DATA MINING & KNOWLEDGE DISCOVERY
(CSEN 3132)

Time Allotted : 3 hrs **Full Marks : 70**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 5 (five) from Group B to E, taking at least one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A
(Multiple Choice Type Questions)

1. Choose the correct alternative for the following: **10 × 1 = 10**

(i) Suppose your model is overfitting. Which of the following is NOT a valid way to try and reduce the overfitting?
(a) Increase the amount of training data
(b) Improve the optimisation algorithm being used for error minimisation
(c) Decrease the model complexity
(d) Reduce the noise in the training data.

(ii) If a transaction set consist of 1000 transactions, 300 transactions contain bread, 350 transactions contain butter, 150 transactions contain both bread and butter. Then the confidence of buying bread with butter (butter ⍰ bread) is
(a) 30% (b) 42.86% (c) 50% (d) 65%

(iii) The binary entropy is maximum when p(a) =
(a) 1.00 (b) 0.25 (c) 0.50 (d) 0

(iv) Classification rules are extracted from
(a) Root node (b) Decision tree
(c) Siblings (d) Branches.

(v) Removing duplicate records is a process called
(a) Recovery (b) Data cleaning
(c) Data dredging (d) Data pruning

(vi) Which one of the following clustering techniques needs the merging approach?
(a) Partitioned (b) Naïve Bayes
(c) Hierarchical (d) Both (a) and (c)

(vii) To detect fraudulent usage of credit cards, which one of the following data mining tasks should be used?
(a) Outlier analysis (b) Prediction
(c) Association analysis (d) Feature selection.

(viii) In SVM, when the C parameter is set to infinite, which of the following holds true?
(a) The optimal hyperplane, if exists, will be the one that completely separates the data
(b) The soft-margin classifier will separate the data
(c) Both (a) and (b)
(d) None of the above.

(ix) The Support Vector Machines are less effective when
(a) The data is linearly separable
(b) The data is clean and ready to use
(c) The data is noisy and contains overlapping points
(d) None of the above.

(x) Suppose that X_1, ..., X_m are categorical input attributes and Y is categorical output attribute. Suppose we plan to learn a decision tree without pruning, using the standard algorithm. The maximum depth of the decision tree must be
(a) less than m+1 (b) greater than m+1
(c) both (a) and (b) can be true (d) Neither (a) nor (b) is true.

Group – B

2. It is required to discriminate between Martians (M) and Humans (H) based on the following characteristics: Green ∈ {N, Y}, Legs ∈ {2, 3}, Height ∈ {S, T}, Smelly ∈ {N, Y}. The available training data is as follows:

Sl No	Species	Green	Legs	Height	Smelly
1	M	N	3	S	Y
2	M	Y	2	T	N
3	M	Y	3	T	N
4	M	N	2	S	Y
5	M	Y	3	T	N
6	H	N	2	T	Y
7	H	N	2	S	N
8	H	N	2	T	N
9	H	Y	2	S	N
10	H	N	2	T	Y

Learn a decision tree by building a decision tree by selecting the best attribute that yields maximum Information Gain (IG). **Build the decision tree only for the first two levels (the root level and the next level).** [(CO1, CO2, CO4) (Remember/LOCQ) (Apply/IOCQ)

12

3. (a) Consider the following data set for a binary class problem.

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

(i) Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

(ii) Calculate the gain in the Gini index when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

(b) Describe one method of post pruning by drawing a decision tree.
[CO3 Understand/LOCQ][CO5 Evaluate/HOCQ]

(4 + 4) + 4 = 12

Group – C

4. (a) Consider the data set in the following Table.

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	-
3	0	1	1	-
4	0	1	1	-
5	0	0	1	+
6	1	0	1	+
7	1	0	1	-
8	1	0	1	-
9	1	1	1	+
10	1	0	1	+

Estimate the conditional probabilities for P(A|+), P(B|+),
P(C|+), P(A|−), P(B|−), and P(C|−).

(a) Use the conditional probabilities, estimated in part (a) to predict the class label for a test sample (A = 0, B = 1, C = 0) using the Naive Bayes approach.

(b) In which cases Naïve Bayes classifier may provide undecisive results (hint: posterior probability may become zero)? How can we use Naïve Bayes classifier in those circumstances?

