DSC 40A - Homework 4

Due: Friday, Feb 16 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. You can use a slip day to extend the deadline by 24 hours.

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

Problem 1. Reflection and Feedback Form

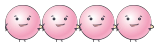
 Make sure to fill out this Reflection and Feedback Form, linked here for two points on this homework! This form is primarily for your benefit; research shows that reflecting and summarizing knowledge helps you understand and remember it.

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 $\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_m$ 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

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

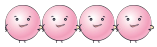
Hint: First, show that we can write $\vec{w}^T X^T X \vec{w}$ as a dot product of two vectors. Then, try and re-write those vectors in terms of $\vec{r}_1, \vec{r}_2, ..., \vec{r}_m$ and $\vec{w}$.

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

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

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 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.

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 $\vec{b}$.

$$A^T\vec{b} = \begin{bmatrix} 8 & 4 & -2 \\ 3 & 5 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix} = \begin{bmatrix} 8(1) + 4(-1) + (-2)(2) \\ 3(1) + 5(-1) + 2(1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Since $A^T\vec{b} = \vec{0}$, 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, which is the space of all linear combinations of the columns of A. As a good exercise in linear algebra, see if you can prove this result!)

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

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

b) Now, consider the typical multiple regression scenario where our prediction rule has an intercept term (w_0):

$$H(\vec{x}) = w_0 + w_1x^{(1)} + w_2x^{(2)} + ... + w_dx^{(d)}.$$

For this scenario, X is a $n \times (d+1)$ design matrix, $\vec{y} \in \mathbb{R}^n$ is an observation vector, and $\vec{w} \in \mathbb{R}^{(d+1)}$ is the parameter vector. We'll use $\vec{w}^*$ to denote the optimal parameter vector, or the one that satisfies the normal equations.

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

Hint: Use the normal equations and the definition of orthogonality to the columns of a matrix given in the problem description.

c) 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

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

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

Using what you learned in parts (a) and (b), show that for multiple linear regression with an intercept term, the sum of the residuals is zero, that is

$$\sum_{i=1}^{n} e_i = 0.$$

d) Now suppose our multiple linear regression prediction rule does not have an intercept term:

$$H(\vec{x}) = w_1x^{(1)} + w_2x^{(2)} + ... + w_dx^{(d)}.$$

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 not.

Problem 4. Multiple Linear Regression and Ridge Regression

a) Under what condition is X^TX invertible? Consider a linear regression model with two predictors and the following dataset. How do you construct the design matrix X (including an intercept term) and response vector Y? How can you calculate the weights w using the formula $w^* = (X^TX)^{-1}X^TY$?

$$X^{(1)} = [1 \quad 1 \quad 2 \quad 2]$$
$$X^{(2)} = [1 \quad 2 \quad 1 \quad 2]$$
$$Y = [1 \quad 1 \quad 2 \quad 2]$$

b)  In scenarios where $X^T X$ is not invertible or when dealing with multicollinearity, ridge regression provides a solution by adding a penalty to the size of the weights. The ridge regression formula is:

$$w^*_{ridge} = (X^T X + \lambda I)^{-1} X^T Y$$

where $\lambda$is the regularization parameter that controls the strength of the penalty, and I is the identity matrix of appropriate size. This approach ensures that $X^T X + \lambda I$ is always invertible, allowing for the estimation of w even in under-determined systems or when $X^T X$ is nearly singular.

Given an under-determined system with three predictors and the following dataset with $\lambda = 1$:

$$X^{(1)} = [1 \quad 2 \quad 3]$$
$$X^{(2)} = [2 \quad 3 \quad 4]$$
$$X^{(3)} = [3 \quad 4 \quad 5]$$
$$Y = [4 \quad 5 \quad 6]$$

How do you construct the design matrix X and response vector Y? How can you calculate the weights w using the given ridge regression formula? Feel free to use a calculator on this one.

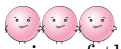
Problem 5. Real Estate

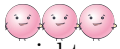
You are given a data set containing information on recently sold houses in San Diego, including

- 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)  Suppose you standardize all six variables and fit a linear prediction rule to predict the sale price of the house based on all five of the other variables. Which feature would you expect to have the largest magnitude weight? Without standardizing, which feature would you expect to have the largest magnitude weight? Explain why.

b)  Suppose you use multiple linear regression on the original (unstandardized) data and the weight associated with Year is α. Suppose you replace Year with a new predictor variable, Age, which is 0 if the house was built in 2023, 1 if the house was built in 2022, 2 if the house was built in 2021, etc. If we do multiple linear regression again using Age instead of Year, what will be the weight associated with Age in terms of α?

c)  Suppose you add a new feature called Rooms, which is the total number of bedrooms and bathrooms in the house. Would multiple linear regression with this extra feature enable you to make better predictions?

Problem 6. Competition: Predicting Energies of Elementary Particles!

