DSC 40A

Theoretical Foundations of Data Science I

In This Video

Which prediction minimizes the mean error?

Recommended Reading

Course Notes: Chapter 1, Section 1

The Best Prediction

- We want the best prediction, h^*.
- Goal: find h that minimizes the mean error:

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

- This is an optimization problem.

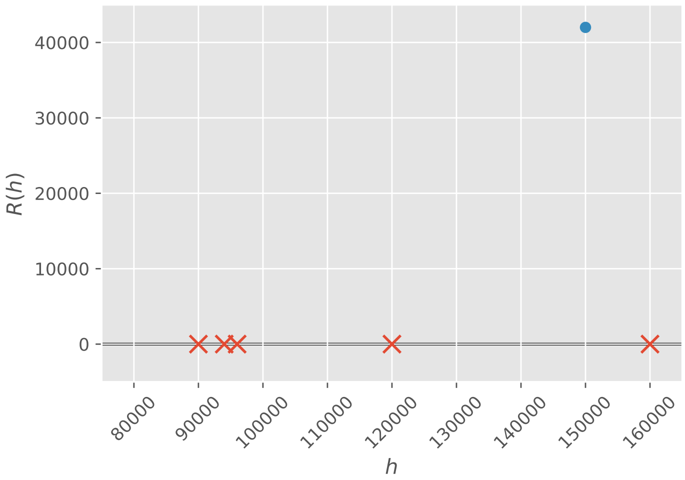
Question

Can we use calculus to minimize R?

Minimizing with Calculus

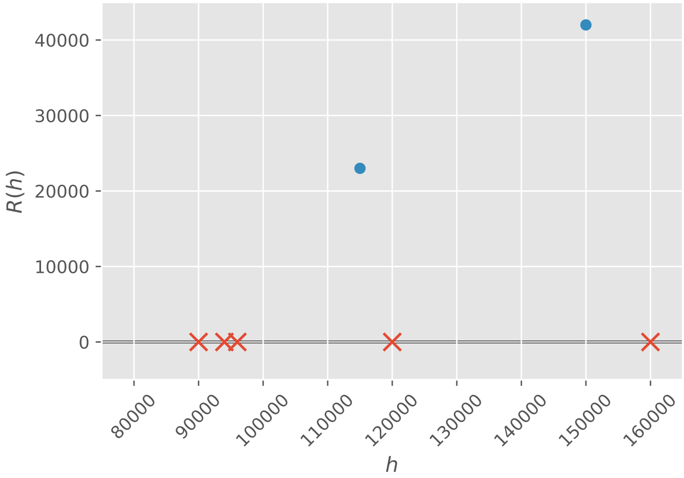
- Calculus: take derivative, set equal to zero, solve.

