DSC 40A - Homework 4

Due: Friday, November 4, 2022 at 11:59pm PDT

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 PT on the due date. You can use a slip day to extend the deadline by 24 hours. Make sure to correctly assign pages to Gradescope when submitting.

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

Note: Problems 2, 3, and 4 refer to a supplemental Jupyter Notebook, which can be found **at this link**.

Problem 1. Vector Calculus Involving Matrices

Let X be a fixed matrix of dimension $m \times n$, and let $\vec{w} \in \mathbb{R}^n$. In this problem, you will show that the gradient of $\vec{w}^T X^T X \vec{w}$ with respect to $\vec{w}$ is given by

$$\frac{d}{d\vec{w}}(\vec{w}^T X^T X \vec{w}) = 2X^T X \vec{w}$$

as we used in Lecture 8.

Let $\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_m$ be the column vectors in $\mathbb{R}^n$ that come from **transposing the rows of** X. For example, if $X = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 3 & 1 \end{bmatrix}$, then $\vec{r}_1 = \begin{bmatrix} 1 \\ 4 \\ 7 \end{bmatrix}$ and $\vec{r}_2 = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$.

a) 🥑🥑🥑🥑 Show that, for arbitrary X and $\vec{w}$, we can write

$$\vec{w}^T X^T X \vec{w} = \sum_{i=1}^{m} (\vec{r}_i^T \vec{w})^2.$$

Hint: First, show that we can write $\vec{w}^T X^T X \vec{w}$ as a dot product of two vectors. Then, try and re-write those vectors in terms of $\vec{r}_1, \vec{r}_2, ..., \vec{r}_m$ and $\vec{w}$.

Now that we have written

$$\vec{w}^T X^T X \vec{w} = \sum_{i=1}^{m} (\vec{r}_i^T \vec{w})^2$$

we can apply the chain rule, along with the result of part (a) above, to conclude that

$$\begin{aligned}\frac{d}{d\vec{w}}(\vec{w}^T X^T X \vec{w}) &= \sum_{i=1}^{m} 2(\vec{r}_i^{\,T}\vec{w})\frac{d}{d\vec{w}}(\vec{r}_i^{\,T}\vec{w}) \\ &= \sum_{i=1}^{m} 2(\vec{r}_i^{\,T}\vec{w})\vec{r}_i\end{aligned}$$

b) 🥑🥑🥑🥑 Next, show that, for arbitrary X and $\vec{w}$, we can write

$$2X^T X\vec{w} = \sum_{i=1}^{m} 2(\vec{r}_i^{\,T}\vec{w})\vec{r}_i$$

Hint 1: Suppose A is a matrix and $\vec{v}$ is a vector, and that the product $A\vec{v}$ is well-defined. How can we write the product $A\vec{v}$ as a sum involving the columns of A and elements of $\vec{v}$?

Hint 2: It is likely that you'll need to use one of your intermediate results from part (a).

Since you've shown that $\frac{d}{d\vec{w}}(\vec{w}^T X^T X\vec{w})$ and $2X^T X\vec{w}$ are both equal to the same expression, $\sum_{i=1}^{m} 2(\vec{r}_i^{\,T}\vec{w})\vec{r}_i$, you have proven that they are equal to one another, i.e. that

$$\frac{d}{d\vec{w}}(\vec{w}^T X^T X\vec{w}) = 2X^T X\vec{w}$$

as desired.

Problem 2. Billy's Back!

This problem is formatted slightly differently. It is entirely contained in the supplemental Jupyter Notebook, which can be found **at this link**. However, you will not submit your code — instead, each subpart tells you what to include in your PDF writeup.

Note that the problem is worth 18 points total, split across 6 parts.

Problem 3. Sums of Residuals

Let's start by recalling the idea of orthogonality from linear algebra. This will allow us to prove a powerful result regarding linear regression, starting in part (b).

Two vectors are **orthogonal** if their dot product is 0, i.e. for $\vec{a}, \vec{b} \in \mathbb{R}^n$:

$$\vec{a}^T\vec{b} = 0 \implies \vec{a}, \vec{b} \text{ are orthogonal}$$

Orthogonality is a generalization of perpendicularity to multiple dimensions. (Two orthogonal vectors in 2D meet at a right angle.)

