

# Lecture 3 – Mean Squared Error and Empirical Risk Minimization



DSC 40A, Spring 2023

# Announcements

- ▶ Homework 1 is due **Tuesday at 11:59pm.**
  - ▶ Come to office hours on Monday or Tuesday for help!
  - ▶ See [dsc40a.com/calendar](http://dsc40a.com/calendar) for the Office Hours schedule.
- ▶ Solutions to Groupwork 1 are available on Campuswire.

# Agenda

- ▶ Recap from Lecture 2 – minimizing mean absolute error and formulating mean squared error.
- ▶ Minimizing mean squared error.
- ▶ Comparing different minimizers.
- ▶ Empirical risk minimization.

## Recap from Lecture 2

# The median minimizes mean absolute error

- ▶ Our problem was: find  $h^*$  which minimizes the mean

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

- ▶ **Regardless of if  $n$  is odd or even**, the answer is  $h^* = \text{Median}(y_1, \dots, y_n)$ . The **best prediction**, in terms of mean absolute error, is the **median**.
  - ▶ When  $n$  is odd, this answer is unique.
  - ▶ When  $n$  is even, any number between the middle two data points also minimizes mean absolute error.
  - ▶ We define the median of an even number of data points to be the mean of the middle two data points.

## The mean absolute error is **not differentiable**

- ▶ We can't compute  $\frac{d}{dh} |y_i - h|$ .
- ▶ 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 mean absolute error is **not differentiable**

- ▶ We can't compute  $\frac{d}{dh} |y_i - h|$ .
- ▶ 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)  $e^{|y_i - h|}$

c)  $|y_i - h|^3$

b)  $|y_i - h|^2 = (y_i - h)^2$

d)  $\cos(y_i - h)$

## The squared error

- ▶ Let  $h$  be a prediction and  $y$  be the true value (i.e. the “right answer”). The **squared error** is:

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

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

$$\begin{aligned} \frac{d}{dh}(y - h)^2 &= 2(y - h) \cdot (-1) \\ &= -2(y - h) \\ &= 2(h - y) \end{aligned}$$



## The new idea

- ▶ Find  $h^*$  by minimizing the **mean squared error**:

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

→ always some notion of error

- ▶ Strategy: Take the derivative, set it equal to zero, and solve for the minimizer.

**Minimizing mean squared error**

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

### Discussion Question

Which of these is  $dR_{\text{sq}}/dh$ ?

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

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

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

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

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

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

Set to zero and solve for minimizer  $= h$

$$R_{sq}^2(h) = \frac{2}{n} \sum_{i=1}^n (h - y_i) = 0$$

$$\rightarrow \sum_{i=1}^n (h - y_i) = 0$$

$$(h - y_1) + (h - y_2) + \dots + (h - y_n) \left( \sum_{i=1}^n h \right) - \left( \sum_{i=1}^n y_i \right) = 0$$

$$n \cdot h - \sum_{i=1}^n y_i = 0$$

the mean!  $\rightarrow h = \frac{1}{n} \sum_{i=1}^n y_i$

# The mean minimizes mean squared error

- ▶ Our new problem was: find  $h^*$  which minimizes the mean squared error,  $R_{sq}(h) = \frac{1}{n} \sum_{i=1}^n (y_i - h)^2$ .
  - ▶ The answer is:  $\text{Mean}(y_1, \dots, y_n)$ .
  - ▶ The **best prediction**, in terms of mean squared error, is the **mean**.
  - ▶ This answer is always unique!

## Discussion Question

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

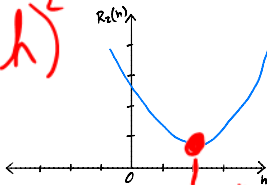
$$\frac{1}{n} \sum_{i=1}^n (y_i - h)^2$$

ex.)

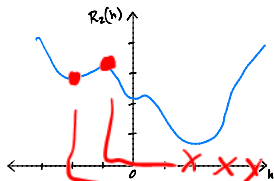
$$y_1 = 5, y_2 = 7, y_3 = 10$$

$$\frac{1}{3} \left( (5-h)^2 + (7-h)^2 + (10-h)^2 \right)$$

still a quadratic

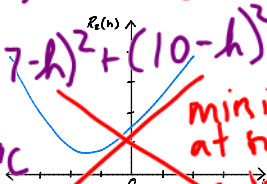


(a) mean of salaries



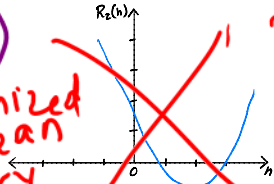
(b)

not quadratic



(c)

minimized at mean salary > 0



(d)

$R_{sq}(h) \geq 0$  for any h

## Comparing the median and mean



# Outliers

- ▶ Consider our original dataset of 5 salaries.

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

med  $\triangle$

- ▶ As it stands, the **median is 96,000** and the **mean is 112,000**.

- ▶ What if we add 300,000 to the largest salary?

