The University of Hong Kong
FACULTY OF ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE

COMP7103 DATA MINING
Date: 21st December 2021 Time: 9:30am – 11:30am

Only approved calculators as announced by the Examinations Secretary can be used in this examination. It is the candidates' responsibility to ensure that their calculator operates satisfactorily, and candidates must record the name and type of the calculator used on the front page of the examination script.

Answer ALL 4 questions.

1. Short questions (25%)

   (a) (5%) Briefly explain the term "curse of dimensionality". Describe three techniques that are commonly used to handle high-dimensional data in data analysis.

   (b) (5%) Explain "min-max normalization" and "z-score normalization". Discuss why normalization is often done in processing multi-dimensional data.

   (c) (5%) Briefly explain two techniques in SVM for handling data that is not linearly separable.

   (d) (5%) Define "maximal frequent itemset" and "closed frequent itemset". Prove that any maximal frequent itemset must be closed frequent.

   (e) (5%) Explain the "partitioning problem" and the "fragmented rule problem" in quantitative association rule mining. Also discuss how those problems can be handled.

2. (28%) Consider the training dataset D_{train} of labeled objects shown in Table 1. In the table, each row shows an object id, a nominal attribute value (a_1), a numerical attribute value (a_2), and a class label.

Object	a_1	a_2	Class
X_1	X	6	C_1
X_2	X	15	C_3
X_3	X	23	C_1
X_4	X	24	C_2
X_5	Y	8	C_3
X_6	Y	12	C_2
X_7	Y	17	C_2
X_8	Y	21	C_3
X_9	Y	22	C_2
X_{10}	Y	26	C_3

Table 1: Dataset D_{train}

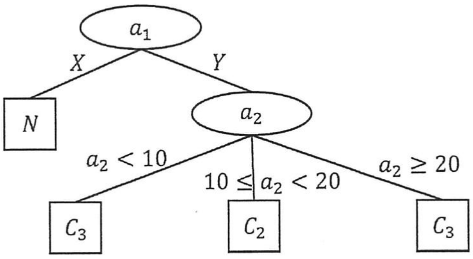

Figure 1: Decision tree

(a) (3%) Suppose a sample of three objects: $S = \{X_1, X_5, X_8\}$ is drawn from D_{train}. For each sampling strategy below, discuss whether it is possible that S is obtained based on such strategy.

i. Sampling without replacement.

ii. Sampling with replacement.

iii. Stratified sampling where D_{train} is partitioned according to the class label.

(b) (3%) Compute the GINI index of D_{train}.

(c) (3%) Figure 1 shows a decision tree built with D_{train}. Determine the class label of leaf node N. Briefly explain your answer.

(d) (3%) Calculate the gain ratio of the test using a_2 in the decision tree.

(e) (3%) Describe the advantage of using gain ratio as the impurity measure compared with using entropy.

(f) (5%) Use the test dataset D_{test} shown in Table 2 to evaluate the decision tree shown in Figure 1. Construct the confusion matrix and then calculate the precision, recall and F measure for the target class C_2.

Object	a_1	a_2	Class
X_{11}	Y	9	C_2
X_{12}	Y	14	C_2
X_{13}	Y	18	C_2
X_{14}	Y	25	C_1
X_{15}	Y	27	C_3

Table 2: Dataset D_{test}

(g) (3%) Explain the *independent assumption* in naïve Bayesian classification. What is the rationale of making that assumption?

(h) (5%) Consider applying naïve Bayesian classification to classify the object X_{12} in Table 2. The numerical attribute a_2 is discretized into three intervals with split points 10 and 20. (This discretization is the same as that shown in Figure 1.) Use D_{train} (Table 1) as the training data, apply naïve Bayesian classification to classify object X_{12}.

3. (17%) Figure 2 shows an FP-tree that is constructed from a dataset of 8 transactions. In this question, assume a minimum support count of 3 as the support requirement.

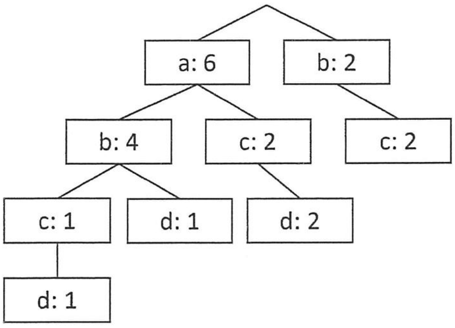

Figure 2: An FP-tree

(a) (5%) Construct the conditional FP-tree that is conditional on item d.

(b) (5%) Explain whether you can reconstruct the original set of 8 transactions given the FP-tree. Use the tree shown in Figure 2 to illustrate your explaination if appropriate.

(c) (3%) Is itemset $\{a, c\}$ a closed itemset? Explain your answer.

(d) (4%) Find the *lift* of the association rule "$\{a, c\} \rightarrow \{d\}$". Comment on the rule's quality.

4. (30%) Table 3 shows the locations (x-y coordinates) of 6 data objects, as well as the corresponding proximity matrix calculated using Euclidean distance.

	Location		Distance to					
Object	x	y	O_1	O_2	O_3	O_4	O_5	O_6
O_1	-30	0	0	6	32	47	47	60
O_2	-24	0	6	0	26	42	42	54
O_3	2	0	32	26	0	30	30	28
O_4	6	30	47	42	30	0	60	38
O_5	6	-30	47	42	30	60	0	38
O_6	30	0	60	54	28	38	38	0

Table 3: Dataset and proximity matrix for cluster analysis

(a) (10%) Based on the given proximity matrix, perform clustering on the objects using simple-link as inter-cluster similarity. Show your results by drawing a dendrogram. Mark the dendrogram with the proximity values of the merge points.

(b) (10%) Suppose two clusters, $Cluster_1 = \{O_1\}$ and $Cluster_2 = \{O_2, O_3\}$ are formed as a clustering result. Explain how you would evaluate the quality of the clustering by suggesting an inter-cluster measure and an intra-cluster measure, and how these measures are applied to the clustering ($Cluster_1$, $Cluster_2$).

(c) (10%) In the process of finding 3 clusters using k-means, 3 centroids, $C_1 = (-30, 0)$, $C_2 = (-11, 0)$, and $C_3 = (14, 0)$, are found at the end of an iteration of the algorithm. Table 4 shows the distances between the centroids and the 6 data objects.

i. Show the clustering result given by the algorithm at the end of that iteration.

ii. Show the clustering result given by the algorithm at the end of the next iteration. State any assumptions you made.

	Location		Distance to					
Centroid	x	y	O_1	O_2	O_3	O_4	O_5	O_6
C_1	-30	0	0	6	32	47	47	60
C_2	-11	0	19	13	13	34	34	41
C_3	14	0	44	38	12	31	31	16

Table 4: Distances between 3 centroids and the 6 data objects

END OF PAPER