DATA MINING
(CSEN 3105)

Time Allotted : 2½ hrs **Full Marks : 60**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 4 (four) from Group B to E, taking one from each group.

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

Group – A

1. Answer any twelve: **12 × 1 = 12**

Choose the correct alternative for the following

(i) Two documents are said to be very close to each other when the Cosine Similarity measure between their term frequency vectors is
(a) Close to 0 (b) Exactly 0 (c) Close to 1 (d) Exactly 1

(ii) Which of the following can be appropriately represented as an ordinal attribute?
(a) Gender of a student (b) Ratings in a survey
(c) Hair colour (d) Probability of rain.

(iii) Overfitting occurs when a model gives
(a) Low accuracy for both training data and new data (b) High accuracy for new data but not for training data
(c) High accuracy for training data but not for new data (d) High accuracy for both training data and new data.

(iv) If pi is the probability of the ith symbol in a language, then the mathematical expression of Entropy is given by
(a) $-\sum p_i * \log_2(p_i)$ (b) $\sum p_i * \log_2(p_i)$ (c) $-\log_2(p_i)$ (d) $\log_2 p_i$

(v) Out of 10000 transactions, data show that 7000 transactions included bread, 5000 transactions included milk and 4000 transactions included both bread and milk. Which of the following statements is true?
(a) Bread and milk are uncorrelated
(b) Bread and milk are positively correlated
(c) Bread and milk are negatively correlated
(d) Insufficient data to comment about correlation between bread and milk.

(vi) The total number of possible rules, extracted from a dataset containing 4 distinct items is
(a) 20 (b) 30 (c) 40 (d) 50.

(vii) DBSCAN uses k-nearest neighbour distance to find
(a) eps (b) minpts (c) Core points (d) Border points

(viii) The Frequent itemsets obtained in FP-Growth
(a) Can be greater than that by Apriori
(b) Must be same as that by Apriori
(c) Can be smaller than that by Apriori
(d) Can be greater than, equal to or smaller than that by Apriori.

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

(x) k-means clustering does not depend on
(a) Selection of distance metric (b) Selection of k
(c) Initial guess as to cluster centroids (d) Dimension of data points.

Fill in the blanks with the correct word

(xi) In a binary classification problem, the probability of one class is 0.65. Its entropy is _____.

(xii) When True Positive value is 437 and False Positive value is 63, the precision is _____.

(xiii) If there are 3 points (2,5), (3,2) and (4,5) in a cluster, the cluster centre would be ____.

(xiv) A separating hyperplane is represented as $\mathbf{W} \cdot \mathbf{X} + b = 0$, the formula for maximal margin is _____.

(xv) A dataset contains r distinct items. The total number of possible rules, extracted from the dataset is 12. What is the value of r? _____.

Group - B

2. (a) Consider the following 2-D dataset

	A1	A2
X1	1.5	1.7
X2	2.2	1.5
X3	1.8	1.6
X4	1.1	1.4
X5	1.9	1.1

Use Euclidean distance to rank the above data points with respect to a new data point X = (1.4, 1.2). *[(CSEN3105.1)(Apply/IOCQ)]*

(b) Can we say Cosine similarity measure is a metric? Justify your answer. *[(CSEN3105.2)(Understand/LOCQ)]*

(c) The following table shows the number of instances of 8 different words in 2 documents. Find out the cosine similarity measure of the two documents. *[(CSEN3105.2)(Apply/IOCQ)]*

Document	Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8
Doc1	3	0	5	0	0	4	0	1
Doc2	4	1	3	0	6	0	0	4

(d) What is Pearson's Correlation Coefficient? What is its use? *[(CSEN3105.2)(Understand/LOCQ)]*

3 + 3 + 3 + 3 = 12

3. (a) Extract a rule-based system from the training sample given below.

Instance	Classification	A1	A2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

[(CSEN3105.2)(Apply/IOCQ)]

(b) For each of the rules generated, compute the accuracy and coverage of the rules. Which rule(s) appear to be better? *[(CSEN3105.2)(Apply/IOCQ)]*

7 + [(2 × 2) + 1] = 12

Group - C

4. (a) Compare Information Gain, Gain Ratio and Gini Index as attribute selection measure to create decision tree. *[(CSEN3105.3)(Understand/LOCQ)]*

(b) Explain with an example what you understand by Gini Index. *[(CSEN3105.1)(Remember/LOCQ)]*

(c) Find out, using gain in Gini Index, what will be the root node in the decision tree for classifying whether an unknown person is Male or Female. The training data is provided in the following table. *[(CSEN3105.3, CSEN3105.6)(Evaluate/HOCQ)]*

