

**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.

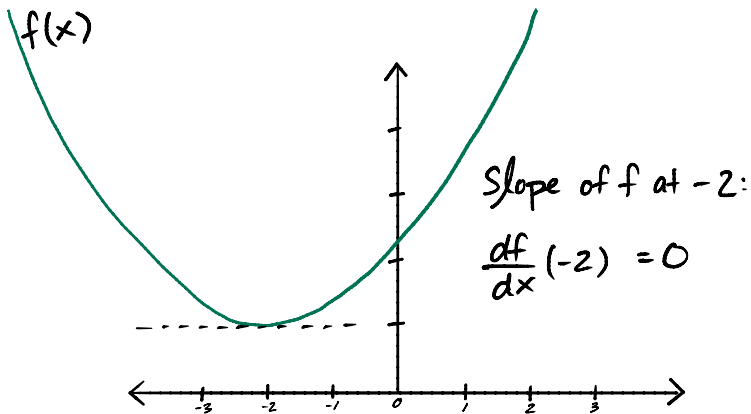
# Today

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

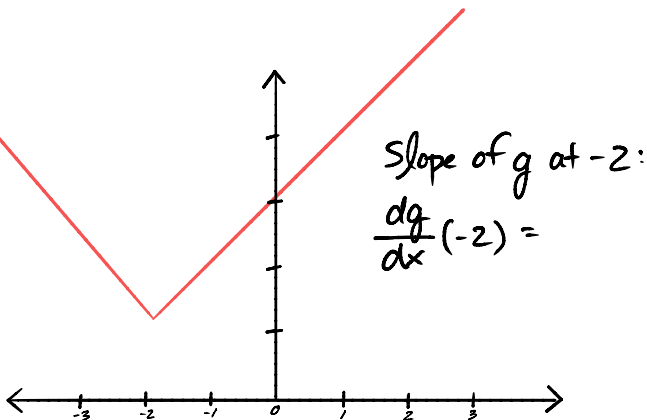
# Minimizing via calculus

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

## Example: differentiable

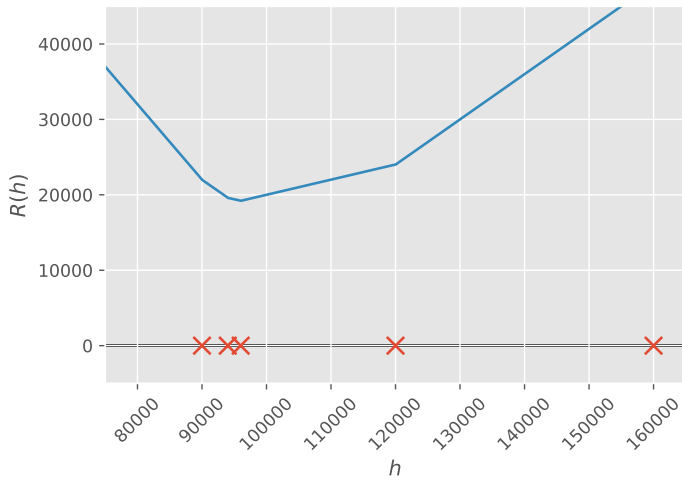


Example: **not differentiable**





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:  $1.0652 \times 10^{10}$

**mean squared error:**  $2.13 \times 10^9$

- 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:  $3.47 \times 10^9$

**mean squared error:**  $6.95 \times 10^8$

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

## The New Idea

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

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

- ▶ 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  $dR_{\text{sq}}/dh$ ?

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

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

**Set to zero and solve for minimizer**

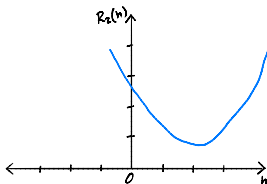
## The **mean** minimizes the **mean squared error**

- That is:

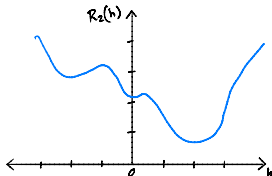
$$\arg \min_{h \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n (y_i - h)^2 = \text{Mean}(y_1, \dots, y_n)$$

## Discussion Question

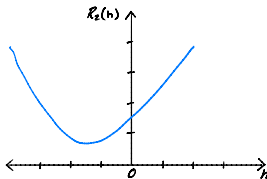
Suppose  $y_1, \dots, y_n$  are salaries. Which plot could be  $R_{sq}(h)$ ?



