DSC 40A - Homework 3

Due: Friday, August 16th 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.

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. We encourage you type your solutions in LaTeX, 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 26 points. The point value of each problem or sub-problem is indicated by the number of avocados shown.

- This homework involves some long calculations. You may use a calculator (Python is recommended!), but you may not use any tools that perform regression for you. Show your work by showing the mathematical expression you're evaluating with a calculator, and the numerical result; you don't need to show every intermediate step.
- For Problems 6(b) and 6(c), 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 submit your answers PDF to the Homework 3 assignment on Gradescope, and submit your completed notebook `hw03-code.ipynb` to the Homework 3, Problems 6(b) and 6(c) autograder on Gradescope.

Problem 1. Reflection and Feedback Form

🥑🥑🥑 Make sure to fill out this Reflection and Feedback Form, linked here for three 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. Shout for Stroud

In the National Football League (NFL), the highest paid position is the quarterback. The job of the quarterback on an (American) football team is, among other things, to throw (or "pass") the football to other players, who then score "touchdowns." Each time a quarterback throws the ball to another player and that other player scores, we say the quarterback made a "touchdown pass."

Suppose that we have access to a dataset containing information about a random sample of 50 quarterbacks. For each quarterback, we have the number of touchdown passes they threw, along with their salary in 2023. In the 2023 dataset, the number of touchdown passes for all quarterbacks has a mean of 17 and a standard deviation of 3.

We minimize mean squared error to fit a linear hypothesis function, $H(x) = w_0 + w_1 x$, to this dataset. We will use the hypothesis function to help other players predict their 2023 salary in millions of dollars (y) based on their number of touchdown passes (x).

a) 🥑🥑🥑🥑 CJ Stroud was one of the quarterbacks in our 2023 dataset. Suppose that in 2023, he had 26 touchdown passes and his salary was only 10 million, the smallest salary in our sample.

In 2024, Stroud signed a new contract based on his performance. In 2024, he again threw 26 touchdowns, but his salary shot up to 50 million!

Suppose we create two linear hypothesis functions, one using the dataset from 2023 when Stroud had a salary of 10 million and another using the dataset from 2024 when Stroud had a salary of 50 million. Assume that all other players threw the same amount of touchdowns and had the same salary in both datasets. **That is, only this one data point is different between these two datasets.**

Suppose the optimal slope and intercept fit on the first dataset (2023) are w_1^* and w_0^*, respectively, and the optimal slope and intercept fit on the second dataset (2024) are w_1' and w_0', respectively.

What is the difference between the new slope and the old slope? **That is, what is $w_1' - w_1^*$?** The answer you get should be a number with no variables.

Note: Since we want to salary in millions of dollars, use 10 instead of 10,000,000 for Stroud's salary in 2023.

Hint: There are many equivalent formulas for the slope of the regression line. We recommend using this one for this problem:

$$w_1^* = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

b) 🥑🥑🥑 Let $H^*(x)$ be the linear hypothesis function fit on the 2023 dataset (i.e. $H^*(x) = w_0^* + w_1^* x$) and $H'(x)$ be the linear hypothesis function fit on the 2024 dataset (i.e. $H'(x) = w_0' + w_1' x$).

Consider two other players, Davis Mills and Tom Brady, neither of whom were part of our original sample in 2023. Suppose that in 2021, Mills had 24 touchdowns and Brady had 43 touchdowns.

Both Mills and Brady want to try and use one of our linear hypothesis functions to predict their salary for next year.

Suppose they both first use $H^*(x)$ to determine their predicted yields as per the first rule (when Stroud had a salary of 10 million). Then, they both then use $H'(x)$ to determine predicted yields as per the second rule (when Stroud had a salary of 50 million).

Whose prediction changed more by switching from $H^*(x)$ to $H'(x)$ – Mills' or Brady's?

Hint: You should draw a picture of both prediction rules, $H^*(x)$ and $H'(x)$. You already know how the slopes of these lines differ from part (a). Can you identify a point that each line must go through?

c) 🥑🥑 In this problem, we'll consider how our answer to part (b) might have been different if Stroud had fewer touchdowns in both 2023 and 2024.

