
DSC 40A - Homework 3

due Friday, October 24th at 11:59 PM

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. **Only handwritten solutions will be accepted (use of tablets is permitted). Do not typeset your homework (using L^AT_EX or any other software).**

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

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.

Prove or disprove the following statements.

- a) 🥑🥑 Recall: a set of vectors is said to be *linearly independent* if there exists no vector in the set that is equal to a linear combination of the other vectors in the set.

Let $\vec{x}, \vec{y}, \vec{z} \in \mathbb{R}^n$ be given vectors. If $\{\vec{x}, \vec{y}\}$ is linearly independent, and $\{\vec{x}, \vec{z}\}$ is linearly independent, then \vec{x} cannot be written as a linear combination of \vec{y} and \vec{z} .

- b) 🥑 Let $\vec{x} \in \mathbb{R}^n$ and suppose there exists $c \in \mathbb{R}$ with $c \neq 0$ such that $\vec{x}^\top (c\vec{x}) = 0$. Then \vec{x} is the zero vector.

- c) 🥑🥑 Let $A \in \mathbb{R}^{n \times n}$ be any matrix and then let

$$S = \{\vec{x} \in \mathbb{R}^n : \vec{x}^\top A \vec{x} = 0\},$$

i.e. S is the set of all vectors x in \mathbb{R}^n such that $\vec{x}^\top A \vec{x} = 0$. Then S is a subspace of \mathbb{R}^n .

Problem 3. Nullspace to the max

Let $A \in \mathbb{R}^{n \times d}$ be any given matrix, where $n, d \geq 1$ are fixed dimensions. As a reminder, the *nullspace* or *kernel* of a matrix is defined as

$$\text{null}(A) = \{\vec{x} \in \mathbb{R}^d : A\vec{x} = \vec{0}\}.$$

- a) 🥑🥑 Prove that $\text{null}(A) = \text{null}(A^\top A)$ by completing the following steps.
- (i) If $\vec{x} \in \text{null}(A)$, then $\vec{x} \in \text{null}(A^\top A)$.
 - (ii) If $\vec{x} \in \text{null}(A^\top A)$, then $\vec{x} \in \text{null}(A)$.
- b) 🥑🥑 Use the rank-nullity theorem (see the [course notes, Appendix B](#)) and (a) to show that $\text{rank}(A) = \text{rank}(A^\top A)$.
- c) 🥑🥑🥑 Now suppose $\vec{y} \in \mathbb{R}^n$ is a fixed vector, and consider the linear system of equations $A\vec{x} = \vec{y}$ in the unknown $\vec{x} \in \mathbb{R}^d$. Prove that there exists a unique solution \vec{x}^* to this equation if and only if $A^\top A \in \mathbb{R}^{d \times d}$ is an invertible matrix and \vec{y} belongs to the column space of A . (*Note: To prove an “if and only if” statement, there are two steps. See [IF AND ONLY IF PROOFS](#)*)

Problem 4. Making Connections... and Projections

Suppose we have a dataset of n points, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. In Groupwork 3, we proved that the optimal parameter m^* that minimizes mean squared error for the hypothesis function $H(x) = mx$ is:

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

(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 recent lectures (see [lecture 9](#)), to start making the connections between linear algebra and empirical risk minimization more clear.

Moving forward, consider the dataset of two points, $(2, 1)$ and $(3, 2)$. We can store the 2D coordinates of our two points in vectors, \vec{x} and \vec{y} , as follows:

$$\vec{x} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \quad \vec{y} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

- a) 🥑🥑🥑 Our goal is to find the vector in $\text{span}(\vec{x})$ that is closest to \vec{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 in [Lecture 10](#), we looked at how to project a vector \vec{y} onto the span of two or more vectors. Let's explore that concept here.

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

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

$$\text{Let } \vec{y} = \begin{bmatrix} 7 \\ 2 \end{bmatrix}.$$

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:

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

In Lecture 10 we expressed $w_1 \vec{x}^{(1)} + w_2 \vec{x}^{(2)}$ 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 allowed us to more efficiently solve for the values w_1, w_2, \dots, w_d that minimize projection error when we have several spanning vectors, $\vec{x}^{(1)}, \vec{x}^{(2)}, \dots, \vec{x}^{(d)}$. Moving forward we will use this approach for multiple linear regression.

