LUMS
A Not-for-Profit University

School of Science and Engineering

AI 501 Mathematics for Artificial Intelligence

ASSIGNMENT 1 – SOLUTIONS

Due Date: 11 am, Saturday, October 5, 2024.
Format: 11 problems for a total of 100 marks
Instructions:

- You are allowed to collaborate with your peers but copying your colleague's solution is strictly prohibited. This is not a group assignment. Each student must submit his/her own assignment.
- Solve the assignment on blank A4 sheets and staple them before submitting.
- Submit in-class or in the dropbox labeled AI-501 outside the instructor's office.
- **Write your name and roll no. on the first page.**
- Feel free to contact the instructor or the teaching assistants if you have any concerns.

- You represent the most competent individuals in the country, do not let plagiarism come in between your learning. In case any instance of plagiarism is detected, the disciplinary case will be dealt with according to the university's rules and regulations.
- We require you to acknowledge any use or contributions from generative AI tools. Include the following statement to acknowledge the use of AI where applicable.

  I have used [insert Tool Name] to [write, generate, plot or compute; explain specific use of generative AI] [number of times].

Problem 1 (6 marks)

Farmer's Dilemma - Predicting Crop Yield

A farmer in a rural village has been struggling to predict his wheat crop yield every season. He believes that the amount of rainfall and the number of sunny days during the growing season significantly affect his harvest. To better understand this relationship, he collects data over 8 seasons, recording the amount of rainfall (in inches) and the number of sunny days, along with the corresponding wheat yield (in tons).

Season	Rainfall (in inches)	Sunny Days	Wheat Yield (in tons)
1	12	50	6.0
2	15	55	7.0
3	10	48	5.0
4	18	60	8.0
5	20	65	9.0
6	14	53	6.5
7	16	57	7.5
8	17	62	8.2

() Calculate the Pearson correlation coefficient between the amount of rainfall and wheat yield. [5 marks]

Solution: The Pearson correlation coefficient is calculated using the formula:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Where:

- X_i are the rainfall values,
- Y_i are the wheat yield values,
- $\bar{X}$ and $\bar{Y}$ are the means of the rainfall and wheat yield values, respectively.

Using the given data for rainfall and wheat yield:

$$X = [12, 15, 10, 18, 20, 14, 16, 17]$$
$$Y = [6.0, 7.0, 5.0, 8.0, 9.0, 6.5, 7.5, 8.2]$$

Calculate the means of X and Y:

$$\bar{X} = \frac{12 + 15 + 10 + 18 + 20 + 14 + 16 + 17}{8} = 15.25$$
$$\bar{Y} = \frac{6.0 + 7.0 + 5.0 + 8.0 + 9.0 + 6.5 + 7.5 + 8.2}{8} = 7.15$$

Using the Pearson formula, compute the correlation coefficient r. After performing the calculations, you will find:

$$r_{\text{rainfall, yield}} = 0.953$$

Therefore, the Pearson correlation coefficient between rainfall and wheat yield is approximately 0.953, indicating a strong positive relationship.

() Calculate the Pearson correlation coefficient between the number of sunny days and wheat yield. Based on your findings, which factor seems to have a stronger linear relationship with wheat yield? What recommendations would you give the farmer? [5 marks]

Solution: Similarly, calculate the Pearson correlation coefficient between sunny days and wheat yield using the same formula.
Using the given data for sunny days and wheat yield:

$$X = [50, 55, 48, 60, 65, 53, 57, 62]$$

$$Y = [6.0, 7.0, 5.0, 8.0, 9.0, 6.5, 7.5, 8.2]$$

Calculate the mean of sunny days:

$$\bar{X} = \frac{50 + 55 + 48 + 60 + 65 + 53 + 57 + 62}{8} = 56.25$$

Using the Pearson formula, the correlation coefficient is calculated to be:

$$r_{\text{sunny days, yield}} = 0.979$$

Therefore, the Pearson correlation coefficient between sunny days and wheat yield is approximately 0.979, indicating a very strong positive relationship.

Conclusion: Both rainfall and sunny days have a strong positive linear relationship with wheat yield, but the number of sunny days has a slightly stronger correlation. Based on these results, the farmer should focus more on tracking sunny days when predicting wheat yield but should not neglect the importance of adequate rainfall.

Problem 2 (8 marks)

Vector Space and Gram-Schmidt Orthogonalization in Engineering Stability

Imagine you are part of an engineering team working on stabilizing the foundation of a futuristic building. The structural integrity of the building depends on three support columns represented by vectors in $\mathbb{R}^3$. These columns are initially described by the vectors:

$$\mathbf{v_1} = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}, \quad \mathbf{v_2} = \begin{pmatrix} 4 \\ -2 \\ 2 \end{pmatrix}, \quad \mathbf{v_3} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

However, upon inspection, you realize that some of the columns may be providing redundant support, leading to inefficiencies and potential instability.

