## Midterm 1 - DSC 40A, Winter 2024

## Instructions

- This is a 50-minute exam consisting of 5 questions worth a total of 40 points.
- The only allowed resource is the provided reference sheet.
- No calculators.
- Please write neatly and stay within the provided boxes.
- You may fill out the **front page only** until you are instructed to start.

## Statement of Academic Integrity

By submitting your exam, you are attesting to the following statement of academic integrity.

I will act with honesty and integrity during this exam.

| Name:            |  |
|------------------|--|
| PID:             |  |
| Seat you are in: |  |

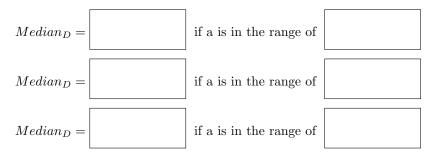
Version - A

1. (10 points) Consider a dataset D with 5 data points  $\{7, 5, 1, 2, a\}$ , where a is a positive real number. Note that a is not necessarily an integer.

a) (2 points) Express the mean of D as a function of a, simplify the expression as much as possible.



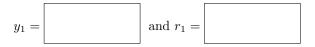
b) (3 points) Depending on the range of a, the median of D could assume one of three possible values. Write out all possible median of D along with the corresponding range of a for each case. Express the ranges using double inequalities, e.g., i.e.  $3 < a \leq 8$ :



c) (5 points) Given that  $Mean_d < Median_D$ , determine the range of a that satisfies this condition. make sure to show your work

| Range of a:      |  |  |  |
|------------------|--|--|--|
|                  |  |  |  |
|                  |  |  |  |
| ~                |  |  |  |
| Supporting Work: |  |  |  |
|                  |  |  |  |
|                  |  |  |  |
|                  |  |  |  |
|                  |  |  |  |
|                  |  |  |  |
|                  |  |  |  |
|                  |  |  |  |
|                  |  |  |  |
|                  |  |  |  |
|                  |  |  |  |
|                  |  |  |  |
|                  |  |  |  |
|                  |  |  |  |
|                  |  |  |  |
|                  |  |  |  |
|                  |  |  |  |
|                  |  |  |  |
|                  |  |  |  |
|                  |  |  |  |

**2.** (4 points) Let  $R_{sq}(h)$  represent the mean squared error of a constant prediction h for a given dataset. For the dataset  $\{3, y_1\}$ , the graph of  $R_{sq}(h)$  has its minimum at the point  $(5, r_1)$ . Find out the value of  $y_1$  and  $r_1$ 



**3.** (10 points) The hyperbolic cosine function is defined as  $cosh(x) = \frac{1}{2}(e^x + e^{-x})$ . In this problem, we aim to prove the convexity of this function using power series expansion.

a) (3 points) Prove that  $f(x) = x^n$  is convex if n is an even integer.

| Proof: |  |  |  |
|--------|--|--|--|
|        |  |  |  |
|        |  |  |  |
|        |  |  |  |
|        |  |  |  |
|        |  |  |  |
|        |  |  |  |
|        |  |  |  |
|        |  |  |  |
|        |  |  |  |

**b)** (2 points) Power series expansion is a powerful tool to analyze complicated functions. In power series expansion, a function can be written as an infinite sum of polynomial functions with certain coefficients. For example, the exponential function can be written as:

$$e^{x} = \sum_{n=0}^{\infty} \frac{x^{n}}{n!} = 1 + x + \frac{x^{2}}{2} + \frac{x^{3}}{6} + \frac{x^{4}}{24} + \dots$$
(1)

where n! denotes the factorial of n, defined as the product of all positive integers up to n, i.e.  $n! = 1 \times 2 \times 3 \times ... \times (n-1) \times n$ . Given the power series expansion of  $e^x$  above, write the power series expansion of  $e^{-x}$  and explicitly specify the first 5 terms, i.e., similar to the format of Equation 1:

 $e^{-x} = \sum_{n=0}^{\infty} =$ 

c) (5 points) Using the conclusions you reached in a) and b), prove that  $cosh(x) = \frac{1}{2}(e^x + e^{-x})$  is convex.

**Proof:** 

4. (10pt) Note that we have two simplified closed form expressions for the estimated slope w in simple linear regression that you have already seen in discussions and lectures:

$$w = \frac{\sum_{i} (x_i - \overline{x}) y_i}{\sum_{i} (x_i - \overline{x})^2} \tag{1}$$

$$w = \frac{\sum_{i} (y_i - \overline{y}) x_i}{\sum_{i} (x_i - \overline{x})^2} \tag{2}$$

where we have dataset  $D = [(x_1, y_1), \ldots, (x_n, y_n)]$ , sample means  $\overline{x} = \frac{1}{n} \sum_i x_i$ ,  $\overline{y} = \frac{1}{n} \sum_i y_i$ . Without further explanation,  $\sum_i$  means  $\sum_{i=1}^n y_i$ .

a) (6pt) Are (1) and (2) equivalent? That is, is the following equality true? Prove or disprove it.

**Proof:** 

$$\sum_{i} (x_i - \overline{x}) y_i = \sum_{i} (y_i - \overline{y}) x_i$$

b) (2pt) True or False: If the dataset shifted right by a constant distance a, that is, we have the new dataset  $D_a = (x_1 + a, y_1), \ldots, (x_n + a, y_n)$ , then will the estimated slope w change or not?

 $\bigcirc$  True  $\bigcirc$  False

c) (2pt) True or False: If the dataset shifted up by a constant distance b, that is, we have the new dataset  $D_b = [(x_1, y_1 + b), \dots, (x_n, y_n + b)]$ , then will the estimated slope w change or not?

 $\bigcirc$  True  $\bigcirc$  False

## **5.** (6 points)

Answer:

Suppose the following information is given for a linear regression:

$$X = \begin{bmatrix} 1 & 2 \\ 1 & -1 \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} a \\ b \end{bmatrix} \qquad \vec{w^*} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Where X is the design matrix,  $\vec{y}$  is the observation vector, and  $\vec{w}^*$  is the optimal parameter vector. Solve for parameter a and b using the normal equation, show your work.

Supporting Work: