(vi) K-means clustering suffers from
(a) Bad initialization of centroids.
(b) Bad selection of K.
(c) Selection of only round shaped clusters.
(d) All of the above.

(vii) For representing a hierarchical clustering scheme, a tree-like representation method named ______________ can be used.
(a) Decision tree (b) Dendrogram
(c) Spanning tree (d) FP Growth tree.

(viii) Point of inflection can be found if $\frac{\partial^2 y}{\partial x^2}$ of y=f(x) is
(a) greater than 0, (b) less than 0,
(c) equal to zero, (d) all of the above.

(ix) If there are n unique items in a market basket, a lattice for generating all the possible combinations of items that a buyer can buy can be built in the order of,
(a) n (b) n^2 (c) 3^n (d) 2^n

(x) In K- nearest neighbor the input is translated to __________.
(a) values (b) points in multidimensional space
(c) strings of characters (d) nodes.

Group – B

2. (a) Draw a decision tree to predict whether a student will be accepted in the post-graduate program using the data provided in Table 1.

ID	GATE qualified	Publications	Written Test qualified	Interview performance	Decision
1	Yes	Yes	No	Bad	Reject
2	No	Yes	Yes	Bad	Reject
3	Yes	No	No	Good	Accept
4	No	No	Yes	Bad	Reject
5	No	Yes	No	Bad	Reject
6	Yes	No	Yes	Good	Accept
7	No	Yes	Yes	Good	Accept
8	Yes	Yes	No	Good	Accept

9	A , B , C
10	C , D, E

12

7. (a) Consider a training set that contains 100 positive examples and 400 negative examples. For each of the following candidate rules,
$R1$: $A \rightarrow +$ (covers 4 positive and 1 negative examples),
$R2$: $B \rightarrow +$ (covers 30 positive and 10 negative examples),
$R3$: $C \rightarrow +$ (covers 100 positive and 90 negative examples),

Determine which is the best and worst candidate rule according to:
i) Rule accuracy, ii) FOIL's information gain and iii) Likelihood ratio statistic.

4 × 3 = 12

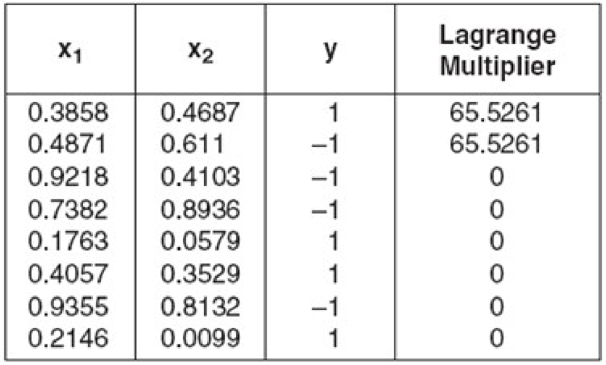
Group - E

8. (a) Define a core point and a noise point in DBSCAN density based clustering.

(b) Describe the DBSCAN algorithm and clearly mention how to choose the parameters of the algorithms (MinPts and Eps).

(2 + 2) + (5 + 3) = 12

9. A linearly separable dataset is given in Table 3. Predict the class of (0.6, 0.8) using a support vector machine classifier.



x_1	x_2	y	Lagrange Multiplier
0.3858	0.4687	1	65.5261
0.4871	0.611	−1	65.5261
0.9218	0.4103	−1	0
0.7382	0.8936	−1	0
0.1763	0.0579	1	0
0.4057	0.3529	1	0
0.9355	0.8132	−1	0
0.2146	0.0099	1	0

Table 3: Linearly separable dataset

12

M.TECH/CSE/2ND SEM /CSEN 5237/2016
2016

DATA MINING AND KNOWLEDGE DISCOVERY
(CSEN 5237)

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 alternatives for the following: **10 × 1=10**

(i) The process of selecting good hypothesis and improving the theory based on this is called _______.
(a) heuristic search (b) hill climbing algorithm
(c) incremental search (d) Apriori algorithm.

(ii) Association rules are always defined on________.
(a) binary attribute. (b) single attribute.
(c) relational database. (d) multidimensional attributes.

(iii) A security dog in an airport detects an unidentified luggage as a potential bomb threat, later it is opened and no bomb was found. This is known as _______________ in evaluating predictive systems.
(a) FALSE positive (b) TRUE positive
(c) FALSE negative (d) TRUE negative.

(iv) The goal in Naïve Bayes classifier is to predict class label using
(a) posterior probability (b) prior probability
(c) likelihood (d) evidence.

(v) Maximum and minimum values for misclassification error (in a binary classification) are
(a) 0.5, -0.5 (b) 1, 0 (c) 1, 0.5 (d) 0.5, 0.

ID	GATE qualified	Publications	Written Test qualified	Interview performance	Decision
9	No	No	Yes	Good	Reject
10	No	Yes	No	Bad	Reject
11	No	No	No	Good	Reject
12	No	Yes	No	Good	Accept
13	Yes	Yes	Yes	Bad	Accept
14	Yes	No	No	Bad	Reject
15	Yes	No	Yes	Bad	Accept

Table 1. Decision data for 15 candidates, who applied for post-graduate program.

12

3. (a) Define Data Mining.

(b) What are different techniques used in data mining to handle missing values or data.

(c) What is difference between DBMS and Data Mining.