(a) **(3 points)** Prove that the vectors $\mathbf{v_1}$ and $\mathbf{v_2}$ are linearly dependent. Explain the significance of this linear dependence in the context of the structural stability of the building.

(b) **(5 points)** Apply the Gram-Schmidt orthogonalization process to the set $\{\mathbf{v_1}, \mathbf{v_2}, \mathbf{v_3}\}$. Discuss how the process helps to break the linear dependence and provide new, orthogonal directions that improve the structural support for the building.

Solution:

(a) Linear Dependence Proof:

The vectors $\mathbf{v_1} = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}$ and $\mathbf{v_2} = \begin{pmatrix} 4 \\ -2 \\ 2 \end{pmatrix}$ are linearly dependent if there exists a scalar c such that $\mathbf{v_2} = c\mathbf{v_1}$.

We can check this by comparing the components of $\mathbf{v_1}$ and $\mathbf{v_2}$:

$$\mathbf{v_2} = 2\mathbf{v_1}.$$

This shows that $\mathbf{v_2}$ is indeed a scalar multiple of $\mathbf{v_1}$, proving linear dependence.

Significance: The linear dependence between $\mathbf{v_1}$ and $\mathbf{v_2}$ means that they lie along the same line in $\mathbb{R}^3$. This leads to redundancy in the structural support since both vectors provide support in the same direction. As a result, the foundation lacks stability in other directions, which is crucial for maintaining the overall balance and strength of the structure.

(b) Gram-Schmidt Process:

Now, we apply the Gram-Schmidt process to the set of vectors $\{\mathbf{v_1}, \mathbf{v_2}, \mathbf{v_3}\}$.

Step 1: Set $\mathbf{u_1} = \mathbf{v_1} = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}$.

Step 2: Project $\mathbf{v_2}$ onto $\mathbf{u_1}$:

$$\text{proj}_{\mathbf{u_1}} \mathbf{v_2} = \frac{\mathbf{v_2} \cdot \mathbf{u_1}}{\mathbf{u_1} \cdot \mathbf{u_1}} \mathbf{u_1} = \mathbf{v_2}.$$

Therefore, $\mathbf{u_2} = \mathbf{v_2} - \text{proj}_{\mathbf{u_1}} \mathbf{v_2} = \mathbf{0}$, confirming that $\mathbf{v_2}$ provides no additional information as it is redundant.

Step 3: For $\mathbf{v_3} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, compute the projection onto $\mathbf{u_1}$:

$$\text{proj}_{\mathbf{u_1}} \mathbf{v_3} = \frac{\mathbf{v_3} \cdot \mathbf{u_1}}{\mathbf{u_1} \cdot \mathbf{u_1}} \mathbf{u_1} = \frac{2}{6} \mathbf{u_1} = \begin{pmatrix} \frac{2}{3} \\ -\frac{1}{3} \\ \frac{1}{3} \end{pmatrix}.$$

Now, subtract this projection from $\mathbf{v_3}$ to get $\mathbf{u_3}$:

$$\mathbf{u_3} = \mathbf{v_3} - \text{proj}_{\mathbf{u_1}} \mathbf{v_3} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} - \begin{pmatrix} \frac{2}{3} \\ -\frac{1}{3} \\ \frac{1}{3} \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ -\frac{1}{3} \end{pmatrix}.$$

Breaking the Linear Dependence:

Before applying the Gram-Schmidt process, $\mathbf{v_1}$ and $\mathbf{v_2}$ were linearly dependent, collapsing the subspace. By applying Gram-Schmidt, we remove this redundancy, transforming the set of vectors into an orthogonal set. This introduces new, independent directions that span the full subspace and provide more stable structural support for the building. The orthogonal vectors $\mathbf{u_1}$ and $\mathbf{u_3}$ form the basis for the improved foundation.

Problem 3 (10 marks)

The Mystery of the Hidden Dimensions

In the ancient kingdom of AI501, the wise King had three legendary artifacts: the Scepter of Independence, the Crown of Basis, and the Shield of Rank. These powerful artifacts were said to protect the kingdom from chaos by maintaining the balance of the royal Vector Space, V. One day, a sinister sorcerer, known as Ibrahim, stole the artifacts and scattered them across different subspaces, threatening the very structure of the kingdom.

