
DSC 40A - Homework 2
due Friday, October 18th at 11:59PM

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

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.

Homework 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. We encourage you to type your solutions in L^AT_EX, using the Overleaf template on the course website.


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 49 points. The point value and difficulty 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 L^AT_EX, using the Overleaf template on the course website.

Note: For Problem 6, you'll need to code your answers in Python. More detailed instructions are provided in Problem 6. Note that to submit the homework, you'll have to add your plots as figures in your solution PDF (instructions on how to do this are provided below and can be googled as well). You do not need to submit your code.


Problem 1. 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 2. Mean Absolute Error - Adding a Data Point

Suppose we have a dataset $y_1 \leq y_2 \leq \dots \leq y_n$ and want to minimize absolute loss on it for the prediction h . As we've seen before, the corresponding empirical risk is mean absolute error,

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

- a)  Suppose that $R_{\text{abs}}(\alpha) = R_{\text{abs}}(\alpha + 3) = M$, where M is the minimum value of $R_{\text{abs}}(h)$ and α is some constant. Suppose we add to the dataset a new data point y_{n+1} whose value is $\alpha + 1$. For this new larger dataset: What is the minimum of $R_{\text{abs}}(h)$? At what value of h is this minimum achieved?

Your should only involve the variables n, M, α , and constants.

- b)  Let n be odd. 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 3. 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

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

- 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

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

Problem 4. Simple Linear Regression

x	-3	-2	-1	0	1	2	3
y	9	4	1	0	1	4	9

- a) 🥑 For the data above, how does the scatterplot (x on horizontal axis and y on vertical) look like? Hand-drawn is ok.
- b) 🥑🥑 Find the best fitting line using linear regression with x as predictor and y response. What is the intercept and slope?
- c) 🥑🥑 Now, find the best fitting line using linear regression with x^2 as predictor and y response. What is the intercept and slope?
- d) 🥑 Now that we are using x^2 as feature, how does the scatterplot (x^2 on horizontal axis and y on vertical) look like? Hand-drawn is ok.
- e) 🥑 Were the slopes calculated in (b) and (c) different from each other? If yes, please explain the reason for the change.

Problem 5. Formula 1

You wish to establish a linear relationship between the distances driven by two Formula 1 drivers, Driver A and Driver B, during a specific Grand Prix.

The distances are given in the table below:

Driver A (miles)	Driver B (km)
186.4	300
199.5	320
211.3	340
223.7	360
236.1	380
248.5	400
260.9	420
273.4	440

Table 1: Distances driven by Driver A (in miles) and Driver B (in kilometers).

Driver A's distances are measured in miles, and Driver B's distances are measured in kilometers. You are asked to model Driver B's distances as a linear function of Driver A's distances. However, you are also interested in converting Driver A's distances into kilometers to make the relationship clearer.

- a) 🥰🥰🥰🥰 Your friend Zoe suggests that instead of converting all the distances for Driver A from miles to kilometers before performing least squares regression, you could perform the regression with Driver A's distances in miles and then convert the slope and intercept afterward.

Recall that distance in miles can be converted to kilometers using the formula: 1 mile = 1.60 km

Is Zoe correct that you'll get the same regression coefficients either way? Show your work.

- b) 🥰🥰🥰🥰 More generally, suppose we want to perform least squares regression for a linear relationship $y = w_1x + w_0$, where x is Driver A's distance in miles and y is Driver B's distance in kilometers. How do the slope w_1 and intercept w_0 of the regression line change if we replace x with a linear transformation $f(x) = ax + b$, where a and b are constants? Prove your answer by expressing the new slope and intercept in terms of the original slope w_1 , intercept w_0 , and the constants a and b .
- c) 🥰🥰🥰🥰 Prove in general that the mean squared error (MSE) does not change if we use a linear transformation of x as a predictor. Specifically, for an arbitrary data set y_1, \dots, y_n representing distances driven by Formula 1 drivers, show that if $z = ax + b$, then the MSE when using z as a predictor is the same as when using x .

Hint: Use the result from part b) and express the MSE for z in terms of the MSE for x . Show that both are equivalent by simplifying the expressions.





- d) 🥰🥰 Please state if the following statement is True/False. Provide explanation for full points.

The least squares line always passes the center of the data (\bar{x}, \bar{y}) .

Problem 6. Does more data help (for linear regression)?

Here we will generate a dataset and implement simple linear regression.

The question is provided at [this supplementary Jupyter Notebook](#). The code that you write in that notebook will not be graded and you do not need to submit the code on gradescope. You do need to add the plots you generate in the notebook below and answer the two questions. For placing figures please read [this tutorial](#).

- a)  Add the plots you generated for fixed δ and varying n .
- b)  Add the plots you generated for fixed n and varying δ .
- c)  How does adding more points to the dataset improve the fit of the model?
- d)  How does increasing noise impact the model's accuracy?