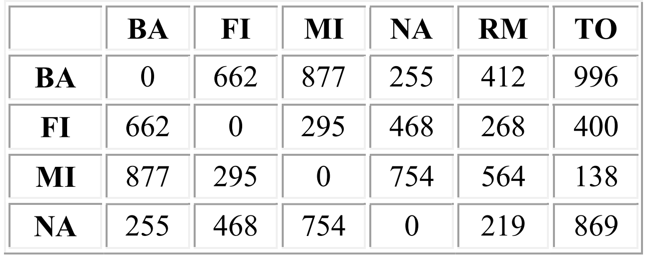
(d) "Data Mining is applicable for any kind of Information repository"-Justify.

(e) What is meant by outlier and how it is detected by Data Mining.

2 + 3 + 2 + 3 + 2 = 12

Group – C

4. (a) Distances between six Italian cities are given by their distances as provided in the distance matrix in Table 2. Use MAX (complete link) agglomerative clustering algorithm to form clusters. Clearly draw the dendrogram and sequence of agglomeration.



	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869

M.TECH/CSE/2ND SEM /CSEN 5237/2016

ID	GATE qualified	Publications	Written Test qualified	Interview performance	Decision
9	No	No	Yes	Good	Reject
10	No	Yes	No	Bad	Reject
11	No	No	No	Good	Reject
12	No	Yes	No	Good	Accept
13	Yes	Yes	Yes	Bad	Accept
14	Yes	No	No	Bad	Reject
15	Yes	No	Yes	Bad	Accept

Table 1. Decision data for 15 candidates, who applied for post-graduate program.

12

3. (a) Define Data Mining.

(b) What are different techniques used in data mining to handle missing values or data.

(c) What is difference between DBMS and Data Mining.

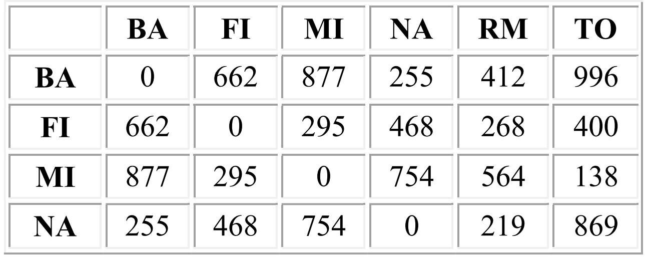
(d) "Data Mining is applicable for any kind of Information repository"-Justify.

(e) What is meant by outlier and how it is detected by Data Mining.

2 + 3 + 2 + 3 + 2 = 12

Group – C

4. (a) Distances between six Italian cities are given by their distances as provided in the distance matrix in Table 2. Use MAX (complete link) agglomerative clustering algorithm to form clusters. Clearly draw the dendrogram and sequence of agglomeration.



	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869

CSEN 5237 3

RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

Table 2: Distances between Italian cities

12

5. (a) Explain the working principle of Naïve Bayesian Classification. In addition, find the Class(X) using Naïve Bayes on the following Dataset, where X= (Age=30; Income=high ; Student=No ; Credit Rating= Fair)

Age	Income	Student	Credit_rating	Buys_laptop
≤ 30	High	No	Fair	No
≤ 30	High	No	Excellent	No
31.40	High	No	Fair	Yes
> 40	Medium	No	Fair	Yes
> 40	Low	Yes	Fair	Yes
> 40	Low	Yes	Excellent	No
31.40	Low	Yes	Excellent	Yes
≤ 30	Medium	No	Fair	No
≤ 30	Low	Yes	Fair	Yes
> 40	Medium	Yes	Fair	Yes
≤ 30	Medium	Yes	Excellent	Yes
31.40	Medium	No	Excellent	Yes
31.40	High	Yes	Fair	Yes
> 40	Medium	No	Excellent	No

12

Group – D

6. Design all Frequent Itemsets using apriori algorithm from the following transaction data given minimum support = 30%. In addition design all association rules from the above Frequent Sets at min Confidence 60%

Transaction Id	Data Items
1	A ,B , C , E
2	B , D , E
3	B , C
4	A , B ,D
5	A , C
6	B , C
7	A , C, E
8	A , B , C , E

RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

Table 2: Distances between Italian cities

12

5. (a) Explain the working principle of Naïve Bayesian Classification. In addition, find the Class(X) using Naïve Bayes on the following Dataset, where X= (Age=30; Income=high ; Student=No ; Credit Rating= Fair)

Age	Income	Student	Credit_rating	Buys_laptop
≤ 30	High	No	Fair	No
≤ 30	High	No	Excellent	No
31.40	High	No	Fair	Yes
> 40	Medium	No	Fair	Yes
> 40	Low	Yes	Fair	Yes
> 40	Low	Yes	Excellent	No
31.40	Low	Yes	Excellent	Yes
≤ 30	Medium	No	Fair	No
≤ 30	Low	Yes	Fair	Yes
> 40	Medium	Yes	Fair	Yes
≤ 30	Medium	Yes	Excellent	Yes
31.40	Medium	No	Excellent	Yes
31.40	High	Yes	Fair	Yes
> 40	Medium	No	Excellent	No

12

Group – D

6. Design all Frequent Itemsets using apriori algorithm from the following transaction data given minimum support = 30%. In addition design all association rules from the above Frequent Sets at min Confidence 60%

Transaction Id	Data Items
1	A ,B , C , E
2	B , D , E
3	B , C
4	A , B ,D
5	A , C
6	B , C
7	A , C, E
8	A , B , C , E