Lecture 2 – Minimizing Mean Absolute Error

DSC 40A, Winter 2024

Announcements

- Look at the readings linked on the course website!
- Homework 1 is out today and due Wed., January 17 midnight.
- See Calendar on course website for office hours.
    - Plan to come to office hours at least once a week for help on homework.

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

What is good/bad, intuitively?

- The data:

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

- Consider these hypotheses:

  $h_1 = 150{,}000 \qquad h_2 = 115{,}000$

Quantifying our intuition

- Intuitively, a good prediction is close to the data.

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

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

Quantifying our intuition

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

salary	absolute error of h_2
90,000	25,000
94,000	21,000
96,000	19,000
120,000	5,000
160,000	45,000
	sum of absolute errors: 115,000
	mean absolute error: 23,000

Mean absolute error (MAE)

- Mean absolute error on data:

$$h_1 : 42{,}000 \qquad h_2 : 23{,}000$$

- Conclusion: h_2 is the better prediction.

- In general: pick prediction with the smaller mean absolute error.

We are making an assumption...

- We're assuming that future salaries will look like present salaries.
- That a prediction that was good in the past will be good in the future.

Discussion Question

Is this a good assumption?

Which is better: the mean or median?

- Recall:

  mean = 112,000   median = 96,000

- We can calculate the mean absolute error of each:

  mean : 22,400   median : 19,200

- The median is the best prediction so far!

- But is there an even better prediction?

Finding the best prediction

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

A formula for the mean absolute error

- We have data:

$$90{,}000 \quad 94{,}000 \quad 96{,}000 \quad 120{,}000 \quad 160{,}000$$

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

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

A formula for the mean absolute error

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

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

A formula for the mean absolute error

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

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

A formula for the mean absolute error

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

$$R(\pi) = \frac{1}{5}\Big(|90{,}000 - \pi| + |94{,}000 - \pi| + |96{,}000 - \pi| + |120{,}000 - \pi| + |160{,}000 - \pi|\Big) = \mathbf{111{,}996.8584...}$$

Discussion Question

Without doing any calculations, which is correct?

A. $R(50) < R(100)$
B. $R(50) = R(100)$
C. $R(50) > R(100)$

A *general* formula for the mean absolute error

- Suppose we collect n salaries, $y_1, y_2, \dots, y_n$.

- The mean absolute error of the prediction h is:

- 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 $R(h)$.

Summary

- We started with the learning problem:

  Given salary data, predict your future salary.

- We turned it into this problem:

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

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

- **Now**: we solve this math problem.

Minimizing mean absolute error

Many possible predictions

- So far, we have 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:

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

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 with calculus

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

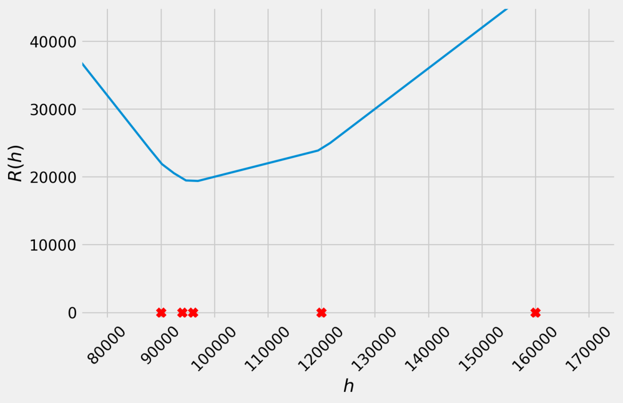
Minimizing with calculus

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

Uh oh...

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

Plotting the mean absolute error

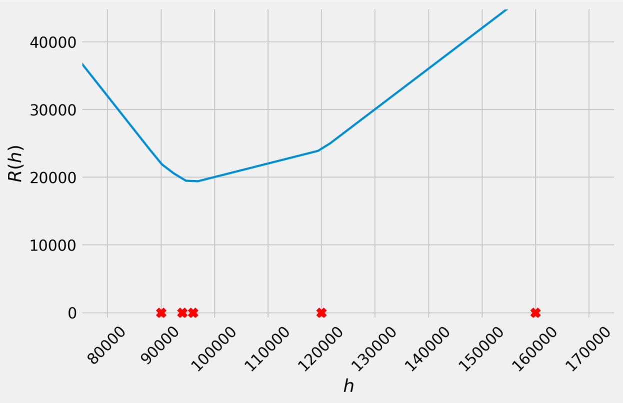


Discussion Question

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.

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

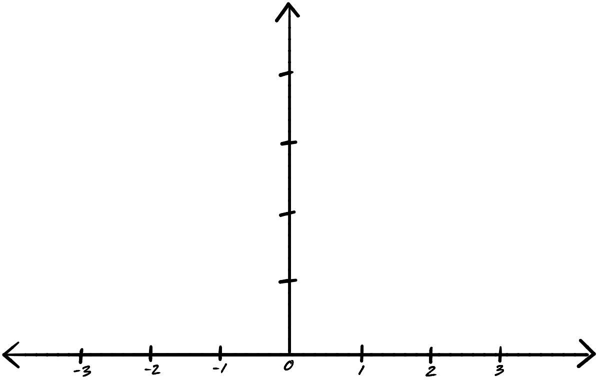
- Let

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

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

Absolute value functions

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

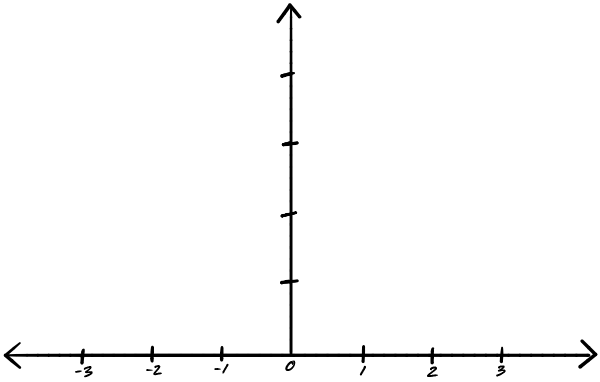


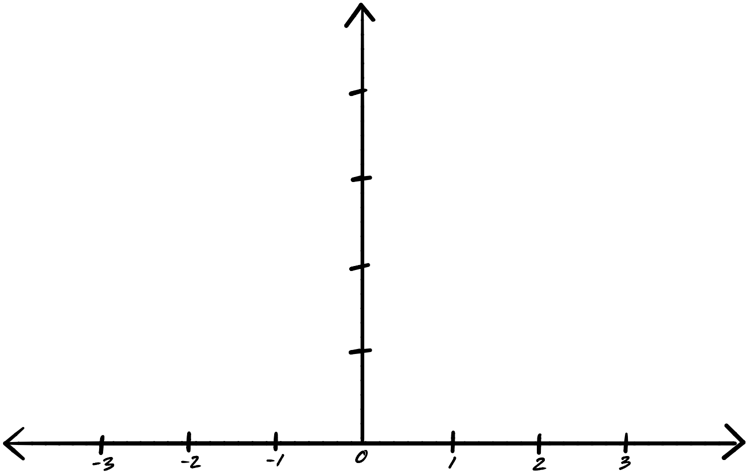
Sums of absolute values

- Let

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

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




-3
-2
-1
0
1
2
3

The slope of the mean absolute error

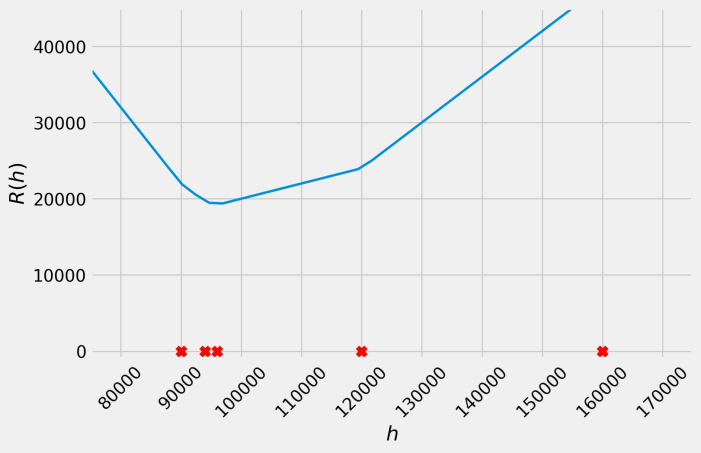
$R(h)$ is a sum of absolute value functions (times $\frac{1}{n}$):

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

The slope of the mean absolute error

The slope of R at h is:

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



Where the slope's sign changes

The slope of R at h is:

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

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 $y_1, \ldots, y_n$

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.

$$90{,}000 \quad 94{,}000 \quad 96{,}000 \quad 120{,}000 \quad 160{,}000$$

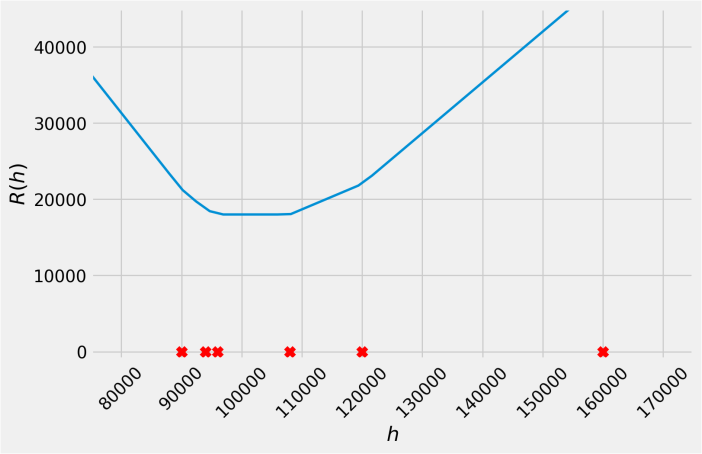
Suppose we collect a 6th salary, so that our data is now

$$90{,}000 \quad 94{,}000 \quad 96{,}000 \quad 108{,}000 \quad 120{,}000 \quad 160{,}000$$

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 in the interval [96,000, 108,000]

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 $h^* = \text{Median}(y_1, \ldots, y_n)$. 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 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) $e^{|y_i - h|}$

b) $|y_i - h|^2$

c) $|y_i - h|^3$

d) $\cos(y_i - h)$

Why?

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: $\text{Median}(y_1, \ldots, y_n)$.
    - The **best prediction**, in terms of mean absolute error, is the **median**.
- We then started to consider another type of error that is differentiable and hence is easier to minimize.
- **Next time:** We will find the value of h^* that minimizes this other error, and see how it compares to the median.