Desperate to restore order, the King called upon his most trusted advisors, the students of AI501, to retrieve the artifacts by solving a series of puzzles. The students began by studying the kingdom's primary defense: a mysterious matrix A, which governs the stability of the entire Vector Space. The matrix is of size 4×4, and its columns represent different vectors that describe the forces protecting the kingdom. The matrix A is given by

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 8 \\ 1 & 0 & 1 & 0 \\ 3 & 6 & 9 & 12 \end{bmatrix}$$

The puzzles were as follows:

- The **Scepter of Independence** can only be retrieved if the students identify whether the columns of matrix A are linearly independent. Can the students prove the linear independence of the columns of A? If not, what is the dimension of the subspace they span? [5 marks]
- The **Crown of Basis** was hidden in a secret subspace of the Vector Space V. The dimension of this subspace was equal to the rank of matrix A. The students must determine the rank of matrix A to find the exact location of the subspace. What is the rank of matrix A? [2.5 marks]
- Finally, to retrieve the **Shield of Rank**, the students need to discover a basis for the column space of matrix A. How can the students determine the basis for this subspace using the columns of matrix A? [2.5 marks]

The King depends on the students to solve these puzzles, for, without the artifacts, the kingdom of Mathematics of AI will crumble into chaos. Will the students be able to determine the linear independence, rank, and basis of matrix A in time to save the kingdom?

Solution:

1. Linear Independence and Dimension of the Subspace:

To determine whether the columns of matrix A are linearly independent, we need to find the rank of A. We will do this by row-reducing A to its echelon form.

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 8 \\ 1 & 0 & 1 & 0 \\ 3 & 6 & 9 & 12 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 0 \\ 0 & -2 & -2 & -4 \\ 0 & 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

From the reduced form, we see that there are two non-zero rows, indicating that the rank of A is 2. Since the rank is less than the number of columns, the columns of A are not linearly independent. The dimension of the subspace they span is equal to the rank, which is 2.

2. Rank of Matrix A:

As determined from the row-reduction, the rank of matrix A is 2.

3. Basis for the Column Space of Matrix A:

The basis for the column space can be found by selecting the columns of A corresponding to the pivots in the row-reduced form. The pivots are in the first and second columns.

Therefore, the basis for the column space of A consists of the first and second columns of A:

$$\left\{ \begin{bmatrix} 1 \\ 2 \\ 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \\ 0 \\ 6 \end{bmatrix} \right\}$$

Problem 4 (8 marks)

Story Similarity and Distance Measures

Suppose you are analyzing stories, each represented by a vector with 3 elements describing the following characteristics: drama, comedy, and science fiction. You want to compare pairs of stories using these dimensions, and are given the following examples:

$$v_{\text{Story A}} = \begin{bmatrix} 0.9 \\ 0.05 \\ -0.1 \end{bmatrix}, \quad v_{\text{Story B}} = \begin{bmatrix} 0.8 \\ 0.02 \\ 0.78 \end{bmatrix}, \quad v_{\text{Story C}} = \begin{bmatrix} 0.6 \\ 0.85 \\ 0.89 \end{bmatrix}$$

(a) **(2 points)** Compute the Chebyshev distance between Story A and Story B. What short-coming might arise from using this distance for this comparison?

(b) **(2 points)** Compute the Manhattan distance between Story A and Story B. Discuss why Manhattan distance may not be a good metric for this comparison.

(c) **(2 points)** Compute the Euclidean distance between Story A and Story B. Why might this not be the most appropriate metric in this case?

(d) **(2 points)** Using Cosine Similarity, determine which of Story B or Story C is closer to Story A. What feature/characteristic would you attribute this closeness to? Be sure to show your working.

Solution:

1. Chebyshev Distance:

To compute the Chebyshev distance between Story A and Story B, we take the maximum absolute difference across corresponding elements of the vectors.

$$d_{\text{Chebyshev}}(v_{\text{Story A}}, v_{\text{Story B}}) = \max(|0.9 - 0.8|, |0.05 - 0.02|, |-0.1 - 0.78|)$$

Shortcoming: The Chebyshev distance only considers the largest difference between two elements, ignoring smaller differences that may still be significant in the comparison.

2. Manhattan Distance:

The Manhattan distance is the sum of the absolute differences across corresponding elements of the vectors.

$$d_{\text{Manhattan}}(v_{\text{Story A}}, v_{\text{Story B}}) = |0.9 - 0.8| + |0.05 - 0.02| + |-0.1 - 0.78|$$

Why it's not ideal: The Manhattan distance gives equal weight to all dimensions, which may not always reflect the importance of each characteristic in a story.

3. Euclidean Distance:

The Euclidean distance is the square root of the sum of the squared differences across corresponding elements of the vectors.

$$d_{\text{Euclidean}}(v_{\text{Story A}}, v_{\text{Story B}}) = \sqrt{(0.9 - 0.8)^2 + (0.05 - 0.02)^2 + (-0.1 - 0.78)^2}$$

Why it's not ideal: In this context, the Euclidean distance might overemphasize small differences in some dimensions.

4. Cosine Similarity:

The cosine similarity measures the cosine of the angle between two vectors. We first compute the cosine similarity between Story A and Story B, and then between Story A and Story C.

