
DSC 40A - Homework 1

Due: Wednesday, Jan 17 at 11:59pm

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. 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.

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.

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 50 points. The point value of each problem or sub-problem is indicated by the number of avocados shown.

Problem 1. Welcome Survey


 Please fill out the [Welcome Survey, linked here](#) for two points!

Problem 2. Means


Which of the following statements *must* be true? Remember to justify all answers.

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


Solution: This statement is not always true, for example in the case of the data set 8, 8. The mean is 8 but none of the numbers in the data set are smaller than 8. So we do not need at least half of the numbers to be smaller than the mean.

- b)  Some of the numbers in the data set must be smaller than the mean.

Solution: This statement is not always true. The same example data set 8, 8 shows we do not need some of the numbers to be smaller than mean.

- c)  Exactly half of the numbers in a data set must be smaller than the mean.

Solution: This statement is not always true. The same example data set 8, 8 shows we do not need exactly half of the numbers to be smaller than the mean.

- d)  Not all of the numbers in the data set can be smaller than the mean.

Solution: This statement is always true, because we can always find an element of a data set (namely, the largest one) that is greater than or equal to the mean. So it is impossible to have all the numbers in the data set be smaller than the mean.

Problem 3. Medians

- a) 🤔🤔 Is the following statement true? Remember to justify all answers.

"Half of the numbers in a data set must be smaller than or equal to the median."

Solution: This statement is true.

- If the dataset has an *odd* number of observations, the median is the middle number.
- If the dataset has an *even* number of observations, the median is the average of the two middle numbers.

For example, in the dataset $[3, 5, 7, 9, 11]$, the median is 7, as it is the middle number. In the dataset $[3, 5, 7, 9]$, the median is $\frac{5+7}{2} = 6$, which is the average of the two middle numbers.

In both cases, half of the numbers in the dataset will be smaller than the median.

- b) 🤔🤔🤔 Consider the dataset $D = [1, 2, 3, 4, 5, 6, 1000]$. This dataset includes an outlier (1000) which significantly differs from the other values. Calculate the mean and median of D and compare their difference. Write briefly about what you observed. (This simple example shows why median is "robust" compared with mean.)

Solution:

Mean:

$$\text{Mean} = \frac{1 + 2 + 3 + 4 + 5 + 6 + 1000}{7} = \frac{1021}{7} \approx 145.86$$

Median: The median is the middle value of the ordered dataset. Since there are 7 values, the median is the 4th value (as it is the middle one).

1, 2, 3, 4, 5, 6, 1000

Observation: The mean is significantly higher due to the influence of the outlier (1000). In contrast, the median remains at 4, reflecting the central tendency of the majority of the data points without being skewed by the outlier. This example illustrates that the median is a more robust measure of central tendency in the presence of outliers, as it is less influenced by extreme values compared to the mean.

Problem 4. Linear Functions

Consider the linear function $f(x) = 3x - 7$.

- a) 🤔🤔 If $a \leq b$, show that $f(a) \leq f(b)$.

Solution: If $a \leq b$, then $3a \leq 3b$, and $3a - 7 \leq 3b - 7$ (using properties of inequalities). This says $f(a) \leq f(b)$.

- b) 🤔🤔 Both of the statements below are true, but only one is a consequence of the property you proved in part (a). Which is it? Show that this statement is true, using the result of part (a).

1. $\text{Mean}(f(x_1), \dots, f(x_n)) = f(\text{Mean}(x_1, \dots, x_n))$
2. $\text{Median}(f(x_1), \dots, f(x_n)) = f(\text{Median}(x_1, \dots, x_n))$ [Hint: separately discuss the case where n is even and n is odd]

Solution: Part (a) implies the property

$$\text{Median}(f(x_1), \dots, f(x_n)) = f(\text{Median}(x_1, \dots, x_n))$$

because part (a) says that the order of the data values x_1, x_2, \dots, x_n on a number line is the same as the order of the data values $f(x_1), f(x_2), \dots, f(x_n)$. This means that if x_i is the middle data value of x_1, x_2, \dots, x_n , then $f(x_i)$ is the middle data value of $f(x_1), f(x_2), \dots, f(x_n)$. This shows that when n is odd,

$$\text{Median}(f(x_1), \dots, f(x_n)) = f(\text{Median}(x_1, \dots, x_n)).$$

If n is even and there are two middle values, x_i and x_j , then part (a) implies that the two middle values among $f(x_1), f(x_2), \dots, f(x_n)$ are $f(x_i)$ and $f(x_j)$. So

$$\begin{aligned} \text{Median}(f(x_1), \dots, f(x_n)) &= \frac{f(x_i) + f(x_j)}{2} \\ &= \frac{3x_i - 7 + 3x_j - 7}{2} \\ &= 3 \left(\frac{x_i + x_j}{2} \right) - 7 \\ &= f(\text{Median}(x_1, \dots, x_n)) \end{aligned}$$

- c) 🤔🤔 Now, prove the other statement.

