DSC 40A - Homework 4

due Friday, October 31th at 11:59 PM

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

Note: Slip days cannot be used for Homework 4! We plan to release the solutions on Saturday after the deadline so you can use them to study for the midterm exam.

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 LaTeXor 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 **65 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. 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

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

Let $\vec{r}_1, \vec{r}_2, \dots, \vec{r}_m$ 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: Use the column-mixing interpretation of matrix-vector multiplication from Lecture 10.

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, $\sum\limits_{i=1}^{m} 2(\vec{r}_i^T \vec{w})\vec{r}_i$, 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 3. Regrssion Equation MCQ

You have a dataset of real features $x_i \in \mathbb{R}$ and observations y_i, and you propose the following linear prediction rule:

$$H_1(x_i, \alpha_0, \alpha_1) = \alpha_0 + \alpha_1 x_i.$$

a) 🎃🎃 Your friend Reggie decides to use $z_i = -\frac{1}{2}x_i$ and the prediction rule

$$H_2(z_i, \beta_0, \beta_1) = \beta_0 + \beta_1 z_i.$$

H_2 achieves the same minimal MSE as H_1 for

○ $\beta_1^* = -\frac{1}{2\alpha_1^*}$

○ $\beta_1^* = -2\alpha_1^*$

○ $\beta_1^* = \frac{1}{2\alpha_1^*}$

○ $\beta_1^* = \frac{2}{\alpha_1^*}$

○ H_2 cannot achieve the same minimum as H_1

Justify your response.

b) 🎃🎃🎃 Your friend Essie proposes to use $v_i = (x_i)^2$ and the prediction rule

$$H_3(v_i, \gamma_0, \gamma_1) = \gamma_0 + \gamma_1 v_i.$$

○ $\gamma_1 = \alpha_1^2$

○ $\gamma_1 = \sqrt{\alpha_1}$

○ $\gamma_1 = -\sqrt{\alpha_1}$

○ The optimal γ_1 depends on the dataset (x_i, y_i)

○ H_3 cannot achieve the same minimum MSE as H_1

Justify your response.

Problem 4. Real Estate

You are given a data set containing information on recently sold houses in San Diego, including

- square footage
- number of bedrooms
- number of bathrooms
- year the house was built
- asking price, or how much the house was originally listed for, before negotiations
- sale price, or how much the house actually sold for, after negotiations

The table below shows the first few rows of the data set. Note that since you don't have the full data set, you cannot answer the questions that follow based on calculations; you must answer conceptually.

House	Square Feet	Bedrooms	Bathrooms	Year	Asking Price	Sale Price
1	1247	3	3	2005	500,000	494,000
2	1670	3	2	1927	1,000,000	985,000
3	716	1	1	1993	335,000	333, 850
4	1600	4	2	1962	830,000	815,000
5	2635	4	3	1993	1,250,000	1,250,000
⋮	⋮	⋮	⋮	⋮	⋮	⋮

a) 🎃🎃🎃 First, suppose we fit a multiple linear regression model to predict the sale price of a house given all five of the other variables. Which feature would you expect to have the largest magnitude weight? Why? (Note that the weight of a feature is the value of w^* for that feature.)

Then, suppose we standardize each variable separately, i.e. we convert each variable to standard units. (Recall, to convert a variable to standard units, we replace each value x_i with $\frac{x_i - \bar{x}}{\sigma_x}$.) Suppose we fit another multiple linear regression model to predict the sale price of a house given all five of the other standardized variables. Now, which feature would you expect to have the largest magnitude weight? Why?

b) 🎃🎃🎃 Suppose we fit a multiple linear regression model to predict the sale price of a house given all five of the other variables in their original, unstandardized form. Suppose the weight for the Year feature is α.

Now, suppose we replace Year with a new feature, Age, which is 0 if the house was built in 2024, 1 if the house was built in 2023, 2 if the house was built in 2022, and so on. If we fit a new multiple linear regression model on all five variables, but using Age instead of Year, what will the weight for the Age feature be, in terms of α?

c) 🎃🎃 Now, suppose we fit a multiple linear regression model to predict the sale price of a house given all five of the other variables, plus a new sixth variable named Rooms, which is the total number of bedrooms and bathrooms in the house. Will our new regression model with an added sixth feature make better predictions than the models we fit in (a) or (b)?

d) 🎃🎃🎃🎃 Now, suppose we fit two multiple linear regression models to predict the sale price of a house. The first uses the features "Square feet" and "Bedrooms":

$$H(\gamma_0, \gamma_1, \gamma_2) = \gamma_0 + \gamma_1 x^{(1)} + \gamma_2 x^{(2)}$$

The second model uses the features “Square feet” and “Bedrooms” and a new sixth feature named “Length of street name”, which is the number of letters in the name of the street that the house is on:

$$H'(\lambda_0, \lambda_1, \lambda_2, \lambda_3) = \lambda_0 + \lambda_1 x^{(1)} + \lambda_2 x^{(2)} + \lambda_3 x^{(6)}$$

Prove that $\text{MSE}(H') \leq \text{MSE}(H)$ always.

