DSC 40A - Homework 2
Due: Tuesday, April 18th 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; 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.

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

For Homework 2, it is *required* that you type your solutions in $\LaTeX$, using the Overleaf template on the course website.

Problem 1. Reflection and Feedback Form

🥑🥑🥑🥑 Make sure to fill out this Reflection and Feedback Form, linked here for four 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 2. Command+C Command+V

Consider a dataset of n numbers $y_1, y_2, ..., y_n$ with mean M and standard deviation S. Suppose we introduce k new values to the dataset, $y_{n+1}, y_{n+2}, ..., y_{n+k}$, all of which are equal to M.

Let the new mean and standard deviation of all $n + k$ values be M' and S', respectively.

a) 🥑🥑 Find M' in terms of M, n, k, and S. (You may not need to use all of these variables in your answer.)

b) 🥑🥑🥑🥑 Find S' in terms of M, n, k, and S. (You may not need to use all of these variables in your answer.)

Problem 3. Living in Harmony

In Lecture 2, we found that using the squared loss function, $L_{\text{sq}}(y_i, h) = (y_i - h)^2$, the constant prediction that minimizes empirical risk is $h^* = \text{Mean}(y_1, y_2, ..., y_n)$.

In this problem, we will look at a new loss function, the "relative squared loss function" $L_{\text{rsq}}(y_i, h)$:

$$L_{\text{rsq}}(y_i, h) = \frac{(y_i - h)^2}{y_i}$$

Throughout this problem, assume that each of $y_1, y_2, ..., y_n$ is positive.

a) 🥑🥑 Determine $\frac{\partial}{\partial h} L_{\text{rsq}}$, the partial derivative of the relative squared loss function with respect to h.

b) 🥑🥑🥑 What value of h minimizes empirical risk for the relative squared loss function — that is, what is h^*? Your answer should only be in terms of the variables $n, y_1, y_2, ..., y_n$, and any constants.

c) 🥑🥑🥑 Let $M(y_1, y_2, ..., y_n)$ be your minimizer h^* from the part (b). That is, for a particular dataset $y_1, y_2, ..., y_n$, $M(y_1, y_2, ..., y_n)$ is the value of h that minimizes empirical risk for relative squared loss on that dataset.

What is the value of $\lim_{y_4 \to \infty} M(1, 3, 5, y_4)$ in terms of $M(1, 3, 5)$? Your answer should involve the function M and/or one or more constants.

Hint: To notice the pattern, evaluate $M(1, 3, 5, 100)$, $M(1, 3, 5, 10000)$, and $M(1, 3, 5, 1000000)$.

d) 🥑🥑 What is the value of $\lim_{y_4 \to 0} M(1, 3, 5, y_4)$? Again, your answer should involve the function M and/or one or more constants.

e) 🥑🥑 Based on the results of parts (c) and (d), when is the prediction $M(y_1, y_2, ..., y_n)$ robust to outliers? When is it not robust to outliers?

Problem 4. Slippery Slope

In Lecture 2, we found that $h^* = \text{Median}(y_1, y_2, ..., y_n)$ 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** and the values are arranged in increasing order. That is, $y_1 \leq y_2, \leq ... \leq y_n$.

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.

b) 🥑🥑🥑 Let y_a and y_b be two values in our dataset such that $y_a < y_b$ and that the slope of $R_{\mathrm{abs}}(h)$ is the same between $h = y_a$ and $h = y_b$. Specifically, let d be the slope of $R_{\mathrm{abs}}(h)$ 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 $R_{\mathrm{abs}}(h)$ 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 $R_{\mathrm{abs}}(h)$ 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 5. Happy Family

In class, we defined the *mean absolute deviation from the median* as a measure of the spread of a dataset. This measure takes the absolute deviations, or distances, of each value in the dataset from the median, and computes the mean of these absolute deviations. We can think of this one measure of spread as a member of a family of analogously defined measures of spread:

- mean absolute deviation from the median
- median absolute deviation from the median
- mean absolute deviation from the mean
- median absolute deviation from the mean

While all four of these measures capture the notion of spread, they do so in different ways, and so they may have different values for the same dataset.

a) 🥑🥑🥑🥑 Consider the histogram shown below.
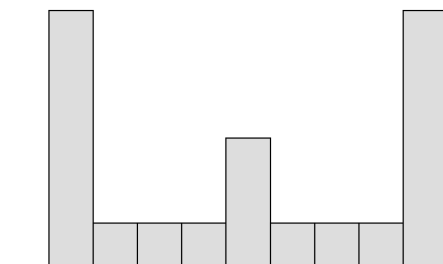
1. Draw a histogram showing the rough shape of the distribution of the absolute deviations from the mean in the histogram above.

2. In the above histogram, which of these two measures is greater, or are they about the same? Justify your answer, identifying a feature in the graph that helped you get there.

   - mean absolute deviation from the mean
   - median absolute deviation from the mean

b) 🥑🥑🥑🥑 Consider the histogram shown below.
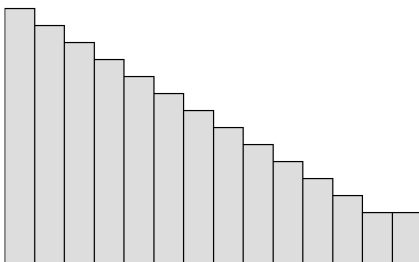
1. Draw a histogram showing the rough shape of the distribution of the absolute deviations from the median in the histogram above.

