Lecture 2 – Minimizing Mean Absolute Error

DSC 40A, Fall 2021 @ UC San Diego

Suraj Rampure, with help from **many others**

Announcements

- Look at the readings linked on the course website!
- Homework 1 is out and is due on **Monday, 10/4 at 11:59pm**.
  - Survey 1 will come out on Thursday after lecture and will be due with the homework.
- Groupwork 1 is out and is due on **Thursday, ~~10/30~~ 9/30 at 11:59pm**.
- Come to discussion **tomorrow** to work on groupwork!
- See Calendar on course website for office hours locations and Zoom links.
  - In-person office hours are now in SDSC.

Agenda

1. Recap from Lecture 1 – learning from data.
2. Minimizing mean absolute error.
3. Identifying another choice of error.

Recap from Lecture 1 – learning from data

Last time

- **Question:** How do we turn the problem of learning from data into a math problem?
- **Answer:** Through optimization.

A formula for the mean absolute error

- We have data:

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

- Suppose our prediction is h.

- The **mean absolute error** of our prediction is:

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

|actual − prediction|

Many possible predictions

- Last time, we considered four possible **hypotheses** for future salary, and computed the mean absolute error of each.
  - $h_1 = 150{,}000 \implies R(150{,}000) = 42{,}000$
  - $h_2 = 115{,}000 \implies R(115{,}000) = 23{,}000$
  - $h_3 = \text{mean} = 112{,}000 \implies R(112{,}000) = 22{,}400$
  - $h_4 = \text{median} = 96{,}000 \implies R(96{,}000) = 19{,}200$
- Of these four options, the median has the lowest MAE. But is it the **best possible prediction overall**?

A *general* formula for the mean absolute error

- Suppose we collect n salaries, $y_1, y_2, \ldots, y_n$.
- The mean absolute error of the prediction h is:

$$R(h) = \frac{1}{n}\left[|y_1 - h| + |y_2 - h| + \ldots + |y_n - h|\right]$$

- Or, using **summation notation**:

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

The best prediction

- We want the best prediction, h^*.
- The smaller $R(h)$, the better h.
- Goal: find h that minimizes $R(h)$.

Discussion Question

Can we use calculus to minimize R?

Minimizing mean absolute error

Minimizing with calculus

- Calculus: take derivative with respect to *h*, set equal to zero, solve.

① $h(x) = f(x) + g(x) \Rightarrow h'(x) = f'(x) + g'(x)$

② $h(x) = c f(x) \Rightarrow h'(x) = c f'(x)$

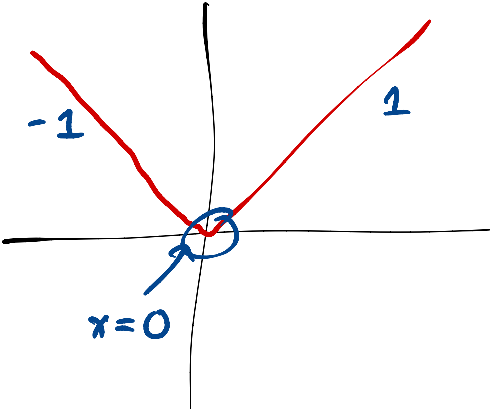
$$R(h) = \frac{1}{n} \sum_{i=1}^{n} |y_i - h|$$

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

Minimizing with calculus

- Calculus: take derivative with respect to h, set equal to zero, solve.

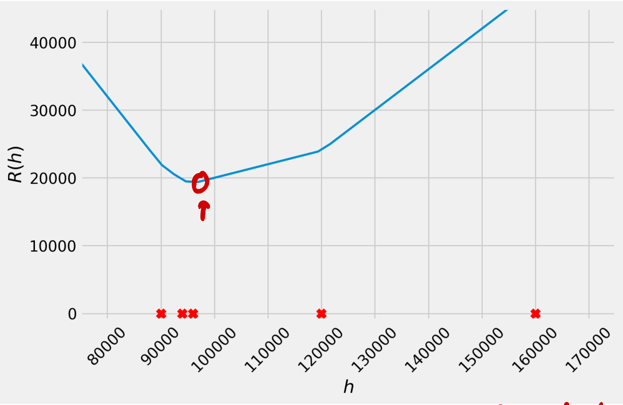
$$f(x) = |x|$$



Uh oh...