Problem 5. Billy the Waiter

🎃🎃🎃🎃🎃🎃🎃🎃🎃🎃🎃🎃🎃🎃

This problem is contained in a supplemental Jupyter Notebook, which you can access **at this link**. Once you've finished, make sure to merge either a pdf copy (the notebook contains some helper links for exporting a Jupyter notebook as a PDF) or screenshots of your work with the remainder of your assignment, and then upload this to the designated parts 5(a)-5(f) in Gradescope.

Note that this problem is worth a total of 14 points, split across 6 parts.

Problem 6. All About That Grade

You are studying the relationship between the number of hours studied and exam scores for a group of students. You collect the following data points:

Hours studied (x)	Exam score (y)
1	50
2	65
3	70
4	85
5	90
6	92
7	95
8	96

a) 🎃 Suppose, you are using polynomial regression of degree 2 to model the relationship between hours studied (x) and exam scores (y). Write down the polynomial equation. You can consider w_0, w_1, and w_2 as the coefficients of the model.

b) 🎃 Define the design matrix.

c) 🎃🎃 Without calculating $X^T X$ explain why it is invertible.

d) 🎃🎃🎃🎃🎃 Calculate the coefficients w_0, w_1, and w_2 using the normal equations (you can use a calculator / code for calculations but show your work by writing out the equations you are solving step by step).

e) 🎃🎃 Calculate the mean squared error of your model.

f) 🎃🎃 Use your polynomial model to predict the exam score for a student who studies for 4.5 hours.

Problem 7. Normal Equations for Simple Linear Regression

You are given n training pairs $\{(x_i, y_i)\}_{i=1}^n$ with $x_i, y_i \in \mathbb{R}$. Consider the simple linear regression (SLR) model with intercept w_0 and slope w_1

$$\widehat{y}_i = w_0 + w_1 x_i, \qquad \vec{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \in \mathbb{R}^2.$$

In this problem you will use the *normal equations* to derive the optimal (w_0^*, w_1^*) that minimize the mean squared error (MSE), and show these match the formulas previously obtained via calculus in week 2.

a) 🎃🎃 The design matrix and vector notation for this model are defined as

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \in \mathbb{R}^{n\times 2}, \quad \vec{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}, \quad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

Write the prediction vector $\widehat{\vec{y}}$ and the empirical risk (average squared loss) $R(w_0, w_1)$ using only terms involving matrices, vectors, and vector lengths.

b) 🎃 Review the lecture slides on multiple regression, and state the *normal equation* in this setting. Your answer should be an equation for the optimal regression parameters $\vec{w}^*$ in terms of X, X^T, and $\vec{y}$. *Note: You don't need to derive the equation from scratch, refer to the lecture notes for the derivation.*

c) 🎃🎃🎃🎃 Compute $X^\top X$ and $X^\top \vec{y}$ explicitly *as matrix–matrix and matrix-vector products of columns* whose elements are expressed in terms of the sums $\sum x_i$, $\sum x_i^2$, $\sum y_i$, and $\sum x_i y_i$. (Hint: write $X = [\vec{1}\ ,\ \vec{x}]$ with $\vec{1} = (1, \dots, 1)^\top$ and $\vec{x} = (x_1, \dots, x_n)^\top$, and calculate the products).

d) 🎃🎃🎃🎃 Using your answers from part (c), compute the matrix $(X^T X)^{-1}$ and plug it into equation for the optimal parameters $\vec{w}^*$ you reviewed from the lecture slides in part (b). Then simplify your formulas for w_0^* and w_1^* using the sample means $\overline{x} = \frac{1}{n}\sum x_i$ and $\overline{y} = \frac{1}{n}\sum y_i$. As shorthand, we suggest you use the notation

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2, \qquad S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

Show that

$$w_1^* = \frac{S_{xy}}{S_{xx}}, \qquad w_0^* = \overline{y} - w_1^*\,\overline{x}.$$

Note: It may help to recall the 2×2 matrix inverse formula: $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$.