Course: **Data Mining [A]** 27-Jan-2023 **(Fall 2022)**

Resource Person: **Dr. Muhammad Faheem** **QUIZ-2 (Data Preprocessing)**

Total Points: 20 **Time Allowed: 20 Minutes**

Solution

Question 1: [5]

Given two objects represented by the tuples (32, 7, 52, 20) and (30, 6, 46, 18):

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

Solution:

The Euclidean distance is computed using the equation (2.16). Therefore, we have

$$=\sqrt{(32-30)^2 + (7-6)^2 + (52-46)^2 + (20-18)^2}$$
$$= 6.7082$$

The Manhattan distance is computed using the equation 2.17. Therefore, we have

$$= |32 - 30| + |7 - 6| + |52 - 46| + |20 - 18|$$
$$= 11$$

Question 2: [10]

	Pakistan	**Nepal**	**Total**
Western Culture	500 (180)	400 (720)	900
Eastern Culture	100 (420)	2000 (1680)	2100
Total	600	2400	3000

Compute Correlation Analysis of nominal attributes using chi-square. Assume that the chi-square value needed to reject the hypothesis at the 0.001 significance level is 10.828.

The expected frequency in above table is plotted (see numbers with red font color) using equation 3.2.
The x^2 is computed using equation 3.1. Therefore, we have

$$= 586.89 + 243.81 + 142.22 + 60.95 = 1033.87.$$

Since the value is above 10.828, we can reject the hypothesis that both nominal attributes are independent and can conclude that the two attributes are strongly correlated.

Question 3: [5]

Differentiate and explain the difference between min-max normalization and z-score normalization.

The min-max normalization and z-score normalization are methods for data transformation by normalization. These both methods are used for decimal scaling.

min-max Normalization: This method performs linear transformation using minimum and maximum values of an attribute. This techniques map the values in the defined range e.g. [0.0-1.0]. This normalization is performed using equation 3.8.

z-score Normalization: This method normalize the values of an attribute using mean and standard deviation. This method is useful when actual minimum and maximum of attribute are unknown or when there are outliers that dominate the min-max normalization. This normalization is performed using equation 3.9.