
DSC 40A - Homework 1 - Solutions

due Friday, October 10th at 11:59PM

Homeworks are due to Gradescope by 11:59PM on the due date.

You can use a slip day to extend the deadline by 24 hours; you have four slip days to use in total throughout the quarter.

Homework will be evaluated not only on the correctness of your answers, but on your ability to present your ideas clearly and logically. You should **always explain and justify** your conclusions, using sound reasoning. Your goal should be to convince the reader of your assertions. If a question does not require explanation, it will be explicitly stated.

Homeworks should be written up and turned in by each student individually. You may talk to other students in the class about the problems and discuss solution strategies, but you should not share any written communication and you should not check answers with classmates. You can tell someone how to do a homework problem, but you cannot show them how to do it. **Only handwritten solutions will be accepted (use of tablets is permitted). Do not typeset your homework (using L^AT_EX or any other software).**

For each problem you submit, you should **cite your sources** by including a list of names of other students with whom you discussed the problem. Instructors do not need to be cited.

This homework will be graded out of **61 points**. The point value of each problem or sub-problem is indicated by the number of avocados shown.

Problem 1. Syllabus



Please confirm you have read the course [syllabus](#).

Problem 2. Welcome Survey



Make sure to fill out the [Welcome Survey, linked here](#) for two points on this homework!

Problem 3. Reflection and Feedback Form



Make sure to fill out this Reflection and Feedback Form, linked [here](#), for two points on this homework! This form is primarily for your benefit; research shows that reflecting and summarizing knowledge helps you understand and remember it.

Problem 4. The Proof is in the Pudding

In this problem, you'll prove or disprove various statements about a dataset of numbers, y_1, y_2, \dots, y_n . But first, let's discuss the general approach. (Note that this problem looks long, but most of it is us explaining *how* to answer it!

To prove that a statement is always **true**, you must provide some sort of reason as to why it is always true, no matter what the values y_1, y_2, \dots, y_n are. For example, consider the statement:

“Suppose we add 5 to each of y_1, y_2, \dots, y_n . The mean of the new dataset must be greater than the mean of the original dataset.”

This statement is always true, but it's not enough just to say “This statement is always true; since we're adding a positive number to each value, the mean will also increase.” That's good intuition to have, but we need to provide a more rigorous justification. Here's what a more rigorous justification might look like:

“The mean of the original dataset is $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. The mean of the new dataset is:

$$\frac{1}{n} \sum_{i=1}^n (y_i + 5) = \frac{1}{n} \left(\sum_{i=1}^n y_i + \sum_{i=1}^n 5 \right) = \frac{1}{n} \left(\sum_{i=1}^n y_i + 5n \right) = \frac{1}{n} \sum_{i=1}^n y_i + 5 = \bar{y} + 5$$

Therefore, the mean of the new dataset is equal to the original dataset’s mean plus 5, so the mean of the new dataset is greater than the mean of the original dataset, and so the statement is always true.”

Note that in the argument above, we didn’t assume anything specifically about the numbers in the original dataset — we didn’t use a specific example. Just because a statement holds true for one example, doesn’t mean it always holds true!

On the other hand, to disprove a statement, what you need to show is that it is **not** always true. The easiest way to do this is to provide a counterexample, i.e. a set of values y_1, y_2, \dots, y_n where the statement is false. For example, consider the statement:

“The smallest number in the dataset must be less than the mean.”

Valid justification might look like:

“This statement is not always true. For example, consider the case where our dataset only contains one unique number, like 8, 8, 8. Here, the mean is 8 and the smallest number is 8, so the smallest number is not less than the mean, and so the statement is not always true.”

This is a counterexample, and is a sufficient disproof. (Fun fact: there exist [entire books](#) about counterexamples!)

Note that in both of the examples above, our answers clearly stated whether or not we thought the statement was always true. Your answers should do the same.

Now it’s your turn! Consider a dataset of numbers y_1, y_2, \dots, y_n . Prove or find a counterexample to disprove each of the following statements.

- a) 🥝🥝 At least half of the numbers in the dataset must be smaller than the mean.

Solution: This statement is not always true. As a counterexample, consider the dataset 8, 8. The mean is 8 but none of the numbers in the dataset are smaller than 8. So we do not need at least half of the numbers to be smaller than the mean.

- b) 🥝🥝 Suppose that all of the elements in the dataset are unique. Then, removing the largest element in the dataset must increase the mean.

