DSC 40A - Homework 1

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 type-set your homework (using LaTeXor 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, $y_1, y_2, ..., y_n$. 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 $y_1, y_2, ..., y_n$ are. For example, consider the statement:

"Suppose we add 5 to each of $y_1, y_2, ..., y_n$. 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 $y_1, y_2, ..., y_n$ 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 $y_1, y_2, ..., y_n$. 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.

b) 🥑🥑 Suppose that all of the elements in the dataset are unique. Then, removing the largest element in the dataset must 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.

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.

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 $y_1, \ldots, y_n$ is the prediction with the smallest mean error. Your friend Max has many ideas for other ways to make predictions.

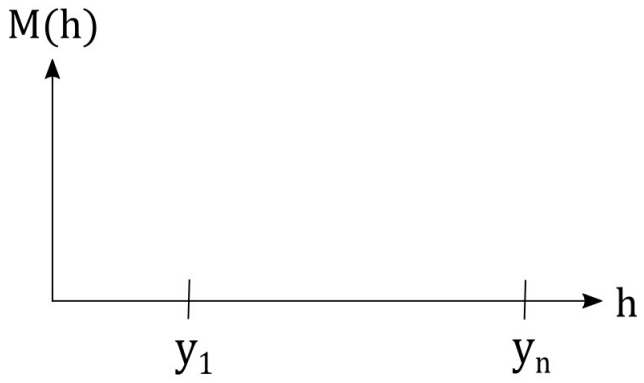
Max suggests that instead of minimizing the mean error, we could minimize the *maximum error*:

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

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

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

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



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.

Problem 6. Slippery Slope

As you learned from the lectures, $h^* = \text{Median}(y_1, y_2, ..., y_n)$ 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 $y_1, y_2, ..., y_n$ such that n **is odd**, all values y_i are **distinct**, and that the values are arranged in increasing order. That is, $y_1 \leq y_2, \leq \ldots \leq y_n$.

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 $\beta =$ (smallest value greater than α) $- \alpha$.

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.

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.

Problem 7. New loss function

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

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

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

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 $R_Q(h)$.

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

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

Problem 8. An Alternative

In the lectures, we found that $h^* = \text{Mean}(y_1, y_2, ..., y_n)$ 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 $y_1, y_2, ..., y_n$ 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.

b) 🥑🥑 Show that:

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

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.

c) 🥑🥑 Show that:

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

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

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

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 $\{y_1, y_2, \ldots, y_{n-1}\}$ are any *fixed* data points, and $y_n = z$ for arbitrary $z \in \mathbb{R}$, prove $\lim_{z\to\infty} \overline{y} = \infty$.

Assume the data $\{y_1, y_2, \ldots, y_n\}$ 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 $y'_{\lceil n/2\rceil+1}, \ldots, y'_n = z$ for $z \in \mathbb{R}$. Show that

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

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

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

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

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