BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

I SEMESTER 2022-2023 CS F415 – DATA MINING, Mid Semester (closed book)

Time: 90 Minutes **Weightage: 30%** **Date: 4th November 2022**

1. (a) How do outliers effect the mean and median?

   (b) What are the differences between feature selection and feature reduction?

   (c) Find which pair of documents are the most similar? Show the dissimilarity/similarity calculations.
   D1: new home sales top forecast; D2: home sales rise in July; D3: increase in home sales in July

   (d) Given data points with different types of attributes (nominal, ordinal, interval scale, symmetric and a symmetric binary. Write single formula to calculate dissimilarity between them. [10]

2. (a) What is a stable classifier? Comment on ID3 which is an unpruned Decision Tree classifier.
   (b) Why do we use gain ratio over gain in decision Tree?
   (c) Name four limitations of Decision Tree classifier. [8]

3. (a) Compare the applicability of k-means and k-medoid when data has (a) different size clusters and (b) different density clusters
   (b) Compare k-means and k-medoid with respect to space and time complexities. [8]

4. Consider glass dataset which has 5 classes {building windows, vehicle windows, tableware, containers, headlamps} with the distribution {146, 17, 9, 13, 29}.
   (a) How would you make the training and testing setup for the problem?
   (b) Which classifier would you use and why?
   (c) How would you calculate final precision of the classifier? [8]

5. Consider the following data

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
A	3.5	2	9.1	2	1.5	7	2.1	8
B	4	4	4.5	6	7	6.5	2.5	4
Class	H	H	L	H	H	H	L	L

   (a) Compare a nonlinear split $AB\text{-}B^2<=0$ with axis parallel binary splits on *A* and *B* to construct a Decision tree using information gain.
   (b) Use Naïve Bayes classifier to classify the tuple {5,5}. Discretize attribute *A* as [1.5, 4.3), [4.3, 6.8), [6.8, 9.1] and *B* with equi-width binning for 3 bins.
   (c) Apply bisecting means clustering on the data given in the above table (do not use class) using Manhattan distance for 3 iterations.

   [9+8+9]