Solution: This statement is not always true. As a counterexample, consider the dataset 1, 2, 3, which has a mean of $\frac{1+2+3}{3} = 2$. If we remove the largest element, 3, then the mean of the resulting values is $\frac{1+2}{2} = 1.5$, which is lower than the original mean. So, removing the largest element in the dataset does not necessary increase the mean.

- c) 🥝🥝 Suppose that all of the elements in the dataset are unique, that n is odd, and that the mean of the dataset is not equal to the median of the dataset. Then, if we remove the median value from the dataset, the median of the new dataset must be different from the median of the original dataset.

Solution: This statement is not always true. As a counterexample, consider the dataset 1, 2, 3, 4, 10, which has a median of 3. If we remove the median, 3, then the resulting values are 1, 2, 4, 10, which has a median of 3. So, removing the median from the original dataset doesn’t necessarily change the median of the new dataset.

A common mistake was using counter-examples where the mean equaled the median. The reason we specified that the original dataset's mean had to be different from its median was so that you couldn't use the counterexample 1, 2, 3 but had to dig a little deeper.

- d) 🥑🥑 Suppose we introduce a new number to the dataset that is greater than the mean of the existing dataset. The mean of the new dataset must be greater than the mean of the original dataset.

Solution: This statement is always true.

The mean of the original dataset is $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. Note that this means that $\sum_{i=1}^n y_i = n\bar{y}$. We will need to use this fact later on.

Let the new value, y_{n+1} , be equal to $\bar{y} + b$, where $b > 0$ is some positive number. This represents the fact that the new number that we've introduced to the dataset is larger than the original dataset's mean.

The new dataset's mean is:

$$\begin{aligned} \frac{1}{n+1} \sum_{i=1}^{n+1} y_i &= \frac{1}{n+1} \left(\sum_{i=1}^n y_i + y_{n+1} \right) \\ &= \frac{1}{n+1} \left(\sum_{i=1}^n y_i \right) + \frac{y_{n+1}}{n+1} \\ &= \frac{n\bar{y}}{n+1} + \frac{\bar{y} + b}{n+1} \\ &= \frac{n\bar{y} + \bar{y} + b}{n+1} \\ &= \frac{(n+1)\bar{y} + b}{n+1} \\ &= \bar{y} + \frac{b}{n+1} \end{aligned}$$

Since $b > 0$ and $n > 0$, $\frac{b}{n+1} > 0$, and so the mean of the new dataset is greater than the mean of the original dataset.

A common mistake was dividing by n instead of $n+1$, i.e. not adjusting the average to the fact that we have an additional number in our dataset.

Problem 5. Max's Idea

In the lectures, we argued that one way to make a good prediction h was to minimize the mean absolute error:

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

We saw that the median of y_1, \dots, y_n is the prediction with the smallest mean error. Your friend Max has many ideas for other ways to make predictions.

Max suggests that instead of minimizing the mean error, we could minimize the *maximum error*:

$$M(h) = \max_{i=1, \dots, n} |y_i - h|$$

In this problem, we'll see if Max has a good idea.

- a) 🥑🥑🥑🥑 Suppose that the data set is arranged in increasing order, so $y_1 \leq y_2 \leq \dots \leq y_n$. Argue that $M(h) = \max(|y_1 - h|, |y_n - h|)$.

Solution: Since

$$M(h) = \max_{i=1, \dots, n} |y_i - h|,$$

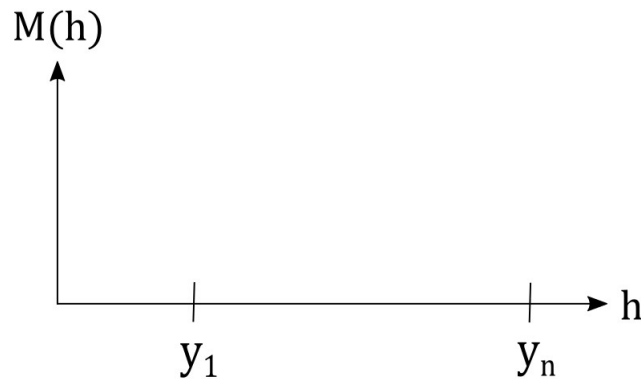
note that the maximum error is the distance from h to the furthest data point. For any h , the furthest data point must be either y_1 or y_n . We can show this by contradiction where we assume that one of the other data points, say y_k with $k \in \{2, \dots, n-1\}$ was the furthest data point from h , then we show that this must be impossible. Consider two cases.