Solution:

$$\begin{aligned} \text{Mean}(f(x_1), \dots, f(x_n)) &= \frac{f(x_1) + f(x_2) + \dots + f(x_n)}{n} \\ &= \frac{(3x_1 - 7) + (3x_2 - 7) + \dots + (3x_n - 7)}{n} \\ &= \frac{3(x_1 + x_2 + \dots + x_n) - 7n}{n} \\ &= 3 \left(\frac{x_1 + x_2 + \dots + x_n}{n} \right) - 7 \\ &= f(\text{Mean}(x_1, \dots, x_n)) \end{aligned}$$

- d) Suppose we consider a different linear function $g(x) = -5x + 4$. Prove or find a counterexample to disprove each of the following:

1. 😊😊 If $a \leq b$, then $g(a) \leq g(b)$.

Solution: This is false. A counterexample is $a = 1, b = 2$. Then $g(a) = g(1) = -5*1 + 4 = -1$ and $g(b) = g(2) = -5*2 + 4 = -6$. So $a \leq b$ but $g(a) > g(b)$.

2. 😊😊 $\text{Mean}(g(x_1), \dots, g(x_n)) = g(\text{Mean}(x_1, \dots, x_n))$

Solution: This is true. A proof is shown here.

$$\begin{aligned} \text{Mean}(g(x_1), \dots, g(x_n)) &= \frac{g(x_1) + g(x_2) + \dots + g(x_n)}{n} \\ &= \frac{(-5x_1 + 4) + (-5x_2 + 4) + \dots + (-5x_n + 4)}{n} \\ &= \frac{-5(x_1 + x_2 + \dots + x_n) + 4n}{n} \\ &= -5 \left(\frac{x_1 + x_2 + \dots + x_n}{n} \right) + 4 \\ &= g(\text{Mean}(x_1, \dots, x_n)) \end{aligned}$$

3. 😊😊 $\text{Median}(g(x_1), \dots, g(x_n)) = g(\text{Median}(x_1, \dots, x_n))$

Solution: This is true. A proof relies on the fact that if $a \leq b$, then $g(a) \geq g(b)$. This comes from the property of inequalities that says the inequality reverses when we multiply by a negative number. So if $a \leq b$, then $-5a \geq -5b$ and $-5a + 4 \geq -5b + 4$.

This fact implies that the order of the data values x_1, x_2, \dots, x_n on a number line is the opposite as the order of the data values $g(x_1), g(x_2), \dots, g(x_n)$. In other words, applying the function g reverses the order of the data. This means that if x_i is the middle data value of x_1, x_2, \dots, x_n , then $g(x_i)$ is still the middle data value of $g(x_1), g(x_2), \dots, g(x_n)$. This shows that when n is odd,

$$\text{Median}(g(x_1), \dots, g(x_n)) = g(\text{Median}(x_1, \dots, x_n)).$$

If n is even and there are two middle values, x_i and x_j , then since the order of the data gets reversed by g , this means that the two middle values among $g(x_1), g(x_2), \dots, g(x_n)$ are $g(x_j)$ and $g(x_i)$. So

$$\begin{aligned} \text{Median}(g(x_1), \dots, g(x_n)) &= \frac{g(x_j) + g(x_i)}{2} \\ &= \frac{-5x_j + 4 + -5x_i + 4}{2} \\ &= -5 \left(\frac{x_i + x_j}{2} \right) + 4 \\ &= g(\text{Median}(x_1, \dots, x_n)) \end{aligned}$$

