive} \mid \text{contains "love"})$$

$$= \frac{P(\text{contains "love"} \mid \text{positive}) \cdot P(\text{positive})}{P(\text{contains "love"})}$$

$$P(\text{positive}) \approx$$

$$P(\text{contains "love"}) \approx$$

Classification

- ▶ Now we are given a tweet we have never seen before; it does not contain “love”.
- ▶ We wish to classify its sentiment **automatically**.
- ▶ We will compute:

$P(\text{positive} \mid \text{does not contain “love”})$

$P(\text{negative} \mid \text{does not contain “love”})$

$P(\text{neutral} \mid \text{does not contain “love”})$

- ▶ The classification is determined by which probability is highest.

Estimating Probabilities

Sentiment	Contains “Love”
positive	yes
positive	no
negative	no
negative	no
neutral	yes
neutral	no
positive	yes

$P(\text{positive} \mid \text{does not contain “love”})$

≈

Estimating Probabilities

Sentiment	Contains “Love”
positive	yes
positive	no
negative	no
negative	no
neutral	yes
neutral	no
positive	yes

$P(\text{negative} \mid \text{does not contain “love”})$

≈

Estimating Probabilities

Sentiment	Contains “Love”
positive	yes
positive	no
negative	no
negative	no
neutral	yes
neutral	no
positive	yes

$P(\text{neutral} \mid \text{does not contain “love”})$

≈

Choosing the Most Likely Sentiment

$P(\text{positive} \mid \text{does not contain "love"}) \approx$

$P(\text{negative} \mid \text{does not contain "love"}) \approx$

$P(\text{neutral} \mid \text{does not contain "love"}) \approx$

- ▶ Since $P(\text{neutral} \mid \text{does not contain "love"})$ is largest, we assign tweet the sentiment of **neutral**.

Better Classifications

- ▶ We are making classification based on the presence/absence of one word.
- ▶ We'll get better results if we use more words:
 - ▶ w_1 = "love";
 - ▶ w_2 = "terrible";
 - ▶ w_3 = "angry";
- ▶ Suppose a tweet contains w_1 , doesn't contain w_2 , contains w_3 . We want to compute:
 - $P(\text{positive} \mid w_1 = \text{yes} \ \& \ w_2 = \text{no} \ \& \ w_3 = \text{yes})$
 - $P(\text{negative} \mid w_1 = \text{yes} \ \& \ w_2 = \text{no} \ \& \ w_3 = \text{yes})$
 - $P(\text{neutral} \mid w_1 = \text{yes} \ \& \ w_2 = \text{no} \ \& \ w_3 = \text{yes})$

A Practical Problem

- ▶ Suppose we use a lot of words, w_1, \dots, w_k .
- ▶ We approximate
 $P(\text{positive} \mid w_1 = \text{yes} \ \& \ w_2 = \text{yes} \ \& \ \dots \ \& \ w_k = \text{no})$
- ▶ There may not be enough data to satisfy such a specific condition.
- ▶ Next time: how do we use conditional independence assumption to help?



Barack Obama

@BarackObama

Follow



I love data science!!!

10:05 PM - 28 Feb 2018

21,976 Retweets 15,772 Likes

212

21K

15K



DSC 40A

Lecture 17

Naïve Bayes, pt II

Sentiment Analysis

- ▶ Goal: given a tweet, determine if it is **positive**, **negative**, or **neutral**.



Barack Obama 
@BarackObama

I love data science!!!

10:05 PM - 28 Feb 2018

21,976 Retweets 15,772 Likes

212 21K 15K

Classification

We have collected data:

Sentiment	“Love”	“Hate”	“Good”
positive	yes	no	yes
positive	no	no	yes
positive	yes	yes	no
negative	yes	yes	no
negative	no	no	yes
neutral	no	no	no
neutral	no	no	no

We want to classify a new tweet as **positive**, **negative**, or **neutral**:

The image shows a simulated Twitter post card. At the top left is a placeholder profile picture of a person. To its right, the word "Person" is displayed above the handle "@RealPerson". To the far right is a blue "Follow" button with a white arrow icon, and a small downward arrow indicating more options. Below this header, the tweet itself is shown in a light gray box: "I hate how good this new movie is!". Underneath the tweet, the timestamp "12:00 PM - 1 Oct 2018" is visible. At the bottom of the card are four small, semi-transparent social media icons: a speech bubble, a retweet symbol, a heart, and an envelope.