- If $h < y_k$ and y_k is the furthest data point from h , then this is a contradiction because y_n is at least as far from h .
- If $h \geq y_k$ and y_k is the furthest data point from h , then this is a contradiction because y_1 is at least as far from h .

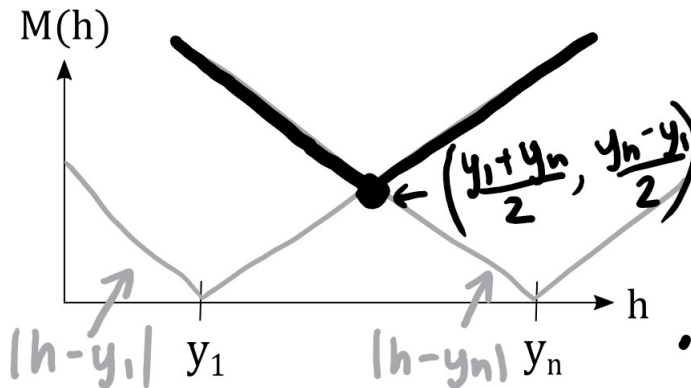
Then, for any h , the furthest data point from h must be y_1 or y_n . Only this furthest data point will be used to determine the maximum error. Therefore, the equation can be simplified as:

$$M(h) = \max(|y_1 - h|, |y_n - h|)$$

- b) 🥑🥑 On the axes below, draw the graph of $M(h) = \max(|y_1 - h|, |y_n - h|)$. Label key points with their coordinates.



Solution:



Note that $M(h)$ is the max of two functions, some students drew the two functions but then didn't indicate $M(h)$ itself.

- c) 🥑🥑🥑🥑 Show that $M(h)$ is minimized at $h^* = \frac{y_1 + y_n}{2}$, which is sometimes called the *midrange* of the data. Then discuss whether Max had a reasonable idea.

Solution: As we can see from the graph, $M(h)$ is minimized at the intersection of the two functions: $|y_1 - h|$ and $|y_n - h|$. Since absolute value functions have slope ± 1 , the intersection point is the midpoint of y_1 and y_n , which is $h^* = \frac{y_1 + y_n}{2}$.

This is a reasonable prediction since it falls in the center of the data set in some sense. Some benefits of using this as a prediction is that it's easy to calculate, and falls between the smallest and largest data values. It's not, however, very sophisticated, and therefore has some big issues. A major drawback is that it is very sensitive to outliers. This prediction works well when the data set is evenly spread out, but it doesn't work well when the dataset is distributed asymmetrically, such as many more high values than low values.

Problem 6. Slippery Slope

As you learned from the lectures, $h^* = \text{Median}(y_1, y_2, \dots, y_n)$ is the constant prediction that minimizes mean absolute error:

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

Suppose that we have a dataset of numbers y_1, y_2, \dots, y_n such that n is **odd**, all values y_i are **distinct**, and that the values are arranged in increasing order. That is, $y_1 \leq y_2 \leq \dots \leq y_n$.

Note: Parts (a) and (b) are independent of each other.

- a) 🥑🥑🥑🥑 Suppose that $R_{\text{abs}}(\alpha) = V$, where V is the minimum value of $R_{\text{abs}}(h)$ and α is one of the numbers in our dataset.

Let $\alpha + \beta$ be the smallest value greater than α in our dataset, where $\beta > 0$. Another way of thinking about this is that $\beta = (\text{smallest value greater than } \alpha) - \alpha$.

Suppose we modify our dataset by replacing the value α with the value $\alpha + \beta + 1$. In our new dataset of n values, what is the new minimum value of $R_{\text{abs}}(h)$ and at what value of h is it minimized? Your answers to both parts should only involve the variables V , α , β , n , and/or one or more constants.

Solution: The new minimum of $R_{\text{abs}}(h)$ is $V + \frac{1}{n}$, and the h^* that minimizes $R_{\text{abs}}(h)$ for the new dataset is $\alpha + \beta$.