- Suppose Stroud instead had 17 touchdowns in both 2023 and 2024. If his salary increased from 2023 to 2024, and everyone else's data stayed the same, which slope would be larger: $H^*(x)$ or $H'(x)$?
- Suppose Stroud instead had 10 touchdowns in both 2023 and 2024. If his salary increased from 2023 to 2024, and everyone else's data stayed the same, which slope would be larger: $H^*(x)$ or $H'(x)$?

You don't have to actually calculate the new slopes, but given the information in the problem and the work you've already done, you should be able to answer the questions and give brief justification.

Problem 3. Correlation Bounds

In both this class and DSC 10, you were told that the correlation coefficient, r, ranges between -1 and 1, where $r = -1$ implies a perfect negative linear association and $r = 1$ implies a perfect positive linear association. However, you were never given a proof of the fact that $-1 \leq r \leq 1$.

Here, you will prove this fact, using linear algebra. Before proceeding, you'll want to review slide 15 onwards in Lecture 6. **Remember to show your work all throughout!**

a) 🥑🥑 Determine the angle between the vectors $\vec{a} = \begin{bmatrix} 5 \\ -7 \\ 14 \end{bmatrix}$ and $\vec{b} = \begin{bmatrix} -13 \\ 2 \\ 9 \end{bmatrix}$. Your answer should involve the function $\cos^{-1}$ (you do not have to find the angle in terms of degrees or radians).

b) 🥑🥑 Let $\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$. We define the "mean-centered" version of $\vec{x}$ to be $\vec{x_c} = \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix}$, where $\bar{x}$ is the mean of the components of $\vec{x}$.

The mean-centered version of $\vec{y}$, named $\vec{y_c}$, is defined similarly. Express $\vec{x_c} \cdot \vec{y_c}$ using summation notation.

c) 🥑🥑🥑 Prove that:

$$r = \frac{\vec{x_c} \cdot \vec{y_c}}{\|\vec{x_c}\| \|\vec{y_c}\|}$$

d) 🥑🥑 Argue why the result in (c) implies that $-1 \leq r \leq 1$.

Hint: If you're completely stuck on how to proceed, try to think about what the purpose of part (a) was — it's in some way related to this part.

Problem 4. Making Connections... and Projections

Suppose we have a dataset of n points, $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. In Lecture 5, we proved that the optimal parameter m^* that minimizes mean squared error for the hypothesis function $H(x) = mx$ is:

$$m^* = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

(There, we used the variable w instead of m; we've used m above to avoid conflicting with a different definition of w below.)

In this problem, we'll derive the same result using our knowledge of vector projections from Lectures 6 and 7, to start making the connections between linear algebra and empirical risk minimization more clear.

Moving forward, consider the dataset of two points, $(4, 1)$ and $(7, 2)$. We can store the x and y coordinates of our two points in vectors, $\vec{x}$ and $\vec{y}$, as follows:

$$\vec{x} = \begin{bmatrix} 4 \\ 1 \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} 7 \\ 2 \end{bmatrix}$$

a) 🥑🥑🥑 Our goal is to find the vector in $\text{span}(\vec{x})$ that is closest to y. The answer is a vector of the form $w\vec{x}$, where $w \in \mathbb{R}$ is some scalar. The w that we choose is one that minimizes the length, $\|\vec{e}\|$ of the error vector, $\vec{e}$:

$$\vec{e} = \vec{y} - w\vec{x}$$

What is w^*, the value w that minimizes $\|\vec{e}\|$? In other words, what value of w minimizes projection error? (Note that the *vector* projection of $\vec{y}$ onto $\text{span}(\vec{x})$ is not w^*, but $w^*\vec{x}$ — however, here we're just asking you for the value of w^*, and of course, to show your work).

b) 🥑🥑 What is the error vector, $\vec{e}$, you found in part (a), and what is its length, $\|\vec{e}\|$?

c) 🥑🥑 The value of w^* you found in part (a) should be equal to the value you find using the formula for m^*. In general, the w^* that minimizes $\|\vec{y} - w\vec{x}\|$ is equal to m^*, the m that minimizes $\frac{1}{n}\sum_{i=1}^{n}(y_i - mx_i)^2$. Explain why this is the case.