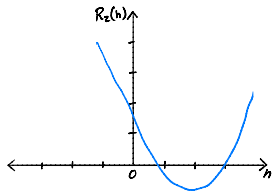
(a)



(b)



(c)



(d)

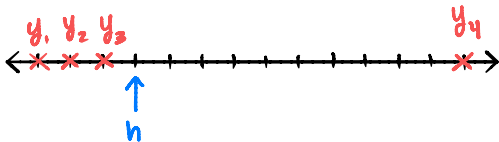
## The **mean** is easy to compute

```
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.



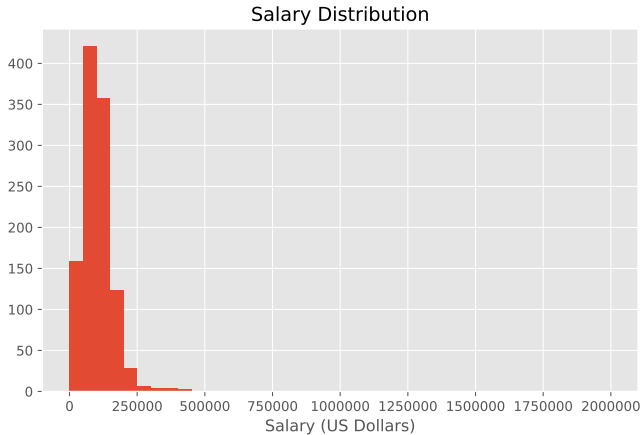
- ▶  $|y_4 - h|$  is 10 times as big as  $|y_3 - h|$ .
- ▶ But  $(y_4 - h)^2$  is 100 times as big as  $(y_3 - h)^2$ .
- ▶ Squared error can be dominated by outliers.

## Example: Data Science Salaries

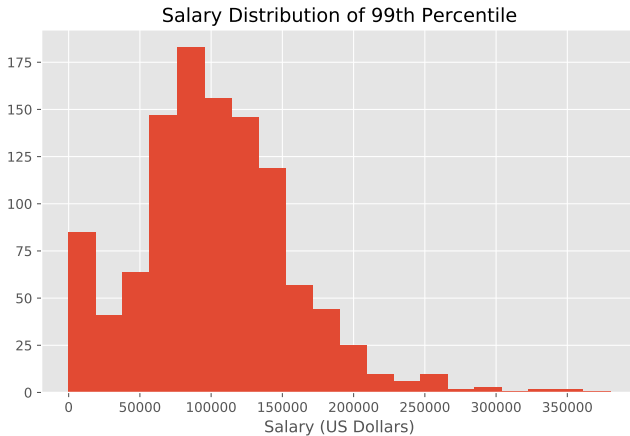
- ▶ Data set of 1121 self-reported data science salaries in the United States from the 2018 StackOverflow survey.
- ▶ Median = \$100,000
- ▶ Mean = \$111,032
- ▶ Max = \$2,000,000
- ▶ Min = \$52
- ▶ 95th Percentile: \$200,000



# 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).

■ Average income in USD ■ Median income

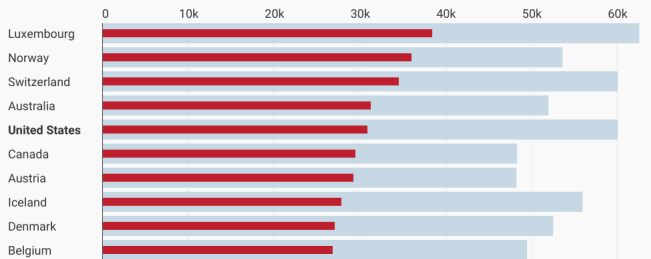
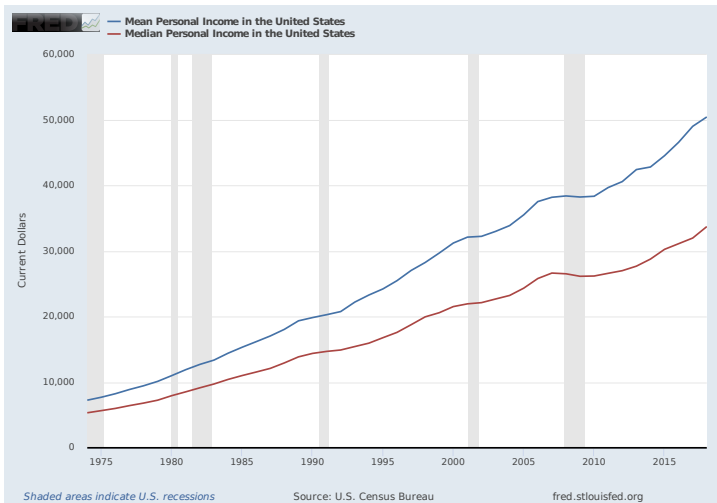


Chart: Lisa Charlotte Rost, Datawrapper

# 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
$L_{\text{abs}}$	median	<b>insensitive</b>	<b>no</b>	<b>not simple</b>
$L_{\text{sq}}$	mean	<b>sensitive</b>	<b>yes</b>	<b>simple, fast</b>

# 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**.