**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI** - **HYDERABAD CAMPUS**

**Data Mining (CS F415) - MID SEMESTER EXAMINATION**

**SECOND SEMESTER 2019 – 2020**

**Date: 06.03.2020 Maximum Marks: 50 Maximum Duration: 90min**

**Name:**

**ID:**

**Please read through the instructions carefully:**

1. Answer all questions.
2. This is a closed book exam.
3. There are total 10 pages and 10 questions in the exam. All questions are compulsory.
4. Answers to all questions must be written in this booklet itself in the designated space provided after each question. Answers written in any other place will not be evaluated.
5. The rough workings need to be done in the supplementary sheet provided and should be attached along with the main booklet. The rough workings are not going to be evaluated.
6. Provide short and clear answers in legible handwriting.

Q1. Can Silhouette coeeficiect be used both as a measure of cluster validity as well as cluster stability. Explain how? 4 marks

Consider the dataset given below where attributes Age and Income have very different scales.



Compute the distance between x­1­ and x2. Next compute the distance between the two samples after range and z normalization. What is the main observation? 4 marks

Q2. Consider the sepal length and sepal width attributes for the Iris dataset, as plotted in the figure shown below.



There are n = 150 points in the d = 2 dimensional attribute space. The sample mean vector is given as



And sample covariance matrix is given as



Provide a comment on how these statistical measures can be used to infer relationship between the two attributes. 2 marks

Q3a). Why are concept hierarchies useful in data mining? 2 marks

3b) Does use of concept hierarchies’ lead to any side effects in mining of multilevel association rules? 2 marks

Q4. Consider the following table



Say the minimum support is 40% and confidence support is 60%. Provide a comment on the interestingness of the association rule A=>B. 4 marks

Q5. The number of candidate sequences generated in sequential pattern mining is larger than number of candidate itemsets. Mention the reasons for the additional number of candidates? 4 marks

Q6. Consider the data set shown in the table below. Suppose the following discretization strategies are applied to the continuous attributes of the data set.



D1:  Partition the range of each continuous attribute into 3 equal-sized bins.

D2:  Partition the range of each continuous attribute into 3 bins; where each bin contains an equal number of transactions.

For each strategy, construct a binarized version of the data set. 6 marks

Q7. Consider the data sequence shown below for a given object. Count the number of occurrences for the sequence ⟨{p}{q}{r}⟩ according to atleast 3 different support counting methods.

****

Assume that ws = 0, mingap = 0, maxgap = 3, maxspan = 5. 6 marks

Q8. Given the following points: 2, 3, 4, 10, 11, 12, 20, 25, 30. Assume K = 2, and that the initial means are randomly picked up as μ1 = 2 and μ2 = 4. Show the clusters obtained using K-means algorithm after each iteration, and show the new means for the next iteration. 6 marks

Q9. For the transactions given in the table below, label each node with the following letter(s) in the lattice structure shown below as follows:

• M if the node is a maximal frequent itemset,

• C if it is a closed frequent itemset,

• N if it is frequent but neither maximal nor closed, and

• I if it is infrequent.

Assume that the support threshold is equal to 30%. 8 marks



Q10. Consider the following transaction database.



Generate the FP tree structure and then generate the conditional pattern bases and conditional FP trees for items p and m. 6 marks

End of Exam