Hint: $\|\vec{y} - w\vec{x}\|$ *and* $\frac{1}{n}\sum_{i=1}^{n}(y_i - mx_i)^2$ *are related, but not exactly the same.*

In parts (a) through (c), we projected $\vec{y}$ onto the span of a single vector, $\vec{x}$. But starting in Lecture 7, we'll look at how to project a vector $\vec{y}$ onto the span of two or more vectors. Let's start to explore that concept here.

d) 🥑🥑 Consider the vectors $\vec{x}^{(1)}$ and $\vec{x}^{(2)}$, defined as follows:

$$\vec{x}^{(1)} = \begin{bmatrix} 0 \\ 3 \end{bmatrix}, \quad \vec{x}^{(2)} = \begin{bmatrix} 5 \\ 23 \end{bmatrix}$$

Again, let $\vec{y} = \begin{bmatrix} 7 \\ 2 \end{bmatrix}$.

What is the vector projection of $\vec{y}$ onto $\text{span}(\vec{x}^{(1)}, \vec{x}^{(2)})$ — that is, what vector in $\text{span}(\vec{x}^{(1)}, \vec{x}^{(2)})$ is closest to $\vec{y}$? Give your answer in the form of a vector.

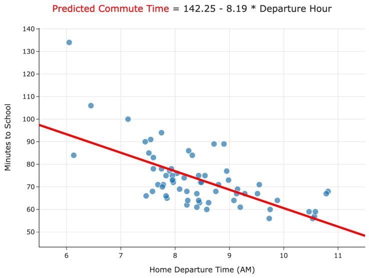
e) 🥑🥑🥑 Let $\vec{h}$ be your answer to the previous part. Find scalars w_1 and w_2 such that:

$$w_1\vec{x}^{(1)} + w_2\vec{x}^{(2)} = \vec{h}$$

Looking ahead: in Lecture 7, we will express $w_1\vec{x}^{(1)} + w_2\vec{x}^{(2)}$ as the matrix-vector product $X\vec{w}$, where $X = \begin{bmatrix} | & | \\ \vec{x}^{(1)} & \vec{x}^{(2)} \\ | & | \end{bmatrix} = \begin{bmatrix} 0 & 5 \\ 3 & 23 \end{bmatrix}$ and $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$. This will allow us to more efficiently solve for the values $w_1, w_2, ..., w_d$ that minimize projection error when we have several spanning vectors, $\vec{x}^{(1)}, \vec{x}^{(2)}, ..., \vec{x}^{(d)}$.

Problem 5. Lagrange Interpolation

So far in this class, our primary tool for making predictions has been simple linear regression — that is, a straight line. However, one issue with linear regression is that a straight line doesn't pass through each data point, leading to prediction errors.



An example simple linear regression line from lecture. While the line fits the trend in the data reasonably well, it doesn't directly pass through many points.

In this problem, we will explore the idea of **polynomial interpolation**, which is a method of constructing a polynomial that passes directly through a given set of points. Interpolation is widely used in numerical analysis, a subfield of mathematics that deals with approximating solutions to equations that (often) don't have solutions that we can solve for by hand, often by writing code.

The specific method for interpolation we'll study in this problem is called Lagrange Interpolation. It solves the following problem:

> Given a set of $n + 1$ points, $(x_1, y_1), (x_2, y_2), ..., (x_{n+1}, y_{n+1})$, what is the equation of the degree n polynomial that passes through all $n + 1$ points?

We've written a detailed guide about how Lagrange Interpolation works, along with a complete example. Click this blue text to see that guide.

a) 🥑🥑🥑 Consider the following dataset of $n + 1 = 3$ points:

$$(1, -3), (-2, 5), (4, 8)$$

The basis polynomial $p_1(x)$, corresponding to the first point we were given, is:

$$p_1(x) = \frac{(x+2)(x-4)}{(1+2)(1-4)} = -\frac{(x+2)(x-4)}{9}$$

Find the other two basis polynomials, $p_2(x)$ and $p_3(x)$.

b) 🥑🥑🥑🥑 Using the basis polynomials in part (a), find the polynomial $p(x)$ of degree $n = 2$ that passes through all three of our points. Make sure to write it in standard form, i.e. in the form $p(x) = a_0 + a_1 x + a_2 x^2$.

c) 🥑🥑 One of the goals of this problem, other than to expose you to the concept of interpolation, is to have you compare interpolation to regression. They both have the same goal of finding a function that *best* passes through the data, but they both use different meanings of "best" and use different approaches.

