Question Paper

Exam Date & Time: 29-Nov-2022 (09:00 AM - 12:00 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

FIFTH SEMESTER B.TECH END SEMESTER EXAMINATIONS, NOV 2022

DATA MINING AND PREDICTIVE ANAYSIS [ICT 3171]

Marks: 50 **Duration: 180 mins.**

A

Answer all the questions. Section Duration: 180 mins

Instructions to Candidates: Answer ALL questions Missing data may be suitably assumed

1) Find all frequent itemsets using Apriori and FP-growth, respectively using the dataset given in table Q.1A. Compare the efficiency of the two mining processes. (5)

A)

Table Q.1A

TID	*items_bought*
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

B) Using Naïve bayes on the dataset given in table Q.1B , classify the tuple 'X' to one of the classes of the dataset. (3)

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

Table Q.1B

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

C) A database contains 80 records on a particular topic. A search was conducted on that topic and 60 records were retrieved. 45 records were relevant out of the 60 records retrieved, (2)

Calculate the precision and recall scores for the search. Also, show that accuracy is a function of sensitivity and specificity

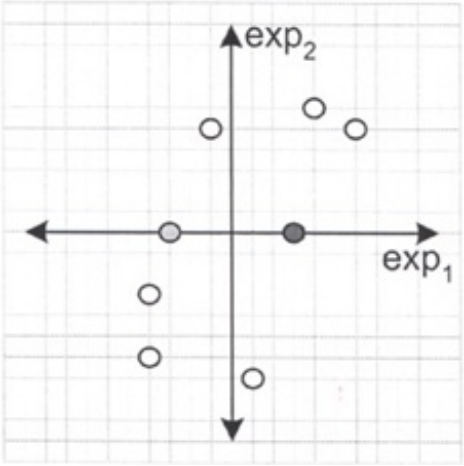
2) Consider the gene expression values shown in table Q.4A from 2 micro array experiments for 8 genes. (5)

A)

Table Q.4A

	exp_1	exp_2
$gene_1$	-4	-3
$gene_2$	6	5
$gene_3$	1	-7
$gene_4$	-4	-6
$gene_5$	4	6
$gene_6$	-1	5
$gene_7$	-3	0
$gene_8$	3	0

Negative expression values in the table mean that the gene is downregulated(i.e.,expressed less) in the experimented cell, and positive values mean that the gene is upregulated(i.e.,over expressed). A biologist is trying to find out whether these 8genes can be separated into two groups based on their behaviour in the experimented conditions. In order to visualize the relationships, she sketched a 2-dimensional plot of the genes shown below.



Use k-means clustering to cluster these 8 genes into 2 clusters. Use gene7 as the initial cluster center for cluster I and use gene 8 as the initial cluster center for cluster 2.Indicate which datapoint belongs to which cluster and also give the coordinates of the centroids at each iteration. Iterate until convergence. When computing distances to centroids, use the squared distance(no need to take the square-root)

B) Illustrate the working of Bagging and Boosting algorithms and give the comparison of the same. (3)

C) Suppose that you are to allocate a number of automatic teller machines (ATMs) in a given region so as to satisfy a number of constraints. Households or places of work may be clustered so that typically one ATM is assigned per cluster. The clustering, however, may be constrained by two factors: (1) obstacle objects (i.e., there are bridges, rivers, and highways that can affect ATM accessibility), and additional user-specified constraints, such as each ATM should serve at least 10,000 households. How can a clustering algorithm such as k-means be modified for quality clustering under both constraints? (2)

3) Consider the data set shown in Table (5)

A) **Table Q.3A**

Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

i. Compute the support for itemsets {e}, {b, d}, and {b, d, e} by treating each transaction ID as a market basket.

ii. Use the results in part (i) to compute the confidence for the association rules {b, d} $\longrightarrow$ {e}

and {e}→ {b, d}. Is confidence a symmetric measure?

iii. Repeat part (i) by treating each customer ID as a market basket.

iv. Use the results in part (iii) to compute the confidence for the association rules {b, d} –→ {e} and {e} –→ {b, d}.

v. Suppose s1 and c1 are the support and confidence values of an association rule r when treating each transaction ID as a market basket. Also, let s2 and c2 be the support and confidence values of r when treating each customer ID as a market basket. Discuss whether there are any relationships between s1 and s2 or c1 and c2.

B) Describe each of the following clustering algorithms in terms of the following (3)

criteria: (i) shapes of clusters that can be determined; (ii) input parameters that must be specified; and (iii) limitations.

(a) *k*-means

(b) DBSCAN

(c) CLARA

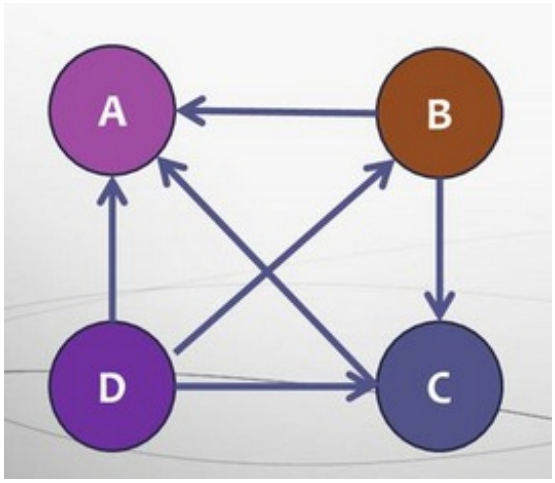
C) Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning. (**Example:** Age in years. **Answer:** Discrete, quantitative, ratio) (2)

i. Brightness as measured by a light meter

ii. Bronze, Silver, and Gold medals as awarded at the Olympics.

4) Categorize various web mining techniques and brief about Web structure mining. Find page rank each of the nodes given in below figure (5)

A)



B) You are given a data set with 100 records and are asked to cluster the data. You use K-means to cluster the data, but for all values of K, 1 ≤ K ≤ 100, the K-means algorithm returns only one non-empty cluster. You then apply an incremental version of K-means, but obtain exactly the same result. How is this possible? How would single link or DBSCAN handle such data? (3)

C) We generally will 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? Justify why association rules with 99% confidence may be interesting? (2)

5) Answer the following with respect to Support Vector Machines. (5)

A) i) List and brief about different types of SVMs

ii) Compare and contrast various linear kernel functions.

B) List the various assumptions of Linear Regression. Also, justify how these assumptions ensure that the Linear Regression gets the best possible result from the given dataset. (3)

C) Illustrate with suitable example, the problems the decision tree suffers from. (2)

-----End-----