Suppose we want to represent the fact that some vector $\vec{b}$ is orthogonal to many vectors $\vec{a}_1$, $\vec{a}_2$, ..., $\vec{a}_d$ all at once. It turns out that we can do this by creating a new $n \times d$ matrix A whose columns are the vectors $\vec{a}_1$, $\vec{a}_2$, ..., $\vec{a}_d$, and writing $A^T\vec{b} = 0$.

For instance, suppose $\vec{a}_1 = \begin{bmatrix} 8 \\ 4 \\ -2 \end{bmatrix}$, $\vec{a}_2 = \begin{bmatrix} 3 \\ 5 \\ 1 \end{bmatrix}$, and $\vec{b} = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$. Then,

$$A = \begin{bmatrix} 8 & 3 \\ 4 & 5 \\ -2 & 1 \end{bmatrix} \implies A^T = \begin{bmatrix} 8 & 4 & -2 \\ 3 & 5 & 1 \end{bmatrix}$$

Note that the product $A^T\vec{b}$ involves taking the dot product of each row in A^T with $\vec{b}$. If $A^T\vec{b}$ is a vector of all 0s, i.e. the 0 vector, then it is the case that $\vec{b}$ is orthogonal to each row of A^T, and hence orthogonal to each column of A.

(We will not use this fact in this class, but if $A^T\vec{b} = 0$, it also means that $\vec{b}$ is orthogonal to the **column space** of A.)

a) 🥑🥑 In the example above, verify that $\vec{b}$ is orthogonal to the columns of A.

b) 🥑🥑 Suppose $\vec{1}$ is a vector in $\mathbb{R}^n$ containing the value 1 for each element, i.e. $\vec{1} = \begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix}$.

For any other vector $\vec{b} = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{bmatrix}$, what is the value of $\vec{1}^T\vec{b}$, i.e. what is the dot product of $\vec{1}$ and $\vec{b}$?

Your answer may involve a summation symbol. Explain what it means in words.

Hint: This subpart should not take much time, so let us know if you're stuck on it. Try making up an example $\vec{b}$ and see what $\vec{1}^T\vec{b}$ evaluates to, before generalizing your result to arbitrary $\vec{b}$.

c) 🥑🥑 Now, consider the multiple regression scenario where X is a $n \times (d+1)$ design matrix, $\vec{y} \in \mathbb{R}^n$ is an observation vector, and $w^* \in \mathbb{R}^{(d+1)}$ is the optimal parameter vector.

Show that the error vector, $\vec{y} - X\vec{w}^*$, is orthogonal to the columns of X.

Hint: Again, this should not take very long. Start with the normal equations, $X^TX\vec{w}^* = X^T\vec{y}$, use the distributive property of matrix multiplication, and use what you learned in part (a).

d) 🥑🥑🥑🥑 We define the ith **residual** to be the difference between the actual and predicted values for individual i in our data set. In other words, the ith residual e_i is

$$e_i = y_i - H^*(\vec{x}_i) = (\vec{y} - X\vec{w}^*)_i$$

(Note that $(\vec{y} - X\vec{w}^*)_i$ is referring to element i of the vector $\vec{y} - X\vec{w}^*$. Also, we use the letter e for residuals because residuals are also known as errors.)

Using what you learned in parts (a), (b), and (c), show that the ***residuals of a multiple linear regression prediction rule with an intercept term sums to 0***, i.e. that $\sum_{i=1}^n e_i = 0$.

Note: If you want to see an example of this fact, look at the "Supplement for Problem 3" portion of the supplemental Jupyter Notebook, which can be found **at this link**.

e) 🥑🥑🥑🥑 Now suppose our multiple linear regression prediction rule does not have an intercept term, i.e. that our prediction rule is of the form $H(\vec{x}) = w_1x^{(1)} + w_2x^{(2)} + ... + w_dx^{(d)}$.