Conditional Probabilities for Classification



- ▶ We will calculate:

$P(\text{positive} | \text{love=no} \& \text{hate=yes} \& \text{good=yes})$

$P(\text{negative} | \text{love=no} \& \text{hate=yes} \& \text{good=yes})$

$P(\text{neutral} | \text{love=no} \& \text{hate=yes} \& \text{good=yes})$

- ▶ We will classify the tweet as whichever sentiment whose conditional probability is largest.

Approximating Probabilities with a Sample

$P(\text{positive} \mid \text{love=no} \& \text{hate=yes} \& \text{good=yes})$

$$= \frac{\# \text{ of positive tweets with hate and good, but not love}}{\# \text{ of tweets with hate and good, but not love}}$$

$$\approx \frac{\# \text{ of positive tweets in sample with hate and good, but not love}}{\# \text{ of tweets in sample with hate and good, but not love}}$$

The Curse of Dimensionality

$P(\text{positive} \mid \text{love}=\text{no} \ \& \ \text{hate}=\text{yes} \ \& \ \text{good}=\text{yes})$

$$\approx \frac{\# \text{ of } \text{positive} \text{ tweets in sample with hate and good, but not love}}{\# \text{ of tweets in sample with hate and good, but not love}}$$

=

Sentiment	“Love”	“Hate”	“Good”
positive	yes	no	yes
positive	no	no	yes
positive	yes	yes	no
negative	yes	yes	no
negative	no	no	yes
neutral	no	no	no
neutral	no	no	no

Naïve Bayes

To approximate $P(\text{positive} | \text{love}=no \ \& \ \text{hate}=yes \ \& \ \text{good}=yes)$:

- ▶ Start with Bayes' Theorem:

$$P(\text{positive} | \text{love}=no \ \& \ \text{hate}=yes \ \& \ \text{good}=yes)$$

$$= \frac{P(\text{love}=no \ \& \ \text{hate}=yes \ \& \ \text{good}=yes \mid \text{positive}) \cdot P(\text{positive})}{P(\text{love}=no \ \& \ \text{hate}=yes \ \& \ \text{good}=yes)}$$

- ▶ Then **assume** conditional independence of features given class:

$$P(\text{love}=no \ \& \ \text{hate}=yes \ \& \ \text{good}=yes \mid \text{positive})$$

$$= P(\text{love}=no \mid \text{positive}) \cdot P(\text{hate}=yes \mid \text{positive}) \cdot P(\text{good}=yes \mid \text{positive})$$

Naïve Bayes

$$P(\text{positive} \mid \text{love=no} \ \& \ \text{hate=yes} \ \& \ \text{good=yes})$$

$$= \frac{P(\text{love=no} \ \& \ \text{hate=yes} \ \& \ \text{good=yes} \mid \text{positive}) \cdot P(\text{positive})}{P(\text{love=no} \ \& \ \text{hate=yes} \ \& \ \text{good=yes})}$$

$$= \frac{P(\text{love=no} \mid \text{positive}) \cdot P(\text{hate=yes} \mid \text{positive}) \cdot P(\text{good=yes} \mid \text{positive}) \cdot P(\text{positive})}{P(\text{love=no} \ \& \ \text{hate=yes} \ \& \ \text{good=yes})}$$

- ▶ We are able to estimate $P(\text{love=no} \mid \text{positive})$, $P(\text{hate=yes} \mid \text{positive})$, and $P(\text{good=yes} \mid \text{positive})$ from the data.

Naïve Bayes

We want to find the biggest out of:

$$P(\text{positive} \mid \text{love=no} \& \text{hate=yes} \& \text{good=yes})$$

$$= \frac{P(\text{love=no} \mid \text{positive}) \cdot P(\text{hate=yes} \mid \text{positive}) \cdot P(\text{good=yes} \mid \text{positive}) \cdot P(\text{positive})}{P(\text{love=no} \& \text{hate=yes} \& \text{good=yes})}$$

$$P(\text{negative} \mid \text{love=no} \& \text{hate=yes} \& \text{good=yes})$$

$$= \frac{P(\text{love=no} \mid \text{negative}) \cdot P(\text{hate=yes} \mid \text{negative}) \cdot P(\text{good=yes} \mid \text{negative}) \cdot P(\text{negative})}{P(\text{love=no} \& \text{hate=yes} \& \text{good=yes})}$$

$$P(\text{neutral} \mid \text{love=no} \& \text{hate=yes} \& \text{good=yes})$$

$$= \frac{P(\text{love=no} \mid \text{neutral}) \cdot P(\text{hate=yes} \mid \text{neutral}) \cdot P(\text{good=yes} \mid \text{neutral}) \cdot P(\text{neutral})}{P(\text{love=no} \& \text{hate=yes} \& \text{good=yes})}$$