We are told that α minimizes the mean absolute error. This indicates that α is the median of the dataset, and since we've given that n is odd, it is the unique minimizer of mean absolute error. Before modifying α , here's how our values look on a number line; note that since n is odd, there are $\frac{n-1}{2}$ values to the left of the median, 1 value equal to the median, and $\frac{n-1}{2}$ values to the right of the median.

$$\underbrace{y_1 \quad y_2 \quad \dots \quad y_{\frac{n-1}{2}}}_{\text{the smallest } \frac{n-1}{2} \text{ values}} \quad \underbrace{\alpha}_{\text{the current median}} \quad \alpha + \beta \quad \underbrace{y_{\frac{n-1}{2}+3} \quad y_{\frac{n-1}{2}+4} \quad \dots \quad y_n}_{\text{the largest } \frac{n-3}{2} \text{ values}}$$

When we modify the value of α to be $\alpha + \beta + 1$, the new median becomes $\alpha + \beta$, which is the value that was immediately to the right of the old median. The **old median** shifts to the right of the **new median**. Note that there might be at least one point between $\alpha + \beta$ and $\alpha + \beta + 1$, i.e. the points don't necessarily "swap" order. For example there might be $y_i = \alpha + \beta + 0.2$ in between them. We don't know if there are such points, and it doesn't affect the solution.

Now that we know that the new minimizer of $R_{\text{abs}}(h)$ is $\alpha + \beta$, we need to calculate $R_{\text{abs}}(\alpha + \beta)$. To do so, we'll find the new **sum** of absolute errors from the median and divide it by n . The old sum of absolute errors from the old median (α) is

$$\sum_{i=1}^n |y_i - \alpha| = Vn,$$

since the old average of absolute errors from the median is V .

To calculate the new $R_{\text{abs}}(h = \alpha + \beta)$, consider that since the median has shifted by β to the right, it has moved β units away from the points that were to the left of the median, and β units closer to the points that were to the right of $\alpha + \beta$.

We can break this into four cases:

- For the smallest $\frac{n-1}{2}$ values, the new median is now β units further away than the old median was. This *adds* $\beta \cdot \left(\frac{n-1}{2}\right)$ to the sum of absolute errors from the median.
- For the $\frac{n-1}{2} - 1$ values that were to the right of the $\alpha + \beta$ before the change — that is, the last bracket above — the new median is now β units closer than the old median was. This *subtracts* $\beta \cdot \left(\frac{n-1}{2} - 1\right)$ from the sum of absolute errors from the median, or equivalently, adds $-\beta \cdot \left(\frac{n-1}{2} - 1\right)$.
- In the old dataset, exactly one of the n values was equal to the median, and that point had an absolute distance of 0 from the median. That's still the case in the new dataset, so this fact alone doesn't change the sum of absolute errors from the median.
- The distance between the median and the point to the right of it used to be $(\alpha + \beta) - \alpha = \beta$, but the distance between these two points is now $(\alpha + \beta + 1) - (\alpha + \beta) = 1$. The difference between these two is $1 - \beta$; if $\beta > 1$, these two points are now closer than they were before, and if $\beta < 1$, these two points are now further than they were before. This adds $1 - \beta$ to the sum of absolute errors from the median.

So, the new sum of absolute errors from the median is the old sum of absolute errors plus all the additions we calculated due to the shift in the median:

$$Vn + \beta \cdot \left(\frac{n-1}{2}\right) + 1 - \beta - \beta \cdot \left(\frac{n-1}{2} - 1\right) = Vn + 1$$

And so the average of absolute errors from the median, $R_{\text{abs}}(\alpha + \beta)$, is:

$$\frac{Vn + 1}{n} = V + \frac{1}{n}$$

- b) 🥑🥑🥑 Let y_a and y_b be two values in our dataset such that $y_a < y_b$ and that the slope of $R_{\text{abs}}(h)$ is the same between $h = y_a$ and $h = y_b$. Specifically, let d be the slope of $R_{\text{abs}}(h)$ between y_a and y_b .

Suppose we introduce a new value q to our dataset such that $q > y_b$. In our new dataset of $n + 1$ values, the slope of $R_{\text{abs}}(h)$ is still the same between $h = y_a$ and $h = y_b$, but it's no longer equal to d . What is the slope of $R_{\text{abs}}(h)$ between $h = y_a$ and $h = y_b$ in our new dataset? Your answer should depend on d , n , q , and/or one or more constants.