On a single scatter plot:

- Plot our dataset, $(1, -3), (-2, 5), (4, 8)$.
- Draw the interpolating polynomial $p(x)$ you found in (b).
- Draw $H^*(x) = w_0^* + w_1^*$, the linear hypothesis function that minimizes mean squared error for this dataset. (You don't have to actually calculate w_0^* and w_1^* — you can estimate the approximate location of the line.)

The process of Lagrange Interpolation finds a polynomial $p(x)$ with a mean squared error of 0, which is generally much lower than the mean squared error of $H^*(x)$ on a given dataset (unless the dataset consists of points that all fall on a straight line). Why isn't Lagrange Interpolation used very often in the context of finding hypothesis functions to use for prediction, and why do we prefer empirical risk minimization in general?

Make sure to include a picture of your scatter plot, with all three of the above steps completed, in your PDF.

Problem 6. Least Absolute Deviation Regression

In lecture, we explored least squares regression, and defined it as the problem of finding the values of w_0 (intercept) and w_1 (slope) that minimize mean squared error:

$$R_{\text{sq}}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))^2.$$

Notice that we used the squared loss function, $(y_i - (w_0 + w_1 x_i))^2$ as our metric for deviation. What if we used a different loss function instead?

In this problem, we are going to introduce another type of linear regression: least absolute deviation (LAD) regression. We will define least absolute deviation regression in terms of the absolute loss function rather than the squared loss function to measure how far away our predictions are from the data. That is, we will try to instead minimize

$$R_{\text{abs}}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} |y_i - (w_0 + w_1 x_i)|$$

Since absolute value functions are not differentiable, we cannot just take the gradient of R_{abs}, set it equal to zero, and solve for the values of w_0 and w_1, as we did to minimize R_{sq}. In order to generate the optimal LAD regression line we are going to leverage a very useful theorem:

If you have a dataset with n data points in $\mathbb{R}^k$, where $k \leq n$, then one of the optimal LAD regression lines must pass through k data points.

Notice that unlike with least squares regression, the LAD regression line may not be unique!

This theorem is useful to us because it allows us to adopt a very conceptually simple, albeit not very efficient, strategy to compute an optimal LAD regression line. Since our data will be in $\mathbb{R}^2$, we will generate all possible unique pairs of points and calculate the intercept w_0 and slope w_1 of the line between each pair. Then we'll just select which (w_0, w_1) pair among these finite options has the smallest value of $R_{\text{abs}}(w_0, w_1)$. This is guaranteed by the theorem to be an optimal LAD regression line.

Parts (b) and (c) of this problem will require you to write code in this supplementary Jupyter Notebook. The code that you write in that notebook is autograded, both using public test cases that you can see in the notebook and hidden test cases that will only be run after you submit on Gradescope.

To submit your homework, in addition to submitting your answers PDF to the Homework 3 assignment on Gradescope, also submit `hw03-code.ipynb` to the Homework 3, Problems 6(b) and 6(c) autograder on Gradescope and wait until you see all public test cases pass!

a) 🥑🥑 If you are given n data points, how many pairs of points are there? Give your answer in terms of n.

Hint: Try it out on some small values of n and look for a pattern. Note that if you have two data points (x_1, y_1) and (x_2, y_2), this counts as only one pair of points because the line from (x_1, y_1) and (x_2, y_2) is the same as the line from (x_2, y_2) to (x_1, y_1).

b) 🥑🥑🥑 First, we'll find the intercept and slope of the regular least squares regression line. In the linked supplementary notebook, read the problem statement and complete the implementation of the function `least_squares_regression`.

c) 🥑🥑🥑🥑🥑 Next, we'll find the intercept and slope of the least absolute deviations line. In the linked supplementary notebook, read the problem statement and complete the implementations of the functions `mean_absolute_error` and `find_best_mad_line`.

d) 🥑🥑 Now that we have calculated the least squares regression line and the least absolute deviation regression line for our data, let's try plotting them together to see the difference! In the linked supplementary notebook, generate a scatter plot with the data in black, the least squares line in blue, and the least absolute deviation line in red. **Include a picture of your plot in your PDF; this problem is not autograded.**

e) 🥑🥑 Given your knowledge of the loss functions behind least absolute deviation and least squares regression, provide one advantage and one disadvantage of using LAD over least squares for regression.