The formula for cosine similarity is:

$$\cos(\theta) = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|}$$

First, we compute the dot product between Story A and Story B:

$$v_{\text{Story A}} \cdot v_{\text{Story B}} = (0.9 \times 0.8) + (0.05 \times 0.02) + (-0.1 \times 0.78)$$

$$v_{\text{Story A}} \cdot v_{\text{Story B}} = 0.72 + 0.001 + -0.078 = 0.643$$

Now, we compute the magnitudes of the vectors:

$$\|v_{\text{Story A}}\| = \sqrt{(0.9)^2 + (0.05)^2 + (-0.1)^2} = \sqrt{0.81 + 0.0025 + 0.01} = \sqrt{0.8225} \approx 0.907$$

$$\|v_{\text{Story B}}\| = \sqrt{(0.8)^2 + (0.02)^2 + (0.78)^2} = \sqrt{0.64 + 0.0004 + 0.6084} = \sqrt{1.2488} \approx 1.1175$$

Now, we compute the cosine similarity:

$$\cos(\theta_{\text{A,B}}) = \frac{0.643}{0.907 \times 1.117} = \frac{0.643}{1.013} \approx 0.6344$$

Next, we compute the cosine similarity between Story A and Story C.

$$v_{\text{Story A}} \cdot v_{\text{Story C}} = (0.9 \times 0.6) + (0.05 \times 0.85) + (-0.1 \times 0.89)$$

$$v_{\text{Story A}} \cdot v_{\text{Story C}} = 0.54 + 0.0425 + -0.089 = 0.4935$$

Now, we compute the magnitude of Story C:

$$\|v_{\text{Story C}}\| = \sqrt{(0.6)^2 + (0.85)^2 + (0.89)^2} = \sqrt{0.36 + 0.7225 + 0.7921} = \sqrt{1.8746} \approx 1.369$$

Now, we compute the cosine similarity:

$$\cos(\theta_{\text{A,C}}) = \frac{0.4935}{0.907 \times 1.369} = \frac{0.4935}{1.241} \approx 0.3973$$

Conclusion: Since the cosine similarity between Story A and Story B (0.6344) is higher than that between Story A and Story C (0.3973), Story B is closer to Story A in terms of cosine similarity. This closeness can be attributed to the similar magnitudes of the characteristics across both stories, especially in drama and comedy.

Problem 5 (10 marks)

Matrix Inverses: A Rigorous Investigation

Let A be an $m \times n$ matrix, and let A have both a left inverse L and a right inverse R. Prove the following statements:

(a) **(3 points)** Show that A must be a square matrix ($m = n$). Specifically, prove that if A has a left inverse L and a right inverse R, then m must equal n.

(b) **(3 points)** Prove that the left inverse L and the right inverse R of A are equal. In other words, show that $L = R$ if both L and R exist.

(c) **(4 points)** Consider a matrix A where $A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix}$.

(i) Compute the left and right inverses of A if they exist.

(ii) Use your results to demonstrate that A does not satisfy the conditions needed for having both a left and right inverse.

(iii) Explain why the specific matrix A you worked with in part (i) does or does not fit the general conditions for having left and right inverses.

Solution:

(a) Proof for Square Matrix Condition:

To show $m = n$ if A has both a left inverse L and a right inverse R:

1. If L is a left inverse, then $LA = I_n$, indicating A must have full column rank ($\text{rank}(A) = n$).

2. If R is a right inverse, then $AR = I_m$, indicating A must have full row rank ($\text{rank}(A) = m$).

3. Since A must have full column and row ranks, $\text{rank}(A) = m = n$. Therefore, A must be a square matrix.

(b) Proof that $L = R$:

Given $LA = I_n$ and $AR = I_n$:

$$L(AR) = LI_n \implies (LA)R = L \implies I_nR = L \implies R = L.$$

Thus, L and R must be equal.

(c) Applied Part:

Consider $A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix}$:

(i) **Compute the Left and Right Inverses of A:**

Since A is a 3×3 matrix, first check its determinant:

$$\det(A) = \begin{vmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{vmatrix} = 0.$$

Since $\det(A) = 0$, A is not invertible, so it does not have a true left or right inverse.

(ii) **Demonstrate Non-Existence of Both Inverses:**

As A is not invertible, it cannot have a right inverse. Additionally, a matrix with a zero determinant cannot have a left inverse. Hence, A does not satisfy the conditions for having both inverses.

(iii) **Explain Why A Does Not Fit General Conditions:**

The matrix A does not fit the general conditions because: - It is not full rank; the rows are linearly dependent (each row is a scalar multiple of the others). - The matrix A is singular, meaning it cannot have a left or right inverse.

Problem 6 (10 marks)

The Adventure of the Regression Model

In a small town, there is a local bakery that wants to predict its daily sales based on the number of customers visiting the store. The bakery owner, Ms. Baker, has collected data from the past month, recording the number of customers and the corresponding daily sales in dollars. Ms. Baker is keen to understand the relationship between the number of customers and the sales using a univariate linear regression model. She needs your help to apply this model to her data and evaluate its performance.