1. Is it still guaranteed that $\sum_{i=1}^n e_i = 0$? Why or why not?
2. Is it still possible that $\sum_{i=1}^n e_i = 0$? If you believe the answer is yes, come up with a simple example where a prediction rule without an intercept has residuals that sum to 0. If you believe the answer is no, state why.

Problem 4. Least Absolute Deviation Regression for Multiple Variables

In Lectures 8, 9, and 10, we explored least squares regression and derived the general solution for least squares regression with multiple parameters, also known as the normal equations:

$$X^T X\vec{w}^* = X^T\vec{y},$$

where X is the design matrix, $\vec{y}$ is the observation vector, and $\vec{w}$ is the parameter vector.

In this problem, we are going to try and expand our concept of least absolute deviation (LAD) regression from Homework 3 to accommodate linear prediction rules with two features.

This time, instead of data in $\mathbb{R}^2$, you will be given data in $\mathbb{R}^3$. Now that we have added an extra dimension to our data, we are no longer solving for a regression line, but rather a regression plane of the form $H(\vec{x}) = w_0 + w_1x^{(1)} + w_2x^{(2)}$. In order to use notation that is more convenient and more similar to what you've seen in earlier coursework, let's suppose we're trying to find a regression plane of the form

$$z = H(x, y) = ax + by + c$$

When performing LAD regression, our loss function is absolute loss. That means that we're trying to find the a, b, and c that together minimize mean absolute error,

$$R_{abs}(a, b, c) = \frac{1}{n}\sum_{i=1}^{n} |z_i - (ax_i + by_i + c)|$$

(Here, z_i represents the ith output value.)

We will adopt a strategy similar to Homework 3 to solve for the LAD regression plane. Recall the theorem from last week: If you have a data set with n data points in $\mathbb{R}^k$, where $k \leq n$, then one of the optimal LAD regression prediction rules must pass through k data points.

Since our data will be in $\mathbb{R}^3$, we will generate all possible unique triplets of points and calculate the coefficients a, b, and c that define a plane of the form $z = ax + by + c$ that goes through the three points in our triplet. Then we'll just select which a, b, c triplet among these finite options has the smallest value of $R_{abs}(a, b, c)$. This is guaranteed by the theorem to be an optimal LAD regression plane.

a) 🥑🥑🥑 Before implementing the above strategy in code, let's get familiar with the process by doing an example by hand. Let's say we have three points, $A = \begin{bmatrix} 1 \\ 4 \\ 5 \end{bmatrix}$, $B = \begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix}$, and $C = \begin{bmatrix} 4 \\ -2 \\ 0 \end{bmatrix}$.

Given these three points in $\mathbb{R}^3$, compute the equation of the plane going through them, giving your answer in the form $z = ax + by + c$.

Note: You may need to review some concepts from multivariable calculus (i.e. Math 20C) to do this problem. Here is a nice summary and some examples, obtained from MIT OpenCourseWare under a Creative Commons License.

Hint: After reading the linked notes, a good first step may be to compute $\vec{AB}$, the difference between points A and B, and $\vec{AC}$, the difference between the points A and C, and take their cross product.

b) 🥑🥑🥑🥑🥑 Now, we'll find the LAD regression plane, again for Billy's tip dataset that you worked on in Problem 2.

Recall from the problem description the procedure outlined to generate an optimal LAD regression plane. In the supplemental Jupyter Notebook, found **at this link**, we've already defined several functions for you:

- `generate_all_combinations`, which takes in a dataset and a combination size (in our case, 3) and returns all subsets of the dataset of that size.
- `plane_mae`, which takes in values of a, b, c and a dataset and returns the mean absolute error of the plane $z = ax + by + c$ on that dataset.
- `find_best_plane`, which takes in a list of planes (i.e. a list of (a, b, c) triplets) and a dataset and finds the plane with the lowest mean absolute error. (You implemented this yourself in Homework 2.)

Your job is to complete the implementation of a function, `generate_all_planes`, that generates all planes given a list of the triplets of points and return a list of these planes; the list of planes should be a list of (a, b, c) triplets. Turn in a screenshot of your function as well as the (a, b, c) triplet corresponding to the LAD regression plane for the given data.