[(CO3,CO3,CO4) (Understand and apply/LOCQ, Analyze/HOCQ)]

6 + 3 + 3 = 12

5. Consider a classification problem in R^2, with two classes, in which the training classes are given by:

x_1	x_2	Class
-2	-2	A
-2	-1	A
1	2	A
2	1	A
-2	2	B
0	2	B
0	-1	B
2	-1	B

Consider the non-linear mapping from input space to a two-dimensional feature space, given by

(x_1, x_2) (x_1^2, x_1x_2)

(a) Plot the training patterns in input space and label them according to the class they belong to. State whether the patterns from the two classes are linearly separable in this space.

(b) Plot separately the training patterns in feature space and label them according to the class they belong to. State whether the patterns from the two classes are linearly separable in this space.

(c) Find the widest-margin classifier in feature space. More specifically, find the equations of the classification boundary and of the two margin boundaries. Plot these three boundaries on the same graph that was used in step b). Also indicate the support vectors in the feature space. Note that the boundaries and support vectors are easy to find by inspection.

(d) Calculate the margin once you identify the margin boundaries.

[(CO1,CO3,CO5)(Remember/LOCQ, Understand/IOCQ, Apply/HOCQ)]

2 + 2 + 6 + 2 = 12

Group – D

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

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	$\{a, b, e\}$

(b) Use the results in part (a) to compute the confidence for the association rules *{b, d}* → *{e}* and *{e}* → *{b, d}*. Is confidence a symmetric measure?

(c) Repeat part (a) by treating each customer ID as a market basket. Each item

should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise.)

(d) Use the results in part (c) to compute the confidence for the association rules *{b, d} → {e}* and *{e} → {b, d}.*
[(CO1,CO2,CO3)(Remember/LOCQ, Understand/IOCQ)]

3 + 3 + 3 + 3 = 12

7. (a) Construct the FP-growth tree for the data provided in Question 6.

(b) Consider the market basket transactions shown in the following Table.

Transaction Id	Items Bought
1.	{Bread, Butter, Jam}
2.	{Bread, Cheese, Coke}
3.	{Bread, Jam, Biscuit}
4.	{Biscuit, Coke, Cheese}
5.	{Biscuit, Butter, Jam}
6.	{Bread, Cheese, Jam, Coke}
7.	{Jam, Coke, Cheese}
8.	{Butter, Jam}
9.	{Coke, Cheese, Bread, Jam}
10.	{Butter, Biscuit}

(i) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?

(ii) What is the maximum size (k) of frequent k-itemsets in this data (assuming minsup > 0)?
[(CO1,CO3,CO6)(Remember/LOCQ, Understand/IOCQ)]

9 + 3 = 12

Group – E

8. (a) Cluster the following eight points (with (x, y) representing locations) into three clusters using k-means clustering technique:
A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)
Consider A1(2, 10), A4(5, 8) and A7(1, 2) as the initial cluster centres and Euclidean distance measure.

(b) Define core point, border point and noise point in DBSCAN with a diagram.
[(CO2,CO3,CO6)(Remember/LOCQ, Understand/IOCQ)]

9 + 3 = 12

9. For the one-dimensional data set {7,10,20,28,35,54}, perform hierarchical clustering and plot the dendrograms to visualize it using:
(a) MAX (complete linkage) distance and
(b) MIN (single linkage).
Note: Draw the dendrogram with merging distance and clearly show the merge sequence. [(CO1,CO3,CO6)(Remember/LOCQ, Apply/LOCQ)]

6 + 6 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	65.62%	25.00%	9.38%

Course Outcome (CO):

After the completion of the course students will be able to

CSEN3132.1 Learn and understand basic knowledge of data mining and related models.

CSEN3132.2. Understand and describe data mining algorithms.

CSEN3132.3. Understand and apply Data mining algorithms.

CSEN3132.4. Suggest appropriate solutions to data mining problems.

CSEN3132.5. Analyse data mining algorithms and techniques.

CSEN3132.6. Perform experiments in Data mining and knowledge discovery using real-world data.

*LOCQ: Lower Order Cognitive Question; IOCQ: Intermediate Order Cognitive Question; HOCQ: Higher Order Cognitive Question

Department & Section	Submission Link
CSE - A + B + C	https://classroom.google.com/c/NDAxMTk3MzY5MDA5/a/NDY0MTU1MDUzMzc5/details