BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

Second Semester 2022-23

COMPREHENSIVE EXAMINATION

PART A (CLOSED BOOK)

CS F415 Data Mining

Date : 06-05-2022 (FN) **Max. Duration : 90 minutes** **Max.Marks : 40 (20%)**

Instructions: Write your answers in the answer sheet provided. DO NOT write anything on this question paper.

I. Multiple Choice Questions. Choose the most appropriate answer. *Write the correct choice (A, C, etc.) in the answer sheet.* [1*10 = 10]

1. A Covariance of zero between two variables necessarily implies independence.
   (A) True (B) False

2. The range of values for silhouette coefficient when
   Silhouette coefficient = (b-a)/max(a,b)
   Where, a= average intra-cluster distance i.e the average distance between each point within a cluster.
   b= average inter-cluster distance i.e the average distance between all clusters.
   (A) 0 to 1
   (B) -1 to 1
   (C) 0 to n
   (D) 1 to n

3. When Gini Index has its minimum value, it denotes:
   (A) Highest level of purity
   (B) Random assignment of classes
   (C) Lowest level of purity
   (D) Non-homogenous distribution

4. Contingency Tables are only applicable to Asymmetric Binary Variables.
   (A) True (B) False

5. Statistical Test for Spatial randomness in cluster analysis
   (A) Grubb's test
   (B) Hopkin's test
   (C) t-test
   (D) discordancy test

6. Nodes resulting in a homogenous class distribution are preferred for splitting in a decision tree
   (A) True (B) False

7. What will be the error rate of an ensemble classifier composed of 50 identical base classifier, each having an error rate $\varepsilon = 0.35$.
   (A) 0.06
   (B) 0.35
   (C) 0.06667
   (D) 0.50

8. Which Agglomerative Clustering technique is often used as a robust method of initializing a K-Means clustering?
(A) MIN
(B) MAX
(C) Ward's method
(D) Group Average

9. In general the time complexity of agglomerative clustering algorithms is
(A) $O(N^3)$
(B) $O(N^2)$
(C) $O(N^2 \log N)$
(D) $O(N \log N)$

10. Which of the following measure is equivalent to Geometric mean between the confidence of association rules?
(A) Interest Factor
(B) Correlation coefficient
(C) IS Measure
(D) Kappa

II. Write your answer in one line only. [1*7 = 7]

1. The Supremum or Chebyshev distance can be defined as …….
2. The "coverage" of a rule is defined as ……..
3. Analysis of multidimensional contingency tables is more complicated because ……
4. ……is required to avoid Simpson's Paradox.
5. The total number of possible rules extracted from a data set that contains 'd' items is ……..
6. When determining multiple Outliers at once, the techniques may suffer with a problem …….
7. **Density-based Outlier:** The outlier score of an object is …….

III. Expand the following Terms: [1*7=7]

A. SSB	
B. LOF	
C. PS Measure	
D. CPCC	
E. OLAP	
F. PCA	
G. OOB error	

IV. Write short answers for the following: [2*8=16]

1. Name the five major tasks in Data Preprocessing.
2. Explain Occam's Razor with a suitable example.
3. Explain and thus differentiate between Discretization and Concept Hierarchies.
4. What are internal and external indices with respect to cluster validation? Give one example of each.
5. What are symmetric and asymmetric measure with respect to Association rules validation? Give two examples of each.
6. Assume I run DBSCAN with MinPoints=5 and epsilon=0.2 for a dataset and I obtain 5 clusters and 7% of the objects in the dataset are classified as outliers. Now I run DBSCAN with MinPoints=10 and epsilon=0.2. How do expect the clustering results to change?
7. Assume we have an association rule
`if Drink_Tea and Drink_Coffee then Smoke`
that has a lift of 2. What does say about the relationship between smoking, and drinking coffee, and drinking tea? Moreover, the support of the above association rule is 1%. What does this mean?
8. While calculating conditional probabilities in Naïve Bayes classifier, one of the probability value is zero. Briefly describe the approach used in this regard to tackle the zero probability value.

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

Second Semester 2022-23

COMPREHENSIVE EXAMINATION

PART B (Open BOOK)

CS F415 Data Mining

Date : 06-05-2022 (FN) **Max.Duration : 90 minutes** **Max.Marks : 40 (20%)**

Instructions: Answer all parts of a question together. Answer each question on a fresh page.

1. We will generally be more interested in association rules with high confidence. However, often we will not be interested in association rules that have a confidence of 100%. Why? Then specifically explain why association rules with 99% confidence may be interesting (i.e., what might they indicate)? **[4]**

2. You are provided a small dataset of real estate transactions wherein a property's no. of rooms, its size (in square feet) and its selling price in lakhs (INR) are provided. Use K-nearest neighbors, with (a) k=2 and (b) k=4, to estimate the selling price of a property having 4 rooms and size 2000 (in square feet).
   Note that both the features have different scales and you must employ necessary preprocessing before giving your prediction results. **[5+5=10]**

#Rooms	Size	Price (lakhs INR)
1	871	106
2	852	79
2	797	81
3	909	109
3	1229	116
3	1207	154
5	2508	245
6	2475	225
6	2555	300
8	3612	382

3. The population S originally consists of 30 instances, 16 of the positive class and 14 of the negative class. We now perform a split which results in *two distributions*: Split one (S1) has 17 instances, with 4 of the positive class and 13 of the negative class, while split two (S2) as 13 instances, with 12 of the positive class and 1 of the negative class. Determine the Information Gain in this scenario. Show all steps and computations related to entropy. **[4]**

4. We want to cluster five documents into two clusters having inter-document distances shown below. Assume that at a certain step, documents C and E are selected as medoids. Which two documents will be selected as medoids in the next iteration using the PAM algorithm? You must explain the steps followed in arriving at your answer. **[10]**

Document	**A**	**B**	**C**	**D**	**E**
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

5. Consider the 5 transactions given below. If minimum support is 20% and minimum confidence is 80%, determine the frequent itemsets and association rules using the *FP-Tree* algorithm. **[12]**

Transaction	**Items**
T1	Bread (500 grams), Jam (1 bottle), Butter (100 grams), Ketchup (1 bottle)
T2	Bread (500 grams), Butter (100 grams), Egg (1 dozen)
T3	Bread (500 grams), Milk (2 litres), Butter (200 grams)
T4	Egg (2 dozens), Bread (500 grams), Ketchup (1 bottle), Milk (3 litres)
T5	Egg (1 dozen), Milk (1 litre), Jam (1 bottle)