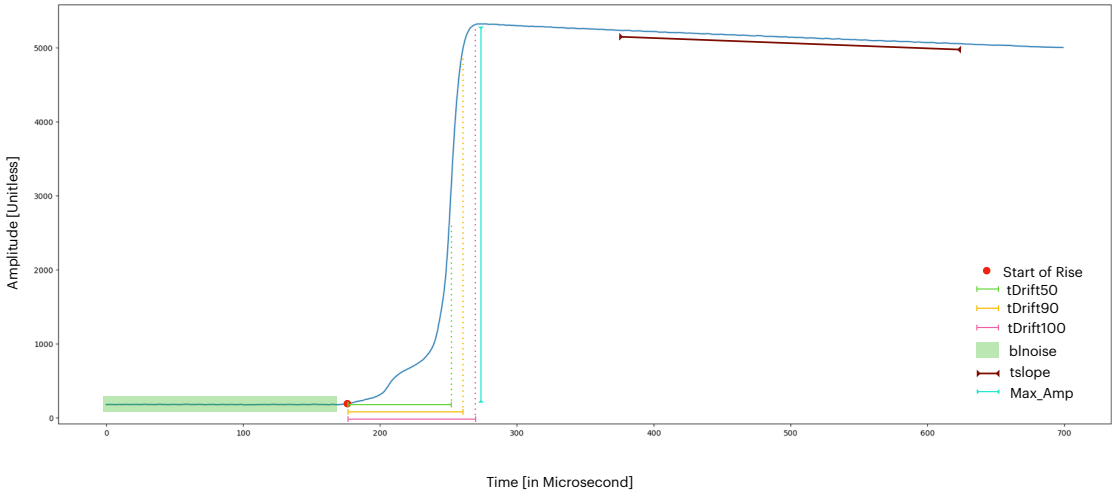
 We are hosting a class-wide competition to see who can make the best predictions! Top predictions can earn extra credit on Midterm 1!

High-Purity Germanium (HPGe) detector is one of the most sensitive detectors human beings have ever manufactured. It is sensitive in a sense that it measures the energy of elementary particles (electron, photon, etc) very accurately. Because of this, HPGe detectors have a wide range of applications, including the search for neutrinos and dark matters, medical imaging, as well as nuclear non-proliferation.

When a particle comes into HPGe detector, it produces a waveform, or time series data, as shown in the picture below. A time series is a sequence of data points that occur in successive order over some period of time. More formally, we can define time series this way: for each data point, a time series contains n pairs of t_i, a_i where t_i is the i^{th} time sample and a_i is the value at i^{th} time sample. To simplify this problem, we will not ask you to directly analyze the time series, but we extract certain features from the HPGe time series for you to build model.

In the supplementary Jupyter notebook (linked), you are given access to a CSV containing training data with information about 400 elementary particles which deposits their energy in a HPGe detector. We read this in as a DataFrame where the columns are different parameters. The columns are:

- **Max_Amp:** Maximum amplitude of the waveform, or the largest number among all a_is
- **tDrift50:** Period from the Start of Rise (t_{SR}) to when the waveform reaches 50% of Max_Amp (t_{50}), can also be written as $t_{50} - t_{SR}$
- **tDrift90:** Period from the Start of Rise (t_{SR}) to when the waveform reaches 90% of Max_Amp (t_{90}), can also be written as $t_{90} - t_{SR}$
- **tDrift100:** Period from the Start of Rise (t_{SR}) to when the waveform reaches Max_Amp, can also be written as $t_{Max_Amp} - t_{SR}$
- **blnoise:** The standard deviation of amplitude values a_i in the green-colored region.
- **tslope:** The slope of the waveform tail.



Your task for this problem is to find the best prediction rule using regression to estimate the energy of each HPGe detector waveforms, given the listed parameters above. The requirements are as follows:

1. You must use regression.
2. The function used for regression is your choice (linear, polynomial, exponential, ...)
3. You may use **up to three variables**. You decide which ones.
4. Your design matrix may have **up to five columns**. You decide what the design matrix looks like.

We've provided you with a function `calculate_MSE` to calculate the mean squared error of your predictions on each waveform in the training data. Your job is to fill in the body of the `predict` function. This function should take as input one row of the DataFrame (corresponding to one particular waveform) and return the predicted energy corresponding to this waveform. How you make this prediction is up to you, subject to the rules above. Feel free to add more cells and functions, and to change the provided `predict` function, but do not change the provided `calculate_MSE` function.

When we grade this question, we will run your prediction function on a hidden test dataset as well, so be mindful of *overfitting* the training data. You'll earn full credit on this homework problem by finding a prediction function whose MSE on the hidden test data is below a certain threshold, which we hope most students will achieve. Additionally, the ten best prediction functions (as determined by the MSE on the hidden test data) will earn some extra credit on the upcoming midterm exam according to the following scheme: for $n \leq 10$, the n^{th} ranked prediction function in the class earns $11 - n$ percentage points as extra credit on Midterm 1.