# B.TECH/CSE/7TH SEM/CSEN 4144/2018
# DATA MINING AND KNOWLEDGE DISCOVERY
# (CSEN 4144)

**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) Data Mining is
(a) the actual discovery phase of a knowledge discovery process
(b) the stage of selecting the right data for a KDD process
(c) a subject-oriented integrated time variant non-volatile collection of data in support of management
(d) none of the above.

(ii) Assuming log base 2, the entropy of a binary feature with p(x = 1) = 0.75 is
(a) 0.1875 (b) 0.8113 (c) 0.1887 (d)2.41.

(iii) In a picture, where 7 cats and 10 dogs are present, your dog detection algorithm has detected 9 entities, out of which only 6 are dogs and remaining are cats. What is the precision of your algorithm?
(a) 0.6 (b) 0.66 (c) 6/17 (d) 0.9.

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

(v) 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.

(vi) Boosting is said to be a good classifier because
(a) it creates all ensemble members in parallel, so their diversity can be boosted
(b) it attempts to minimize the margin distribution
(c) it attempts to maximize the margins on the training data
(d) none of the above.





(vii) Closed frequent itemsets category is a
(a) superset of frequent itemsets
(b) subset of maximal frequent itemsets
(c) superset of maximal frequent itemsets
(d) subset of infrequent itemsets.

(viii) DBSCAN cannot be used (with high accuracy) for datasets that are
(a) convex (b) uniform density
(c) non-uniform density (d) none of the above.

(ix) Suppose that $X_1$, ..., $X_m$ are categorical input attributes and Y is categorical output attribute. Suppose we plan to make a decision tree learn 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) either (a) or (b) can be true (d) none of (a) and (b) are true.

(x) After SVM learning, each Lagrange multiplier $\alpha_i$ takes either zero or non-zero value. What does it indicate in each situation?
(a) A zero $\alpha_i$ indicates that the datapoint i has become a support vector datapoint, on the margin.
(b) A zero $\alpha_i$ indicates that the learning process has identified support for vector i.
(c) A non-zero $\alpha_i$ indicates the datapoint i is a support vector, meaning it touches the margin boundary.
(d) A non-zero $\alpha_i$ indicates that the learning has not yet converged to a global minimum.

## Group – B

2. (a) Define Information gain and gain in the Gini index.

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

| Sl No | A | B | Gender |
|---|---|---|---|
| 1 | T | F | C1 |
| 2 | T | T | C1 |
| 3 | T | T | C1 |
| 4 | T | F | C2 |
| 5 | T | T | C1 |
| 6 | F | F | C2 |
| 7 | F | F | C2 |
| 8 | F | F | C2 |
| 9 | T | T | C2 |
| 10 | T | F | C2 |

(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?

**2 + 10 = 12**

3. (d) 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?

(i) Data space and Feature space

(ii) Overfitting

(iii) Pre and po

(iv) Precision a

**3 × 4 = 12**

| 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 | $\{a, b, e\}$ |

A market basket dataset

4. (a) Define Bayes' classification rule from classification er

(b) Consider the fol e Naïve Bayes' classifier to predict the decision for a weekend having weather = rainy, parents = yes, financial condition = rich.

**3 × 4 = 12**

| Weekend | Weather | Parents | Financial Condition | Decision |
|---|---|---|---|---|
| W1 | Sunny | Yes | Rich | Cinema |
| W2 | Sunny | No | Rich | Play Tennis |
| W3 | Windy | Yes | Rich | Cinema |
| W4 | Rainy | Yes | Poor | Cinema |
| W5 | Rainy | No | Poor | Cinema |
| W6 | Rainy | Yes | Poor | Cinema |
| W7 | Windy | No | Poor | Cinema |
| W8 | Windy | No | Rich | Play Tennis |
| W9 | Windy | Yes | Rich | Cinema |
| W10 | Sunny | No | Rich | Play Tennis |

**6 + 6=12**

5. Sometimes data is just nonlinearly separable or data has errors and one wants to ignore them to obtain a better solution. In fact, this is achieved by relaxing the margin, in other words, using a soft margin.

Derive the Lagrangian for the optimization problem as defined by linear SVM – soft margin classification.

**5 + 7=12**

## Group – D

6. Prove that the total number of possible rules extracted from a market basket data set that contains d unique items is $R = 3^d - 2^{d+1} + 1$.

**2 + 10 = 12**

9. (a) Perform K-means clustering on all the points in the following table, where K=2. Randomly select the initial seeds and perform the algorithm for two iterations.

Explain FP-growth tree and its use briefly.

**9 + 3 =12**

| Points | X co-ordinate | Y co-ordinate |
|---|---|---|
| p1 | 1 | 9 |
| p2 | 2 | 10 |
| p3 | 7 | 4 |
| p4 | 10 | 3 |
| p5 | 5 | 5 |
| p6 | 6 | 11 |
| p7 | 3 | 4 |
| p8 | 4 | 9 |
| p9 | 8 | 1 |
| p10 | 3 | 12 |
| p11 | 7 | 6 |
| p12 | 11 | 2 |

7. Consider the market basket dataset shown in table on next page.

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

(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.)

(b) Describe the DBSCAN Algorithm.

**6 + 6 = 12**