Problem 5. Max's Idea

In the first lecture, 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

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

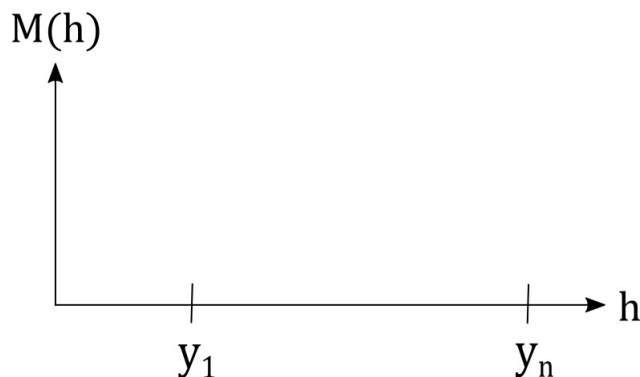
note that the maximum error is the distance from h to the furthest data point. For any h , the furthest data point must be either y_1 or y_n . We can show this by contradiction where we assume that one of the other data points, say y_k with $k \in \{2, \dots, n-1\}$ was the furthest data point from h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

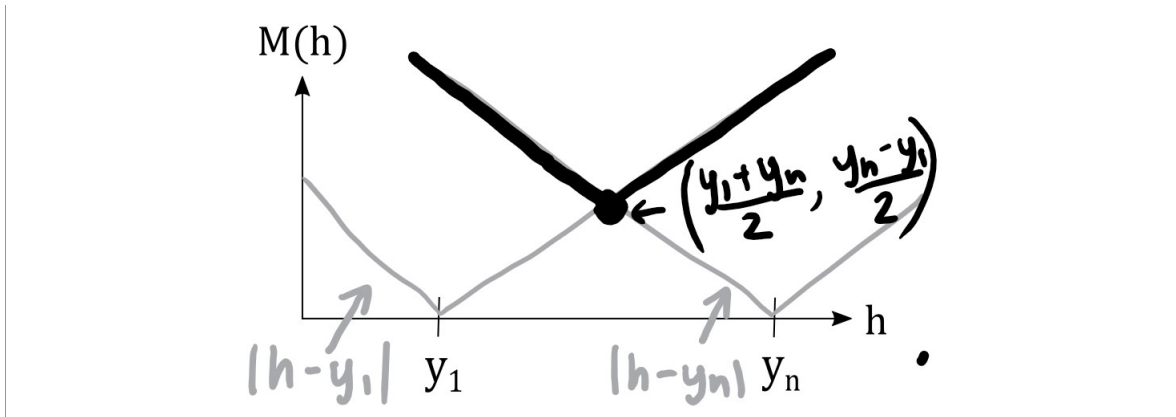
Then, for any h , the furthest data point from h must be y_1 or y_n . Only this furthest data point will be used to determine the maximum error. Therefore, the equation can be simplified as:

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

- 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:



- 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. Max's Other Idea

You friend Max from problem 3 has another idea. He suggests that instead of minimizing mean absolute error, we try maximizing the following quantity:

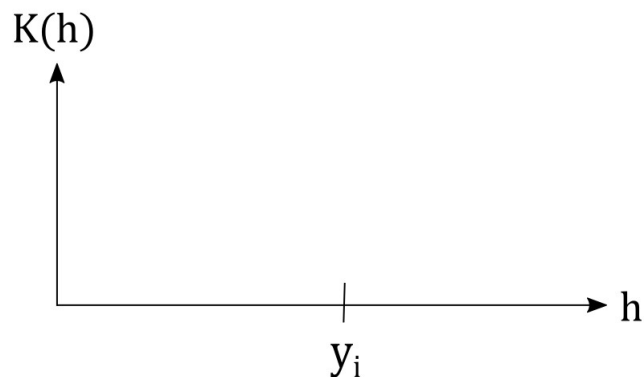
$$P(h) = \prod_{i=1}^n e^{-|h-y_i|}.$$

The above formula is written using product notation, which is similar to summation notation, except terms are multiplied and not added. For example,