(a) **(5 points) Conceptual Understanding:**
Ms. Baker wants to understand the concept of residuals and how they reflect the quality of the model. Describe what residuals are, how they are calculated, and why minimizing the residual sum of squares is important in fitting the model. Provide an intuitive explanation of how the residuals impact the fit of the regression line.

(b) **(5 points) Practical Application:**
Ms. Baker provides you with the following dataset:

Number of Customers(x)	Daily Sales(y)
10	120
15	180
20	240
25	300
30	360

Using the data, calculate the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the linear regression model $y = \beta_0 + \beta_1 x$. The equations you need are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Show all steps involved in your calculations.

Compute the residual sum of squares (RSS) for the model you obtained. Evaluate the goodness of fit by calculating the total sum of squares (TSS) and the coefficient of determination (R^2). Show all your calculations and explain what the R^2 value indicates about the model's performance.

Solution:

(a) Conceptual Understanding:

Residuals are the differences between the observed values and the predicted values from the regression model:

$$\text{Residual}_i = y_i - \hat{y}_i$$

Minimizing the residual sum of squares (RSS) ensures the best fit of the regression line to the data by reducing the discrepancy between observed and predicted values. Smaller residuals indicate a better fit of the model. Residuals are crucial for understanding how well the model captures the relationship between the variables. If the residuals are small and randomly distributed, it suggests that the model is a good fit for the data.

(b) Practical Application:

Given data:

x	y
10	120
15	180
20	240
25	300
30	360

Calculate $\bar{x}$ and $\bar{y}$:

$$\bar{x} = \frac{10 + 15 + 20 + 25 + 30}{5} = 20$$
$$\bar{y} = \frac{120 + 180 + 240 + 300 + 360}{5} = 240$$

Compute $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{5}(x_i - \bar{x})^2}$$

$$\sum_{i=1}^{5}(x_i-\bar{x})(y_i-\bar{y}) = (10-20)(120-240)+(15-20)(180-240)+(20-20)(240-240)+(25-20)(300-240)+(30-20)(360-240)$$
$$= (-10 \cdot -120) + (-5 \cdot -60) + (0 \cdot 0) + (5 \cdot 60) + (10 \cdot 120)$$
$$= 1200 + 300 + 0 + 300 + 1200 = 3000$$
$$\sum_{i=1}^{5}(x_i - \bar{x})^2 = (10-20)^2 + (15-20)^2 + (20-20)^2 + (25-20)^2 + (30-20)^2$$
$$= 100 + 25 + 0 + 25 + 100 = 250$$
$$\hat{\beta}_1 = \frac{3000}{250} = 12$$

Compute $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 240 - 12 \cdot 20 = 240 - 240 = 0$$

Compute Residual Sum of Squares (RSS):

$$\text{RSS} = \sum_{i=1}^{5}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$
$$= (120 - 0 - 12 \cdot 10)^2 + (180 - 0 - 12 \cdot 15)^2 + (240 - 0 - 12 \cdot 20)^2 + (300 - 0 - 12 \cdot 25)^2 + (360 - 0 - 12 \cdot 30)^2$$
$$= (120 - 120)^2 + (180 - 180)^2 + (240 - 240)^2 + (300 - 300)^2 + (360 - 360)^2$$
$$= 0^2 + 0^2 + 0^2 + 0^2 + 0^2 = 0$$

Compute Total Sum of Squares (TSS):

$$\text{TSS} = \sum_{i=1}^{5}(y_i - \bar{y})^2$$
$$= (120 - 240)^2 + (180 - 240)^2 + (240 - 240)^2 + (300 - 240)^2 + (360 - 240)^2$$
$$= (-120)^2 + (-60)^2 + 0^2 + 60^2 + 120^2$$
$$= 14400 + 3600 + 0 + 3600 + 14400 = 32400$$

Compute R^2:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{0}{32400} = 1$$

Explanation:

The R^2 value of 1 indicates that the model perfectly fits the data, meaning that the regression line explains 100% of the variance in the daily sales based on the number of customers.

Problem 7 (10 marks)

Modified Ridge Regression Using Projection

Given a dataset with N observations and d features, with A being the design matrix and y the response vector, the modified ridge regression cost function is:

$$L(w) = \frac{b_m}{2}\|Aw - y\|^2 + \frac{\lambda b_r}{2}\|w\|^2,$$

where b_m is the weight for the mean squared error term, b_r is the weight for the regularization term, and λ is the regularization parameter.

(a) **(6 points)** Derive the closed-form solution for the weight vector w in this modified ridge regression problem using the concept of projection. Also, specify the shape of each matrix involved in the derivation and the final shape of the weight vector w.

(b) **(4 points)** Discuss how varying b_m and b_r independently affects the bias-variance trade-off in the model.