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

Plotting the mean absolute error



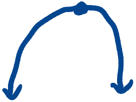
- slope changes where the original data points are
- a bunch of line segments

Discussion Question

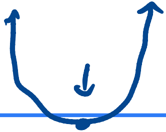
A local minimum occurs when the slope goes from ________________. Select all that apply.

A) ~~positive to negative~~
B) negative to positive
C) positive to zero.
D) negative to zero.

A

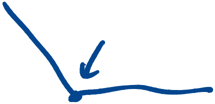


B

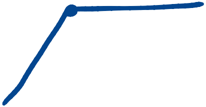


To answer, go to `menti.com` and enter the code 4144 9385.

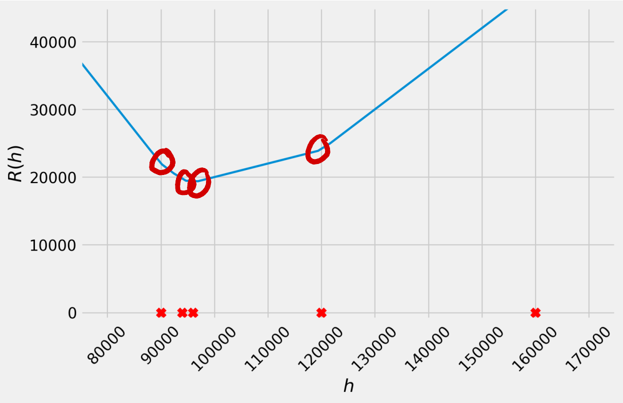
D



C



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

$$f_1(x) = 3x + 7 \qquad f_2(x) = 5x - 4 \qquad f_3(x) = -2x - 8$$

- What is the slope of $f(x) = f_1(x) + f_2(x) + f_3(x)$?

$$f(x) = (3x + 5x - 2x) + C$$

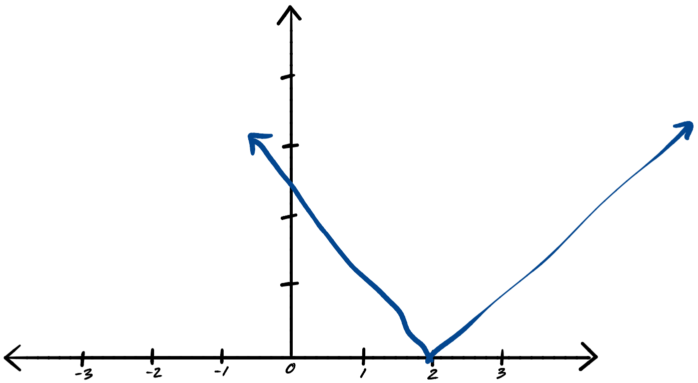
$$f(x) = 6x + C$$

$$\text{slope} = f'(x) = \boxed{6}$$

Absolute value functions

Recall, $f(x) = |x - a|$ is an absolute value function centered at $x = a$.

