# M.TECH/CSE/2ND SEM/CSEN 5237/2015
# 2015

## 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 x 1=10**

(i) Decision trees are appropriate for the problems where:
- (a) attributes are both numeric and nominal.
- (b) target function takes on a discrete number of values.
- (c) data may have errors.
- (d) all of above.

(ii) Prediction of classes using decision tree can be broken into rules in the following form
- (a) disjunction of conjunctions.
- (b) disjunction of disjunctions.
- (c) conjunction of disjunctions.
- (d) conjunction of conjunctions.

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

(iv) Clustering is considered to be
- (a) Unsupervised learning
- (b) Supervised learning
- (c) Semi-Supervised learning
- (d) Reinforcement learning.

(v) DBSCAN uses k-nearest neighbour distance to find the parameter
- (a) Eps (radius)
- (b) MinPts
- (c) Core points
- (d) Noise points.

(vi) Dijkstra's SSSP (single source shortest path) algorithm may be used in calculation of which of the following network centrality?
(a) Degree (b) Closeness (c) PageRank (d) Katz.

(vii) Support vectors can be identified by
- (a) zero value Lagrangian multipliers
- (b) class labels
- (c) non-zero Lagrangian multipliers
- (d) proximity to (0, 0).

(viii) If T consists of 500000 transactions, 20000 transactions contain bread, 30000 transactions contain jam, 10000 transactions contain both bread and jam. Then the confidence of buying bread with jam is
(a) 33.33% (b) 66.66% (c) 45% (d) 50%.

(ix) _____ training may be used when a clear link between input data sets and target output values does not exist.
(a) Competitive (b) Perception (c) Supervised (d) Unsupervised.

(x) Average of all shortest path distances for all pair of vertices in a social network is usually around 6. This feature is known as
(a) scale free feature (b) randomness feature
(c) small world feature (d) none of the above.

**Group – B**

2.(a) Define support and confidence in mining frequent pattern mining.

(b) A market basket dataset is given by dataset in the following table

| 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} |

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

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

**3+4+5=12**

3. Construct the FP-tree for the transaction database provided in question 2 and find all frequent itemsets using FP-growth approach.

**(5+7)=12**

## Group – C

4. Distances between six Italian cities are given by their distances as provided in the distance matrix in the following table. 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 |
| **RM** | 412 | 268 | 564 | 219 | 0 | 669 |
| **TO** | 996 | 400 | 138 | 869 | 669 | 0 |

**12**

5.(a) The entire output of a factory is produced on three machines. The three machines account for 20%, 30%, and 50% of the output, respectively. The fraction of defective items produced is this: for the first machine, 5%; for the second machine, 3%; for the third machine, 1%. If an item is chosen at random from the total output and is found to be defective, what is the probability that it was produced by the third machine?

(b) Briefly outline the major ideas of naïve Bayesian classification for classifying n-dimensional data points drawn from m classes.

**3+9=12**

## Group – D

6. A linearly separable dataset is given in the following table. 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 |

**12**

7. 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 x 3)=12**

**Group – E**

8.(a) Define modularity and what is its use in social networks?

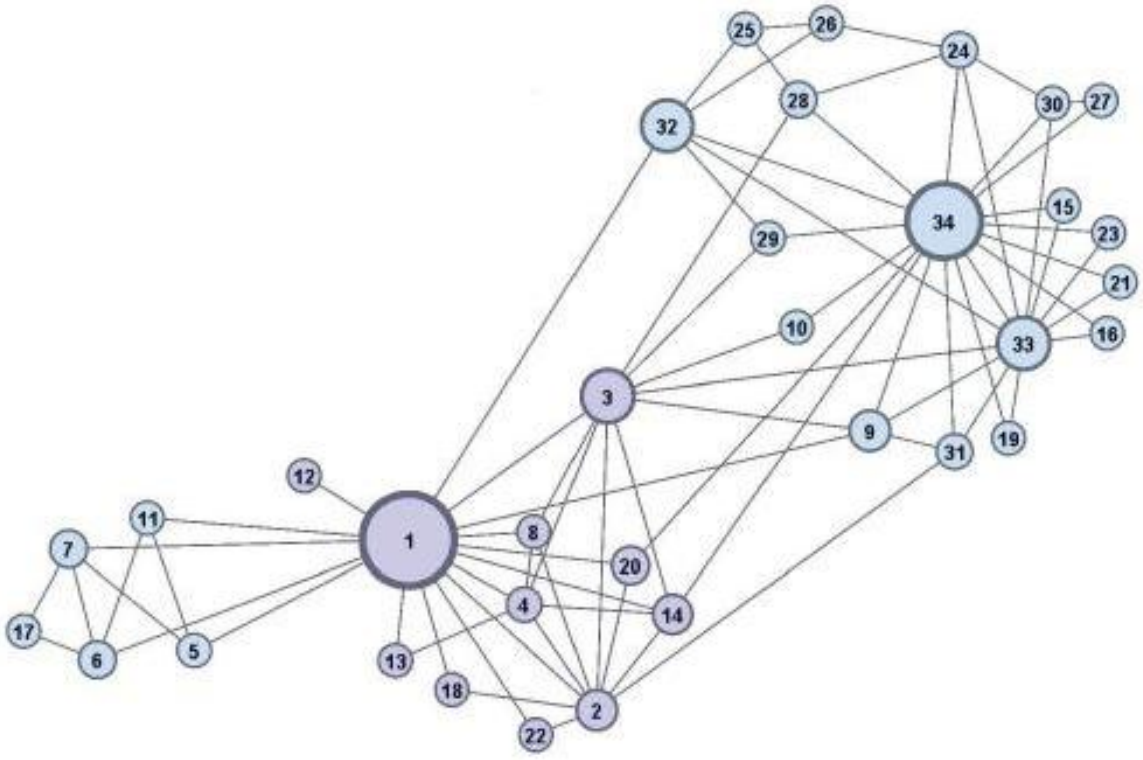
(b) State the modularity maximization problem.

(c) Write a heuristic that solves modularity maximization problem in less than quadratic (in terms of n) running time. Describe it with equations and diagrams wherever needed. You may write any state-of-the-art algorithm or may come up with your own algorithm.

**2+2+8=12**

9.(a) In DBSCAN, how do we choose the thresholds *Eps* and *MinPts*? Explain with illustration.

(b) From the following network




i) Plot the degree distribution of the network.
ii) Find the local clustering coefficient for node 34.
Intuitively suggest the most important broker node from the network.

**4+8=12**