Problem 8 (12 marks)

The Dilemma of the Automated Drone

You are developing an automated drone system to monitor a secure area for potential intrusions. The drone uses a machine learning model to classify detected movements as either "Intruder" or "Non-Intruder." The performance of your model is evaluated using the following confusion matrix:

	Actual Intruder (Positive)	Actual Non-Intruder (Negative)
Predicted Intruder	a_{11}	a_{12}
Predicted Non-Intruder	a_{21}	a_{22}

(a) **(3 points)** Suppose you are concerned about missing actual intruders (i.e., failing to detect real threats). Which evaluation metric would you prioritize to minimize this issue and why? Explain the importance of this metric in the context of your application.

(b) **(3 points)** If you had to choose between maximizing precision or maximizing specificity, which would you choose for this drone system and why? Provide a detailed explanation of how each metric impacts the performance of the system in this scenario.

(c) **(6 points)** Given the confusion matrix representation above, derive expressions for the following metrics:

- Accuracy
- Precision
- Recall
- Specificity

Additionally, answer the following:

(i) Given that the model never predicts the negative class, and the Accuracy is equal to the F1 score, what should be the value of a_{12}?

Solution:

(a) Metric Selection:

To minimize the issue of missing actual intruders (i.e., failing to detect real threats), you should prioritize **Recall** (Sensitivity). Recall measures the proportion of actual positives (intruders) that are correctly identified by the model. High recall ensures that most actual intruders are detected, which is crucial for reducing the number of false negatives (missed intruders).

(b) Precision vs. Specificity:

For this drone system, you should prioritize **Specificity**. Specificity measures the proportion of actual negatives (non-intruders) that are correctly identified. High specificity ensures that non-intruders are correctly classified, avoiding false positives (incorrectly identifying non-intruders as intruders). This is important to prevent unnecessary alarms and ensure the reliability of the system.

(c) Confusion Matrix Derivations:

- **Accuracy:**

$$\text{Accuracy} = \frac{a_{11} + a_{22}}{a_{11} + a_{12} + a_{21} + a_{22}}$$

- **Precision:**

$$\text{Precision} = \frac{a_{11}}{a_{11} + a_{12}}$$

- **Recall (Sensitivity):**

$$\text{Recall} = \frac{a_{11}}{a_{11} + a_{21}}$$

- **Specificity:**

$$\text{Specificity} = \frac{a_{22}}{a_{22} + a_{12}}$$

Additional Questions:

(i) Given that the model never predicts the negative class, this implies $a_{22} = 0$ and $a_{21} = 0$. The F1 score is given by:

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where

$$\text{Precision} = \frac{a_{11}}{a_{11} + a_{12}}$$

and

$$\text{Recall} = \frac{a_{11}}{a_{11} + a_{21}} = \frac{a_{11}}{a_{11}} = 1$$

Since the accuracy is equal to the F1 score:

$$\text{Accuracy} = \frac{a_{11} + a_{22}}{a_{11} + a_{12} + a_{21} + a_{22}} = \frac{a_{11}}{a_{11} + a_{12}}$$

The F1 score simplifies to:

$$\text{F1 Score} = \frac{2 \times 1 \times \frac{a_{11}}{a_{11}+a_{12}}}{1 + \frac{a_{11}}{a_{11}+a_{12}}} = \frac{2a_{11}}{a_{11} + a_{12} + a_{11}} = \frac{2a_{11}}{2a_{11} + a_{12}}$$

Setting the accuracy equal to the F1 score:

$$\frac{a_{11}}{a_{11} + a_{12}} = \frac{2a_{11}}{2a_{11} + a_{12}}$$

Cross-multiplying, solving and simplifying, we get:

$$a_{11} + a_{12} = a_{11}$$

Therefore:

$$a_{12} = 0$$

Since $a_{12} = 0$, it confirms that $a_{12} = 0$.

Problem 9 (10 marks)

The Tale of the Expanding Universe

In a distant mathematical universe, a group of interdimensional explorers discover a peculiar phenomenon: as they traverse from one dimension to another, their perception of distance and volume within a *unit ball* dramatically shifts.

Hint: The volume V_n of a ball of radius R in n-dimensional space is given by:

$$V_n = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)} R^n,$$

where $\Gamma(x)$ is the Gamma function, and for positive integers x, $\Gamma(x) = (x-1)!$.

(a) **(4 points)** Upon entering the universe, they encounter a spherical planet which has a unit radius in their familiar 3D world. They observe that the volume of the planet (which corresponds to the volume of a unit ball in 3 dimensions) is approximately 4.19 cubic units. The explorers, intrigued by how this might change in higher dimensions, decide to explore unit balls in various dimensions n. Use the given formula for V_n to calculate the volume of the unit ball in $n = 10$ and $n = 100$ dimensions.

(b) **(3 points)** The explorers then notice something strange: as they increase the dimensionality beyond 10, the majority of the volume seems to *disappear* from the center of the unit ball and shifts to its surface. Explain this phenomenon mathematically, and show how the proportion of the volume near the center diminishes as n increases.
Hint: Analyse the fraction of the volume of unit radius ball that lies inside a shell between unit radius ball and a ball of radius $0 < r_0 \leq 1$, and show that for most of the volume lies inside the shell even when $r_0 \to 1$.