$|x-2|$



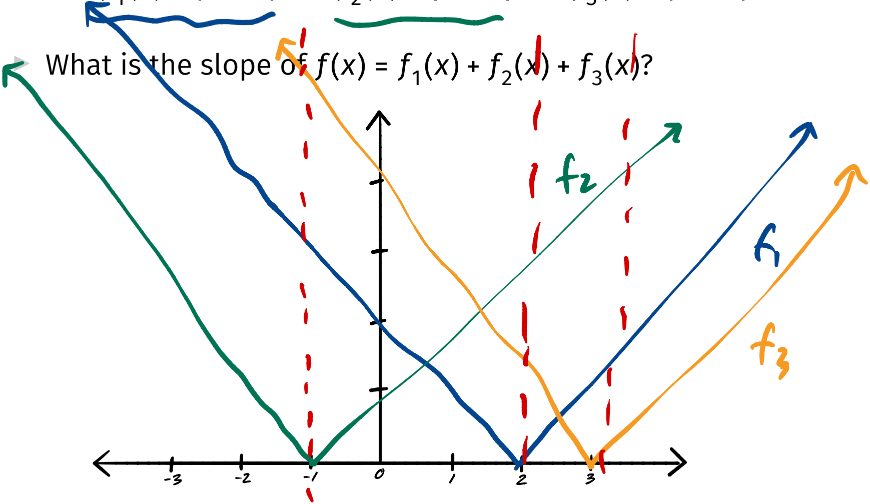
Sums of absolute values

region	slope
$x < -1$	-3
$-1 < x < 2$	-1
$2 < x < 3$	1
$x > 3$	3

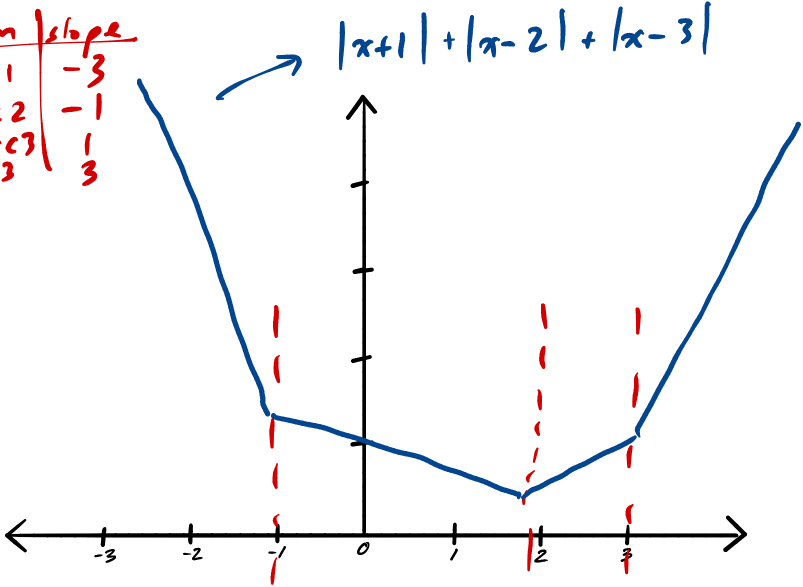
- Let

$$f_1(x) = |x - 2| \qquad f_2(x) = |x + 1| \qquad f_3(x) = |x - 3|$$

- What is the slope of $f(x) = f_1(x) + f_2(x) + f_3(x)$?



region	slope
$x < -1$	-3
$-1 < x < 2$	-1
$2 < x < 3$	1
$x > 3$	3



The slope of the mean absolute error

$R(h)$ is a sum of absolute value functions (times $\frac{1}{n}$):

Fact: $a > b$, or $a = b$, or $a < b$

$|-2| = -(-2)$

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

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

$$= \frac{1}{n}\left[\sum_{y_i < h} |h - y_i| + \sum_{y_i = h} |h - y_i| + \sum_{y_i > h} |h - y_i|\right]$$

$$= \frac{1}{n}\left[\sum_{y_i < h} (h - y_i) + \sum_{y_i > h} -(h - y_i)\right]$$

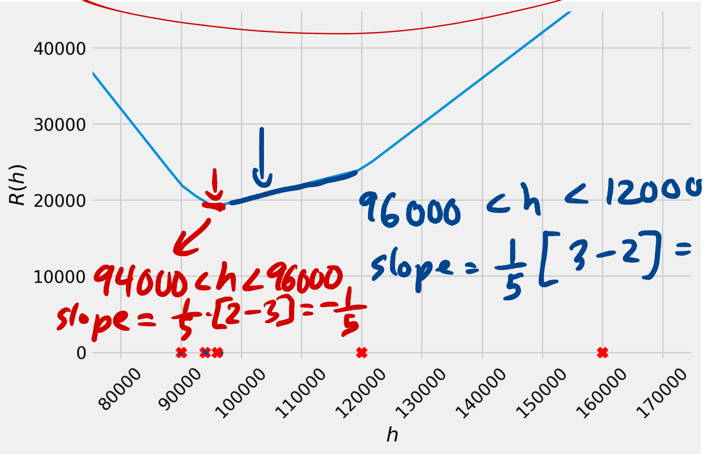
slope of R at $h = \frac{1}{n}\left[\sum_{y_i < h} 1 + \sum_{y_i > h} (-1)\right]$

The slope of the mean absolute error

The slope of R at h is:

$$\frac{1}{n} \cdot \left[(\#\text{ of } y_i\text{'s} < h) - (\#\text{ of } y_i\text{'s} > h) \right]$$

$\sum_{y_i < h} 1$ $\qquad$ $\sum_{y_i > h} 1$



$94000 < h < 96000$

$\text{slope} = \frac{1}{5} \cdot [2-3] = -\frac{1}{5}$

$96000 < h < 120000$

$\text{slope} = \frac{1}{5} [3-2] = \frac{1}{5}$

Where the slope's sign changes

not defined when $h = y_i$!!!