90,000 94,000 96,000 120,000 460,000

med  $\triangle$

- ▶ Now, the **median is still 96,000** but the **mean is 172,000!**

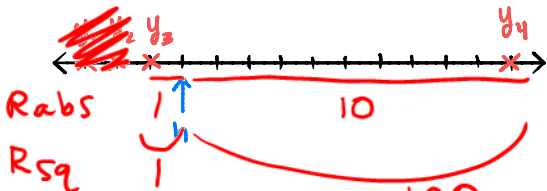
- ▶ **Key Idea:** The mean is quite **sensitive** to outliers.

median = robust

$$\begin{aligned} &= 112,000 \\ &+ \frac{300,000}{5} \end{aligned}$$

# Outliers

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



outlier

means  
abs  
error

$$\frac{1 + 10}{2}$$

- ▶  $|y_4 - h|$  is 10 times as big as  $|y_3 - h|$ .

- ▶ But  $(y_4 - h)^2$  is 100 times as big as  $(y_3 - h)^2$ .
  - ▶ This “pulls”  $h^*$  towards  $y_4$ .

mean sq  
error

$$\frac{1 + 100}{2}$$

- ▶ Squared error can be dominated by outliers.



mean  
sq error  
 $(25 + 36) / 2$

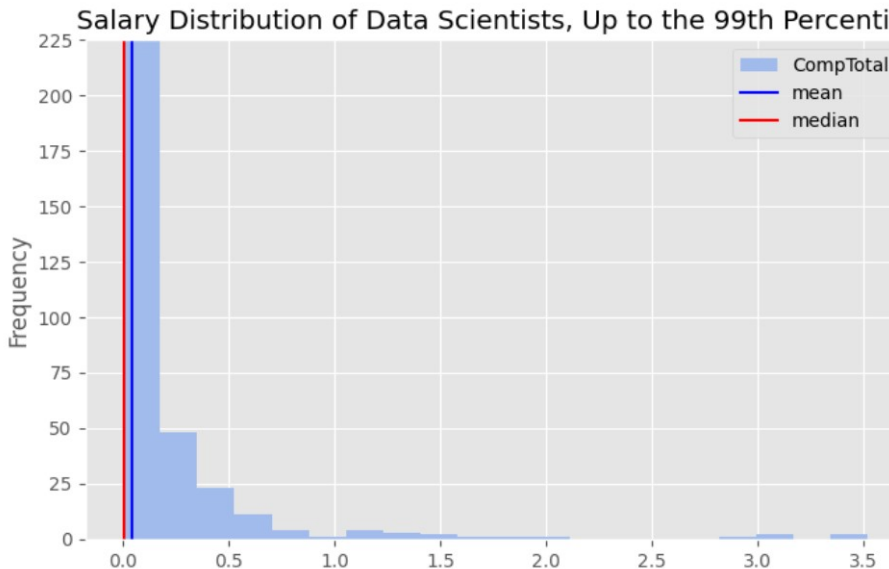
## Example: Data Scientist Salaries

- ▶ Dataset of 2,016 self-reported data science salaries in the United States from the 2022 StackOverflow survey.
- ▶ Median = \$86,700.
- ▶ Mean = \$501,425,531.
- ▶ Min = \$20.
- ▶ Max = \$1,000,000,000,000.  
*1 trillion*
- ▶ 90th Percentile: \$700,000.

*outlier*

*→ certain error in data*

## Example: Data Scientist 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).

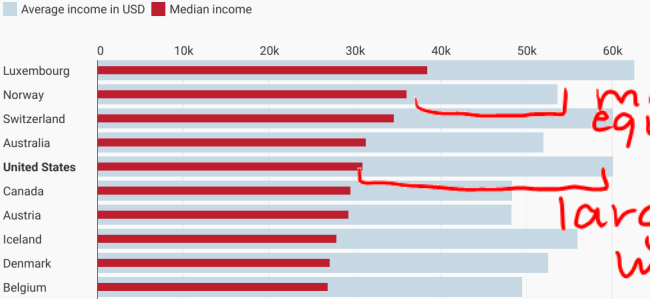
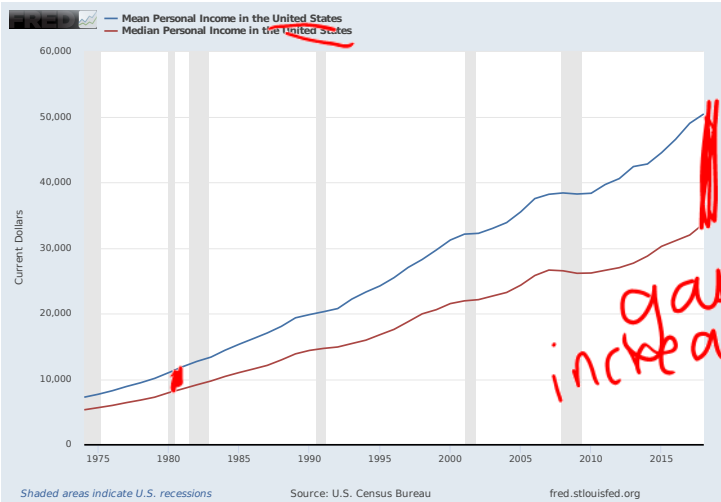


Chart: Lisa Charlotte Rost, Datawrapper

more equitable?  
large wealth gap

# Example: Income Inequality



gap increasing

over time

# Empirical risk minimization

## A general framework

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

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

- ▶ Then we introduced the **mean squared error**:

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

- ▶ They have the same form: both are averages of some measurement that represents how different  $h$  is from the data.

