DATA MINING & KNOWLEDGE DISCOVERY
(MCAP 2251)

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) Support Vector machines is _____
(a) classification learning (b) unsupervised machine learning
(c) supervised machine learning (d) reinforcement learning.

(ii) Naive Bayes classifiers are a collection of classification algorithms based on____
(a) Naive Therom (b) Bayes Therom
(c) Both (a) and (b) (d) none of these.

(iii) PCA is a
(a) linear method (b) non linear method
(c) continuous method (d) repeated method.

(iv) In Hyper plane, f(x) = sign(w*x + b), where 'w' is a?
(a) Constant (b) Vector
(c) Distance (d) None of the above.

(v) Which statement about outliers is true?
(a) Outliers should be part of the training dataset but should not be present in the test data
(b) Outliers should be part of the test dataset but should not be present in the training data
(c) The nature of the problem determines how outliers are used
(d) More than one of (a), (b) or (c) is true.

(vi) What does Apriori algorithm do?
(a) It mines all frequent patterns through pruning rules with lesser support
(b) It mines all frequent patterns through pruning rules with higher support
(c) Both (a) and (b)
(d) None of the above.

(vii) Which of the following is finally produced by Hierarchical Clustering?
(a) Final estimate of cluster centroids
(b) Tree showing how close things are to each other
(c) Assignment of each point to clusters
(d) All of the Mentioned.

(viii) When you find noise in data, which of the following option would you consider in k-NN?
(a) I will increase the value of k
(b) I will decrease the value of k
(c) Noise cannot be dependent on value of k
(d) None of these.

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

(x) You are given data about seismic activity in Japan, and you want to predict a magnitude of the next earthquake, this is an example of
(a) Dimensionality Reduction (b) Supervised Learning
(c) Unsupervised Learning (d) Reinforcement Learning.

Group- B

2. (a) Define the following terms in view of Market Basket Analysis with some example.
(i) Support.
(ii) Confidence.
(iii) Lift. [(CO4)(Remember/LOCQ)]

(b) Consider the following Table:

Transaction	T1	T2	T3	T4	T5	T6	T7	T8	T9
List of Item_Id	I1,I2,I5	I2,I4	I2,I3	I1,I2,I4	I1,I3	I2,I3	I1,I3	I1,I2,I3,I5	I1,I2,I3

Find all frequent item sets by using Apriori Algorithm where the minimum support Count = 2.
Write down the Apriori Algorithm. Discuss the main drawbacks of Apriori Algorithm. [(CO7)(Evaluate/HOCQ)]

6 + 6 = 12

3. (a) Let us consider the following transactions with different Item Sets

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

Applying the FP Growth algorithm find out an association rule between different Items where minimum count value = 3. [(CO7)(Evaluate/HOCQ)]

(b) Given the following data, using PCA reduce the dimension from 2 to 1. [(CO3)(Evaluate/HOCQ)]

Feature	Example-1	Example-2	Example-3	Example-4
x	4	8	13	7
y	11	4	5	14

6 + 6 = 12

Group – C

4. (a) Why naïve Bayesian classification is called naïve? Briefly outline the major ideas of naïve Bayesian classification. [(CO4)(Understand/LOCQ)]

(b) Use Naïve Bayes' classifier to predict whether a person defined by the tuple (age = youth, income = medium, student = yes, credit_rating = fair) buys a computer or not. The training data is as follows: [(CO2)(Analyze/IOCQ)]

RID	Age	Income	Student	Credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle-aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle-aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle-aged	medium	no	excellent	yes
13	middle-aged	high	yes	fair	Yes
14	senior	medium	no	excellent	no

(2 + 2) + 8 = 12

5. (a) Define gain in the Gini index. [(CO2) (Remember/LOCQ)]

(b) Consider the following data set for a binary class problem. [(CO5)(Analyze/IOCQ)]

Sl No	Color	Size	Act	Age	Inflated
1	Yellow	Small	Stretch	Child	T
2	Yellow	Small	Stretch	Child	T
3	Yellow	Small	Stretch	Child	T
4	Yellow	Small	Stretch	Child	T
5	Yellow	Small	Stretch	Adult	T
6	Yellow	Small	Stretch	Child	F
7	Purple	Large	Dip	Adult	F
8	Purple	Large	Dip	Child	F
9	Purple	Small	Stretch	Adult	T
10	Purple	Small	Stretch	Child	F
11	Purple	Small	Dip	Adult	T
12	Purple	Small	Dip	Child	T
13	Purple	Large	Stretch	Adult	F
14	Purple	Large	Stretch	Child	F
15	Purple	Large	Dip	Adult	F
16	Purple	Large	Dip	Child	T

Calculate the information gain when splitting on different attributes. Which attribute would the decision tree induction algorithm choose?

3 + 9 = 12

Group – D

6. Explain the following terms (any four) **(4 × 3) = 12**
   (i) Support Vectors.
   (ii) Hyper planes.
   (iii) Marginal Distances
   (iv) Linear separable.
   (v) Nonlinear separable. [(CO5) (Remember/LOCQ)]

7. (a) What is the difference between logistic regression and SVM? [(CO5) (Remember/LOCQ)]
   (b) Explain how kernel function is used in non-linear support vector machines. Also justify the statement that "One can use infinite-dimensional spaces with the kernel trick" in the perspective of non-linear SVM classification. [(CO5) (Analyze/IOCQ)]

6 + 6 = 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. [(CO6)(Analyze/IOCQ)]
   (b) Define core point, border point and noise point in DBSCAN with a diagram. [(CO7)(Analyze/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:
   - MAX (complete linkage) distance and
   - MIN (single linkage).

   Note: Draw the dendrograms with merging distance and clearly show the merge sequence. [(CO6)(Analyze/IOCQ)]

(6 + 6) = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	32.29	48.95	18.75

Course Outcome (CO):

After the completion of the course students will be able to
1. Describe basic concept of data mining and related models.
2. Store and use data for online processing
3. Pre process the data for mining applications
4. Apply the association rules for mining the data
5. Design and deploy appropriate classification techniques
6. Cluster the high dimensional data for better data organization
7. Implement the data mining algorithms for real-world data.

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