Solution:

The new slope is

$$\frac{dn - 1}{n + 1}$$

Recall the formula for the slope (or “derivative”) of $R_{\text{abs}}(h)$:

$$\frac{d}{dh} R_{\text{abs}}(h) = \frac{1}{n} (\# \text{ points left of } h - \# \text{ points right of } h)$$

Let c be some number between y_a and y_b . Since in both the original dataset and new dataset, the slope of $R_{\text{abs}}(h)$ between $h = y_a$ and $h = y_b$ is the same, we can just consider the slope of $R_{\text{abs}}(h)$ at the point $h = c$. (This slope will always be defined because $h = c$ isn't one of our data points.)

Then, we have that:

$$d = \frac{1}{n} (\# \text{ points left of } c - \# \text{ points right of } c)$$

Or, equivalently:

$$\# \text{ points left of } c - \# \text{ points right of } c = nd$$

In our new dataset, since the added value q is greater than y_b , it is also greater than c , which means the number of points to the right of c increases by 1. So, the new slope of $R_{\text{abs}}(h)$ between $h = y_a$ and $h = y_b$ is:

$$\begin{aligned} & \frac{1}{n+1} (\# \text{ points left of } c - (\# \text{ points right of } c + 1)) \\ &= \frac{1}{n+1} (\# \text{ points left of } c - \# \text{ points right of } c - 1) \\ &= \frac{1}{n+1} (nd - 1) \quad (\text{using the result from above}) \\ &= \frac{nd - 1}{n + 1} \end{aligned}$$

So, the new slope of $R_{\text{abs}}(h)$ between $h = y_a$ and $h = y_b$ is $\frac{nd-1}{n+1}$, where d is the original slope of the segment and n is the original number of data points.

A common mistake was using an incorrect formula for derivative of the empirical risk:

$$\frac{d}{dh} R_{\text{abs}}(h) = \frac{1}{n} (\# \text{ points } \mathbf{right} \text{ of } h - \# \text{ points } \mathbf{left} \text{ of } h),$$

i.e. swapping left and right. Also, some students averaged over n points instead of $n + 1$.

Problem 7. New loss function


Suppose we are given a data set of size n with $0 < y_1 \leq y_2 \leq \dots \leq y_n$.

Define a new loss function by

$$L_Q(h, y) = (h^2 - y^2)^2$$

and consider the empirical risk

$$R_Q(h) = \frac{1}{n} \sum_{i=1}^n L_Q(h, y_i).$$

- a)  Show that $R(h)$ has critical points at $h = 0$ and when h equals the **quadratic mean** of the data, defined as

$$QM(y_1, y_2, \dots, y_n) = \sqrt{\frac{y_1^2 + y_2^2 + \dots + y_n^2}{n}}.$$


Solution: We start by finding the critical points of $R_Q(h)$.

$$\begin{aligned} R'_Q(h) &= \frac{d}{dh} \left(\frac{1}{n} \sum_{i=1}^n (h^2 - y_i^2)^2 \right) \\ &= \frac{1}{n} \sum_{i=1}^n \frac{d}{dh} (h^2 - y_i^2)^2 \\ &= \frac{1}{n} \sum_{i=1}^n 2(h^2 - y_i^2) \cdot 2h \\ &= \frac{4h}{n} \sum_{i=1}^n (h^2 - y_i^2) \\ &= \frac{4h}{n} \left(nh^2 - \sum_{i=1}^n y_i^2 \right) \end{aligned}$$

Let $R'_Q(h) = 0$, then either $\frac{4h}{n} = 0$, or $nh^2 - \sum_{i=1}^n y_i^2 = 0$.

Therefore, $R_Q(h)$ has critical points when:

$$h = 0 \text{ and } h = \sqrt{\frac{y_1^2 + y_2^2 + \dots + y_n^2}{n}} = QM(y_1, y_2, \dots, y_n).$$

- b)  Recall from single-variable calculus the **second derivative test**, which says that for a function f with critical point at x^* ,

- if $f''(x^*) > 0$, then x^* is a local minimum, and
- if $f''(x^*) < 0$, then x^* is a local maximum.

Use the second derivative test to determine whether each critical point you found in part (a) is a maximum or minimum of $R_Q(h)$.

Solution: For a critical point h to be a minimum, we must show that $R''_Q(h) > 0$. For a critical point to be a maximum, we must show that $R''_Q(h) < 0$.