Naïve Bayes

Since they all have the same denominator, we can just pick that with the largest numerator:

$$\begin{aligned} & P(\text{love=no} \mid \text{positive}) \cdot P(\text{hate=yes} \mid \text{positive}) \cdot P(\text{good=yes} \mid \text{positive}) \cdot P(\text{positive}) \\ & P(\text{love=no} \mid \text{negative}) \cdot P(\text{hate=yes} \mid \text{negative}) \cdot P(\text{good=yes} \mid \text{negative}) \cdot P(\text{negative}) \\ & P(\text{love=no} \mid \text{neutral}) \cdot P(\text{hate=yes} \mid \text{neutral}) \cdot P(\text{good=yes} \mid \text{neutral}) \cdot P(\text{neutral}) \end{aligned}$$

This is **Naïve Bayes** classification.

Running Naïve Bayes

$$P(\text{love}=\text{no} \mid \text{positive}) \cdot P(\text{hate}=\text{yes} \mid \text{positive}) \cdot P(\text{good}=\text{yes} \mid \text{positive}) \cdot P(\text{positive})$$

Sentiment	“Love”	“Hate”	“Good”
positive	yes	no	yes
positive	no	no	yes
positive	yes	yes	no
negative	yes	yes	no
negative	no	no	yes
neutral	no	no	no
neutral	no	no	no

Running Naïve Bayes

$P(\text{love}=\text{no} \mid \text{negative}) \cdot P(\text{hate}=\text{yes} \mid \text{negative}) \cdot P(\text{good}=\text{yes} \mid \text{negative}) \cdot P(\text{negative})$

Sentiment	“Love”	“Hate”	“Good”
positive	yes	no	yes
positive	no	no	yes
positive	yes	yes	no
negative	yes	yes	no
negative	no	no	yes
neutral	no	no	no
neutral	no	no	no

Running Naïve Bayes

$$P(\text{love}=\text{no} \mid \text{neutral}) \cdot P(\text{hate}=\text{yes} \mid \text{neutral}) \cdot P(\text{good}=\text{yes} \mid \text{neutral}) \cdot P(\text{neutral})$$

Sentiment	“Love”	“Hate”	“Good”
positive	yes	no	yes
positive	no	no	yes
positive	yes	yes	no
negative	yes	yes	no
negative	no	no	yes
neutral	no	no	no
neutral	no	no	no

The Classification

We have:

$$P(\text{love=no} \mid \text{positive}) \cdot P(\text{hate=yes} \mid \text{positive}) \cdot P(\text{good=yes} \mid \text{positive}) \cdot P(\text{positive}) \\ \approx$$

$$P(\text{love=no} \mid \text{negative}) \cdot P(\text{hate=yes} \mid \text{negative}) \cdot P(\text{good=yes} \mid \text{negative}) \cdot P(\text{negative}) \\ \approx$$

$$P(\text{love=no} \mid \text{neutral}) \cdot P(\text{hate=yes} \mid \text{neutral}) \cdot P(\text{good=yes} \mid \text{neutral}) \cdot P(\text{neutral}) \\ \approx$$

So we classify the tweet as: **positive, negative, neutral.**

Example

Suppose that today's humidity is $> 50\%$, the temperature is hot, and the air pressure is low. Use Naïve Bayes to predict whether tomorrow will be rainy, cloudy, or sunny.

Next Day's Weather	Humidity	Temperature	Air Pressure
Rainy	$> 50\%$	Cool	Low
Rainy	$> 50\%$	Hot	Low
Rainy	$> 50\%$	Cool	Low
Rainy	25%-50%	Hot	High
Rainy	25%-50%	Hot	Low
Rainy	25%-50%	Cool	Low
Rainy	25%-50%	Cool	Low
Rainy	$< 25\%$	Cool	Low
Rainy	$< 25\%$	Hot	Low
Rainy	$< 25\%$	Hot	High
Cloudy	$> 50\%$	Cool	Low
Cloudy	$> 50\%$	Cool	Low
Cloudy	25%-50%	Hot	High
Cloudy	$< 25\%$	Cool	High
Cloudy	$< 25\%$	Cool	Low
Sunny	$> 50\%$	Cool	Low
Sunny	$> 50\%$	Hot	High
Sunny	$> 50\%$	Cool	High
Sunny	25%-50%	Hot	High
Sunny	$< 25\%$	Hot	High

