Course: **Data Mining** 23-Jan-2023 **(Fall 2022)**

Resource Person: **Dr. Muhammad Faheem**

ASSIGNMENT–2 (Introduction and Getting to know Data)

Total Points: 20
Submission Due: Saturday Jan 27, 2023
(Google Classroom Course Page)

Instructions: Please Read Carefully!

- **This is an individual assignment. Everyone is expected to complete the given assignment on their own, without seeking any help from any website or any other individual.** There will be strict penalties for any work found copied from any source and the university policy on plagiarism will be strictly enforced.
- Assignment is to be submitted via Google Classroom.
    - You should already have created your account on Google Classroom as per my earlier email. If not, then follow the link in that email to create your account.
    - Submit your assignment on or before due date. **No late submissions will be possible**.
    - The viva of this assignment will be conducted.

Question 1: **[5]**

Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):

(a) Compute the *Euclidean distance* between the two objects.
(b) Compute the *Manhattan distance* between the two objects.
(c) Compute the *supremum distance* between the two objects.

(a) Compute the Euclidean distance between the two objects.

The Euclidean distance is computed using Equation (2.6).

Therefore, we have Euclidean distance 6.7082.

(b) Compute the Manhattan distance between the two objects.

The Manhattan distance is computed using Equation (2.7). Therefore, we have $|22 - 20| + |1 - 0| + |42 - 36| + |10 - 8| = 11$

(c) Compute the supremum distance between the two objects.

The supremum distance is computed using Equation (2.8). Therefore, we have a supremum distance of 6.

Question 2: **[5]**

Explain with examples the following concepts in Data Integration:

(a) Entity Identification Problem
(b) Redundancy and Correlation Analysis for Nominal and Numeric data

The metadata from the different data sources must be integrated in order to match up equivalent real-world entities. This is referred to as the entity identification problem. Refer to 3.3.1 Section.

Derived attributes may be redundant, and inconsistent attribute naming may also lead to redundancies in the resulting data set. Also, duplications at the tuple level may occur and thus need to be detected and resolved. Refer to 3.3.2 section.

Question 3: **[10]**

Consider the data for the attribute marks: 33, 35, 36, 36, 39, 40, 40, 41, 42, 42, 45, 45, 45, 45, 50, 53, 53, 55, 55, 55, 55, 56, 60, 65, 66, 72, 90.

(a) Use smoothing by bin means to smooth these data, using a bin depth of 3.

(b) How might you determine outliers in the data?

(c) Use min-max normalization to transform the value 55 for marks onto the range [0.0, 1.0].

(d) Use z-score normalization to transform the value 35 for marks.

(e) Plot an equal-width histogram of width 10.

(f) Sketch examples of each of the following sampling techniques: SRSWOR, SRSWR, cluster sampling, and stratified sampling. Use samples of size 5 and the strata "Below-Average", "Average" and "Exceptional". Compute an approximate median value for the data.

(a)

The following steps are required to smooth the above data using smoothing by bin means with a bin depth of 3.

Step 1: Sort the data. (This step is not required here as the data are already sorted.)

Step 2: Partition the data into equidepth bins of depth 3.

Step 3: Calculate the arithmetic mean of each bin.

Step 4: Replace each of the values in each bin by the arithmetic mean calculated for the bin.

This method smooths a sorted data value by consulting to its "neighborhood". It performs local smoothing.

(b)

Outliers in the data may be detected by clustering, where similar values are organized into groups, or 'clusters'. Values that fall outside of the set of clusters may be considered outliers. Alternatively, a combination of computer and human inspection can be used where a predetermined data distribution is implemented to allow the computer to identify possible outliers. These possible outliers can then be verified by human inspection with much less effort than would be required to verify the entire initial data set.

(c) Refer to equation 3.8

(d) Refer to equation 3.9

(e) See Figure 3.4.

(f) See Figure 3.5.