We start by rewriting the first derivative and taking the second derivative:


$$\begin{aligned}
 R'_Q(h) &= \frac{4}{n} \left(nh^3 - h \sum_{i=1}^n y_i^2 \right) \\
 R''_Q(h) &= \frac{4}{n} \left(\frac{d}{dh} (nh^3) - \frac{d}{dh} \left(h \sum_{i=1}^n y_i^2 \right) \right) \\
 &= \frac{4}{n} \left(3nh^2 - \sum_{i=1}^n y_i^2 \right) \\
 &= 12h^2 - \frac{4}{n} \sum_{i=1}^n y_i^2
 \end{aligned}$$

When $h = 0$, $R''_Q(h) = -\frac{4}{n} \sum_{i=1}^n y_i^2 < 0$ because $0 < y_1 \leq y_2 \leq \dots \leq y_n$. Therefore $h = 0$ is a local maximum.

When h is the quadratic mean, $h^2 = \frac{1}{n} \sum_{i=1}^n y_i^2$, so

$$\begin{aligned}
 R''_Q(h) &= 12h^2 - \frac{4}{n} \sum_{i=1}^n y_i^2 \\
 &= \frac{12}{n} \sum_{i=1}^n y_i^2 - \frac{4}{n} \sum_{i=1}^n y_i^2 \\
 &= \frac{8}{n} \sum_{i=1}^n y_i^2 \\
 &> 0 \text{ because } 0 < y_1 \leq y_2 \leq \dots \leq y_n.
 \end{aligned}$$

Therefore, $h = QM(y_1, y_2, \dots, y_n)$ is a local minimum.

- c)  Show that the quadratic mean always falls between the smallest and largest data values, which is a property that any reasonable prediction should have. This amounts to proving the inequality

$$y_1 \leq QM(y_1, y_2, \dots, y_n) \leq y_n.$$

Solution: We start by proving $y_1 \leq QM(y_1, y_2, \dots, y_n)$.

Given $0 < y_1 \leq y_2 \leq \dots \leq y_n$, then we know $y_1 \leq y_i$ for $i = 1, 2, \dots, n$. This means $y_1^2 \leq y_i^2$ for $i = 1, 2, \dots, n$.

Therefore,

$$\begin{aligned}
 QM(y_1, y_2, \dots, y_n) &= \sqrt{\frac{\sum_{i=1}^n y_i^2}{n}} \\
 &\geq \sqrt{\frac{y_1^2}{n}} \\
 &= \sqrt{\frac{n \cdot y_1^2}{n}} \\
 &= \sqrt{y_1^2} \\
 &= y_1
 \end{aligned}$$

This shows $QM(y_1, y_2, \dots, y_n) \geq y_1$.

The second part of the inequality can be proved similarly. We are given that $0 < y_1 \leq y_2 \leq \dots \leq y_n$, which implies $y_n \geq y_i$ for $i = 1, 2, \dots, n$. This means $y_n^2 \geq y_i^2$ for $i = 1, 2, \dots, n$.

Therefore,

$$\begin{aligned}
 QM(y_1, y_2, \dots, y_n) &= \sqrt{\frac{\sum_{i=1}^n y_i^2}{n}} \\
 &\leq \sqrt{\frac{y_n^2}{n}} \\
 &= \sqrt{\frac{n \cdot y_n^2}{n}} \\
 &= \sqrt{y_n^2} \\
 &= y_n
 \end{aligned}$$

This shows $QM(y_1, y_2, \dots, y_n) \leq y_n$.

Problem 8. An Alternative

In the lectures, we found that $h^* = \text{Mean}(y_1, y_2, \dots, y_n)$ is the constant prediction that minimizes mean squared error:

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

To arrive at this result, we used calculus: we took the derivative of $R_{\text{sq}}(h)$ with respect to h , set it equal to 0, and solved for the resulting value of h , which we called h^* .

In this problem, we will minimize $R_{\text{sq}}(h)$ in a way that **doesn't** use calculus. The general idea is this: if $f(x) = (x - c)^2 + k$, then we know that f is a quadratic function that opens upwards with a vertex at (c, k) , meaning that $x = c$ minimizes f . As we saw in class (see [Lecture 2, slide 16](#)), $R_{\text{sq}}(h)$ is a quadratic function of h !

Throughout this problem, let y_1, y_2, \dots, y_n be an arbitrary dataset, and let $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ be the mean of the

y 's.

- a) 🥑🥑 What is the value of $\sum_{i=1}^n (y_i - \bar{y})$? Justify your answer.