*your choice - should encapsulate closeness of  $h$  to  $y_i$*



## A general framework

- ▶ Definition: A loss function  $L(h, y)$  takes in a prediction  $h$  and a true value (i.e. 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 **squared loss**:

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

higher loss = higher error  
= worse prediction

## 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 on the data set:

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

← all data values

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

# The learning recipe

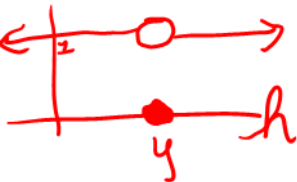
1. Pick a loss function.
  2. Pick a way to minimize the average loss (i.e. empirical risk) on the data.
- ▶ **Key Idea:** The choice of loss function determines the properties of the result. **Different loss function = different minimizer = different prediction!**
    - ▶ Absolute loss yields the median.
    - ▶ Squared loss yields the mean.
    - ▶ The mean is easier to calculate but is more sensitive to outliers.

## Example: 0-1 Loss

1. Pick as our loss function the **0-1 loss**:

$$L_{0,1}(h, y) = \begin{cases} 0, & \text{if } h = y \\ 1, & \text{if } h \neq y \end{cases}$$

↑  
data



2. Minimize empirical risk:

$$R_{0,1}(h) = \frac{1}{n} \sum_{i=1}^n L_{0,1}(h, y_i)$$

## Example: 0-1 Loss

1. Pick as our loss function the **0-1 loss**:

$$L_{0,1}(h, y) = \begin{cases} 0, & \text{if } h = y \\ 1, & \text{if } h \neq y \end{cases}$$

2. Minimize empirical risk:

$$R_{0,1}(h) = \frac{1}{n} \sum_{i=1}^n L_{0,1}(h, y_i)$$

$$h = y_1$$

$$\Rightarrow \frac{1}{n} \sum_{i=1}^n L_{0,1}(y_1, y_i) =$$



### Discussion Question

Suppose  $y_1, \dots, y_n$  are all distinct. Find  $R_{0,1}(y_1)$ .

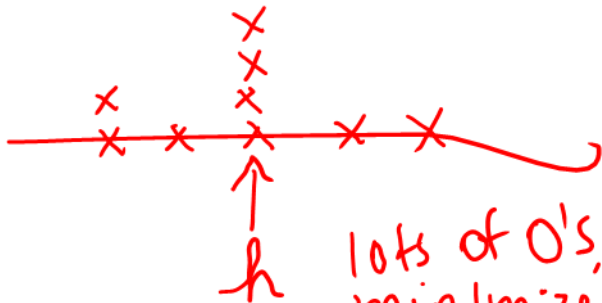
- a) 0   b)  $\frac{1}{n}$    c)  $\frac{n-1}{n}$    d) 1

$$\frac{1}{n} (0 + 1 + 1 + \dots + 1)$$

## Minimizing empirical risk

$$R_{0,1}(h) = \frac{1}{n} \sum_{i=1}^n \begin{cases} 0, & \text{if } h = y_i \\ 1, & \text{if } h \neq y_i \end{cases}$$

more generally



lots of 0's, few 1's  
minimize  $R_{0,1}$

## Different loss functions lead to different predictions

| Loss             | Minimizer | Outliers    | Differentiable |
|------------------|-----------|-------------|----------------|
| $L_{\text{abs}}$ | median    | insensitive | no             |
| $L_{\text{sq}}$  | mean      | sensitive   | yes            |
| $L_{0,1}$        | mode      | insensitive | no             |

- ▶ The optimal predictions are all **summary statistics** that measure the **center** of the data set in different ways.

## Summary



## Summary

- ▶  $h^* = \text{Mean}(y_1, \dots, y_n)$  minimizes  $R_{sq}(h) = \frac{1}{n} \sum_{i=1}^n (y_i - h)^2$ , i.e. the mean minimizes mean squared error.
- ▶ The mean absolute error and the mean squared error fit into a general framework called **empirical risk minimization**.
  - ▶ Pick a loss function. We've seen absolute loss,  $|y - h|$ , squared loss,  $(y - h)^2$ , and 0-1 loss.
  - ▶ Pick a way to minimize the average loss (i.e. empirical risk) on the data.
- ▶ By changing the loss function, we change which prediction is considered the best.

## Next time

- ▶ **Spread** – what is the meaning of the value of  $R_{abs}(h^*)$ ?  
 $R_{sq}(h^*)$ ?
- ▶ Creating a new loss function and trying to minimize the corresponding empirical risk.
  - ▶ We'll get stuck and have to look for a new way to minimize.