

---

## 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.
- b)  Some of the numbers in the data set must be smaller than the mean.
- c)  Exactly half of the numbers in a data set must be smaller than the mean.
- d)  Not all of the numbers in the data set can 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."
- 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.)

### Problem 4. Linear Functions

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

- a) 🤔🤔 If  $a \leq b$ , show that  $f(a) \leq f(b)$ .
- 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]
- c) 🤔🤔 Now, prove the other statement.
- 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  $g(a) \leq g(b)$ .
  2. 🤔🤔  $\text{Mean}(g(x_1), \dots, g(x_n)) = g(\text{Mean}(x_1, \dots, x_n))$
  3. 🤔🤔  $\text{Median}(g(x_1), \dots, g(x_n)) = g(\text{Median}(x_1, \dots, x_n))$

### 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  $M(h) = \max(|y_1 - h|, |y_n - h|)$ .
- 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. Max's Other Idea


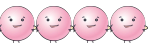
Your friend Max from problem 3 has another idea. He suggests that instead of minimizing mean absolute error, we try maximizing the following quantity:

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

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.

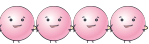
- a)  On the axes below, sketch a graph of the basic shape of the likelihood function  $K(h) = e^{-|h-y_i|}$ . Label key points with their coordinates. Explain, based on the graph, why larger values of  $K(h)$  correspond to better predictions  $h$ .
- b)  Recall from Groupwork 1 that for a function 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  $x^*$ .

- 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:*

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

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