Problem 5. Visualizing Spans and Projections in \mathbb{R}^3

In this problem, you will use the Desmos 3D Graphing Calculator (<https://www.desmos.com/3d>) to visualize how a vector projects onto the span of two other vectors. The exercise will help connect the algebraic definition of a projection with its geometric meaning.

Consider the following three vectors in \mathbb{R}^3 :

$$\vec{v}_1 = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}, \quad \vec{v}_2 = \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix}, \quad \vec{v}_3 = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}.$$

- a) 🥑🥑 Plot \vec{v}_1 and \vec{v}_2 in the Desmos 3D calculator as arrows from the origin. Label each vector and include a screenshot of your plot.
- b) 🥑🥑🥑 Plot a picture of $\text{span}(\{\vec{v}_1, \vec{v}_2\})$. Provide a screenshot of your plot, including the original vectors \vec{v}_1, \vec{v}_2 . *Hint: It may help to find an equation of the form $ax + by + cz = 0$ for $a, b, c \in \mathbb{R}$ which describes the plane when setting up your Desmos plot.*
- c) 🥑🥑 Plot the third vector \vec{v}_3 in the same Desmos diagram. Label it clearly and provide a screenshot.
- d) 🥑🥑🥑🥑 Compute and plot the projection of \vec{v}_3 onto the plane $\text{span}(\vec{v}_1, \vec{v}_2)$. Plot the projection vector in Desmos as an arrow segment from the origin. Provide a screenshot.

Problem 6. Vector-valued constant model

Suppose we fix a collection of training data $\{(\vec{x}_i, \vec{y}_i)\}_{i=1}^n$, where $\vec{x}_i \in \mathbb{R}^d$ are feature vectors and $\vec{y}_i \in \mathbb{R}^d$ are vector-valued targets. Consider the *constant vector hypothesis* that predicts the same vector for every input:

$$H(\vec{x}) = \vec{h}, \quad \vec{h} \in \mathbb{R}^d.$$

This is a similar setting to the constant model we have encountered throughout the past few weeks, except that the function is *vector* valued. In this problem we will identify the *vector* of parameters \vec{h}^* which minimizes the associated MSE.

Note that this problem will serve as a walkthrough of Section 2.2 in [the course notes](#), so if you get stuck, you may find that discussion helpful.

- a) 🥑 Using the squared vector length loss (as we did in Homework 2), write the loss for a single training example (\vec{x}_i, \vec{y}_i) under the constant hypothesis $H(\vec{x}_i) = \vec{h}$.
- b) 🥑🥑🥑 Write down the empirical risk function of the vector \vec{h} over the dataset of n examples. By expanding the vector lengths used in Part (a), show that the risk can be expressed as a nested sum

$$R_{sq}(h) = \frac{1}{n} \sum_{i=1}^n \sum_{t=1}^d (\vec{y}_i^{(t)} - h^{(t)})^2.$$

- c) 🥑🥑🥑 Next we will compute the gradient $\nabla R_{sq}(h)$ of the risk function with respect to \vec{h} . For a *single* coordinate direction $t \in \{1, \dots, d\}$, compute the partial derivative $\frac{\partial}{\partial h^{(t)}} R(\vec{h})$.
- d) 🥑🥑 Stack the coordinate-wise derivatives you found in the previous part to obtain $\nabla R_{sq}(\vec{h})$ in vector form. Show that the gradient takes the form

$$\nabla R_{sq}(\vec{h}) = 2 \left(\vec{h} - \frac{1}{n} \sum_{i=1}^n \vec{y}_i \right)$$

- e) 🥑🥑 Solve $\nabla R(\vec{h}^*) = \vec{0}$ for \vec{h}^* . Write a couple sentences comparing your result to the equivalent version for the scalar-valued constant model we have explored over the past two weeks.

Problem 7. 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.

You will need to include screenshots of your code and plots when you submit your Homework 3 to Gradescope.

- 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. Submit a picture of your plot.

- 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.