# DEPARTMENT OF COMPUTER SCIENCE

## Faculty of Mathematics and Computer Sciences

## South Asian University (A University established by SAARC Nations)

*Akbar Bhawan, Chanakyapuri, New Delhi - 110021, India*

**Semester-End Examination, 3rd July 2020**
**M. Sc. (CS) II Semester**
**Data Mining**

***The scanned copy of the answer sheet must be sent to me latest by 3pm***

Time: 3 Hours Max. Marks: 40

**1.** What do you mean by *support* and *confidence* of an Association Rule? Explain with the help of suitable examples. Consider the *min-support* value as 2 and following set of transactions, and generate top-five association rules with respect to confidence value. Show all the steps. **[10]**

T1: ACDEF, T2: ABEF, T3: ABCD, T4: BCDEF, T5: ABCDEF

**2.** Considering the following benchmark dataset and system predictions, generate confusion matrix and determine the values of Precision, Error-rate, Specificity, and Sensitivity. **[10]**

Benchmark:

| Instance ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Class label | C1 | C2 | C1 | C2 | C1 | C1 | C2 | C1 | C1 | C2 | C1 | C1 | C2 | C2 | C1 |

System prediction:

| Instance ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Class label | C1 | C1 | C1 | C2 | C2 | C1 | C1 | C2 | C2 | C1 | C1 | C2 | C1 | C2 | C2 |

**3.** Consider the following three-dimensional data points and assume that k = 2, and initial centroid of the clusters as C1 = {x1} and C2 = {x6}. **[10]**

| | A1 | A2 | A2 |
|---|---|---|---|
| $x_1$ | 0 | 1 | 1 |
| $x_2$ | 1 | 1 | 0 |
| $x_3$ | 1 | 0 | 1 |
| $x_4$ | 0 | 1 | 1 |
| $x_5$ | 0 | 0 | 0 |
| $x_6$ | 0 | 1 | 0 |

Answer the following questions:

(i) Generate 2-dimesnional distance (dissimilarity) matrix.

(ii) Apply the K-means algorithm for two iterations, assuming the Manhattan distance or the L1-norm

(iii) Find the radius and dimeter of the final clusters obtained in step (ii)

(iv) Find the complete-link, single-link, and mean distance between the final clusters obtained in step (ii)

**4.** Considering the following dataset in which “play” is the class attribute. For the feature set <Rainy, Hot, High, True>, what will be class label using the Naïve Bayes classifier? **[10]**

| Outlook | Temperature | Humidity | Windy | Play |
|---|---|---|---|---|
| Sunny | Hot | High | False | ***No*** |
| Sunny | Hot | High | True | ***No*** |
| Overcast | Hot | High | False | ***Yes*** |
| Rainy | Mild | High | False | ***Yes*** |
| Rainy | Cool | Normal | False | ***Yes*** |
| Rainy | Cool | Normal | True | ***No*** |
| Overcast | Cool | Normal | True | ***Yes*** |
| Sunny | Mild | High | False | ***No*** |
| Sunny | Cool | Normal | False | ***Yes*** |
| Rainy | Mild | Normal | False | ***Yes*** |
| Sunny | Mild | Normal | True | ***Yes*** |
| Overcast | Mild | High | True | ***Yes*** |
| Overcast | Hot | Normal | False | ***Yes*** |
| Rainy | Mild | High | True | ***No*** |

***