The slope of R at h is:

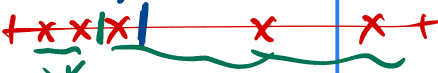
$$\frac{1}{n} \cdot \Big[(\text{\# of } y_i\text{'s} < h) - (\text{\# of } y_i\text{'s} > h)\Big]$$

Discussion Question

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

A) h = mean of $y_1, \ldots, y_n$
B) h = median of $y_1, \ldots, y_n$
C) h = mode of $y_1, \ldots, y_n$

$\frac{1}{5}(3-2)$



$\frac{1}{5}(2-3)$

To answer, go to `menti.com` **and enter the code 4144 9385.**

The median minimizes mean absolute error, when n is odd

- Our problem was: find h^* which minimizes the mean absolute error, $R(h) = \frac{1}{n}\sum_{i=1}^{n} |y_i - h|$.
- We just determined that when n is odd, the answer is $\text{Median}(y_1, \ldots, y_n)$. This is because the median has an equal number of points to the left of it and to the right of it.
- But wait — what if n is **even**?

Discussion Question

Consider again our example dataset of 5 salaries.

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

Suppose we collect a 6th salary, so that our data is now

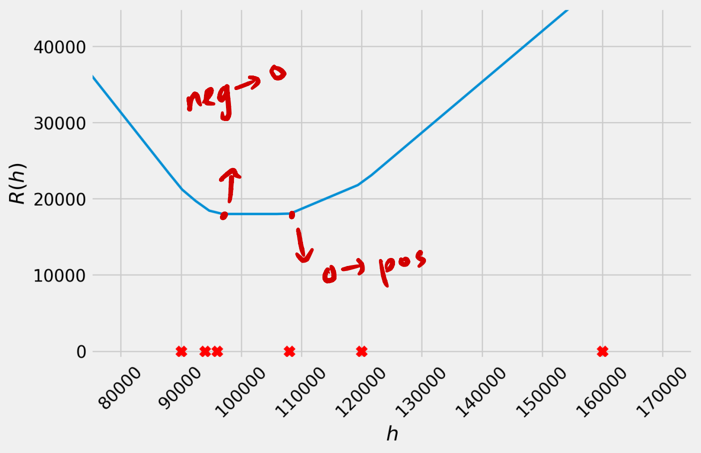
90,000 94,000 96,000 108,000 120,000 160,000

Which of the following correctly describes the h^* that minimizes mean absolute error for our new dataset?

A) 96,000 only
B) 108,000 only
C) 102,000 only
D) Any value between 96,000 and 108,000, inclusive

To answer, go to `menti.com` and enter the code 4144 9385.

Plotting the mean absolute error, with an even number of data points



- What do you notice?

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.

Identifying another type of error

Two things we don't like

1. **Minimizing** the mean absolute error wasn't so easy.

2. Actually **computing** the median isn't so easy, either.

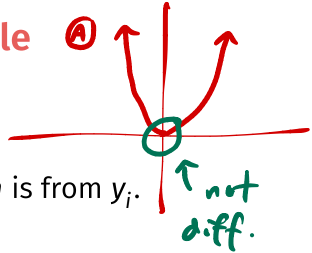
- **Question**: Is there another way to measure the quality of a prediction that avoids these problems?

The mean 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 error is not differentiable

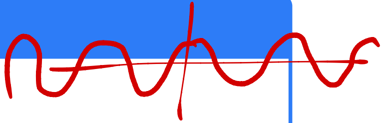
Ⓐ



- 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**?

Ⓒ not diff

Ⓓ



Discussion Question

Which of these would work?

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

b) $|y_i - h|^2$

Ⓑ $|y_i - h|^2 = (y_i - h)^2$

c) $|y_i - h|^3$

d) $\cos(y_i - h)$

The squared error

$x^2 = (-x)^2$

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

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

- Like absolute error, measures how far h is from y.

- But unlike absolute error, the squared error is **differentiable**:

$$\frac{d}{dh}(y - h)^2 = 2(y-h)(-1)$$
$$= -2(y-h)$$
$$= 2(h-y)$$

The mean squared error

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

salary	absolute 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	absolute 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**:

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

squared err for indiv. points

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

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

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

To answer, go to `menti.com` **and enter the code 4144 9385.**

Summary

Summary

- Our first problem was: find h^* which minimizes the mean absolute 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**, in terms of mean absolute error, is the **median**.
- We then started to consider another type of error, squared error, that is differentiable and hence is easier to minimize.
- **Next time:** We will finish determining the value of h^* that minimizes mean squared error, and see how it compares to the median.