Sl No	Over 170 cm (Yes / No)	Eye (Blue or Brown)	Hair (Short or Long)	Gender (Male / Female)
1	No	Blue	Short	Male
2	Yes	Brown	Long	Female
3	No	Blue	Long	Female
4	No	Blue	Long	Female
5	Yes	Brown	Short	Male
6	No	Blue	Long	Female
7	Yes	Brown	Short	Female
8	Yes	Blue	Long	Male

3 + 3 + 6 = 12

5. (a) What is the main assumption made while solving the problem of classification using Naïve Bayes Classifier? *[(CSEN3105.3)(Understand/LOCQ)]*

(b) "In Naïve Bayes Classification technique, prior probability of each class is required to be computed, whereas not the prior probability of the tuple to be classified" – Explain. *[(CSEN3105.3)(Understand/LOCQ)]*

(c) Given the following table, classify the tuple X, who has a long hair, whose height is between 155 and 165 cm, who is in job but cannot cook. Solve the problem using Naïve Bayesian Classifier. *[(CSEN3105.3, CSEN3105.6)(Apply/IOCQ)]*

Sl. No.	Hair	Height	In job	Can cook	Gender
1	Long	Less than 155 cm	No	No	Female
2	Long	Less than 155 cm	No	Yes	Female
3	Medium	Less than 155 cm	No	No	Male
4	Short	155 –165 cm	No	No	Male
5	Short	Greater than 165 cm	Yes	No	Male
6	Short	Greater than 165 cm	Yes	Yes	Female

Sl. No.	Hair	Height	In job	Can cook	Gender
7	Medium	Greater than 165 cm	Yes	Yes	Male
8	Long	155 – 165 cm	No	No	Female
9	Long	Greater than 165 cm	Yes	No	Male
10	Short	155 – 165 cm	Yes	No	Male
11	Long	155 – 165 cm	Yes	Yes	Male
12	Medium	155 – 165 cm	No	Yes	Male
13	Medium	Less than 155 cm	Yes	No	Male
14	Short	155 – 165 cm	No	Yes	Female

2 + 2 + 8 = 12

Group - D

6. Food items ordered to a Fast-Food shop in 9 different occasions are given in the following table:

Order Id	Food items
1	Chop, Cutlet, Vada
2	Cutlet, Dosa
3	Cutlet, Idli
4	Chop, Cutlet, Dosa
5	Chop, Idli
6	Cutlet, Idli
7	Chop, Idli
8	Chop, Cutlet, Idli, Vada
9	Chop, Cutlet, Idli

(i) Assuming the value of minimum support count = 2, find out all the frequent itemsets by confined candidate generation algorithm (Apriori Algorithm).

(ii) Assuming the value of minimum confidence threshold = 75%, find out all the strong association rules, generated from any one of the frequent 3-itemsets. *[(CSEN3105.4, CSEN3105.6)(Apply/IOCQ)]*

(8 + 4) = 12

7. A transaction dataset is given in the following. Assume minimum support count as 2.

TID	Items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

(i) Draw the FP-Growth Tree.

(ii) Find out the prefix paths for the suffix ‘e’ and draw the corresponding conditional FP Tree. *[(CSEN3105.4)(Apply/HOCQ]*

7 + (3 + 2) = 12

Group - E

8. Perform K-means clustering (using Euclidean distance as distance function) on 2-dimensional data points, given in the following table. Assume that the initial centroids are P5, P7 and P9. Show the centroids and clusters in first two iterations.

Points	X coordinate	Y coordinate
P1	1	10
P2	10	2
P3	7	3
P4	2	3
P5	9	5
P6	4	11
P7	3	9
P8	3	5
P9	4	3

[(CSEN3105.3, CSEN3105.6)(Apply/IOCQ)]

12

9. (a) Define Core Point, Border Point and Noise Point in the perspective of DBSCAN clustering algorithm. *[(CSEN3105.3)(Remember/LOCQ)]*

(b) Explain why DBSCAN does not work well for data having varying density. *[(CSEN3105.3)(Understand/LOCQ)]*

(c) Perform hierarchical clustering, considering complete link method (MAX distance) on the distance matrix given below, to generate a cover. Show the nested clusters and the dendrogram. *[(CSEN3105.3, CSEN3105.6)(Apply/IOCQ)]*

Points	**P_1**	**P_2**	**P_3**	**P_4**	**P_5**
P_1	0				
P_2	2	0			
P_3	10	5	0		
P_4	8	3	7	0	
P_5	6	9	4	1	0

3 + 2 + 7 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	21.88	59.37	18.75