Solution:

To proceed, we'll use the fact that \bar{y} , by definition, is $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$, meaning that $\sum_{i=1}^n y_i = n\bar{y}$.

$$\begin{aligned}\sum_{i=1}^n (y_i - \bar{y}) &= \sum_{i=1}^n y_i - \sum_{i=1}^n \bar{y} \\ &= n\bar{y} - \bar{y} \sum_{i=1}^n 1 \\ &= n\bar{y} - n\bar{y} \\ &= \boxed{0}\end{aligned}$$

So, $\sum_{i=1}^n (y_i - \bar{y}) = 0$.

- b) 🥑🥑 Show that:

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

Hint: To proceed, start by rewriting $y_i - h$ in the definition of $R_{sq}(h)$ as $(y_i - \bar{y}) + (\bar{y} - h)$. Why is this a valid step? Make sure not to expand unnecessarily.

Solution:

We know that $R_{sq}(h) = \frac{1}{n} \sum_{i=1}^n (y_i - h)^2$. We can write this out as

$$\begin{aligned}&\frac{1}{n} \sum_{i=1}^n (y_i + \bar{y} - \bar{y} - h)^2 \\ &= \frac{1}{n} \sum_{i=1}^n ((y_i - \bar{y}) + (\bar{y} - h))^2 \\ &= \frac{1}{n} \sum_{i=1}^n ((y_i - \bar{y})^2 + 2(y_i - \bar{y})(\bar{y} - h) + (\bar{y} - h)^2)\end{aligned}$$

(Expanding the square)

- c) 🥑🥑 Show that:

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

Hint: At some point, you will need to use your result from part (a).

Solution:

From part b, we know $R_{sq} = \frac{1}{n} \sum_{i=1}^n ((y_i - \bar{y})^2 + 2(y_i - \bar{y})(\bar{y} - h) + (\bar{y} - h)^2)$

$$\begin{aligned}
\Rightarrow R_{sq} &= \frac{1}{n} \left(\sum_{i=1}^n (y_i - \bar{y})^2 + \sum_{i=1}^n 2(y_i - \bar{y})(\bar{y} - h) + \sum_{i=1}^n (\bar{y} - h)^2 \right) \\
&= \frac{1}{n} \left(\sum_{i=1}^n (y_i - \bar{y})^2 + 2(\bar{y} - h) \sum_{i=1}^n (y_i - \bar{y}) + \sum_{i=1}^n (\bar{y} - h)^2 \right) \\
&= \frac{1}{n} \left(\sum_{i=1}^n (y_i - \bar{y})^2 + 2(\bar{y} - h) \cdot 0 + \sum_{i=1}^n (\bar{y} - h)^2 \right) \quad (\text{From part a } \sum_{i=1}^n (y_i - \bar{y}) = 0) \\
&= \frac{1}{n} \left(\sum_{i=1}^n (y_i - \bar{y})^2 + 0 + \sum_{i=1}^n (\bar{y} - h)^2 \right) \\
&= \frac{1}{n} \left(\sum_{i=1}^n (y_i - \bar{y})^2 + \sum_{i=1}^n (\bar{y} - h)^2 \right) \\
&= \frac{1}{n} \left(\sum_{i=1}^n (y_i - \bar{y})^2 + n \cdot (\bar{y} - h)^2 \right) \\
&= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 + (\bar{y} - h)^2
\end{aligned}$$

A common mistake was plugging in $y_i - \bar{y} = 0$, disregarding that what we showed in a) was that it is the sum that is equal to zero: $\sum_{i=1}^n y_i - \bar{y} = 0$.

- d) 🥒 Why does the result in (c) prove that $h^* = \text{Mean}(y_1, y_2, \dots, y_n)$ minimizes $R_{sq}(h)$?

Solution:

From part c, we know $R_{sq} = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 + (\bar{y} - h)^2$. The term $\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$ is the variance, which is a constant that does not depend on h , so we only need to minimize $(\bar{y} - h)^2$. The minimum possible value for this is 0, since it is a squared term and cannot have a negative value. We set $(\bar{y} - h) = 0$ which gives us the equation $h = \bar{y}$. Thus the minimizing value is $h^* = \bar{y}$.

Problem 9. Breakdown Point

Consider the constant model for real-valued scalar input data. For a given loss function, define the *breakdown point* p to be the smallest proportion of data that, when modified at will, can cause the minimizer of the empirical risk to diverge to infinity.