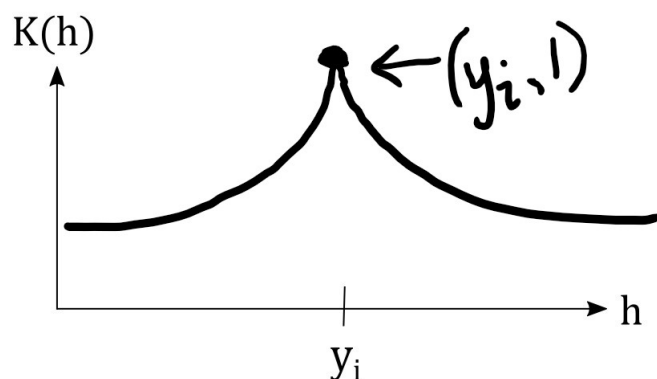
$$\prod_{i=1}^n a_i = a_1 \cdot a_2 \cdot \dots \cdot a_n.$$

Max's reasoning is that for some models, $K(h) = e^{-|h-y_i|}$ is used to compute how likely prediction h will appear given the observation y_i – hence it is called “likelihood.” Then, we should attempt to maximize the chance of getting the prediction h , given the set of observations. In this problem, we'll see if Max has a good idea.

- a) 🤔🤔🤔🤔 On the axes below, sketch a graph of the basic shape of the likelihood function $K(h) = e^{-|h-y_i|}$. Label key points with their coordinates. Explain, based on the graph, why larger values of $K(h)$ correspond to better predictions h .



Solution:



Because larger values of $K(h)$ are closer to the data point x_i , they correspond to better predictions h .

- b) 🤔🤔🤔🤔 Recall from Groupwork 1 that for a function of one variable $f(x)$, a value x^* is said to be a **minimizer** of $f(x)$ if

$$f(x^*) \leq f(x) \quad \text{for all } x.$$

Similarly, x^* is said to be a **maximizer** of $f(x)$ if

$$f(x^*) \geq f(x) \quad \text{for all } x.$$

Suppose that $f(x)$ is a function that is minimized at x^* and c is a positive constant. Show that the function $g(x) = c \cdot f(x)$ is minimized at x^* and the function $q(x) = -c \cdot f(x)$ is maximized at x^* .

Solution: To show that x^* (the minimizer of f) is also a minimizer of g , we need to argue that $g(x^*) \leq g(x)$ for every possible x .

We'll start with the fact that x^* is a minimizer of f . We have, for every x , that $f(x^*) \leq f(x)$. For an arbitrary positive constant c , we have

$$\begin{aligned} f(x^*) &\leq f(x) \\ c \cdot f(x^*) &\leq c \cdot f(x) \\ g(x^*) &\leq g(x) \quad \text{for every possible } x. \end{aligned}$$

Following similar argument for $q(x)$, we have:

$$\begin{aligned} f(x^*) &\leq f(x) \\ -c \cdot f(x^*) &\geq -c \cdot f(x) \\ q(x^*) &\geq q(x) \quad \text{for every possible } x. \end{aligned}$$

This proves that x^* is a minimizer of g and maximizer for q .

- c) 🤔🤔🤔🤔 In physics and some machine learning models, people prefer to minimize **Negative Log Likelihood (NLL)** over maximize likelihood. Calculate the NLL of $P(h)$, and use it to explain whether or not Max had a reasonable idea.

Hint: assume all logs are natural logarithm \ln (i.e log base e). some useful log equalities:

$$\ln(a * b) = \ln(a) + \ln(b)$$

$$\ln(e^a) = a * \ln(e) = a$$

Solution: Starting from

$$P(h) = \prod_{i=1}^n e^{-|h-y_i|}$$

Adding a negative log sign to both sides of the equation, we have:

$$\begin{aligned} -\ln P(h) &= -\ln \prod_{i=1}^n e^{-|h-y_i|} \\ &= -\ln(e^{-|h-y_1|} * e^{-|h-y_2|} * \dots * e^{-|h-y_n|}) \\ &= -\ln(e^{-|h-y_1|}) - \ln(e^{-|h-y_2|}) - \dots - \ln(e^{-|h-y_n|}) \\ &= |h-y_1| + |h-y_2| + \dots + |h-y_n| \\ &= \sum_{i=1}^n |h-y_i| \end{aligned}$$

Recall from the lecture that the mean absolute error is:

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

Clearly, $R(h)$ and the NLL of $P(h)$ is only different by a constant $1/n$. Since n represents the size of our dataset, $1/n$ can never be negative. We know that $R(h)$ is minimized at the median of the data, and by part (b), therefore $P(h)$ is maximized at the median of the data. As a result, Max's idea of minimizing the total error isn't any different than our original idea of minimizing the mean absolute error, which is a very reasonable way to make predictions.