2. In the above histogram, which of these two measures is greater, or are they about the same? Justify your answer, identifying a feature in the graph that helped you get there.

   - mean absolute deviation from the median

   - median absolute deviation from the median

Problem 6. Zoe's Bakery

Zoe owns a bakery and wants to figure out how to sell the most baked goods.

a) 🥑🥑🥑 For each of her five baked goods, Zoe recorded the cost in dollars for making the baked good, x, and the number of orders for that baked good on a particular day, y.

baked good	cost (x)	number of baked goods sold (y)
Cookies	4	70
Brownies	11	80
Croissants	8	40
Cupcakes	7	57
Muffins	5	43

For example, Zoe spent \$11 making brownies, and sold 80 brownies.

Find the optimal parameters c_0^* and c_1^* that minimize mean squared error for the hypothesis function $H(x) = c_0 + c_1 x$, which predicts the number of baked goods sold of a particular item as a function of the cost of baking that item. Give exact values for c_0^* and c_1^*; do not round. You may use a calculator, but you must show all of your work directly in LaTeX.

b) 🥑🥑 Let's interpret the meaning of the hypothesis function $H(x) = c_0^* + c_1^* x$ that you found in part (a).

   - What does $50 \cdot c_1^*$ represent in terms of Zoe's bakery?

   - What does the reciprocal of the slope, $\frac{1}{c_1^*}$, represent in terms of Zoe's bakery?

c) 🥑🥑 What is the mean squared error, MSE_x, for this dataset, using the line you found in part (a)? Round your final answer to three decimal places. Again, you may use a calculator, but you must show all of your work directly in LaTeX.

d) 🥑🥑🥑 Zoe knows that baking each baked good takes a significant amount of time. She decides to quantify the value of baking time in terms of the number of baked goods sold. For each baked good she baked, Zoe recorded the number of hours to bake one unit of the good, z, and the number of items sold on a particular day, y.

baked good	baking time (z)	number of baked goods sold (y)
Cookies	30	70
Brownies	44	80
Croissants	38	40
Cupcakes	36	57
Muffins	32	43

Find the optimal parameters d_0^* and d_1^* that minimize mean squared error for the hypothesis function $H(z) = d_0 + d_1 z$, which predicts the number of baked goods sold of a particular item as a function of

4

the baking time of that item. Give exact values for d_0^* and d_1^*; do not round. You may use a calculator, but you must show all of your work directly in LaTeX.

e) 🥑🥑 What is the mean squared error, MSE_z, for this dataset, using the line you found in part (d)? Round your final answer to three decimal places. Again, you may use a calculator, but you must show all of your work directly in LaTeX.

f) 🥑🥑🥑🥑🥑 You should have found that $\text{MSE}_x = \text{MSE}_z$, which says that for this data, the mean squared error is the same if we use the variable x or the variable z to make our hypothesis function H. This happens because the number of hours required to bake one unit of a baked good (z) is linearly related to the cost of baking that baked good (x) by the formula:

$$\text{baking time} = 22 + 2 \cdot \text{cost}$$

In the rest of this problem, we'll verify some general properties concerning the scenario where we predict some variable y based on x, as compared to predicting y based on z, when z is a linear transformation of x. We'll no longer use the bakery data given above, but we'll prove properties in general.

First, suppose we have a dataset $\{x_1, x_2, \ldots, x_n\}$ and we define a dataset $\{z_1, z_2, \ldots, z_n\}$ by the linear transformation:

$$z_i = ax_i + b$$

Suppose also we have a dataset $\{y_1, y_2, \ldots, y_n\}$.

Let c_0^* and c_1^* be the optimal intercept and slope of the regression line (that is, the optimal linear hypothesis function) for y with x as the predictor variable,

$$H(x) = c_0^* + c_1^* x$$

Similarly, let d_0 and d_1 be the intercept and slope of the regression line for y with z as the predictor variable,

$$H(z) = d_0^* + d_1^* z$$

Express d_0^* and d_1^* in terms of c_0^*, c_1^*, a, and b, and/or one or more constants.

Hint: In Homework 1, Question 3, you proved that $\text{Mean}(f(x_1), \ldots, f(x_n)) = f(\text{Mean}(x_1, \ldots, x_n))$ for the linear function in the problem. This actually holds true for any linear function, so you can use the fact that if $y_i = ax_i + b$, then $\bar{y} = a\bar{x} + b$ without proof.

g) 🥑🥑🥑 Let MSE_x be the mean squared error for the dataset $\{y_1, y_2, \ldots, y_n\}$ using the hypothesis function:

$$H(x) = c_0^* + c_1^* x.$$

Similarly, let MSE_z be the mean squared error for the dataset $\{y_1, y_2, \ldots, y_n\}$ using the hypothesis function:

$$H(z) = d_0^* + d_1^* z$$

Show that $\text{MSE}_x = \text{MSE}_z$.