CSCI 4380/6380 – Data Mining Fall 2023

Assignment 2 Due date: September 22th, 11:59pm, EST

---

1. (20 points) It is important to define or select similarity measures in data analysis. However, there is no commonly accepted subjective similarity measure. Results can vary depending on the similarity measures used. Nonetheless, seemingly different similarity measures may be equivalent after some transformation.
Suppose we have the following 2-D data set:

	A_1	A_2
$\boldsymbol{x_1}$	1.5	1.7
$\boldsymbol{x_2}$	2	1.9
$\boldsymbol{x_3}$	1.6	1.8
$\boldsymbol{x_4}$	1.2	1.5
$\boldsymbol{x_5}$	1.5	1.0

(a) Consider the data as 2-D data points. Given a new data point, x = (1.4,1.6) as a query, rank the database points based on similarity with the query using Euclidean distance, Manhattan distance, supremum distance, and cosine similarity.
(b) Normalize the data set to make the norm of each data point equal to 1. Use Euclidean distance on the transformed data to rank the data points.

2. (20 points) What are the value ranges of the following normalization methods?
(a) min-max normalization
(b) z-score normalization
(c) z-score normalization using the mean absolute deviation instead of standard deviation
(d) normalization by decimal scaling

3. (20 points) Recall the Question 1 in Assignment 1, gave the following data (in increasing order) for the attribute age: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70, answer the following:

(e) Use min-max normalization to transform the value 35 for age onto the range [0.0, 1.0].

(f) Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years.

(g) Use normalization by decimal scaling to transform the value 35 for age.

(h) Comment on which method you would prefer to use for the given data, giving reasons as to why.

4. (20 points) Using the data for age and body fat given in Question 2 in Assignment 1, answer the following:

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(a) Normalize the two attributes based on z-score normalization.

(b) Calculate the correlation coefficient (Pearson's product moment coefficient). Are these two attributes positively or negatively correlated? Compute their covariance.

5. (20 points) Programming

(a) Write a program (your own function) to perform principle component analysis (PCA); (10 points)

(b) Given the following data (n (number) x d (dimension)), apply your PCA function to perform dimensionality reduction and reduce the dimension to 2; (5 points)

$$\begin{bmatrix} 5.1 & 3.5 & 1.4 & 0.2 \\ 4.7 & 3.2 & 1.3 & 0.3 \\ 4.9 & 3.6 & 1.2 & 0.2 \\ 6.0 & 3.0 & 5.2 & 2.3 \\ 5.9 & 3.1 & 5.1 & 1.8 \\ 5.8 & 2.9 & 5.3 & 2.2 \end{bmatrix}$$

(c) Obtain the scatter plot of the data in 2-dimensional space. (5 points)