Plotting the Mean Error



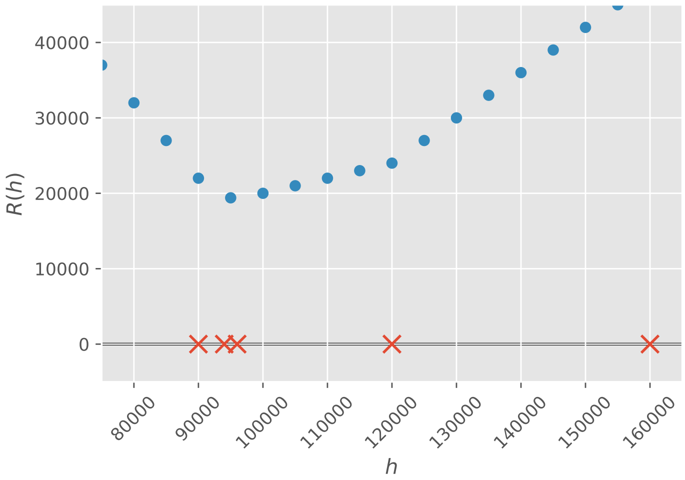
Recall: $R(150{,}000) = 42{,}000$

Plotting the Mean Error

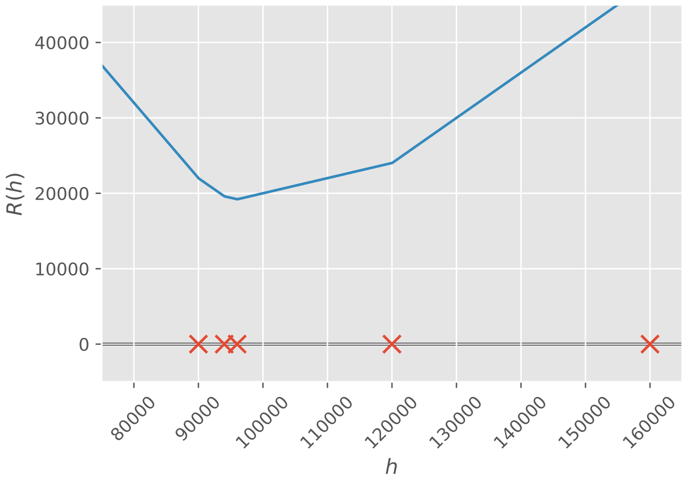


Recall: $R(115{,}000) = 23{,}000$

Plotting the Mean Error



Plotting the Mean Error

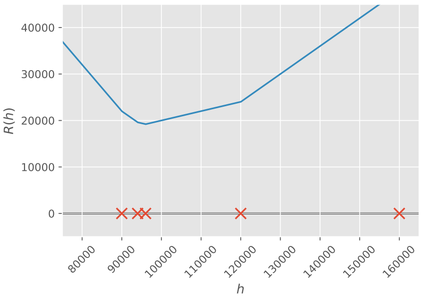


Question

A local minimum occurs when the slope of a function 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

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

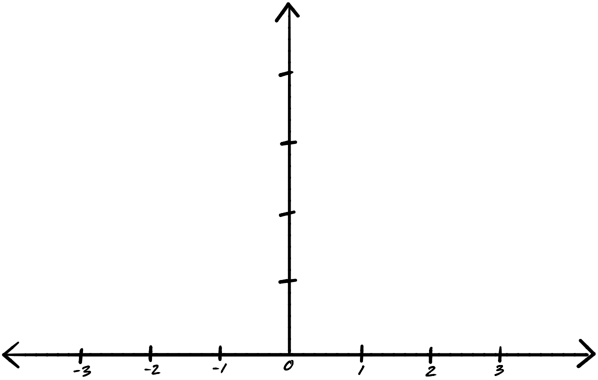
- What is the slope of $f(x) = f_1(x) + f_2(x) + f_3(x)$?

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 $f(x) = f_1(x) + f_2(x) + f_3(x)$?



The Slope of the Mean Error

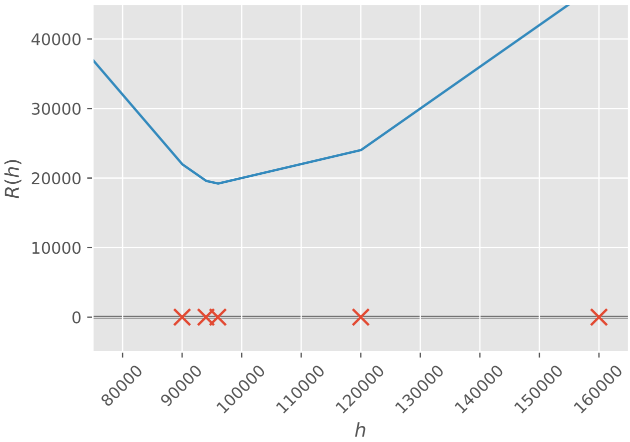
$R(h)$ is a sum of absolute value functions (times $\frac{1}{n}$):

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

The Slope of the Mean 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:

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

Question

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

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

Summary: 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: $\text{Median}(y_1, \ldots, y_n)$.
- The **best prediction**[1] is the **median**.
- **Next time:** We consider a different measure of error that is differentiable.

[1] in terms of mean error