- a) 🥒🥒🥒 Show that the breakdown point of the square loss is $1/n$. That is, if $\{y_1, y_2, \dots, y_{n-1}\}$ are any *fixed* data points, and $y_n = z$ for arbitrary $z \in \mathbb{R}$, prove $\lim_{z \rightarrow \infty} \bar{y} = \infty$.

Solution:

The minimizer of mean squared error is the mean. So we must show that by modifying just 1 of the n data points (such as by making y_n approach infinity), the mean diverges to infinity:

$$\begin{aligned}
\lim_{z \rightarrow \infty} \bar{y} &= \lim_{z \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n y_i \\
&= \lim_{z \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^{n-1} y_i + \frac{1}{n} z \right) \\
&= \frac{1}{n} \sum_{i=1}^{n-1} y_i + \lim_{z \rightarrow \infty} \frac{1}{n} z \\
&= \infty
\end{aligned}$$

A common mistake was taking the limit of the MSE rather than the limit of the mean.

Assume the data $\{y_1, y_2, \dots, y_n\}$ are all distinct and n is odd, i.e., the median is unique. Show that the breakdown point of the absolute loss is 0.5 by completing the following two steps.

- b) 🥑🥑🥑 Suppose you modify strictly less than half of the data by defining $y'_{\lceil n/2 \rceil + 1}, \dots, y'_n = z$ for $z \in \mathbb{R}$. Show that

$$\lim_{z \rightarrow \infty} \text{median}\{y_1, y_2, \dots, y'_{\lceil n/2 \rceil + 1}, \dots, y'_n\} = \max\{y_1, y_2, \dots, y_{\lceil n/2 \rceil}\}$$

In other words, the minimizer of the empirical risk does not diverge to infinity.

Solution:

Notice that there is no assumption that $y_1 \leq y_2 \leq \dots \leq y_n$, so $y_{\lceil \frac{n}{2} \rceil + 1}, y_{\lceil \frac{n}{2} \rceil + 2}, \dots, y_n$ represent $\lfloor \frac{n}{2} \rfloor$ points selected arbitrarily from the dataset.

Let $M = \max\{y_1, y_2, \dots, y_{\lceil n/2 \rceil}\}$, the maximum of the unchanged data points. As soon as $z > M$, there are $\lfloor \frac{n}{2} \rfloor$ points to the left of M (since it's the max of the unchanged points) and $\lfloor \frac{n}{2} \rfloor$ points to the right of M (we moved $\lfloor \frac{n}{2} \rfloor$ points to the right of it). Hence M is the median.

As we continue to increase z , M is preserved as the median by the same reasoning. Thus

$$\lim_{z \rightarrow \infty} \text{median}\{y_1, y_2, \dots, y'_{\lceil n/2 \rceil + 1}, \dots, y'_n\} = M$$

- c) 🥑🥑🥑 Suppose you modify strictly greater than half of the data by defining $y'_{\lceil n/2 \rceil}, \dots, y'_n = z$ for $z \in \mathbb{R}$. Show that

$$\lim_{z \rightarrow \infty} \text{median}\{y_1, y_2, \dots, y'_{\lceil n/2 \rceil}, \dots, y'_n\} = \infty.$$

Solution:

In this case, $\lfloor \frac{n}{2} \rfloor$ points are unchanged and $\lceil \frac{n}{2} \rceil$ points have been moved to z . As soon as z is greater than the maximum of the unchanged points, the $\lfloor \frac{n}{2} \rfloor$ unchanged points $y_1, \dots, y_{\lfloor \frac{n}{2} \rfloor}$ are all less than z , and there are $\lceil \frac{n}{2} \rceil$ points greater than or equal to z . This makes z the median. This property is preserved as z increases, so

$$\lim_{z \rightarrow \infty} \text{median}\{y_1, y_2, \dots, y'_{\lceil n/2 \rceil}, \dots, y'_n\} = \infty$$

- d) 🥑🥑🥑 Explain how you can interpret (a) and (b) in the context of sensitivity to outliers.

Solution: (a) and (b) show that the square loss is much more sensitive to outliers than the absolute loss. Using the square loss, just changing 1 out of the n data points to approach infinity (a ridiculously large outlier) could cause the minimizer to diverge to infinity. Meanwhile, using absolute loss, it requires us to change at least half of the data to approach infinity before the minimizer even changes at all.