DSAA5002 Data Mining & Data Science

(2025 Fall)

Question 1. What is the value range of the following measures?

(a) z-score normalization of a real number

(b) cosine similarity between two real-valued vectors

(c) Manhattan distance between two real-valued vectors

Question 2. For the following group of data: 100, 200, 400, 800, 1500, 3000

(a) Calculate the mean and variance

(b) Normalize all values by z-score normalization

(c) Normalize all values by min-max normalization with min=0 and max=10

(d) Normalize all values by decimal scaling

Question 3. For two vectors x=(2, 3, 1) and y=(4, 2, 8), calculate the following measures between x and y:

(a) cosine similarity

(b) Euclidean distance

(c) Manhattan distance

(d) Supremum distance

Question 4. Consider two vectors $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$**. Suppose** $|x| = \sqrt{\sum_{i=1}^{n} x_i^2} = 1$ and $|y| = \sqrt{\sum_{i=1}^{n} y_i^2} = 1$. **Prove** $d(x, y) = \sqrt{2(1 - \cos(x, y))}$**, where** $d(x, y)$ **is the Euclidean distance between x and y.**