(c) **(3 points)** Realizing that this loss of volume at the center could have serious implications for navigation, the explorers decide to quantify their findings. They calculate the distance from the origin at which half the volume of the unit ball resides. Find and explain how this distance changes as the number of dimensions increases, particularly comparing $n = 3$, $n = 10$, and $n = 100$.

Solution:

(a) Deriving the Volume of a Unit Ball:

The volume V_n of a unit ball in n-dimensional space is given by:

$$V_n = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)},$$

where $\Gamma(x)$ is the Gamma function. Using this formula:

For $n = 3$:

$$V_3 = \frac{4\pi}{3} \approx 4.18879 \text{ cubic units.}$$

For $n = 10$:

$$V_{10} = \frac{\pi^5}{5!} \approx 2.5502 \text{ units.}$$

For $n = 100$:

$$V_{100} \approx 1.704 \times 10^{-40} \text{ units.}$$

(b) Volume Concentration Near the Surface:

To show that the volume is concentrated near the surface for a unit n-ball, let's use the differential proportionality relationship:

$$dV_n \propto r^{n-1} dr.$$

Let $0 < r_0 \leq 1$ be some radius over which we will integrate and find the volume. Then, the Volume V_{n_0} is given by:

$$V_{n_0} \propto \int_0^{r_0} r^{n-1} dr,$$

$$V_{n_0} \propto \frac{r_0^n}{n}.$$

This shows that the volume away from the surface the center is proportional to $\frac{r_0^n}{n}$. But, as we increase n, this value becomes smaller and smaller, as $r_0 \leq 1$, which implies most of the volume must lie near the surface.

(c) Distance for Half the Volume:

To find the radius r_{half} where half the volume resides, we use the same proportionality condition as in part (b):

$$V_{n_{half}} \propto \frac{r_{half}^n}{n}.$$

We also know that:

$$V_n \propto \frac{1}{n},$$

and:

$$V_n = 2V_{n_{half}},$$

which gives us the following relationship:

$$\frac{2r_{half}^n}{n} = \frac{1}{n}.$$

Solving for r_{half}:

$$r_{half} = \left(\frac{1}{2}\right)^n.$$

For $n = 3$, $r_{\text{half}} \approx 0.7937$, For $n = 10$, $r_{\text{half}} \approx 0.933$, For $n = 100$, $r_{\text{half}} \approx 0.993$.

Problem 10 (6 marks)

Linear, Affine and Convex Combination We require you to plot the linear, affine, and convex combinations of vectors in $\mathbf{R}^3$. Write a Python script to generate three random vectors v_1, v_2, v_3, where each coordinate of these vectors is drawn from a standard normal distribution. We require you to submit both the code and the resulting plots.

Problem 11 (10 marks)

Properties of the p-Norm

Let $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ and define the p-norm for $1 \leq p < \infty$ as:

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}.$$

(a) **(4 points)** Prove that for any vector $\mathbf{x} \in \mathbb{R}^n$, the p-norm satisfies the following properties:

- $\|\mathbf{x}\|_p = 0$ if and only if $\mathbf{x} = \mathbf{0}$.
- $\|\alpha\mathbf{x}\|_p = |\alpha|\|\mathbf{x}\|_p$ for any scalar $\alpha \in \mathbb{R}$.
- $\|\mathbf{x} + \mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ (triangle inequality).

Use Minkowski's inequality to prove the triangle inequality:

$$\left(\sum_{i=1}^{n} |x_i + y_i|^p \right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} + \left(\sum_{i=1}^{n} |y_i|^p \right)^{\frac{1}{p}}, \forall p \geq 1.$$

(b) **(3 points)** Show that the p-norm is monotonically decreasing with respect to p, i.e., for $1 \leq p \leq q < \infty$, prove that:

$$\|\mathbf{x}\|_p \geq \|\mathbf{x}\|_q.$$

(c) **(3 points)** We want to investigate the behavior of the p-norm as $p \to \infty$. Prove the following:

$$\lim_{p \to \infty} \|\mathbf{x}\|_p = \|\mathbf{x}\|_\infty = \max_i |x_i|.$$

Solution:

(a) Basic Properties of the p-Norm:

We begin by proving the three properties of the p-norm:

- $\|\mathbf{x}\|_p = 0$ if and only if $\mathbf{x} = \mathbf{0}$: If $\mathbf{x} = \mathbf{0}$, then clearly $\|\mathbf{x}\|_p = 0$. Conversely, if $\|\mathbf{x}\|_p = 0$, then the sum $\sum_{i=1}^{n} |x_i|^p = 0$, which implies that each $x_i = 0$, and hence $\mathbf{x} = \mathbf{0}$.
- $\|\alpha\mathbf{x}\|_p = |\alpha|\|\mathbf{x}\|_p$: We compute:

$$\|\alpha\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |\alpha x_i|^p \right)^{\frac{1}{p}} = |\alpha| \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} = |\alpha|\|\mathbf{x}\|_p.$$

- Triangle inequality $\|\mathbf{x} + \mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$: Consider some $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. These vectors will have some components $x_i, y_i \in \mathbb{R}$, for $0 \leq i \leq n$. Then, the sum of $\mathbf{x}$ and $\mathbf{y}$ will have components $x_i + y_i$ for $0 \leq i \leq n$. We use these components to write out the Minkowski inequality for any $p \geq 1$:

$$\left(\sum_{i=1}^{n} |x_i + y_i|^p \right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} + \left(\sum_{i=1}^{n} |y_i|^p \right)^{\frac{1}{p}}.$$

But from, the definition of the p-norm:

$$\left(\sum_{i=1}^{n} |x_i + y_i|^p \right)^{\frac{1}{p}} = \|\mathbf{x} + \mathbf{y}\|_p.$$

The right hand-side of the inequality can be converted similarly. Hence, $\|\mathbf{x} + \mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$.

(b) Monotonicity of the p-Norm:

We want to show that for $1 \leq p \leq q < \infty$, $\|\mathbf{x}\|_p \geq \|\mathbf{x}\|_q$. Let us first define a new vector:

$$\mathbf{y} = \frac{\mathbf{x}}{\|x\|_q},$$

which has components $y_i = x_i / \|\mathbf{x}\|_q$. This implies a unit q-norm for $\mathbf{y}$; $\|\mathbf{y}\|_q = 1$. We can deduce that the magnitude of the individual components of $\mathbf{y}$ are less than 1 i.e:

$$\sum_{i=0}^{n} |y_i|^q = \sum_{i=0}^{n} \frac{|x_i|}{\|\mathbf{x}\|_q} = 1 \implies |y_i| \leq 1.$$

Now, as $p \leq q$, we get $|y_i|^q \leq |y_i|^p$. We can sum these to obtain the following inequality:

$$\sum_{i=0}^{n} |y_i|^q \leq \sum_{i=0}^{n} |y_i|^p$$

But the LHS is simply 1:

$$1 \leq \sum_{i=0}^{n} |y_i|^p \implies 1 \leq \sum_{i=0}^{n} \frac{|x_i|^p}{\|\mathbf{x}\|_q^p}$$

We are at the final step. We now use the fact that term in the denominator of the RHS is a simple scalar and bring it to the left side to obtain:

$$\|\mathbf{x}\|_q^p \leq \sum_{i=0}^{n} |x_i|^p$$

We simply take the pth root of the entire inequality and we've done it:

$$\|\mathbf{x}\|_q \leq \left(\sum_{i=0}^{n} |x_i|^p \right)^{\frac{1}{p}} = \|\mathbf{x}\|_p$$

Hence, shown. The p-norm is monotonically decreasing with respect p.

(c) Limit as $p \to \infty$:

We want to show that as $p \to \infty$, $\|\mathbf{x}\|_p$ converges to the infinity norm $\|\mathbf{x}\|_\infty = \max_i |x_i|$.

The p-norm is defined as:

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}.$$

We can bound this sum from above and below. Without loss of generality, we can exponentiate the equality and proceed from there:

$$\|\mathbf{x}\|_p^p = \left(\sum_{i=1}^{n} |x_i|^p \right).$$

Let $M = \max_i |x_i|$. For the lower bound, we know that the sum on the RHS will always be greater than or equal to the **largest** component of the vector x raised to pth power i.e:

$$M^p \leq \left(\sum_{i=1}^{n} |x_i|^p \right).$$

For the upper bound, we know that $M \geq |x_i|$ and then by exponentiating the inequality and summing over all i:

$$\left(\sum_{i=1}^{n} |x_i|^p \right) \leq \sum_{i=1}^{n} M^p = nM^p.$$

We can put both the lower and upper bound together to obtain:

$$M^p \leq \left(\sum_{i=1}^{n} |x_i|^p \right) \leq nM^p.$$

We take pth root of this inequality:

$$M \leq \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} \leq n^{\frac{1}{p}} M.$$

Now, when we take the limit as $p \to \infty$, the $n^{\frac{1}{p}} \to 1$ which implies that:

$$M \leq \|\mathbf{x}\|_p \leq M.$$

Hence, by the squeeze theorem, we have that:

$$\lim_{p \to \infty} \|\mathbf{x}\|_p = \|\mathbf{x}\|_\infty = M.$$

This makes sense, as the largest term would dominate over all others as $p \to \infty$.

— End of Assignment —

Question	Points	Score
1	6	
2	8	
3	10	
4	8	
5	10	
6	10	
7	10	
8	12	
9	10	
10	6	
11	10	
Total:	100	